FORECASTING NEARSHORE WAVE FEATURES: A COMPARISON OF DETERMINISTIC AND DATA DRIVEN MODELS

By

John R. Holloway

A paper submitted in partial fulfillment of the
requirements to complete Honors in the Department
of Physics and Physical Oceanography

Examining Committee:

Dylan E. McNamara, PhD

Fredrick M. Bingham,PhD

Scott L. Nooner, PhD

Honors Council Representative

Director of the Honors Scholars College

Approved By:

Dylan E. McNamara, PhD
Associate Professor, Department
of Physics & Physical Oceanography

Dr. Lynn Leonard
Chair, Department of Physics &
Physical Oceanography

University of North Carolina Wilmington

Wilmington, North Carolina

April 2016

TABLE OF CONTENTS

ABSTRACT

Many operational, research, and recreation activities rely on accurate forecasts of nearshore wave information. Common forecasting methods rely on models built using equations representing the underlying physics. These models contain parameterized representations of dissipative processes leaving forecasts prone to error. This thesis investigates whether data driven techniques can be used to improve wave forecasts relative to forecasts provided by a state-of-the-art wave model. Specifically, nonlinear time series forecasting and artificial neural networks are applied to data collected from historical buoy records to make predictions on wave data collected from a nearshore ADCP. These data driven methods are compared to wave forecasts provided by a new deterministic wave model, the Nearshore Wave Prediction System (NWPS). Results show that, even with the limited data provided, short term forecast accuracy is increased with data driven techniques, thus providing hope that these methods could improve wave forecasting over all time horizons.

1 INTRODUCTION

Forecasting wave information in the nearshore zone is essential to society for a variety of operational, research and recreation based activities. Improvements in real time wave height forecasts would increase overall efficiency for operational activities in the nearshore zone such as surveying, and deployment and maintenance of instruments. More accurate forecasts would improve efficiency by reducing the amount of missed or delayed time at sea due to rough conditions. In addition to improving efficiency of operational activities, more accurate wave forecasts in the nearshore zone would improve current research focusing on forecasting wave energy, surf zone currents, and shoreline evolution, all of which incorporate these types of wave models. Recreational activities like surfing and sport fishing, which attract tourism and are vital components of many coastal economies, would also benefit from improved nearshore wave forecasts.

Early attempts to forecast wave information in the nearshore zone began in the 1960's. These attempts relied on equations derived from the underlying physical processes of generation, interference, refraction and shoaling. In the generation and propagation of waves, as with nearly any macroscopic system, friction and other dissipative process play a significant role in the dynamics. However, due to their inherent complexity, these processes can not be understood from first principles. As such, the parameterizations used in dynamical equations are limited in their applicability and therefore cause model-based deterministic forecasting techniques to be prone to error. Despite technological advances in computation and larger simulation models, that include more and more physical processes, forecasting with high accuracy at time scales outside auto-correlated behavior remains a challenge.

Recent advances in instrumentation and storage capabilities have allowed for the collection of enormous data sets from a variety of natural systems. This is particularly true in the ocean where a complex network of buoys and other autonomous instruments have been collecting and recording wave information for decades. These buoy and instrument records contain hourly measurements of wave height, period and direction. These extensive records

offer hope that increased understanding and forecasting accuracy of system dynamics might be possible.

Data driven forecasting using attractor reconstruction and nonlinear time series analysis [19] has been used recently in ecological [3], meteorological [5], and oceanographic systems [1]. Machine learning is another data driven approach to forecasting that has received considerable attention in recent years [20]. These methods train on large data sets of chosen system features to discover patterns that help to forecast target aspects of a system. One of the most widely used machine learning techniques is Artificial Neural Networks, which have been successful in forecasting wave heights in deep water from leading in situ observations [11].

The focus of this study is to explore whether artificial neural networks or nonlinear time series forecasting can achieve more accurate forecasts of wave features in the near shore zone compared with a new state-of-the-art deterministic wave model. The new deterministic wave model created by the National Weather Service(NWS), the Nearshore Wave Prediction System (NWPS), will be used as a basis of forecasting comparison. For both the wave model and the data driven forecasting algorithms, the target data for prediction is nearshore significant wave height, dominant wave period, and mean wave direction collected by an Acoustic Doppler Current Profiler, that was set at 5m depth off the coast of Emerald Isle, North Carolina. Results will compare common error metrics between forecasted and measured data.

2 METHODS

2.1 Near Shore Wave Prediction System

Centralized deterministic models that are run by NOAA 's National Center for Environmental Prediction (NCEP) are impractical for nearshore forecasts due to their low spatial resolution. The demand for high resolution forecasts of coastal process, including wave information, has increased in the past decade. The National Weather Service (NWS) has been working on decentralizing common deterministic wave models in order to provide higher resolution forecasts for nearshore zones [8]. This effort by NWS has produced the Near Shore Wave Prediction System (NWPS) model which is run by regional Weather Forecast Offices(WFO). The NWPS is one part of larger system known as AWIPS-II that seeks to integrate high resolution wave and storm surge models.

The decentralized NWPS model contains two of NOAA's core deterministic models, WAVEWATCH-III and SWAN. These larger models are used to provide the boundary conditions and forcing for NWPS. The NOAA WAVEWATCH-III (NWW3) operational model is based on the Energy or Action Balance Equation [13]. NWW3 consists of one global and five regional wave models: Alaskan Waters(AKW),Western North Atlantic(WNA), North Atlantic Hurricane (NAH), Eastern North Pacific(ENP) and North Pacific Hurricane (NPH). All the regional models obtain hourly boundary data from the global NWW3 model. Models are run every six hours starting with a 6hr hindcast to assure swell continuity and they provide 126hr forecasts with the exception of NAH which provides forecasts only to 72hrs. Models are based on shallow water physics neglecting wave interactions with mean currents [13]. The global model computes wave fields over the globe on a 1.25 deg x 1 deg grid scale of $\approx 7.5km$ and has a minimum depth of 25m. The model has both graphical and binary outputs available. The NWW3 models use a combination of bathymetric and obstruction grids. Model inputs include [13]:

- Winds from the Global Forecast System(GFS), available at 3hr intervals

- For the NAH and NPH models, the GFS wind fields are blended with hourly Geophysical Fluid Dynamics Laboratory (GFDL) hurricane winds when available
- Ice concentrations obtained from NCEP's automated sea ice concentration analysis
- Sea Surface Temperatures as needed in the stability correction for wave growth obtained from Golal Data Assimilation System
- Boundary data for regional models obtained from global model and updated hourly
- models use combination of bathymetric and landmass obstruction grids

SWAN, Simulating Waves Nearshore, is a third generation core deterministic model that forecasts wave processes in coastal regions. SWAN differs from other third generation models in its Eulerian formulation of the Action Balance Equation [12]. SWAN has been tested, validated and implemented and seems to perform well in representing coastal wave processes. To the extent that model parameterizations are found to perform poorly, the model is easily adaptable to include new representations of underlying physical processes [12]. Some of the specific dynamics included in SWAN are:

- Wind input and white capping according to Komen (1984).
- Quadruplet interactions using Discrete Interaction Approximation.
- Joint North Sea Wave Project (JONSAWP) formulation of Hasselman(1985) for bottom friction, and depth induced wave breaking.
- Triad interactions using Lumped Triad Approximation formulation.

Local wind fields are taken from within a given WFO's domain and used to force the wave models. Bathymetric and topographic inputs are taken from Digital Elevation Models (DEM) by the National Geophysical Data Center [8]. Once the local forecasted wind fields and information from the supplementary larger scale models have been received, the NWPS

model is run. The model outputs high resolution nearshore wave information that is post processed into wave field information every three hours out to ninety hours [8]. Frequency spectra (continuous signals) is available for select WFO locations. The wave field information outputted by NWPS contains significant wave height (H_s), peak period (T_p), and mean wave direction($\bar{\theta}$).

The NWPS model has been validated at many of its regional output locations. Validation of the model has been done through statistical analysis comparing model outputs to observations collected from instruments like an Acoustic Doppler Profiler(ADCP) [8]. The NWS is continuously working to extend the NWPS model to new locations throughout the East, West, and Gulf Coasts. Recently the NWPS model was configured for simulations benefitting the WFO in Morehead City, NC. Specifically the NWPS model was run for locations offshore of Emerald Isle, N.C. (Figure 1) and those simulations serve as the backdrop for the model forecast comparisons used in this study.

2.2 In situ Data

In situ wave data was collected with an ADCP deployed at 5m depth by RPS Evans-Hamilton, Inc. The site was chosen to coordinate with one of the NWPS output locations in Emerald Isle NC, near the Morehead City WFO. Wave data was collected by the ADCP from August 27 through November 11, 2015. Forecasting the bulk wave features significant wave height, peak wave period, and swell direction (H_s,T_p,$\bar{\theta}$) (Figure 2) as taken from the ADCP record was the aim of the wave and data driven models.

2.3 Nonlinear Time Series Analysis Technique

Current wind generated wave models like NWPS, NWW3 and SWAN are computationally intensive, requiring large processing power to run real time forecasts. In addition to being computationally expensive, these deterministic models require parametrization of most physical processes such as friction and other dissipative processes, which can not be

Figure 1: A map of the Emerald Isle N.C. study site. (Red box) represents the nearest 5m NWPS output grid that was used for comparison with the (Yellow pin) 5m nearshore ADCP

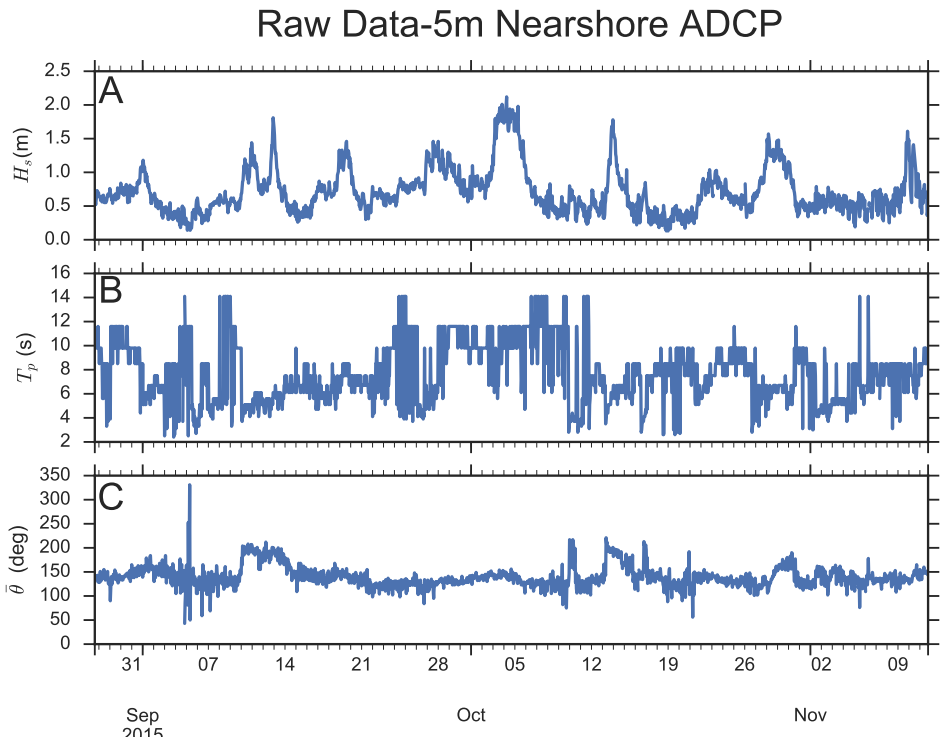

Figure 2: A plot of the raw (A)H_s significant wave height , (B) T_p peak wave period , and (C) $\overline{\theta}$ mean wave direction collected from the 5m Nearshore ADCP at the Emerald Isle study site.

derived from first principles. These parameterizations are often only approximately correct and even then, only for a limited range of conditions. Furthermore, inherent strong nonlinearities make long term forecasts (greater than 24 hours) unreliable. In nonlinear ecological systems [2] , there has been some recent success in using data driven models for forecasting, specifically nonlinear time series forecasting.

