ELSEVIER

Ecological Modelling 116 (1999) 33–44

ECOLOGICAL MODELLING

Statistical modeling of total phosphorus concentrations measured in south Florida rainfall

Hosung Ahn *

Resources Assessment Division, South Florida Water Management District, 3301 Gun Club Road, West Palm Beach, FL 33406, USA

Received 25 February 1998; accepted 19 August 1998

Abstract

Atmospheric deposition can be a significant source of phosphorus to South Florida's aquatic system. The weekly total phosphorus (TP) concentrations in rainfall have been measured routinely in the region since 1974, but the historical data set has significant gaps due to instrumental failures and sample contamination. This study attempts to develop a statistical model of rainfall-borne TP concentration to estimate missing data. The model is based on a multivariate stochastic time-series theory. The model parameters and noise covariances were calibrated using the expectation maximization algorithm which is known to be efficient for data sets with many gaps. Model verification demonstrates that the calibrated model provides unbiased data estimates while preserving the statistics of the raw data. The data with gaps filled in are useful for computing the weekly TP loads. © 1999 Elsevier Science B.V. All rights reserved.

Keywords: Atmospheric deposition; Total phosphorus; Missing data; Kalman filter; Time series model; Expectation–maximization algorithm

1. Introduction

Phosphorus concentrations of aquatic systems are directly related to eutrophication and to the structure of the aquatic vegetation community. The management of phosphorus inputs to the South Florida ecosystem has become an increasing concern resulting in the need for accurate monitoring and analyses of phosphorus distribution in the region. The South Florida water management district (District) has collected atmospheric deposition data in the region since 1974. The monitoring program was significantly improved in 1992 with the deployment of wet/dry collectors (Aerochem Metrics Model 301 automatic wet/dry sampler) and adoption of a standard operating procedure for atmospheric data

* Tel.: +1-561-6876516; fax: +1-561-6876442; e-mail: hosung.ahn@sfwmd.gov.

0304-3800/99/$ - see front matter
PII: S0304-3800(98)00159-8

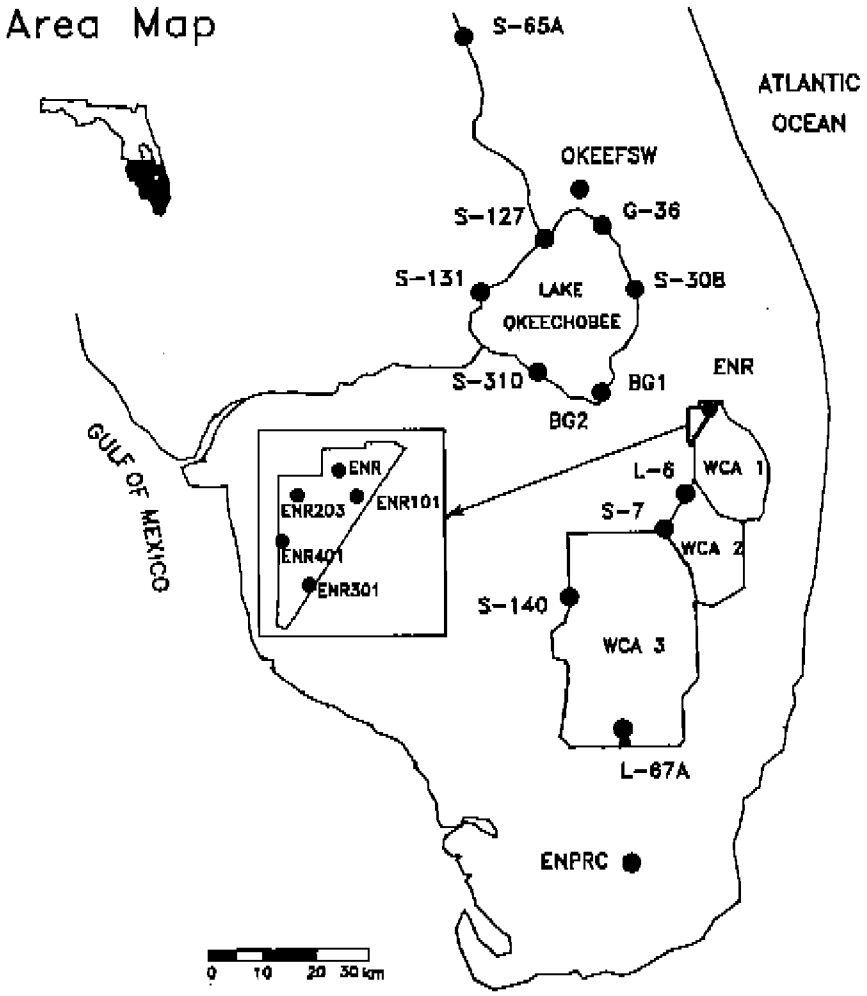


Fig. 1. Locations of the atmospheric deposition monitoring sites operated by the South Florida water management district.

collection and processing. Currently, there are a total of 19 atmospheric deposition monitoring sites operated by the District (Fig. 1). Wet and dry deposition data have been collected in weekly intervals and analyzed at the District's laboratory to determine the level of nutrients and major ions.

