Short Term Forecasting of the Near-shore

Davoud Ataee Tarzanagh[1], Zheming Gao.[2], Li Liu[3], Angelo Marney[4], Chathurangi Pathiravasan. [5], David Robinson [6], Caoxin Sun [7]

Mentors: Ty Hesser and Matthew Farthing [8], Lea Jenkins [9]

Abstract

A new methodology for forecasting near-shore wave conditions is proposed. Near shore features of interest for beach landing craft include maximum wave height, wave direction, wave period, beach slope, sandbar location, and water depth at the sandbar location. Using a combination of remotely sensed topography data and images of the near shore taken from cameras at the beach, bathymetry data was inferred using an improved cBathy algorithm. Historical wave height, wave period, and wave direction was directly measured from a single sensor 900 meters offshore. ARIMA time series models were used to forecast the wave period, wave direction, and wave height using the data from the offshore sensor. These forecasted values were then used to compute near shore wave height using a validated wave transformation model. Beach slope, sandbar locations, and sandbar depth were determined from the bathymetry data. We tested our methods using data from the U.S. Army Corps of Engineers Field Research Facility at Duck, North Carolina.

1 Introduction

Providing an accurate characterization of near shore features is important for planning beach landing. For example, those piloting ocean craft in the near shore region may need to know about certain wave parameters such as maximum wave height, wave direction, and wave period. Ocean craft that needs to land on a beach may also need information about the sandbar location, sandbar depth, and beach slope. These are just a few examples of when knowledge of near shore features is important.

The research for this report focused on forecasting potential parameters of interest to be used in determining beach landing conditions. The specific goal is to predict the maximum wave height, the wave period, the wave direction, the sandbar location, the depth at the sandbar, and the beach slope. One difficulty in forecasting these parameters is the physical complexity of shallow water waves. Shallow water waves are greatly affected by the bathymetry, but direct measurement of bathymetry may be costly and time prohibitive. Near shore parameters are also greatly influenced by the tide, so tide also needs to be taken into account.

Wave height, wave period, and wave angle were measured in situ from a single sensor 900 meters offshore at the USACE Research Facility at Duck, North Carolina. Additionally, remotely sensed topographic surveys and image data was collected at this location, and an improved version of the popular cBathy algorithm [8] was used to determine bathymetry. Using this data, a time series model was built to forecast the parameters of interest. A wave transformation model was then used to determine a wave height profile according to the forecasting parameters.

In the second section, we describe the problem in greater detail, including our proposed solution to the model using what data is available to us. The third section includes the details about how the problem is solved. The fourth section is dedicated to presentation of our solution. We provide a summary of the entire project and directions for future research in the fifth section. The sixth section is acknowledgements.

[1] Department of Mathematics and UF Informatics Institute(UFII), University of Florida
[2] Department of Operations Research, North Carolina State University
[3] Department of Mathematics and Statistics, University of North Carolina at Charlotte
[4] Department of Mathematics, Virginia Tech
[5] Department of Mathematics, Southern Illinois University, Carbondale, IL
[6] Department of Scientific Computing, Florida State University
[7] Department of Computer Science and Statistics, University of Rhode Island
[8] USACE-Engineer Research and Development Center, Vicksburg, MS
[9] Clemson University, Clemson SC

2 The Problem

The exact problem is to determine the following near shore parameters along a one dimensional transect (we shall now refer to this one dimensional transect as the problem domain) aligned normal to the shore:

1. H_{max}: maximum wave height measured in meters
2. α: wave direction measured clockwise from the shore normal direction
3. T: the wave period measure in seconds
4. b: the beach slope
5. x_{bar}: the sand bar location measured in meters away from the onshore reference point
6. h_{bar}: the sand bar depth measured in meters

The wave parameters H_{max}, α, and T are to be forecast at hourly intervals starting at April 11, 2017 at 12:00 pm for 72 hours. In situ measurements from an offshore direction array of sensors can be used to test the accuracy of the forecast. It is assumed that the wave period remains constant along the problem domain at any particular time. This assumption is justified because it can be demonstrated that as a wave propagates from deep water towards the shore, the number of waves passing sequential points in a given time period must be constant [10]. It is also assumed that the bathymetry h does not change over the forecast period. In reality, the bathymetry changes with time due to the movement of sediment along the sea floor, but this assumption that h does not change is useful in simplifying the model, and it is still reasonable since the forecast period is short. In order to solve this problem, we leverage the following data resources described in the next subsection.

2.1 Data resources

This research project relied on data collected at the U.S. Army Corps of Engineers (USACE) Field Research Facility (FRF) in Duck, NC [10]. The USACE Coastal Observation and Analysis Branch (COAB) is responsible for collecting and maintaining these large data sets. The data used in our project consists of two different sets of data: offshore direction array data and bathymetry data.

2.1.1 Offshore direction array data

The offshore direction array data consists of in situ measurements of wave parameters. These measurements are taken hourly and can be accessed from a database. COAB collected the directional wave data using an array of pressure sensors at the FRF in Duck, NC. The array consists of 15 pressure gauges mounted approximately 0.5 meters off the bottom in the vicinity of the 8 meter isobath almost 900 meters offshore. Table 1 shows the position of the sensors. Each sensor is located in a array with a fixed local y = 950 meter coordinate, with x coordinates varying as the sensor gets further from shore. Wave direction is determined from these gauges using an iterative maximum likelihood estimator. Wave height and wave period are also measured at these sensors. Only the data from the 8m-array (furthest from shoreline) was used for time series analysis.

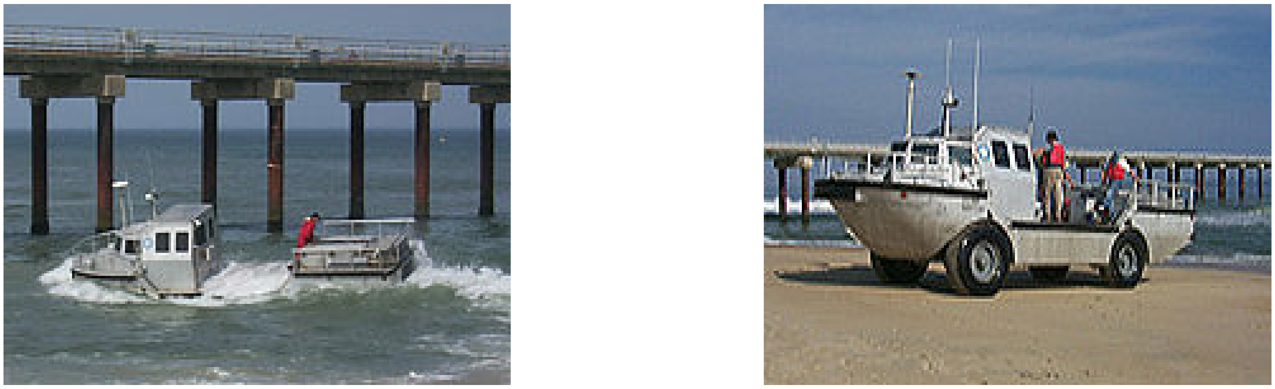
[10] Data from FRF data website https://chlthredds.erdc.dren.mil/

sensor	x(m)
8m-array	900
awac-6m	550
awac-4.5m	400
adop-3.5m	300
xp 150	150
xp 125	125

Table 1: The sensor names and their distances from a reference point on the shore. The distance x is measured in meters. The 8m-array sensor is the furthest from the shoreline, while the xp 125 array was closest to the shoreline.

2.1.2 Topographic survey data

Topographic survey data was collected at the FRF in Duck, NC. The surveys were done monthly, providing accurate topography data with discrete local coordinates (x, y) within the grid $[50, 950] \times [-100, 1100]$, with a 12 meter step length on x, and 24 meter step length on y. These topographic surveys were done by the USACE using the LARC Survey System [2]. The images 1a and 1 show some of the equipment needed to carry out these surveys. Figure 2 shows the topography data obtained from a single survey. On shore topography data may be more easily collected using remote sensing techniques, such as light detection and ranging (LiDAR) or photogrammetry, but these remote sensing techniques may be very limited when it comes to determining underwater topography, due to their reliance sea water clarity.



(a) Survey boat

(b) Survey boat

Figure 1: LARC survey boats used to carry out topographic surveys. Running these surveys can be expensive and time consuming.

2.1.3 Tide data

Tide has a significant effect on many near-shore parameters, so tidal data needs to be taken into account when making predictions. Tidal data is taken from the National Oceanic and Atmospheric Administration (NOAA) tides and currents database [11]. The NOAA tides and currents database provides short term predictions of the tide in many locations throughout the United States, including at the FRF. Tide data for each hour is stored in comma separated value format and .mat files. Figure 3 illustrate the tide data.

2.1.4 Preprocessed image data

In many scenarios, getting underwater topography data from surveys may not be possible, for example, due to high costs, lack of equipment, or lack of time. On shore topography data may be reliably obtained via

[11] https://tidesandcurrents.noaa.gov/waterlevels.html

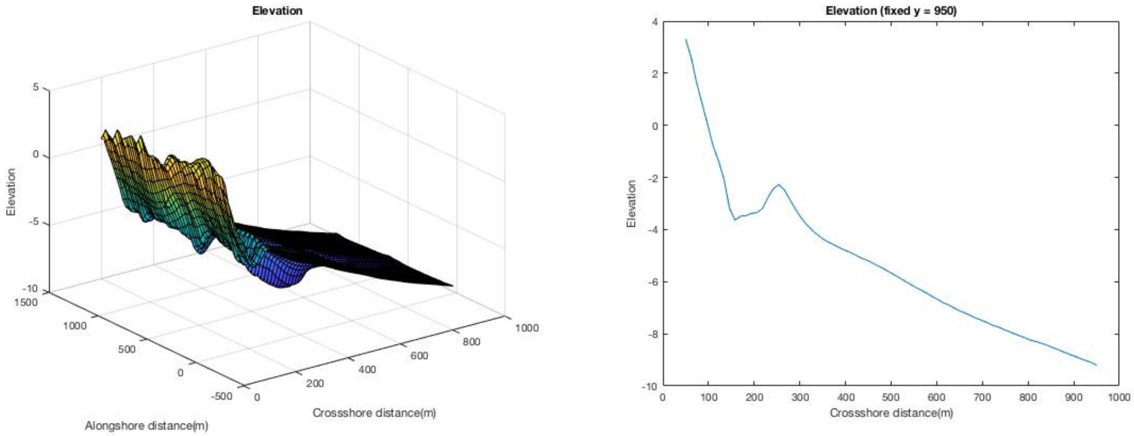


Figure 2: Bathymetry data obtained from a LARC survey on April, 2017. The left image is a surface interpolation of the surveyed bathymetry. The right image is a one dimensional transect of the bathymetry surface, obtained by fixing a single y coordinate.

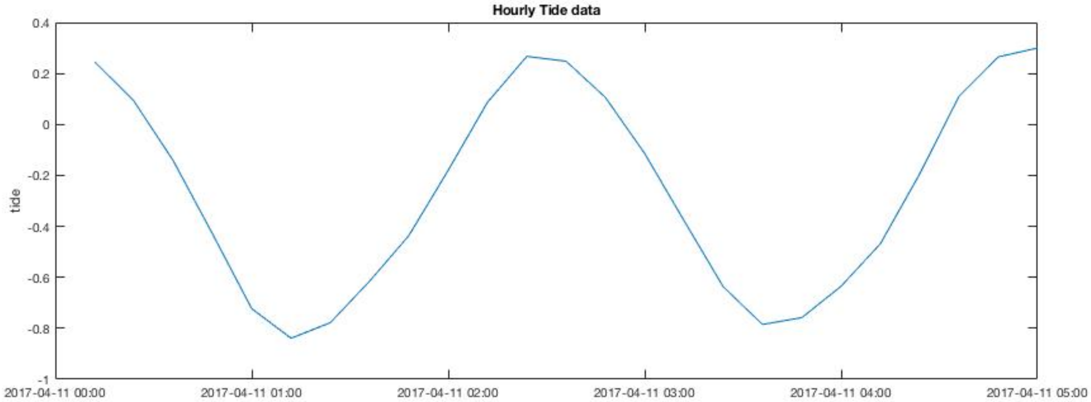


Figure 3: The hourly tide data on April 11, 2017.

remote sensing techniques. One way to obtain near shore bathymetry data is to infer it from ocean surface measurements collected from image data. In particular, a dispersion relationship based on linear wave theory relates water depth to surface properties, and these surface properties can be estimated from processed image data from the FRF provided by the USACE. The processed image data is taken as an input into the an improved version of the cBathy algorithm [8], and the improved cBathy algorithm outputs the estimated bathymetry.

