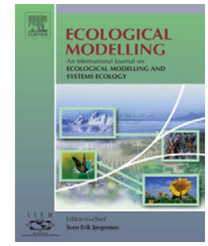
available at www.sciencedirect.com

ScienceDirect

journal homepage: www.elsevier.com/locate/ecolmodel



Modelling algal blooms using vector autoregressive model with exogenous variables and long memory filter

Gilbert C.S. Lui[a,*], *W.K. Li*[a], *Kenneth M.Y. Leung*[b], *Joseph H.W. Lee*[c], *A.W. Jayawardena*[c]

[a] *Department of Statistics & Actuarial Science, The University of Hong Kong, Pokfulam, Hong Kong, China*
[b] *Department of Ecology & Biodiversity, The University of Hong Kong, Pokfulam, Hong Kong, China*
[c] *Department of Civil Engineering, The University of Hong Kong, Pokfulam, Hong Kong, China*

ARTICLE INFO

Article history:
Received 9 August 2005
Received in revised form
16 June 2006
Accepted 20 June 2006
Published on line 13 October 2006

Keywords:
Red-tide
Algal blooms
VARX modelling
Time series forecasting
Long range dependence
Early warning system

ABSTRACT

Algal blooms (ABs), which commonly occur in urbanised coastal marine environments worldwide, often result in hypoxia and even fish kills. Understanding the mechanism and providing accurate prediction of ABs' formation and occurrence is of foremost importance in relation to the protection of sensitive marine resources. In this paper, a multivariate time series model, namely the vector autoregressive model with exogenous variables (VARX) and the long memory filter is proposed to model and predict ABs. To evaluate the effectiveness of this VARX model, both daily and 2-h field monitoring data of chlorophyll fluorescence (CHL), dissolved oxygen (DO), total inorganic nitrogen (TIN), water temperature (TEMP), solar radiation (SR) and wind speed (WS) obtained at Kat O, Hong Kong, between February 2000 and March 2003 were employed. Unlike the other data driven approaches, this VARX model not only provides more interpretable effects of specific lags of environmental factors, but also sheds light on the feedback effects of AB on these variables. In general, daily CHL measurements up to 4 days can provide crucial information for predicting algal dynamics, while the VARX model is able to explicitly reveal ecological relationships between CHL and other environmental factors. In addition, the application of long-memory filter can further extract patterns of seasonal variations which is thought to be correspondent to the variation of algal species in Hong Kong water. With a view to providing an early warning signal of AB to fishermen and regulatory authorities, an alarming system was developed based on the VARX model; it could achieve 83% correct prediction of AB occurrences with a lead time of 2.5 days. Concerning the forecast performance of the VARX model, daily forecasting performance is comparatively better than that of artificial neural network models.

1. Introduction

Algal blooms (ABs), which are caused by rapid growth and accumulation of microscopic phytoplankton, have been increasingly observed in coastal waters throughout the world (Anderson, 1997; Wong, 2003; Kirkpatrick et al., 2004; Qi et al., 2004). As many of these algal species contain red or brown pigments, such ABs are commonly termed as 'red-tide' (Anderson, 1994). Algal blooms not only can result in severe depletion of dissolved oxygen (DO), but also can kill fish and other marine organisms (e.g. corals) through DO depletion (i.e. anoxia) and poisonous toxins generated from

* *Corresponding author.*
E-mail address: csglui@hkusua.hku.hk (G.C.S. Lui).
0304-3800/$ – see front matter
doi:10.1016/j.ecolmodel.2006.06.017

toxic algal species (Anderson, 1994; Lu and Hodgkiss, 2004). In 1998, for example, a devastating red-tide event led to the massive fish kills in both open-sea-cage farms and open waters in Hong Kong, and resulted in an estimated loss of US$ 40 million (equivalent to 3400 tonnes of live fish) (Yang and Hodgkiss, 2004). With a view to minimizing the impacts of ABs on coastal marine resources, an early warning system for accurate and robust prediction of AB's occurrence is highly desirable, especially for fish farmers to have enough time to implement any preventive counter measures (e.g. artificial aeration, relocation of the fish rafts).

The forecast of AB remains a very difficult task, as the bloom can occur and subside over rather short time scales (days to few weeks) under the right environmental condition such as favourable temperature, solar radiation, nutrient concentration and turbulence (linked to wind speed and tidal movement) (Lee et al., 2003; Basterretzea et al., 2005). Over the last two decades, a number of predictive models have been developed to forecast the dynamics of ABs in the natural environment (e.g. Ouchi, 1982; Schofield et al., 1999; Perez et al., 2000; Walsh et al., 2001; Baird et al., 2004). Many of these models are, nonetheless primarily based on some well-established biological theories and physical processes in relation to algal growth (Donaghay and Osborn, 1997), resulting in a dynamical system which is often deterministic in nature. However, the natural environment is usually subject to a lot of noise and the conditions required by a deterministic model are often violated. For example, the open boundary conditions, pollution loading, and ecological rate parameters required to drive the deterministic models are rarely known with a high degree of certainty. A statistical model is therefore worthy of consideration because the data will be allowed to speak for themselves without any constraints imposed by related theories.

Previous works in this direction include Lee et al. (2003, 2004) and Muttil et al. (2004). In the first and second papers, artificial neural networks are used to model a dataset collected in Hong Kong consisting of the algal biomass data as measured by chlorophyll-*a* concentration, i.e. chlorophyll fluorescence (CHL), dissolved oxygen (DO), water temperature (TEMP), solar radiation (SR), wind speed (WS), total inorganic nitrogen (TIN), phosphorus, secchi-disc depth, rainfall and tidal range (Lee et al., 2003, 2004). In the third paper, real-time prediction of the same dataset is considered using genetic programming (Muttil et al., 2004) (see also Muttil and Lee, 2005). While genetic programming and artificial neural network can often come up with reasonable forecasts, the interpretation of the resulting model is often difficult because of the black-box type approach of the methods. In this paper, we considered the modelling of CHL data using a multivariate time series (VARX) approach. The advantage of such an approach is that the relevant estimates of the model parameters are often interpretable and may give insights on the actual algal dynamics process. Unlike the other data driven approaches, the VARX approach not only provides more interpretable effects of specific lags of environmental factors, but also sheds light on the feedback effects of AB on these variables. Coupled with the application of long-memory filter the VARX model can further summarize the algal dynamics during seasonal transitional periods, indicating a clearer pattern of seasonal variations which is correspondent to the observed variation of algal species in Hong Kong waters. Furthermore, when used together with process-based dynamical system model, the time series model should provide information that could be assimilated into the theoretical model.

