

# Accounting for Model Uncertainty in Prediction of Chlorophyll *a* in Lake Okeechobee

E. Conrad LAMON III and Merlise A. CLYDE <sup>1</sup>

Long term eutrophication data along with water quality measurements (total phosphorous and total nitrogen) and other physical environmental factors such as lake level (stage), water temperature, wind speed and direction were used to develop a model to predict chlorophyll *a* concentrations in Lake Okeechobee. The semi-parametric model included each of the potential explanatory variables as either linear predictors, regression spline predictors or product spline interactions allowing for nonlinear relationships. A Gibbs sampler was used to traverse the model space. Predictions that incorporate uncertainty about inclusion of variables and their functional forms were obtained using Bayesian model averaging (BMA) over the sampled models. Semi-parametric regression with Bayesian model averaging and spline interactions provides a flexible framework for addressing the problems of nonlinearity and counterintuitive total phosphorus function estimates identified in previous statistical models. The use of regression splines allows nonlinear effects to be manifest, while their extension allows inclusion of interactions for which the mathematical form cannot be specified *a priori*. Prediction intervals under BMA provided better coverage for new observations than confidence intervals for ordinary least squares models obtained using backwards selection. Also, BMA was more efficient than OLS in terms of predictive mean squared error for overall lake predictions.

**Key Words:** Algal Blooms; Bayesian Model Averaging; Gibbs Sampler; Semi-parametric Regression; Variable Selection.

## 1 INTRODUCTION

Accurate models for the prediction of algal biomass (as measured by chlorophyll *a* concentrations) in Lake Okeechobee are critical for management planning to improve water quality. Specification of a model for the prediction of chlorophyll *a* requires the analyst to consider many possible environmental factors, their relationships and interactions. Uncertainty regarding the most important set of predictor variables from previous studies and the structural form of their relationship to chlorophyll *a* concentrations has highlighted the need to consider model uncertainty when developing a model to predict chlorophyll *a* concentrations in Lake Okeechobee.

Internal loading of phosphorus (such as wind resuspension of bottom sediments) has been shown to be important in other lakes (Carper and Bachmann 1984), and several researchers (McCauley et al. 1989, Prairie et al. 1989, Watson et al. 1992) have proposed a sigmoid relationship between  $\log_{10}$  total phosphorus and  $\log_{10}$  chlorophyll *a*. Phlips et al. (1993) identified five environmentally

---

<sup>1</sup>E. Conrad Lamon III is Assistant Professor, Institute for Environmental Studies, Louisiana State University, 42 Atkinson Hall, Baton Rouge, LA 70803. Merlise A. Clyde is Assistant Professor, Institute of Statistics and Decision Sciences, Duke University, 210A Old Chemistry, BOX 90251 Durham, NC 27708-0251.

distinct zones within Lake Okeechobee using multivariate procedures. Lamon (1995) found evidence of such a sigmoid relationship for some of the zones of Lake Okeechobee identified by Philips et al. (1993), and Lamon et al. (1996) present generalized additive models showing this nonlinear relationship. Seasonal and localized nutrient limitation by both phosphorus and nitrogen occurs in Lake Okeechobee (Schelske 1989, Aldridge et al. 1995). In-lake total phosphorus is not correlated with loading rate when normalized by either depth or residence time (as is done with many nutrient loading models), but is correlated with lake level (Canfield and Hoyer 1988). No significant correlation exists between annual average in-lake phosphorus concentrations and their respective phosphorus loading rates (Canfield and Hoyer 1988). Prediction of annual average total phosphorus was better by wind speed than by stage in Lake Okeechobee (Maceina and Soballe 1990), though lake stage is a slightly better predictor of chlorophyll *a* than wind when no annual averaging is done (Lamon 1995). The many environmental factors interacting to stimulate algal growth and bloom condition formation in Lake Okeechobee make variable selection and specification of an empirical model a complicated task.

Development of a model that takes into account the nonlinear relationships between environmental factors and chlorophyll *a* concentration while seeking out important interactions among these factors is the main goal of this paper. Specifically, interactions between wind velocity and lake stage are examined, since their interaction can be important in sediment resuspension and therefore light limitation (Carper and Bachmann 1984). Such a model not only improves prediction of chlorophyll *a* in Lake Okeechobee, but provides insight into mechanisms that control algal biomass as measured by chlorophyll *a*. We use regression splines and product spline bases to allow for flexible nonlinear functions and interactions in these variables. Because of the large number of possible models, we use Stochastic Search Variable Selection (SSVS) (George and McCulloch 1993, 1997, Smith and Kohn 1996) to identify important models, and use Bayesian Model Averaging (BMA) (Raftery, Madigan, and Hoeting 1997, Clyde 1999) to incorporate uncertainty about which predictor variables should be incorporated into the model. Rather than selecting a single model to make predictions, as is common practice, predictions under BMA are based on a weighted average of several models, where weights are based on the degree to which the data support each model as measured by the posterior probability of the model. Hoeting et al. (1999) provide a review of BMA and present several case studies where BMA leads to improved predictive performance.

The outline of the paper is as follows. In Section 2, we describe the data collection and variables that are used in the model. In Section 3, we describe the semi-parametric model, the SSVS algorithm used to search for high probability models, and the estimators used in model averaging. Section 4 describes the results for Lake Okeechobee, and presents graphical methods for viewing model uncertainty. We also compare the predictive performance of BMA to backward model selection, a standard approach used in practice, on an independent validation subset of the data.

## 2 MONITORING DATA

The South Florida Water Management District (SFWMD) has monitored physical and chemical conditions in Lake Okeechobee since 1973, on at least a monthly frequency. The spatial extent of sampling was restricted to eight pelagic stations until 1986, when 29 near-littoral stations were added and subsequently sampled on at least a monthly frequency. A multi-agency study initiated in 1988, hereafter the Lake Okeechobee Ecosystem Study (LOES), included monthly or semi-monthly sampling at over 100 locations in the pelagic and littoral regions. Data from both sources are considered here (Figure 1), and no averaging over space or time is done prior to model development. Samples collected by the SFWMD were near surface, and in situ measurements of temperature (TEMP) were made using a Hydrolab Surveyor water analyzer. Analytical methods used for the samples collected by the SFWMD are described in James et al. (1995) and follow the SFWMD QA/QC plan (SFWMD 1993). Methods used to analyze the LOES samples were described in Philips et al. (1993a). Total phosphorus (TP) and total nitrogen (TN) were determined by methods similar to those used by the SFWMD. Chlorophyll *a* was determined using the protocol described in Standard Methods (SM 1002 G) (A.P.H.A. 1985). Due to changes in the method used by SFWMD for chlorophyll *a* determination, only data collected since 1980 were used in this study.

*(Figure 1 about here)*

Sediment types of sand, peat, rock, mud, and littoral in Lake Okeechobee were determined by Reddy et al. (1991). A sampling grid was established on the lake which determined 171 sampling locations 3.2 km apart. Principal sediment zones were determined based on physical descriptions of surface deposits. Due to the paucity of data and sampling stations located in the rock sediment type, these data were combined with that of the peat sediment type. Use of sediment based zones is done because of uncertainty associated with the ecological zones of Philips et al. (1993) which have been used in previous modeling studies of Lake Okeechobee (Lamon 1995, Lamon et al. 1996). Sediment type is an “on the ground” classifier of ecological region derived from in lake measurements and is not subject to the uncertainty associated with model choice as in the discriminant functions used by Philips et al. (1993). Other variables used in the analysis include lake stage, wind speed and wind direction. Lake level (STAGE) is obtained from three stations by the US Army Corps of Engineers and is a daily average over the three stations for the sampling day. Hourly wind measurements (km/hr day) were taken at Moorehaven, Florida. The variable WIND is the summation of the hourly wind data for every 24 hour period beginning at 12:01 am and is a measure of wind energy for the sampling day. Wind direction (DIR) is a daily average in degrees.

Log transformations of variables are commonly used in water quality studies to linearize relationships and stabilize variance. Water quality measurements are bounded at zero on the left and usually contain large tails on the right. Previous inference regarding the relationship between chlorophyll *a* and TP have by and large been done on a log-log scale, and have tried to characterize the log-log relationship as linear. Others have characterized the relationship as being sigmoidal.

By using semi-parametric models we should be able to capture either situation, and will use log transformations to allow us to test if the linear log-log relationship holds. Log transformations were applied to all variables except wind direction.

