J. Ind. Geophys. Union (April 2006)
Vol.10, No.2, pp.79-84

Analysis of Triple Collocation Method for validation of model predicted significant wave height data

G.Muraleedharan, A.D.Rao, Mourani Sinha and D.K.Mahapatra
Centre for Atmospheric Sciences, Indian Institute of Technology, Delhi, Hauz Khas, New Delhi -110 016.
E-mail: muralg@yahoo.com

ABSTRACT
The validation of the model (Wam, Swan and Nested-Swan) predicted significant wave height data using the new triple collocated statistical method suggests that the predicted values are sufficiently accurate when compared with the buoy measurements. The Wam and Nested-Swan significant wave height estimations give significant positive correlation with deep and shallow water buoy measurements respectively. The linear regression (LR) method is inconsistent and the new method (Functional Relationship, FR) paves way for estimating the variances of the errors in measuring and predicting the physical truth (here significant wave height). The larger the random errors, the larger are the deviations between FR and LR lines. The FR model lines align with the best fit-line (x=y) while comparing the model results and buoy measurements. Also the deviations of the data from the FR model lines are a minimum. Thus we can say FR model as compared to LR model is more realistic when inherent error exists in both cases, measurement by instruments and model predictions.

INTRODUCTION

Caires & Sterl (2003) opined that all the observing systems including numerical models illustrate the truth with random errors and with some outliers. Hence errors-in-variables models are suitable for comparison between systems. Errors-in-variables model is proposed to make triple data set comparisons. The model assumes that each data set corresponds to observations of a linear systematic deviation from the physical truth plus a random error. The limitations of using standard statistical methods for comparing data with inherent random errors have been acknowledged in several works (Bauer & Staabs 1998; Sterl, Komen & Cotton 1998).

This method cannot be applied for two sets of variables because the functional relationship between the variables and the variances of the random errors can be estimated once the ratio of these variances is known, which is often not the case.

For the readers' awareness, the actual derivation of the Functional Relationship Model (FR) or errors-in-variables and the variances of the random errors (Caires and Sterl 2003) are described here. In this work we consider the triple collocation of significant wave height observations of buoy (both deep and shallow water), WAM, SWAN and NESTED-SWAN wave models.

MATERIALS AND METHODS

The models WAM and SWAN

Although WAM and SWAN are both third generation wave models which compute random short-crested wind-generated waves, WAM was primarily developed for deep water ocean waves, whereas SWAN has been developed specifically for coastal and inland waters. As per requirement for comparison of data sets both codes can be used for shallow and deep water calculations.

The numerical model SWAN is able to characterize sea waves in nearshore regions (up to the breaking zone) taking into account the presence of currents and wind generated waves. This model was developed at the Technical University of Delft (Booij, Ris & Holthuijsen.1999) and it can be considered the nearshore version of the WAM model (WAMDI group 1988). The following basic physics are accounted for both codes:

1) Wave propagation in time and space, 2) Wave generation by wind, 3) Shoaling and refraction due to current and depth, 4) White capping and bottom friction and 5) Quadruplet wave-wave interactions.

The WAM model, as described in (Gunther et al.1992) computes the two dimensional wave variance spectrum through integration of the transport equation

$$\frac{dE}{dt}+\frac{\partial}{\partial\phi}(\dot{\phi}E)+\frac{\partial}{\partial\lambda}(\dot{\lambda}E)+\frac{\partial}{\partial\theta}(\dot{\theta}E)+\frac{\partial}{\partial\sigma}(\dot{\sigma}E)=S_{in}+S_{wc}+S_{nl}+S_{bf} \quad \text{(i)}$$

where E represents the spectral density with respect to (f,l,s,q), s- frequencies, q-directions, f-latitudes, and l- longitudes.

$\dot{\phi}, \dot{\lambda}, \dot{\sigma}$ and $\dot{\theta}$ are the rate of change of the position and propagation direction of a traveling wave packet. The source terms are a superposition of the wind input S_{in}, dissipation due to whitecapping S_{wc}, bottom friction affects S_{bf}, and nonlinear transfer processes S_{nl}. WAM uses an implicit scheme for source function integration and an explicit first-order upwind flux scheme for the propagation terms and can propagate the solution on a Cartesian mesh or a spherical grid. The wind time step can be chosen arbitrarily. In most cases the propagation time step is larger or equal to the source time step, usually taken 20 minute. A consequence of the explicit scheme is that the time step must be reduced if the mesh is highly refined (i.e. the spatial step size ≤ 20 km) to maintain the CFL criterion, thus increasing the computational effort. WAM is one of the most extensively tested wave models in the world and is well documented. A detailed description of the WAM code is given by (Gunther et al. 1992) and (Komen et al. 1993).

