Available online at www.sciencedirect.com
SCIENCE DIRECT®
ELSEVIER
Ecological Modelling 189 (2005) 363–376
ECOLOGICAL MODELLING
www.elsevier.com/locate/ecolmodel

# Genetic programming for analysis and real-time prediction of coastal algal blooms

Nitin Muttil[a], Joseph H.W. Lee[b,*]

[a] *Research Associate, Department of Civil Engineering, The University of Hong Kong, Pokfulam Road, Hong Kong, China*
[b] *Chair Professor, Department of Civil Engineering, The University of Hong Kong, Pokfulam Road, Hong Kong, China*

Received 6 August 2004; received in revised form 15 February 2005; accepted 29 March 2005
Available online 23 June 2005

**Abstract**

Harmful algal blooms (HAB) have been widely reported and have become a serious environmental problem world wide due to its negative impacts to aquatic ecosystems, fisheries, and human health. A capability to predict the occurrence of algal blooms with an acceptable accuracy and lead-time would clearly be very beneficial to fisheries and environmental management. In this study, we present the first real-time modelling and prediction of algal blooms using a data driven evolutionary algorithm, Genetic Programming (GP). The daily prediction of the algal blooms is carried out at Kat O station in Hong Kong using 3 years of high frequency (two-hourly) chlorophyll fluorescence and related hydro-meteorological and water quality data. The results for the prediction of chlorophyll fluorescence, a measure of algal biomass, are within reasonable accuracy for a lead-time of up to 1 day. The results generally concur with those obtained with artificial neural network. As compared to traditional data-driven models, GP has the advantage of evolving an equation relating input and output variables. A detailed analysis of the results of the GP models shows that GP not only correctly identifies the key input variables in accordance with ecological reasoning, but also demonstrates the relationship between the auto-regressive nature of bloom dynamics and flushing time. This study shows GP to be a viable alternative for algal bloom modelling and prediction; the interpretation of the results is greatly facilitated by the analytical form of the evolved equations.


*Keywords:* Harmful algal blooms; Red tides; Data-driven models; Genetic programming; Real-time prediction; Water quality modelling; Hong Kong

## 1. Introduction

Harmful algal blooms (HABs) refer to the explosive growth and accumulation of harmful microscopic algae (phytoplankton). The well-known form of algal

* Corresponding author. Tel.: +852 2859 2672; fax: +852 2559 5337.
*E-mail address:* hreclhw@hkucc.hku.hk (J.H.W. Lee).

0304-3800/$ – see front matter © 2005 Elsevier B.V. All rights reserved.
doi:10.1016/j.ecolmodel.2005.03.018

bloom – the red tide – has been widely reported and has become a serious environmental problem due to its negative impacts on human health and aquatic life (e.g. anoxia or shellfish poisoning). In the past two decades there is an increasing trend in the occurrence of harmful algal blooms throughout the world. In particular, in April 1998, a devastating red tide resulted in the worst fish kill in Hong Kong's history, it destroyed over 80% (3400 tonnes) of cultured fish stock, with estimated loss of more than HK$312 million (Lee and Qu, 2004). Thus, a capability to analyze and predict the occurrence of algal blooms with an acceptable accuracy and lead-time would clearly be very beneficial to fisheries and environmental management.

Traditionally, models of phytoplankton dynamics are based on theories of the dependence of growth and decay factors on physical and biotic environmental variables (e.g. solar radiation, nutrients, flushing)—expressed mathematically and incorporated in advective diffusion equations in a water quality model. Such deterministic models are normally referred to as process-based models. Nowadays, with the availability of large amounts of data and with development of artificial intelligence techniques, a new paradigm of modelling called "data-driven modelling" has emerged.

Data-driven models are ideally suited to model the algal dynamics since such models can be set up rapidly and is known to be effective in handling dynamic, nonlinear and noisy data, especially when underlying physical relationships are not fully understood, or when the required input data needed to drive the process-based models are not available. In the recent past, various data-driven models, such as artificial neural network (ANN) and fuzzy logic models, have been applied to model the water quality variables with different degrees of success (Recknagel et al., 1997, 2002; Maier et al., 1998; Chen and Mynett, 2003; Lee et al., 2003). In the present study, we employ an evolutionary based data-driven model, the Genetic Programming (GP) for analysis of the high frequency (two-hourly) chlorophyll fluorescence and related hydro-meteorological and water quality data. In the following sections, we first outline the key principles of genetic programming, followed by its application to modelling of algal dynamics. The optimal GP model for real time prediction of algal dynamics is then presented along with a comparison of model performance with those of other data driven models. Finally, the relationship between the auto-regressive nature of the revealed algal dynamics and flushing time is investigated using long-term water quality data of a similar semi-enclosed coastal water (Tolo Harbour).

## 2. Genetic programming

Genetic Programming (GP) is a relatively new automatic programming technique for evolving computer programs to solve problems (Koza, 1992). In engineering applications, GP is frequently applied to model structure identification problems. In such applications, GP is used to infer the underlying structure of either a natural or experimental process in order to model the process numerically. A number of applications of GP have been reported, which include sediment transport modelling, salt-water intrusion in estuaries (Babovic and Abbott, 1997b); effect of flexible vegetation on flow in wetlands (Harris et al., 2003; Babovic and Keijzer, 2000) and emulating the rainfall-runoff process (Whigham and Crapper, 2001; Liong et al., 2002). GP inferred models have the advantages of generating simple expressions and thus offering some possible interpretations to the underlying process.

Genetic Programming is a member of the Evolutionary Algorithm (EA) family. EAs are engines simulating grossly simplified processes occurring in nature. The fundamental idea is that of emulating the Darwinian theory of evolution, where a population is progressively improved by selectively discarding the not-so-fit population and breeding new children from better populations. EAs work by defining a goal in the form of a quality criterion and then use this goal to measure and compare solution candidates in a stepwise refinement of a set of data structures and return an optimal or near-optimal solution after a number of generations. Evolutionary Strategies (ES) (Schwefel, 1981), Genetic Algorithms (GA) (Holland, 1975) and Evolutionary Programs (EP) (Fogel et al., 1966) are three early variations of evolutionary algorithms whereas GP is a relatively new approach (Koza, 1992). These techniques have become extremely popular due to their success at searching complex non-linear spaces and their robustness in practical applications.