Nonlinear time series analysis (NLTSA) as used for forecasting is based on attractor reconstruction. Specifically, Taken's Theorem [4] states that a time series of a single system variable can be embedded in a multi-dimensional space and the trajectories within the embedded space are topologically similar to the attractor of the underlying system. The embedding is performed by choosing a time lag and embedding dimension to construct positions in the embedded space as,

$$\overrightarrow{y_t(x)} = (x_t, x_{t-\tau}, \ldots, x_{t-(m-1)\tau}) \tag{1}$$

where x is the time series, t is the time, τ is the lag value and m is the embedding dimension of the reconstructed phase space. As the dynamical system steps through time, a trajectory is traced out by these position vectors in the reconstructed phase space. Systems that are governed by smooth attractors have correlated neighboring trajectories and information from those neighbor trajectories can be used to make predictions of future evolution [1] [2] [6]. To test the predictability of a given time series, the series is split into a training and testing set. The training set is embedded to reconstruct the system attractor. Next, individual embedded points from the test set are chosen for forecasting. Near neighbors in the embedded space to the testing point in question, are used to generate predictions. The future trajectory of the near neighbor points is the predicted future evolution of a given test point. The prediction performance is calculated over a range of possible near neighbors in the phase space to choose the amount of neighbors that yields optimum predictability.

A similar extended NLTSA approach is used in this study in order to generate nearshore wave feature forecasts. The 1D NLTSA method is extended to include historical spatial

wave data collected from surrounding offshore National Data Buoy Center (NDBC) Buoys. (http://www.ndbc.noaa.gov/) The four spatial wave datasets from NDBC include significant wave height (H_s), peak period (T_P) and mean wave direction ($\bar{\theta}$) from: Onslow Bay (41159), Diamond Shoals (41025), W. Bermuda(41048) and NE Bahamas (41047) (Figure 3).These four particular datasets were selected out of this region's immense NDBC historical data base due to the fact they had the least amount of missing data corresponding to the wave features collected from the ADCP, and not necessarily due their proximity to the location of that ADCP. The use of closer buoy locations like South Hatteras (41002) or ILM-2 (41110) may have been a better choice than the distant W. Bermuda(41048) and NE Bahamas (41047) locations, had they not lacked extended periods of data from August through November.

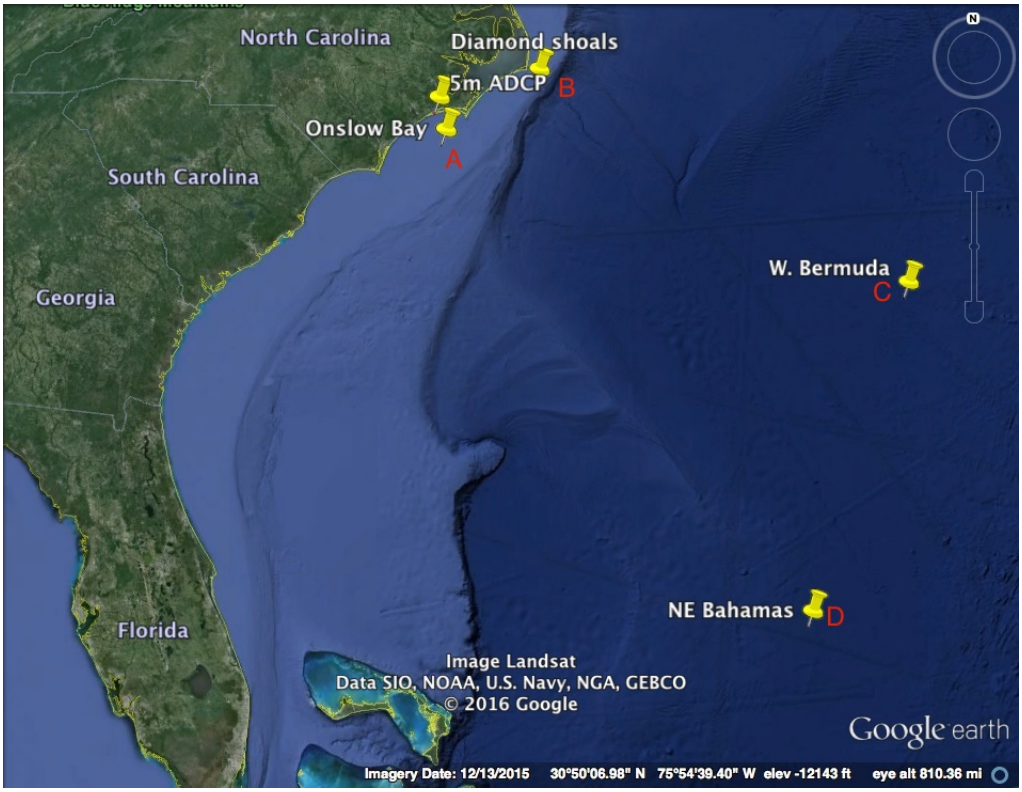

Figure 3: A map of NOAA'S National Data Buoy Center, (A) Onslow Bay Buoy 41159 (B) Diamond Shoals Buoy 41025 (C)West Bermuda Buoy 41048 (D)Northeast Bahamas Buoy 41047

The 1D NLTSA equation (1) is extended for the four buoy's to construct a feature set to be used for attractor reconstruction below,

$$\overrightarrow{Z_t(x^i)} = \sum_{i=1}^{N} (x_t^i, x_{t-\tau}^i, \ldots, x_{t-(m-1)\tau}^i) \tag{2}$$

where i is a particular buoy feature, N is the total number of buoys, x is the time series, t is the time, τ is the lag value and m is the embedding dimension of the reconstructed phase space. Once the reconstructed attractor for the surrounding buoys is determined, each point in the phase space is then mapped to corresponding point on the ADCP wave feature time series. ADCP wave forecasts are then generated by time pairing the trajectories of near neighbors in the buoy feature phase space to point of interest on the ADCP time series. Similarly prediction performance is calculated over various numbers of near neighbors in the reconstructed space in order to determine optimum predictability.

.

2.4 Artificial Neural Network

An Artificial Neural Network (ANN) is another type of data driven model that does not require knowledge of the underlying physical principles in order to make predictions. ANN's are a form of machine learning (artificial intielligence) loosely based on the network structure of neurons in the human brain (Figure 4) [18]. In a training phase that sets up for a specific network, input data is sent to neurons in the input layer. With known outputs for a given input, the goal of the training phase is to determine the weights of each neuron to neuron connection across layers. Training is complete once a desired input-output relationship is reached [10]. The particular algorithm used to adjust weights during training can vary, but the goal is to reduce the global error E between the network predictions and the actual observations as defined below [10]

$$E = \frac{1}{P} \sum_{i=1}^{P} E_i \tag{3}$$

where P is the number of training patterns and E_i is the error for that training algorithm given as,

$$E_i = \frac{1}{2}\sum_{k=0}^{N}(o_k - t_k)^2 \tag{4}$$

where N is the total number of training features associated with a given pattern, o_k is the network output and t_k is the target output for the same kth output.

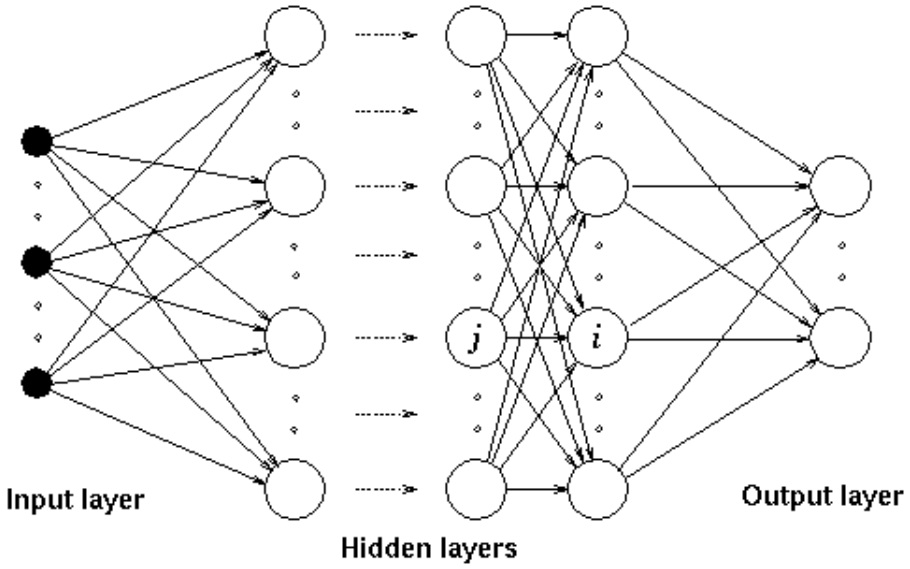

Figure 4: A schematic of a common Neural Network Architecture. Each layer is comprised of a series of neurons that are connected to neurons in other layers.

ANN's have been used to forecast wave data such as: significant wave height (H_S), mean wave period($\bar{T}$), and wave energy Flux. M.C Deo(1990) used ANN's to forecast significant wave height and mean wave period off the coast of India using nearby wind speed data. Additional research using ANN's had the objective to make real time wave forecasts with varying lead times from observed waves at a given site [10]. In these studies the networks used a three layer structure that was forward fed [10]. The networks were trained on 80% of the data using a variety of algorithms: Back propagation, Cascade Correlation, Conjugate Gradient.

Previous ANN studies found that satisfactory results can be achieved with a proper

choice of training algorithms [10]. Results showed that a properly trained network can achieve 77 % accuracy in predictions of wave properties in deep water with large sample sizes and prediction areas [10]. These studies also concluded that due to the complexity of the underlying phenomena highly accurate forecasts may not be possible.

More recent work on wave forecasting has looked at direct comparisons between a physics based model and data driven ANN models. Tests were run over thirteen data sets from the NDBC that span the Atlantic, Pacific and Gulf regions. The data sets contained hourly values of H_s, T_p, and $\overline{\theta}$ from 2007-2010. The ANN in the study contained three hidden layers, was trained using the back propagation algorithm, and made predictions on H_s, T_p and, wave energy flux. The study showed that increasing the number of hidden layers within the network yielded no increase in prediction accuracy and only increased computation time [7] .

From the forecasting studies using ANN in comparison with physical wave models [7], errors from the different forecasting approaches showed different properties. Errors associated with the physics based model forecasts increased gradually over a period of two days, while errors from the ANN models were initially smaller and increased very rapidly [7]. For shallow coastal sites the ANN model worked well over short forecast horizons (1-5hr). The physics based model preformed better over longer forecast horizons (+6hr).

