Statistical Analysis and Modeling of Red Tide Blooms

R. D. Wooten and C. P. Tsokos

Department of Mathematics and Statistics, University of South Florida, Tampa, Florida 33620, USA
E-mail: rwooten@cas.usf.edu

Abstract

The present study considers the random phenomenon that is Red Tide as found around the State of Florida. Among the many organism that make up Red Tide, Karenia Brevis is the organism commonly associated with an outbreak, the probability distribution which best describes the behavior of Karenia Brevis is the Weibull probability distribution. There are regional differences as well as regional relationships including delay effects. Recursion rates indicate a logistic growth model; however, additional information is needed before research can determine the effects of runoff on Red Tide blooms.

1 Introduction

Red Tide is becoming a matter of concern in the State of Florida. There are the environmental and health aspects that result from red tide that are of significant importance. In addition, red tide plays a major role in the fishing industry and tourism of the State of Florida which have a significant impact on the State's economy. There are many different microorganisms responsible for red tide release toxins (Steidinger, 1983) into both the water and air. These neurotoxins can affect the respiratory and cardiac system (Rounsefell, Nelson, 1966), reducing blood flow and slowing down the heart. Documented as early as the 1800s, large blooms killing thousands of fish. Moreover, with the modern advances in marine technology, blooms are becoming easier to monitor as well as storage of the gathered information and sophisticated computational powers to run statistical algorithms to analyze and interpret the data is continuously increasing.

In the present study, we begin with available to perform descriptive statistics, which are simple statistics, which describe the number of organism within a sample and established that the best way to analyze the data is through a logarithmic filter that reduces the scale and homogenizes the variance. Utilizing historical data gathered sporadically over the past several decades, we determined that the Weibull probability distribution

[0] Key Words: Karenia Brevis, Weibull probability distribution, Logistic Growth, Delay Effects, Recursion Analysis.

best characterize the magnitude of a bloom; that is, the probabilistic behavior of the logarithmic transformation of the count of the organism *Karenia Brevis*, the primary organism found in Red Tide (Dixon, 2003).

Secondly, we use inferential statistic to determine regional differences (Steidinger, 1973). Next, we used recursion to estimate logistically — according to the logistic growth model — the rate at when a bloom grows. That is, estimating the subject response (magnitude of a bloom) depending on the percent of the total capacity taken up by the current bloom and the remaining percent of the total capacity available.

Furthermore, we proceed to establish a relationship between nutrient runoff from the State of Florida and the magnitude of a Red Tide Bloom. That is, we developed a statistical model of the subject response (magnitude of bloom) as a function of soil nutrients that wash into the oceans: *Sulfate* SO_4, *Nitrate Ion* NO_3 and *Ammonium Ion* NH_4; these minerals are common in fertilizers used in agriculture. The present study can be extended to include more precise statistical models on the subject response once consistent concurrent data is gathered; that is, we need to establish a data bank where not just the organism count and date are recorded but salinity, water temperature, and other contributing variables on the same temporal scale.

The present analysis is important to the State of Florida on both an environmental and economical point of view. Accurately estimating the size of a bloom enable us to accurate post warnings in areas affect by and outbreak, and a better understanding of the contributing entities that the subject response; namely, the size of a bloom will lead to understanding of the cause and effect of Red Tide.

In the present study we will address the following issues:

1. *What constitutes Red Tide?*
2. *What is the probability distribution of the magnitude of an outbreak that describes its behavior?*
3. *Recursion analysis: Logarithm Logistic Growth.*
4. *What is the relative growth rate of a bloom in terms of magnitude?*
5. *What are the regional differences in terms of magnitude of the bloom? Evaluate and make statistical inferences regional; is the mean magnitude of a bloom in Pensacola the same as the mean magnitude of bloom in the St. Petersburg area.*
6. *Determine the key contributing variables that drive Red Tide blooms.*
7. *Identify interactions that exist between the key attributing variables and higher order terms.*
8. *Determine additional information needed to be gathered by scientist on new or existing data.*

2 Analysis of the Various Organisms Measured in Red Tide

There are 57 various generalized organisms found in over $56,000$ samples taken over a forty-eight year time spanning, 1954 through 2002; only thirty-one organisms are recorded in at least ten samples and twenty-one organisms are recorded in at least one-hundred samples; Table 1 includes a general group of “other plankton”. Only ten organisms found in at least one thousand samples that we are considering: *Karenia* (more specifically *Karenia Brevis*), Diatom, Other Plankton, Gymnodinium, Dinoflagellates, Micro-flagellates, Gyrodinium, Ciliates, Gonyaulax, and *Peridinium*.

In the present study, we will concentrate on *Karenia Brevis* (formerly *Gymnodinium breve*). This organism, when present in sufficient numbers (thousands or millions of cells per milliliter) turns the water red invoking the name Red Tide. Little is known about Red Tide and its cause and effects. One question to be addressed in this study is whether Red Tide blooms around the State of Florida are correlated, possibly with a time delay, to the run-off from the State of Florida. First, we must analyze the main organism associated with Red Tide: namely *Karenia Brevis*.

Descriptive Statistics for Karenia Brevis — Let $c(t)$ be the concentration of *Karenia Brevis* at time t, where t is measured in days since January 1, 1954. We have a sample of these concentrations; namely, $x_i = c(t_i)$ for the various samples taken at various times t_i. If we consider the raw count of this data, then there is an extreme skew in the data as shown in Fig. 1. Fig. 1 is important because it illustrates that a common condition found is that no *Karenia Brevis* exist in most of the samples. There are numerous samples with zero count; that is, $x = 0$ and therefore we consider $x > 0$ and the natural logarithm of the count (concentration[1]), $\ln c(t)$ will be considered to adjust the scale and bring the underlying distribution into focus. Hence, this study analyzes the conditional probability distribution of the magnitude of a bloom, given there is a count of *Karenia Brevis* is greater than zero; that is, this organism is in fact present (at least one), see Fig. 2.

