Using Multi-Layer Perceptrons to Predict the Presence of Jellyfish of the Genus *Physalia* at New Zealand Beaches

David R. Pontin, Michael J. Watts, *Member, IEEE*, and S. P. Worner

***Abstract*— The apparent increase in number and magnitude of jellyfish blooms in the worlds oceans has lead to concerns over potential disruption and harm to global fishery stocks. Because of the potential harm that jellyfish populations can cause and to avoid impact it would be helpful to model jellyfish populations so that species presence or absence can be predicted. Data on the presence or absence of jellyfish of the genus Physalia was modelled using Multi-Layer Perceptrons (MLP) based on oceanographic data. Results indicated that MLP are capable of predicting the presence or absence of Physalia in two regions in New Zealand and of identifying significant biological variables.**

I. Introduction

Jellyfish blooms have the potential to change the species composition in an ecosystem through altering the availability of food resources, and therefore, threatening fisheries [8]. Furthermore, it has been reported that jellyfish populations are increasing in both the intensity and frequency of blooms [12], [17]. To begin to understand potential impacts to marine ecosystems it is important to first understand factors that contribute to the formation of a bloom and once formed predicting where and when the bloom occurs. By being able to predict where and when jellyfish are likely to occur it is possible to safeguard fisheries and mitigate the threat of jellyfish stings on swimmers at beaches in coastal regions.

The genus *Physalia* is one of the most commonly found jellyfish on New Zealand beaches, and is the most commonly found stinging jellyfish. *Physalia* is considered to be one of the more primitive living jellyfish as it lacks many of the morphological characteristics associated with species that evolved later [4], [5]. In particular *Physalia* only have a pneumatophore (float) and lack a swimming bell [4] causing them to permanently inhabit the surface of the ocean [10]. Also the lack of any swimming mechanisms means that *Physalia* is completely dependant on ocean winds and currents for movement. The only adaptation for movement *Physalia* possess is the float, in that there are two morphs one with a left hand sail and one with a right hand sail, allowing individuals to move at slightly different angles in the same wind condition [1]. These characteristics mean that potentially any *Physalia* population movements can be modelled based on wind, current and swell information. For this reason *Physalia* are an ideal target species to investigate the problem of predicting the occurrence of jellyfish populations based on oceanographic data..

Because detailed scientific datasets on jellyfish are virtually non-existent we used a data set that has been collected for non-scientific purposes. The dataset was sourced from Surf Lifesaving New Zealand (SLSNZ). Surf Lifesaving New Zealand is a volunteer organisation that provides surf lifeguards on beaches throughout New Zealand. Because SLSNZ is a volunteer organisation it is reliant on community funding to operate and subsequently has developed sophisticated recording systems to document all aspects of their service to the community. The result is that there are detailed records in electronic format of every patrol that has occurred on the 72 patrolled beaches in New Zealand over at least seven years. The unique aspect of this dataset is that incidents involving jellyfish stings have been recorded. Based on investigation of the data and the fact that *Physalia* is the only stinging species regularly recorded we regarded the data held by SLSNZ as a proxy presence/absence dataset for *Physalia* in New Zealand.

Clearly such data is noisy with non-linear patterns. Artificial Neural Networks (ANN), and Multi-Layer Perceptrons (MLP) in particular have shown great promise in their application to identify factors that influence biological populations, particularly in a complex environment [11], [14], [6] however, their use for this purpose in ecology is still not widely accepted despite having been shown to out perform more conventional techniques [11], [2], [13]. The combination of high model performance and the ability to determine variable contributions to the model makes ANN a valuable tool for understanding the underlying factors that drive the presence of *Physalia* at New Zealand beaches.

The aim of this study was to investigate the potential of an ANN model to predict the presence of *Physalia* on New Zealand beaches based on oceanographic data, and to use the model to determine factors that may cause or inhibit the occurrence of *Physalia*.

II. Method

A. Data

As the goal of this work was to predict the presence of *Physalia* jellyfish on New Zealand beaches from oceanographic data, two data sets were sourced and combined into the final modelling data set. These sets were oceanographic data and data from Surf Lifesaving New Zealand (SLSNZ)

1) Oceanographic data: Oceanographic data was sourced from the National Institute of Water and Atmosphere (NIWA). The data contained time series outputs from

David Pontin and S. P. Worner are with the Bio-Protection and Ecology Division, PO Box 84, Lincoln University, Lincoln 7647, New Zeland (emails:{pontin,worner}@lincoln.ac.nz)

Michael J. Watts is with the School of Biological Sciences, University of Sydney, NSW 2006, Australia (email: michael.watts@bio.usyd.edu.au).