For this study, an ANN was constructed in Python 2.7 using the Skflow Neural Network Package. Skflow serves as simplified interface for Google's Tensor Flow Open Source Software Library used for deep learning and mimics the commonly used Python package ScikitLearn. This study's ANN was trained on the same four lagged and embedded NDBC's historical buoy time series used previously in the NLTSA. These time buoy time series contained the same time paried hourly measurements of H_S, T_P, and $\bar{\theta}$. Network training on the buoy records was used in order to make predictions on the same Emerald Isle ADCP used for the other model validations discussed in the previous sections. The ANN was trained using Skflow's Regressor Algorithm. This Regressor Algorithm uses a multi layered perception

modeled after Linear Regression.

Linear Regression is the simplest form of regression and models systems using a linear combination of weighted features to produce one output. Consider a single layer ANN with one neuron that uses a linear activation function to predict the output y_i ,

$$y_i = h(x_i, w) = w^t x_i \tag{5}$$

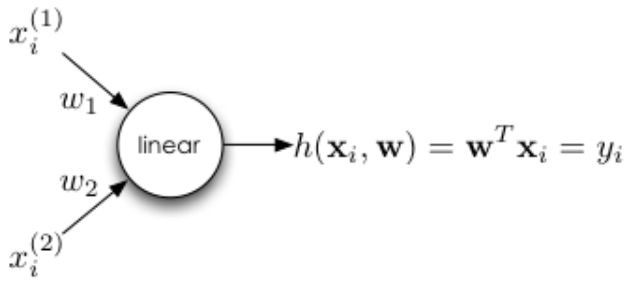

Figure 5:

In order to find the weights that provide the best fit for the training data, the least square error, $L(w)$, must be calculated over the data set [18] [17]:

$$L(w) = \sum_i (h(x_i, w) - y_i)^2 \tag{6}$$

Weights are found by minimizing $L(w)$ using back propagation and gradient descent. Back propagation is the simplest but slowest training algorithm for an ANN. It seeks to minimize E using a gradient descent approach. In a gradient descent approach weights and biases are adjusted by moving small increments in the direction negative of the gradient of the global error during each iteration until a specified number of iterations is completed [11]. This is first done by computing the change in error with respect to a particular weight connecting layer j to layer k, $w_{j\to k}$, as [18]

$$\frac{\partial}{\partial w_{j\to k}} L(w) = \frac{\partial}{\partial w_{j\to k}} \sum_i (h(x_i, w) - y_i)^2 \tag{7}$$

$$= \sum_i 2(h(x_i, w) - y_i) \frac{\partial}{\partial w_{j \to k}} h(x_i, w) \tag{8}$$

Now the gradient of the network function can be calculated with respect to the weight in question [18]. For this single layer example our network function is $h(x_i, w) = w_1 x_i^1 + w_2 x_i^2$. The gradient with respect to w_1 is x_1 and with respect to w_2 is just x_2. The full gradient is

$$\nabla_w L(w) = \left(\frac{\partial L(w)}{\partial w_1}, \frac{\partial L(w)}{\partial w_2} \right) = \left(\sum_i 2x_i^1 h(x_i, w), \sum_i 2x_i^2 h(x_i, w) \right) \tag{9}$$

The weights are then updated using gradient descent with a carefully chosen step size η.

$$w = w - \eta \nabla_w L(w) \tag{10}$$

After a set amount of iterations the weights are used to define network connections between neurons [18].

2.5 Performance Metrics for Error Analysis

In order to quantify the differences between the forecasts and the in situ observations each forecast is time paired to its corresponding buoy measurements(H_s, T_p, and $\overline{\theta}$). The error metrics used for performance score calculations are defined below according to Hanson (2009) [9]. For n values of buoy measurements m_i and n forecasts f_i the error metrics include the bias, which calculates the average error between the forecasted values of interest and their corresponding measurements,

$$b = \frac{1}{n} \sum_{i=1}^{n} (f_i - m_i). \tag{11}$$

The root-mean-square(RMS) error is used to measure the differences between the forecasts values and their time paired measurements by,

$$E_{RMS} = \sqrt{\frac{\sum_{i=1}^{n}(f_i - m_i)^2}{n}}. \tag{12}$$

Additional error metrics include the Scatter Index,

$$SI = \frac{\sigma_d}{\overline{m}} \tag{13}$$

where σ_d known as the standard deviation of difference is given by,

$$\sigma_d = \sqrt{\frac{\sum_i (f_i - m_i - b)^2}{n-1}}. \tag{14}$$

For directional data, the directional difference $\Delta\theta = |\theta_f - \theta_m|$ is first used to calculate S and C. Where S and C are defined as,

$$S = \sum_{i=1}^{n} \sin(\Delta\theta_i) \tag{15}$$

and

$$C = \sum_{i=1}^{n} \cos(\Delta\theta_i). \tag{16}$$

S and C are then used to calculate the angular bias. Angular bias, as defined below, is used to quantify the directional differences ($\Delta\theta$) between the forecast values of interest and the actual measurements.

$$b_a = \begin{cases} \arctan(\frac{S}{C}) & \text{for } S > 0, C > 0; \\ \arctan(\frac{S}{C}) + \pi & \text{for } C < 0; \text{ and} \\ \arctan(\frac{S}{C}) + 2\pi & \text{for } S < 0, C > 0 \end{cases} \tag{17}$$

The last error metric, Circular Correlation or *circor*,

$$circor = \frac{\sum_{i=1}^{n} \sin(\theta_m - \bar{\theta_m}) \sin(\theta_f - \bar{\theta_f})}{\sqrt{\sum_{i=1}^{n} [\sin(\theta_m - \bar{\theta_m})]^2 \sum_{i=1}^{2} [sin(\theta_f - \bar{\theta_f}]^2}} \tag{18}$$

is used to quantify the overall directional correlation between the forecasts and the measurement.

These error metrics were used for performance score calculations. This is done by normalizing wave component metics to mean quantities. The non dimensional performance scores range from 0 to 1 and includes the RMS error performance,

$$\hat{RMS} = \left(1 - \frac{E_{RMS}}{m_{RMS}}\right) \tag{19}$$

where the root-mean-square of the measurements is given by,

$$m_{RMS} = \sqrt{\frac{\sum m^2}{n}}. \tag{20}$$

The bias performance calculation is done by subtracting the ratio of the absolute value of the $\hat{b}$ to the m_{RMS} from 1.

$$\hat{b} = \left(1 - \frac{|b|}{m_{RMS}}\right) \tag{21}$$

Similarly, the Scatter index performance is calculated by subtracting the Scatter Index error calculation defined above from 1.

$$\hat{SI} = (1 - SI). \tag{22}$$

For the directional data, angular bias performance is calculated similarly to the bias performance calculation by subtracting the ratio of the absolute values of the anugular bias which will be in deg over 180

$$\hat{b_a} = \left(1 - \frac{|b_a|}{180}\right) \tag{23}$$

and circular correlation performance is already normalized so

$$\hat{circor} = circor. \tag{24}$$

3 RESULTS

In order to test the performance of the NWPS model, the model was run over a 75 day time frame spanning from August 27th through November 11th 2015. The model outputs were then time paired to the corresponding ADCP observations for comparison.(Figure 6) From the comparison of the NWPS model to the raw data, NWPS appeared to do well in forecasting the significant wave height, and the mean wave direction in low energy intervals between storm events. The model appeared to under predict the significant wave height in both magnitude and duration during storm events. This is possibly a result of the corresponding under forecasted wave periods. Incorrect wave periods would make shoaling and refraction related changes to significant wave height very difficult to capture. The model forecasts also frequently missed the significant but small interval directional variations. NWPS did not appear to do well in forecasting the peak wave period. This was likely due to Peak period being too crude of measure for the wave field due to the rapid changes that occurred in peak period as two separate wave trains interacted. For example, when a longer period swell is on the decline while a shorter period swell is also present, the peak period can rapidly decrease as the longer period swell loses energy and the shorter period swell becomes dominant.

Before generating forecasts with the NLTSA technique, time lag values were needed to embed the system in the reconstructed phase space. Lag values for the buoy features were selected by calculating the average mutual information at a series of time lags. The first minimum value in the average mutual information between the four buoys was chosen for the lag values to use for the feature and target training and testing sets [19]. Lag values of 100, 75, and 50 hours were selected respectively(Figure7) for significant wave height, peak

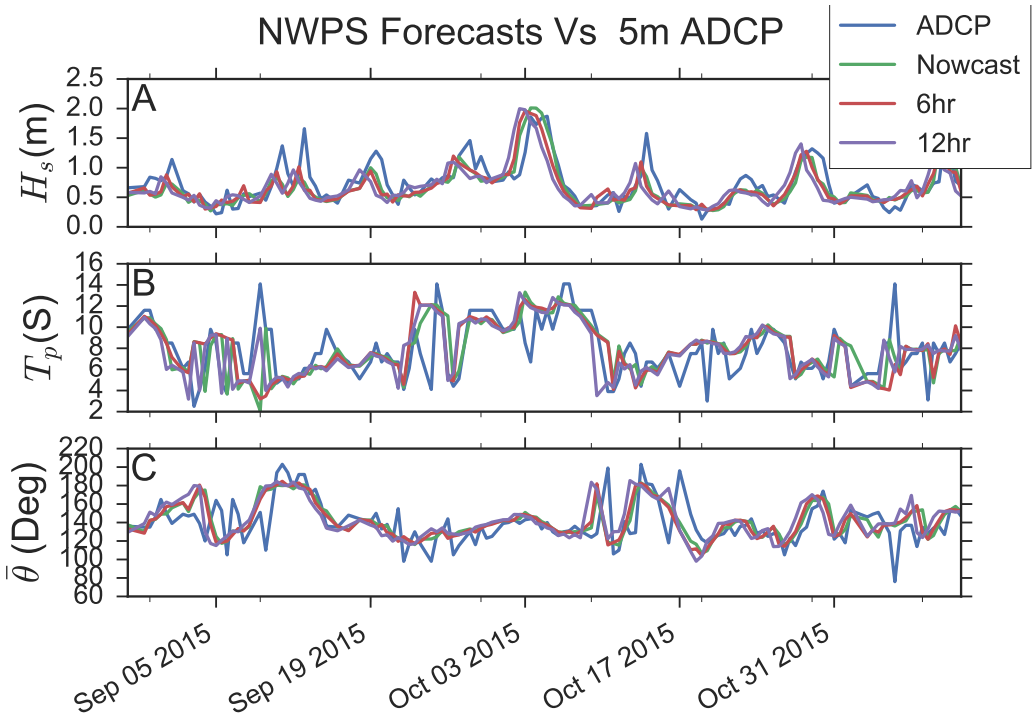

Figure 6: A plot of the dynamical wave model (NWPS) against the in-situ ADCP (blue)data for (A) H_s significant wave height, (B) T_p peak period and (C)$\bar{\theta}$ mean wave direction forecasts. Dynamical Model Outputs: Nowcast(green) , 6 hour forecast(red),12 hour forecast(violet)

period, and mean wave direction. The embedding dimension was determined by exploring the correlation coefficient R^2 from resulting forecasts at various time scales as function for different number of embedding dimensions. An embedding dimension of 4 provided the most accurate forecasts for the nonlinear forecasting technique (Figure 8). After exploring the lag values and embedding dimensions for the attractor reconstruction, forecasts were generated and prediction performance is calculated over various numbers of Near Neighbors in the reconstructed space (Figure 9). Twenty five Near Neighbors were chosen for this analysis to generate accurate medium horizon (between 12 and 24hrs) ADCP forecasts (Figure 9).