However, there is a significant amount of missing data in the measured nutrient data sets caused by instrumental failure and sample contamination due to bird droppings and other foreign matter. About 64% of rainfall total phosphorus (TP) concentration data collected on a weekly basis in the region is missing from the historical data sets (Ahn, 1997). The amount of rainfall-borne phosphorus loads to the ecosystem can be accurately estimated at weekly time intervals since corresponding rainfall is highly variable in space and time. Although the gaps in the data are not necessarily detrimental to quantifying monthly or yearly summary statistics, they do preclude calculation of weekly rainfall TP loads. If the physical processes driving the occurrence and transportation mechanism of atmospheric deposition are known, one could build a mathematical model to estimate the data gaps. However, neither a mathematical model nor supporting input data for the model on a regional scale are available. Alternatively, one can adopt an empirical approach using

a statistical model based on currently available data to estimate the missing data. Thus the objective in this study was to develop a statistical model to estimate the missing data in the phosphorus concentration of rainfall.

This study uses a multivariate time-series model with covariate terms since the data are measured at multiple sites. For complete (no missing) data sets, a variety of numerical algorithms, such as Gauss–Newton method and scoring method (Box and Jenkins, 1976; Brockwell and Davis, 1987; Harvey, 1990), are available to estimate parameters in the time-series models, but they are not applicable for an incomplete data set. For incomplete data sets which have some data gaps in them, it has been known (Dempster et al., 1977; Shumway and Stoffer, 1982; Stoffer, 1985, 1986) that an expectation–maximization (EM) algorithm is suitable for estimating parameters of time-series models. A pre-condition to applying the EM algorithm is to set the model into state–space form to estimate Kalman filtering and smoothing estimates. The Kalman filter and smoother recursions provide a convenient means for calculating the conditional expectations of both state and error vectors. The reason for using smoothing in this case is to take advantage of the forward measurement information and to give a fast convergence in the EM algorithm. The Kalman filtering in conjunction with a stochastic time series model has been widely applied for ecological modeling and data analyses (Padgett and Papadopoulos, 1979; Chen and Papadopoulos, 1988; Tiwari and Dienes, 1994; Boudjema and Chau, 1996), but no specific work has been found for dealing with incomplete ecological data sets. Thus the overall EM algorithm applied to the TP data observed from multiple sites is introduced in the next section.

2. Method

2.1. Autoregressive model with covariate

Consider a multi-variate state vector $x_t = (x_{t,1},...,x_{t,nx})'$ at time t $(= 1,...,T)$, where (nx) is the number of state sizes, T is the time span, and $(\,)'$ denotes transpose of a matrix. With the (nz) multiple covariate vector $z_t = (z_{t,1},...,z_{t,nz})'$ which is measured completely and concurrently, the order-q multivariate autoregression model is given by:

$$x_t = \sum_{i=1}^{q} \phi x_{t-i} + \psi z_t + w_t \tag{1}$$

where $\phi(\mathrm{nx} \times \mathrm{nx})$ and $\psi(\mathrm{nx} \times \mathrm{nz})$ are the regression parameters, and w_t is the Gaussian white noise with $w_t \approx N(0, Q)$. For wet TP concentration data, x_t could represent a TP vector measured from nx multiple sites at time t, while z_t may be a concurrently measured covariate vector having a size of nz.

To estimate the parameters $\{\phi, \psi, Q\}$, an EM algorithm can be applied in conjunction with the modified Kalman smoother estimators to derive a simple recursive procedure. The EM algorithm is known to be an alternative non-linear optimization algorithm suitable for estimating missing data (Dempster et al., 1977; Shumway, 1988). To apply the Kalman filter recursion, Eq. (1) should be set into a state–space form which consists of state and measurement equations. With an augmented vector $X(t) = [x_t,...,x_{t\ q-1}]'$, the state equation of Eq. (1) is the form of:

$$x(t) = \begin{bmatrix} \phi_1 & \cdots & \phi_{q-1} & \phi_q \\ 1 & \cdots & 0 & 0 \\ \vdots & & \vdots & \vdots \\ 0 & \cdots & 1 & 0 \end{bmatrix} \begin{bmatrix} x_{t-1} \\ x_{t-2} \\ \vdots \\ x_{t-q} \end{bmatrix} + \sum \begin{bmatrix} \psi \\ 0 \\ \vdots \\ 0 \end{bmatrix}$$

$$[z_t] + \begin{bmatrix} w_t \\ 0 \\ \vdots \\ 0 \end{bmatrix} = \Phi x(t-1) + \Psi z(t) + w(t) \tag{2}$$

where $\Phi(\mathrm{nc} \times \mathrm{nc})$ and $\Psi(\mathrm{nc} \times \mathrm{nz})$ are the augmented matrices with a dimension of nc$(= \mathrm{nx} \times q)$, and I and 0 in the parameter matrices denote the one and zero diagonal matrices. To allow for estimating the missing data, the measurement equation is written by:

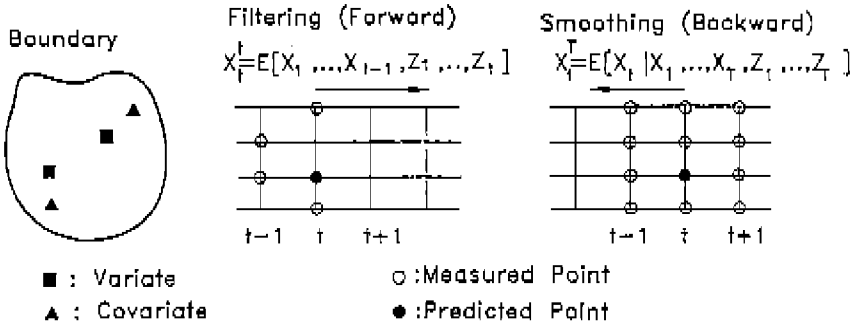


Fig. 2. Schematics of filtering and smoothing of a multivariate autoregressive model with covariate.

$$y_t = [m_t 0 \ldots 0] \begin{bmatrix} x_t \\ \vdots \\ x_{t-q+1} \end{bmatrix} + v_t = M(t)x(t) + v_t \tag{3}$$

where $y_t(\mathrm{nx} \times 1)$ is the state measurement vector at t, $M(t)$ is the $(\mathrm{nx} \times \mathrm{nc})$ measurement matrix in which the diagonal element $\{m_t\}$ is 1 if y_t is measured, or 0 otherwise. The measurement noise v_t is the Gaussian white noise having $v_t \approx N(0, R)$.