Fig. 2 gives more insight into the true nature of *Karenia Brevis*, but there are many records detecting a single organism; illustrated in Fig. 2, over 1500 samples contained exactly one organism. Hence, consider when there is a bloom — meaning the count is more than one as shown in Fig. 3. In Fig. 2, the calculated sample mean magnitude of Red Tide bloom is 9.097, whereas in Fig. 3, given that there is a bloom (more than one organism recorded) the sample mean magnitude of Red Tide bloom is calculated as 10.118. Furthermore, the significantly reduced variance allows for a better estimate of the mean magnitude of a bloom.

Consider the three subgroups: (a) no organism found (N), (b) exactly one (a single) organism present (P) and (c) finally, when the organism is in bloom (B). Rarely is there a single organism present (P); barely 3% of the samples recorded a count of one, see Table 2. Normally, that is in the majority (approximately 70%) of the samples, there are not even a single *Karenia Brevis* present. Only an estimated 27% of the samples recorded is or contains a bloom; that is, more than one organism. Table 2 below gives estimates of the percent of the data in the defined subgroups.

[1] Count per sampled liter.

When we further consider the count, there is a disparity between the minimum statistic, $\min_i \{\ln x_i\} = 0.693$ and the bulk of the remaining values as illustrated by the gap in the histogram shown by Fig. 2. Consider outliers defined by Chebyshev's inequality: $P\{|x-\mu| \geq t\} \leq \frac{\sigma^2}{t^2}$ which for $t = 8.667$ yields $P\{|x-10.118| \geq 8.667\} \leq 0.1$, which implies that twenty-one outliers are present, of which only two are extreme highs leaving nineteen lower level outliers. All nineteen of these extreme outliers are $x \in \{2,3,4\}$ whereas the upper outliers are $x \in \{197656000, 358000000\}$. Empirically, these outliers constitute approximately 0.14% of the data count in bloom.

Hence, we will redefine the categories: first, NO organism found (N) $x = 0$, organisms Present (P), but few, $0 < x \leq 5$ and finally when the organism is in full bloom (B) $x > 5$. This redefining does not significantly affect the percentages in each category as given in Table 3, but does remove the gap in the histogram Fig. 3 and illustrated by Fig. 4. In Table 3, more precise estimates of the percent of the data in the redefined subgroups; that is, the redefinition of a bloom to be $x > 5$ instead of $x > 1$, yields the descriptive statistics given the chart along with the histogram illustrated by Fig. 4. While these statistics, the mean, the median, the standard deviation and the variance are all extremely close, the main difference is illustrated in that there is no gap in the data in Fig. 4.

Parametric Inferential Analysis — Consider $\ln(\ln(x))$ for the samples where *Karenia Brevis* is in full bloom and x is the concentration of *Karenia Brevis* in a given sample. This transformation of the data helps indicates some form of an extreme value distribution. The curvature of the normal probability plot given in Fig. 5 indicates that the Weibull probability distribution would be a good fit. In fact, that the Weibull probability distribution as shown by Fig. 6 is the only distribution at the 0.01 levels that cannot be rejected.

Thus, the best probabilistic characterization of the existing data designated by x is the logarithmically transformed Weibull probability distribution function.

Logarithmic transformation and its properties — The data that characterizes as the organism count in *Karenia Brevis* have such a large scale that a logarithmic transformation must first be taken to consider the probability distribution to be useful. This transformed data will be referred to as the magnitude of the data; that is, $\{\ln x_i\}_{i=1}^N$ which need not be based on the natural logarithm — this can be adjusted as needed to a general base $\{\log_b x_i\}_{i=1}^N$. In this study, we will use the natural logarithm.

Assuming $x_i \geq 1$, defined $\{y_i\}_{i=1}^N$ as the magnitude of the original data set, where $y_i = \ln x_i$. Further, assume this transformed data is best fit by the two-parameter Weibull; that is, $y \sim W(\theta = 0, \lambda, \alpha)$, than consider the cumulative probability density function given by

$$F_Y(y) = \left\{ \begin{array}{cc} 1 - \exp\left\{-\left(\frac{y}{\lambda}\right)^\alpha\right\}, & y \geq 0 \\ 0, & \text{otherwise} \end{array} \right. , \tag{1}$$

where α is the shape parameter and λ is the scale parameter.

Then in terms of the original data, the transformed cumulative probability distribution function, given by equation (2), yields the transformed probability distribution function given by equation (3). That is, we have

$$\begin{aligned} F_X(x) &= P\{X \le x\} = P\{\ln X < \ln x\} \\ &= P\{Y < \ln x\} = F_Y(\ln x) \\ &= 1 - \exp\left\{ - \left(\frac{\ln x}{\lambda} \right)^{\alpha} \right\}, \end{aligned}$$

and

$$F_X(x) = 1 - \exp\left\{ - \left(\frac{\ln x}{\lambda} \right)^{\alpha} \right\}. \tag{2}$$

The probability density function of x is given by

$$f_X(x) = \frac{\alpha}{\lambda x} \left(\frac{\ln x}{\lambda} \right)^{\alpha - 1} \exp\left\{ - \left(\frac{\ln x}{\lambda} \right)^{\alpha} \right\}, \tag{3}$$

This satisfies the condition for being a probability density function,

$$0 \le f_X(x) \le 1, \text{ for } x \ge 1$$

and

$$\int_1^{\infty} f_X(x)\, dx = 1.$$

Proof of Property 1. Since $f_Y(y)$ is a probability density function such that $0 \le f_Y(y) \le 1$ for $y \ge 0$. Using the defined transformation, we have $0 \le f_Y(\ln x) \le 1$ for $\ln x \ge 0$; that is $x \ge 1$. Hence, dividing by x, $0 \le \frac{f_Y(\ln x)}{x} \le \max_{x \ge 1} \frac{1}{x} = 1$; that is, $0 \le f_X(x) \le 1$, for $x \ge 1$. □