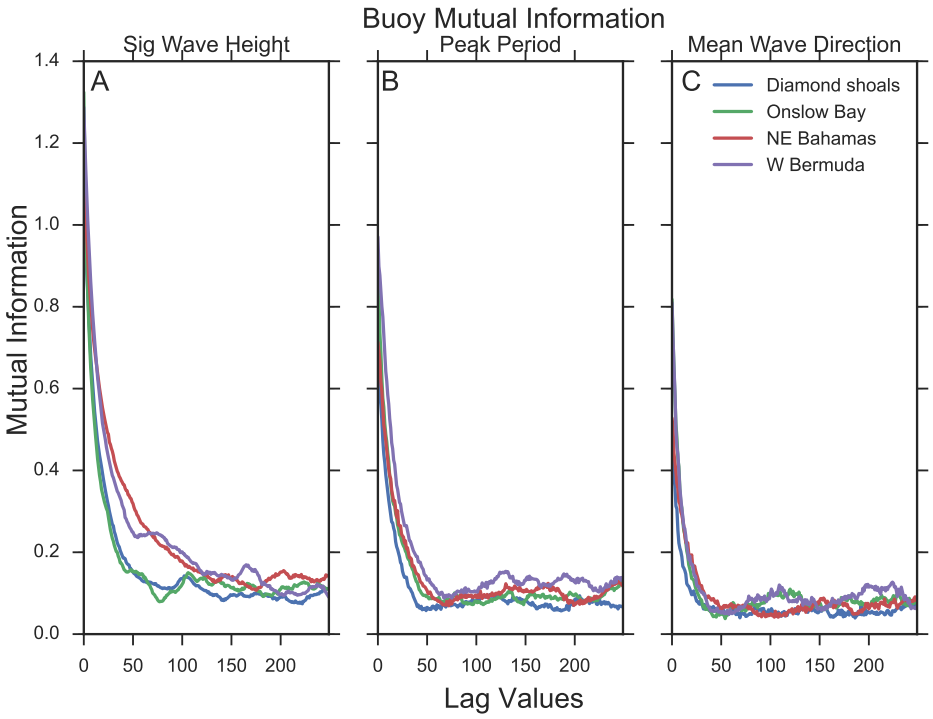

Figure 7: A plot of the average mutual information between the Diamond shoals(blue), Onslow Bay(green) NE Bahamas(Red) and W Bermuda(violet) buoys for (A)significant wave height,(B) peak period and (C)mean wave direction

The forecasts generated by the NLTSA technique spanned 15 days from October 28th to November 11th. The peak period testing window was limited due to missing peak period data from the historical buoy records(Figure 10). The NLTSA technique does not forecast wave heights and peak periods well even at short forecast horizons. This technique under forecasts

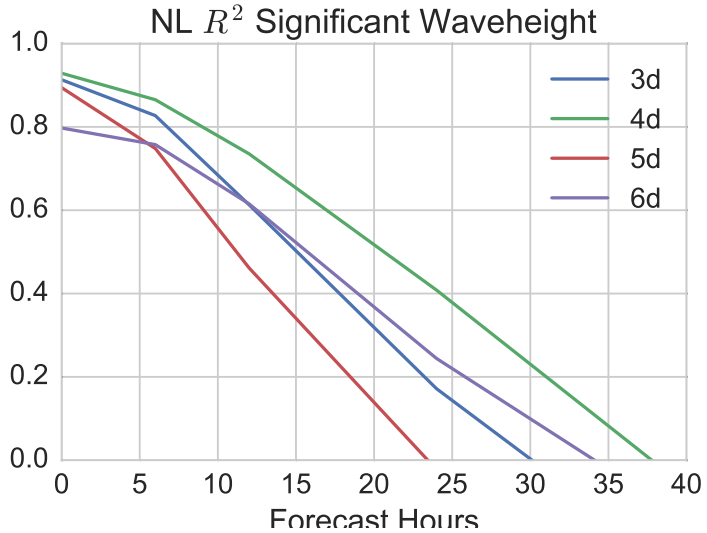

Figure 8: A plot of the correlation coefficient R^2 for significant wave height forecasts for, 3(blue) 4 (green),5(red), and 6(violet) embedding dimensions.

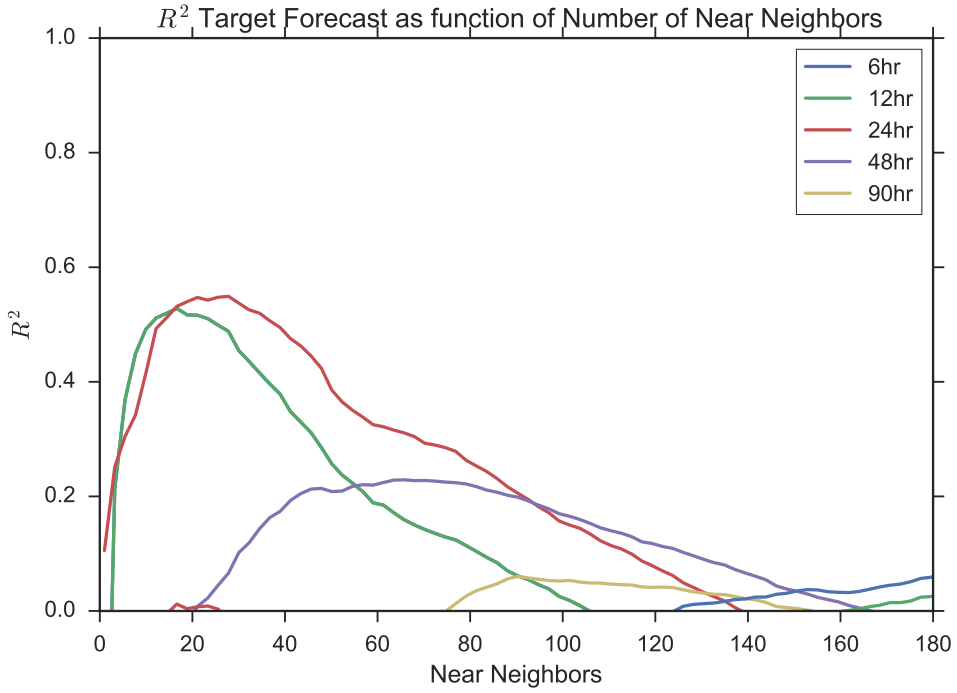

Figure 9: A plot of the R^2 Correlation Coefficient for the Target Forecasts as a function of the Number of Near Neighbors.

significant wave height when under forecasting periods and similarly tends to over forecast wave height when over forecasting corresponding wave periods. Low forecasted wave heights relative to the target occurred during the storm event (40hrs to100hrs) (Figure10). Poor peak period forecasting could again be due to variations in peak period as numerous swell events coincide at the ADCP.(Figure10 B) The nonlinear technique faired well in forecasting the mean swell direction in the initial testing period when there was little variation. (Figure10 C) NLTSA loses forecasting skill when larger short interval variations appear between 100 hrs to 200 hrs.

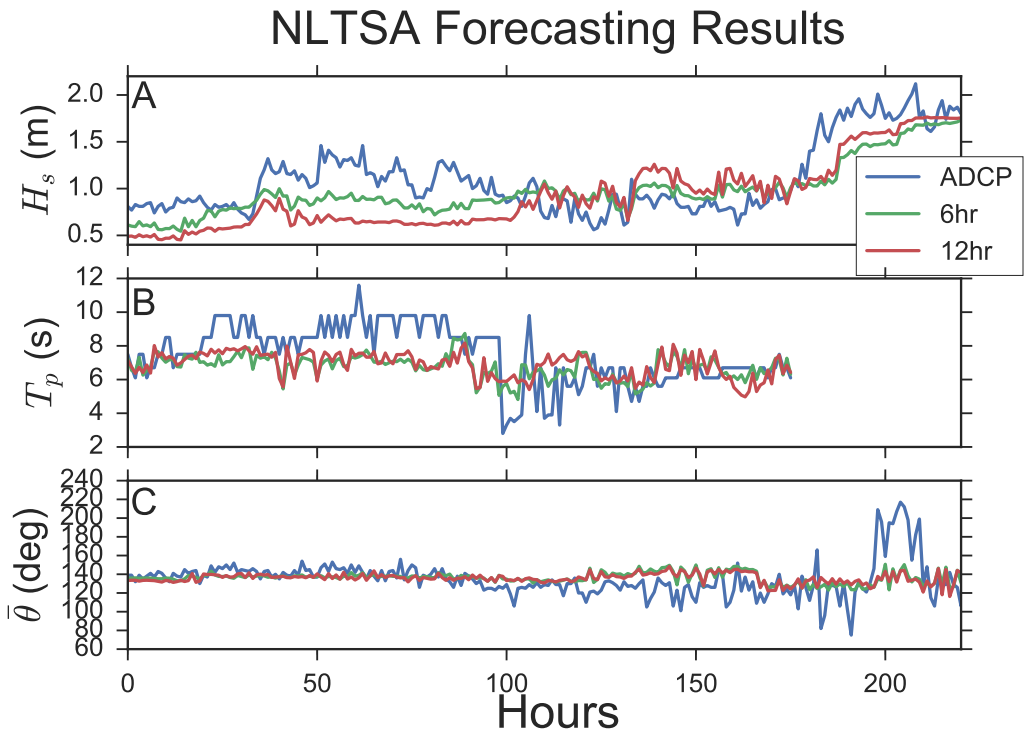

Figure 10: A plot of the Nonlinear forecasting model (NLTSA) against the in-situ ADCP (blue-ADCP)data for A) H_s significant wave height, (B) T_p peak period and (C) $\bar{\theta}$ mean wave direction forecasts. 6 hour forecast(green),12 hour forecast(red)

Next the ANN was used to generate predictions on the same data record from the ADCP that was used for testing the NLTSA. The peak period testing window was again limited due to same missing peak period data from the historical buoy records(Figure 11 B). Before generating forecasts, an optimum network architecture was determined by exploring different

combinations of the number of hidden layers and the number of hidden units in each of the layers. The optimum network configuration chosen for this study consisted of three hidden layers each with 75 hidden units(Figure 12). The ANN does well in forecasting the significant wave heights out to 6 hrs(Figure 11). These predictions show similar under forecasting during the same storm event seen with the NLTSA. The mean wave direction ANN forecasts have more variation then the NLTSA, but still appear to capture the longer trends in swell direction. The ANN does a very poor job in forecasting the wave period with significant under forecasting during the storm event. Once again this is likely due to variations in peak periods measured by eh ADCP during coinciding swell events.

