Proceedings of OMAE04
23rd International Conference on Offshore Mechanics and Arctic Engineering
June 20-25, 2004, Vancouver, British Columbia, Canada

OMAE2004-51161

THE INFLUENCE OF HINDCAST MODELING UNCERTAINTY ON THE PREDICTION OF HIGH RETURN PERIOD WAVE CONDITIONS

Daniel C. Brooker
School of Oil and Gas Engineering
University of Western Australia
Nedlands, W.A. 6907
E-mail: daniel.brooker@uwa.edu.au

Geoffrey K. Cole
School of Oil and Gas Engineering
University of Western Australia
Nedlands, W.A. 6907
E-mail: geoff.cole@uwa.edu.au

Jason D. McConochie
Woodside Energy Ltd
1 Adelaide Terrace
Perth, W.A. 6000
E-mail: jason.mcconochie@woodside.com.au

Keywords: Extreme Value Analysis, Design Wave, Hindcasting, Uncertainty, Data Analysis, Monte-Carlo, Bootstrap, Probability

ABSTRACT

Extreme value analysis for the prediction of long return period met-ocean conditions is often based upon hindcast studies of wind and wave conditions. The random errors associated with hindcast modeling are not usually incorporated when fitting an extreme value distribution to hindcast data. In this paper, a modified probability distribution function is derived so that modeling uncertainties can be explicitly included in extreme value analysis. Maximum likelihood estimation is then used to incorporate hindcast uncertainty into return value estimates and confidence intervals.

The method presented here is compared against simulation techniques for accounting for hindcast errors. The influence of random errors within modeled datasets on predicted 100 year return wave estimates is discussed.

INTRODUCTION

The standard approach to estimation of high return period wave heights is to fit an extreme value distribution (e.g. Weibull, Gumbel, Generalised Pareto) to existing wave data and predict the wave height for a given return period using the fitted distribution. When hindcast data is used instead of actual measured data, the same procedure for distribution fitting and return value prediction is usually used.

Wave hindcasting involves some degree of error due to simplifications/assumptions in the analytical model used and measurement errors in the input data. Whilst comparison against measured data is often used for calibration of a hindcast model, the effect of hindcast modeling errors on return value estimates is not always considered. An important point to consider is that a hindcast model can be calibrated to give unbiased estimates of wave height, but still result in biased return values due to random variation. Given that hindcasting uncertainties are known and can be quantified, there is a question as to whether their neglect during extreme value analysis is appropriate.

In this paper, a method by which modeling uncertainty can be explicitly included in extreme value analysis is presented and compared with results from Monte-Carlo and bootstrap studies. This new method is based upon the use of a modified probability distribution function which incorporates both environmental and hindcast modeling uncertainty. Maximum likelihood estimation of this modified distribution is used to incorporate hindcast uncertainty into return value estimates and confidence intervals.

The method presented here is used to check the robustness of 100 year return wave estimates for some typical datasets subject to a known hindcast error.

BACKGROUND

Hindcasting of met-ocean data using meteorological charts is commonly used as an alternative for direct measurement of wind and wave values during storms. Such hindcasting studies typically show some degree of bias and random variation when compared against measured values. Bias is usually eliminated by calibrating hindcast model results using any available measurements in the area (e.g. via regression methods). The random variation however remains; typically a mean variation between measurements and model predictions of around 1.0m is observed for significant wave height values [1,2]. A scatter index of 10-15% is representative of modern hindcasts, where the scatter index (*S.I.*) is defined as:

$$S.I. = \frac{\sigma_m}{\mu_{meas}}$$

where σ_m is the standard deviation of the difference between measured and modeled values and μ_{meas} is the mean of the measured wave heights used for comparison.

Direct measurement of ocean data (e.g. via wave buoys) also involves random errors, particularly the measurement of wave heights. These errors are, however, generally much smaller than those associated with hindcasting and will not be considered directly in this paper.