### 3 MODEL DEVELOPMENT

We adopt the following model for the chlorophyll  $a$  concentrations:

$$Y = f(x_1, \dots, x_p) + \epsilon \quad (1)$$

where  $Y$  is the  $n \times 1$  vector of  $\log_{10}$  chlorophyll  $a$  concentrations and  $x_1, \dots, x_p$  represent the possible predictor variables: logTP, logTN, logTEMP, logSTAGE, logWIND, and DIR. The function  $f$  is assumed to be a smooth function of these predictor variables, and  $\epsilon$  is a  $n \times 1$  vector of independent normal errors with mean 0 and variance  $\sigma^2$ . In general, the curse of dimensionality prevents us from fitting non-parametric models with high order interactions. The model developed in this paper is an extension of the work of Smith and Kohn (1996) inspired by the MARS (multivariate adaptive regression splines) models of Friedman (1991) and allows flexible two-way interactions among the variables. Smith and Kohn presented a framework for semi-parametric multiple regression using additive splines in a Bayesian hierarchical model for variable selection (George and McCulloch 1993, 1997). Friedman developed product spline basis functions for use in multiple regression with interactions among predictors.

In the following model for  $f$ , we consider three ways in which a variable can be used for prediction of the response. First, a variable  $x_j$  can be used as a linear predictor. This corresponds to the usual log-log models. Second, a variable  $x_j$  may enter nonlinearly if the log-log relationship does not hold. This nonlinear function can be approximated by a piece-wise polynomial using an additive regression spline predictor,  $s_j(x_j)$ , as in generalized additive models (Hastie and Tibshirani 1990, Smith and Kohn 1996). In this case,  $s_j()$  is approximated by a piece-wise linear regression spline,

$$s_j(x_j) = \sum_{k=1}^{K_j} \beta_k (x_j - \tilde{x}_{jk})_+$$

where  $(z)_+ = \max(0, z)$ , which is a one sided power spline basis function. The  $\tilde{x}_{jk}$  represent  $K_j$  knots placed along the domain of the variable  $x_j$  such that

$$\min(x_j) < \tilde{x}_{j1} < \dots < \tilde{x}_{jK_j} < \max(x_j).$$

The number and placement of the knots determines the amount of smoothing, with many, closely placed knots yielding functions with high local variance. In order to avoid over-fitting the data, it is usually necessary to remove some knots. Third, smooth interactions between variables  $x_j$  and  $x_l$  are represented by the use of products of spline basis elements (Friedman 1991),

$$s_{jl}(x_j, x_l) = \sum_{k=1}^{K_j} \sum_{m=1}^{K_l} \beta_{km} (x_j - \tilde{x}_{jk})_+ (x_l - \tilde{x}_{lm})_+$$

where the knots are defined as before. The sum is over all  $K_j \times K_l$  knots for the pairs of variables  $(x_j, x_l)$ , where  $K_j$  is the number of knots placed along the domain of the first variable  $x_j$  and  $K_l$  is the number of knots for the second variable  $x_l$ .

Using these terms, the function  $f$  is expressed as a semi-parametric model

$$f(x_1, \dots, x_p) = \beta_0 + \sum_{j=1}^p (\beta_j x_j) + \sum_{j=1}^p \left( s_j(x_j) + \sum_{l < j} s_{jl}(x_j, x_l) \right)$$

and, while this is not linear in terms of the original measurements  $x_j$ , it is linear in terms of the unknown regression parameters and the basis functions,  $x_j$ ,  $(x_j - \tilde{x}_{jk})_+$  and  $(x_j - \tilde{x}_{jk})_+(x_l - \tilde{x}_{lm})_+$  for a fixed set of knot locations. We can construct a  $n \times r$  design matrix  $X$  which consists of all linear terms for each of the regressors and all basis elements for the smooth and interaction terms, and rewrite the model in (1) as a linear model,

$$Y = X\beta + \epsilon \tag{2}$$

where  $\beta = (\beta_1, \dots, \beta_r)'$  is a  $r \times 1$  vector of all regression coefficients in the linear predictor components, plus the coefficients in the smoothing spline components  $s_j(\cdot)$  and  $s_{jl}(\cdot)$ .

### 3.1 HIERARCHICAL MODEL FOR VARIABLE SELECTION

Using a large number of variables and knots can lead to over-fitting, and thus some method of variable selection is typically used to eliminate variables that are not important. This corresponds to the *a priori* belief that some of the regression coefficients  $\beta_j$  are 0, which is equivalent to dropping the corresponding columns from the design matrix  $X$ . We can represent this belief by building a hierarchical model that allows for uncertainty about which of the  $\beta_j$  coefficients are 0. The vector  $\gamma$  is a  $r \times 1$  vector of indicator variables, which represent different models where  $\beta_j = 0$  if  $\gamma_j = 0$  and  $\beta_j \neq 0$  if  $\gamma_j = 1$ . Given  $\gamma$ , let  $\beta_\gamma$  consist of all the nonzero elements of  $\beta$  and let  $X_\gamma$  be the columns of  $X$  corresponding to those elements of  $\gamma$  that are equal to one.

In building the hierarchical model we make the following prior assumptions:

1. Given  $\gamma$  and  $\sigma^2$ ,

$$\beta_\gamma | \gamma, \sigma^2 \sim N(\mathbf{0}, c\sigma^2(X'_\gamma X_\gamma)^{-1})$$

where  $c$  is a positive scale factor. Choosing a large value of  $c$  means our prior contains very little information about  $\beta_\gamma$  compared to the likelihood; we use  $c = 1000$ , the choice recommended by Smith and Kohn (1996). This choice for  $c$  is on the order of the average sample size over all ecological zones ( $\bar{n} = 917$ ), and is comparable to the recommended default value for the unit-information prior (Kass and Raftery 1995) which gives an approximation to the Bayes Information Criterion (BIC). BIC-like priors penalize adding variables more than AIC (which often leads to over-fitting), and often lead to better predictive performance for large sample

sizes than with AIC (Hoeting et al. 1999). Other choices for  $c$  can be based on calibrating  $c$  with other classical model selection methods (George and Foster 1997).

2. The prior of  $\sigma^2$  given  $\gamma$  is  $p(\sigma^2|\gamma) \propto 1/\sigma^2$ . Under this default improper prior distribution,  $\log(\sigma^2)$  has a uniform distribution and as long as there are at least  $r$  observations, leads to a proper joint posterior distribution for  $\beta, \gamma$ , and  $\sigma^2$ .
3. The indicator variables  $\gamma_j$  are assumed to be *a priori* independent with

$$p(\gamma_j = 1) = \pi_j \quad 0 \leq \pi_j \leq 1$$

for  $j = 1, \dots, r$ . We take  $\pi_j = 0.5$ , which places a uniform prior distribution over models.

### 3.2 POSTERIOR DISTRIBUTIONS

Because of the conjugate structure in the prior distributions, the posterior distributions for  $\beta$  and  $\sigma^2$  conditional on a model  $\gamma$  are known in closed form (including normalizing constants).

The posterior distribution for  $\beta_\gamma$  given  $\gamma$  and  $\sigma^2$  is normal

$$\beta_\gamma | \sigma^2, \gamma, Y \sim N \left( \hat{\beta}_\gamma, \frac{c}{1+c} \sigma^2 (X'X)^{-1} \right) \quad (3)$$

with mean

$$\hat{\beta}_\gamma = \frac{c}{1+c} (X'_\gamma X_\gamma)^{-1} X'_\gamma Y$$

which corresponds to shrinking the usual ordinary least squares estimates under model  $\gamma$  towards 0 with shrinkage factor  $c/(1+c)$ .

The posterior distribution for  $\sigma^2$  given  $\gamma$  is an inverse Gamma distribution, or in other words  $1/\sigma^2$  has a Gamma distribution,

$$\sigma^{-2} | \gamma, Y \sim \text{Gamma} \left( \frac{n}{2}, \frac{S(\gamma)}{2} \right) \quad (4)$$

where

$$S(\gamma) = Y'Y - \frac{c}{1+c} Y'X_\gamma (X'_\gamma X_\gamma)^{-1} X'_\gamma Y. \quad (5)$$

denotes the residual sum of squares for the Bayesian linear model.