978-1-4244-1821-3/08/$25.00©2008 IEEE

NOAA/NCEP Wavewatch III model hindcast [19] representing eighty 1.25×1 degree global grid cells surrounding New Zealand. Each cell contained three-hourly measurements of five variables (significant wave height (m), peak wave period (s), peak wave direction (°N) and U and V wind vector components (ms^{-1})). MATLAB® was used to transform and manipulate the files so that they were able to be incorporated in the models. All variables were transformed to daily data points, by averaging each of the eight data points for each day. Furthermore, from the U and V wind vector components, wind velocity (ms^{-1}) and direction were calculated. The circular mean was used for all directional variables. Once the transformations had been completed each file contained daily data for significant wave height (m), peak period (s), peak direction (°N), wind velocity (ms^{-1}) and wind direction (°N). For each region, data from a cell was included if the cell was less than 250km distant from the centre of the region. For this work the oceanographic data for two regions in New Zealand were extracted, West Auckland and the Bay of Plenty. The oceanographic cells associated with each of these regions are shown in Figures 1 and 2, respectively.

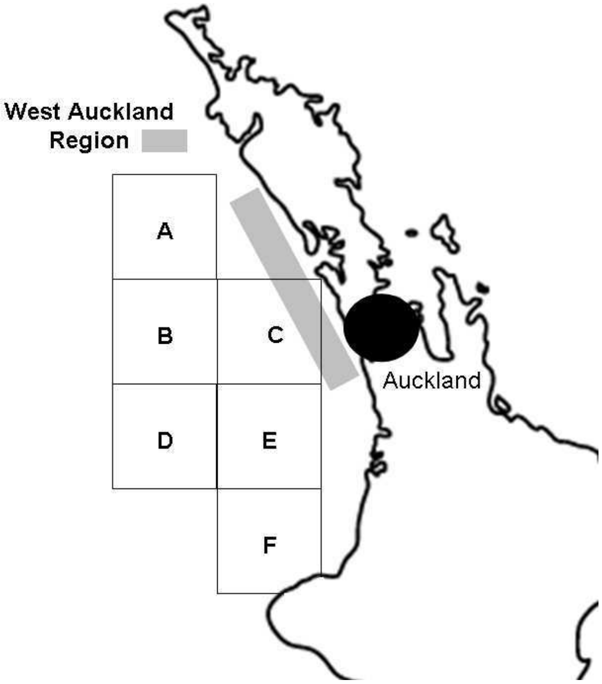

Fig. 1. Oceanic cells associated with the West Auckland region

2) Surf Lifesaving data: Data concerning jellyfish incidents was sourced from Surf Lifesaving New Zealand (SLSNZ). SLSNZ maintains an electronic database of all patrol records. We accessed the records of patrols carried out from the 2000/2001 season to the 2004/2005 season. The database recorded all incidents of jellyfish stings. In addition all patrol records carried out during this time period were also extracted. Records that showed a beach headcount of zero (that is, there were no people on the beach) were excluded, as clearly there will be no jellyfish incidents if no one is swimming at the time. The use of the SLSNZ data restricted the study to dates from late southern hemisphere spring to early autumn as this is the time when lifeguards patrol the beaches.