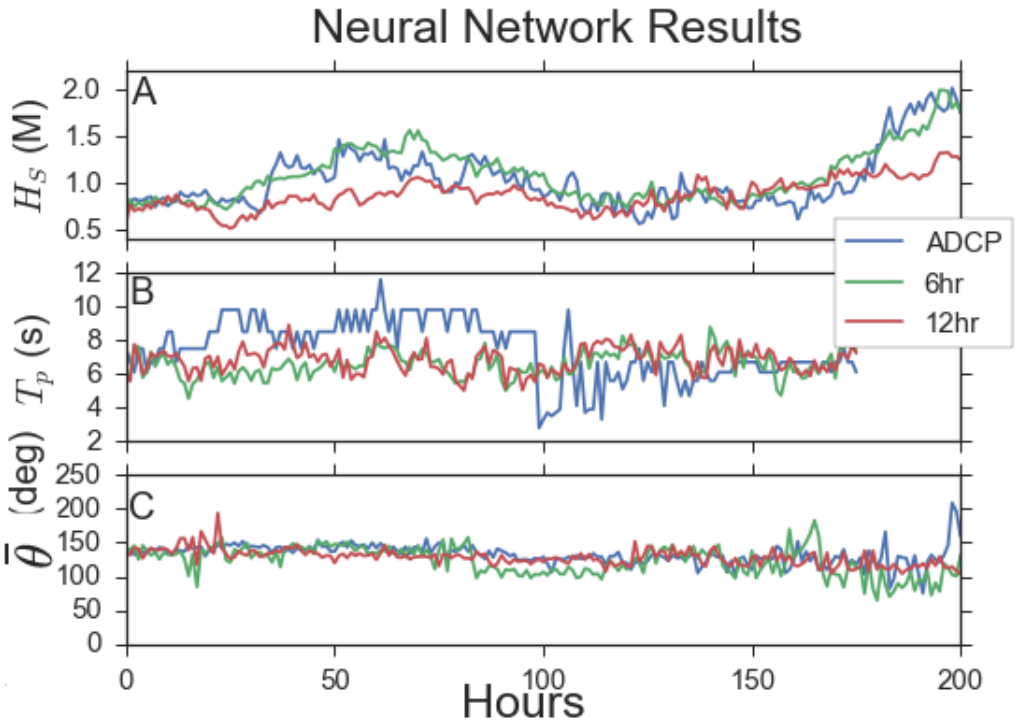

Figure 11: A plot of the Neural Network forecasting model (ANN) against the in-situ ADCP (blue-ADCP)data for (A) H_s significant wave height, (B) T_p peak period and (C) $\bar{\theta}$ mean wave direction forecasts. 6 hour forecast(green),12 hour forecast(red)

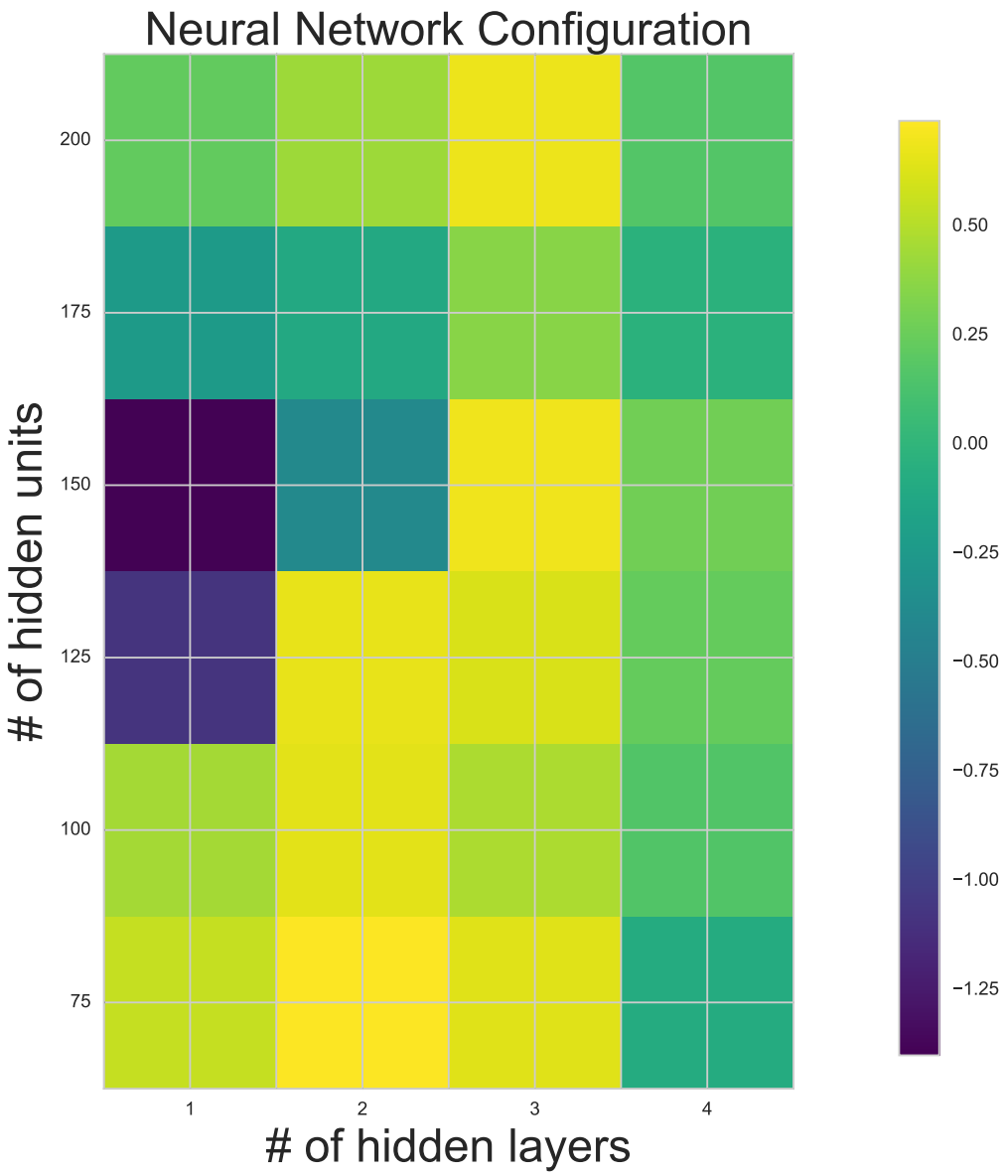

Figure 12: A heat map that shows the Coefficient of Determination for combinations of Differing numbers of hidden layers and numbers of hidden units in each layer.

To explore the capability of the ANN in forecasting wave information from a more offshore location, the technique was used to forecast a slightly longer Historical NDBC data set from the buoy at Onslow Bay (41159). Wave data inputs to the ANN for this scheme were taken from 4 offshore buoys: South Hatteras (41002), Diamond Shoals (41025), NE Bahamas (41047) and Nantucket (44008). The purpose of this forecasting exercise was to explore whether a longer data record and less dynamic oceanographic environment, as found further offshore, would yield higher forecast skill. Figure 13 shows that with more training data the ANN was in fact able to have improved accuracy in forecasting both significant wave height and peak periods at the Onslow Bay buoy.

Despite the improved accuracy over the nearshore ADCP this offshore buoy ANN resulted in similar error trends. The ANN again under forecasted significant wave heights when under forecasting the corresponding peak periods during high energy storm events. The improved accuracy however provides hope that with more extensive data records this technique could be used to make better forecasts and in the case of this particular buoy, which was decommissioned, ANN forecasts might provide a means to fill in missing data for historical NDBC buoy records.

3.1 Performance Scores

The performance metrics (Figure 14) for forecasting significant wave height show that the NWPS model is better at forecasting on longer horizons with the forecasting ability decreasing over time. Figure 14 (A), (C) and (D) show that both NLTSA and the ANN achieved peak forecasting skill at 6 hr forecast horizons and their forecast performance declined over time. The data driven models and the deterministic model tend to have high bias scores which indicates that forecasting is not skewed to over or under prediction (Figure14 B).

Other performance metrics (Figure 15) show that all of the models have difficultly forecasting peak period. Despite all three of the models low performance in peak period forecasting, the NLTSA technique does the best at short horizons while the NWPS model does best

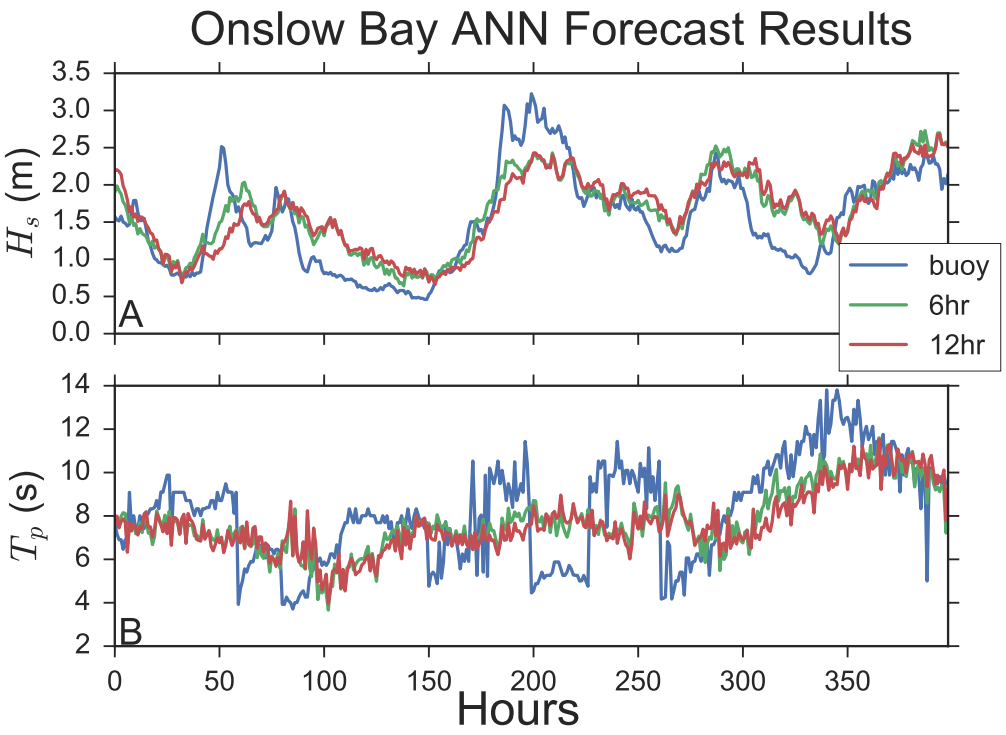

Figure 13: A plot of the Neural Network model (ANN) results against the Historical NDBC Onslow bay buoy (blue- Buoy) data for (A) H_s significant wave height , and (B) T_p peak period . 6 hour forecast(green),12 hour forecast(red)

for 12 hr forecasts. Forecasting ability decreases after 6 hours for the NLTSA and decreases after 12 hrs for the NWPS model. Figure 15 (A) shows that both data driven systems have poor forecast skill. All models tend to have high bias scores which indicates that they each over predict just as much as they under predict (Figure 15 B). The root-mean-square error and scatter index (Figure 14 and Figure 15 (C) and (D)) show the trend of model forecasts initially being well correlated to the measured data with correlation trends declining with increasing forecast horizons .

Wave direction performance scores (Figure 16) show that when models have low angular bias there tends to be anti-correlation. Anti-correlation results from the measured direction being greater than the mean of the measured direction ,while the corresponding forecast direction is less than the mean forecast direction. On the contrary, low angular bias tend to have directional forecasts and measurements that are well correlated, which can be inferred when the measured forecast is greater than the mean, and the corresponding forecast will also be greater than the mean of the forecasts.