2.1.5 Data preprocessing

All of the data available on the USACE servers were originally saved in different formats, so to access the data, a lot of work had to be done to import it into the appropriate format. For example, topography survey data was available as a text file, so it had to be downloaded and converted into a .mat file in order to be used in MATLAB. This conversion also required some organizing the arrays so that the topography survey data could be more easily used. The offshore direction array data was also downloaded and converted into a .mat file which produced structures when ran in MATLAB. Each sensor in the array had its own .mat file, and the relevant wave parameters were organized in matrices belonging to a field of the structure. The sensor data needed to be cleaned, since there were missing values that took the form of NaN in MATLAB. These NaN values were set to zero, to signify that no measurements were taken at the sensor.

2.2 The approach

To determine and forecast the near shore parameters of interest, we combine a data-driven statistical approach with a deterministic physics-based approach. A flowchart for solving the problem is shown in Figure 4.

We obtain an estimate for the bathymetry h from two sources of data. The first source of data comes from the remotely sensed preprocessed image data of the near shore region. These preprocessed images are then input into an improved cBathy algorithm to produce an estimate of the bathmetry, which produce fairly accurate further offshore estimates. On shore topography data is obtained from a survey, which can in theory be taken remotely using LiDAR or photogrammetry techniques. An adjusted bathymetry is determined by merging the onshore topography and the cBathy bathymetry estimate. This adjusted bathymetry is then used to determine sandbar location x_{bar}, sandbar depth h_{bar}, and beach slope b, very close to the forecast period, giving us three of the parameters of interest.

A single offshore sensor (8m-array) is used to measure historical wave height H_0, historical wave period T, and historical wave direction α_0, all at a single point. This sensor stores the data hourly, and this hourly data is used to build an autoregressive integrated moving average (ARIMA) time series model for each of the three wave parameters. Approximately three months worth of historical data is used to build each model. The ARIMA models are used to forecast the three wave parameters at the location of the offshore sensor. Since T is assumed to be constant throughout the problem domain, this gives us one more parameter of interest.

The three forecasted parameters, along with the bathymetry profile $h(x)$, are used in the one dimensional wave transformation model. The wave transformation model takes the form of an ordinary differential equation, with the forecasted wave height H_0 acting as a boundary condition for the model. The wave direction α_0 is propagated towards shore using Snell's Law, giving is one more parameter of interest. The bathymetry profile $h(x)$ and period affect some of the parameters in the model. The ordinary differential equation is solved using a first order upwind scheme, from which the wave height $H(x)$ over the problem domain can be inferred, and from which we can determine maximum wave height H_{max}, giving us the final parameter of interest.

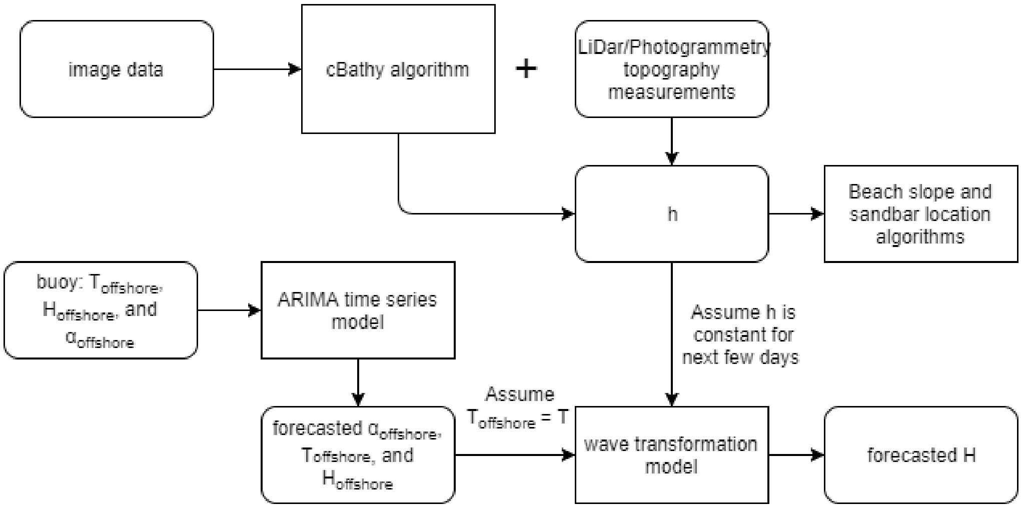


image data
cBathy algorithm
+
LiDar/Photogrammetry topography measurements
h
Beach slope and sandbar location algorithms
buoy: $T_{offshore}$, $H_{offshore}$, and $\alpha_{offshore}$
ARIMA time series model
Assume h is constant for next few days
forecasted $\alpha_{offshore}$, $T_{offshore}$, and $H_{offshore}$
Assume $T_{offshore}$ = T
wave transformation model
forecasted H

Figure 4: The flowchart for solving the problem.

3 Methodology

3.1 cBathy

The problem of estimating ocean wave properties from optical signals can be surprisingly challenging. Walker [14] showed that for waves outside the surf zone viewed at the typical low-graze angles of coastal cameras, the primary source of light from the ocean comes from specular reflection of skylight by the sea surface and the primary source of wave contrast is variations in sea surface slope and the associated slope dependence of the optical reflection coefficient. Since sea surface slope depends on the wave amplitude, a, times the wave number, k, this mechanism is dominated by high wave number (short wavelength) waves, or ocean chop. This is apparent when viewing any ocean scene. Human observers instinctively filter the observed wave patterns to see the longer, coherent incident wave pattern while ignoring the short wave clutter, but a computer algorithm must understand and properly deal with these sources of noise. In the following algorithm, this will be done through both frequency domain methods (temporal Fourier transforms) and through coherence and EOF-based filtering.

While the disadvantage of optical data is the high-noise level, the advantage is the huge volumes of data that are available at very low cost. A single camera can deliver around 35 MB per second, a data rate well beyond what is needed for wave characterization, and one that requires extensive data reduction. The cBathy [9] will strive estimate bathymetry over a 420 by 1000m region with a spatial resolution of 10 by 25m in the cross shore (x) and alongshore y receptively. Temporal sampling is done at 2 Hz, a further reduction by a factor of 15 over typical 30 Hz video rates, for record lengths of 1024 s, each hour. Even with this reduction of 4 orders of magnitude in available data usage, 17.6 million intensity samples are collected for each data run, so that there are approximately 10,000 degrees of freedom for every individual depth estimate. Thus, robust signal processing opportunities are available.

Next, we review an example implemented in this software. Figure 1 shows the typical pixel sampling array described above with the blue dots each corresponding to the locations, $[x_p, y_p]$, of a pixel time series data. The analysis is carried out sequentially at a series of user-selected analysis points, $[x_m, y_m]$ and is based on data from the immediately surrounding pixels (green points) within a user-specified range,

$$[x_m \pm L, y_m \pm L].$$

Within each such tile, the goal is to estimate the wave number k, for each of a set of candidate frequencies f_b. Estimates may be poor or impossible at times due to weather, sun glare or calm seas, so estimates from hourly data collections are objectively averaged to yield a stable running average depth,$\bar{h}(x_m, y_m)$. The first step is to Fourier transform, `fft`, the optical intensity time series at each pixel, $I(x_p, y_p)$ math formula, such that $G(x_p, y_p, f)$= `fft` $(I(x_p, y_p), t)$. Because our interest is in modeling wave phase and neglecting spatial variations in magnitude, the Fourier coefficients are then normalized, $\hat{G} = G/|G|$. The full data set is then sub-sampled to a local data tile in the region math formula (example green region in Figure 5) and the cross-spectral matrix computed between all possible pixel pairs for each of the desired frequency bands

$$C_{ij} \quad = \quad \langle \hat{G}(p_i, p_i, f), \hat{G}(p_j, p_j, f) \rangle, \tag{1}$$

where superscript $*$ indicates the complex conjugate and the expected value is averaged across each frequency band.

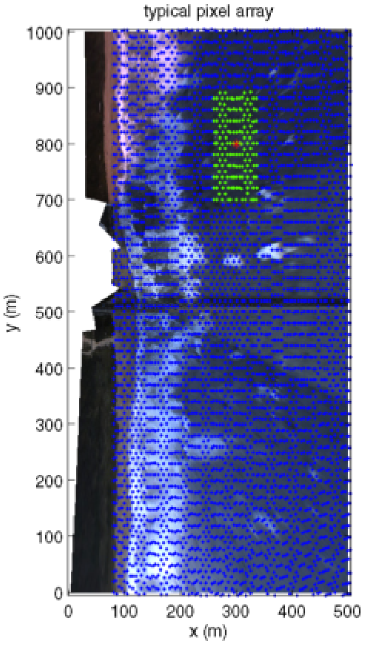


Figure 5: Example pixel array used for cBathy analysis. The 8600 pixels (half shown) span a 420 by 1000 m region with 5 by 10 m resolution. For each analysis point, depth is estimated based on cross-spectral phase within a nearby region (green pixels). The background image is a rectified snapshot that merges views from the five available cameras [9].

For complex natural seas, the cross-spectral matrix can mix the effects of multiple wave trains from different directions. To extract only coherent motions from this mix, the dominant (complex) singular vector,v, and associated eigenvalue, λ, are extracted from C through using PCA. We define the optimum wave number, k, and wave direction, α, as those values that yield the best match between observed and modeled spatial phase structure of v based on a forward model

$$v' = \tan^{-1}(\frac{\text{imag}(v)}{\text{real}(v)}) = \exp(i[k\cos(\alpha)]x_p + k\sin(\alpha)y_p + \phi), \tag{2}$$

where the search is accomplished using MATLAB routines based on the Levenberg-Marquardt algorithm. The scalar phase angle, math formula, is of no geophysical value in the subsequent analysis and simply provides an appropriate phase shift to match the observed spatial structure of v'.

3.1.1 Robust cBathy

PCA is arguably the most widely used statistical tool for data analysis and dimensionality reduction today. However, its brittleness with respect to grossly corrupted observations often puts its validity in jeopardy a single grossly corrupted entry in C could render the estimated v arbitrarily far from the true singular vector. A number of natural approaches to robust PCA have been explored and proposed in the literature over several decades. The representative approaches include influence function techniques, multivariate trimming, alternating minimization, and random sampling techniques and robust PCA [5].

The robust PCA [5] can be considered an idealized version of PCA, in which we aim to recover a low-rank matrix L from highly corrupted measurements C. Unlike the small noise term E in classical PCA, the entries in S can have arbitrarily large magnitude, and their support is assumed to be sparse but unknown.

In this work, we study the tensor robust PCA problem which aims to recover the low rank tensor L and sparse component S from $C = L + S \in \mathbb{R}^{n_1 \times n_2 \times n_3}$ (focuses on the 3-way tensor) by convex optimization

$$\min_{L,S} \quad \operatorname{rank}(L) + \|S\|_0 \qquad s.t. \qquad C = L + S, \tag{3}$$

where $\|S\|_0$ denotes the number of non zeros of tensor S.

3.1.2 Alternating direction methods for robust PCA

The alternating direction methods (ADM) in [4] are based on an augmented Lagrangian framework. Note that given a penalty parameter $\rho > 0$, the augmented Lagrangian function associated with problem (3) is

$$\mathcal{L}_\rho(L, S; \Lambda) := \|L\|_* + \xi\|S\|_1 - \langle \Lambda, L + S - C\rangle + \frac{1}{2\rho}\|L + S - D\|_F^2, \tag{4}$$

where Λ is a matrix of Lagrange multipliers. Note that the penalty parameter ρ can be adjusted dynamically, and this yields the k-th iteration of the augmented Lagrangian method as follows:

$$\begin{cases} (L_{k+1}, S_{k+1}) & := & \operatorname{argmin}_{L,S} \mathcal{L}_{\rho_k}(L, S; \Lambda_k) \\ \Lambda_{k+1} & := & \Lambda_k - (L_{k+1} + S_{k+1} - C)/\rho_k, \\ \rho_{k+1} & := & \eta\rho_k, \end{cases} \tag{5}$$

where $\eta \in (0, 1]$.

The ADMM in [4] is based on (5). Indeed, it is easy to minimize $\mathcal{L}_\rho(L, S; \Lambda)$ with respect to L or S while keeping the other matrix fixed and each minimization has a closed form solution which is easy to compute. Thus, ADMM computes (L_{k+1}, S_{k+1}) by alternatingly minimizing $\mathcal{L}_{\rho_k}(L, S; \Lambda_k)$ repeatedly in L and in S, while fixing the other, until the stopping criterion for the inner loop is met, i.e.,

$$\begin{cases} L_{k,j+1} & := & \operatorname{argmin}_L \mathcal{L}_{\rho_k}(L, S_{k,j}; \Lambda_k), \\ S_{k,j+1} & := & \operatorname{argmin}_S \mathcal{L}_{\rho_k}(L_{k,j+1}, S; \Lambda_k), \end{cases}$$

loop is repeated until $\max\{\|L_{k,j+1} - L_{k,j}\|_F,\ \|S_{k,j+1} - S_{k,j}\|_F\} \leq 10^{-6}\ \|D\|_F$ holds; at that point (L_{k+1}, S_{k+1}) is set to $(L_{k,j+1}, S_{k,j+1})$. Next, Λ_k and ρ_k are updated:

$$\begin{cases} \Lambda_{k+1} & := & \Lambda_k - (L_{k+1} + S_{k+1} - D)/\rho_k, \\ \rho_{k+1} & := & \eta\rho_k. \end{cases}$$

As a result, the iterate (L_{k+1}, S_{k+1}) in EADM only approximately minimizes $\mathcal{L}_{\rho_k}(L, S; \Lambda_k)$.