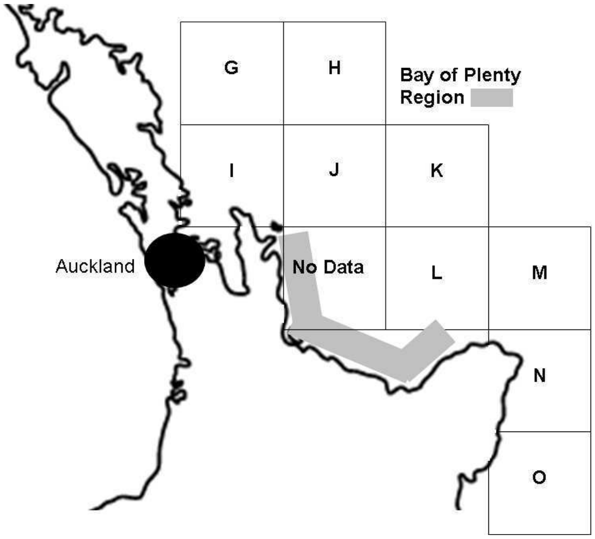

Fig. 2. Oceanic cells associated with the Bay of Plenty region

3) Final Data Sets: The West Auckland data set contained 434 data points of which 100 (23%) represented the presence of *Physalia*. The West Auckland data set contained 36 variables from six ocean cells, with five variables each, and six single month periods. The Bay of Plenty data set contained 411 data points of which 79 (19%) represented the presence of *Physalia*. The Bay of Plenty data set contained 51 variables from nine ocean cells of five variables each, and six single month periods. Months in both regional data sets were represented using an orthogonal binary encoding.

B. Training and Evaluation of MLP

Standard three neuron-layer MLP were used in these experiments, and the learning algorithm used was unmodified back-propagation with momentum. Each network modelled a single region, that is, there was only one output neuron per network, where the output indicated the predicted presence or absence of *Physalia* at the region on that particular day.

The method of training and evaluating the MLP (and also selecting the parameters) was similar to that suggested in [7], [16]. A total of 64 runs were carried out over each region, where each run used a different combination of hidden neuron layer size, learning rate and momentum. Each run consisted of 1000 trials. For each trial, the training and test data set was randomly divided into a training set, consisting of two-thirds of the available data, and a test set consisting of the remaining one-third. A MLP was then created with randomly initialised connection weights and trained over the training data set. The accuracy of the MLP over the training set was then evaluated to determine how well the network had learned the training data. The accuracy of the MLP was then evaluated over the testing data set to determine how well the network generalised. Accuracy was measured as both the percentage of examples correctly classified and using Cohen's Kappa statistic [3]. Whereas percentage accuracy is easily interpreted, it is also easily biased by unbalanced numbers of classes. That is, percentage correct may be misleadingly high when the data set in question has only a small number of examples from one class. The Kappa statistic takes the number of examples of each class into account and thus yields a less biased measure of accuracy than percentages.

For each trial the contributions of each input neuron to the output of the network was also determined, using the method of Olden and Jackson as described in [14]. This method has been experimentally determined to give the least-biased estimate of the contribution of each input neuron [15] and has been used previously in ecological modelling applications [6].

At the completion of the 64 runs, the run with the highest mean kappa over the testing sets was selected as the winner for that region. The accuracy of the networks within this run was then evaluated over the validation data set. A sensitivity analysis was also performed over the significant continuous input variables of the best-generalising network within that run. That is, a sensitivity analysis was performed over each non-binary variable of the MLP with the highest testing Kappa of the winning run. This was to illustrate the response of the network to variations in these variables so that the influence of strongly contributing inputs (as determined above) could be investigated.

III. Results and Discussion

A. Training Parameters

The optimal training parameters for each region, as determined by generalisation accuracy, are presented in Table I. The number of hidden neurons and amount of training required for the Bay of Plenty region was substantially greater than that required for the West Auckland region. Although a general rule of thumb for determining the architecture of MLP is that the number of connections should be less than the number of training examples, in the case of the networks for the Bay of Plenty, reducing the number of hidden neurons such that this rule was observed meant that the performance of the networks was unacceptably low.

TABLE I

Optimal training parameters by region. "Neurons" is the number of hidden layer neurons

Region	Neurons	Epochs	Learning rate	Momentum
West Auckland	5	200	0.05	0.1
Bay of Plenty	15	500	0.1	0.1

B. Accuracies

The accuracies of the MLP for each region are presented in Table II as both overall percentage correct and as Cohen's Kappa statistic. It is apparent that the networks for both regions were able to generalise reasonably well. For the West Auckland region, the validation accuracies were the highest accuracies recorded for that region. While the results for the Bay of Plenty region would seem to indicate that overtraining has occurred, as could be expected from the size of the networks, the high validation accuracy shows that the networks were none the less still able to generalise beyond the training data.

There was a relatively large gap between the percentage accuracies and Kappa values over the test data sets. This indicates that a relatively large number of test presence examples were falsely classified as absences. A large number of false negatives could be expected to yield a high validation accuracy if the number of presences in the validation set is very low. However analysis of the validation data showed that the distribution of occurrences in the validation data set was equal to that of the training and testing set. Also, a large number of false negatives would adversely affect the Kappa statistic for the validation data set, which plainly did not happen.