Proof of Property 2. This transformed distribution $f_X(x)$ is a probability density function. Since the transformation is continuous $f_X(x) = \frac{f_Y(\ln x)}{x}$, we can write

$$\int_1^{\infty} f_X(x)\, dx = \int_1^{\infty} \frac{f_Y(\ln x)}{x}\, dx = \int_0^{\infty} f_Y(y)\, dy = 1.$$

□

Maximum Likelihood Function

$$L(x) = \frac{\alpha^n}{\lambda^n \prod_{i=1}^{n} x_i} \prod_{i=1}^{n} \left(\frac{\ln x_i}{\lambda} \right)^{\alpha - 1} \exp\left\{ - \sum_{i=1}^{n} \left(\frac{\ln x}{\lambda} \right)^{\alpha} \right\}$$

or

$$\ln L(x) = n \ln \alpha - n \ln \lambda - \sum_{i=1}^{n} \ln x_i + (\alpha - 1) \sum_{i=1}^{n} \frac{\ln x_i}{\lambda} - \sum_{i=1}^{n} \left(\frac{\ln x_i}{\lambda} \right)^{\alpha}.$$

The MLEs of λ and α are obtained by solving the following system of two equations:

$$\frac{\partial \ln L(x)}{\partial\lambda} = -\frac{n}{\lambda} - (\alpha-1)\sum_{i=1}^{n}\frac{\ln x_i}{\lambda^2} - \alpha\sum_{i=1}^{n}\frac{1}{\lambda x_i}\left(\frac{\ln x_i}{\lambda}\right)^{\alpha-1} = 0$$

and

$$\frac{\partial \ln L(x)}{\partial\alpha} = \frac{n}{\alpha} + \sum_{i=1}^{n}\ln x_i - \sum_{i=1}^{n}\ln\left(\frac{\ln x_i}{\lambda}\right)\left(\frac{\ln x_i}{\lambda}\right)^{\alpha} = 0.$$

Unfortunately, it is not easy to solve this system of equations to optimize this likelihood function analytically. However, with the advent of recent technologies we can accurately obtain estimates of the solution of the above equation using iterative procedure (Qiao, Tsokos, 1995).

The j^{th} moment of the random variable x is given by

$$\begin{aligned} E_X\left(x^j\right) &= E_X\left(e^{j\ln x}\right) \\ &= \int_0^\infty e^{j\ln x} f_X(x)\,dx = \int_0^\infty e^{j\ln x}\frac{f_Y(\ln x)}{x}\,dx \\ &= \int_1^\infty e^{jy} f_Y(y)\,dy = E_Y\left(e^{jy}\right). \end{aligned}$$

Thus, we can use the above expression to obtain estimates of the basic statistics of the phenomenon of interest.

Let $MGF_Y(t) = E\left(e^{yt}\right)$ be the moment generating function for the standard two-parameter Weibull in terms of the variable y. Under the given transformation, the moment generating function for the transformed probability density function can be expressed into terms of the original distribution. That is, we have

$$\begin{aligned} MGF_X(t) &= E_X\left(e^{tx}\right) \\ &= \int_1^\infty e^{tx} f_X(x)\,dx = \int_1^\infty e^{tx}\frac{f_Y(\ln x)}{x}\,dx \\ &= \int_0^\infty e^{te^y} f_Y(y)\,dy = E_Y\left(e^{te^y}\right) \\ &= MGF_Y\left(e^y\right) \end{aligned}$$

Two-Parameter Weibull Probability Distribution Function — Using numerical schemes, we can estimate the two-parameter Weibull. Applying the MLE yields a scale parameter estimate of $\hat{\lambda} = 11$ and shape parameter estimate of $\hat{\alpha} = 4.2$ as shown below,

$$F(x) = \begin{cases} 1 - \exp\left\{-\frac{x^{4.2}}{11}\right\}, & x > 5 \\ 0, & \text{otherwise.} \end{cases} \tag{4}$$

From the moment generating function, $E\left(x^n\right) = \lambda^n \Gamma\left(1+\frac{n}{k}\right)$, we can estimate the sample mean, sample standard deviation, skewness and kurtosis for each of the probability distribution function and are shown in Table 4.

The three-parameter Weibull probability distribution, which gives a best-fit probability distribution using MLE yields a threshold $\hat{\theta} = 1.693114$, scale parameter $\hat{\lambda} = 9.395226$, and shape parameter $\hat{\alpha} = 3.484966$. This yields the cumulative probability distribution function given by

$$F(x) = \begin{cases} 1-\exp\left\{-\frac{(x-1.693114)^{3.484966}}{9.395226}\right\}, & x > 5 \\ 0, & \text{otherwise.} \end{cases} \tag{5}$$

The difference between the two and three-parameter Weibull probability distributions is insignificant. When considering the simple regress between the empirical probability distribution and the two and three-parameter Weibull, 99.4% of the variation in the empirical probability is explained by the estimated two-parameter Weibull probability distribution whereas 99.5% is explained by the three-parameter Weibull probability distribution. Both of these probability distributions accepted at the same level of significance, results shown in Table 5 below.

Since these two probability distributions are extremely close in their estimate, we will invoke the law of parsimony and continue with the two-parameter Weibull probability distribution function. Thus, we can proceed to estimate, given *Karenia Brevis* is present, the probability of exceeding a given count in a given sample as shown in Fig. 7.

In any given sample in which *Karenia Brevis* is present, as few as 22 organisms could be present. However, in every ten samples in which *Karenia Brevis* is present, this number jumps to $660,000$ in count or $\ln x \approx 13.4$. In every hundred samples in which *Karenia Brevis* is present, this number jumps again to $7,452,052$, which implies $\ln x \approx 15.8$; this is an increase of over 11-fold. Recall, that there were over $56,000$ samples taken over a 48-year period, at this rate there are up to 1200 samples taken in a year. Therefore, consider the return periods assuming 1200 samples per year, then in any given year a sample could contain upwards of $41,243,332$ count of *Karenia Brevis*: that is, $\ln x \approx 17.5$. Up to six times that found in every hundred samples; therefore consider the return periods for $\ln x > 17$. That is, $x > 39,824,784$, see Fig. 7 and Table 6.