An example plot comparing measured to modeled significant wave heights is shown in Figure 1. This data has been obtained on Australia's North West Shelf; each point corresponding to the maximum wave height observed/predicted during passage of a tropical cyclone. Values have been normalized by the predicted 100 year return period significant wave height, H_{100}. The values shown in Figure 1 are those after calibration of the hindcast model via linear regression. i.e. the individual modeled values are corrected using:

$$H_i^{model,corr} = (H_i^{model} - b)/a$$

where the constants a and b have been determined via linear regression of hindcast predictions against measurements[1]:

$$H_i^{model} = aH_i^{meas} + b$$

After correction, the mean error is equal to zero, hence any systematic bias in the hindcast values is eliminated. i.e.

$$\bar{\varepsilon} = \sum_{i=1}^{n} (H_i^{meas} - H_i^{model})/n = 0$$

[1] It would usually be appropriate to set b=0 to ensure that the hindcast model can be sensibly extrapolated beyond the highest available measurement. This has been done in this case.

This type of calibration scheme is most appropriate when the available measurement dataset is inadequate for distribution fitting and hindcasting is being used to significantly extend the measurement dataset.

For this dataset, the distribution of the residuals $\varepsilon_i = (H_i^{meas} - H_i^{model,corr})$ can be well approximated by the normal (Gaussian) distribution and is taken to be independent of wave height (i.e. an absolute error). By inspection of Figure 1, the assumption of independence appears a reasonable for this set of data, with the random variation associated with the larger storms similar to that seen for the smaller ones. Similar comparisons using the same hindcast model in different locations in Australian waters display low correlation between ε_i and wave height, as does the published data in ref. [1] and [2]. It should be noted however that the assumption of independence of random error may not be true for all hindcast models in current use; it may sometimes be more appropriate to describe the error as a function of wave height.

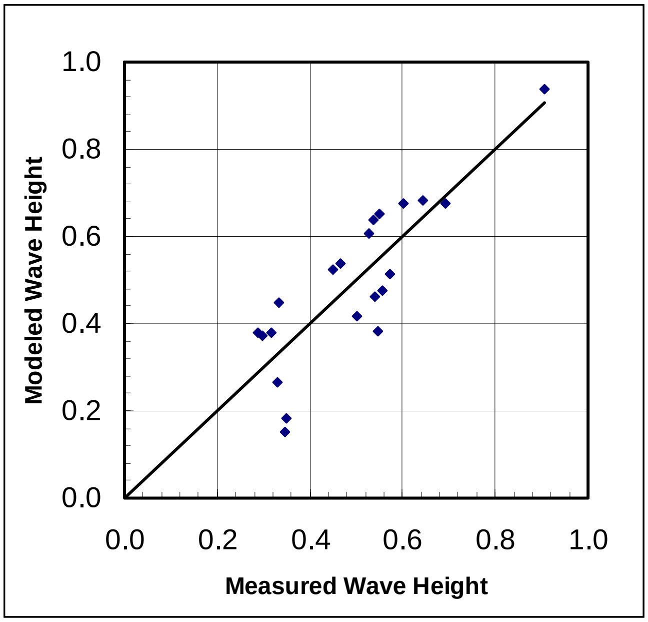

Figure 1 - Comparison of Measured Significant Wave Heights (Normalized by H_{100}) Against Corrected Hindcast Predictions - Australian North West Shelf

Taking independent, normally distributed errors, we can express the hindcast uncertainty via the pdf:

$$f_E(\varepsilon) = \frac{1}{\sqrt{2\pi\sigma_m^2}} e^{-\frac{\varepsilon^2}{2\sigma_m^2}}$$

where σ_m is the standard deviation of ε_i.

Calibration of a hindcast model to eliminate bias does not necessarily result in unbiased extreme value predictions. The addition of any random error to a dataset results in a more highly skewed empirical distribution provided the error is symmetric (Figure 2). Hence, the effect of using a sample of hindcast data is to increase the skewness of the sample and so over-predict return values. This paper is primarily concerned with methods for estimating the degree of over-estimation that is likely given that the magnitude of the hindcast error can be reasonably assessed using the σ_m calculated during calibration.