As a well-known phenomenon in the forecasting literature, a multivariate time series model may not necessarily outperform a univariate time series model in out-of-sample forecasting performances, for example, Riise and Tjøstheim (1984) and Granger and Newbold (1986). A multivariate model usually involves more variables and hence more parameters need to be estimated. More parameter estimates imply more random variations and there could also be problems of measurement errors inherent in some of the ecological data. All these may affect the out-of-sample forecasting performance of a multivariate time series model. However, as in the econometric literature such a model may be viewed as a vehicle for explaining the relationships or testing certain theories about the variables considered. As pointed out by Granger and Newbold (1986) these models are useful in assessing the potential consequence of various policy alternatives or in simulating the behaviour of the dependent variable assuming different realized values of the explanatory variables. In ecology, policy alternatives also play an important part in the future state of an ecological system and a multivariate time series model should be useful in studying their implications. In this paper, the emphasis is on the successful application of multivariate time series models on observations of various frequencies (daily and 2-hourly), with a view to understanding better the underlying the dynamics of ABs and the other ecological variables.

2. The VARX model

The vector autoregressive model with exogenous variables (VARX) allows the variables in the dataset to be modelled jointly over present and past time periods. Let $\mathbf{y}_t = (y_{1t}, y_{2t}, \ldots, y_{kt})'$ be a vector of time series variables such as CHL, DO and TIN and $\mathbf{x}_t = (x_{1t}, x_{2t}, \ldots, x_{rt})'$ be a vector of exogenous variables such as water temperature (TEMP), solar radiation (SR), wind speed (WS) in the present study. Here, we use $'$ to denote the transpose of a vector or matrix. Hence, $\mathbf{y}_t$ and $\mathbf{x}_t$ are $k \times 1$ and $r \times 1$ column vectors respectively. Let Φ_i and Θ_i^* be $k \times k$ and $k \times r$ matrix coefficients. The VARX(p, s) model (Shumway and Stoffer, 2000) is defined as

$$\mathbf{y}_t = \sum_{i=1}^{p} \Phi_i \mathbf{y}_{t-i} + \sum_{i=0}^{s} \Theta_i^* \mathbf{x}_{t-i} + \boldsymbol{\epsilon}_t,$$

where the unobserved noise variable, $\boldsymbol{\epsilon}_t = (\epsilon_{1t}, \epsilon_{2t}, \ldots, \epsilon_{kt})'$, represents a vector noise process that has zero mean and independent over time t. The variables $\mathbf{y}_t$ are referred to as the dependent, response, or endogenous variables, and the variables $\mathbf{x}_t$ are referred to as the independent, input, predictor regressor, or exogenous variables.

The VARX(p, s) model can be written alternatively,

$$\Phi(B)\mathbf{y}_t = \Theta^*(B)\mathbf{x}_t + \boldsymbol{\epsilon}_t,$$

where

$$\Phi(B) = I_k - \Phi_1 B - \cdots - \Phi_p B^p,$$

$$\Theta^*(B) = \Theta_0^* + \Theta_1^* B + \cdots + \Theta_s^* B^s,$$

and B is the backshift operator, $B^i \mathbf{y}_t = \mathbf{y}_{t-i}$. Let $E(\cdot)$ be the expectation operator.

The following assumptions are made:

- $E(\boldsymbol{\epsilon}_t) = 0$, $E(\boldsymbol{\epsilon}_t \boldsymbol{\epsilon}_t') = \boldsymbol{\Sigma}$, which is a positive-definite, $k \times k$ matrix, and $E(\boldsymbol{\epsilon}_t \boldsymbol{\epsilon}_s') = 0$ for $t \neq s$. It is also assumed that $\boldsymbol{\epsilon}_t$ is multivariate normally distributed.
- For stationarity of the VARX process, the roots of $|\Phi(z)| = 0$ are outside the unit circle, where $|\cdot|$ is the determinant function.
- The explanatory or exogenous (independent) variables $\mathbf{x}_t$ are not correlated with residuals $\boldsymbol{\epsilon}_t$, $E(\mathbf{x}_t \boldsymbol{\epsilon}_t') = 0$. In many situations, future values of the exogenous variables can be assumed to be known or predicted fairly accurately because they are deterministic variables.

Various model selection criteria can be used to choose the appropriate model. The following list includes the Akaike Information Criterion (AIC), and the Schwarz Bayesian Criterion (SBC), also referred to as BIC.

$$\text{AIC} = \log(|\tilde{\boldsymbol{\Sigma}}|) + 2L/T,$$
$$\text{SBC} = \log(|\tilde{\boldsymbol{\Sigma}}|) + L\log(T)/T$$

where L denotes the number of parameters estimated, T the sample size and $\tilde{\boldsymbol{\Sigma}}$ is the maximum estimate of $\boldsymbol{\Sigma}$. Often the model with the smallest AIC or SBC value is chosen as the final model. Model diagnostic checks (Li, 2004) can also be applied to further refine the model.

3. The data and modelling results

The data used in this study were collected from February 2000 to March 2003 from the algal bloom (AB) dynamics field monitoring station of the University of Hong Kong at Kat O, Hong Kong. This was the same dataset used in Lee et al. (2004) and Muttil et al. (2004). Both daily and 2-h data were considered. Readers are referred to these two papers for more details about the data. Based on the autocorrelation plot of the CHL data the maximum order of the VARX model was set to 21. For valid statistical inference, the data had to be stationary. Consequently, all variables had been first order differenced, i.e. instead of the original variable y_{it} or x_{it} the variables $\tilde{y}_{it} = y_{it} - y_{it-1}$ and $\tilde{x}_{it} = x_{it} - x_{it-1}$, were used in the model. Here $\mathbf{y}_t = (\text{CHL}_t, \text{DO}_t, \text{TIN}_t)'$ and $\mathbf{x}_t = (\text{TEMP}_t, \text{SR}_t, \text{WS}_t)'$. To take care of the diurnal effect, in the case of 2-h data, $\tilde{x}_{it} = x_{it} - x_{it-12}$ for $i = 1, 2$. In order to obtain a more parsimonious and interpretable model the series $\tilde{y}_{1t}$ was further differenced using the idea of fractional differencing (Hosking, 1981; Li, 2004) which was a common way to handle long range dependence in time series. Essentially, long range dependence with time series $\tilde{y}_{1t}$ was removed by passing it through the filter,

$$\tilde{\tilde{y}}_{1t} = \sum_{j=0}^{M} \Pi_j \tilde{y}_{1t-j} = \tilde{y}_{1t} - d\tilde{y}_{1t-1} - \frac{1}{2}d(1-d)\tilde{y}_{1t-2}$$
$$- \frac{1}{6}d(1-d)(2-d)\tilde{y}_{1t-3} - \cdots$$
$$- \frac{1}{M!}d(1-d)\ldots(M-d-1)\tilde{y}_{1t-M},$$

where $-\frac{1}{2} < d < \frac{1}{2}$.