The maximum $\ln x$, as shown in Table 6 above, has a high of magnitude 19.755 or $379,741,000$ count in a given sample has an estimated return period of between 90 and 100 years.

Mixed Probability Distributions — Further study of the histogram, shown by Fig. 8, indicates a bimodal behavior and therefore, a mixture of the two normal probability distributions might yield an even better fit. The mixture of two normal probability distributions is given by

$$g\left(x \mid \mu_1, \sigma_1, \mu_2, \sigma_2, \alpha\right) = \alpha f\left(x \mid \mu_1, \sigma_1\right) + (1-\alpha) f\left(x \mid \mu_1, \sigma_2\right). \tag{6}$$

The expected value and variance of the mixed probability distribution functions are given, respectively, by

$$E_g(x) = \alpha E_{f_1}(x) + (1-\alpha)E_{f_2}(x)$$

and

$$V_g(x) = \alpha V_{f_1}(x) + (1-\alpha)V_{f_2}(x).$$

Both of these properties follow from the following relationship within the moments, that is,

$$\begin{aligned} E_g\left(x^p\right) &= \int_R x^p g\left(x \mid \mu_1, \sigma_1, \mu_2, \sigma_2, \alpha\right) dx \\ &= \alpha \int_R x^p f\left(x \mid \mu_1, \sigma_1\right) dx + (1-\alpha) \int_R x^p f\left(x \mid \mu_1, \sigma_1\right) dx \\ &= \alpha E_{f_1}\left(x^p\right) + (1-\alpha) E_{f_2}\left(x^p\right). \end{aligned}$$

Hence, we can compute an estimate of the expected value $\hat{E}_g(x) = E(x) = 10.13$ and variance $\hat{V}_g(x) = V(x) = 2.72$. First we can estimate one of the peaks by considering the mode of the magnitudes, $M = 6.9$, which in a normal distribution gives an indication to the potential first mean and since the second peak is more certain and can be estimated as $\hat{\mu}_2 = 12.5$ this will be the initial mean. If we further assume that the sample standard deviations are the same; that is, $\hat{\sigma}_1 = \hat{\sigma}_2 = \hat{\sigma}$, then we can use least squares regression to estimate the mixing factor, α. That is, consider the mixed model given by equation (6) where p_i is the cumulative empirical probability distribution given by

$$p_i = \alpha F\left(x_i \mid \mu_1, \sigma_1\right) + (1-\alpha) F\left(x_i \mid \mu_1, \sigma_2\right) + \varepsilon. \tag{7}$$

If we let $\beta = 1-\alpha$, the least-squares regression yields $\hat{\alpha} = 0.542792$ and $\hat{\beta} = 0.73584$. Using these two estimates of the mixing factor, we have $\hat{\alpha}_1 = 0.542792$ and $\hat{\alpha}_2 = 0.264155$, simply take the average of these two estimate results in $\hat{\alpha} = 0.4034735$. This estimate in a better fit of the initial data to a mixed probability distribution with $\chi^2 = 187.34$, which is very close to the Weibull probability distribution, with $\chi^2 = 186.36$.

Therefore, we further consider the first estimated mixture factor $\hat{\alpha} = 0.4034735$ in conjunction with the relationship given by

$$\hat{\mu} = \alpha\hat{\mu}_1 + (1-\alpha)\hat{\mu}_2 \tag{8}$$

to either re-estimate of the lower peak μ_1 or upper peak μ_2. If the upper peak is fixed and we use the data to re-estimate the lower peak, we have $\hat{\mu}_1 = 6.6$. This yields a worse fit with $\chi^2 = 225.09$.

However, if we fix the lower bound and estimate the upper peak, this yields $\hat{\mu}_2 = 12.3$ and $\chi^2 = 167.4764$.

By considering various values of α and continuously re-estimating, we can reduce the chi-squared statistic as shown in Table 7. Furthermore, once we have established the best sample means that give us the additional adjustment of the sample standard deviations using the relationship given by

$$\hat{\sigma}^2 = \alpha\hat{\sigma}_1^2 + (1-\alpha)\hat{\sigma}_2^2 \tag{9}$$

in order to reduce the chi-squared statistics further, also shown in Table 7.

Note that these estimates for the mixing factor α and the first standard deviation $\hat{\sigma}_1$ are only accurate to the second decimal, but this reduces the chi-squared statistic to $\chi^2 = 15.293$, which indicates a better fit using the mixed statistical model.

This is a significant improvement over both the two and three-parameter Weibull probability distribution functions with the chi-squared statistics, $\chi^2 = 30.2782$ and $\chi^2 = 31.7121$, respectively. However, as illustrated in Fig. 8, all three of these distributions are highly correlated to the empirical probability distribution.

Once we have established the underlying probability structure, we can derive the function that approximates the return period, we can use this information to generate profiles.

For example, magnitudes of bloom between 17 and 18.75 occur between 0 and 10 years. Magnitudes of bloom between 19.715 and 19.755 occur every 90 to 100 years.

Once this information is known, we can estimate the probable organism counts. For example, every 40 to 50 years blooms can reach a count between $260,991,918$ and $284,146,355$.

Recursion Analysis: Logarithmic Logistic Growth — Consider the year 1957 for two reasons: first, there are 116 hourly samples taken over a twenty-nine day period and second, all at the same location near a bridge (Gulf Blvd.) that separates the Gulf of Mexico and Boca Ciega Bay, see Fig. 9. Hence, the remainder of this study will concentrate on this one period of time during which data was collected on a consistent temporal and spatial scale.