The posterior distribution of  $\gamma$  is known up to the normalizing constant,

$$\begin{aligned} p(\gamma|y) &\propto p(y|\gamma)p(\gamma) \\ &\propto (1+c)^{-q_\gamma/2} S(\gamma)^{-n/2} \prod_{i=1}^r \pi_i^{\gamma_i} (1-\pi_i)^{1-\gamma_i} \end{aligned}$$

where  $q_\gamma = \sum_{j=1}^r \gamma_j$  is the number of nonzero elements of  $\beta$ . Since the vector  $\gamma$  completely specifies a model (i.e., which variables are included), the posterior distribution of  $\gamma$  given  $Y$ ,  $p(\gamma|Y)$ , determines

posterior model probabilities. Under the uniform prior distribution over models, the posterior model probabilities simplify to

$$p(\gamma|y) = \frac{(1+c)^{-q_\gamma/2} S(\gamma)^{-n/2}}{\sum_{\gamma' \in \Gamma} (1+c)^{-q_{\gamma'}/2} S(\gamma')^{-n/2}} \quad (6)$$

where the normalizing constant is obtained by summing over all possible models  $\Gamma$ .

### 3.3 BAYESIAN MODEL AVERAGING

For making predictive inferences at a new point  $x^*$ , the predictive mean at  $x^*$  given the current data  $Y$  is obtained by first finding the predictive means for each model  $\gamma$ ,  $\hat{Y}_\gamma^* = x_\gamma^* \hat{\beta}_\gamma$ , where  $x_\gamma^*$  corresponds to the columns of  $x^*$  where the elements of  $\gamma$  are 1. These are then combined to obtain the predictive mean by computing the weighted average of the predictions,

$$\hat{Y}^* = \sum_{\gamma \in \Gamma} x_\gamma^* \hat{\beta}_\gamma p(\gamma|Y) \quad (7)$$

where the weights are the posterior model probabilities and the summation is over the set of all possible models  $\Gamma$ . Other quantities such as the predictive cumulative distribution function

$$P(Y^* \leq y|Y) = \sum_{\gamma \in \Gamma} P(Y^* \leq y|\gamma, Y) p(\gamma|Y) \quad (8)$$

can be obtained similarly via model averaging. The predictive density conditional on a model  $\gamma$  is

$$p(Y^*|\gamma, Y) = \int p(Y^*|\beta, \sigma^2, \gamma) p(\beta|\sigma^2, \gamma, Y) p(\sigma^2|\gamma, Y) d\beta d\sigma^2$$

which is a Student- $t$  distribution, so that the scaled quantity

$$\frac{Y^* - x_\gamma^* \tilde{\beta}_\gamma}{\sqrt{\frac{S(\gamma)}{n} (1 - x_\gamma^* (x_\gamma^{*'} x_\gamma^* + \frac{1+c}{c} X_\gamma' X_\gamma)^{-1} x_\gamma^{*'})^{-1}}}$$

has a standard Student- $t$  distribution with  $n$  degrees of freedom. The unconditional predictive distribution is a mixture of Student- $t$  distributions.

### 3.4 IMPLEMENTING MODEL AVERAGING

When the number of columns  $r$  in the design matrix is large, we cannot enumerate all possible models to calculate the posterior model probabilities or posterior expectations of quantities of interest. We can approximate BMA by estimating the posterior model probabilities by using a sample of models. Markov Chain Monte Carlo (MCMC) methods are one popular way of stochastically searching the model space to identify models that are used in model averaging (Clyde 1999) and lead to simulation consistent estimates of posterior means. In this problem, we will use the Stochastic Search Variable Selection algorithm of George and McCulloch (1997), which is a Gibbs sampler

(Gelfand and Smith 1990), to sample models according to their posterior probabilities. Because the posterior distributions for  $\beta$  and  $\sigma^2$ , and future observations are completely specified conditional on a model, we do not need to use a Gibbs sampler to sample from their posterior distributions as the conditional posterior expectations of interest exist in closed form.

The SSVS algorithm proceeds in the following manner:

- (i) Choose an initial value of  $\gamma$ ,  $\gamma^{[0]} = (\gamma_1^{[0]}, \dots, \gamma_r^{[0]})$
- (ii) At iteration  $k$ : Initialize  $\gamma^{[k]} = \gamma^{[k-1]}$ . For  $j = 1, \dots, r$  successively generate  $\gamma_j^{[k]}$  from a Bernoulli distribution with  $Pr(\gamma_j = 1|Y, \gamma_{(j)}^{[k]})$ , where  $\gamma_{(j)}^{[k]}$  represents the current state of indicators without the  $j$ th element. The probability that  $\gamma_j = 1$  is based on the full conditional distribution:

$$\begin{aligned} Pr(\gamma_j = 1|y, \gamma_{(j)}) &= (1/1 + h) \\ h &= \frac{1 - \pi_j}{\pi_j} (c + 1)^{\frac{1}{2}} \left( \frac{S(\gamma_j^1)}{S(\gamma_j^0)} \right)^{\frac{n}{2}}, \end{aligned}$$

where  $S(\gamma_j^1)$  corresponds to the Bayesian residual sum of squares  $S(\gamma)$  from (5) evaluated with  $\gamma_j = 1$  and  $S(\gamma_j^0)$  corresponds to evaluating  $S(\gamma)$  with  $\gamma_j = 0$ .

- (iii) Repeat step (ii) for an adequate period of time for “burn-in”, and then repeat  $K$  times to generate  $K$  models from the posterior distribution  $p(\gamma|Y)$ .

Each iteration  $k$  corresponds to a “visit” to a potentially new model since each iteration results in a new vector  $\gamma$ , and therefore in a new set of predictors for the model (specification). We ran the Gibbs sampler for a burn in period of 10,000 iterations. Because of the high dimensional storage requirements for posterior inferences for our validation study (storing posterior means, models, and computing the predictive distributions for each point in the validation data set requires large amount of disk space and memory), we used the first 10,000 iterations after burn in for subsequent posterior inference. Standard Gibbs sampler diagnostics did not indicate problems with convergence based on the output  $\gamma$  or quantities of interest such as predictive means under model averaging. To check convergence of predictive means, we ran the Gibbs sampler for an additional 1,000,000 iterations, and obtained virtually identical predictions under the testing and validation data.

### 3.5 ESTIMATING POSTERIOR MODEL PROBABILITIES

There are two main approaches used in the literature for estimating posterior model probabilities used in BMA. Let  $S$  denote the set of unique models that appeared in the Gibbs sampler output. The relative frequency (RF) estimator of the posterior model probabilities is

$$p(\widehat{\gamma|Y})^{RF} = f_\gamma / K \tag{9}$$

where  $f_\gamma$  is the number of times that model  $\gamma$  appeared among the  $K$  models from the Gibbs sampler. This approach has been used by George and McCulloch (1993), Clyde et al. (1996), Smith and Kohn (1996) (among others) and is based on the standard ergodic average of  $\gamma$  from the Markov chain output. In large problems, this may result in many of the probabilities being estimated as 0 (if the model did not appear in the sample) or  $1/K$ , because the model space is very large compared to the typical number of iterations  $K$  and the chain does not return to the model. While this is an unbiased estimator in repeated MCMC sampling, convergence to the posterior distribution of  $\gamma$  may be very slow in large model spaces. This can be a serious problem for model selection based on the highest probability model, but not so much for model averaging (Clyde 1999). However, in the conjugate framework, the posterior model probabilities are known up to the normalizing constant. Using this additional information, we can alternatively estimate the posterior model probabilities by renormalizing over the models in the sample (Clyde et al. 1996, Clyde 1999, Raftery et al. 1997). A simulation consistent estimate of the posterior model probabilities obtained by renormalizing the posterior model probabilities (RPP) within the set of sampled models is

$$p(\widehat{\gamma|Y})^{RPP} = \frac{(1+c)^{-q_\gamma/2} S(\gamma)^{-n/2} p(\gamma)}{\sum_{\gamma' \in S} (1+c)^{-q_{\gamma'}/2} S(\gamma')^{-n/2} p(\gamma')} \quad (10)$$

for models  $\gamma$  in  $S$  and is 0 otherwise. For simulation consistent estimates of (7) and (8) we replace  $p(\gamma|Y)$  by the estimates in (9) or (10) and carry out the summation over  $S$  instead of  $\Gamma$ . While simulation studies indicate that (10) converges to  $p(\gamma|Y)$  faster in moderate sample spaces (it is exact if all models are sampled, unlike the RF estimator), it is not clear which method is better in large model spaces. We will compare the predictive performance of the two estimators in a validation data set in section 4.