The value of d could be chosen by minimizing the total distance between $\tilde{y}_{1t}$ and $\tilde{\tilde{y}}_{1t}$, that is, minimizing $\sum_{t=M+1}^{T}(\tilde{y}_{1t} - \tilde{\tilde{y}}_{1t})^2$ over d. In the present study, M was chosen to be 40 and the value of d was found to be between -0.15 and -0.30 in all cases. Since ordinary differencing ($d = 1$) had already been applied once to the raw CHL data this implied that the overall amount of differencing was around 0.70–0.85.

Two training periods were used resulting in two slightly different VARX models. The first training period was to use all data available in 2000 for the fitting of the model while keeping the data in 2001 and 2002 for checking the accuracy of the out-of-sample forecast. This was also one way to gauge the correctness of the fitted model. The second period was to use all data in the years 2000 and 2001 for model fitting and the data in 2002 for checking the out-of-sample forecast performances. Only coefficients that were statistically significant at the 5% level were retained in the model.

3.1. Analysis of daily data

The whole VARX fitting process was then applied to daily data and the final result for year 2000–2001 data is shown in Table 1. We can see that the high order of lags of CHL did not appear in the equations of CHL_t and DO_t. Note that wind speed had a very poor predictive power and was thus removed from the final model. However, the nutrient (TIN) had a significant positive coefficient at lag 1 in the equation of chlorophyll and the DO had a significant negative coefficient at lag 8. The CHL series was also highly autocorrelated with significant effect up to 4 days. The results in Table 1 were consistent with that of Lee et al. (2004) where a significant lag of 7 days for DO to affect CHL was also observed. The equations of DO and TIN also appeared to reflect the algal dynamics reasonably. For example, DO was negatively related to CHL, suggesting algal respiration and decomposition of settled algae on the sea bed. Conversely, CHL was also negatively related to TIN at a lag of 2 days, suggesting nutrient uptake for algal growth. Both water temperature and solar radiation had not been entered into the CHL equation because their effects were too insignificant.

3.2. Analysis of 2-h data

The above results are compared with those obtained at the higher sampling frequency of 2-h time intervals. As in the case of daily data, the use of the long memory filter resulted in a simplification of the VARX models and Table 2 shows the final result using the year 2000–2001 data. The dynamics of the AB as suggested by the VARX models appear to be more detailed and complicated. In particular, the effects of TIN were initially

Table 1 – Estimation of daily data (2000–2001) by VARX model, fractional differencing $d = 0.72$

Dependent variable	Coefficient estimate	Standard error	t-Value	p-Value
CHL_t				
CHL_{t-1}	0.1145	0.0381	3.01	0.0027
CHL_{t-4}	−0.0765	0.0389	−1.97	0.0498
DO_{t-8}	−36.6896	10.7995	−3.40	0.0007
TIN_{t-1}	1.7478	0.3337	5.24	0.0001
DO_t				
CHL_{t-4}	−0.0004	0.0001	−3.07	0.0022
CHL_{t-8}	−0.0003	0.0001	−2.05	0.0406
DO_{t-4}	−0.0819	0.0386	−2.12	0.0342
DO_{t-17}	0.1036	0.0371	2.79	0.0054
$TEMP_t$	0.3803	0.0717	5.31	0.0001
$TEMP_{t-1}$	−0.3802	0.0717	−5.31	0.0001
TIN_{t-2}	−0.0026	0.0011	−2.30	0.0215
TIN_t				
CHL_{t-2}	−0.0084	0.0041	−2.07	0.0391
DO_{t-20}	4.1815	1.1707	3.57	0.0004
TIN_{t-4}	−0.0609	0.0358	−1.70	0.0892
TIN_{t-5}	−0.1022	0.0357	−2.86	0.0044
TIN_{t-6}	−0.0994	0.0357	−2.78	0.0056
TIN_{t-21}	0.3588	0.0357	10.06	0.0001

Note: CHL_t is fractionally differenced and the other variables are first differenced.

negative but became positive at lags 7 and 11, i.e. after 14 and 22 h respectively. Similarly, the effects of DO were positive initially (i.e. at smaller lags) but became negative at lags 8 and 11 in Table 2. This may be partly due to the diurnal migration movements of dinoflagellates from sea bed (where nutrients are available) to the surface (to fetch energy), and partly due to the interaction between diatoms and dinoflagellates which are often found to coexist in an AB (Flynn et al., 2002). Moreover, increased algal growth would increase DO production during photosynthesis during the day (with solar radiation) but would increase oxygen demand for their respiration during night-time. However, massive algal bloom would dramatically reduce the DO level eventually. We therefore study the day and night effects which are reported in Section 3.4 using the VARX model. The precise biological interaction is clearly a matter for future research.

From Table 1, we can also observe that TEMP did not seem to have a direct effect on CHL given the past values of CHL, DO and TIN. This occurs not just in daily data but also in 2-h data as well. Although TEMP appeared in the equation of CHL_t for 2-h data, the total effect, i.e. the sum of coefficients of TEMP, was approximately zero. For the equation of DO_t, the same phenomenon is also observed. In other words, the water temperature did not play an important role in the VARX models. This was explained by Lee et al. (2003) where the change in water temperature in coastal waters over the time scale of short term forecasting (say a week) was typically small; hence it may not be an important input variable in an artificial neural network model or any other data driven forecast model. In contrast, SR had a more direct and positive effect on CHL for the 2-h results. When we investigated the equation of DO_t, the effect of SR was also transmitted by DO positively on CHL as well. Nevertheless, it is important to sound a note that tidal movement could also be associated with mixing events as well as the advection of relatively clean water from the outer bay. In general, neap tide (i.e., small tidal range) will be associated with weaker vertical mixing and flushing rates, and AB can deplete DO quicker in this situation. On the other hand, DO