In all the years of data collecting, the most sampled year is 1957 in which 4138 samples were taken; however, 52.3% (2165) of these were taken in only four months Fig. 10. Moreover, approximately 11% (233) of these samplings where taken at the same location, denoted in blue in Fig. 11, are gathered on a consistent temporal scale. Hence, we will restrict the following analysis to this time and place.

Few additional samples were taken toward the beginning of the year, but the count was zero; however, once the outbreak was in full bloom, many samples were taken. Therefore, unfortunately, we do not have the samples from this site until October. As illustrated in Figs. 11 and 12, once this bloom is present the magnitude increases quite quickly. However, it is interesting to note that for the data collected on a consistent basis at a single site (Fig. 13), there is an oscillation in the mean daily magnitudes, which might be explained by a logistic growth pattern.

The logistical model defined by a growth constant r, the proportion of space taken (assuming the maximum capacity is $C \geq \max_i x_i$ and C relatively large) by

$$P_n = rP_{n-1}\left(C - P_{n-1}\right).$$

Alternatively, if we define the present proportion as $p_n = \frac{\bar{x}_n}{C}$ where $\bar{x}_n$ is the daily sample mean magnitude for the n^{th} day, then this logistical model becomes $p_n =$

$rp_{n-1}\left(1-p_{n-1}\right)$ and given as a time series yields

$$p(t) = p_0 + rp(t-1)\left[1-p\left(t-1\right)\right] + \varepsilon, \text{ where } p\left(t_i\right) = p_i. \tag{10}$$

The theoretical model we propose, where estimate is given below by equation (10) with $C = 50$ explains 68.72% of the variation in the percent is shown Fig. 13. This is equivalent to saying that the maximum capacity in a liter is a bloom of magnitude 50; or 5.18×10^{21} count is plausible.

The developed statistical model is given by

$$\hat{p}(t) = 1.4077p(t-1)\left[1-p\left(t-1\right)\right] - 0.0227. \tag{11}$$

Finally, the proposed model given by equation (11) yields good estimates of the growth of a bloom as a function of the time; this model should be used to obtain useful information on the subject matter.

Regional Analysis and Delays — There are hundreds of various latitude and longitude locations recorded, these points can be used to generate a contour plot of the $\ln x$ over the various locations given in Fig. 10, but there are two distinct regions where the counts have been in excess of a half million count of *Karenia Brevis*. Consider the magnitude of the bloom as defined by the greatest integer function of the magnitude of a bloom; that is, the least integer below the value, $m = \text{int}[\ln(x)]$, shown in the contour plot in Fig. 14.

There is a slight delay, but not greatly resolved on a daily bases since samplings are not based on a uniform temporal scale. The highest magnitude of bloom is in region 8. With a mean magnitude of 10.872, this is approximately $23,000$ more organism count than the second highest region, the Tampa Bay. This might have to do with the fact that there are significantly fewer samples taken at in the first region; moreover, is the fact that these samples may have been taken when Red Tide was in full bloom and very few samples were taken.

Contributing Variables to Red Tide Blooms — To determine the key contributing variables that drive Red Tide blooms, more information is needed. Possible contributing “wash-off” variables suggested by IMaRS:

1. *Ammonium hydroxide* — a name used to describe the process of mixing an ammonia and water.
2. *Nitrate* — a salt of *nitric acid*; *Nitrate ion* is a polyatomic anion.
3. *Sulfate* — a salt of *sulfuric acid*; *Sulfate ion* is a polyatomic anion.

These nutrient concentrates, C_i, are measured at seven sites in the State of Florida, however distance to the bloom and estimated “wash-off” rates are needed to determine the delay between when the constitutes are recorded on land and their direct affect on the bloom. The difficulty in determining the relationship between runoff for the State of Florida (or other regions) (Duke, Given, Tinoco, 2004) and Red Tide Blooms is both spatial and temporal. Time bias exists between measured runoff and measured Red Tide, once the temporal bias is compensated for, Geographic Information Systems can be used to deal with spatial bias.

Additional attributing variables measured by the National Buoy Data Center include standard meteorological data for atmospheric temperature, T_a, atmospheric pressure, P, dew point, T_a, gusts, g, wind speed, w, wind direction, θ, sea surface (water) temperature, T_w; others include distance between the bloom and these measured variables as well as salinity.

Once data is gathered for these presently lurking variables, we can develop a statistical model for the response variable Red Tide as follows:

$$RT = f\left(M_1, M_2, \ldots, M_{31}, C_1, C_2, C_3, T_a, T_w, T_d, P, w, \theta, d, s\right).$$

This statistical model will include interaction between the organisms $M_i \times M_j$, chemical interaction between the nutrients $C_i \times C_j$, as well as interaction between the organisms and the nutrients $C_i \times M_j$ as well as all other possible interactions and quadratic terms such as T_w^2.

When samples taken from nearby buoys with meteorological data are considered, preliminary studies show that there is correlation between several standard meteorological data and the magnitude of bloom of *Karenia Brevis*. Table 9 gives the correlation coefficients and ranks these variables by explanatory power with wind speed ranked first and atmospheric temperature ranked second.

2.1 Results

The developed statistical model can be used to estimate the magnitude of a Red Tide bloom. This is important in monitoring the magnitude of a bloom as a function of time. The estimate of the number of organisms as a function of time can be used for public safety and advisories. Also, the proposed model and analysis can be easily updated once more data become available.

Furthermore, having a better understanding of regional differences enables us to rank the regions and determine where research efforts should be concentrated. In addition, understanding the time delays between these regions establish preliminary functions which can be built upon.

2.2 Discussion

The main contributor to Red Tide is Karenia Brevis. Even with the limited data available we can determine that the probability distribution of probable bloom magnitude is best characterization of the subject response (the magnitude of the bloom) is the Weibull probability distribution. There is no significant difference between the two and three-parameter Weibull probability distribution function. Using this information we can estimate that blooms can reach a high of 139 million organisms per sample every ten years, but up 2.7 times that every hundred years. According to recursive analysis, the relative logistic growth rate is estimated at 1.4. These results are based on the assumption that the maximum magnitude within a sample (capacity) is 50.