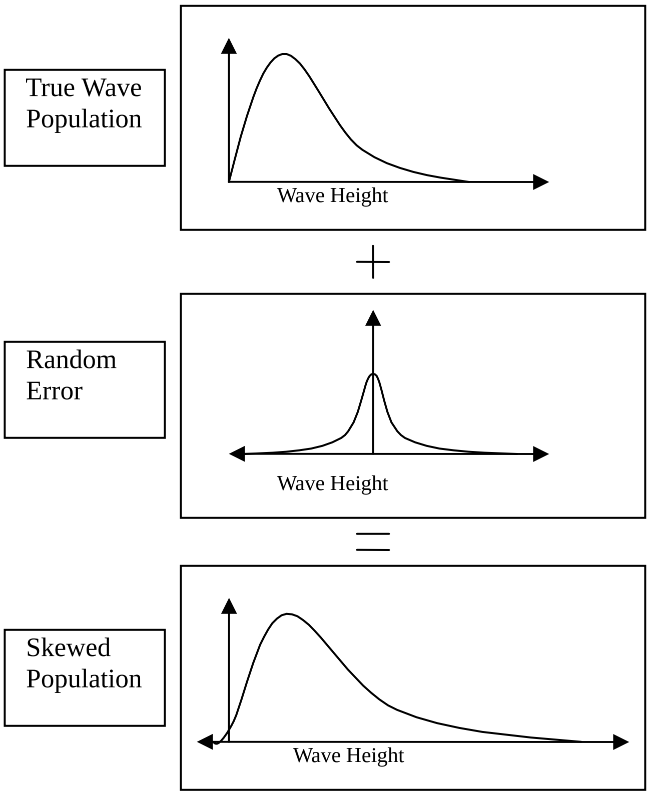

Figure 2 - Effect of Hindcast Uncertainty on Sample Distributions

NUMERICAL SIMULATION METHODS FOR ACCOUNTING FOR MODELING ERROR

Two well established simulation based methods for dealing with a problem of this type are Monte-Carlo methods and bootstrapping.

The Monte-Carlo Approach

An example of the Monte-Carlo approach to account for hindcast error is a numerical simulation of the following steps:

1) An extreme value distribution is fitted to a calibrated hindcast dataset of size *N*.
2) A large number of samples (e.g. 1000) of size *N* are randomly drawn from this fitted distribution.
3) A randomly determined error is added to each sample point, taken from the distribution of hindcast error values (i.e. drawn from a normal distribution of zero mean and standard deviation σ_m).
4) For each sample, a new extreme value fit is conducted, and return period estimates are made.
5) The mean of the resultant return values is then found

An estimate of the amount of bias due to hindcasting errors is obtained by taking the difference between this value and the return value estimate obtained from the original data.

A broadly similar approach for assessing uncertainties is described in [3]. Used in this way, the monte-carlo approach could be equally well described as a “parametric” bootstrap method. Another bootstrap method that makes more direct use of available data is the residual bootstrap method described below.

The Residual Bootstrapping Approach

The residual bootstrap method is similar, but makes use of sampling from the actual data rather than from fitted distributions:

1) For each data point in the set of calibrated hindcast values, add a randomly drawn residual (error) from the set of residuals, ε_t (sampled with replacement).
2) Repeat this a large number of times (e.g. 500), and conduct an extreme value fit on each dataset to obtain return period estimates.
3) Find the mean of the set of return values.

Once again, an estimate of the amount of bias due to hindcasting errors is obtained by taking the difference between this value and the return value found from the original data. Note that this method bootstraps the hindcast errors only, it is also possible to bootstrap the original set of wave heights as well to account for uncertainty due to sample size.

Both these numerical methods are simple to apply, but suffer from some potentially serious drawbacks when dealing with modeling error:

1) In both cases, we are <u>adding error to data that already contains error</u>. We are assuming that the bias due to additional error is equivalent to the inherent bias due to using ‘contaminated’ hindcast data. This may be valid in a lot of statistical applications but is debatable for use in extreme wave analysis due to tail sensitivity, especially for large errors.