The SWAN code contains formulations of two physical processes not present in the WAM code which can play an important role for nearshore calculations; they are depth-induced wave breaking and triad wave-wave interaction. SWAN solves the spectral action balance equation

$$\frac{\partial N}{\partial t}+\frac{\partial}{\partial x}(c_x N)+\frac{\partial}{\partial y}(c_y N)+\frac{\partial}{\partial\sigma}(c_\sigma N)+\frac{\partial}{\partial\theta}(c_\theta N)=\frac{S}{\sigma} \quad \text{(ii)}$$

where t is the time, x and y are the space variables, σ is the intrinsic frequency (relative frequency measured in a system of reference moving with the current), è is the wave direction, c_x, c_y are the propagation velocities in geographical space, $c_ó$ and $c_è$ are the propagation velocities in spectral space (frequency and direction). The right hand side representing the source and sink terms is given as

$$S(\sigma,\theta)=S_{in}(\sigma,\theta)+S_{nl}(\sigma,\theta)+S_{wc}(\sigma,\theta)+S_{bf}(\sigma,\theta)+S_{dib}(\sigma,\theta) \quad \text{(iii)}$$

where $S_{in}(\sigma, \theta)$ is the wind input, $S_{nl}(\sigma, \theta)$ represents the nonlinear resonant quadruplet and triad wave-wave interactions, $S_{wc}(\sigma, \theta)$ stands for dissipation due to whitecapping, $S_{bf}(\sigma, \theta)$ is dissipation due to bottom friction and $S_{dib}(\sigma, \theta)$ is dissipation due to depth induced wave breaking.

SWAN uses the wave action density spectrum N (σ, θ) rather than the energy density spectrum, E (σ, θ), (as in WAM) since in the presence of currents, the wave action density spectrum is conserved whereas the energy density spectrum is not, they are related through the relation

$N(\sigma,\theta) = E(\sigma,\theta) / \sigma$

In deep water, quadruplet wave-wave interactions transfer wave energy from the peak frequency to lower frequencies (which may produce a shift in the spectrum peak towards lower frequencies) and to higher frequencies (where some energy is dissipated by white capping). In very shallow water, triad wave-wave interactions are more important, enabling the energy transfer from lower to higher frequencies, a process that may originate a secondary high-frequency peak.

The propagation in the SWAN model, in both geographic and spectral spaces, is carried out by using fully implicit numerical schemes. These schemes have the advantage of being unconditionally stable, with no CFL-restriction. In this sense the time step is unlimited but for accurate computations the time step should be sufficiently small.

The models WAM and SWAN are forced with QSCAT/NCEP blended wind which is available at http://ncardata.ucar.edu/datasets/ds744.4/data. The data is available globally at 0.5 degree resolution and at six hourly intervals. The boundary condition generated from a coarse WAM run is given input to the SWAN run which is referred as a NESTED SWAN run. The analysis region is divided in to two parts ranging from (1)10°N to 20°N and 66°E to 76°E in the Arabian Sea region and from (2)10°N to 20°N and 80°E to 90°E in the Bay of Bengal Region with 0.25 degree resolution. For this the QSCAT/NCEP wind is interpolated to 0.25 degree with bilinear cubic spline interpolation method. The model is integrated with a time step of 15 minutes and a 5 km. fetch is used.

The deep water buoy OB8(at depth 3545m) is deployed at location 11.5°N, 81.5°E and the shallow water buoys SW6 is located at 13.2°N, 80.4°E in Bay of Bengal and SW4 is located at 12.9°N, 74.7°E in Arabian Sea. The output parameters are wind speed, wind direction and significant wave height. The November and December data of 2003 comprising the post-monsoon season are considered in this work. Triple collocation of significant wave heights of WAM and NESTED-SWAN with buoy OB8 and SWAN and NESTED-SWAN with buoys SW6 (depth of recording=35.3m) and SW4 (depth of recording=51m) are carried out for the application of FR model. The conventional linear regression model (LR) is also fitted to the above mentioned data to know the additional advantage over LR by the implementation of FR model.