There are regional differences mainly northern and southern differences. This may be due to the proximity of the rivers and streams to the open water, or the ebb and flow effect the open oceans have on the Gulf and shorelines surrounding Florida. There is a correlation between the nutrients released into the soil and surface waters with some delay effects, but without more data on a more refined time scale, these exact correlations and delay effects cannot be accurately modeled. For a more detailed analysis, a data bank of periodic data (preferably hourly) of the original response variable (organism count) and the various contributing entities such as salinity at several fixed locations needs to be established.

2.3 Acknowledgment

We wish to say a special thank you to Dr. Dean Martin, Distinguished University Professor at the University of South Florida, for his helpful suggestions.

References

[1] L. K. Dixon, "Red Tide Bloom Dynamics with respect to Rainfall and Riverine Flow," Mote Marine Laboratory Technical Report Number 795, 2003.

[2] L. D. Duke, S. E. Given, and M. Tinoco, Correlation of Storm Characteristics with Constituent Concentrations in Urban Storm Water Discharge, 2004.

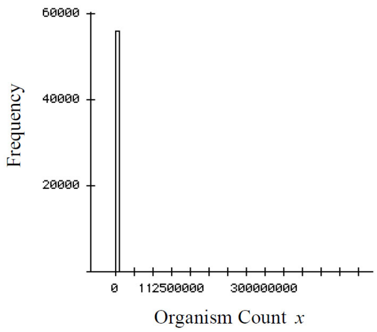
[3] G. A. Rounsefell and W. R. Nelson, Red-tide research summarized to 1964, including an annotated bibliography, U.S. Fish and Wildlife Service, Spec. Scientific Report No. 535, 1966.

[4] K. A. Steidinger, Phytoplankton ecology: a conceptual review based on eastern Gulf of Mexico research, *CRC Critical Review in Microbiology* 3 (1973), 49–68.

[5] K. A. Steidinger, Implications of Dinoflagellates life cycles on initiation of Gymnodinium breve red tides, *Environmental Letters* 9 (1975), 129–139.

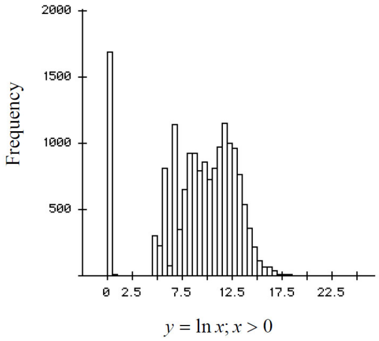
[6] K. A. Steidinger, A re-evaluation of toxic flagellate biology and ecology, pp. 147–188, in "Progress in Phycological Research" (ed. F. E. Round), Vol. 2, Elsevier, New York, 1983.

[7] H. Qiao and C. P. Tsokos, Best Efficient Estimates of the Intensity Function of the Weibull Process, *Journal of Applied Statistics* (1) 25 (1998), 111–120.



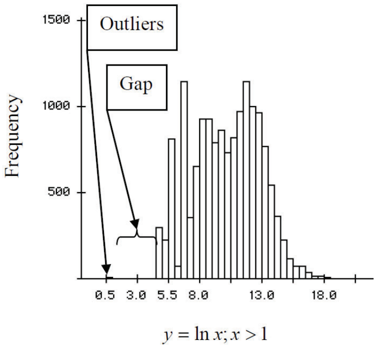
Statistic	Estimate
Count	56272
Mean	155900
Median	0
Std Dev	2594849
Variance	6733242184185
Range	358000000
Min	0
Max	358000000
IQR	333
25th%	0
75th%	333

Figure 1: Histogram of count of *Karenia Brevis* sampled over time.



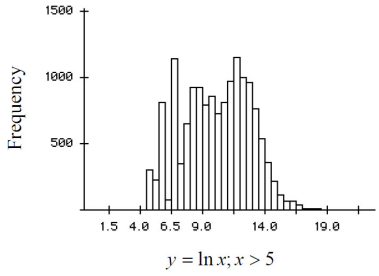
Statistic	Estimate
Count	16731
Mean	9.097
Median	9.798
Std Dev	4.006
Variance	16.046
Range	19.696
Min	0
Max	19.696
IQR	5.124
25th%	6.908
75th%	12.032

Figure 2: Histogram of the natural logarithm of the count of *Karenia Brevis* sampled over time, given the count was at least one.



Statistic	Estimate
Count	15042
Mean	10.118
Median	10.309
Std Dev	2.741
Variance	7.512
Range	19.003
Min	0.693
Max	19.696
IQR	4.215
25th%	8.006
75th%	12.221

Figure 3: Histogram of the natural logarithm of the count of *Karenia Brevis* sampled over time, given the count was at least two.



Statistic	Estimate
Count	15022
Mean	10.131
Median	10.309
Std Dev	2.722
Variance	7.409
Range	17.904
Min	1.792
Max	19.696
IQR	4.22
25th%	8.006
75th%	12.226

Figure 4: Histogram of the natural logarithm of the count of *Karenia Brevis* sampled over time, given the count of five or more.

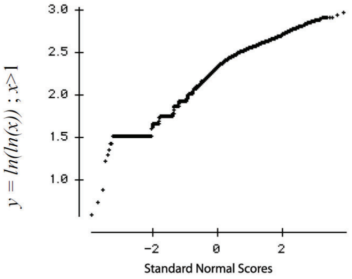


Figure 5: Probability plot of the double natural logarithm of the count of *Karenia Brevis* sampled over time, given the count was greater than five.