2) Extreme wave analysis usually uses the peaks-over-threshold approach, which requires data above a given threshold value only. Either the exceedences over the threshold are modeled as Pareto distributed or a truncated form of a traditional extremal distribution (e.g. Weibull) is used. Both Monte-Carlo and bootstrapping simulation techniques lose their validity when considering such truncated data - the addition of negative error to points near the threshold results in loss of data, artificially reducing the population density near the threshold point.

A method by which these problems can be addressed is by convolving the hindcast error into the extreme value distribution used to conduct the statistical analysis. In this way the effects of hindcast error can be explicitly included, and valid estimates of bias for both truncated and untruncated hindcast datasets can be obtained. This approach is referred to here as the convolution integral approach.

CONVOLUTION INTEGRAL APPROACH

If we assume that the hindcast data is a random variable Z comprised of the sum of two random variables X (distribution of the underlying wave population) and E (hindcast modeling uncertainty), the probability density function of Z (f_Z) is given in terms of the pdfs of X and E by the convolution integral:

$$Z = X + E$$

$$f_Z(z) = \int_{-\infty}^{\infty} f_X(x) f_E(z-x)dx$$

If X and E are both Gaussian, this integral has a closed-form solution, otherwise this integral cannot be simply solved apart from some special cases (see for example [4]). The integral is nonetheless amenable to numerical solution. Maximum likelihood estimation of this convoluted distribution to a set of hindcast data will incorporate the hindcast uncertainty into all return value estimates and confidence intervals.

If there is evidence that hindcast error is proportional to wave height an alternative formulation could be used:

$$Z = XE$$

$$f_Z(z) = \int_{-\infty}^{\infty} f_X(x) f_E(z/x)dx$$

In principle, any assumed distribution for the underlying wave population and any description of hindcast uncertainty can be used with this method. In practice, the maximum likelihood method can give poor estimates for some distributions (e.g. the Generalised Pareto), in this case the convolution method may not be feasible.

To illustrate the application of the convolution integral approach, we will assume that the wave population distribution is two-parameter Weibull in form and the hindcast modeling error is independent and normally distributed.

As discussed before, the hindcast error can be described by:

$$f_E(\varepsilon) = \frac{1}{\sqrt{2\pi\sigma_m{}^2}} e^{\left(-\frac{\varepsilon^2}{2\sigma_m{}^2}\right)}$$

The underlying population distribution is given by the standard Weibull formula with scale parameter *a* and shape parameter *c*:

$$f_X(x) = \left(\frac{c}{a}\right)\left(\frac{x}{a}\right)^{c-1} e^{-\left(\frac{x}{a}\right)^c}$$

Accounting for a threshold value, *t*, below which data is not used, the convolution integral becomes:

$$f_Z(z;a,c,t) = \frac{\int_{-\infty}^{\infty}\left(\frac{c}{a}\right)\left(\frac{x}{a}\right)^{c-1} e^{-\left(\frac{x}{a}\right)^c} \frac{1}{\sqrt{2\pi\sigma_m{}^2}} e^{\left(-\frac{(z-x)^2}{2\sigma_m{}^2}\right)}dx}{\int_t^{\infty}\int_{-\infty}^{\infty}\left(\frac{c}{a}\right)\left(\frac{x}{a}\right)^{c-1} e^{-\left(\frac{x}{a}\right)^c} \frac{1}{\sqrt{2\pi\sigma_m{}^2}} e^{\left(-\frac{(z-x)^2}{2\sigma_m{}^2}\right)}dxdz}$$

$$= \frac{\int_0^{\infty}\left(\frac{x}{a}\right)^{c-1} e^{\left\{-\left(\frac{x}{a}\right)^c - \frac{(z-x)^2}{2\sigma_m{}^2}\right\}}dx}{\int_t^{\infty}\int_0^{\infty}\left(\frac{x}{a}\right)^{c-1} e^{\left\{-\left(\frac{x}{a}\right)^c - \frac{(z-x)^2}{2\sigma_m{}^2}\right\}}dxdz}$$

This can be solved to a reasonable level of accuracy using any numerical integration method (e.g. adaptive Simpson's rule) over some finite bounds (e.g. 0-100m).