Table 2 – Estimation of 2-h data (2000–2001) by VARX model, fractional differencing, $d = 0.75$

Dependent variable	Coefficient estimate	Standard error	t-Value	p-Value
CHL_t				
CHL_{t-2}	−0.0669	0.0106	−6.28	0.0001
CHL_{t-6}	−0.0352	0.0107	−3.28	0.0010
CHL_{t-8}	−0.0315	0.0107	−2.94	0.0032
CHL_{t-9}	0.0454	0.0107	4.25	0.0001
CHL_{t-10}	0.0434	0.0106	4.08	0.0001
CHL_{t-11}	0.0352	0.0107	3.29	0.0010
CHL_{t-12}	0.1443	0.0108	13.42	0.0001
CHL_{t-13}	0.0775	0.0107	7.27	0.0001
DO_{t-1}	7.4419	3.5146	2.12	0.0343
DO_{t-4}	5.3384	3.4578	1.54	0.1227
DO_{t-8}	−10.0990	3.4752	−2.91	0.0037
DO_{t-11}	−8.1301	3.4632	−2.35	0.0189
SR_{t-2}	14.3005	1.5527	9.21	0.0001
$TEMP_t$	−33.6291	6.9267	−4.86	0.0001
$TEMP_{t-1}$	33.2057	6.9208	4.80	0.0001
TIN_{t-5}	−16.3310	2.1947	−7.44	0.0001
TIN_{t-6}	−9.1786	2.2097	−4.15	0.0001
TIN_{t-7}	16.1178	2.2169	7.27	0.0001
TIN_{t-10}	−17.0393	2.1956	−7.76	0.0001
TIN_{t-11}	12.4193	2.2031	5.64	0.0001
DO_t				
CHL_{t-1}	0.0001	0.0000	4.41	0.0001
CHL_{t-5}	−0.0001	0.0000	−2.16	0.0307
CHL_{t-6}	0.0001	0.0000	2.36	0.0181
CHL_{t-9}	−0.0001	0.0000	−2.40	0.0163
CHL_{t-12}	0.0001	0.0000	2.41	0.0161
DO_{t-1}	−0.2360	0.0110	−21.38	0.0001
DO_{t-2}	−0.1124	0.0106	−10.65	0.0001
DO_{t-3}	−0.0806	0.0103	−7.84	0.0001
DO_{t-4}	−0.0905	0.0103	−8.76	0.0001
DO_{t-5}	−0.0666	0.0104	−6.42	0.0001
DO_{t-6}	−0.1009	0.0104	−9.67	0.0001
DO_{t-7}	−0.0847	0.0105	−8.07	0.0001
DO_{t-8}	−0.0716	0.0105	−6.85	0.0001
DO_{t-11}	0.0559	0.0105	5.33	0.0001
DO_{t-12}	0.0907	0.0106	8.56	0.0001
DO_{t-13}	0.0932	0.0105	8.91	0.0001
SR_t	0.0335	0.0067	5.02	0.0001
SR_{t-1}	0.0207	0.0068	3.04	0.0024
$TEMP_t$	0.4790	0.0209	22.91	0.0001
$TEMP_{t-1}$	−0.3458	0.0303	−11.43	0.0001
$TEMP_{t-2}$	−0.1348	0.0214	−6.29	0.0001
TIN_{t-6}	0.0140	0.0065	2.16	0.0309
TIN_{t-7}	0.0221	0.0065	3.42	0.0006
TIN_t				
CHL_{t-2}	0.0002	0.0001	3.03	0.0024
CHL_{t-8}	−0.0002	0.0001	−3.63	0.0003
DO_{t-2}	0.0364	0.0172	2.12	0.0343
TIN_{t-12}	0.0248	0.0111	2.23	0.0261

Note: CHL_t is fractionally differenced. DO_t and TIN_t are first-differenced. $TEMP_t$ and SR_t are seasonally differenced by 12 periods.

level can remain high during spring tide (i.e., large tidal range) due to the above hydrographic factors (Lee et al., 2003).

3.3. Analysis of transitional periods data

Modelling the periods of the year that had the largest amount of harmful ABs, in particular dinoflagellate blooms is considered. These periods were identified to be (i) the Summer–Winter transitional period from 15 October each year to 15 January next year and (ii) the Winter–Summer transitional period from 15 March each year to 15 June the same year (Wong, 1989; Yang, 2000; Yang and Hodgkiss, 2004). Such high incidences of ABs in these two seasonal transitional periods are likely attributable to the unstable features of weather pattern in Hong Kong (Yang and Hodgkiss, 2004). We focused on the 2-h data. The results for the Summer–Winter and Winter–Summer transitional periods for the year 2000–2001 data are summarized in Table 3. For simplicity, we just listed the equation for chlorophyll which was our main focus. There are some immediate observations. First, the Winter–Summer models for chlorophyll were more complicated than those for the Summer–Winter transitional period involving more lags of the variables. Secondly, the effects of DO and nutrient were different for the two periods. In the Summer–Winter period, the nutrient TIN_t effects were all positive while in the Winter–Summer period the effects of TIN_t were first negative and then turned positive at lag 11. The effects of DO on the chlorophyll were also different. For the Summer–Winter period all the coefficients of DO_t were positive. For the Winter–Summer period with data from the period 2000–2001, the effects of DO_t were positive initially but turned negative after lag 12, i.e. 1 day before. Third, the lagged effect of solar radiation (SR) was positive for both transitional periods. The main difference was just the effect of lag one period for Summer–Winter period and that of lag two period for Winter–Summer period. All these differences may shed some lights on the predominance of different kinds of algal species in these two periods, for example, the observed predominance of dinoflagellate blooms in the Winter–Summer transitional periods (Lu and Hodgkiss, 2004; Yang and Hodgkiss, 2004). Further investigation on the possible causes of the differences appears worthwhile in the future.