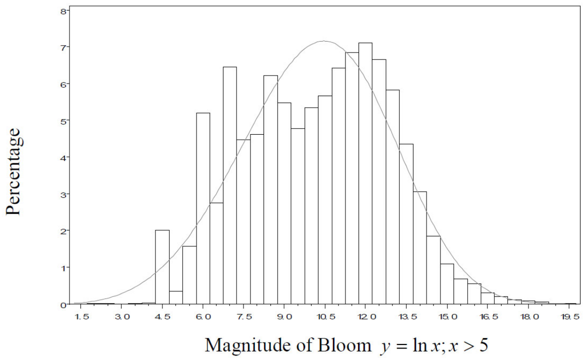


Figure 6: Histogram with best-fit distribution for the natural logarithm of the count of *Karenia Brevis* sampled over time, given the count was greater than five.

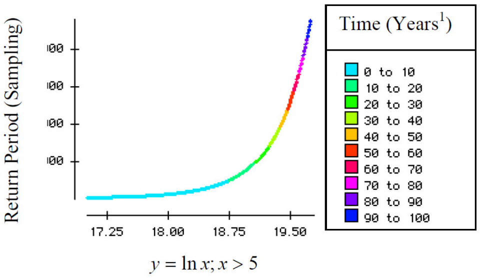


Figure 7: Return periods of the natural logarithm of the count of *Karenia Brevis*.

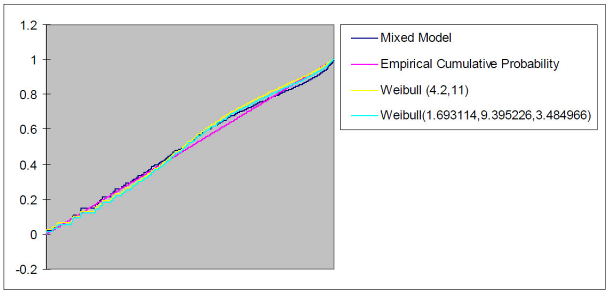


Figure 8: Comparison of the best-fit distributions and the empirical probability distribution.

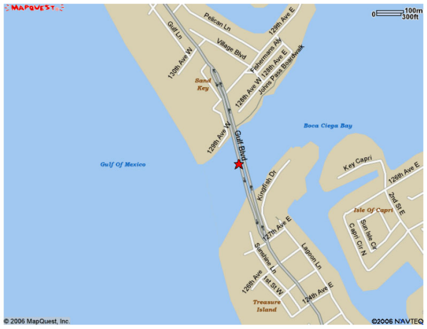


Figure 9: Map of location where data measured hourly.

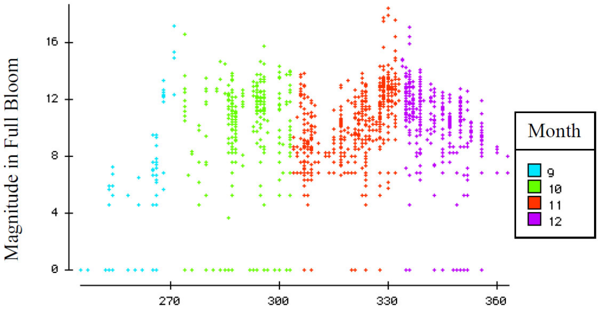


Figure 10: Line graph of magnitude of bloom by month.

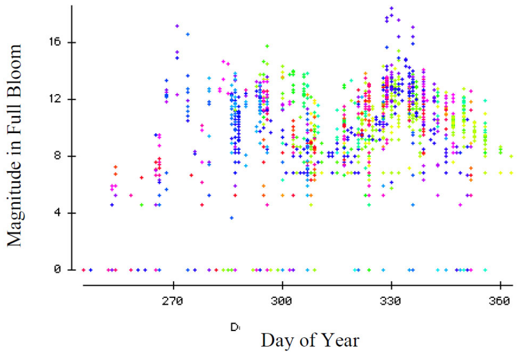


Figure 11: Line graph of magnitude of bloom by location (latitude, longitude).

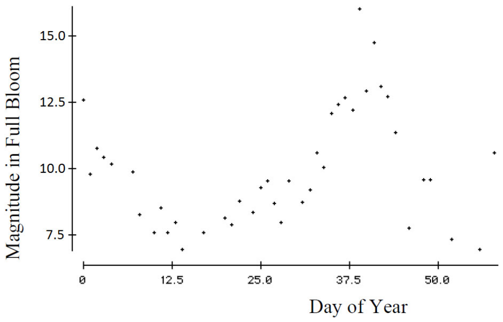


Figure 12: Line graph of magnitude of bloom at a single location at time t_i.

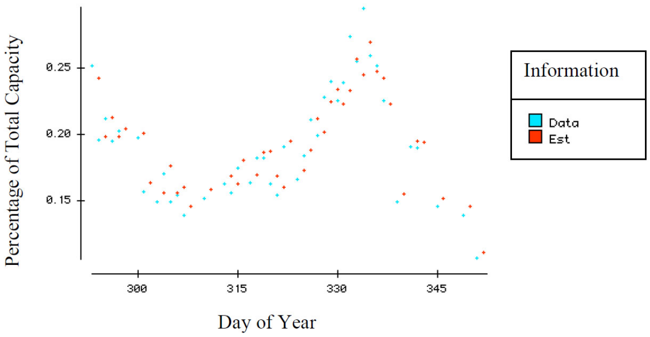


Figure 13: Line graph of percentages based on daily mean of the data (blue) and the estimated percentage based on the previous data.

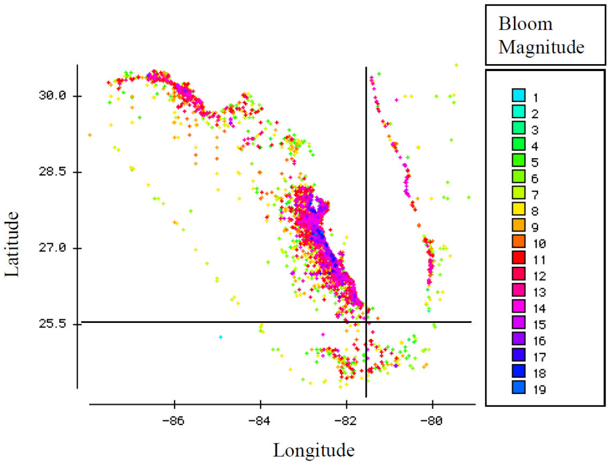


Figure 14: The contour plot of the natural logarithm of the count of *Karenia Brevis* with respect to the sampling location (longitude, latitude).