TABLE II

Mean and standard deviation of accuracies per region. "Train" is the accuracy over the training data sets, "Test" is the accuracy of the test data set and "Validate" is the accuracy over the independent validation data set. Rows labelled "%" are overall percentage accuracies and rows labelled "κ" are accuracies expressed as Cohen's Kappa

Region		Train	Test	Validate
West Auckland	%	80.88/1.82	77.79/3.26	82.0/1.96
	κ	0.35/0.07	0.25/0.08	0.37/0.07
Bay of Plenty	%	95.13/1.77	75.15/3.99	81.89/4.02
	κ	0.83/0.06	0.19/0.09	0.45/0.10

C. Most Contributing Variables

The four variables that positively contributed the most to the networks for each region are presented in Table III, and the four variables that negatively contributed the most for each region are presented in Table IV. It is immediately apparent from both of these tables that the contributions of the inputs for the Bay of Plenty region networks were much larger than for the West Auckland region networks. This is almost certainly because of the greater amount of training that the Bay of Plenty region networks received: as the method of Olden and Jackson [14] is a decompositional, weight-based method, a larger amount of training meant that the magnitudes of the connection weights were able to grow larger than was the case with the West Auckland region networks. Therefore, the contributions were correspondingly higher.

TABLE III

MOST POSITIVELY CONTRIBUTING VARIABLES. "CONTRIBUTION" IS A UNITLESS MEASURE OF INPUT VARIABLE CONTRIBUTION TO THE ACTIVATION OF THE OUTPUT. THE LETTERS AFTER THE VARIABLES "WAVE PERIOD", "WAVE HEIGHT" AND "WIND DIRECTION" INDICATE THE OCEANOGRAPHIC CELLS FROM FIGURES 1 AND 2 THE VARIABLE IS MEASURED FROM

Region	Variable Name	Contribution
West Auckland	January	6.35/1.18
	December	5.67/1.27
	Wave period C	4.37/1.92
	Wave period F	2.79/1.43
Bay of Plenty	January	42.85/6.64
	December	35.58/7.91
	Wind direction G	19.40/12.83
	Wave height J	19.19/6.67

The months of January and December are significant positive variables for both regions. That is, there was a greater probability of *Physalia* being present in these regions during these months than at other months examined. This is considered biologically plausible as December and January are both warm months [9]. This means that there is potentially more food present for the jellyfish during these months and the increase in sea surface temperature allows for more rapid growth and reproduction. Wave period is also significant for the West Auckland region. An increase in wave period denotes that the waves have been generated further away [18] indicting that there had been sustained conditions that would transport the jellyfish into the region and hence increase their probability of occurring. A large wind direction was found to be significant for oceanic cell G in the Bay of Plenty region. If one assumes that the spawning grounds are to the North of the region then wind from this direction is more likely to blow jellyfish into the Bay of Plenty area with local conditions influencing their occurrence at beaches. A larger wave height, especially in combination with wind direction, enables the jellyfish to travel further, faster, increasing the probability of arrival in the region.

TABLE IV

MOST NEGATIVELY CONTRIBUTING VARIABLES. "CONTRIBUTION" IS A UNITLESS MEASURE OF INPUT VARIABLE CONTRIBUTION TO THE ACTIVATION OF THE OUTPUT. THE LETTERS AFTER THE VARIABLES "WIND DIRECTION" AND "WAVE DIRECTION" INDICATE THE OCEANOGRAPHIC CELLS FROM FIGURES 1 AND 2 THE VARIABLE IS MEASURED FROM

Region	Variable Name	Contribution
West Auckland	April	-4.79/1.18
	Wind direction F	-4.75/1.67
	Wind direction E	-3.78/2.60
	March	-3.59/1.64
Bay of Plenty	Wave period L	-34.46/15.0
	Wind direction H	-29.36/13.72
	Wave direction K	-29.10/10.05
	Wind speed G	-24.87/13.83

The months of April and March had a significant negative contribution for the West Auckland region. That is, there was lesser probability of *Physalia* being present in this region during these months than in other months. This is also considered to be biologically plausible as the temperatures during this time decrease significantly [9]. Increases in wind direction in oceanic cells E and F also decreases the probability of *Physalia* being present. As can be seen in Figure 1, as wind direction becomes more northerly, jellyfish may be blown past the West Auckland region or this result may indicate where a *Physalia* spawning ground is located. Unpublished genetic data suggests that there is a possibility of a spawning ground in the Tasman Sea to the south-west of Auckland which supports the model assumption that more northerly winds decrease the probability of *Physalia* presence.