## 4 RESULTS

We used the SSVS algorithm on a model space that included separate spline terms for the environmental factors  $\log_{10}$  total phosphorus (TP),  $\log_{10}$  total nitrogen (TN),  $\log_{10}$  lake level (STAGE),  $\log_{10}$  temperature (TEMP),  $\log_{10}$  wind direction (WIND) and wind direction (DIR) for each sediment type. Before separation into sediment type groups, each predictor was scaled to lie between 0 and 1 to determine the knots, and so that the same potential knots were used in each sediment type. (Table 1). The number of columns of the design matrix for each sediment type is 167, which consists of an intercept, the six linear predictors, ten knots/predictor for the additive components, and 100 knots for the two way interactions between wind and stage.

*(Table 1 about here)*

### 4.1 MODEL CHOICE

Perhaps the best way to understand the Bayesian mixture model resulting from the Gibbs sampling procedure described above is through use of the vector  $\gamma$ . As stated before,  $\gamma_i$  is a binary 0-1 variable

with a value of one if variable  $i$  is included in model  $\gamma$ , and is 0 otherwise. We can view model uncertainty by constructing an image of the matrix of all the unique models  $\gamma$  in  $S$ . In such a matrix, the rows correspond to predictor variables and the columns to sampled models. If a position (row, column) in the matrix is filled, its corresponding  $\gamma_i$  is equal to one, i.e., the predictor variable in that row is included for the corresponding model in that column. For a given predictor variable, the frequency of filled positions in its row is an indication of the importance of that variable in the Bayesian mixture model. The images in Figure 2 (a-d) have been constructed in this manner using columns which correspond to models with the highest posterior probabilities, and are presented grouped according to sediment type. The overall importance of a variable can be determined by combining rows involving the predictor variable. The probability of no nonlinear effect for a given variable is the probability that all of the regression coefficients associated with the spline terms for that variable are zero, or equivalently that all of the corresponding  $\gamma_i$  terms associated with the knots for that variable are all zero. The probability that a variable has no effect on chlorophyll  $a$  production is the probability that all of the  $\gamma_i$ 's for the linear and spline terms are all zero. Table 2 summarizes the posterior probabilities of no effect and probability of no nonlinear effect for each of the variables broken down by ecological zones. The numerical summaries indicate the overall importance of a variable (marginally), while the model-space matrices illustrate uncertainty in knot location, and reflect the joint distribution of  $\gamma$ .

*(Figure 2 about here)*

*(Table 2 about here)*

Figure 2 and Table 2 indicate that  $\log_{10}$  TP is an important predictor of  $\log_{10}$  chlorophyll  $a$ , as the linear term and/or knots occur in most of the models visited by the Gibbs sampler; probability of an effect due to TP is greater than 0.999 in all zones. Also, the probability of no nonlinear effect is very small (probability less than 0.001), indicating that the log-log relationship alone is not adequate for explaining the relationship between  $\log_{10}$ TP and chlorophyll  $a$ .

STAGE and WIND both have high posterior probabilities of inclusion ( $> 0.999$ ). STAGE and WIND enter almost all models through their interaction terms, with the posterior probability of no interaction ranging from 0.02 for the sand zone, up to 0.001 for the littoral zone; probabilities are 0.006 and 0.001 for the sand and peat zones respectively. Lamon (1995) removed  $\log_{10}$  wind speed as a predictor of  $\log_{10}$  chlorophyll  $a$  from an OLS model using a backwards testing strategy based on Type III sums of squares. Maceina and Soballe (1990) used annual, whole lake average values to determine that wind was a better predictor of chlorophyll  $a$  than stage, but taking the annual average loses important information regarding the variability at time scales below one year, as discussed by Lamon et al. (1996). While WIND does not appear as often as STAGE in the model-space plots (Figure 2), the importance of the interaction terms suggests that both are useful predictors of chlorophyll  $a$ . This is further supported by the interaction surface plots for WIND and STAGE discussed later.

The other common feature for all zones is that Wind Direction has a weak effect, with a

probability of no effect ranging between 0.38 (data may be inconclusive) in the peat zone to 0.71 (positive evidence of no effect) for the mud zone. The probability of no relationship between  $\log_{10}$ TN and chlorophyll  $a$  is small for all zones (posterior probability  $< 0.023$ ), however, for the littoral zone, there is only moderate evidence that the spline terms (evidence of a nonlinear effect) are needed (posterior probability = 0.73). The evidence of a temperature effect varies with zone. In the sand zone, there is strong evidence to support a nonlinear  $\log_{10}$ TEMP effect. For the mud and littoral zones there is strong evidence in favor of a  $\log_{10}$ TEMP effect, but only moderate evidence in favor of a nonlinear relationship in the log-log scale. For the peat zone, there is moderate evidence supporting a  $\log_{10}$ TEMP effect (posterior probability = 0.714) with the posterior probability that the effect is only linear in the log-log scale equal to 0.296.

## 4.2 FUNCTIONAL RELATIONSHIPS

To understand the functional relationships between the predictors and the chlorophyll  $a$  concentration response, it is helpful to view partial residuals plots of the splined terms (Figure 3). The six terms in each of these plots are made by summing the effects of all eleven piece-wise linear basis functions comprising each term. Partial residual plots for the mud sediment region model (Figure 3a) indicate a distinct nonlinear relationship in the  $\log_{10}$  TP -  $\log_{10}$  chlorophyll  $a$  relationship, and indicate a decreasing effect as the scaled  $\log_{10}$  TP concentration increases above 0.5 (corresponding to a TP concentration of about  $0.07 \text{ mgL}^{-1}$ ). In the log-log scale, there is not a linear relationship between chlorophyll  $a$  concentrations and phosphorus, as many researchers have assumed. The  $\log_{10}$  TN -  $\log_{10}$  chlorophyll  $a$  plot indicates a significant contribution to the prediction of chlorophyll  $a$  in the mud sediment type. The effect on  $\log_{10}$  chlorophyll  $a$  concentration of  $\log_{10}$  STAGE is of smaller magnitude than  $\log_{10}$  TP,  $\log_{10}$  TN and  $\log_{10}$  TEMP in the mud sediment type, and less than the  $\log_{10}$  STAGE effect in other regions of the lake, probably because the depth of the lake in the mud sediment region stations is large relative to the change in lake stage due the fact that this is the deepest portion of the lake. Partial residuals plots for the sand sediment region model (Figure 3b) indicate a large  $\log_{10}$  STAGE and  $\log_{10}$  TN effect, nearly equal in magnitude to that of  $\log_{10}$  TP. The decreasing effect of total phosphorus beyond a scaled value of about 0.5 (TP about  $0.07 \text{ mgL}^{-1}$ ) in the sand sediment region mirrors that of the mud region model, however an increasing effect is apparent at total phosphorus concentrations below this level. A slight decreasing effect of increased wind velocity on  $\log_{10}$  chlorophyll  $a$  is noted in this region, that is absent in all but the peat sediment region model.

*(Figure 3 about here)*

The magnitude of the  $\log_{10}$  TP effect is very large in the peat sediment region model, as evidenced by the partial residual plot (Figure 3c). It is interesting to note the range of total phosphorus values in the peat sediment type is narrower than in other sediment types. This is likely an effect of the adsorption capacity of the peat sediment. The magnitude of the  $\log_{10}$  STAGE effect is second only to the  $\log_{10}$  TP effect in this region. As is the case with the peat sediment

region model, a small decreasing effect of increasing  $\log_{10}$  wind velocity on  $\log_{10}$  chlorophyll *a* is noted at transformed  $\log_{10}$  wind velocities above about 0.5 (corresponding to a wind velocity of  $15 \text{ m s}^{-1}d$ ). In the littoral sediment region, the partial residuals plots (Figure 3d) indicate a large  $\log_{10}$  TP effect on  $\log_{10}$  chlorophyll *a* concentration. As was the case in the mud and sand sediment region models, there is a peak in this function near a transformed  $\log_{10}$  TP value of 0.5 (TP about  $0.07 \text{ mgL}^{-1}$ ), and there is apparently an increasing effect at TP concentration below this level similar, though different in magnitude, to those found for the sand and peat sediment types. At scaled  $\log_{10}$  TP concentrations above 0.5 ( $0.07 \text{ mgL}^{-1}$ ), this term more closely resembles that of the sand sediment type.