Updating the matrix of Lagrangian multipliers Λ at every iteration after minimizing $\mathcal{L}_{\rho_k}(L, S; \Lambda_k)$ first in L and then in S leads to the following alternating direction method of multipliers (ADMM). In the k-th iteration of ADMM, one computes,

$$\begin{cases} L_{k+1} & := & \operatorname{argmin}_L \mathcal{L}_{\rho_k}(L, S_k; \Lambda_k), \\ S_{k+1} & := & \operatorname{argmin}_S \mathcal{L}_{\rho_k}(L_{k+1}, S; \Lambda_k), \\ \Lambda_{k+1} & := & \Lambda_k - (L_{k+1} + S_{k+1} - C)/\rho_k, \\ \rho_{k+1} & := & \eta\rho_k, \end{cases} \tag{6}$$

where $\eta \in (0, 1]$.

Actually, as will be seen later, if $n_3 = 1$ (X is a matrix in this case), our tensor reduces to Robust PCA in (3). Another advantage of (3) is that it can be solved by polynomial-time algorithms, e.g., the standard Alternating Direction Method of Multipliers (ADMM) [4]. We obtain the alternating linearization method given in **Algorithm 1** that was analyzed in [4] by Boyd et. al for minimizing the sum of two convex functions.

Algorithm 1 Alternating Linearization Method (ALM)

```
input: $L_0 \in \mathbb{R}^{n_1 \times n_@ \times n_3}$, $\rho > 0$
$k \leftarrow 0$, $S_0 \leftarrow L_0$
while not converged do
    $L_{k+1} \leftarrow \text{argmin}_L \, \mathcal{L}_{\rho_k}(L, S_k)$
    $S_{k+1} \leftarrow \text{argmin}_S \, \mathcal{L}_{\rho_k}(L_{k+1}, S)$
    $\Lambda_{k+1} := \Lambda_k - (L_{k+1} + S_{k+1} - C)/\rho_k,$
    $k \leftarrow k + 1$
end while
return $(L_k, S_k)$
```

3.1.3 Performance Evaluation

In this section, we investigate the efficiency of the proposed algorithms, especially the `Robust cBathy`, on the real data sets. We have implemented the following algorithms in the MATLAB R2015b environment on a PC with a 1.8 GHz processor and 6 GB RAM memory and double precision format:

- **cBathy**, the original cBathy algorithm with an original PCA [9];
- **Robust cBathy**, the robust tensor factorization following a PCA;

Proposed robust cBathy is being terminated either the relative square error (RSE),

$$\text{RSE} := \frac{\|L_0 - \tilde{L}\|_F}{\|L_0\|_F} \leq 10^{-3},$$

or the number of iterations exceed 1,00 and respectively.

An example of our result is shown in Figure 6 and Figure 7. It can be seen that the `Robust cBathy` outperforms its competitors in terms of accuracy.

3.2 Determining bathymetry, beach slope, and sandbar location

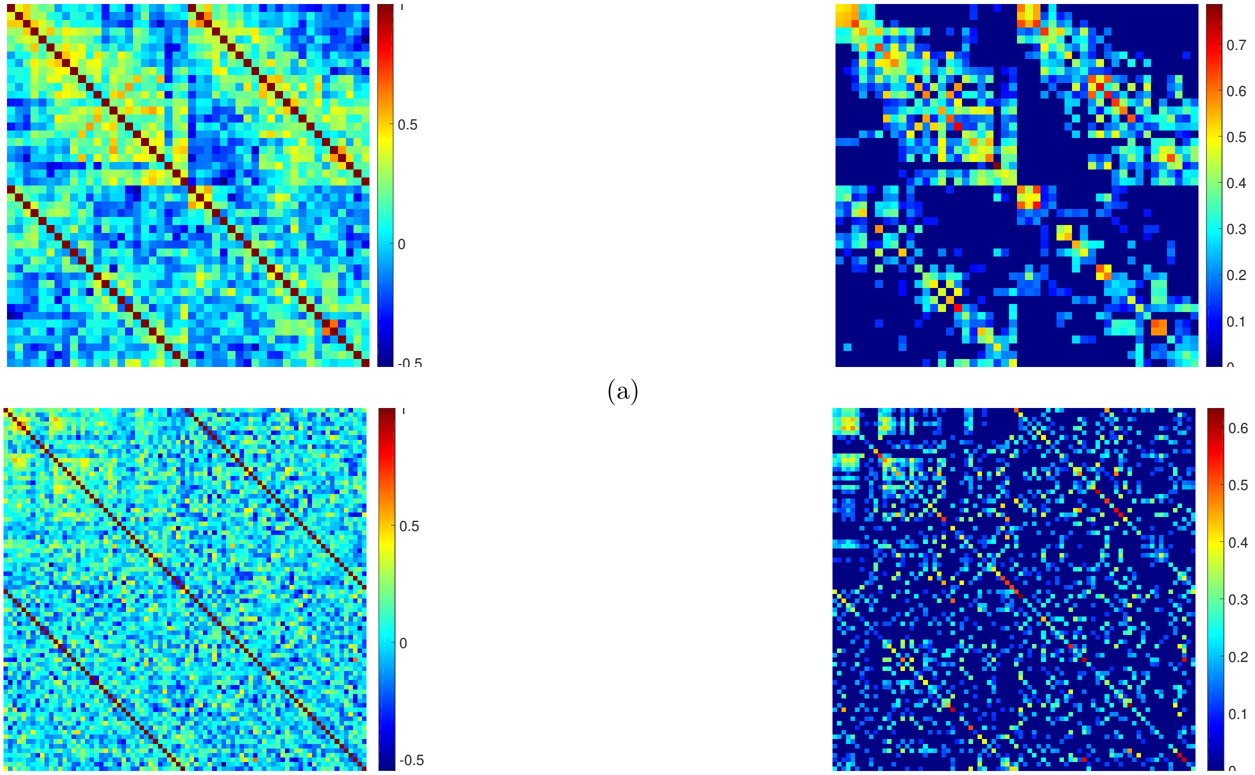
Due to cBathys domain not encompassing the offshore region where the boundary condition data is collecting for the forecast model and extrapolation method is needed to determine the bathymetry of the unkown region. Also, merging of the onshore survey data and cBathy bathymetry is needed to get a better estimate of the shallow water bathymetry due to cBathys increase in error in shallow water regions, usually failing to predict the correct location of the shore. Once the adjusted bathymetry is calculated the sandbar location, depth at the sandbar, and beach slope can all be determined.

3.2.1 Adjusted bathymetry

Shallow water bathymetry improvements Since topography data of the beach shore can be attained remotely via UAV equipped with LiDAR equipment we use survey data to attain topology values of the beach shore [2]. A piecewise cubic hermite interpolating routine (pchip), due to its shape preserving properties, is used with the onshore data points to interpolate between the region where cBathys error is reliably low and the shore.

Beach slope and sandbar location Once the near shore bathymetry is improved as outlined above the region where depth goes to zero is determined. The pchip routine is then used to build an interpolation function to sample points around the shore. A linear least squares fit is determined from the sample points to determine the slope of the beach at the shore. This avoids estimating the slope from the sometimes sharp variations in the near shore bathymetry; see Figure 9.

Sandbar location is determined using a simple local maximum finding algorithm starting from the shore. Once all possible local maximum points are found their depth is evaluated and the sandbar location is taken to be the shallowest of these points. To determine the hourly depth at this point in the forecast tidal data is used to adjust the depth to show the variation at this point. Figure 10 depicts these results.



(a)

(b)

Figure 6: Two examples of the cross-spectral matrix C, computed between all possible pixel pairs. **Left**: cBathy [9]. **Right**: Roboust cBathy.

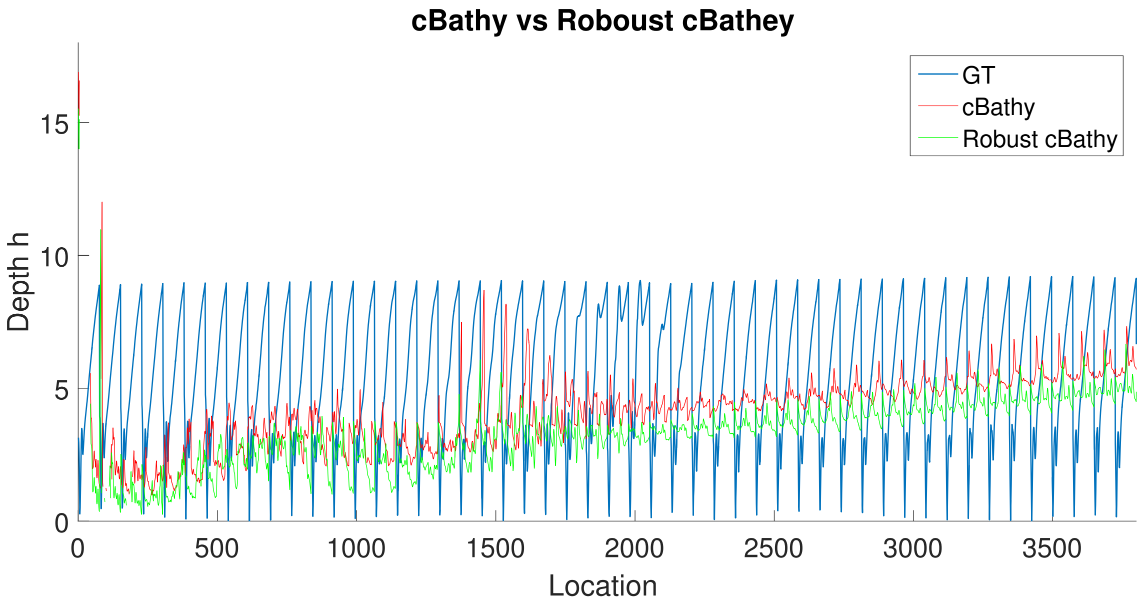


Figure 7: cBathy vs robust cBathy for estimating depth, h.

Offshore bathymetry extrapolation The domain of cBathys bathymetry estimation stops at 800 meters when matched to the offshore sensor array coordinate system. The issue with this is that the forecasted

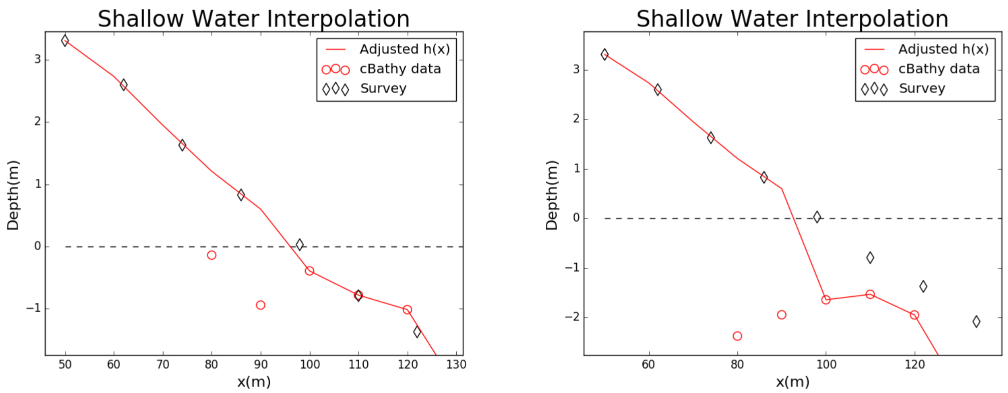


Figure 8: These show the adjustment in near shore bathymetry for two different cBathy bathymetries being merged with the same survey data. These figures show that regardless of the sporadic behavior of cBathys output in the near shore, an improved bathymetry estimate is determined.