Application of the maximum likelihood estimation method then involves maximizing the loglikelihood function for parameters *a* and *c* for the set of data z_i…..z_n

$$\text{Loglikelihood fn.} = \sum_{i=1}^{n} \log f_Z(z_i; a, c, t)$$

We can use a multivariate numerical maximization scheme such as the Quasi-Newton line search method to solve for the parameters. This procedure may be slow, but has the advantage that the Hessian matrix[2] of the likelihood function is approximated as part of the maximization algorithm. Due to the asymptotic properties of maximum likelihood estimators the (approximate) Hessian matrix can be used directly to provide estimates of the variance of the scale and shape parameters.

During the line search procedure the Hessian is approximated as:

$$A \approx \begin{bmatrix} \dfrac{\partial^2 LogL}{\partial a^2} & \dfrac{\partial^2 LogL}{\partial a \partial c} \\ \dfrac{\partial^2 LogL}{\partial a \partial c} & \dfrac{\partial^2 LogL}{\partial c^2} \end{bmatrix}$$

The covariance matrix of the parameters is then given by:

$$\begin{bmatrix} Var(\hat{a}) & Cov(\hat{a},\hat{c}) \\ Cov(\hat{a},\hat{c}) & Var(\hat{c}) \end{bmatrix} = \left[-A\right]^{-1}$$

The 'true' return value (y) for a given return period (RP) and annual rate of occurrence of wave heights exceeding the threshold (λ) is found by solving:

$$y = F_X^{-1}\left(1 - \frac{1}{\lambda RP}\right)$$
$$= \hat{a}\left[-\log\left(\frac{1}{\lambda RP}\right) + \left(\frac{t}{\hat{a}}\right)^{\hat{c}}\right]^{1/\hat{c}}$$

where F_X is the cumulative density function of x.

Variance in the return value is then obtained using the delta method:

$$Var(y) = \begin{bmatrix} \dfrac{\partial y}{\partial a} & \dfrac{\partial y}{\partial c} & \dfrac{\partial y}{\partial \lambda} \end{bmatrix} \begin{bmatrix} Var(\hat{a}) & Cov(\hat{a},\hat{c}) & 0 \\ Cov(\hat{a},\hat{c}) & Var(\hat{c}) & 0 \\ 0 & 0 & Var(\hat{\lambda}) \end{bmatrix} \begin{bmatrix} \dfrac{\partial y}{\partial a} & \dfrac{\partial y}{\partial c} & \dfrac{\partial y}{\partial \lambda} \end{bmatrix}^T$$

This method is an approximate one and should be used with a degree of caution, as calculation of the Hessian is prone to numerical error. An alternative approach is to use the profile likelihood method to derive confidence intervals.

This method is applicable when a large number of comparisons between the hindcast model and measurements have been made and the estimate of σ_m can be reasonably expected to be accurate. If, however, there is a scarcity of information available for calibration of the hindcast model, we can instead use the available residual errors, ε, directly in the fitting procedure. i.e. we maximise the likelihood of obtaining the two samples z_i and ε_j:

$$\text{Loglikelihood fn.} = \sum_{i=1}^{n} \log[f_Z(z_i; a, c, t, \sigma_m)]$$
$$+ \sum_{j=1}^{m}\left(\frac{1}{2}\log[2\pi\sigma_m{}^2] - \frac{\varepsilon_j}{2\sigma_m{}^2}\right)$$

In this formulation, the hindcast standard deviation, σ_m, becomes one of the parameters to be calculated along with a and c. In this way, we not only account for the uncertainty in the modeled wave heights but also the uncertainty in the amount of error involved with hindcasting.