To view the interaction effects it is useful to construct response surfaces of the interaction contribution to chlorophyll *a* concentration prediction. The response surface in such plots are analogous to the function estimates in the spline term plots presented in Figure 3 and combine the interaction spline terms in addition to all other terms involving WIND and STAGE. Such plots for the  $\log_{10}$  STAGE  $\times$   $\log_{10}$  WIND speed interactions are presented in Figure 4. By comparison of the plots in Figure 4 to the main effects in Figure 3 for  $\log_{10}$  STAGE and  $\log_{10}$  wind speed, it appears that the interaction surfaces draw most of their shape and magnitude from the main effects for the mud, sand and littoral sediment types. The peat sediment is an exception to this, since the magnitude of this interaction effect in Figure 4 is greater than the sum of the main effects in Figure 3. For the mud and sand sediments, most of the variation occurs along the  $\log_{10}$  STAGE axis, regardless of the value of  $\log_{10}$  wind speed. Most of the variation occurs along the  $\log_{10}$  STAGE axis for the littoral sediments also, but only for large values of  $\log_{10}$  wind speed. The magnitude of the effect is smallest in the mud sediment, which may be attributed to the fact that since this is the deepest portion of the lake, a unit change in  $\log_{10}$  STAGE changes the total depth less here than in the other sediment zones. The magnitude of the effect is similar in the sand and littoral zones. For the peat sediment, the effect is small only for small values of  $\log_{10}$  wind speed, and the magnitude of the effect is largest of all sediment types. The peat sediments are located in a relatively small portion of the southern end of the lake, whereas the sand and littoral sediment have a much larger spatial extent, and are located primarily in southwest to northwest extremes. For this reason, the fetch length for the dominant wind direction is greatest for the peat sediments, which may be a reason for the larger magnitude there.

*(Figure 4 about here)*

### 4.3 PREDICTIVE MODEL VALIDATION

To compare the predictive performance of Bayesian model averaging we set aside just under 50% of the data in each sediment zone for validation. We compared BMA to model selection with ordinary least squares using 1) coverage of 90% prediction intervals for each observation in the validation sample and 2) predictive mean squared error. The OLS model was obtained by backwards selection starting with the same basis as used on BMA, and potentially contains the same knots for the

nonlinear and interaction terms. This is a common strategy used in water quality studies to simplify models.

The coverage calculations were based on calculating a 90% prediction interval for each case in the validation set, and determining the percentage of intervals that contained the observed  $\log_{10}$  chlorophyll  $a$  concentrations. The predictive intervals for BMA were obtained by finding the 0.05 and 0.95 quantiles of the predictive cumulative distribution function in (8) using  $p(\gamma|Y)$  estimated from the Gibbs sampler output as in (9) (BMA RF) and in (10) (BMA RPP). The predictive MSE for each method was calculated as

$$\frac{1}{N^*} \sum_{i=1}^{N^*} (Y_i^* - \hat{Y}_i^*)^2$$

where  $\hat{Y}_i^*$  is the predicted mean from the corresponding method, using the relative frequencies and the renormalized posterior model probabilities and  $N^*$  is the sample size for the validation sample (Table 3).

*(Table 3 about here)*

*(Table 4 about here)*

Overall the coverage for BMA is higher than OLS, with the exception of the mud sediment zone (Table 3). There is very little difference between using the relative frequencies and the renormalized posterior model probabilities to estimate the posterior model probabilities, with slightly better performance using the relative frequencies in all cases except the peat zone. These coverage values are in line with results from Raftery et al. (1997), who found predictive coverage of 90% predictive intervals with BMA to be between 72% and 92% in both real and simulated data.

Similar results are obtained with predictive MSE (Table 4). BMA using the relative frequency model probabilities is 16% more efficient than OLS in the littoral zone, 20% more efficient in the sand zone, and up to 149% more efficient in the peat zone. As with coverage, OLS is slightly more efficient (6%) in the mud zone.

Overall, the BMA procedure is almost 30 percent more efficient in terms of MSE than OLS with model selection. While the coverage does fall short of the stated 90% levels for all methods, this could be due to the effect of the prior distribution (in general Bayes procedures will not have the stated coverage) or possible evidence that there are other variables, such as seasonal or yearly effects, beyond those measured by the functions of temperature, wind speed, etc. that could explain additional model uncertainty. As in Raftery et al. (1997), overall BMA has better predictive performance than model selection in terms of both coverage and predictive MSE.

## 5 DISCUSSION

We have identified important, nonlinear relationships between  $\log_{10}$  TP and  $\log_{10}$  chlorophyll  $a$  in all sediment zones of Lake Okeechobee. This is not necessarily in conflict with the findings of previous researchers who have documented the lack of a significant linear relationship between these

two variables. Maceina (1990) found no significant ( $\alpha = 0.05$ ) *linear* relationship (emphasis added) between  $\log_{10}$  TP and  $\log_{10}$  chlorophyll *a* in the open-water pelagic region of Lake Okeechobee ( $r = 0.1$ ), which was attributed to light limitation. The BMA spline approach presented here has identified an important *nonlinear* association, with levels of chlorophyll *a* increasing and then decreasing as TP levels increase for three of the ecological zones, which may explain why evidence of a significant linear association was lacking in other studies. Further, this nonlinear relationship is consistent with the light limitation hypothesis. In much of Lake Okeechobee,  $\log_{10}$  chlorophyll *a* increases (nearly) linearly with increasing  $\log_{10}$  TP only up to a rescaled  $\log_{10}$  TP value of around 0.5, or a TP value of about  $0.07 \text{ mgL}^{-1}$  (Figure 3). Beyond this value of TP, additional TP does not result in additional chlorophyll *a* (littoral and peat sediment), and may result in a decrease. We believe that the nonlinear functional relationship identified here provides empirical evidence that algae are limited at times by factor(s) other than TP. Researchers proposing a sigmoid relationship between  $\log_{10}$  total phosphorus and  $\log_{10}$  chlorophyll *a* for other lakes (McCauley et al. 1989, Prairie et al. 1989, Watson et al. 1992) used a parametric model that was restricted to representation of the functional relationship as strictly increasing, therefore the decrease in chlorophyll *a* for TP above  $0.07 \text{ mgL}^{-1}$  suggested by the plots in Figure 3 could not be found.

Functional forms similar to those found for the relationship between  $\log_{10}$  TP and  $\log_{10}$  chlorophyll *a* are present for  $\log_{10}$  TN and  $\log_{10}$  temperature (Figure 3). We interpret these in a similar fashion- factors other than TN or temperature become limiting as these functions begin to level off or decrease. Lamon et al. (1996) suggested, in the case of the  $\log_{10}$  TP- $\log_{10}$  chlorophyll *a* relationship, that this was due to light limitation brought about by resuspended sediment associated with high TP values. This is supported by empirical evidence in the form of a negative correlation ( $r = -0.45$ ) between  $\log_{10}$  secchi depth v.  $\log_{10}$  TP. Secchi depth is a measure of water clarity that is not independent of chlorophyll *a* concentration, with high values indicating relatively clear water. As  $\log_{10}$  TP increases, secchi depth values decrease, not only in the flocculent mud sediments, but in the other sediment types as well. Since low secchi depths can indicate conditions unfavorable for algae growth (i.e., light limitation) as well as high algal densities (since algal cells impede light penetration into the water column), secchi depth would be a poor predictor of chlorophyll *a*. Since there was no direct measure of non-algal turbidity, which could have been used as a regressor to explain the bend in the  $\log_{10}$  TP- $\log_{10}$  chlorophyll *a* relationship, we attempted to explain this behavior using interactions between other (measured) environmental factors, such as sediment type, wind velocity and lake stage.

The  $\log_{10}$  STAGE- $\log_{10}$  chlorophyll *a* relationship has a sigmoid shape in the sand and peat sediment zones (Figure 3) that is lacking in the mud and littoral zones. The magnitude of this effect is also greater in the sand and peat zones. The littoral and mud zones have the shallowest and deepest average depths, respectively, with the averages depth of the sand and peat zones being in between these extremes. In the mud zone, the average depth may still be shallow enough that available wave energy is sufficient to resuspend the loosely consolidated flocculent mud sediments

there. In the littoral zone, increases in stage of the magnitude seen here may not be sufficient to have an effect on resuspension. However, when stage is at its highest in the littoral zone, there is an effect (decreasing) on  $\log_{10}$  chlorophyll  $a$ , which may be attributable to transport of sediment laden waters from the mud zone. Transport of sediment laden waters from the mud zone is suspected to be responsible for the shape of the STAGE/WIND interactions in the mud, sand and littoral zones (Figure 4).

In general, there are three potential sources of model uncertainty: uncertainty arising from *misspecification* of the functional form of the relationship between the predictors and the response, and uncertainty resulting from either the *inclusion* of insignificant predictors or the *omission* of important predictors, and uncertainty regarding *parameter* values given a set of included predictors and the specification of their functional form. While the method used here can account for uncertainty from misspecification, inclusion of unimportant predictors and from parameter values, it cannot account for uncertainty deriving from errors of omission of predictors.