Table 3 – Estimation of 2-h data (2000-2001) by VARX model, fractional differencing

Dependent variable	Coefficient estimate	Standard error	*t*-Value	*p*-Value
Summer–Winter period ($d = 0.85$)				
CHL_t				
CHL_{t-1}	−0.0677	0.0217	−3.11	0.0019
CHL_{t-11}	0.0516	0.0204	2.53	0.0116
CHL_{t-12}	0.1430	0.0210	6.79	0.0001
CHL_{t-13}	0.0668	0.0206	3.24	0.0012
DO_{t-1}	39.0510	7.6537	5.10	0.0001
DO_{t-2}	20.7824	7.5377	2.76	0.0059
DO_{t-3}	18.3130	7.2938	2.51	0.0121
SR_{t-1}	11.7800	2.7263	4.32	0.0001
$TEMP_t$	−0.2737	0.1091	−2.51	0.0122
TIN_{t-7}	50.7866	25.8435	1.97	0.0495
TIN_{t-8}	110.7788	25.8897	4.28	0.0001
Winter–Summer period ($d = 0.74$)				
CHL_t				
CHL_{t-6}	−0.0697	0.0162	−4.29	0.0001
CHL_{t-9}	0.0783	0.0160	4.89	0.0001
CHL_{t-10}	0.0741	0.0159	4.65	0.0001
CHL_{t-12}	0.1478	0.0162	9.14	0.0001
CHL_{t-13}	0.0391	0.0160	2.44	0.0148
DO_{t-4}	20.2326	5.6687	3.57	0.0004
DO_{t-5}	11.1726	5.6903	1.96	0.0497
DO_{t-6}	14.8423	5.7486	2.58	0.0099
DO_{t-12}	18.0843	5.7878	3.12	0.0018
DO_{t-13}	−16.9150	5.7912	−2.92	0.0035
SR_{t-2}	16.6640	2.3690	7.03	0.0001
$TEMP_t$	−0.5124	0.1051	−4.88	0.0001
TIN_{t-5}	−18.3375	2.3750	−7.72	0.0001
TIN_{t-6}	−8.8641	2.3821	−3.72	0.0002
TIN_{t-7}	18.9093	2.3829	7.94	0.0001
TIN_{t-10}	−16.8506	2.3745	−7.10	0.0001
TIN_{t-11}	13.3919	2.3961	5.59	0.0001

Note: CHL_t is fractionally differenced. DO_t and TIN_t are first-differenced. $TEMP_t$ and SR_t are seasonally differenced by 12 periods.

3.4. Analysis of day time and night time data

In order to study the possible differences between the day and the night, the 2-h data were remodelled using day time and night time data separately. The day time data were defined to be those from 6 a.m. each day to 6 p.m. of the same day, inclusive and the night time data were those from 6 p.m. each day to 6 a.m. of the next day, exclusive. The data were seasonally differenced using the filter $(1 - B^6)$ and $(1 - B^5)$ respectively. Again, two different time periods of the years 2000 and 2000–2001 were considered. The whole VARX modelling process was repeated. The main results for day time data CHL_t are shown in the upper panel of Table 4. It can be seen that the pattern of the estimated coefficients for the chlorophyll equation was very similar to those reported in Section 3.2 for 2-h data, in particular, (i) that the signs of DO were positive for the nearer (lower) lags and became negative at the further (higher) lags, and (ii) the signs for the coefficients for the TIN variable were negative for nearer (lower) lags and became positive at further lags (higher lags). Intuitively, SR had a positive total effect on CHL in the case of day time data as shown in Table 4.

Interestingly, when the night time data were considered, it was a different story (lower panel of Table 4). The coefficients of DO in the chlorophyll equation were negative at the nearer (lower) lags but they became positive later. For the nutrient data TIN the coefficients were also negative initially at the earlier lags but these became positive later. The separate analysis of night time and day time data seemed to support the conjecture in Section 3.2 about the diurnal migration of dinoflagellates from the sea bed to the surface during the day. The increased photosynthesis in the day seemed able to compensate for its consumption leading to a positive increase in DO and this further increased the CHL_t. However, at night light reaction of photosynthesis stopped and the DO was being consumed without fresh supply. This led to the negative sign of the DO parameters. However, when we combined the night and day data the effect of photosynthesis seemed to have dominated and we saw positive coefficients of DO in the chlorophyll equation, at least, for the lower lags. A similar explanation went for the nutrient variable TIN. In this case

Table 4 – Estimation of 2-h data (2000–2001) by VARX model, fractional differencing

Dependent variable	Coefficient estimate	Standard error	*t*-Value	*p*-Value
Day time period ($d = 0.75$)				
CHL_t				
CHL_{t-1}	−0.0420	0.0144	−2.92	0.0035
CHL_{t-2}	−0.0623	0.0136	−4.58	0.0001
CHL_{t-7}	0.1802	0.0141	12.76	0.0001
CHL_{t-8}	0.1142	0.0139	8.20	0.0001
CHL_{t-13}	0.0390	0.0136	2.87	0.0042
CHL_{t-14}	0.0600	0.0141	4.27	0.0001
DO_{t-1}	15.6160	4.2106	3.71	0.0002
DO_{t-6}	−14.1324	4.0477	−3.49	0.0005
DO_{t-11}	−10.1419	4.0507	−2.50	0.0123
DO_{t-14}	28.9368	4.2014	6.89	0.0001
SR_t	−8.6418	2.2523	−3.84	0.0001
SR_{t-8}	13.9926	2.2380	6.25	0.0001
$TEMP_{t-10}$	−0.2281	0.1212	−1.88	0.0600
TIN_{t-2}	−15.2875	2.3993	−6.37	0.0001
TIN_{t-3}	−11.3470	2.4082	4.71	0.0001
TIN_{t-4}	15.1377	2.4548	6.17	0.0001
TIN_{t-9}	7.4573	2.4073	3.10	0.0020
TIN_{t-10}	10.7069	2.4136	4.44	0.0001
Night time period ($d = 0.78$)				
CHL_t				
CHL_{t-3}	−0.0342	0.0165	−2.07	0.0386
CHL_{t-5}	0.1173	0.0172	6.84	0.0001
DO_{t-1}	−8.1330	4.4114	−1.84	0.0653
DO_{t-5}	20.3493	4.5080	4.51	0.0001
TIN_{t-3}	−22.7369	2.5753	−8.83	0.0001
TIN_{t-4}	13.1623	2.5330	5.20	0.0001

Note: CHL_t is fractionally differenced. DO_t and TIN_t are first-differenced. $TEMP_t$ and SR_t are seasonally differenced by 6 periods in day time period. $TEMP_t$ and SR_t are seasonally differenced by 5 periods in night time period.

the data provided more consistent results, namely the coefficients of TIN were basically negative for lower lags, for both day time and night time data, reflecting that nutrients were basically consumed in the growth of the AB. Not surprisingly, solar radiation and water temperature play an insignificant role in Table 4 due to the low level of solar radiation at night.