COMPARISON STUDY USING KODIAK WAVE DATA

A publicly available dataset of storm significant wave heights, measured over a period of 19 years at the Kodiak site off the Alaskan coast, has been distributed by the International Association for Hydraulic Research (IAHR) [6]. We shall artificially introduce random error to this dataset in order to test the accuracy of the proposed convolution integral approach for assessing bias, and to compare it against numerical simulation methods.

The Kodiak dataset is shown in Figure 3, along with a truncated 2-parameter Weibull fit of the data. A threshold level of 6.85m is used here, corresponding with a total of 49 wave height measurements. A storm occurrence frequency of 49/19=2.579 storms/year is used to obtain return period estimates. Based on this data, the nominal 100 year return period wave height is 11.878m. 90% confidence intervals due to sampling variability are included on the plot (asymptotic normal bounds).

In order to assess the degree of bias in the 100 year return values due to random error in the data, modified datasets were created. Normally distributed random errors of zero mean and 0.8m standard deviation were added to the Kodiak dataset to provide 10000 modified datasets[3]. Refitting of the truncated Weibull distribution to each modified dataset to obtain 100 year return values produced the results shown in Figure 4.

As expected, the general trend is a positive bias in the predicted 100 year wave value, with the amount of bias ranging from -1.6m to +3.4m. The mean bias due to the addition of random error was +0.388m. In other words, for this set of data, a 0.8m standard deviation in the accuracy of each measurement

[2] Matrix of second partial derivates

[3] These errors are used for illustrative purposes only and do not necessarily reflect the hindcast uncertainty of the Kodiak dataset.

will, on average, result in a 0.388m overestimate of the 100 year wave value (an overestimation of around 3.3%).

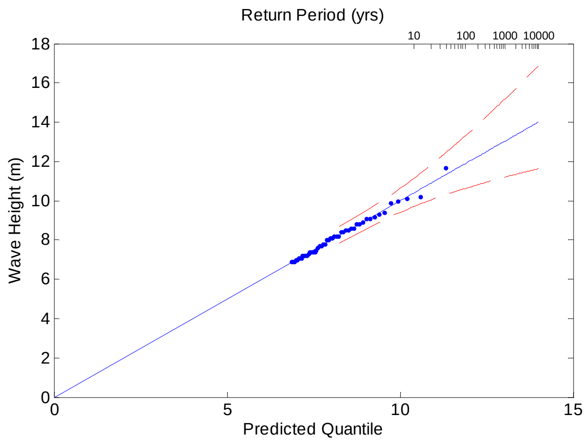

Figure 3 - Truncated Weibull Plot for Kodiak Data

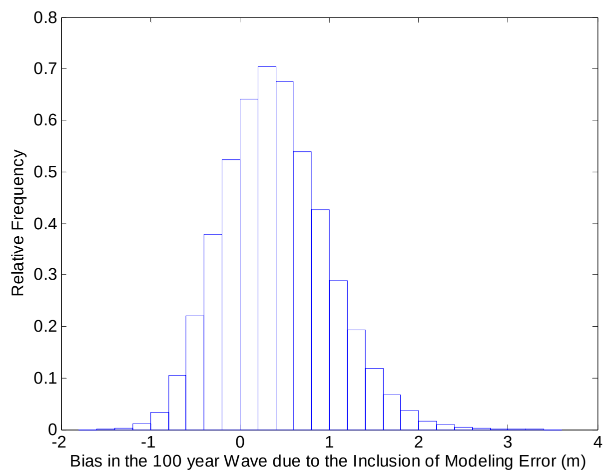

Figure 4 - Histogram of Calculated Bias in H_{100} due to the Inclusion of Modeling Error (Results from 10000 Datasets, Consisting of Normally Distributed Error of Zero Mean and 0.8m Standard Deviation added to the Kodiak Data)

A test sample with a bias near +0.388m was selected for use in a comparison study of the Monte-Carlo, residual bootstrap, and convolution integral methods for assessing bias. The results are presented in Table 1. 1000 iterations were used for both the Monte-Carlo and the residual bootstrap analyses. The error residuals were used directly in the convolution integral analysis.