The coefficients of the functional relationships between the variables and the variances of their associated random errors are computed as follows (Caires and Sterl 2003). Assume that there are three sets of n observations (x_i, y_i, z_i), $i=1,2,\ldots\ldots.n$ and these observations correspond to the measurement of the physical truth with certain systematic deviations and subject to zero mean random errors. i.e. $\frac{1}{n}\sum_{i=1}^{n}(e_{xi}, e_{yi}, e_{zi}) = 0$. Let the true measurements X, Y, Z of the physical truth corresponding to the observations x, y, z are linearly related to the deterministic variables T_i, $i=1,\ldots.n$ after dropping the subscripts as

$$x=X+e_x=T+e_x \quad (1)$$
$$y=Y+e_y=\alpha_1+\beta_1 T+e_y \quad (2)$$
$$z=Z+e_z=\alpha_2+\beta_2 T+e_z \quad (3)$$

The averages of which are denoted by $<x>$, $<xy>$ etc. Now we want to estimate the unknown parameters α_1, α_2, β_1, and β_2 and the variances of the errors. Removing the mean from each of the variables and denoting the results by x^*, y^*, z^* and T^*, the model is simplified to

$$x^*=T^*+e_x$$
$$y^*=\beta_1 T^*+e_y$$
$$z^*=\beta_2 T^*+e_z$$

Noting that $<Te_x>=<Te_y>=<Te_z>=0$, and the errors are independent, so that $<e_x e_y>=<e_x e_z>=<e_y e_z>=0$, we can estimate β_1 and β_2 as follows.

$$y^*z^*=\beta_1\beta_2(T^*)^2+\beta_1 T^* e_z+\beta_2 T^* e_y+e_y e_z$$
$$<y^*z^*>=\beta_1\beta_2<(T^*)^2>+\beta_1<(T^* e_z)>+\beta_2<T^* e_y>+<e_y e_z>$$
$$=\beta_1\beta_2<(T-<T>)^2>+\beta_1<(T-<T>)e_z>+\beta_2<(T-<T>)e_y>$$
$$=\beta_1\beta_2<T^2>-2\beta_1\beta_2<T>^2+\beta_1\beta_2<T>^2$$
$$<y^*z^*>= \beta_1\beta_2(<T^2>-<T>^2), \text{ and} \quad (4)$$
$$<x^*z^*>=\beta_2(<T^2>-<T>^2) \quad (5)$$

Dividing (4) by (5)

$$(4)/(5)= \beta_1=<y^*z^*>/<x^*z^*> \quad (6)$$
$$\text{Similarly } <x^*y^*>=\beta_1(<T^2>-<T>^2) \quad (7)$$
$$(4)/(7)= \beta_2=<y^*z^*>/<x^*y^*> \quad (8)$$

Using eq (6) and (8), we obtain the variances of the errors as given below.

$$(x^*)^2=(T^*+e_x)^2 =(T^*)^2+2T^*e_x+e_x^2$$
$$=(T-<T>)^2+2(T-<T>)e_x+e_x^2$$
$$<(x^*)^2>=<T^2>-2<T><T>+<T>^2+2<Te_x>-2<T><e_x>+<e_x^2>$$

$<(x^*)^2>=<T^2>-<T>^2+<e_x^2>$ and multiplying (6) by (8)

$$\beta_1\beta_2=<y^*z^*>^2/(<x^*z^*><x^*y^*>)$$
$$<x^*z^*><x^*y^*>/<y^*z^*>=<y^*z^*>/\beta_1\beta_2$$
$$<(x^*)^2>-<x^*z^*><x^*y^*>/<y^*z^*>=<T^2>-<T>^2+<e_x^2>-<y^*z^*>/\beta_1\beta_2$$

Substituting for $\beta_1\beta_2$ from (4) gives,

$$<(x^*)^2>-<x^*z^*><x^*y^*>/<y^*z^*>=<T^2>-<T>^2+<e_x^2>-<y^*z^*>(<T^2>-<T>^2)/<y^*z^*>$$