2.2. The Kalman filtering and smoothing

The problem of estimating $x(t)$ in Eq. (2) can be approached by the expectation of it conditioning on the measurements $(y_1,...,y_s)$ as:

$$x_t^s = E[x(t)|y_1,...,y_s,z_1,...,z_s] \tag{4}$$

where s is the span of the measurement. With the state estimation error vector $\tilde{x}_t$ defined as the true value minus the above estimated value, the error covariance can be estimated by:

$$p_t^s = E[\tilde{x}_t \tilde{x}_t' \mid y_1,...,y_s, z_1,...z_s]. \tag{5}$$

The following problems occur when estimating the x_t^s and p_t^s: if $t = s$, it is called a filtering problem; if $t < s$, it is a smoothing problem. Fig. 2 shows a schematic of filtering and smoothing for the given time-series model expressed by Eq. (1).

Based on the state–space form, the forward recursion $(t = 1,...,T)$ of the state and error covariance are given (Jazwinski, 1970) by

$$x_t^{t-1} = \Phi x_{t-1}^{t-1} + \Psi z(t) \tag{6}$$

$$p_t^{t-1} = \Phi p_{t-1}^{t-1} \Phi' + \begin{bmatrix} Q & 0 \\ 0 & 0 \end{bmatrix} \tag{7}$$

$$K_t = p_t^{t-1} M(t)' [M(t) p_t^{t-1} M(t)' + R]^{-1} \tag{8}$$

$$x_t^t = x_t^{t-1} + K_t [y_t - M(t) x_t^{t-1}] \tag{9}$$

$$p_t^t = p_t^{t-1} + K_t M(t) p_t^{t-1} \tag{10}$$

Defining the expectations of state and error covariances conditioning on all available measurements $(y_1,...,y_T)$ as:

$$x_t^T = E[x(t) \mid y_1,...,y_T, z_1,...,z_t] \tag{11}$$

$$p_t^T = E[\tilde{x}_t \tilde{x}_t' \mid y_1,...,y_T, z_1,...z_T], \tag{12}$$

respectively, the corresponding Kalman smoother in backward recursion $(t = T, T-1,...,1)$ is then given (Jazwinski, 1970) by:

$$J_{t-1} = p_{t-1}^{t-1} \Phi' (p_t^{t-1})' \tag{13}$$

$$p_{t-1,t-2}^T = E[(x_{t-1}^* - \hat{x}_{t-1}^T)(x_{t-2}^* - \hat{x}_{t-2}^T)' \mid y_1,...,y_T] \tag{14}$$

$$x_{t-1}^T = x_{t-1}^{t-1} + J_{t-1}(x_t^T - \Phi x_t^{t-1}) \tag{15}$$

$$p_{t-1}^T = p_{t-1}^{t-1} + J_{t-1}(p_t^T - p_t^{t-1}) J_{t-1}' \tag{16}$$

$$p_{t-1,t-2}^T = p_{t-1}^{t-1} J_{t-2}' + J_{t-1}[p_{t,t-1}^T - \Phi p_{t-1}^{t-1}] J_{t-2}' \tag{17}$$

where the Eq. (17) provides the lag-one smoothed error covariances needed for the expectation step (Shumway and Stoffer, 1981).

2.3. Expectation–maximization algorithm

A log likelihood, $\ln L(x_1,...,x_T|\theta)$, based on a complete data set $(x_1,...,x_T)$ can be estimated by

an expectation conditioning on θ which is the parameter set of the model to be estimated. The EM algorithm is designed to find θ iteratively by maximizing the expectation of the complete-data log likelihood conditioned on the measured data. That is, the expectation step at the i-th iteration computes the log-likelihood function of:

$$\text{Fct.}(\theta \mid \theta_i) = E_i[\ln L\{x(1),\ldots,x(T), v_1,\ldots,v_T; \theta \mid y_1,\ldots,y_T,z_1,\ldots,z_T; \theta_i\}]. \quad (18)$$

The maximization step then chooses θ_{i+1} to maximize Fct.$(\theta|\theta_i)$ using one of the optimization techniques. Since the $(x_1,\ldots,x_T)$ process cannot be measured directly, Eq. (18) can be written in terms of the Kalman smoothed estimators. The following expectation terms (Stoffer, 1985) are needed to compute the maximization step:

$$A = \sum_{t=1}^{T} (p_{t-1}^{t} + x_{t-1}^{T} x_{t-1}^{T\prime}),$$

$$B = \sum_{t=1}^{T} (p_{t,t-1}^{T} + x_{t-1}^{T} x_{t-1}^{T\prime}),$$

$$C = \sum_{t=1}^{T} (p_{t}^{T} + x_{t,1}^{T} x_{t,1}^{T\prime}),$$

$$F = \sum_{t=1}^{T} [x_{t-1}^{T} z(t)'],$$

$$G = \sum_{t=1}^{T} [x_{t,1}^{T} z(t)'] \quad \text{and}$$

$$H = \sum_{t=1}^{T} [z(t) z(t)'] \quad (19)$$

where $x_{t,1}^{T}$ is the first sub-vector in the $x_t^{T\prime} = [x_{t,1}^{T\prime},\ldots,x_{t,q+1}^{T\prime}]$, and the corresponding dimensions are: $A(\text{ns} \times \text{ns})$; $B(\text{nx} \times \text{ns})$; $C(\text{nx} \times \text{nx})$; $F(\text{ns} \times \text{nc})$; $G(\text{nx} \times \text{nc})$; and $H(\text{nc} \times \text{nc})$, with $\text{nc} = (\text{nx} \times \text{q})$. An incomplete-data log-likelihood function is calculated (Gupta and Mehra, 1974; Shumway and Stoffer, 1981) by:

$$2\ln L(Y) \approx \sum_{t=1}^{T} \ln\left|M(t) p_t^{t-1} M(t)'\right| + \sum_{t=1}^{T} e_t^{t-1\prime} [M(t) p_t^{t-1} M(t)' + R]^{-1} e_t^{t-1} \quad (20)$$

with $e_t^{t-1} = (y_t - M(t)x_t^{t-1})$, and x_t^{t-1} and p_t^{t-1} are obtained from the forward recursion.

The maximization step is then obtained by maximizing Eq. (18) with respect to the parameters θ, Q, and R. The resulting estimators (Shumway and Stoffer, 1982; Stoffer, 1985) are given by:

$$\theta_{i+1} = [B \quad G] \begin{bmatrix} A & F \\ F' & H \end{bmatrix}^{-1} \quad (21)$$

$$Q_{i+1} = \left(C - \theta_{i+1} \begin{bmatrix} B \\ G \end{bmatrix} - [B \quad G]\theta_{i+1}' + \theta_{i+1} \begin{bmatrix} A & F \\ F' & H \end{bmatrix} \theta_{i+1}' \right) / T \quad (22)$$

$$R_{i+1} = \frac{1}{T} \sum_{t=1}^{T} \{e_t e_t' + M(t) p_t^{T} M(t)'\} \quad (23)$$

where e_t is the estimation error vector expressed by $e_t = y_t - M(t)x_t'$.

The iterative procedure of the EM algorithm starts with an assumed initial parameter set, $\{\Phi(0), Q(0), R(0)\}$, where (0) indicates the initial time step before iterations. On the i-th iteration, the Kalman filtered and smoothed estimators are computed using Eq. (6) through Eq. (17), with the expectation step given by Eq. (19), the log-likelihood function by Eq. (20), and the maximization step by Eq. (21) through Eq. (23). The procedure continues until the minimum log-likelihood function was obtained.

3. Application

3.1. Formulating model structure

Four of the sites located in the Everglades nutrient removal (ENR) project area (ENR-101, -203, -301, -401) were not included in the modeling because of having an excessive amount of missing data. This study used the weekly rainfall-borne TP concentration data collected from 15 other monitoring sites for the maximum period of record from April 1992–November 1996. The actual periods of record vary from site to site due to periodic expansion of the monitoring program. Generally, it would not be technically difficult to build one multivariate time-series model for all 15 sites,

Table 1
Composition of time-series models and periods of record of the historical data used for model calibration

Model ID	Sites (Number of sites)	Periods of record (month/day/year)	Number of data points	Covariate sites
Model-I	S-65A, S-7 (2)	4/7/92–11/5/96	240	ENR, OKEEFS, S-140 (3)
Model-II	ENR, OKEEFS, S-140 (3)	4/7/92–11/5/96	240	S-65A, S-7 (2)
Model-III	S-308, S-310 (2)	4/7/92–11/5/96	240	ENR, OKEEFS, S-7 (3)
Model-IV	S-127, S-131, BG1, BG2, ENPRC (5)	9/7/93–10/22/96	164	OKEEFS, S-308, S-310 (3)
Model-V	G-36, L-67A, L-6 (3)	8/22/95–10/22/96	62	ENR, ENPRC, OKEEFS (3)

making it somewhat more efficient to estimate model parameters and values for missing data. However, developing such a model in this case was not possible because of the varying periods of record and the lack of covariate information other than the TP data from adjacent sites. Because of these limitations, this study constructed five separate models (Table 1).

In the time-series model described by Eq. (1), the state vector x_t at time t is a function of both the time-lagged state x_{t-q} and the concurrently measured covariate z_t. Without knowing proper covariates for the wet TP data, the concurrently measured wet TP data collected from sites adjacent to the one being modeled were used as the covariates in each model. While inter-site correlation of concurrently measured data is stronger than auto-correlation (correlation in time), selecting sites for the z_t vector is very important. Because z_t does not allow for gaps in the data, the model structure was designed to estimate model parameters and missing data in x_t sequentially by taking the state estimates (with filled-in data) of the previous model in Table 1 and applying it to z_t of the current model. Considering the cross-covariance, periods of record, and the distance from the state site, several alternative models with different combinations of covariates were tried from which an optimal model was selected for each case by maximizing the log-likelihood function. In particular, to obtain a complete covariate data set for Model-I, Model-II without the z_t term was initially used to estimate the missing data in $x_t = \{\text{ENR, OKEEFS, S-140}\}$.

The order q in Eq. (1) was determined using Akaike information criteria (AIC) (Shumway, 1988) which chooses the model order q that minimizes:

$$\mathrm{AIC}(q) = \ln\left(\sum_{t=1}^{T} w_t' w_t / T\right) + 2nx^2 q/T. \tag{24}$$

Based on the AIC statistics, $q = 1$ was dominant in all five models. For example, the computed AIC's for Model-II with q ranging from one to three are $\mathrm{AIC}(1) = -3714$; $\mathrm{AIC}(2) = -3324$; and $\mathrm{AIC}(3) = -3449$, from which $q = 1$ was selected.