Additional performance scores were calculated for the offshore buoy ANN forecasts. The offshore buoy ANN showed improved accuracy in significant wave forecasts at horizons greater than 6 hrs compared with the nearshore ADCP forecasts (Figure 17) with the increase in training data. Peak period performances (Figure 18) also saw dramatic improvements over the nearshore ADCP wave forecasts. The offshore buoy ANN also differed by showing little to no gradual decrease in bias ($\hat{b}$) for both significant wave heights and peak periods at longer forecasting horizons (Figure 17 B). The $R\hat{MS}_{error}$ and $\hat{SI}$ (Figure 17 and 18 C and D) show similar trends in gradual decrease forecasting correlation with increasing forecast horizons.

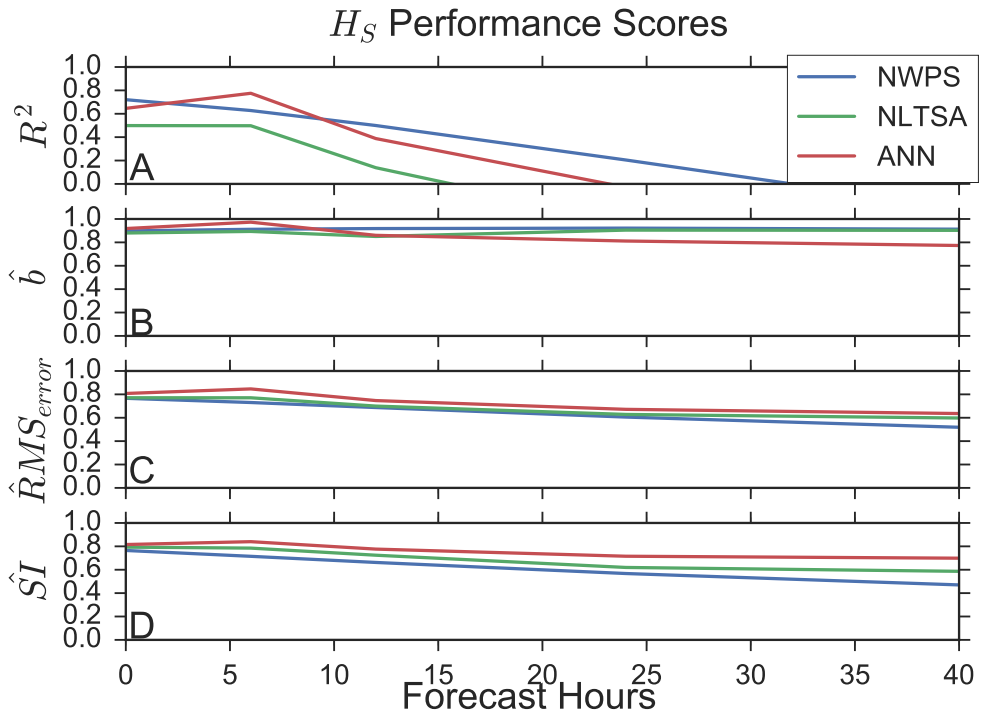

Figure 14: A plot of the (A) coefficient of determination (R^2),(B) $\hat{b}$, (C)$RM\hat{S}_{error}$, and (D)Scatter Index $\hat{SI}$ in wave height forecasts versus forecast hours for the dynamical wave model (NWPS - blue),the nonlinear forecasting (NLTSA - green), and the neural network (ANN-red) for the nearshore ADCP .

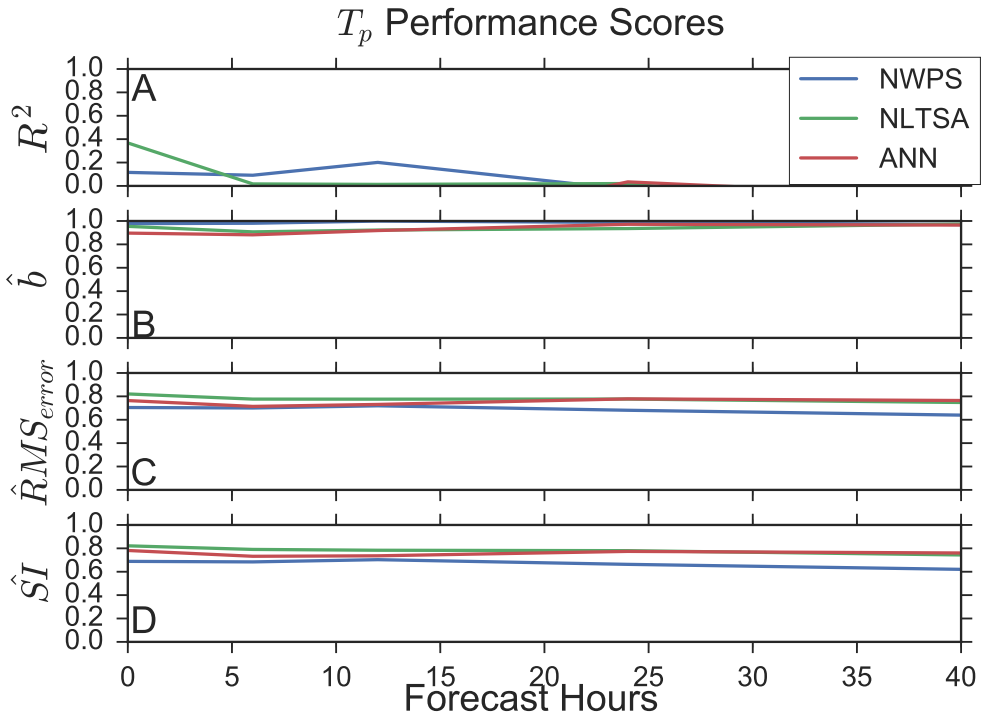

Figure 15: A plot of the (A) coefficient of determination (R^2),(B) $\hat{b}$, (C)$RM\hat{S}_{error}$, and (D)Scatter Index $\hat{SI}$ in peak period wave forecasts versus forecast hours for the dynamical wave model (NWPS - blue), the nonlinear forecasting (NLTSA - green) and the neural network (ANN-red) for the nearshore ADCP.

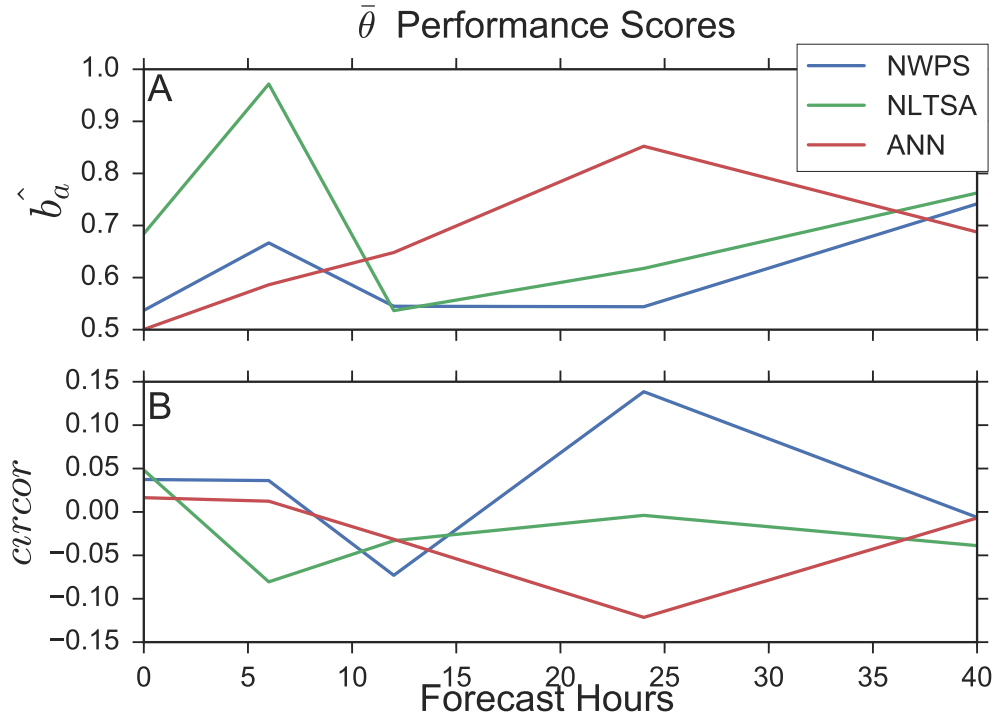

Figure 16: A plot of the (A) Angular Bias $\hat{b}$ and (B) Circular Correlation *circor* in mean wave direction $\bar{\theta}$ forecasts versus forecast hours for the dynamical wave model (NWPS - blue), nonlinear forecasting (NLTSA - green) and neural network (ANN-red) techniques.

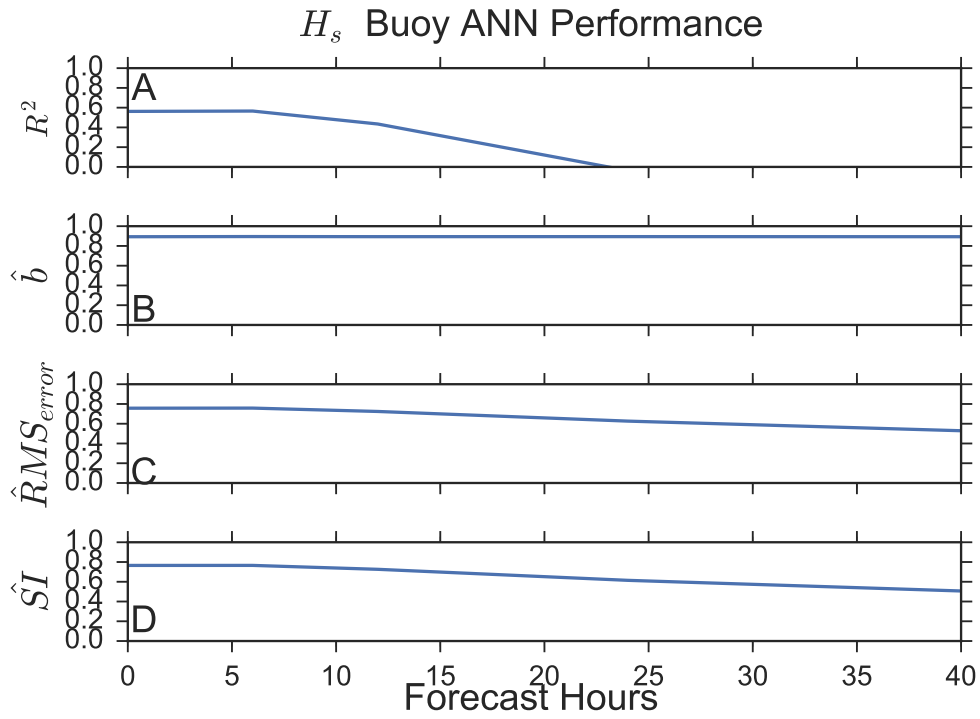

Figure 17: A plot of the (A) coefficient of determination R^2,(B) $\hat{b}$, (C)$RM\hat{S}_{error}$, and (D)scatter index $\hat{SI}$ in significant wave height H_s forecasts versus forecast hours for the Onslow Bay historical buoy record.