Table 1 - Comparison of Different Methods of Estimating Bias in 100 year Return Period Waveheights for a Known Hindcasting Error (σ_m=0.8m)

	Actual Kodiak Data	Average Effect of Error	Monte-Carlo	Residual Bootstrap	Convolution Integral
Predicted H_{100}	11.89m	12.28m	12.17m	12.02m	11.84m
Bias	-	0.388m	0.046m	0.199m	0.370m
% Error in Bias Estimate	-	-	-88 %	-49 %	-4.6 %

It can be seen that the convolution integral approach appears to assess modeling bias reasonably well. The Monte-Carlo and Residual Bootstrap methods appear less accurate.

Whilst these results demonstrate that hindcast uncertainty can have an effect on the prediction of design conditions, the variation due to this uncertainty remains quite small in comparison to the sampling uncertainty (see for example the confidence interval in Figure 3). For 100 year conditions, the added complexity of using the convolution integral approach does not appear warranted given that the degree of bias is likely to be less than 0.5m for typical levels of hindcast uncertainty. However, in cases where there is a large degree of uncertainty in the hindcast model, perhaps due to there being only a few reliable measurements of storm conditions available in the region, or there is a need to examine very long return period values, such as in a reliability study, the effect of hindcast uncertainty is much more important.

This point is illustrated by Table 2, which shows results for the 10000 year condition for a hindcast standard deviation of 1.2m. Note that if hindcast uncertainty is neglected, the 10000 year wave height is overestimated by over a metre.

Table 2 - Comparison of Different Methods of Estimating Bias in 10000 year Return Period Waveheights for a Known Hindcasting Error (σ_m=1.2m)

	Actual Kodiak Data	Average Effect of Error	Monte-Carlo	Residual Bootstrap	Convolution Integral
Predicted H_{10000}	14.04m	15.13m	14.88m	14.12m	14.01m
Bias	-	1.093m	0.257m	1.021m	1.118m
% Error in Bias Estimate	-	-	-77 %	-6.6 %	+2.3 %

CONCLUSIONS

It has been shown here that the use of hindcast data to conduct extreme value analysis results in a small but significant positive bias in the calculation of return values of met-ocean parameters such as significant wave height. Given that hindcast uncertainty can be well estimated via calibration studies, there is no reason why the effect of hindcast uncertainty cannot be incorporated into an extremal analysis. This paper presents a method to accomplish this, and demonstrates that this new method performs better than existing numerical simulation methods in predicting bias, particularly when considering truncated datasets. Tests using a standard ocean wave dataset indicate that the effect of hindcast uncertainty on return values is of particular importance for reliability studies, and in regions where the measurement database available for model calibration is small.

ACKNOWLEDGMENTS

The authors wish to gratefully acknowledge the support and assistance provided by Woodside Energy Ltd. and by Prof. Ian James of Murdoch University. Financial support from the Cooperative Research Centre for Welded Structures is also acknowledged.

REFERENCES

[1] Reece A.M. and Cardone V.J. 1982, Test of Wave Hindcast Model Results Against Measurements During Four Different Meteorological Systems, Offshore Technology Conference, Houston, USA, OTC 4323

[2] Haver S., 1994, Uncertainties Related to Hindcast Wave Data, Proc. Int. Conf. Offshore Mechanics and Arctic Eng., Vol. II, p57-65

[3] Earle M.D. and Baer, L. 1982, Effects of Uncertainties on Extreme Wave Heights, J. Waterw. Port Coastal Ocean Div. Proc. ASCE, **108**, WW4, p 456-478

[4] Van Gelder P.H.A.J.M. 1999, Statistical Methods for the Risk-Based Design of Civil Structures, Comm. on Hydraulic and Geotechnical Eng.

[6] van Vledder G., Goda Y., Hawkes P., Mansard E., Jesus Martin M., Mathiesen M., Peltier E. and Thompson E. 1993, Case Studies of Extreme Wave Analysis: A Comparative Analysis, Proc. WAVES ’93 Conf., 26-28 July, New Orleans, USA p 978-992