A variety of regression methods have been used for the purpose of prediction of chlorophyll  $a$  in Lake Okeechobee, and have followed a progression of increasing complexity and insight into the functional relationships between environmental factors and chlorophyll  $a$  concentrations. The ordinary least squares model (Lamon 1995) provides prediction at a finer spatial scale than a previous mechanistic model developed by SFWMD (James and Bierman 1995) by fitting different slopes and intercepts for each of five previously described ecological zones of Lake Okeechobee. The early mechanistic model used to predict chlorophyll  $a$  predicted only whole lake average concentrations. Residual analysis of the OLS model suggested nonlinear relationships between the chlorophyll  $a$  concentration and some of the predictors, which led to the use of generalized additive models (Lamon et al. 1996). The model developed here addresses conceptual shortcomings of the OLS (Lamon 1995) and GAM models (Lamon et al. 1996). Semi-parametric regression with Bayesian variable selection and product spline interactions provides a flexible framework for addressing the problems identified with the previous work. The use of regression splines allows nonlinear effects to be manifest, while their extension allows inclusion of interactions for which the mathematical form cannot be specified *a priori*. Also, the use of Bayesian model averaging deals with the problem of model uncertainty that was not addressed in the previous models. This approach provides the water quality modeler with an effective means of communicating the importance of various candidate predictors to decision makers, and can therefore be useful in the monitoring design (or redesign) process. As BMA leads to more accurate predictions and prediction intervals, this in turn should lead to better decisions and policy.

## ACKNOWLEDGMENTS

This research was partially supported by NSF grants DMS-96.26135 and DMS-97.33013. We would like to thank the South Florida Water Management District for providing data and support of the

first author and the National Research Center for Statistics and the Environment at the University of Washington for partial support of the second author.

## REFERENCES

- Aldridge, F. J., Philips, E. J., and Schelske, C. L. (1995), "The Use of Nutrient Enrichment Bioassays to Test for Spatial and Temporal Distribution of Limiting Factors Affecting Phytoplankton Dynamics in Lake Okeechobee, Florida," *Archiv für Hydrobiologie Beiheft Ergebnisse der Limnologie* 45, 177-190.
- A.P.H.A. (1971-1989), *Standard Methods for the Analysis of Water and Wastewater*. 13th - 17th Editions. American Public Health Association, Washington, D.C.
- Canfield, D. E. and Hoyer, M. V. (1988), "The Eutrophication of Lake Okeechobee," *Lake and Reservoir Management* 4, 91-99
- Canfield, D. E., Jr., Linda, S. B., and Hodgson, L. M. (1985), "Chlorophyll-Biomass-Nutrient Relationships for Natural Assemblages of Florida Phytoplankton," *Water Resources Bulletin* 21, 381-391.
- Carper, G. L. and Bachmann, R. W. (1984), "Wind Resuspension of Sediments in a Prairie Lake." *Canadian Journal of Fisheries and Aquatic Sciences* 41, 1763-1767.
- Clyde, M., DeSimone, H., and Parmigiani, G. (1996), "Prediction via Orthogonalized Model Mixing," *Journal of the American Statistical Association* 91, 1197-1208.
- Clyde, M. (1999), "Bayesian Model Averaging and Model Search Strategies," In *Bayesian Statistics 6*, eds. Bernardo, J. M., Berger, J. O., Dawid, A. P. and Smith, A. F. M. Oxford University Press, pages 157-185.
- Friedman, J. H. (1991), "Multivariate Adaptive Regression Splines," *The Annals of Statistics* 19, 1-141.
- Gelfand, A. E. and Smith, A. F. M. (1990), "Sampling-based Approaches to Calculating Marginal Densities," *Journal of the American Statistical Association* 85, 398-409.
- George, E. I. and Foster, D. P. (1997), "Calibration and Empirical Bayes Variable Selection," Tech Report, University of Texas, Austin.
- George, E. I. and McCulloch, R. (1993), "Variable Selection via Gibbs Sampling", *Journal of the American Statistical Association* 88, 881-889.
- George, E. I. and McCulloch, R. (1997), "Approaches for Bayesian Variable Selection," *Statistica Sinica* 7, 339-374.
- Hastie, T. J. and Tibshirani, R. J. (1990), *Generalized Additive Models*. Chapman and Hall, New York.
- Hoeting, J. A., Madigan, D., Raftery, A. E., and Volinsky, C. T. (1999), "Bayesian Model Averaging: A Tutorial". To appear in *Statistical Science*.
- James, R. T. and Bierman, V. J. Jr. (1995), "A Preliminary Modeling Analysis of Water Quality

- in Lake Okeechobee, Florida: Calibration Results,” *Water Research* 29, 2755–2766.
- James, R. T., Smith, V. H. and Jones, B. L. (1995), “Historical Trends in the Lake Okeechobee Ecosystem III: Water Quality,” *Archiv für Hydrobiologie Supplement-Band* 107, 49-69.
- Kass, R. E. and Raftery, A. E. (1995), “Bayes Factors,” *Journal of the American Statistical Association*, 90, 773-795.
- Lamon, E. C. (1995), “A Regression Model for the Prediction of Chlorophyll *a* in Lake Okeechobee,” *Lake and Reservoir Management* 11, 283-290.
- Lamon, E. C., Reckhow, K. H. and Havens, K. E. (1996), “Using Generalized Additive Models for Prediction of Chlorophyll *a* in Lake Okeechobee, Florida,” *Lakes and Reservoirs: Research and Management* 2, 37-46.
- Maceina, M. J. and Soballe, D. M. (1990), “Wind-related Limnological Variation in Lake Okeechobee, Florida,” *Lake and Reservoir Management* 6, 93-100.
- Maceina, M. J. (1993), “Summer Fluctuations in Planktonic Chlorophyll *a* Concentrations in Lake Okeechobee, Florida: The Influence of Lake Levels,” *Lake and Reservoir Management* 8, 1-11.
- McCauley, E., Downing, J. A. and Watson, S., (1989), “Sigmoid Relationships Between Nutrients and Chlorophyll Among Lakes,” *Canadian Journal of Fisheries and Aquatic Sciences* 46, 1171-1175.
- Phlips, E. J., Aldridge, F. J., Hansen, P., Zimba, P. V., Inhat, J., Conroy, M. and Ritter, P. (1993), “Spatial and Temporal Variability of Trophic State Parameters in a Shallow Subtropical Lake (Lake Okeechobee, Florida, USA),” *Archiv für Hydrobiologie* 128, 437-458.
- Prairie, Y. T., Duarte, C. M. and Kalff, J. (1989), “Unifying Nutrient-Chlorophyll Relationships in Lakes,” *Canadian Journal of Fisheries and Aquatic Sciences* 46, 1176-1182.
- Raftery, A. E., Madigan, D. and Hoeting, J. A., (1997), “Bayesian Model Averaging for Linear Regression Models,” *Journal of the American Statistical Association* 92, 179-191.
- Reckhow, K. H., (1988), “Empirical Models for Trophic State in Southeastern U.S. Lakes and Reservoirs,” *Water Resources Bulletin* 24, 723-734.
- Reddy, K. R., (1992), “Lake Okeechobee Phosphorus Dynamics Study, Volume III: Biogeochemical Processes in the Sediments,” South Florida Water Management District.
- Schelske, C., (1989), “Assessment of Nutrient Effects and Nutrient Limitation in Lake Okeechobee,” *Water Resources Bulletin* 25, 1119-1130.
- Smith, M. and Kohn, R. (1996), “Nonparametric Regression Using Bayesian Variable Selection,” *Journal of Econometrics* 75, 317-367.
- South Florida Water Management District, (1989), “Interim Surface Water Improvement (SWIM) Plan for Lake Okeechobee, Part I: Water Quality,” February 9, 1989.
- South Florida Water Management District, (1992), “Draft Surface Water Improvement (SWIM) Plan Update for Lake Okeechobee, Volume 1: Planning Document,” June 11, 1992.
- Strickland, J. D. H. and Parsons, T. R., (1968), *A Practical Handbook of Seawater Analysis*. Fisheries Research Board of Canada.

Watson, S., McCauley, E. and Downing, J. A., (1992), "Sigmoid Relationships between Phosphorus, Algal Biomass and Algal Community Structure," *Canadian Journal of Fisheries and Aquatic Sciences* 49, 2605-2610.