Table 5 – In-sample and out-sample RMSEs of forecasting models (Non-fractionally differenced case)

Training period	Variable	RMSE (2002)	RMSE (2001.07–2001.10)
Daily (2000)	CHL-AR	72.7716	175.5023
Daily (2000)	CHL-VARX	75.1641	174.8866
Daily (2000–2001)	CHL-AR	70.9099	172.4904
Daily (2000–2001)	CHL-VARX	71.8470	164.9690
2-h (2000)	CHL-AR	83.2438	65.5900
2-h (2000)	CHL-VARX	84.3869	67.7819
2-h (2000–2001)	CHL-AR	83.8826	64.6789
2-h (2000–2001)	CHL-VARX	84.6627	64.8796

Note: CHL-VARX, VARX model for CHL; CHL-AR, univariate autoregressive model for CHL.

3.5. *Forecasting performances and an alarming system*

The forecasting performances of the VARX models were studied both using the in-sample (training) and out-of-sample (testing) periods. The forecasting performance was measured by the root mean square error, RMSE. The forecasts were 1-step ahead forecasts with values of TEMP and SR assumed known. In order to compare the forecasting results more completely, the estimation results of non-fractionally differenced CHL data for the period 2000–2001 were also considered and they are presented in Appendix A. Due to the limited space, only the results of the equation of CHL_t are shown. The estimation results for the year 2000 are also available upon request.

Since the forecasting performances for fractionally differenced and non-fractionally differenced CHL data were about the same, only the results of non-fractionally differenced CHL data are reported in Table 5 for convenience. From the upper panel of Table 5, the daily results of RMSE values for 2002 using 2 years of training data were somewhat smaller than those using 1 year data only. Intuitively, a 2 year dataset could give more information for estimation and might be useful for prediction as well. In order to compare with the result in Lee et al. (2004) the out-of-sample forecast RMSE for the period 1 July, 2001 to 31 October, 2001 is also reported. For this period Lee et al. (2004) reported a RMSE of 184.00 while the RMSEs for both of our VARX models were somewhat less than 184.00. The numerical results here showed the superiority of our models over the approach of artificial neural network in Lee et al. (2004). As followed from Lee et al. (2004), most significant bloom events were found in 2000 and 2001. Hence, the observations during July 2001 to October 2001 are chosen in Fig. 1 to compare the forecasting performance qualitatively. The model of non-fractionally differenced data seemed to capture the blooms in 2001 July pretty well. During August to October of 2001, the forecasted CHL values were close to the actual observations qualitatively. This result is consistent with the numerical result of Table 5.

The corresponding 2-h data results are reported in the lower panel of Table 5. Note that the RMSE values were much smaller with 2-h data because of smaller observed values of

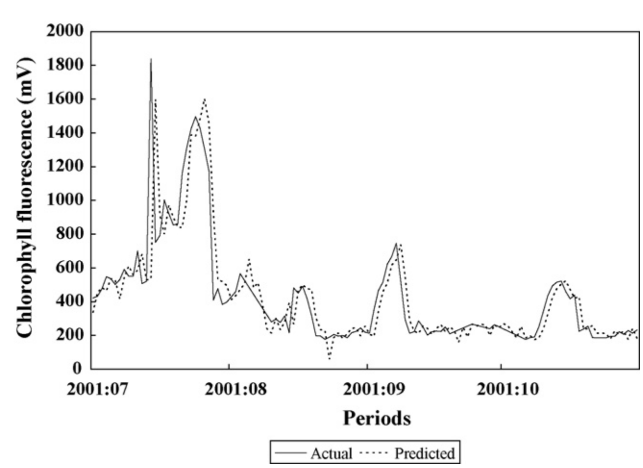


Fig. 1 – Out-of-sample forecasting of daily CHL by VARX model during July, 2001–October, 2001, no fractional differencing.

Table 6 – Results for alarming system of forecasting models

Cases	Variable	Threshold	Average lead time 2002	Correct prediction (%)
Daily naïve case	CHL	400	2.25	75
Daily non-fractional (2000)	CHL-AR	400	2.29	86
Daily non-fractional (2000–2001)	CHL-AR	400	3.43	71
Daily non-fractional (2000)	CHL-VARX	400	2.67	67
Daily non-fractional (2000–2001)	CHL-VARX	400	2.89	67
Daily naïve case	CHL	450	2.00	100
Daily non-fractional (2000)	CHL-AR	450	2.17	83
Daily non-fractional (2000–2001)	CHL-AR	450	3.00	80
Daily non-fractional (2000)	CHL-VARX	450	2.50	83
Daily non-fractional (2000–2001)	CHL-VARX	450	2.33	83

Note: CHL-VARX, VARX model for CHL; CHL-AR, univariate autoregressive model for CHL.

CHL. Again both VARX models had very similar performances and the differences in RMSE values were not significant. The out-of-sample forecasting performance of a high order AR model of lag 34 fitted to the first differenced CHL data is also reported in Table 5. As explained in the Introduction section multivariate model could be inferior in its forecasting performance relative to a univariate model. This seems to be the case here and all the VARX models are somewhat inferior to the AR model in terms of RMSE although the values are very close. The findings here are consistent with those in Lee et al. (2003). Fig. 2 gives the corresponding plots of out-of-sample forecasts for the models based on 2-h data during the same forecasting period.

A warning system for a possible red-tide event was also considered. A red-tide event here was defined to be an observation of CHL value at 500 mV or above. There might not be enough events for testing the warning system if a value higher than 500 mV is chosen. As in all warning systems there was an indicator variable and an alarm was triggered when a certain threshold was crossed by the indicator variable. Here, one-step ahead forecasts from time series models in Section 3 were considered as potential indicator variables. The choice of the threshold value and the indicator was a compromise between rates of correct alarms and the lead-time from an alarm to the actual event. Normally, the frequency of no alarms in the case of an event should also be considered. For the threshold values and the testing period considered here the frequency of no alarm was very small and so this had not been taken into account in this study. Two threshold values, one at 400 mV and the other at 450 mV were considered. The indicator variables were one-step ahead forecasts obtained from the daily VARX models in Section 3 for the periods 2000 and 2000–2001. Only daily data were considered as they are deemed most useful in practice. The results from the VARX model were compared with that of a high order univariate autoregression used at the beginning of this section and a naïve forecast procedure that used the raw data only. The average lead time from sounding of the alarm to the red-tide event given that the event actually occurred was recorded together with the rate of correct prediction. The results of warning system are reported in Table 6.