$<e_x^2>$=Variance of (e_x) since $<e_x>=0$

ie. $Var(e_x)=<e_x^2>-<e_x>^2= <e_x^2>= <(x^*)^2>-<x^*y^*><x^*z^*>/<y^*z^*>$

Similarly we get,

$$Var(e_y)= <(y^*)^2>-<x^*y^*><y^*z^*>/<x^*z^*>$$
$$Var(e_z)= <(z^*)^2>-<x^*z^*><y^*z^*>/<x^*y^*>$$

Then the coefficients α_1 and α_2 can be estimated as:

$$y= \alpha_1+\beta_1 T+e_y$$
$$<y>=\alpha_1+\beta_1<T>+0$$
$$\therefore \alpha_1=<y>-\beta_1<T>$$

But from (1), $<x> = <T>$

$$\therefore \alpha_1=<y>-\beta_1<x> \quad (9)$$

Similarly

$$\alpha_2=<z> -\beta_2< x> \quad (10)$$

The result of applying the model to data is independent of which variables are chosen to be x, y or z. The coefficients in the relationship between Y and Z, ie.,

$Y=\alpha_3+\beta_3 Z$ can be estimated as follows:

From (1), X=T. Then from (2) and (3),

$$Y=\alpha_1+\beta_1 X \quad (11)$$
$$Z=\alpha_2+\beta_2 X \quad (12)$$

Equating the expressions for X obtained from (11) and (12) we get,

$$Y=\alpha_3+\beta_3 Z , \quad (13)$$

where $\alpha_3 = \alpha_1-\alpha_2\beta_1/\beta_2$

$\beta_3 = \beta_1/\beta_2$

The relationship between what the instrument and the numerical models observe(each with its inherent errors) on a physical truth is of interest here. The collocated data sets were formed by clubbing space and time data together.

Another of the assumption used above and in this study is that the errors associated with the different systems are uncorrelated. This may not be always the case. If the errors between two of the data sets are correlated, say $< e_y e_z>\neq 0$, the value of which is known,

then (4) will become,

$$<y^*z^*>= \beta_1\beta_2(<T^2>-<T>^2)+<e_y e_z> \quad (14)$$

From (5) and (7) we obtain

$$\beta_2(<T^2>-<T>^2)=<x^*z^*> \quad (15)$$
$$\beta_1(<T^2>-<T>^2)=<x^* y^*> \quad (16)$$

Substituting (15) and (16) in (14), we get

$$\beta_1=(<y^*z^*> - <e_y e_z>) /<x^*z^*>$$
$$\beta_2=(<y^*z^*> - <e_y e_z>) /<x^*y^*>$$

Then

$$<e_x^2>=<(x^*)^2>-<x^*y^*><x^*z^*>/(< y^*z^*> -<e_y e_z>)$$
$$<e_y^2>=<(y^*)^2>-(<x^*y^*><y^*z^*>-<e_y e_z><x^*y^*>)/< x^*z^*>$$
$$<e_z^2>=<(z^*)^2>-(<x^*z^*><y^*z^*>-<e_y e_z><x^*z^*>)/< x^*y^*>$$

Besides obtaining point estimates of the variables the estimates of their standard errors, reflect their range of variability. The bootstrap method of estimating standard errors of estimators is simple to apply.

The bootstrap estimate of the standard error of the estimate is given by

$$\hat{S}_B\left(\hat{\theta}\right) = \sqrt{\frac{1}{B-1}\sum_{b=1}^{B}\left(\hat{\theta}_b^* - \hat{\theta}^*\right)^2} \quad \text{with} \quad \hat{\theta}^* = \frac{1}{B}\sum_{b=1}^{B}\hat{\theta}_b^*$$

where B is the number of bootstrap samples x^* obtained by randomly sampling n times with replacement from the original sample $x=\{x_i, i=1,.......n\}$.

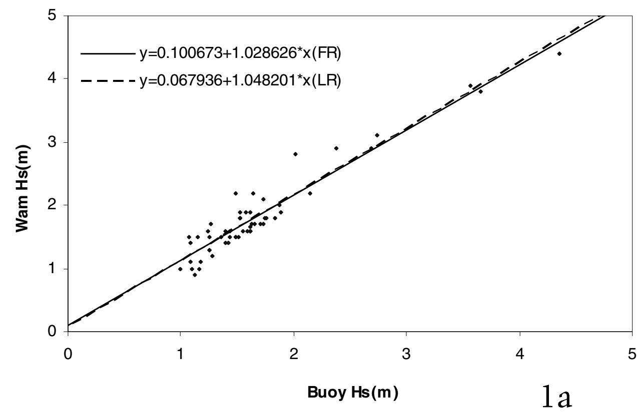
The ideal bootstrap estimate would be $\hat{S}_\infty(\hat{\theta})$, but this is not possible to achieve; a limit on B is applied. According to Efron & Tibshirani (1993), more than B=200 bootstrap samples are needed for estimating the standard error. We can also obtain 95% confidence intervals for a parameter θ, by calculating upper and lower limits of the form