Table 1: Minimum, maximum and range of data values used to scale predictors from 0 to 1 for use in regression spline basis functions. Data values were subtracted from the minimum and divided by the range. For  $\log_{10}$  TP observation *i*, transformed  $\log_{10}TP = (\log_{10}TP_i - (-2.301))/2.301$ .

Parameter	Minimum	Maximum	Range
logTEMP	1.1048	1.5997	0.4948
logTP	-2.301	0.0043	2.3054
logTN	-0.8239	1.0253	1.8492
logSTAGE	1.022	1.2167	0.1947
logWIND	0	2.3662	2.3662
Direction	0	331.24	331.24

Table 2: Posterior probability of no effect and no nonlinear effect for each variable by sediment type

Zone	Variable	P(no effect   Y)	P(no nonlinear effect   Y)
littoral	logTP	< 0.001	< 0.001
	logTN	0.023	0.270
	logTEMP	0.100	0.375
	logSTAGE	< 0.001	< 0.001
	logWIND	0.001	0.001
	Direction	0.402	0.442
	peat	logTP	< 0.001
logTN		< 0.001	< 0.001
logTEMP		0.286	0.296
logSTAGE		< 0.001	0.002
logWIND		0.005	0.005
Direction		0.383	0.519
sand		logTP	< 0.001
	logTN	< 0.001	< 0.001
	logTEMP	< 0.001	0.002
	logSTAGE	< 0.001	< 0.001
	logWIND	< 0.001	< 0.001
	Direction	0.696	0.750
	mud	logTP	< 0.001
logTN		< 0.001	0.007
logTEMP		< 0.001	0.218
logSTAGE		< 0.001	< 0.001
logWIND		0.002	0.002
Direction		0.714	0.725

Table 3: Coverage of 90% prediction intervals for the validation data for OLS and BMA with posterior model probabilities estimated using relative frequencies (RF) and renormalized posterior model probabilities (RPP). The best method is indicated in bold.

Zone	Total sample size	Training sample size	Validation sample size	BMA RF	BMA RPP	OLS
littoral	957	497	460	<b>86.5</b>	85.2	80.2
peat	966	492	474	85.2	<b>86.2</b>	80.7
sand	1,881	1,497	1,384	<b>79.5</b>	79.0	76.1
mud	2,349	1,183	1,166	76.0	74.8	<b>76.9</b>

Table 4: Predictive Mean Square Error (MSE) for the validation data for OLS and BMA with posterior model probabilities estimated using relative frequencies (RF) and renormalized posterior model probabilities (RPP). The best method is indicated in bold.

Zone	BMA RF	BMA RPP	OLS	Relative Efficiency MSE(OLS)/MSE(RF)	Relative Efficiency MSE(OLS)/MSE(RPP)
littoral	<b>0.1688</b>	0.1712	0.1965	1.16	1.15
peat	0.1435	<b>0.1323</b>	0.3575	2.49	2.70
sand	<b>0.1307</b>	0.1328	0.1569	1.20	1.18
mud	0.1944	0.2026	<b>0.1834</b>	0.94	0.91



Figure 1: Location of monitoring stations in Lake Okeechobee coded by principal sediment types of Reddy et al. (1991). The mud sediment zone corresponds roughly to the pelagic zone of Philips et al. (1993).

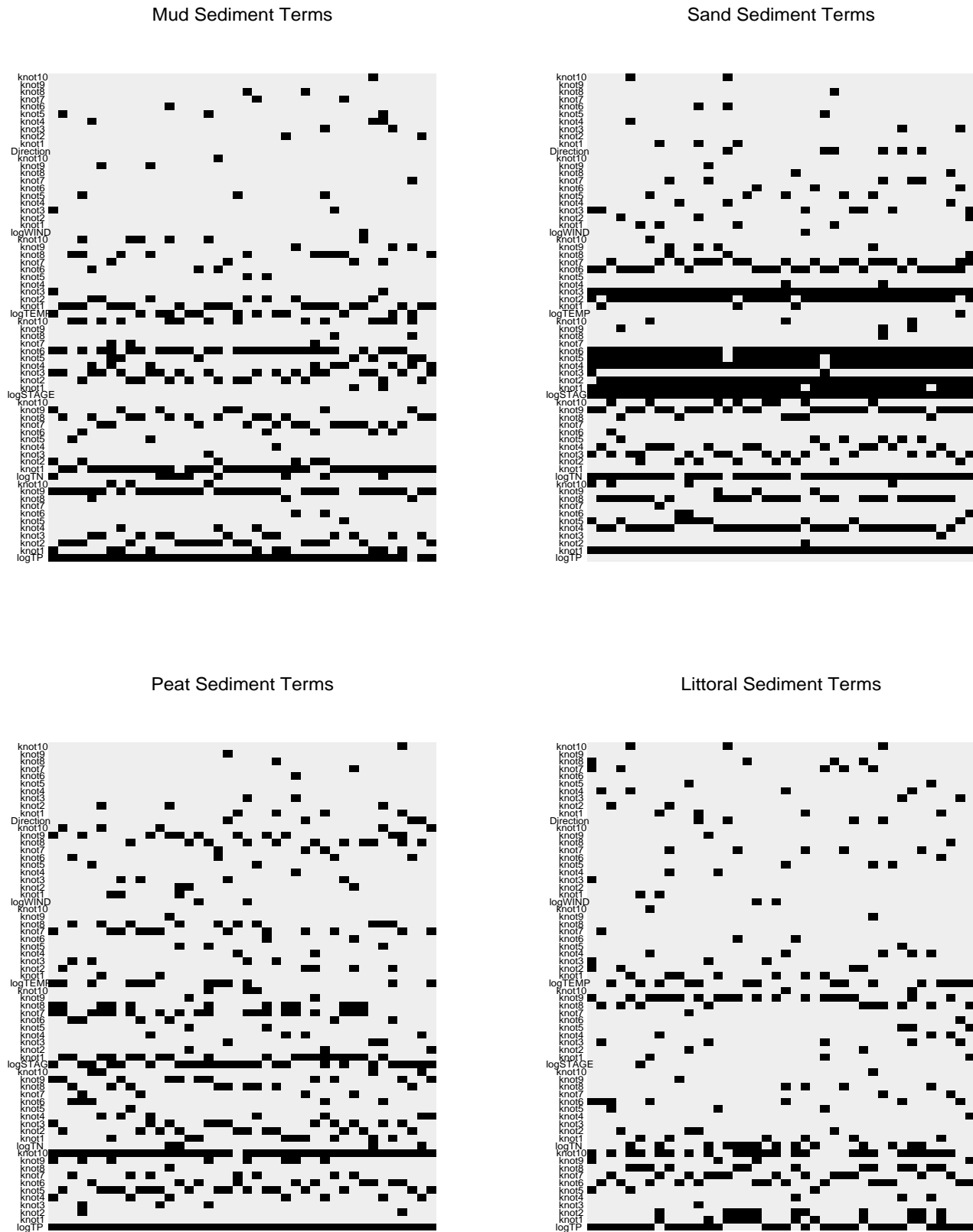
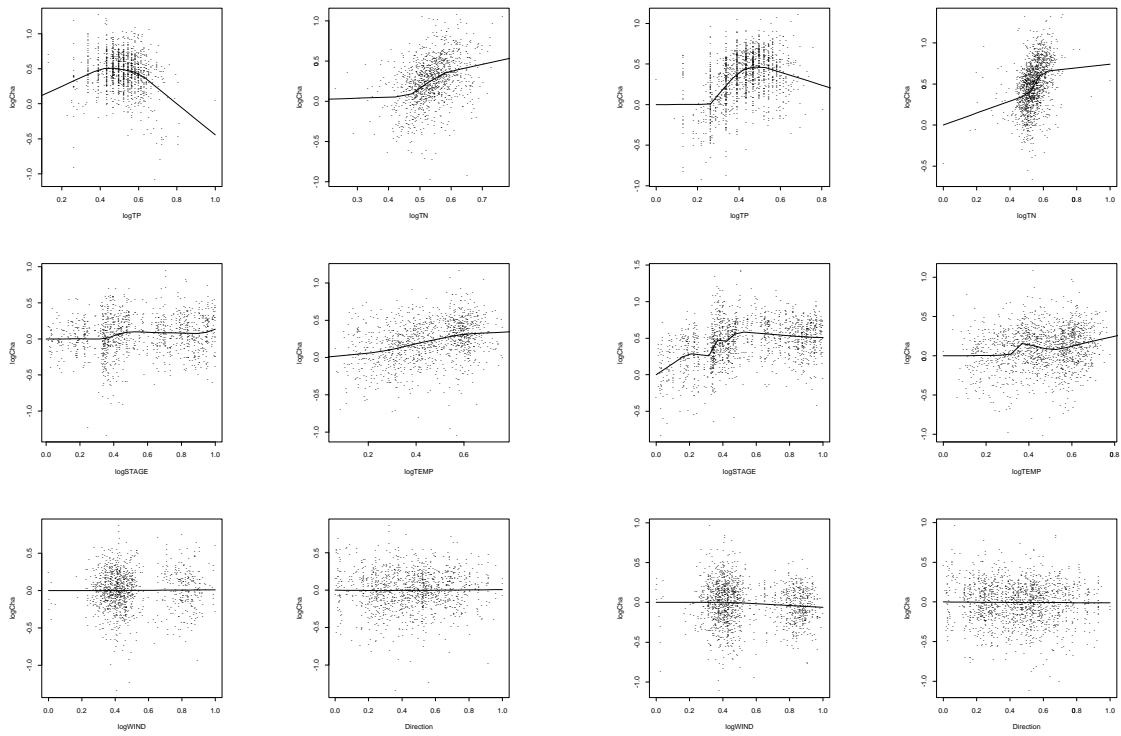
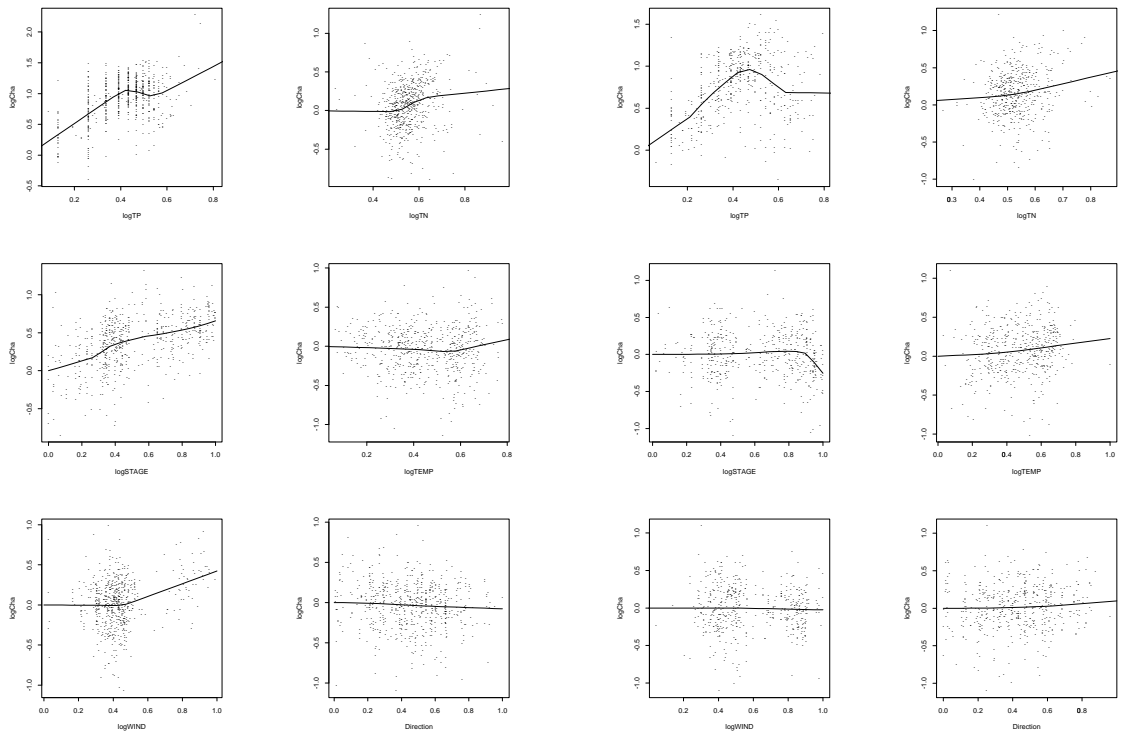