3.2. Parameter estimation

After setting up the measurement matrix in each model based on the availability of data, the parameters of the five models were sequentially estimated using the EM algorithm. Since the distribution of the data before estimating values for gaps were positively skewed (with skewness coeffi-

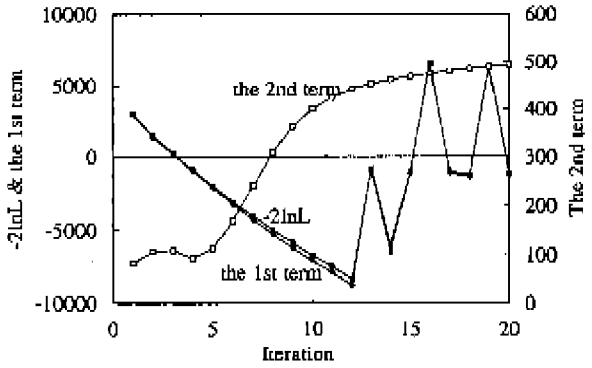
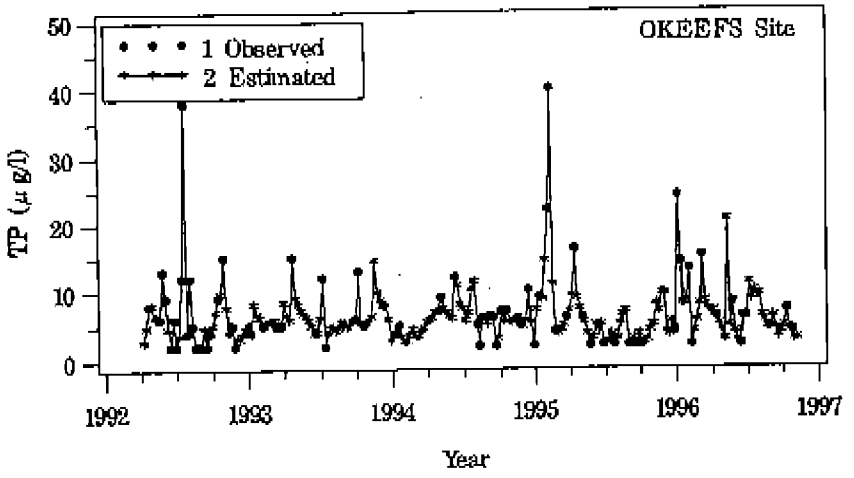
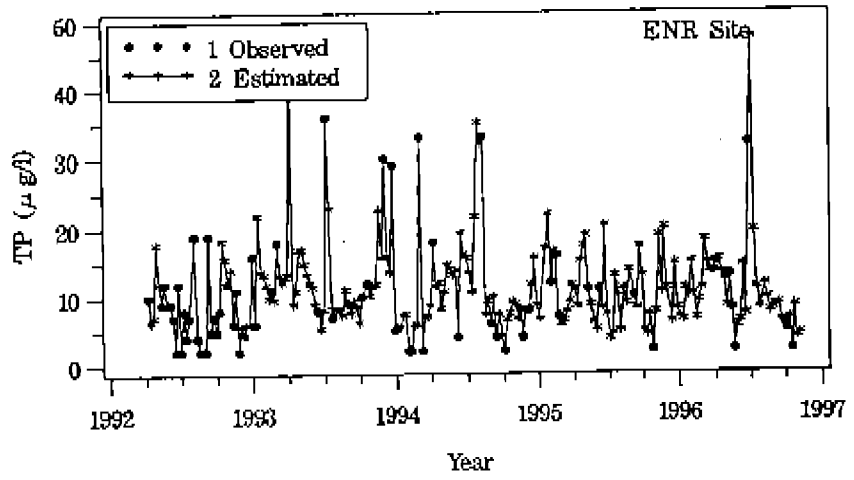


Fig. 3. Convergence of the log-likelihood function (■) for Model-I by the expectation–maximization algorithm, with those for the first (+) and the second (□) terms.





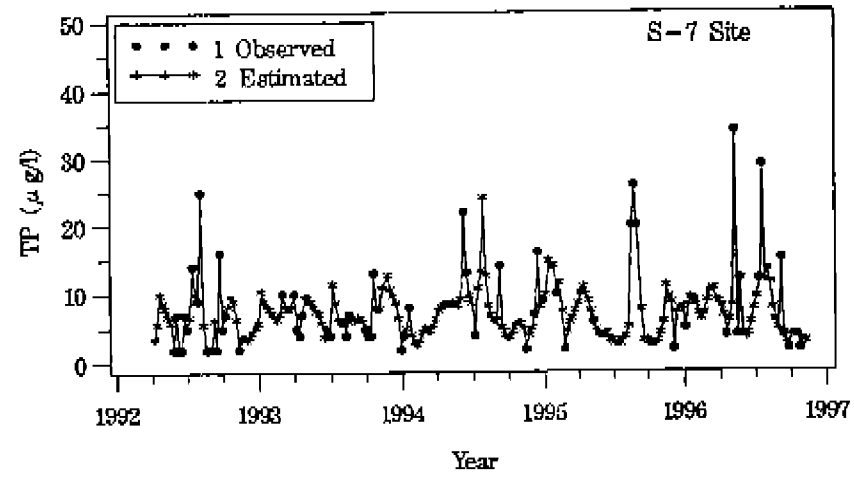


Fig. 4. Time-series of weekly wet total phosphorus concentration data after filling in gaps.