RESULTS AND DISCUSSION

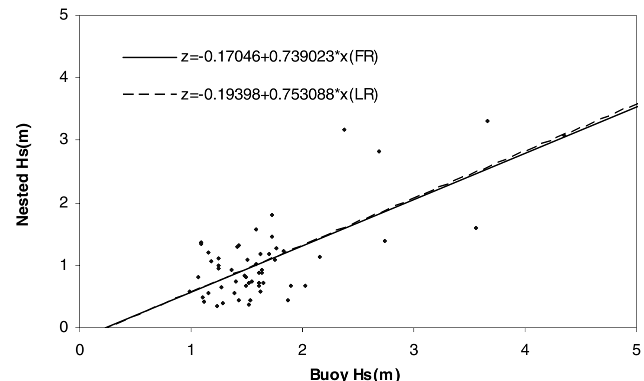
Validation

Buoy(OB8), Wam and Nested-Swan

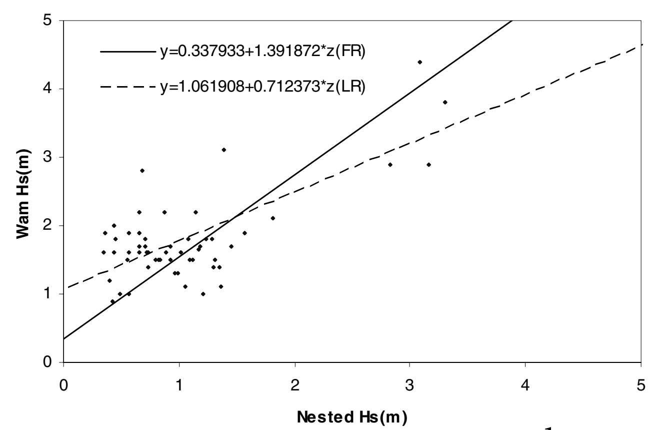
First, the deep water buoy-OB8, Wam and Nested-Swan significant wave heights were compared. The scatter diagrams of the comparisons between the three data sets in December'03 are presented in figs.1a-1c. The functional relationship (discussed above) lines are fitted and no correlations are assumed between the errors. In fig.(1a) the Wam-buoy scattered data fall almost along the line of perfect match (where x, y values are equal for equal scales) and the scatter is less compared to Nezted-Swan-buoy (Fig.1b) and Wam-Nested-Swan (Fig.1c). The functional relationship lines are aligning with the best fit-line in all the cases. The estimate of the variances of the random errors in the Nested-Swan model is much higher than those for the buoy and Wam model, as it is evidenced by the scatter of the plots (Table1). The scatter of the buoy-Wam comparison (Fig.1a) is much smaller than those in the comparisons with the Nested-Swan model (Fig.1b & 1c). For a height range of up to 1.5m (most of the data are within this range, Fig.1c) the FR model prediction is fair whereas LR predicted WAM wave heights are rather higher than the respective Nested-SWAN wave heights



1a



1b



1c

Scatter Diagrams with estimated FR and LR for significant wave height (deep water) triple buoy (OB8), Wam and Nested collocated data (55 data points) for December 2003

To know how the conventional statistical method influences the results, the linear regression analysis was also performed (Figs.1 & 2). The differences are larger between the triple method and the linear regression when the random errors are high.

These studies reveal that the Wam is performing better in predicting the significant wave heights than the Nested-Swan in deep waters.

Buoy (SW6), Swan and Nested-Swan

To test the performance of the Swan and Nested-Swan in predicting the significant wave heights in shallow waters, the outputs are compared with the Buoy (SW6) data for December, figs.2a to 2c.