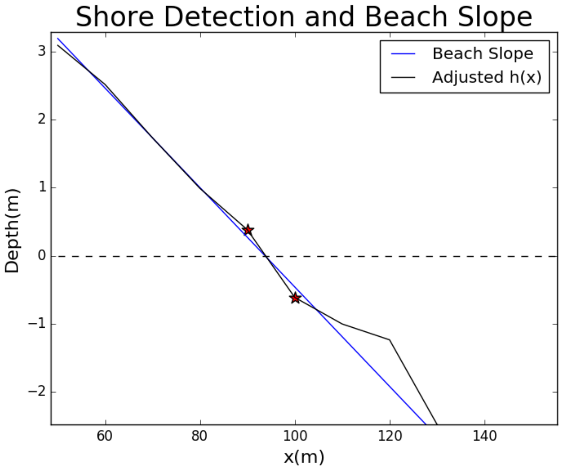


Figure 9: The starred points represent data points that encapsulate the zero depth region. As seen from the figure the LLS produces and accurate representation of b at the shore. The location of the shore can sometimes be encapsulated by different data points due to variation in the shore location from varying tide depth. Even with this variation taken into account b remains almost constant through the forecasting period and only changes on the order of 10^{-2}. This amounts to a change in elevation of about a $\frac{cm}{m}$ so the mean of the beach slopes is taken. The average was taken of the beach slopes and it was found that $\bar{b} = 0.070462$ which is a steep/moderate slope according to the Army field manual for water transport operations.[1]

boundary condition is determined at 900 meters, therefore, extrapolation of the bathymetry is needed to have an estimated depth at this location for the wave model. This is determined using a linear least squares fit to fit a line to all bathymetry data 200 meters farther offshore from the determined sandbar location. This is done since a linear relationship with offshore distance is observed in the farther offshore region. This linear fit is then used to extrapolate to the 900 meter point as shown in Figures 11-12.

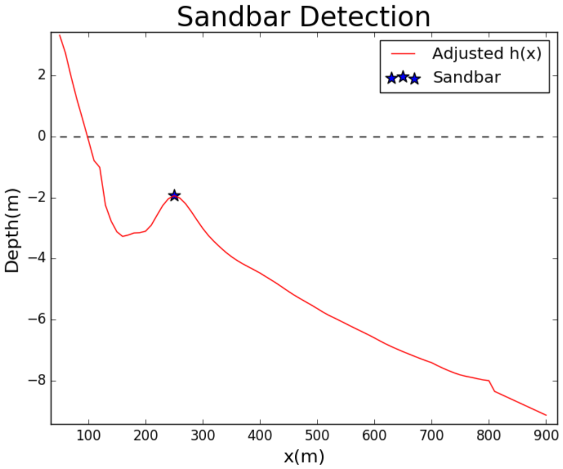


Figure 10: The starred points represent data points that encapsulate the zero depth region. As seen from the figure the LLS produces and accurate representation of the beach slope at the shore.

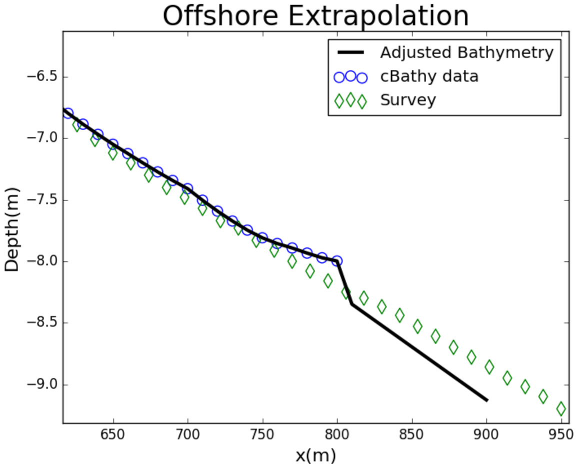


Figure 11: Extrapolation to the 900 meter mark compared with survey data.

3.3 Wave model

The following differential equation was used to model the wave transformation to determine the near shore wave height:

$$\frac{d}{dx}\left(EC_g \cos\left(\alpha\right)\right) = -\delta. \tag{7}$$

The wave energy is represented as

$$E = \frac{1}{8}\rho g H_{\text{rms}}^2 . \tag{8}$$

The group celerity (group velocity) is represented as

$$C_g = \frac{C}{2}\left(1 + \frac{2kh}{\sinh\left(2kh\right)}\right), \tag{9}$$

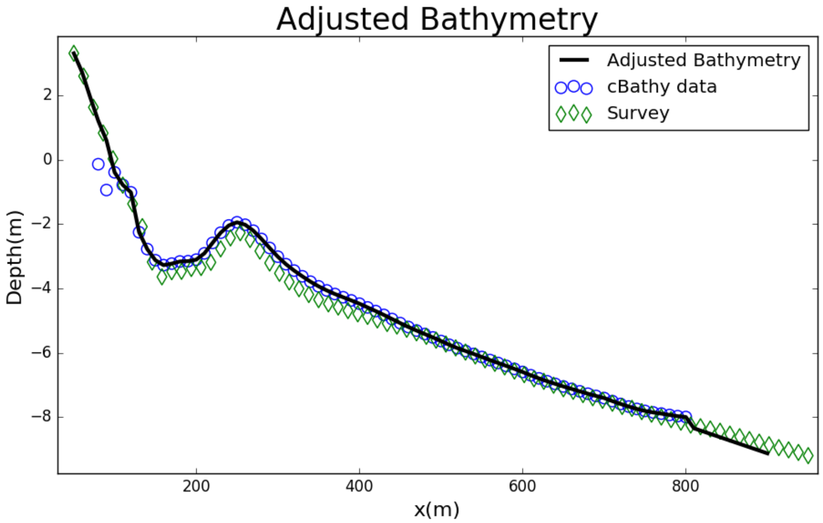


Figure 12: A comparison between the survey ground truth, initial cBathy bathymetry estimate, and the adjusted bathemetry. It's clear to see that the methodology used here produces a reasonable bathymetry to be used with the wave model.

where $c = \frac{\sigma}{k}$. Also, the wave breaking function for this model is

$$\delta = \frac{1}{4h} B \rho g f H_{\text{rms}}^3 \left[(R^3 + \frac{3}{2} R) \exp(-R^2) + \frac{3}{4} \sqrt{pi} (1 - \text{erf}\,(R)) \right] \tag{10}$$

which helps controls the rate at which energy dissipates[11]. We use water density $\rho = 1000$ kg$/m^3$, gravitation constant $g = 9.81$ m$/s^2$, and energy dissipation parameter $B = 1$. Also, the root mean square wave height is defined as $H_{rms} = 0.707H$, and $R = \frac{H_b}{H_{rms}}$, where the maximum wave height is set to $H_b = .78h$.

The wave height profile can be obtained directly from the solution to equation 7. Equation 7 depends on k and σ, and these two parameters must satisfy the following dispersion relation

$$\sigma^2 = gk \tanh{(kh)}. \tag{11}$$

It is also assumed that α propagates through the solution domain according to Snell's law [7]

$$\frac{\sin \alpha_i}{c_i} = \frac{\sin \alpha_{i-1}}{c_{i-1}}. \tag{12}$$

3.3.1 Numerical solution of the wave transformation model

To solve equation 7, a first-order forward finite difference scheme is used. Using $F = EC_g \cos{(\alpha)}$, equation 7 is converted into

$$\frac{d}{dx}(F) = -\delta, \tag{13}$$

and the derivative of F is approximated by

$$\frac{d}{dx} F \approx \frac{F_i - F_{i-1}}{\Delta x}, \tag{14}$$

so that our forward finite difference expression is calculated as

$$F_i = -\delta_{i-1} \Delta x + F_{ii-1}. \tag{15}$$

The initial energy is determined from H at the boundary.

Algorithm 2

```
Calculate $E_0 = \frac{1}{8}\rho g H_{rms,0}^2$
Solve for $k(x)$ for entire domain via $fsolve[\sigma^2 = gk\tanh(kh_0)]$, where $\sigma = \frac{2\pi}{T}$
Evaluate $C_g$ for entire domain using $k(x)$ Evaluate $\alpha(x)$ using: $\alpha_i = \sin^{-1}\left(\frac{C_i \sin(\alpha_{i-1})}{C_{i-1}}\right)$ Evaluate $\delta_0$
Iterate until end of domain is reached using:
    • $E_i = \frac{-\delta_{i-1}\Delta x + F_{ii-1}}{C_{g,i}\cos(\alpha_i)}$
    • $H_i = \frac{1}{0.707}\sqrt{\frac{8E_{i-1}}{\rho g}}$
    • Calculate $\delta_i$ using $H_i$
```

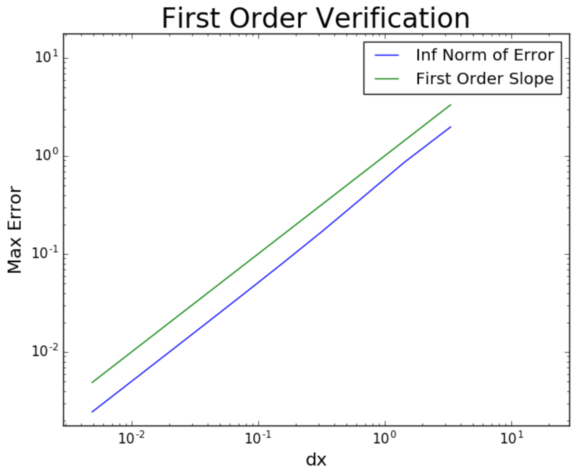


Figure 13: Method of manufactured solutions was used to verify the first order numerical scheme.

3.3.2 Validation of numerical method

The model was validated by using historical boundary conditions from the offshore sensor array and comparing the resulting wave height profile to historical data from the array at the same period. Figure 14 shows that the wave model agrees fairly well with the measured wave height. Figure 15 shows that the wave model agrees fairly well at many time periods for when compared to the 400 meter offshore sensor.

3.4 Time Series Analysis

3.4.1 Forecasting by Time Series

Conventional statistical inference methods (e.g. regression analysis) assume that measurements are independent and identically distributed. However, this is not true in observations which are measured at adjacent time points [13]. Time series is a common approach to model dependency between data points across time and forecast into the future based on historical information.

Problem can arise in time series forecasting when the true distribution is distant from a normal distribution which is commonly assumed in other statistical inferences. Since in time series forecasting one single time point is needed other than the expected mean, the Central Limit Theorem has no chance to operate and the assumption of normality may lead to seriously wrong conclusions. On top of this, it is hard to study the future when random events occur (e.g. hurricane) and this event is not studied through the training data [3].

In the time series case, it is desirable to allow the dependent variable to be influenced by the past values of the independent variables and possibly by its own past values. If the present can be plausibly modeled in

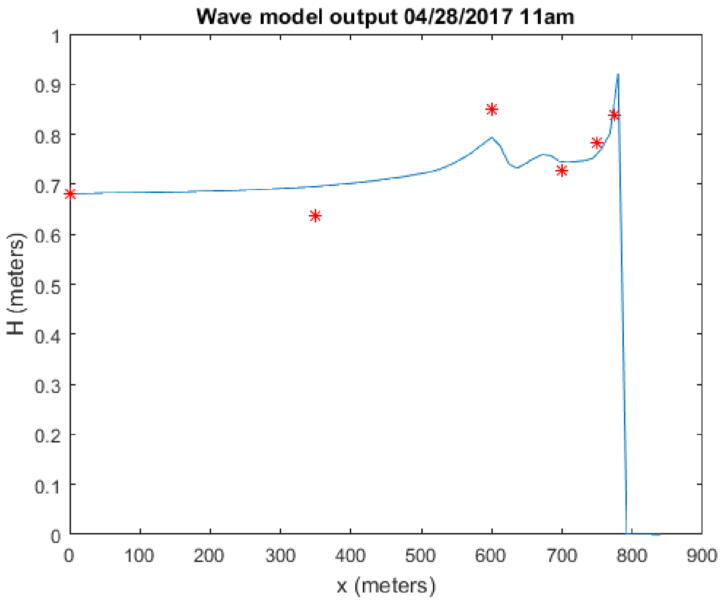


Figure 14: The blue line represents the H profile output obtained from the wave transformation model using measured boundary conditions and survey bathymetry. The red stars represent the measured H values. The computed H is very close to the measured H. The x values start from furthest offshore, and increase as they move on shore.

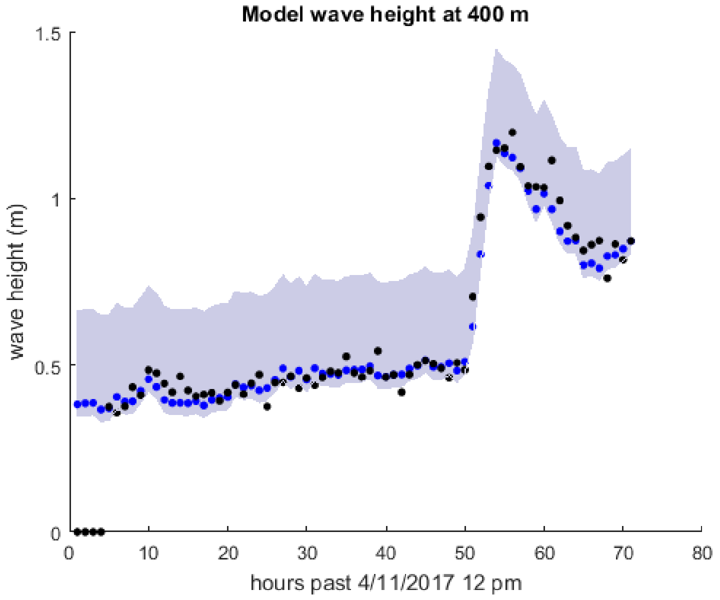


Figure 15: The blue dots represent the computed H at the awac-4.5m sensor from the wave transformation model using measured boundary conditions. The black dots represent the measured wave height values measured at the sensor.

terms of only the past values of the independent inputs, we have the enticing prospect that forecasting will be possible.