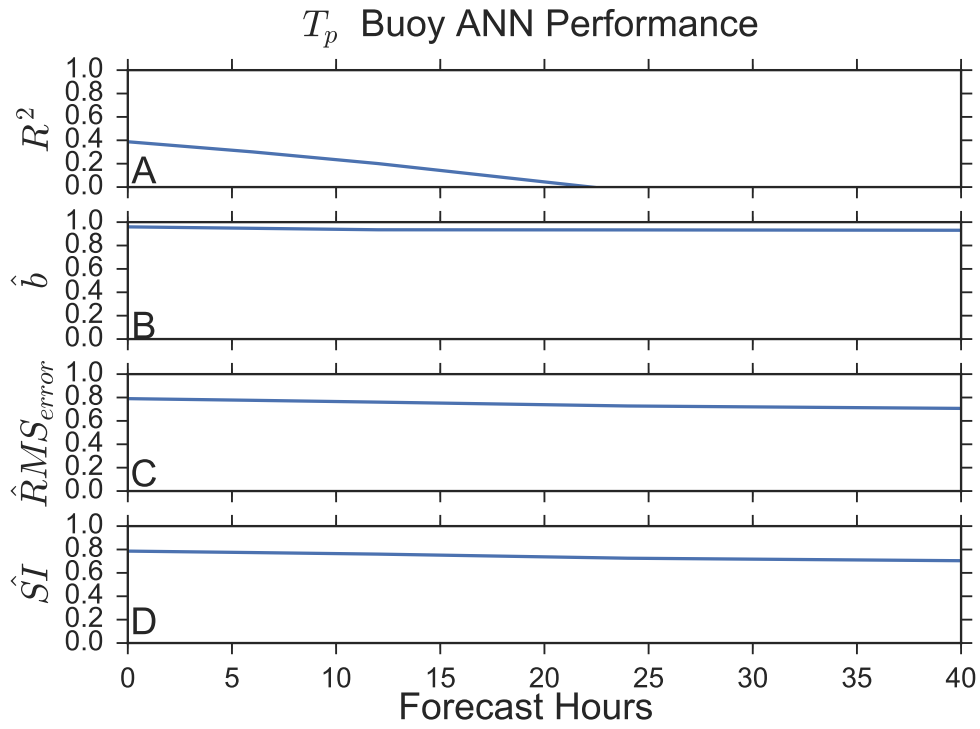

Figure 18: A plot of the (A) coefficient of determination R^2 ,(B) $\hat{b}$, (C)$R\hat{MS}_{error}$, and (D)scatter index $\hat{SI}$ in peak period T_p forecasts versus forecast hours for the Onslow Bay historical buoy record.

4 CONCLUSION

The NLTSA and ANN forecasting techniques were constrained by a fairly small amount of training data. Both techniques benefit greatly with large amounts of data, to help the attractor reconstruction achieve high density for NLTSA and to help constrain network weights for the ANN. Many of the previous ANN studies used data sets that spanned from a few years up to a decade. The ADCP and NDBC buoy data used for this study consisted of only 75 days of hourly measurements. The first 60 days (80% of the data) were used for training to complete the data driven model construction that was then used to make predictions on the small window of the last 15 days. Additional caveats for the data driven models include the parameters used within the models. These parameters are not able to be derived from first principles. For example, the number of layers and neurons selected for the optimum ANN architecture, and the embedding dimension used for attractor reconstruction in the NLTSA are determined by parameter exploration. The values used in this thesis may not be applicable for making predictions at different locations and may not even be the parameters that yield best performance if more data were used in this study.

Despite these short comings, the ANN results show superiority in forecasting significant wave height at short forecast horizons (near 6 hrs) relative to the NWPS wave model. Model performance for both data driven techniques show rapid declines in accuracy after 6 hrs. The NLTSA and ANN forecasts have very low forecast skill after 12 and 24 hrs respectively. The NWPS model out performs both the NLTSA and ANN data driven models in forecasting wave height at longer horizons. These peak forecast horizons and error trends coincide with results of past studies [7]. All three of the models showed very poor peak period forecasting performance, which as discussed, is likely due to the rapidly changing nature of the peak period record.

Future work using these data driven models, should include gathering larger data sets for training. In addition to getting more data, these data driven techniques should be tested at more locations both in the the near shore and deep oceans. Another avenue for future work,

that could be used to improve forecast accuracy, is the exploration of other ANN training algorithms. Previous studies have shown that highly accurate forecasts can be achieved with the proper selection of training algorithm. [10] The NLTSA and ANN models could also see improved forecast skill by training on NDBC's raw continuos wave spectra rather than using bulk wave features. The results presented here do provide hope that data driven forecasting techniques can improve forecasts of wave information in the nearshore zone.

5 APPENDIX: ANN Forecasting code with skflow

First we must load the necessary packages for mathematical functions, running the Neural Network and for creating nice looking plots.

```
# numpy  and pandas for math functions and array manipulations
import numpy as np
import pandas as pd
from pandas import set_option
set_option('display.max_rows',5)
# matplotlib and seaborn for nice looking plots
import matplotlib.pyplot as plt
import os
import seaborn as sns
import datetime as dt
sns.set_style('whitegrid')
cmap='viridis'
%matplotlib inline
# Nueral network packages
import sklearn as sk
import sys
import skflow
from sklearn import cross_validation
from sklearn import preprocessing
from sklearn.preprocessing import StandardScaler
from sklearn.metrics import accuracy_score
# function file for calcutiong error analysis and
# performance scores
import Wavestt as ws
```

Now that the necessary packages are loaded the next step is to load the Features (Buoy data) and the Target (ADCP) data, which will be used for training and testing. Here we are importing the cleaned time paired data from NOAA NDBC Buoys to used as the features

in the NN to predict the significant wave heights at the Emerald Isle nearshore ADCP.

```
#loading  the   pre split features  for training and testing
f_train=pd.read_pickle('Xtrain.pkl')
f_test=pd.read_pickle('Xtest.pkl')
#loading  the   pre split targets  for training and testing
t_train=pd.read_pickle('ytrain.pkl')
t_test=pd.read_pickle('ytest.pkl')
```

Now that the data is loaded we can define the split feature and target datasets, and move onto building the three layered network. Each layer will contain 75 units and will be trained on skflow's regressor algorithm. Once the regressor is built the network is fit to the feature and target test sets After fitting the predictions are generated and R^2 values are calculated.

```
score=[]
# Split dataset into train / test
X_train, X_test= f_train.values,f_test.values
y_train y_test = t_train[0].values,t_test[0].values
# Build 3 layer fully connected DNN
#with 75, 75 and 75 units
#in each layer respecitvely.
regressor = skflow.TensorFlowDNNRegressor(hidden_units=
 [75,75,75],steps=50000, learning_rate=0.001,num_cores=4)

# Fitting the NN
regressor.fit(X_train, y_train)

# calculating the predictions
t=regressor.predict(X_test)
# scoring the  R^2 between predtictions and test set
s = regressor.score(X_test,y_test)
print('Score on Testing Set: ' + str(s))
score.append(s)
```

The 6 hr predictions are then plotted (Figure 19) against the target buoy data.

```
#setting the plotting style
sns.set_context('notebook',font_scale=1 )
sns.set_style('ticks')
# plot the test set and the  NN 6 hr forecast
plt.plot(t_test[0])
plt.plot(t)

plt.title("NN 6 hr predictions vs ADCP ",fontsize=25)
plt.ylabel('Signigicant wave height(m)',fontsize=15)
plt.xlabel('Hours',fontsize=15)
plt.savefig('clean_NN.PDF')
```

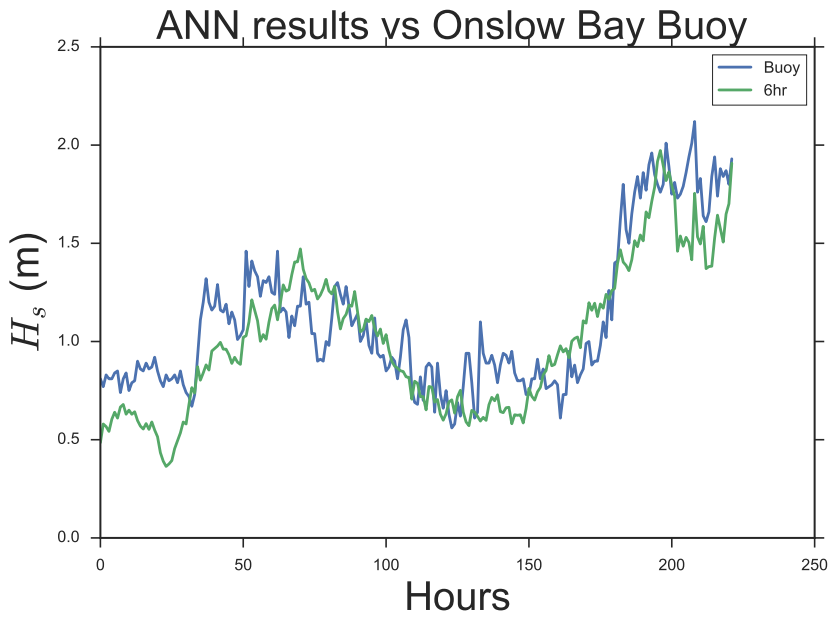

Figure 19: A plot of the ANN 6 hr Forecast results (green) against ADCP (Blue)

6 APPENDIX: NLTSA Forecasting codes

Like the ANN, we must first begin by loading the necessary packages for mathematical functions, array manipulations, running the NLTSA and creating nice looking plots for the result. The heart of NLTSA method lies within the nonlinpy-stripped function in line 15.

```
1  import numpy as np
2  import pandas as pd
3  import matplotlib.pyplot as plt
4  import os
5  import seaborn as sns
6  import datetime as dt
7  sns.set_style('whitegrid')
8  cmap='rainbow'
9  %matplotlib inline
10
11 import sklearn as sk
12 import Wavestt as ws
13 import sys
14
15 import nonlinpy_stripped as nlp
16 import rolling_train_test_split as rtt
```

Now that the necessary packages are loaded, the next step is to load the predetermined lagged and 4 embedding dimension buoy feature and target ADCP datsets. The features and testing sets are then split, 80 % for training and the remaining 20 % for testing.

```
1
2  XX, yy=NL4FEATURES,NL4TARGETWVHT
3
4  RTT = rtt.rolling_train_test_split(
5    NL4FEATURES.values ,NL4TARGETWVHT.values ,.2 ,skip=100)
6
7  Xtrain ,ytrain ,Xtest ,ytest = RTT.get_train_test()
```

The split data sets are now ready to be run through the NLTSA. Forecasts are generated and the performance (CC= R^2) is calculated for various number of near neighbors (nn-range). (Figure 20)

```
nn_range,cc = nlp.nnEfficient(Xtrain[0],ytrain[0] \
  ,Xtest[0],ytest[0],weights='distance')

fig,ax = plt.subplots()
sns.set_context('notebook',font_scale=1.2)
plt.plot(nn_range,cc)
plt.title(' $R^2$  for Target Forecasts \
  as function of Number of Near Neighbors')
plt.xlabel('Near Neighbors')
plt.ylabel('$R^2$')
plt.legend(['6hr','12hr','24hr','48hr','90hr'],frameon=True)
plt.ylim(0,1)
plt.xlim(0,180)
```