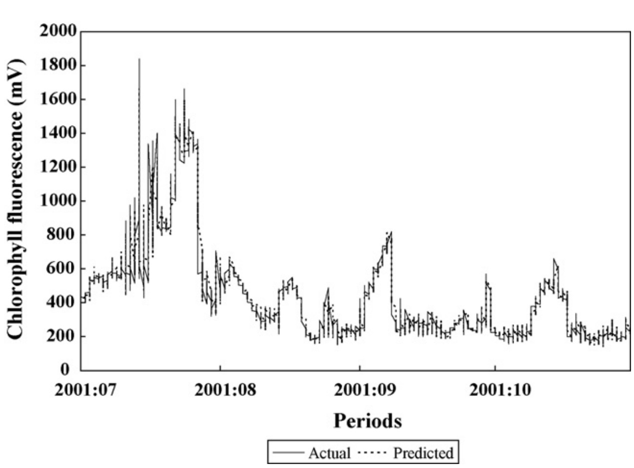


Fig. 2 – Out-of-sample forecasting of 2-h CHL by VARX model during July, 2001–October, 2001, no fractional differencing.

Suppose that the rate of correct prediction was taken as priority over lead time. Then, the champion for the 450 mV case was the VARX model based on the non-fractionally differenced 2000 data. It had an average lead time of 2.5 days from triggering of the alarm to the actual event and an 83% of correct predictions of ABs. Furthermore, the VARX model based on 2000–2001 data provided the same accuracy of the alarm system although the average lead time was shorter than that of the champion by a small amount. As a benchmark, the naïve approach based on the chlorophyll series scored 100% in terms of correct predictions with an average lead time of 2 days. This showed that there is an improvement of earlier warnings by our VARX models over the naïve approach. On the other hand, when the lower threshold value of 400 mV was used, the average lead time was longer than that in the case of higher threshold value. However the percentage of correct predictions of AB was lowered because the warning system became more sensitive to the occurrence of AB. Under this threshold value, the best VARX model had a lead time that was almost 3 days and this might be significant in allowing more time for the fish farms or relevant organizations to take preventive actions. The price, of course, was that there might be a bigger chance of a false alarm. The cost of this might be insignificant when compared with the actual damage due to a harmful AB. Certainly, more future studies in such a warning system will be valuable.

4. Conclusions

The successful application of the multivariate time series models was demonstrated by fitting VARX models to both daily and 2-h chlorophyll data with water temperature and solar radiation as exogenous variables. The VARX models have the advantage that useful information can be obtained for a better understanding of the underlying mechanism of algal bloom (AB) in relation to ecological parameters (e.g. nutrients, temperature, solar radiation and wind speed). Currently, an attempt is being made to include other ecological parameters in the VARX model, such as wind direction that has an implication on turbulence and mixing, and hence the sustainability of the AB.

Daily forecasting performance using RMSE suggested that the VARX models could predict better than the artificial neural network models described in Lee et al. (2004). The estimation results of 2-h observations suggested that the AB dynamics could be simplified by the application of long-memory filter but the forecasting performance could not be improved significantly. Interestingly, this study demonstrated that different modelling strategies for the seasonal transitional periods could lead to better interpretation and understanding of the bloom mechanisms. Finally, we advocate that more future field study and modelling work will be needed in further resolving some of the anomalies observed in the multivariate time series models.

Acknowledgements

The authors would like to thank Professor John Hodgkiss, editor and two anonymous referees for helpful comments. The work described in this paper was supported by the Area of Excellence Scheme under the University Grants Committee of the Hong Kong Special Administration Region, China (Project No. AoE/P-04/2004) and partially supported by Research Grants Council of Hong Kong Special Administration Region, China (Project No. HKU7110/04E).

Appendix A

See Tables A.1 and A.2.

Table A.1 – Estimation of daily data (2000–2001) by VARX model, no fractional differencing

Dependent variable	Coefficient estimate	Standard error	*t*-Value	*p*-Value
CHL_t				
CHL_{t-1}	−0.1613	0.0374	−4.31	0.0001
CHL_{t-2}	−0.0890	0.0382	−2.33	0.0202
CHL_{t-3}	−0.1011	0.0377	−2.68	0.0075
CHL_{t-4}	−0.1379	0.0389	−3.55	0.0004
CHL_{t-5}	−0.1132	0.0389	−2.91	0.0037
CHL_{t-6}	−0.1095	0.0383	−2.86	0.0044
CHL_{t-7}	−0.1078	0.0382	−2.82	0.0049
CHL_{t-8}	−0.1127	0.0396	−2.85	0.0046
CHL_{t-9}	−0.0964	0.0375	−2.57	0.0105
CHL_{t-12}	−0.0606	0.0376	−1.61	0.1080
CHL_{t-14}	−0.0640	0.0365	−1.75	0.0804
DO_{t-8}	−34.7887	10.9376	−3.18	0.0015
TIN_{t-1}	1.7269	0.3317	5.21	0.0001

Note: All series are first-differenced.

Table A.2 – Estimation of 2-h data (2000–2001) by VARX model, no fractional differencing

Dependent variable	Coefficient estimate	Standard error	*t*-Value	*p*-Value
CHL_t				
CHL_{t-1}	−0.2391	0.0109	−21.93	0.0001
CHL_{t-2}	−0.2026	0.0109	−18.51	0.0001
CHL_{t-3}	−0.1300	0.0111	−11.71	0.0001
CHL_{t-4}	−0.1018	0.0112	−9.11	0.0001
CHL_{t-5}	−0.0964	0.0112	−8.64	0.0001
CHL_{t-6}	−0.1057	0.0112	−9.41	0.0001
CHL_{t-7}	−0.0920	0.0112	−8.25	0.0001
CHL_{t-8}	−0.1012	0.0110	−9.19	0.0001
CHL_{t-9}	−0.0297	0.0108	−2.75	0.0061
CHL_{t-12}	0.1073	0.0106	10.12	0.0001
CHL_{t-13}	0.0697	0.0106	6.56	0.0001
DO_{t-4}	6.1241	3.5522	1.72	0.0847
DO_{t-8}	−9.7275	3.5333	−2.75	0.0059
DO_{t-10}	6.0571	3.5117	1.72	0.0846
DO_{t-11}	−6.0791	3.5314	−1.72	0.0852
DO_{t-12}	7.3670	3.5456	2.08	0.0378
SR_{t-2}	14.7275	1.5711	9.37	0.0001
$TEMP_t$	−33.8388	7.0305	−4.81	0.0001
$TEMP_{t-1}$	33.4056	7.0235	4.76	0.0001
TIN_{t-5}	−16.2358	2.2205	−7.31	0.0001
TIN_{t-6}	−9.0745	2.2413	−4.05	0.0001
TIN_{t-7}	16.9944	2.2419	7.58	0.0001
TIN_{t-10}	−16.9670	2.2361	−7.59	0.0001
TIN_{t-11}	12.8372	2.2425	5.72	0.0001

Note: CHL_t, DO_t and TIN_t are first-differenced. $TEMP_t$ and SR_t are seasonally differenced by 12 periods.