3.4.2 Different versions of ARIMA Models

We chose to use an autoregressive integrated moving average model (ARIMA) model to forecast α_t, h_t, and H_t at the boundary for wave equation. Here, subscription t refers to time points that the measurements were taken. ARIMA(p,d,q) model is one of the most commonly used time series models. It added non-stationary model to the ARMA model by taking d th difference of series. ARMA(p,q) model is a mix of AR and MA model, it is based on the idea that the current value of the series, y_t , can be explained as a function of p past values, y_{t-1}, y_{t-2} . . . , y_{t-p} and a linear combination of q errors e_{t-1},e_{t-2}...,e_{t-q}, where p determines the number of steps into the past needed to forecast the current value, and q determines the number steps into past errors. The general formula for ARIMA(p,d,q) model is:

$$x_t = \tau + \sum_{j=1}^{p} \phi_j(x_{t-j}) + \sum_{j=1}^{q} \Phi_j(e_{t-j}) + e_t \tag{16}$$

Let $E(e_t) = \mu$. For a white noise process e_t, the e_t are independent and identically distributed and mean zero. ARMA(p,q) model requires stationary condition, while not all the observation satisfied this condition, so ARIMA(p,d,q) broadens this idea by looking at the difference of d continuous time observations to see if the stationary condition is satisfied. Let $x_t = y_t$ if $d = 0$, $x_t = y_t - y_{t-1}$ if $d = 1$, and if $d = 2$ then $x_t = y_t - 2y_{t-1} + y_{t-2}$. For ARIMA(0,0,q) model has $\tau = \mu$ given the "intercept" line. For $d = 0$ with $p \geq 1$, the "intercept" line gives $\mu = \tau$, and τ needs to be given otherwise $\tau = 0$. The response plot of fitted $\hat{x_t}$ vs x_t should scatter about the identity line with unit slope and zero intercept with no pattern if the model is adequate. The vertical deviation of x_t from the identity line is the residual $\hat{x_t}$ vs $x_t = \hat{e_t}$. A residual plot of $\hat{x_t}$ vs $\hat{e_t}$ or of t vs $\hat{e_t}$ should scatter about the $\hat{e_t} = 0$ line,with no pattern if the model is adequate.

Seasonal patterns often occur in time series. Let s be the seasonal period. Then $s = 12$ for monthly data and $s = 4$ for quarterly data are common.

$x_t \sim$ ARIMA $(p, d, q) \times (P, D, Q)_s$ is the multiplicative seasonal ARIMA model (known as SARIMA) where D is difference of the seasonal effect. Finding a good SARIMA = ARIMA(p,d,q) $\times(P, D, Q)_s$ model given time series y_t is given in the section 3.4.3.

Autoregressive integrated moving average (ARIMAX) models extend ARIMA models through the inclusion of exogenous variables z. We write an $ARIMAX(p, d, q)$ model for some time series data x_t and exogenous data z_t where p is the number of autoregressive lags, d is the degree of differencing and q is the number of moving average lags as:

$$x_t = \tau + \sum_{j=1}^{p} \phi_j(x_{t-j}) + \sum_{j=1}^{q} \Phi_j(e_{t-j}) + \sum_{i=1}^{m} \beta_i(z_{i,t}) \tag{17}$$

Incorporating seasonal effect to ARIMAX models gives SARIMAX models. These four different versions of ARIMA models were compared based on MSE for forecasting wave height. (see section: 4.1.3)

3.4.3 Fitting a good Model

We have to consider many concepts, before selecting ARIMA models. Specially. Autocorrelation function (ACF) at lag k, this is the correlation between series values that are k intervals apart.Partial autocorrelation function (PACF). At lag k, this is the correlation between series values that are k intervals apart, accounting for the values of the intervals between.

For ARIMA(p,d,0) time series the population ACF decays exponentially or like a damped sinusoidal to 0 rapidly. The population PACF has a spike at lag p and usually at lag 1 to $p - 1$. For ARIMA(0,d,q) time series the population PACF decays exponentially or like a damped sinusoidal to 0 rapidly. The population ACF has a spike at lag q and usually at lag 1 to $q - 1$. For ARIMA(p,d,q) time series, the ACF and PACF decay exponentially or like a damped sinusoidal to 0 rapidly. In these cases mostly $d = 0$. Often, if both ACF and PACF do not decay fast or does not have constant mean or for non stationary case $d = 1$.

The response plot of fitted value vs data values should scatter about the identity line with unit slope and zero intercept with no other pattern if the model is adequate. The vertical deviation of x_t from the identity line is the residual. A residual plot of x_t vs residual should scatter about the 0 line, with no other pattern if the model is adequate.

The Akaike information criterion (AIC) is a measure of the relative quality of statistical models for a given set of data. Given a collection of models for the data, AIC estimates the quality of each model, relative to each of the other models. Hence, AIC provides a means for model selection. The model I$_{min}$ with the smallest AIC is always of interest but often over fits. To find out the other best times series model I$_I$ and I$_{min}$ we use the aicmatrix which computes diff(AIC)=AIC(I$_{min}$) - AIC(I$_I$). The initial model to look at is the model I$_I$ with the smallest number of predictors such that diff(AIC)$\leq$ 2. Interesting other candidate models have p+q $\leq$ 3. with entries $\leq$ 7.

An RR plot is a plot of residuals from candidate model1 vs residual from model 2. An FF plot is a plot of fitted values from model 1 vs fitted values from model 2. If the plotted points in the both plots look like the identity line, the simpler model which has less parameters may be better.Ljung-Box statistic is a type of statistical test of whether any of a group of autocorrelations of a time series are different from zero. Instead of testing randomness at each distinct lag, it tests the "overall" randomness based on a number of lags.

Based on theses concepts, good model were fitted and chosen the best fit among other candidate models for each versions of ARIMA, SARIMA, ARIMAX and SARIMAX models. Finding a good ARIMA(p,d,q) $\times(P, D, Q)_s$ model for given time series y_t is similar to getting a good ARIMA(p,d,q) except few things. Following is the process of finding the best ARIMA(p,d,q) $\times(P, D, Q)_s$ model.

1. Plot y_t and determine whether a transformation is needed (Typically the log transformation is used if the variability of response increases with y_t or y_t is right skewed). Also determine if there is a seasonal pattern and y_t is clearly non-stationary so $d = 1$.

2. Want to find d,P, D, and Q. Typically $d \leq 1$, $D \leq 1$, $P \leq 2$ and $Q \leq 2$. Get the candidate difference time series x_t and candidate value of D by the ACF and PACF of x_t.

3. Suppose $D = 0$. Fitting the ARIMA(0,1,0)$\times(1, 0, 1)_s$ model is useful. If the SARIMA parameter is close to 1, take $D = 1$. If 1 seasonal parameter is significant but the other is not, delete the parameter that is not significant.

4. Suppose d and D give the differenced time series x_t. Then plot x_t. If the plot of the differenced time series x_t is clearly non zero, then the R arima function will fit the model wrong since it will not include the an intercept.

5. Use the commands to get the ACF and PACF of x_t. Look at lags s, $2s$, $3s$...Get tentative values of Q and P.

6. Fit a few ARIMA(0,d,0)$\times(P, D, Q)_s$ models with $P, Q \leq 2$ to get P and Q. Look at the output and AIC of the models. Note the coefficient should be at least twice the se in magitude to have a p-value a bit smaller to a lot smaller than 0.05.

7. Suppose d, D, P, Q are set. Now uses the function *saics* to try to find p and q.

8. Find I_{min}, I_I and other models to look at

9. For each candidate model, get the output table, and the response and residual plots. Get the ACF and PACF of the residuals. Check the Ljung Box p-values are above the pval = 0.05 line. Use FF and RR plots to compare candidate models.

10. After getting a good potential ARIMA(p,d,q)$\times(P, D, Q)_s$ model, fit the ARIMA(p+1,d,q)$\times(P, D, Q)_s$ model and the ARIMA(p,d,q+1)$\times(P, D, Q)_s$model that increase p by 1 and q by 1. Check that the estimated parameters do not change much. If the estimated parameters do change much, then the model with $p + 1$ or $q + 1$ may be better than the model with p and q.

3.4.4 Splitting data set to training, testing and forecasting

Data were split into three sections:
Training set (Beginning 01/01/2017 12:00 pm Ending 04/08/2017 11:00 pm)
Testing set (Beginning 04/08/17 12:00 pm Ending 04/11/2017 11:00 pm)
Forecasting (Beginning 04/11/2017 12:00 pm Ending 04/14/2017 11:00 pm)

Table 2: AIC values for different ARIMA(p,1,q) models. ARIMA(2,1,1) has the lowest AIC value. It seems that there is no other candidate models

(p,q)	0	1	2	3
0	478.47	129.93	12.82	8.63
1	16.40	9.44	5.06	4.81
2	7.44	0.00	1.87	3.79
3	4.03	1.87	3.93	5.81

$$\frac{1}{k}\sum_{i=1}^{k}(\widehat{x}_{t+i} - x_{t+i})^2 \tag{18}$$

Models were fit on the training set to find candidate models and has been chosen the best fit for each version ARIMA, SARIMA, ARIMAX, SARIMAX. The chosen model of each version has been tested in the testing period. Then Mean squared error (MSE) was calculated for prediction from each version of the model using Equation 18 , where k is the length of validation set.

3.4.5 Imputing Missing Values

Typically there are three kinds of missing data: missing completely at random (MCAR), missing at random (MAR) and missing not at random (MNAR).

In our case, both MCAR and MNAR occur in training and testing set.

we found wave direction had 42 missing values marked as "NAN" while wave height and period of the same time are present in the same file. Those values are not missing for more than 2 continuous time hours, so we consider them as MCAR, and replace "NAN" with k nearest neighborhood method (KNN) to fit model and make prediction. This solves the problem.

However, we also have MNAR case, i.e. wave height/period/direction of multiple continuous time hours (totally 54 hours) are missing together. For this kind of missing data, we need to survey for the reason of missing. Unfortunately, due to limited time, we are unable to perform the investigation. So we just fit model by ignoring those points.

4 Results

4.1 Results from the time series analysis

4.1.1 Preliminary analysis and fitting ARIMA models

Wave height is used as an example in this section to show how time series model is fitted and validated. The other two variables were fitted in a similar way as wave height. In the next section, forecast results of all three variables wave height, wave period and wave direction are shown.

Time plots of wave height are displayed in Figure 16. Before fitting ARIMA model, log transformation was taken such that the distribution is more symmetric than before transformation (see Fig. 17a and 17b). Following time series ARIMA models were all fitted on log transformed wave height data.

ACF and PACF plots were used to generate rough estimates of the possible values for p,d,q in ARIMA models. ACF plot (Figure 18) reveal slow decreasing trend, which means that the time series is likely to be non-stationary and differencing is needed to make the series stationary. PACF (Figure 19) reveals possible AR lag of 3. After taking the first difference, ACF plot showed faster decaying trend so that stationary assumption holds.(see figure:20) and (see figure:21) Therefore, when fitting the basic ARIMA model, d value was held constant as 1.