For the Bay of Plenty region, the wind direction in oceanic cells H and K makes a significant negative contribution as shown in Table IV. In other words, as wind direction in these cells becomes more northerly, the probability of *Physalia* presence decreases. This contradicts the interpretation of what happens in oceanic cell G but is reasonable as both oceanic cells H and K are located further away from the coast and only winds from the north-east would cause jellyfish to be pushed towards the bay. The situation with wave period for the Bay of Plenty region is the exact opposite to the West Auckland region. This result indicates that local conditions are more important for the occurrence of *Physalia* in the Bay of Plenty region.

D. Sensitivity Analysis

Sensitivity analysis is a way to visualise how an ANN responds to the variation of a single variable. To perform a sensitivity analysis over variable n, all other input variables are set to their mean values, while the values of n are varied across the range of n, and the output of the ANN recorded. The advantage of a sensitivity analysis is that it allows for a more detailed investigation of the importance of a particular variable. Whereas an analysis of the importance of each input will yield a single overall value for the contribution of each input, a sensitivity analysis shows how the network reacts to that variable across its range. Results of the sensitivity analysis are shown in Figure 3 for Auckland and 4 for the Bay of Plenty. Variables analysed from the West Auckland region showed that the networks response to variation from all variables examined was linear. The variables analysed from the Bay of Plenty region showed more of a curved nature as would be expected from the greater amount of training and subsequent greater contributions of the variables to the network, in particular wind direction from cell H strongly indicated that winds greater than $180°$ were not conducive to the presence of *Physalia*. Sensitivity analyses were not performed over binary variables, as this was not appropriate. Therefore, even though months such as January and December were found by contribution analysis to be very significant for the West Auckland Region, no sensitivity analysis was performed for these variables.

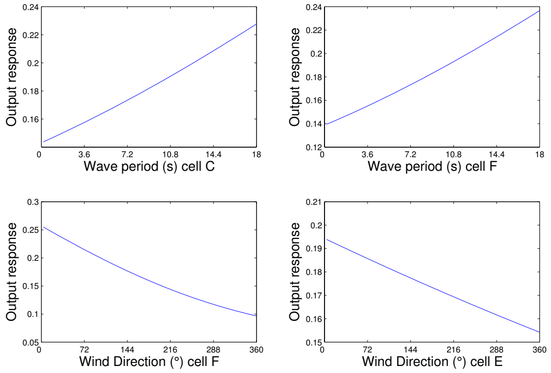

Fig. 3. Sensitivity analysis of the most significant continuous variables for the West Auckland region

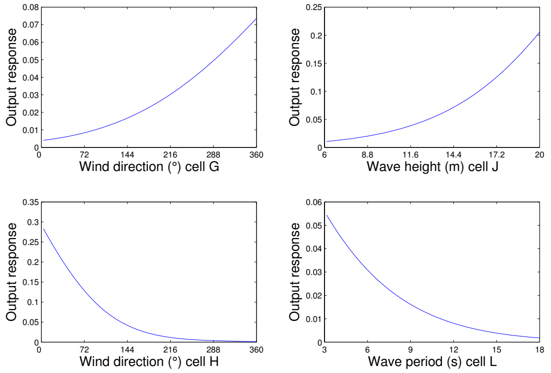

Fig. 4. Sensitivity analysis of the most significant continuous variables for the Bay of Plenty region

E. Issues and Improvements

As is often the case with ecological data sets, the data used in this study is likely to have been very noisy. This is because the presence and absence of jellyfish were inferred from reported jellyfish stings of swimmers. This leads to several potential gaps in the data set: firstly, because it is quite possible for jellyfish to be present and not sting people; secondly, it is possible that some stings were not reported; thirdly, because beaches are not uniformly patronised during the week, as there are far more swimmers during the weekend and public holidays than there are during the working week. However, absent a large, expensive and multi-year survey of *Physalia* populations, this data is all that there is to work with.