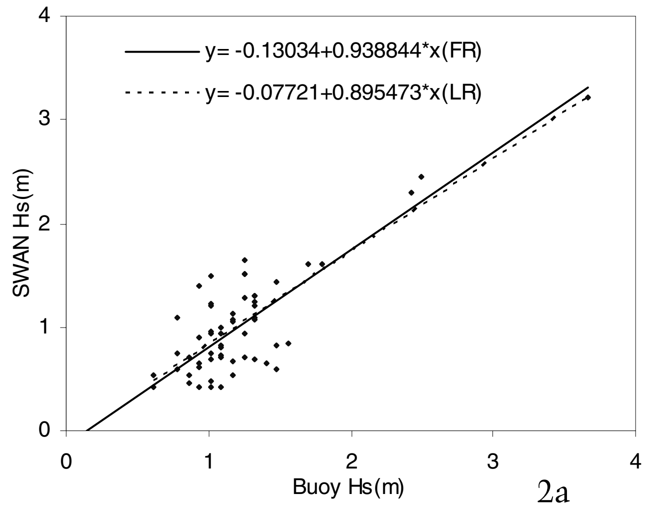
In Fig.2b, the Buoy-Nested-Swan scattered data fall almost along the line of perfect match and the scatter is less compared to Swan-buoy (Fig.2a) and Swan-Nested-Swan (Fig.2c). The estimate of the variances of the random errors in the Swan model is much higher than those for the buoy and Nested-Swan model, as it is evidenced by the scatter of the plots (Table.1). The scatter of the Buoy-Nested-Swan (Fig.2b) is less than those with the Swan model (Figs 2a and 2c).

The triple collocated method and the linear regression method show greater difference when the variances of the random errors are large (fig.2a). The Nested-Swan model seems to be estimating the shallow water significant wave height better than the Swan.

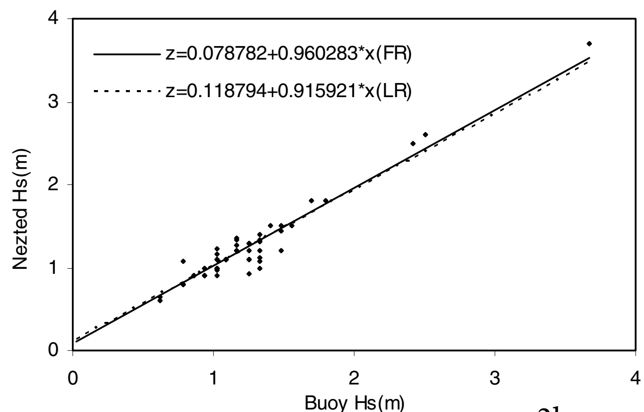
The Buoy, Wam and Nested-Swan triple collocated significant wave height data in deep water for east coast are also compared for November, 2003 to test the consistency in predictions by the models. In this case the scattering of the data is comparatively high and alignment with the best-fit line is also less. This may be due to high variances of the random errors for buoy data (ie.0.06).

The scatter diagrams of the triple collocated data (shallow water); buoy (SW6) measurement and predicted Swan and Nested-Swan significant wave heights are also plotted for November 2003. The scattering is low as evidenced from the values of the variances of the errors (Table1).

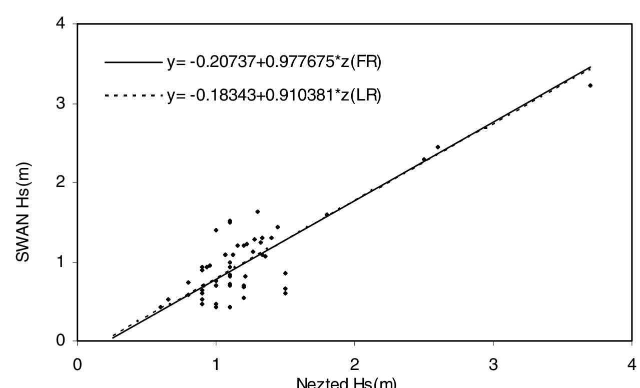
In order to show the collocated data to be independent of space and time, the above analysis is also carried out for the west coast in shallow waters for the same months. The buoy (SW4), Swan and Nested-Swan triple collocated significant wave heights



2a



2b



2c

Scatter Diagrams with estimated FR and LR for significant wave height (shallow water) triple buoy (SW6), Swan and Nested collocated data(56 data points) for December 2003