Several ARIMA models were fitted and derived AIC values were summarized in Table 2. From this table, ARIMA(2,1,1) has the lowest AIC value. It seems that there is no other candidate models because all the other models which have less parameters than I_{min} have entries greater than 7. Fitted coefficient for this

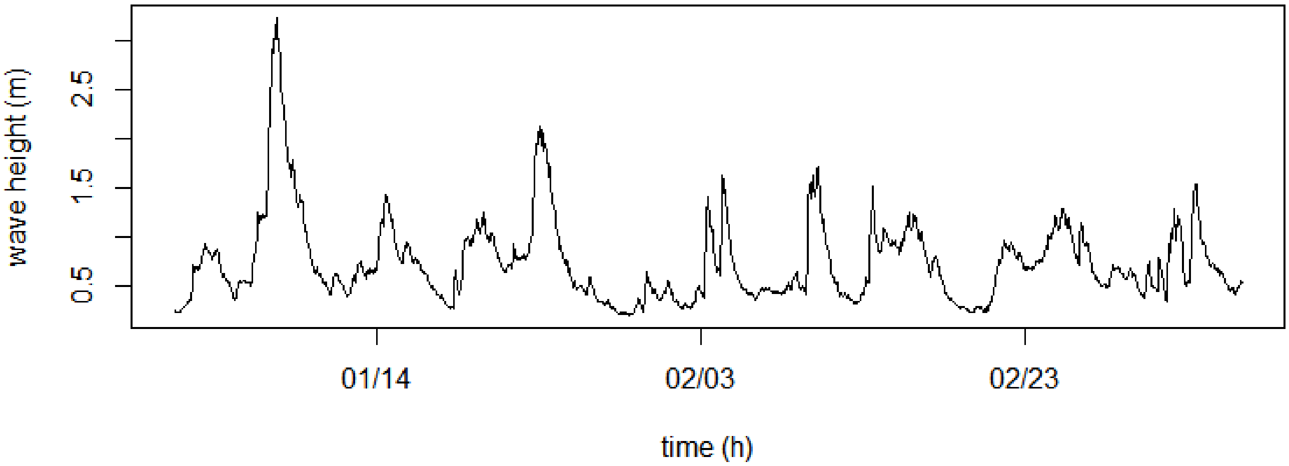
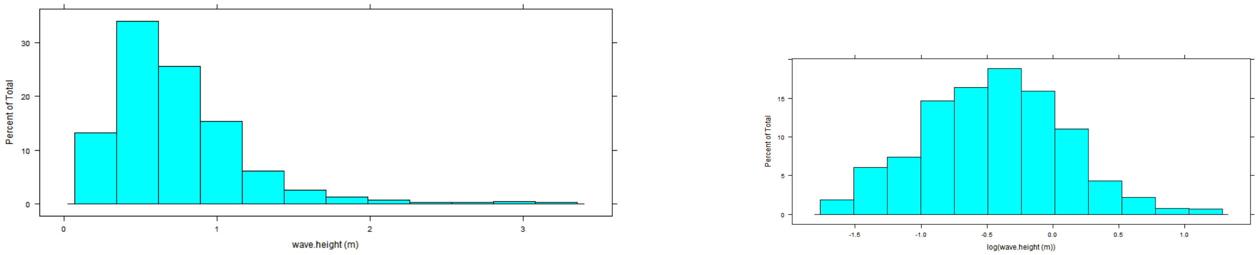


Figure 16: Time plot for wave height between 2017/1/1-2017/4/8 based on the training dataset. Note the trend in wave heights over this time period; the maximum wave height is roughly 3 m.



(a) Frequency histogram for wave height in training set.

(b) Frequency histogram for log-transformed wave height in training set

Figure 17: Frequency histograms of wave height before and after log transformation.

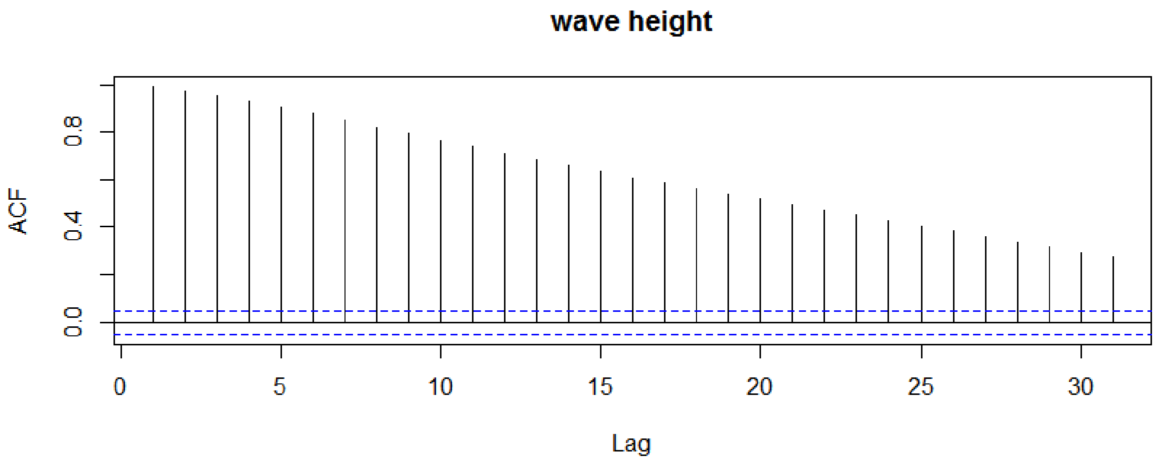


Figure 18: ACF for wave height between 2017/1/1-2017/4/8 (training set).It reveal slow decreasing trend, which means that the time series is likely to be non-stationary

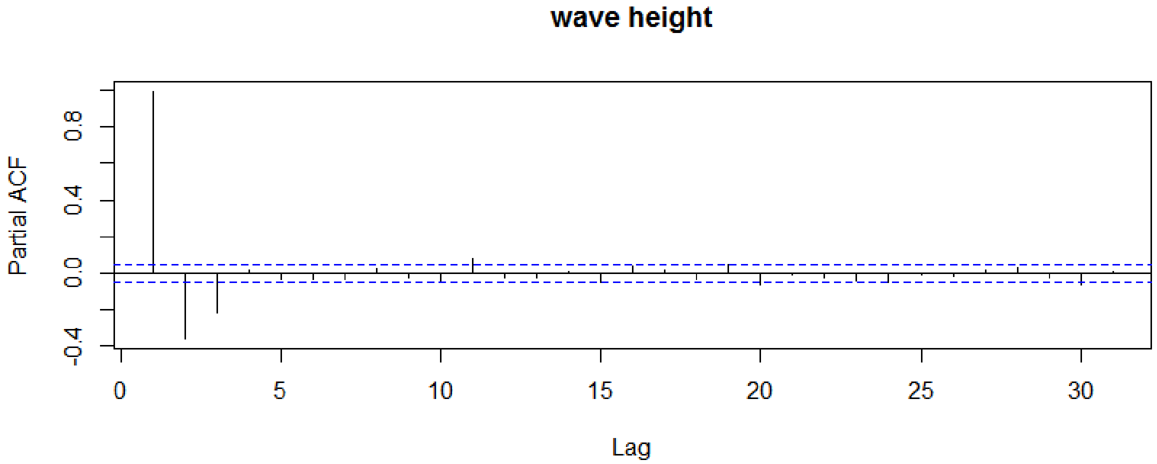


Figure 19: PACF for wave height between 2017/1/1-2017/4/8 (training set)

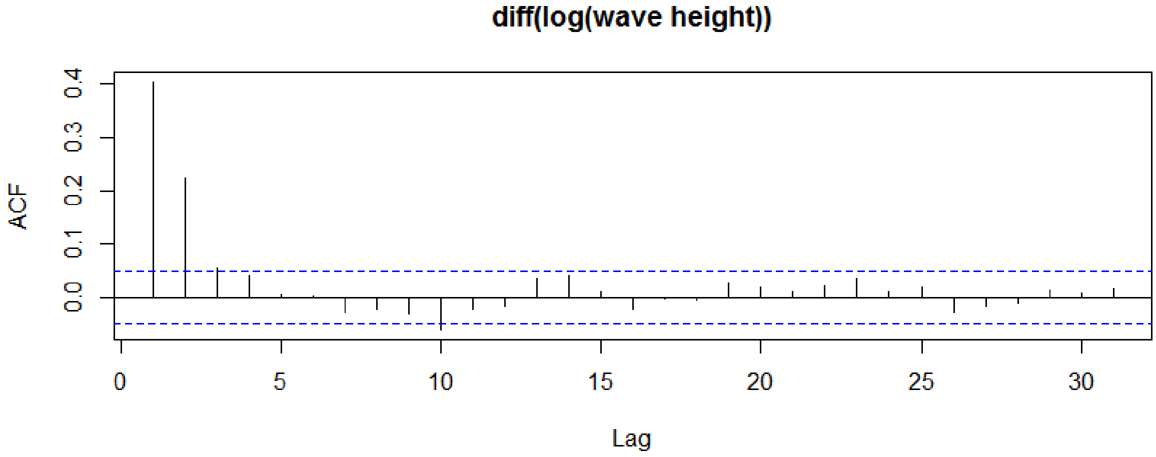


Figure 20: ACF plot for diff(log(wave height)) between 2017/1/1-2017/4/8 (training set). After taking the differencing, ACF plot showed faster decaying trend so that stationary assumption holds.

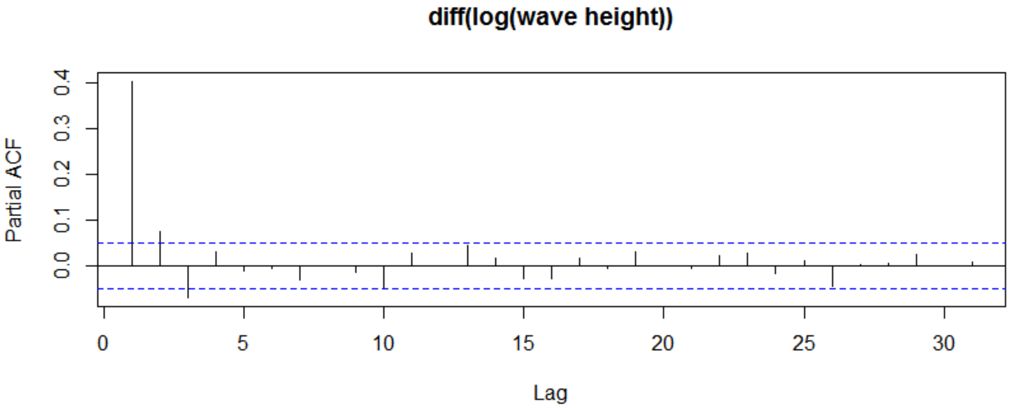


Figure 21: pacf plot for diff(log(wave height)) between 2017/1/1-2017/4/8 (training set). After taking the differencing, PACF plot showed faster decaying trend so that stationary assumption holds.

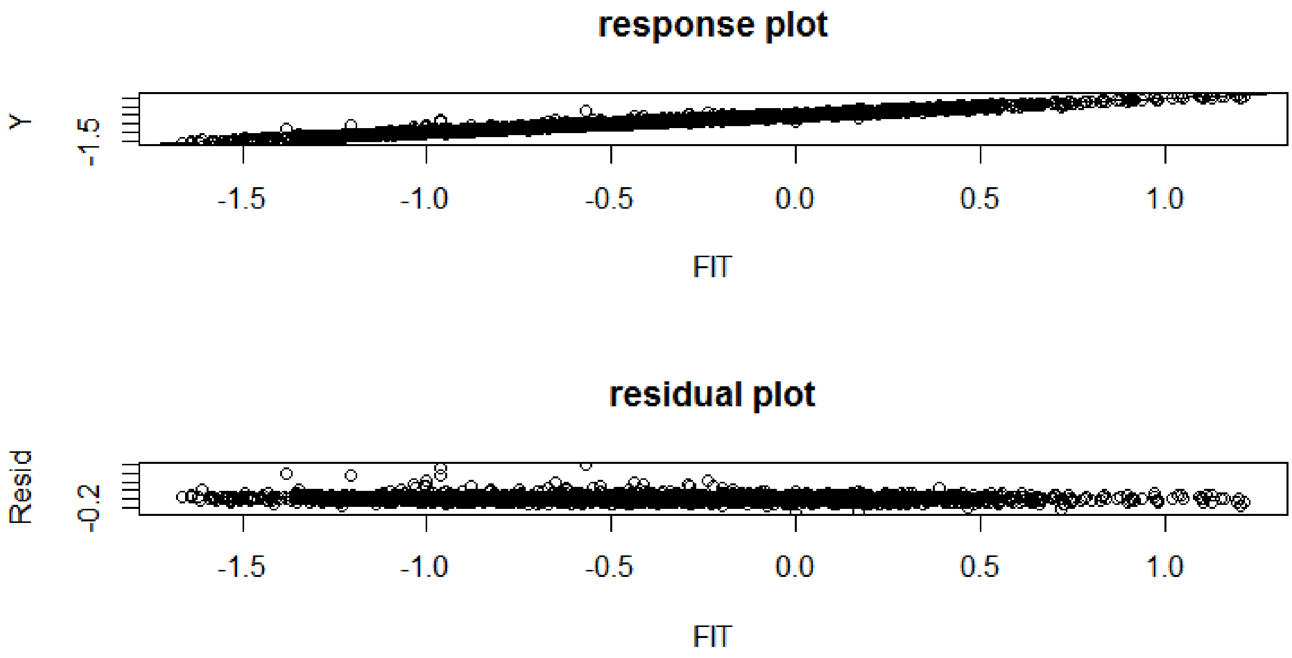


Figure 22: Response for diff(log(wave height)) between 2017/1/1-2017/4/8 (training set).