cients ranging from 0.34 to 1.45), the data were log-transformed before the modeling was applied. For instance, for Model-II (where $x_t' =$ [ENR, OKEEFS, S-140] and $z_t' =$ [S-65A, S-7] are the log-transformed wet TP concentrations in μg/l), the calibrated model is given by:

$$\begin{bmatrix} x_{1,t} \\ x_{2,t} \\ x_{3,t} \end{bmatrix} = \begin{bmatrix} 0.09 & 0.02 & 0.02 \\ 0.02 & 0.03 & -0.03 \\ 0.04 & 0.04 & -0.06 \end{bmatrix} \begin{bmatrix} x_{1,t-1} \\ x_{2,t-2} \\ x_{3,t-1} \end{bmatrix} + \begin{bmatrix} 0.30 & 0.49 \\ 0.26 & 0.59 \\ 0.12 & 0.68 \end{bmatrix} \begin{bmatrix} z_{1,t} \\ z_{2,t} \end{bmatrix} + \begin{bmatrix} w_{1,t} \\ w_{2,t} \\ w_{3,t} \end{bmatrix} \quad (25)$$

with the diagonal terms in Q and R matrices being [0.49, 0.67, 0.30] and [0.0021, 0.0035, 0.0029], respectively. As shown by Eq. (25) and by the results of the other four models (which are not presented here), the regression coefficients for z_t are higher than those of x_t. That is, the inter-site correlation of concurrently measured TP values are higher than the time-lagged correlation of the data.

An interesting observation made during parameter estimation by the EM algorithm was that the values of the ML function diverged after certain convergence was achieved (Fig. 3). That is, the ML function decreased constantly at the initial iterations, after which it began to oscillate with the amplitude of oscillation increasing dramatically. The optimal parameter set in each model was obtained at the minimum ML value. Smoothing estimates, x_t^T and p_t^T, were considered optimal at this minimum ML. As shown in Fig. 3 and the other cases which are not presented here, the second term in Eq. (23) which represents measurement error covariance is not significant to the overall ML function. It was also observed during the parameter calibration that the larger the size of missing data, the faster the divergence comes. As a result, the model for a small state dimension (probably 2–4 sites) gives more accurate estimates for missing data than a larger one. This fact also justified development of five separate models instead of one lumped model. The initial parameter set $\{\Phi(0), Q(0), R(0)\}$ was not sensitive to the final estimation result, which was considered to be another advantage of the EM algorithm as a parameter estimation method for a time-series model.

4. Summary statistics and trends

After filling in the data gaps with estimated values, the summary statistics for each site were computed and compared. That is, the final data consisted of direct observations if they were available and smoothing estimates given by Eq. (15) if they were missing (Fig. 4). Plots in Fig. 5 compare some statistics of the data before and after estimation, where the censoring ratio is the probability of the data being < below detection limit (BDL) of 3.5 μg/l. R^2's of the censoring

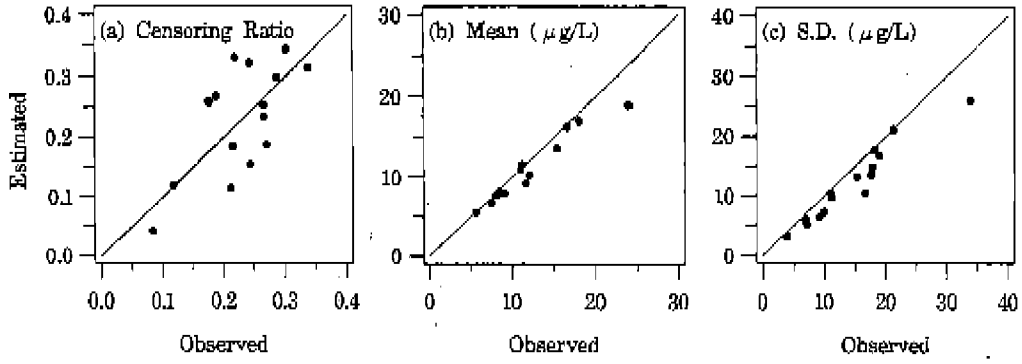


Fig. 5. Comparison of rainfall total phosphorus concentration values before and after filling-in missing data for (a) censoring ratio; (b) mean; and (c) standard deviation.

Table 2
Summary statistics for wet total phosphorus concentration (μg/l) data after filling in gaps

Site Name	Sample size	BDL mean	Mean	S.D.	Estimated Error
BG1	134	2.24	8.25	9.66	1.69
BG2	134	2.28	10.96	13.20	1.75
ENPRC	143	2.33	7.72	10.49	2.67
ENR	206	2.76	10.35	10.49	1.89
G-36	38	1.97	16.26	21.16	1.76
L-67A	43	2.73	5.52	3.17	1.37
L-6	42	2.59	7.77	6.02	1.49
OKEEFS	204	2.70	6.78	5.17	1.67
S-127	132	2.22	19.02	25.98	2.53
S-131	132	2.37	10.75	17.64	2.11
S-140	197	2.64	8.00	7.33	1.47
S-308	139	2.65	17.04	14.87	2.44
S-310	137	2.38	9.31	13.50	1.97
S-65A	196	2.64	13.07	16.64	2.13
S-7	191	2.68	8.00	6.37	1.92

ratio, mean, and standard deviation for the data before and after filling-in are 0.70, 0.93, 0.92, respectively. These comparisons demonstrate that both the censoring ratios and means were preserved in average sense (unbiased) after gaps in the data were filled in; however the variance at each site was slightly lower than that of the original data. This underestimation was mainly caused by the increased sample size of the data. Unlike other sites, the mean and variance of the data from S-127 site (the right-most dot at each plot) were quite underestimated because of the presence of unusually high TP concentration values in the data set.