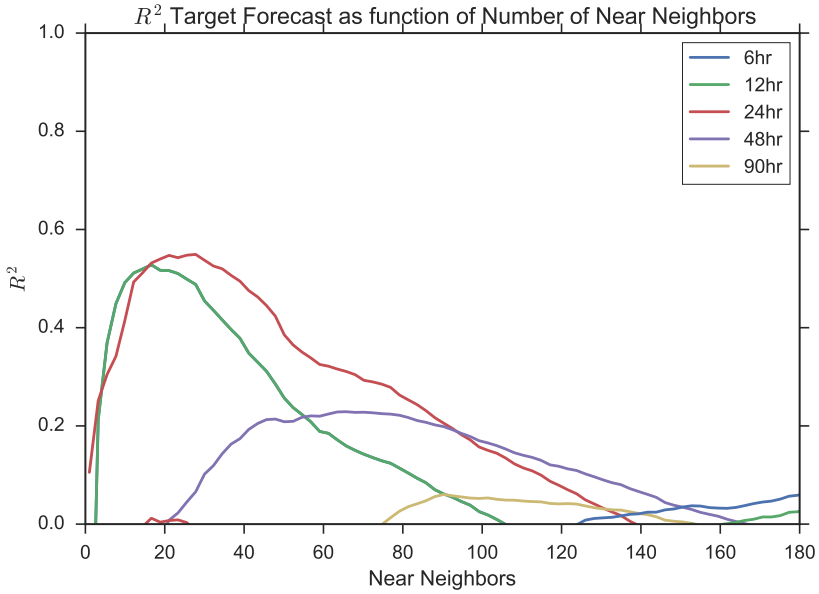

Figure 20: A plot of the R^2 Correlation Coefficient for the Target Forecasts as a function of the Number of Near Neighbors.

The maximum performance is then used to select the to the Number of Near Neighbors

for generating the forecasts of interest.

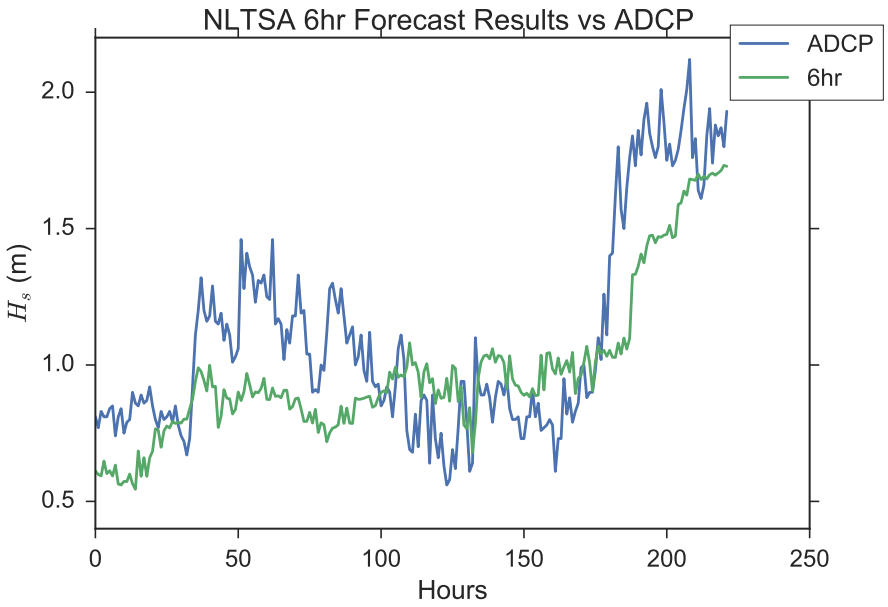

Figure 21: A plot of the NLTSA 6 hr Forecast results (green) against ADCP (Blue)

```
#max predictions at roughly 25 near neighbors
near_neighbors = 25

preds4 = nlp.get_preds(Xtrain[0],ytrain[0],Xtest[0],near_neighbors,weights='
    distance')
col=[0,1,6,12,18,24,30,36,42,48,54,60,66,72,78,84,90]

NL4_HSpreds=pd.DataFrame(preds4,columns=col)

sns.set_context('notebook',font_scale=1)

NL4_HSpreds

fff=pd.DataFrame(ytest[0])
```

The 6hr forecasts are then plotted against the ADCP target data. (Figure 21)

```
sns.set_context('notebook',font_scale=1.5)
```

```
sns.set_style('ticks')

plt.plot(fff[0])
plt.plot(NL4_HSpreds[6])

plt.title('NLTSA 6hr Forecast Results vs ADCP ')
plt.ylabel('$H_s$ (m)')
plt.xlabel('Hours')
plt.legend(['ADCP','6hr'],bbox_to_anchor=(1.12,1.05),frameon=True)
```

7 Acknowledgments

First I want to thank the North Carolina Sea Grant Foundation for funding this project and Greg Dusek of NOAA for coordinating the NWPS model outputs and data collection by RPS Evans Hamilton. I want to give a special thanks to Nick Cortale for teaching me all about python ,allowing me to use his codes, and for all the general coding and LaTex trouble shooting throughout the entire process. Another special thanks to Dr. McNamara ,for giving me the opportunity to work along his side in the Complex Adaptive System Lab and for always being stoked about physical oceanography. I also want to thank Dr. Bruce and the Honors Office for all their logistical support and the helpful honors cafe seminars held through out the past year.

8 Reflection

My Honors Project incorporated two significant theories and techniques that I learned from taking Dr. McNamara's computational physics class and from working along his side in the Complex Adaptive Systems Lab. From both experiences I learned a lot about coding in both MATlab and Python and insight into physical and data driven modeling in general. These skills and techniques played key roles throughout the entire honors project. Specifically my focus was geared towards data driven modeling through Nonlinear Time Series Analysis and Artificial Neural Networks. These data driven techniques can be used to make predictions on the evolution of physical systems without requiring knowledge of the underlying physical processes that govern that particular system.

Collecting the data for the data driven techniques became a very tedious and frustrating process. In order to have the most accurate forecasts, I had to assure that the buoys chosen were within a desirable region in relation to the target location and had the fewest missing data entries during the period in which the ADCP collected the wave features. I spent days sifting through NOAA's immense National Buoy Data Center's historical records in order to try and find the best data set. The "best" data set needed a balance between relative location to the forecasting target and completeness(few missing entries). Had other closer buoys with more complete data sets been chosen, I believe both data driven techniques would have seen significant increases in forecasting accuracy .

Additional caveats of the data driven models could have also yielded higher forecasting accuracy. Specifically those parameters that were determined by trial. Perhaps other network configurations or embedding dimensions could have been explored to improve forecasting at various horizons. Despite the data constraints and parameters determined by trial ,the results from this project showed that these data driven modeling techniques can be used to make more accurate wave forecasts in the nearshore zone compared to those created by complex deterministic models.

The data driven models also show promise in creating hind casts from other buoys records.

This could be very useful for filling in missing data from any of the NDBC's historical records by using any of the other nearby buoy's records. There are many avenues for future work and research using these data driven modeling techniques. I hope to continue furthering my understanding through applied learning about these data driven models and continue trying to improve nearshore wave forecasts.

REFERENCES

[1] D. J. Grimes, N. Cortale, K. Baker, and D. E. McNamara *Nonlinear forecasting of intertidal shoreface evolution* Chaos, 25, 103116 (2015)

[2] Hsieh, C.H., Glaser, S.M., Lucas, A.J., Sugihara G. *Distinguishing random environmental fluctuations from ecological catastrophes for the North Pacific Ocean.* Nature,435(7040):336-40, May 5,(2005)

[3] C.T. Perretti, S.B Munch, and G. Sugihara *Model-Free forecasting outperforms the correct mechanistic model for simulated and experimental data* PNAS 2013 110 (13) 5253-5257; published ahead of print February 25, (2013)

[4] Takens, F., *Detecting Strange Attractors in Turbulence* Dynamical Systems and Turbulence, Lecture Notes Math., Vol. 898, pp. 366-381, Springer, New York, (1981).

[5] Gardner, M. W., Dorling, S. R., *Artificial Neural Networks (The Multilayer Perceptron)-A Review of Applications in the Atmospheric Sciences* Atmospheric Environment, Vol. 32, No. 14/15, 2627-2636. (1998)

[6] Casdagli, Martin, (1991) *Chaos and Deterministic versus Stochastic Non-Linear Modeling.* J. R. Statist. Soc. B, No. 2, pp. 303-328, (1991)

[7] G, Reikard , P.Pinson , J.R.Bidlot *Forecasting ocean wave energy, The ECMWF wave model and time series methods* Ocean Engineering, 38 (2011) 1089-1099

[8] A. Westhuysen, R. Padilla, P. Santos, A. Gibbs, D. Gaer, T. Nicolini, S. Tjden, E.M. Delvaliere, H, Tolman *Development and validation of the Nearshore Wave Prediction system* Presented at the 93rd AMS Annual Meeting, Austin, TX, January 5-10, 2013

[9] J.L, Hanson, B. A. Tracy, H, Tolman R. D. Scott, *Pacific Hind cast Performance of Three Numerical Wave Models* Journal of Atmospheric and Ocean Technology, Vol. 26 (2009)

[10] M.C. Deo, C. Sridhar Naidu *Real time wave forecasting using neural networks* Ocean Engineering 26 (1999) 191-203

[11] M.C. Deo , A. Jha, A.S. Chaphekar, K. Ravikant *Neural networks for wave forecasting* Ocean Engineering 28 (2001) 889-898

[12] N. Booij, R. C. Ris, and L. H. Holthuijsen *A third-generation wave model for coastal regions 1. Model description and validation* Journal of Geophysical Research, VOL. 104, NO. C4, PAGES 7649-7666,(1999)

[13] Tolman, Hendrik L. *The 2002 release of WAVEWATCH III.* 7th International Workshop on Wave Hindcasting and Forecasting. 2002.

[14] Komen, G. J., K. Hasselmann, and K1 Hasselmann. *On the existence of a fully developed wind-sea spectrum.* Journal of physical oceanography 14.8 (1984): 1271-1285.

[15] Hasselmann, D. E., M. Dunckel, and J. A. Ewing. *Directional wave spectra observed during JONSWAP 1973.* Journal of physical oceanography 10.8 (1980): 1264-1280.

[16] R. A. Luettich, J.J Westerink, Norman W, Scheffner *An Advanced Three-Dimensional Circulation Model for Shelves, Coasts, and Estuaries, Report 1. Theory and Methodology of ADCIRC-2DDI and ADCIRC-3DL* Department of the Army US Army Corps of Engineers Washington, DC 20314-1000 Under Work Unit No. 32466

[17] Gardner, Matt W., and S. R. Dorling. *Artificial neural networks (the multilayer perceptron) a review of applications in the atmospheric sciences.* Atmospheric environment 32.14 (1998): 2627-2636.

[18] B. Dolhansky *Artificial Neural Networks: Linear Regression (Part 1)* ml primers, neural networks

[19] Kantz, Holger, and Thomas Schreiber. *Nonlinear time series analysis* Vol. 7. Cambridge university press, 2004.

[20] Shrestha, Durga L., and Dimitri P. Solomatine. *Machine learning approaches for estimation of prediction interval for the model output.* Neural Networks 19.2 (2006): 225-235.