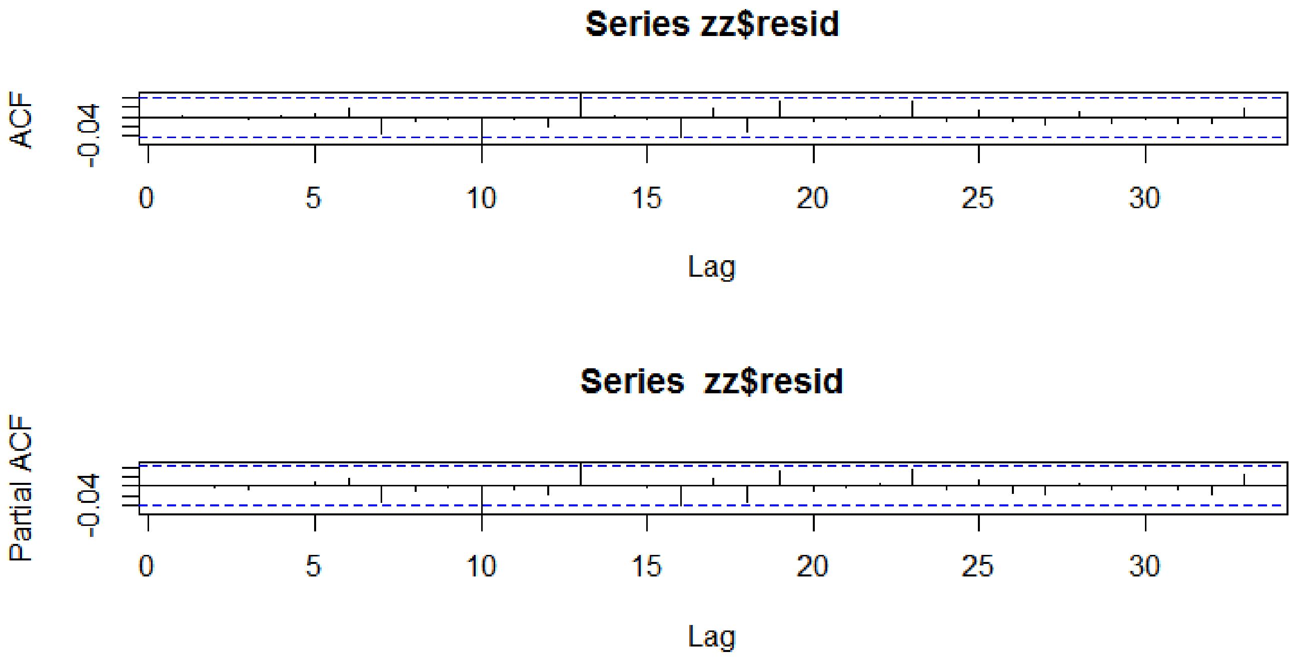


Figure 23: acf/pacf for diff(log(wave height)) between 2017/1/1-2017/4/8 (training set).

Table 3: Coefficients for ARIMA(2,1,1) model. All coefficients have p-values smaller than 0.05, suggesting all coefficients are significant and ARIMA(2,1,1) is potentially good fit for training data.

	estimation	standard error	lower C.I.	upper C.I.	Zvalue	pvalue
ar1	-0.229	0.103	-0.434	-0.024	-2.23	0.0256
ar2	0.340	0.0416	0.256	0.423	8.17	<0.005
ma1	0.630	0.105	0.420	0.840	6.01	<0.005

model was displayed in Table 3. All coefficients have p-values smaller than 0.05, suggesting all coefficients are significant and ARIMA(2,1,1) is potentially good fit for training data.

Moreover, response plot (Figure 22) of actual value Y vs fitted values scatter about the identity line with unit slope and zero intercept with no other pattern. Hence the model we pick is adequate. The residual plot scatter about the zero line with no other pattern if the model is adequate.

residual ACF and PACF plot (Figure 23) suggest that residual seems to roughly approximate a white noise because lags in ACF and PACF are in between $\pm 2SE$ lines.

ARIMA(2,1,1) model tend to have all the p values from the Ljung-Box statistics large and above the dotted line. There fore it can be considered as a good model.

4.1.2 Model Testing

To test the selected ARIMA(2,1,1) model, we considered data from 04/08/2017 at 12 pm to 04/11/2017 at 11.00 pm. According to the Figure:24, It gives a clear picture of how wave height actual values and predicted values varies within that time period. Predicted values from the model were transformed back to exponential values in order to compare with actual values. As a result of that it can be seen that wave height confidence interval (which is in the blue colored region) increases as time index increases. More importantly, it can be conclude that prediction are more accurate for short term horizons than longer term horizons. This is true for predicting wave period (see Fig:25)and wave direction (See Fig: 26). We have considered ARIMA(1,1,1)$\times(1,1,1)_{12}$ for wave period forecasting and ARIMA(4,1,3) for wave direction.

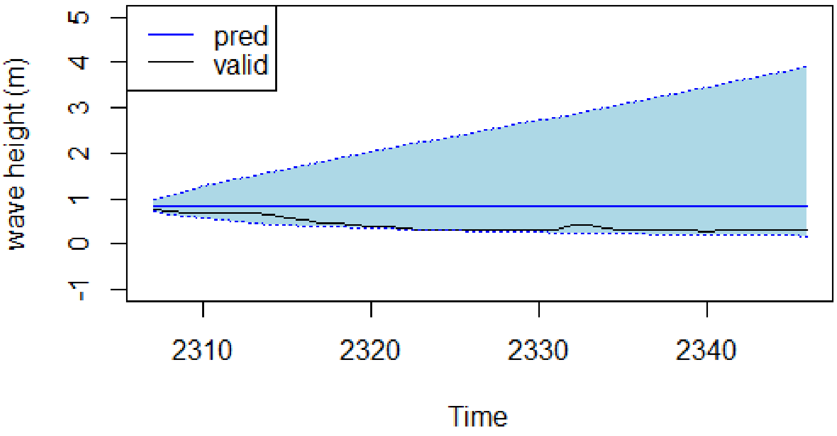


Figure 24: Wave height forecasting 04/08/2017 12 p.m to 04/11/2017 11.00 a.m. Output generated from model ARIMA(2,1,1).

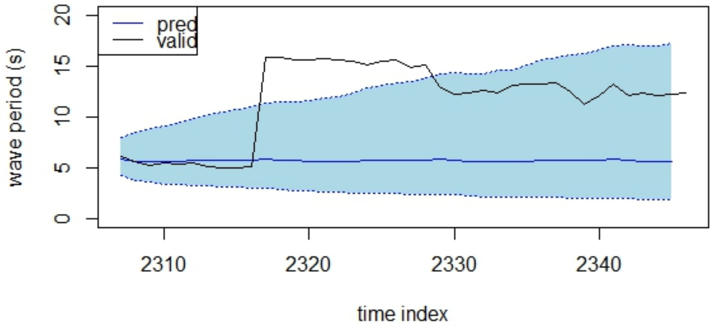


Figure 25: Wave period forecasting 04/08/2017 12 p.m to 04/11/2017 11.00 a.m. Output generated from model SARIMA(1,1,1)×$(1, 1, 1)_{12}$.

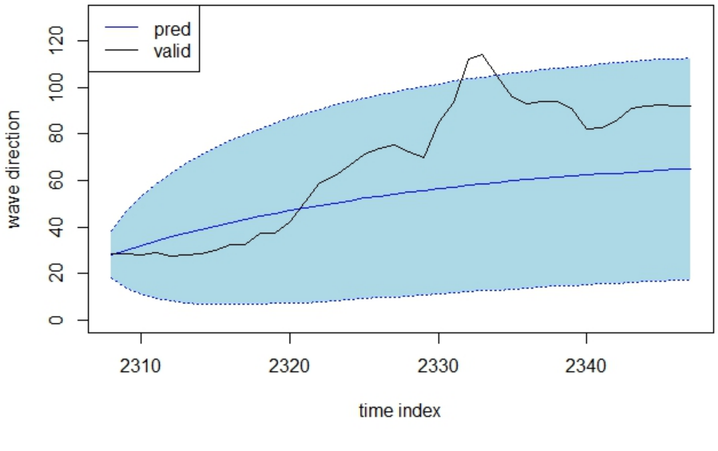


Figure 26: Wave direction forecasting 04/08/2017 12 p.m to 04/11/2017 11.00 a.m. Output generated from model ARIMA(4,1,3).

4.1.3 Comparison of ARIMAX, SARIMA, SARIMAX models

Since wave height is influenced by tide which has cyclical period, incorporating seasonal trend and tide effect into ARIMA could potentially improve model fitting. Thus, ARIMAX models (considering height together with tide), SARIMA models (considering height together with seasonal and dynamic behavior) and SARIMAX models(considering height together with tide and seasonal effects) were developed as an extension of the existing ARIMA model. p,q values changed after taking account into seasonal effect and tide effect. Model selection is mostly similar as in previous section 4.1.1.

Table 4: Comparison of ARIMA, SARIMA, ARIMAX, SARIMAX models.

model	MSE
ARIMA(2,1,1)	0.2031
SARIMA=ARIMA(2,1,1) $\times(0,0,1)_{12}$	0.1983
ARIMAX(1,1,2)	0.2091
SARIMAX=ARIMAX(1,1,2) $\times(0,0,1)_{12}$	0.1995

For the ARIMA and SARIMA fitted models we used log transformation because of the skewness on the other hand, ARIMAX and SARIMAX does not require for log transformation. Table 4 lists of MSE values for these four kind of models.Adding seasonal effect of 12hr period helps to improve the forecasting power slightly, while this may not true with adding tide effects. Although tide data shows some correlation with wave height, including tide data may not help for forecasting. One explanation is that two variables are correlated just from time, not from each other or there may be other variables may affect for wave height forecasting. It can be seen that forecasting is better for SARIMA models because it has lowest MSE value. Figure:27 reveals about graphical representation of wave height forecasting over the period of 04/11/2017 at 12 pm to 04/14 11 am of different version of ARIMA models. It illustrates that SARIMA model is more closer to the actual data trend and therefore it has better forecasting ability compared to other models.

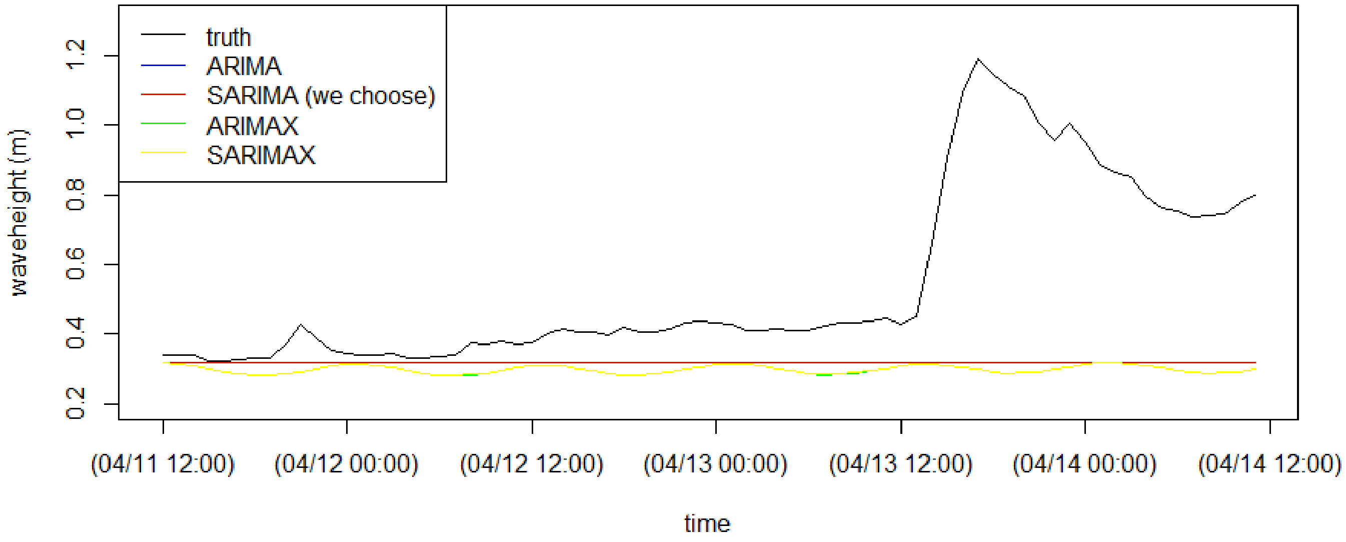


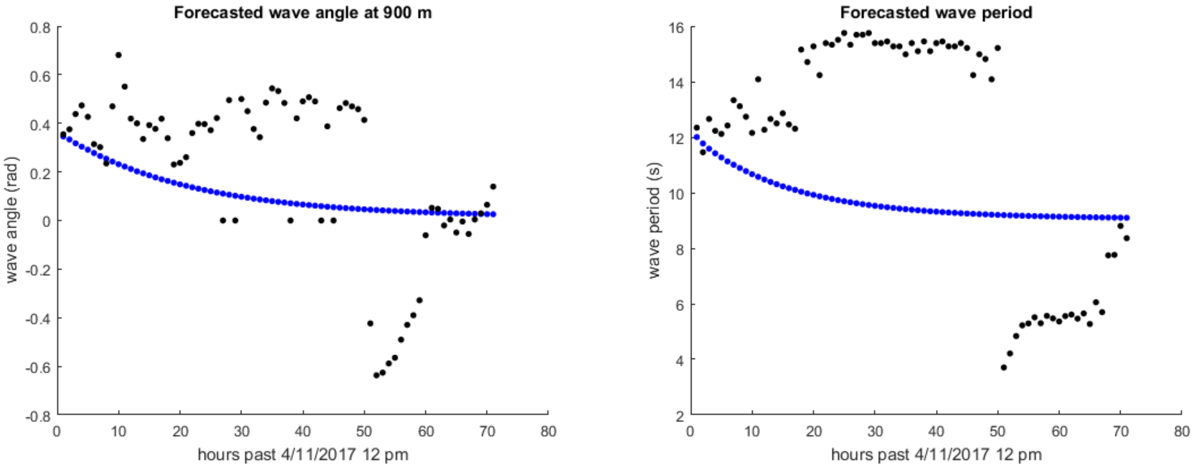
Figure 27: Forecast plot of wave height for all candidate models between 2017/4/11-2017/4/14,where the performance of ARIMA and SARIMA are almost the same, ARIMAX and SARIMAX have only a little difference, hence red and blue, yellow and green, are very close.