Table 1: Data compiled by number of times recorded in samplings. Includes total count over time and mean count per sample as well as the organism and how it these organisms are coded.

Code	Total Count	Samples	Mean Count per Sample	Organism
KARE	8772810000	56272	155900.0924	*Karenia* (1)
DIAT	152712300	2557	59723.23035	*Diatom* (2)
OTHE	1043334	2424	430.4183168	*Other Plankton* (3)
GYMN	108634500	2088	52028.01724	*Gymnodinium* (4)
DINO	42049600	1883	22331.17366	*Dinoflagellates* (5)
MICR	3287360000	1865	1762659.517	*Micro-flagellates* (6)
GYRO	8020570	1791	4478.26354	*Gyrodinium* (7)
CILI	22979410	1449	15858.80607	*Ciliates* (8)
GONY	47679000	1445	32995.84775	*Gonyaulax* (9)
PERI	145862172	1081	134932.629	*Peridinium* (10)
CERA	7063375	857	8241.97783	*Ceratium*
PROR	5692300	755	7539.470199	*Prorocentrum*
NAUP	125476.9	439	285.8243736	*Nauplii*
OSCI	12907406	348	37090.24713	*Oscillatoria*
TRIC	688423720	347	1983930.029	*Trichodesmium*
FLAG	267044	296	902.1756757	*Flagellates*
BLUE	75459800	268	281566.4179	*Blue Green Algae*
COPE	42321	201	210.5522388	*Copepods*
POLY	268663	190	1414.015789	*Polykrikos*
COCH	7443913	134	55551.58955	*Cochlodinium*
RHIZ	95.992	114	0.842035088	*Rhizosolenia*

Table 2: Percent by category

Group	Count	%
B	15042	26.731
N	39541	70.268
P	1689	3.001

Table 3: Percent by category (redefined)

Group	Count	%
B	15022	26.695
N	39541	70.268
P	1709	3.037

Table 4: Statistics based on the two and three-parameter Weibull.

Statistic	Estimate 2 Parameter	Estimate 3 Parameter
Number of Data	15022	15022
Mean	10.13	10.14
Standard Deviation	2.72	2.68
Variance	7.40	7.18
Skewness	-0.0359	-0.0359
Kurtosis	-0.6884	-0.6884

Table 5: Goodness-of-Fit Test for Weibull.

Test	Statistics		*p*-value	
Kolmogorov–Smirnov	D	0.0602309	Pr>D	< 0.001
Cramer–von Miser	W-sq	12.5414661	Pr>W-sq	< 0.001
Anderson–Darling	A-sq	71.3164327	Pr>A-sq	< 0.001

Table 6: Estimations for various return periods.

Return Period (Years^2)	$\ln x$		x	
	Min	Max	Min	Max
0 to 10	17	18.75	24,154,952	139,002,155
10 to 20	18.755	19.07	139,698,906	191,423,727
20 to 30	19.075	19.25	192,383,243	229,175,810
30 to 40	19.255	19.375	230,324,559	259,690,215
40 to 50	19.38	19.465	260,991,918	284,146,355
50 to 60	19.47	19.545	285,570,645	307,812,072
60 to 70	19.55	19.61	309,354,986	328,484,430
70 to 80	19.615	19.66	330,130,965	345,326,187
80 to 90	19.665	19.71	347,057,142	363,031,439
90 to 100	19.715	19.755	364,851,142	379,741,000

Table 7: Estimations of parameters and associated chi-squared statistic.

Trial	μ_1	μ_2	σ_1	σ_2	α	χ^2
1	6.9	12.5	2.72	2.72	0.4034735	186.359
2	6.9	12.3	2.72	2.72	0.4	167.476
3	6.9	11.5	2.72	2.72	0.3	76.650
4	6.9	10.9	2.72	2.72	0.2	34.596
5	6.9	10.7	2.72	2.72	0.15	26.500
⋮						
30	6.9	10.5	1.27	2.90	0.11	15.328
31	6.9	10.5	1.26	2.90	0.11	15.318
32	6.9	10.5	1.25	2.90	0.11	15.309
33	6.9	10.5	1.24	2.90	0.11	15.302
34	6.9	10.5	1.23	2.91	0.11	15.298
35	6.9	10.5	1.22	2.91	0.11	15.296
36	6.9	10.5	1.21	2.91	0.11	15.296
37	6.9	10.5	1.20	2.91	0.11	15.298
38	6.9	10.5	1.19	2.91	0.11	15.303
39	6.9	10.5	1.18	2.91	0.11	15.310
40	6.9	10.5	1.17	2.91	0.11	15.320

Table 8: Correlations given various maximum capacities.

C	R^2	Growth Rate r	Error ε
20	15.43	-3.3277	1.3025
30	67.28	2.6848	-0.2507
40	68.66	1.7204	-0.0679
50	68.72	1.4077	-0.0227
60	68.68	1.2543	-0.0059
70	68.64	1.1635	0.0016
80	68.61	1.1034	0.0052
90	68.58	1.0607	0.0070
100	68.56	1.0288	0.0080
150	68.49	0.9437	0.0082
200	68.45	0.9061	0.0071
1000	68.37	0.8270	0.0017
$1E+14$	68.35	0.8093	0.0000

Table 9: Correlation between Red Tide Bloom and Attributing Variables.

Rank	Variable	Correlation Coefficient
1	w	-0.41
2	T_a	-0.40
3	T_w	-0.29
4	T_d	-0.27
5	θ	-0.19
6	P	-0.19
7	g	-0.07