Table 1. Variances of the errors in different months and locations for the year 2003 (S: Shallow water, D: Deep water)

	Nov						Dec					
	$<e^2_x>$		$<e^2_y>$		$<e^2_z>$		$<e^2_x>$		$<e^2_y>$		$<e^2_z>$	
	S	D	S	D	S	D	S	D	S	D	S	D
East Coast	0.03	0.06	-0.01	0.02	0.01	0.03	0.01	-0.01	0.08	0.05	0.02	0.22
West Coast	0.01		0.00		0.00		0.03		0.00		0.00	

are taken in pairs and the scattering are observed and modeled using FR and LR. They are scattered along the X-axis except for Swan-Nested-Swan combination, which lies along the line of perfect match.

It is interesting to see that the FR model lines are trying to align with the line of perfect match at the same time they are fitted in such a way that the deviations of the scattered points from the lines are a minimum. These seems to be more realistic in a situation where there exists inherent errors in measuring a physical truth both by instruments and predictive models. Whereas the LR model lines completely skip from the best-fit line alignment while trying to fit the data.

The overall analysis suggests that the Wam and Nested-Swan models predicts the significant wave heights more accurately in deep and shallow waters respectively, taking into account the random errors in measuring and predicting the physical truth(here significant wave height).

CONCLUSIONS

The FR model is a novel method of validating the observations by measurements and models, incorporating the inherent errors involved while representing the physical truth. It also paves way for estimating the variances of the random errors. The scattering of the data from the line of perfect match is significant when the variances of the random errors are high. Atleast 3 sets of collocated (w.r.t space and time) data are required to apply this method.

ACKNOWLEDGEMENT

The authors are thankful to The Director, Indian National Centre for Ocean Information Services (INCOIS), for providing the recorded wave data.

REFERENCES

Bauer, E. & Staabs C., 1998. Statistical properties of global significant wave heights and their use for validation, J Geophys Res, 103, 1153-1166.

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

Caires, S. & Sterl A., 2003. Validation of ocean wind and wave data using triple collocation, J Geophys. Res, 108(C3)3098, doi:10.1029/2002JC001491.

Efron, B. & Tibshirani, R. J., 1993. An introduction to the bootstrap,(Chapman and Hall, New York).

Gunther, H.P., Hasselmann, K., Hasselmann, S. & Janssen, P.A.E.M., 1992. The WAM model Cycle 4, (DKRZ Report No. 4, Hamburg).

Komen, G.J., Cavaleri, L., Donelan, M., Hasselmann, K., Hasselmann, S. & Janssen, P.A.E.M., 1994. Dynamics and Modelling of Ocean Waves, (Cambridge University Press, Cambridge).

Sterl, A., Komen, G. J. & Cotton, P.D., 1998. Fifteen years of global wave hindcasts using winds from the European Centre for medium-range weather forecasts reanalysis: validating the reanalyzed winds and assessing the wave climate, J Geophys Res, 103, 5477-5494.

G. Muraleedharan is Research Associate at the Centre for Atmospheric Sciences, IIT Delhi. He has many years research experience in ocean wave climatology including operational wave conditions and extreme risk analysis and developed several statistical tools in this field. Received his M.Sc. and Ph.D in Oceanography from Cochin University of Science and Technology and also submitted post-doctoral thesis for the award of Doctor of Science of the University. Authored about 30 national/international peer reviewed articles. His research interests include application of statistical tools on numerical ocean models for design wave estimations, return periods of severe cyclones and extreme risk assessments.

A.D. Rao is Professor at the Centre for Atmospheric Sciences, IIT Delhi. He has wide experience in numerical modeling including waves, storm surges, ocean and estuarine circulation. He received his Ph.D from IIT Delhi. He has published more than 100 papers in various national and international journals of repute. He is very actively involved in several sponsored R & D projects.

Mourani Sinha did her M.Sc in Applied Mathematics from Calcutta University. She is at present doing her Ph D at the Centre for Atmospheric Sciences, IIT Delhi. Her research interests are numerical ocean wave prediction system in the northern Indian Ocean and model output analysis using statistical techniques for design and operational conditions.

Debasis Kumar Mahapatra is at present working as a Project Scientist in the Centre for Atmospheric Sciences. He is doing his Ph D in the field of internal waves and also keeps keen interest in sea surface gravity waves and coastal upwelling phenomena.