4.1.4 Forecasting

Chosen model was refit over the training and validation period of time with fixed p,d, and q values. Final forecast for 04/11/2017 12:00 - 04/14/2017 11:00 were summarized in Results were incorporated into the wave model as boundary condition to generate wave condition .

4.2 Results from wave forecast

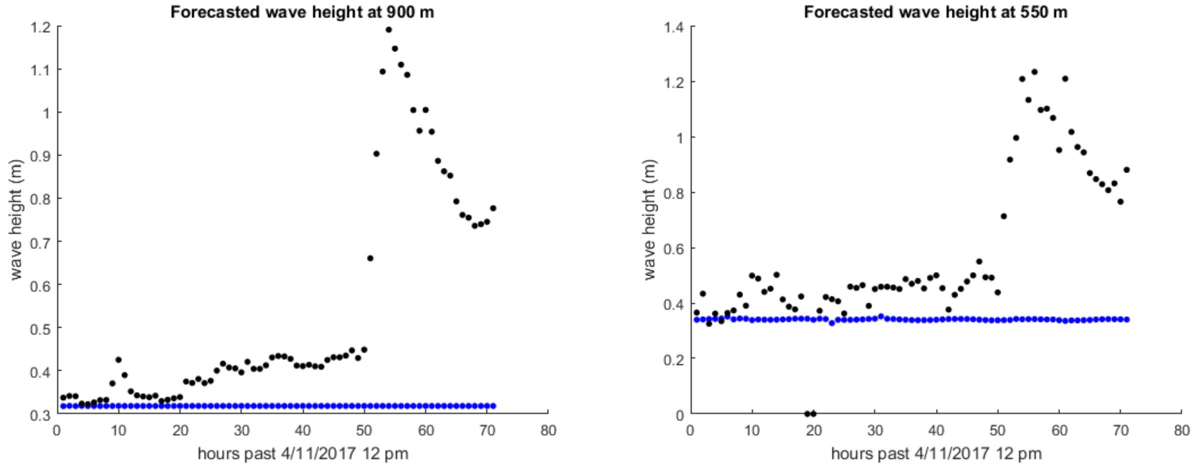
Figure 28 contains the results from the time series forecast for wave period and wave direction at the boundary.



(a) forecastBoundaryAngle

(b) forecastBoundaryPeriod

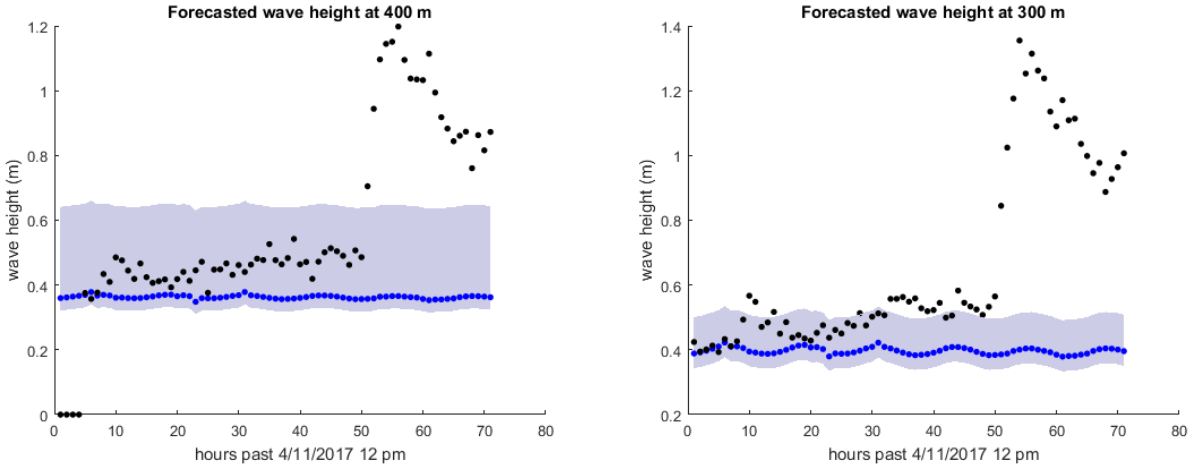
Figure 28: Forecasts of wave direction and wave period at the boundary sensor (offshore 900m).



(a) forecastBoundaryHeight

(b) forecastWave550m

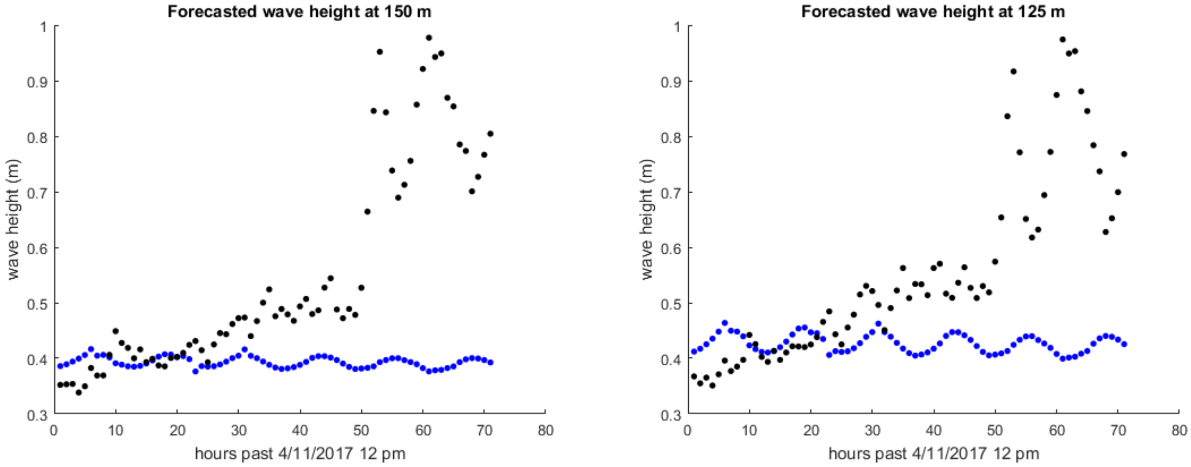
Figure 29: Forecasts of wave height at the boundary sensor (offshore 900m) and at AWAC6 sensor (offshore 550m)



(a) forecastWave400m

(b) forecastWave300m

Figure 30: Forecasts of wave height at AWAC45 sensor (offshore 400m) and at ADOP35 sensor (offshore 300m)



(a) forecastWave150m

(b) forecastWave125m

Figure 31: Forecasts of wave height at XP150 sensor (offshore 150m) and at XP125 sensor (offshore 125m)

5 Summary and Future Work

This project focused on predicting parameters to describe near shore features for beach landing using data collected by the USACE at the FRF. We use a time series model to forecast the wave conditions at 900 meters offshore and use these wave conditions in the wave transformation model to forecast wave height and angle in the problem domain. Our results are only for a one dimensional transect normal to the shore. The maximum wave height can easily be inferred from the output of the wave transformation model.

A robust cBathy algorithm was used to tackle the noisy image problem by making use of suitable convex relaxations. Perhaps the most successful strategy in this sense is provided by trace (or nuclear) norm regularization [5]. However, solving the corresponding optimization problem is computationally expensive for two main reasons. First, many commonly used algorithms require the computation of the proximity operator [4] which entails performing the singular value decomposition at each iteration. Second, and most importantly, the space complexity of convex solvers grows with the matrix size, which makes it prohibitive to employ them for large scale applications. Hence, we may provide an efficient algorithm with running time that matches SVD while nearly matching the best-known robustness guarantees. By combining this robust cBathy output and survey data, bathymetry was reconstructed. This has the benefit in that all data to infer bathymetry could be obtained via remote sensing. From this bathymetry, beach slope, sandbar location, and sandbar depths were determined using simple algorithms.

As far as statistical forecasting is concerned, it would be more accurate if we could incorporate inversion methods such as the Bayesian Markov Chain Monte Carlo approach, which samples a posterior distribution of parameters. There is a range of potential applications of Machine Learning, and it may be more useful to predict the probability distribution for a variable rather, predicting the most likely value for that variable [6]. Quantifying the uncertainty in each prediction of the parameters would be ideal, so that a likelihood range of wave parameters could be provided. Moreover, cBathy also gives values for wave number, wave direction, depth and wave period as vectors. By combining theses values together into a matrix, we can do time series Multivariate ARIMA and ARIMA-X Analysis using the software package 'marima' in R, which may provide more accurate predictions. On the other hand, a time series model for cBathy output can be solved as a least square problem. Since it has considerably high dimension, general methods of solving least-Squares may not work. Therefore Sketching methods can be used to transform high dimension to low dimension[12]. Another way to improve statistical inference is to allow for time-varying parameters in the model. ARIMA model can be easily extended to Dynamic Linear Models (DLM) along with kalman filter after represented in the state-space format. DLM is more flexible than ARIMA models that assume constant coeffcients across time. Local variation can be better captured by DLM than traditional ARIMA methods.

6 Acknowledgements

We acknowledge National Science Foundation (NSF) for funding. We thank SAMSI and NCSU for support and Thomas Gehrmann for organizing. We also appreciate the guidance and patience of our mentors Ty Hesser, Matthew Farthing of the USACE and Lea Jenkins of Clemson University. Finally, we thank the U.S.Army Corps of Engineers for providing a novel problem and interesting data from which to process image data and build forecast models.

References

[1] Field manual 55-50: Beach and weather characteristics. `http://www.globalsecurity.org/military/library/policy/army/fm/55-50/index.html`.

[2] An introduction to larc survey system. `http://frf.usace.army.mil/larc/larcsystem.shtml`.

[3] Introduction to time-series regression. `http://node101.psych.cornell.edu/Darlington/series/series1.htm`. Accessed: 2017-07-22.

[4] Stephen Boyd, Neal Parikh, Eric Chu, Borja Peleato, and Jonathan Eckstein. Distributed optimization and statistical learning via the alternating direction method of multipliers. *Foundations and Trends® in Machine Learning*, 3(1):1–122, 2011.

[5] Emmanuel J Candès, Xiaodong Li, Yi Ma, and John Wright. Robust principal component analysis? *Journal of the ACM (JACM)*, 58(3):11, 2011.

[6] M. Carney, P. Cunningham, J. Dowling, and C. Lee. Predicting probability distributions for surf height using an ensemble of mixture density networks. In *Proceedings of the 22nd International Conference on Machine Learning*, pages 113–120. ACM, 2005.

[7] Jean-Christophe Gonzato and Bertrand Le Saëc. *A phenomenological model of coastal scenes based on physical considerations*, pages 137–148. Springer Vienna, Vienna, 1997.

[8] R. Holman, N. Plant, and T. Holland. cbathy: A robust algorithm for estimating nearshore bathymet-javascript:void(0);ry. *Journal of Geophysical Research: Oceans*, 118(3):2595–2609, 2013.

[9] R. Holman, N. Plant, and T. Holland. cBathy: A robust algorithm for estimating nearshore bathymetry. *Journal of Geophysical Research: Oceans*, 118(5):2595–2609, 2013.

[10] Arthur T Ippen. Estuary and coastline hydrodynamics. 1966.

[11] TT Janssen and JA Battjes. A note on wave energy dissipation over steep beaches. *Coastal Engineering*, 54(9):711–716, 2007.

[12] Mert Pilanci and Martin J Wainwright. Iterative hessian sketch: Fast and accurate solution approximation for constrained least-squares. *The Journal of Machine Learning Research*, 17(1):1842–1879, 2016.

[13] R.H. Shumway and D.S. Stoffer. *Time series analysis and its applications: with R examples*. Springer Science & Business Media, 2010.

[14] Ronald E Walker. *Marine light field statistics*, volume 21. Wiley-Interscience, 1994.