IV. Conclusion

The paper has presented an investigation into the potential for using MLP to predict the presence or absence of jellyfish of the genus *Physalia* at the beaches in two regions of New Zealand. It has also presented the results of input variable contribution analysis of the resulting networks. The results have shown that MLP are able to learn to predict the presence of *Physalia* in the two target regions from oceanographic data, to a reasonable degree of accuracy. The results of the contribution analysis appear to be biologically feasible. Furthermore, the contribution analysis enables further optimisation generating and investigating additional hypotheses concerning *Physalia* presence and absence based on oceanographic data.

Future work will expand the study to other regions of New Zealand, and will examine methods of improving performance, such as reducing the number of input variables used by removing variables that are highly correlated.

References

[1] Barnes, R. 1980. Invertebrate Zoology. Saunders College, Philadelphia, US.

[2] Brosse, S., J.-F. Guegan, J.-N. Tourenq, and S. Lek. 1999. The use of artificial neural network to assess fish abundance and spatial occupancy in the littoral zone of a mesotrophic lake. Ecological Modelling 120:299-311.

[3] Cohen, J., 1960. A coefficient of agreement for nominal scales, Educational and Psychological Measurement 20, 3746.

[4] Collins, A. G. 2002. Phylogeny of Medusozoa and the evolution of cnidarian life cycles. Journal of Evolutionary Biology 15:418-432.

[5] Dunn, C. W., P. R. Pugh, and S. H. D. Haddock. 2005. Molecular phylogenetics of the siphonophora (Cnidaria), with implications for the evolution of functional specialization. Systematic Biology 54:916-935.

[6] Joy, M. K., and R. G. Death. 2004. Predictive modelling and spatial mapping of freshwater fish and decapod assemblages using GIS and neural networks. Freshwater Biology 49:1036-1052.

[7] Flexer, A. Statistical Evaluation of Neural Network Experiments: Minimum Requirements and Current Practice. In: Trappl, R., Cybernetics and Systems '96, Proceedings of the 13th European Meeting on Cybernetics and Systems Research. Austrian Society for Cybernetic Studies, 1005-1008. 1996.

[8] Graham, W. M., F. Pages, and W. M. Hammer. 2001. A physical context for gelatinous zooplankton aggregations: a review. Hydrobiologia 451:199-212.

[9] Greig, M.J., Ridgway, N.M., and Shakespeare, B.S. (1988) Sea surface temperature variations at coastal sites around New Zealand. New Zealand Journal of Marine and Freshwater Research, 22, 391-400.

[10] Lane, C. E. 1960. The Portuguese Man-of-War. Scientific American 2002:158-168.

[11] Lek, S., M. Delacoste, P. Baran, I. Dimopoulos, J. Lauga, and S. Aulagnier. 1996. Application of neural networks to modelling nonlinear relationships in ecology. Ecological Modelling 90:39-52.

[12] Mills, C. E. 2001. Jellyfish blooms: are populations increasing globally in response to changing ocean conditions? Hydrobiologia 451:55-68.

[13] Mutanga, O., and A. K. Skidmore. 2004. Integrating imaging spectroscopy and neural networks to map grass quality in the Kruger National Park, South Africa. Remote Sensing of Environment 90:104-115.

[14] Olden, J. D., and D. A. Jackson. 2002. Illuminating the "black box": a randomization approach for understanding variable contributions in artificial neural networks. Ecological Modelling 154:135-150.

[15] Olden, J.D., Joy, M.K. and Death, R.G. An accurate comparison of methods for quantifying variable importance in artificial neural networks using simulated data. Ecological Modelling 178:389-397. 2004.

[16] Prechelt, L. A Quantitative Study of Experimental Evaluations of Neural Network Learning Algorithms: Current Research Practice. Neural Networks 9(3) 457-462. 1996.

[17] Purcell, J. E. 2005. Climate effects on formation of jellyfish and ctenophore blooms: a review. Journal of the Marine Biological Association of the United Kingdom 85:461-476.

[18] Toba T., Iida. N., Kawamura H., Educhi N. and Jones, I.S.F. Wave dependence of sea-surface wind stress Journal of Physical Oceanography 1990 vol 20 pg 705-721

[19] Tolman, H. L. 1998. Validation of a new global wave forecast system at NCEP. Pages 777-786 in B. L. Edge and J. M. Helmsley, editors. Ocean Wave Measurements and Analysis. ASCE.