REFERENCES

Anderson, D.M., 1994. Red tides. Sci. Am. 271, 62–68.

Anderson, D.M., 1997. Turning back the harmful red tide—commentary. Nature 388, 513–514.

Baird, M.E., Oke, P.R., Suthers, I.M., Middleton, J.H., 2004. A plankton population model with biomechanical descriptions of biological processes in an idealized 2D ocean basin. J. Mar. Syst. 50, 199–222.

Basterretzea, G., Garces, E., Jordi, A., Maso, M., Tintore, J., 2005. Breeze conditions as a favoring mechanism of Alexandrium taylori blooms at a Mediterranean beach. Estuar. Coast. Shelf Sci. 62, 1–12.

Donaghay, P.L., Osborn, T.R., 1997. Toward a theory of biological–physical control of harmful algal bloom dynamics and impacts. Limnol. Oceanogr. 42, 1283–1296.

Flynn, K.J., Clark, D.R., Owens, N.J.P., 2002. Modelling suggests that optimization of dark nitrogen-assimilation need not be a critical selective feature in phytoplankton. New Phytol. 155, 109–119.

Granger, C.W.J., Newbold, P., 1986. Forecasting Economic Time Series, second ed. Academic Press, New York, pp. 265–293.

Hosking, J.R.M., 1981. Fractional differencing. Biometrika 68, 165–176.

Kirkpatrick, B., Fleming, L.E., Squicciarini, D., Backer, L.C., Clark, R., Abraham, W., Benson, J., Cheng, Y.S., Johnson, D., Pierce, R., Zaias, J., Bossart, G.D., Baden, D.G., 2004. Literature review of Florida red tide: implications for human health effects. Harmful Algae 3, 99–115.

Lee, J.H.W., Fernando, T.M.K.G., Wong, K.T.M., 2004. Real time prediction of coastal algal blooms using artificial neural

networks. In: Liong, S.Y., Phoon, K.K., Babovic, V. (Eds.), Proceedings of the Sixth International Conference on Hydroinformatics, vol. 2. World Scientific, Singapore, pp. 1465–1472.
Lee, J.H.W., Huang, Y., Dickman, M., Jayawardena, A.W., 2003. Neural network modelling of coastal algal blooms. Ecol. Model. 159, 179–201.
Li, W.K., 2004. Diagnostic Checks in Time Series. CRC/Chapman & Hall, Boca Raton, pp. 24–25.
Lu, S., Hodgkiss, I.J., 2004. Harmful algal bloom causative collected from Hong Kong waters. Hydrobiologia 512, 231–238.
Muttil, N., Lee, J.H.W., 2005. Genetic programming for analysis and real-time prediction of coastal algal blooms. Ecol. Model. 189, 363–376.
Muttil, N., Lee, J.H.W., Jayawardena, A.W., 2004. Real-time prediction of coastal algal blooms using genetic programming. In: Liong, S.Y., Phoon, K.K., Babovic, V. (Eds.), Proceedings of the Sixth International Conference on Hydroinformatics, vol. 1. World Scientific, Singapore, pp. 890–897.
Ouchi, A., 1982. Forecast of a red tide occurrence in the northern part of Hiroshima Bay—II. Prediction of red tide occurrence by means of multiple linear regression model. Bl. Jpn. Soc. Sci. Fish. 48, 1245–1250.
Perez, E., Booth, F.A., Martin, D.F., 2000. Modelling organism–organism interactions as a means of predicting Florida red tide management potential. J. Environ. Sci. Health A 35, 219–227.
Qi, Y., Chen, J., Wang, Z., Xu, N., Wang, Y., Shen, P., Lu, S., Hodgkiss, I.J., 1998. Some observations on harmful algal bloom (HAB) events along the coast of Guangdong, southern China in 1998. Hydrobiologia 512, 209–214.
Riise, T., Tjøstheim, D., 1984. Theory and practice of multivariate ARMA forecasting. J. Forecast. 3, 245–368.
Schofield, O., Grymski, J., Bissett, W.P., Kirkpatrick, G.J., Millie, D.F., Moline, M., Roesler, C.S., 1999. Optical monitoring and forecasting systems for harmful algal blooms: possibility or pipe dream?. J. Phycol. 35, 1477–1496.
Shumway, R.H., Stoffer, D.S., 2000. Time Series Analysis and Its Applications. Springer, New York, pp. 383–384.
Walsh, J.J., Penta, B., Dieterie, D.A., Bissett, W.P., 2001. Predictive ecological modeling of harmful algal blooms. Hum. Ecol. Risk Assess. 7, 1369–1383.
Wong, P.P.S., 1989. Occurrence and distribution of red tides in Hong Kong. In: Okaichi, T., Anderson, D.M., Nemoto, T. (Eds.), Red Tides: Biology, Environmental Science and Toxicology. Elsevier, Amsterdam, pp. 125–128.
Wong, P.P.S., 2003. Review of red tide research and management in Hong Kong. In: Shin, P.K.S. (Ed.), Turning the Tides—A Festschrift in Honour of Professor Brian Morton. The Marine Biological Association of Hong Kong, Hong Kong, pp. 299–330.
Yang, Z.B., 2000. Harmful algal blooms in selected Hong Kong coastal waters. Ph.D. Thesis. The University of Hong Kong, Hong Kong, pp. 35–57.
Yang, Z.B., Hodgkiss, I.J., 2004. Hong Kong's worst "red tide"—causative factors reflected in a phytoplankton study at Port Shelter station in 1998. Harmful Algae 3, 149–161.