Table 2 summarizes the statistics of the data after filling in data gaps. The mean and standard deviation for each site, as well as the BDL means, were computed by the censored statistical method (Ahn, 1998) because the data were censored. Especially, the estimated BDL means can be useful for computing weekly TP loads based on the wet TP concentrations, where all BDL data points could be replaced by the BDL mean to get unbiased load estimates. From this table, the pooled mean and standard deviation for the 15 sites after filling-in data gaps are 10.6 and 12.1 μg/l, respectively, while an average estimation error (square root of the smoothed error covariance) of missing portion is ≈1.9 μg/l. The mean TP concentrations in rainfall were very low in the water conservation areas (WCA 1, WCA 2, and WCA 3) and increased slightly from the southern rim of Lake Okeechobee to the north.

Plots in Fig. 6 show the monthly average time-series of the TP data after filling in missing data at three arbitrarily selected sites, along with a linear trend line and a 6-month moving average series. The linear trend line in each plot shows that there is no temporal trend in the data during the period of record, while the 6-month moving average fluctuates due to abnormal high TP concentrations that appear randomly in time. The other sites have the same patterns but are not presented here. To investigate the seasonality in the data, the monthly TP concentration values from all 15 sites were pooled, and the statistics for each month of the year were computed (Fig. 7). This analysis confirms that the month-to-month variation of the data is very weak, almost negligible, compared to the estimation error (the last column in Table 2).

5. Summary

Since the rainfall phosphorus concentration data sets in South Florida have numerous data

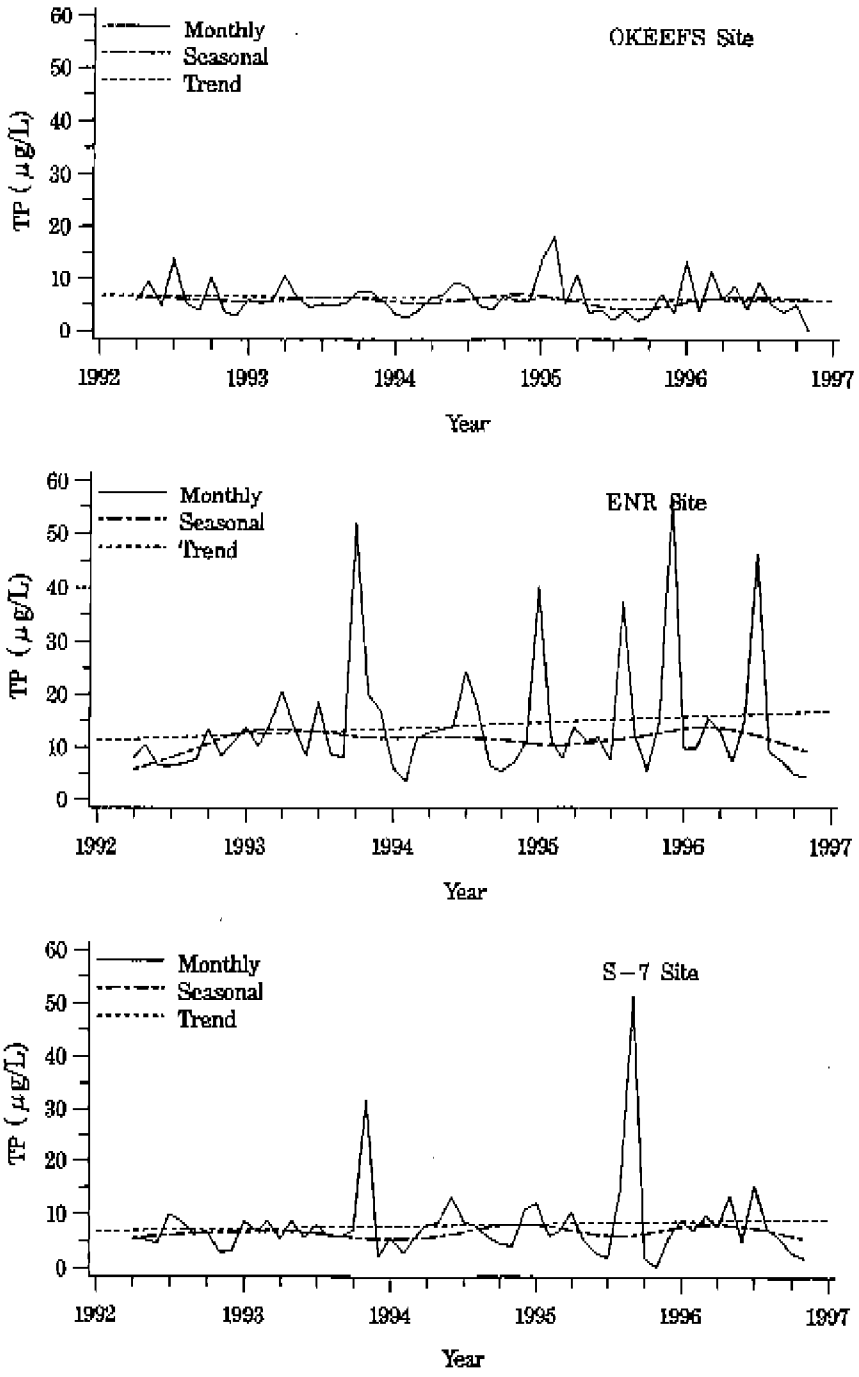


Fig. 6. Seasonal and yearly trends of the wet total phosphorus data for three selected sites.