Figure 2: Variable importance matrix for the 4 zones. The more filled boxes in each row of the matrix, the more important the variable (basis function), i.e. the more it was included in the 40 most likely models visited by the Gibbs sampler.



(a) Mud Zone

(b) Sand Zone



(c) Peat Zone

(d) Littoral Zone

Figure 3: *Partial residuals plots for the splined terms in (a) the mud zone, (b) the sand zone, (c) the peat zone and (d) the littoral zone.*

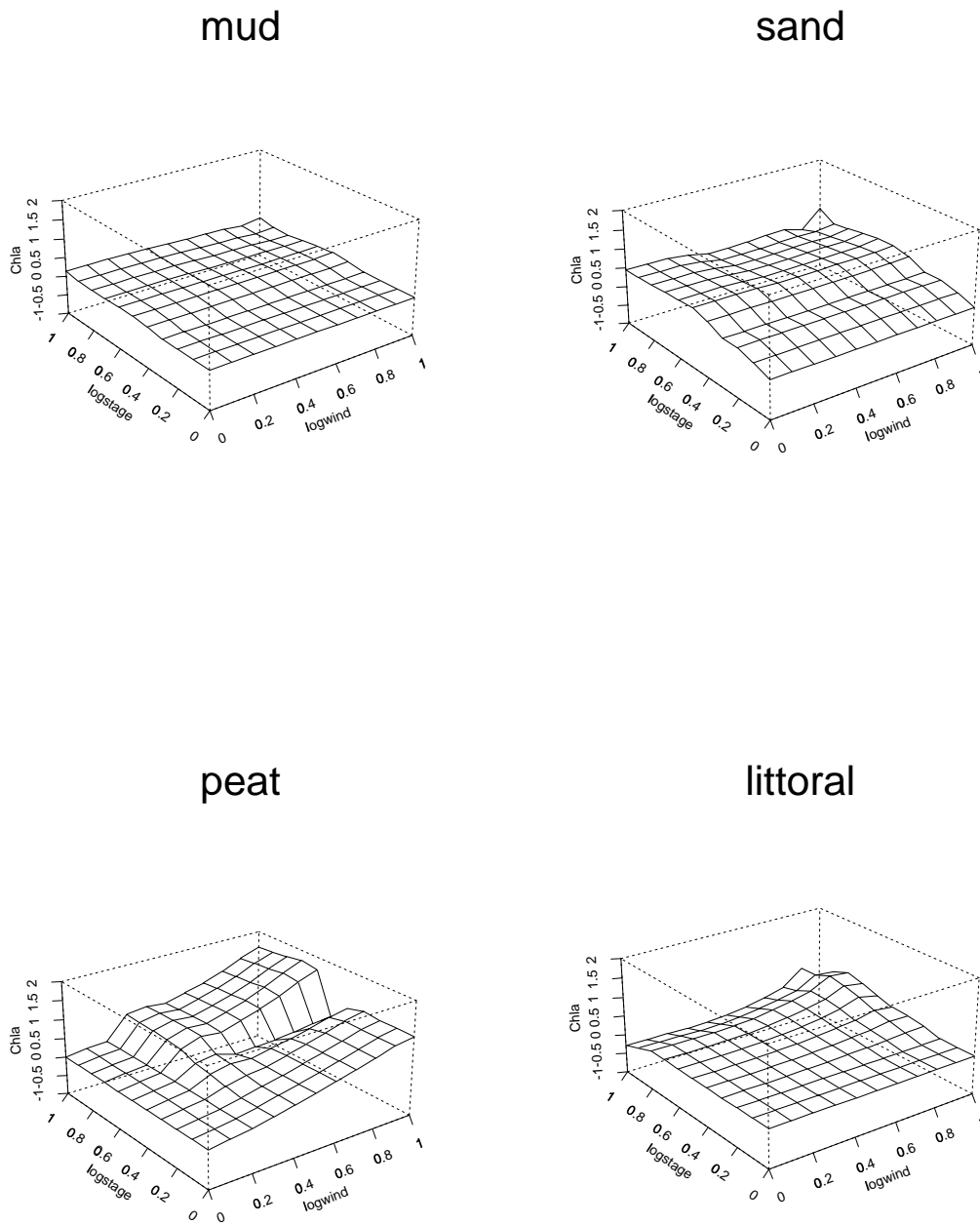


Figure 4: *Plots of the predicted chlorophyll a concentration on the log10 wind - log10 stage plane for each sediment type. The surfaces here are the product spline interaction terms for log10 wind and log10 stage.*