gaps as the result of sample contamination, an attempt was made to estimate values for the missing information with a statistical model. Five multivariate time-series models were developed from historical data collected from 15 monitoring sites. The model parameters and the missing data

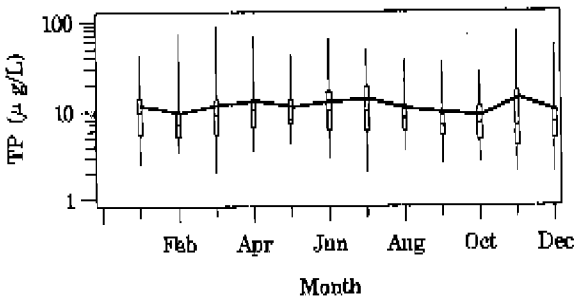


Fig. 7. Box and whisker plots of the monthly total phosphorus concentrations from 15 sites in South Florida. The solid line represents the monthly means of 15 sites, while the middle, bottom, and top edges of each box are the median, 25, and 75% percentiles, and the bottom and top of whiskers are the low and high extremes, respectively.

were estimated simultaneously by an expectation–maximization algorithm. In order to compute the expectation step, the time-series model was set into a state–space form and the Kalman filtering and smoothing algorithms were applied.

As a verification of the model, the statistics of the data after filling-in the gaps were computed and compared with those for the original data set. The results were quite satisfactory in that the censoring ratio and mean of the data (after filling-in gaps) were not biased. However, the variance was slightly underestimated compared to that of the original data. The average concentration ($\mu \pm \sigma$) of wet TP data collected from the 15 sites was estimated to be 10.6 ± 12.1 μg/l, with an average estimation error of 1.9 ($1.6 \sim 3.7$) μg/l. There is neither a temporal trend nor a seasonality in the wet TP concentration data. Instead, random noise in the data appears to be the main cause of long-term irregular fluctuations in the data. In general, the inter-site correlation of the data is stronger than temporal correlation.

Undoubtedly, the TP concentrations resulting from applying this methodology to estimate missing data can be useful for calculating the weekly TP load input from the atmosphere. Alternatively, the load could be calculated for a longer time interval (monthly or yearly), but it would be less accurate than weekly since the spatial and temporal variability of the weekly rainfall is very significant.

Acknowledgements

The author is grateful to Cheol Mo and Maria Loucraft-Manzano for discussions early in the work; Garth Redfield, Thomas James, Susan Gray, and Linda Lindstrom for constructive comments on the draft manuscript; and referees for their constructive comments.

References

Ahn, H., 1997. Outlier detection in total phosphorus concentration data from South Florida rainfall. Technical Publication. WRE-352. South Florida Water Management District, WPB, Florida, pp. 1–35.

Ahn, H., 1998. Estimating the mean and variance of censored phosphorus concentrations in Florida rainfall. J. Am. Water Resour. Assoc. 34 (3), 583–593.

Boudjema, G., Chau, N.P., 1996. Revealing dynamics of ecological systems from natural recordings. Ecol. Model. 91, 15–23.

Box, G.E.P., Jenkins, G.M., 1976. Time Series Analysis: Forecasting and Control. Prentice Hall, Englewood Cliffs, NJ, p. 575.

Brockwell, P.J., Davis, R.A., 1987. Time Series Theory and Methods. Springer-Verlag, New York, p. 519.

Chen, K.W., Papadopoulos, A.S., 1988. A non-parametric method for estimating the joint probability density of BOD and DO. Ecol. Model. 41, 183–191.

Dempster, A.P., Laird, N.M., Rubin, D.B., 1977. Maximum likelihood from incomplete data via the EM algorithm. J. Rol. Stat. Soc. Ser. B39, 1–38.

Gupta, N.K., Mehra, R.K., 1974. Computational aspects of maximum likelihood estimation and reduction in sensitivity function calculation. IEEE Trans. on Automatic Contl., AC-19: 774–783.

Harvey, A.C., 1990. The Econometric Analysis of Time Series. 2nd ed., MIT, Cambridge, MA, p. 387.

Jazwinski, A.H., 1970. Stochastic Processes and Filtering Theory. Academic Press, New York, p. 376.

Padgett, W.J., Papadopoulos, A.S., 1979. Stochastic models for prediction of BOD and DO in streams. Ecol. Model. 6, 289–303.

Shumway, R.H., 1988. Applied Statistical Time Series Analysis. Prentice Hall, Englewood Cliffs, NJ, p. 379.

Shumway, R.H., Stoffer, D.S., 1981. Time series smoothing and forecasting using the EM algorithm, Technical Report No. 27, Division of Statistics, University of California, Davis, CA, pp. 22.

Shumway, R.H., Stoffer, D.S., 1982. An application of time series smoothing and forecasting using the EM algorithm. J. Time Series Anal. 3, 253–264.

Stoffer, D.S., 1985. Maximum likelihood fitting of STARMAX models to incomplete space–time series data. In: Anderson, O.D., Ord, J.K., Robinson, E.A. (Eds.), Time Series Analysis: Theory and Practice. North Holland, Amsterdam, pp. 283–296.

Stoffer, D.S., 1986. Estimation and identification of space–time ARMAX models in the presence of missing data. J. Am. Stat. Assoc. 81, 762–772.

Tiwari, R.C., Dienes, T.P., 1994. The Kalman filter model and Bayesian outlier detection for time series analysis of BOD data. Ecol. Model. 73, 159–165.