DATA DRIVEN MODELLING FOR ENVIRONMENTAL WATER MANAGEMENT

Mofazzal Syed

Thesis submitted for the degree of
Doctor of Philosophy

Division of Civil Engineering, School of Engineering
Cardiff University

July 2007

ProQuest LLC
789 East Eisenhower Parkway
P.O. Box 1346
Ann Arbor, MI 48106-1346

In the name the almighty,
the greatest scientist and engineer of all,
who created myself and the universe.

To my wife **Tuni** with love,
for making me happier
than I ever thought I could be

and

our son **Abudllah**,
whose birth is the best thing
that could ever happen to me.

ACKNOWLEDGEMENTS

I would like to express my gratitude and appreciation to my supervisors Professor Roger Falconer and Dr Binliang Lin for their continued helpful advice, support and encouragement throughout the period of this research project. Professor Falconer not only persuaded me to take this strenuous path but had also instilled a true passion about the subject in myself. He is far more then a research supervisor to me and will remain so for ever. I lost count of the numerous occasions when I demanded his precious time in a busy day to discuss issues ranging from academic to personal and he never declined.

I would like to thank all my teachers, present and past, for instilling the proper knowledge in me, which has enabled me to pursue this path.

I am also indebted to my friends in the Hydro-environmental research group and beyond for responding to my sometimes undaunted demands and to the research office staffs in School of Engineering their unfailing and continuous support.

I will be eternally grateful for the support of my family, particularly my parents, whose encouragement and belief helped me to continue.

And finally I would like to thank my beautiful wife Tuni who supported me through some extremely difficult times and without whom the project would never have been completed.

Abstract

Management of water quality is generally based on physically-based equations or hypotheses describing the behaviour of water bodies. In recent years models built on the basis of the availability of larger amounts of collected data are gaining popularity. This modelling approach can be called data driven modelling. Observational data represent specific knowledge, whereas a hypothesis represents a generalization of this knowledge that implies and characterizes all such observational data.

Traditionally deterministic numerical models have been used for predicting flow and water quality processes in inland and coastal basins. These models generally take a long time to run and cannot be used as on-line decision support tools, thereby enabling imminent threats to public health risk and flooding etc. to be predicted. In contrast, Data driven models are data intensive and there are some limitations in this approach. The extrapolation capability of data driven methods are a matter of conjecture. Furthermore, the extensive data required for building a data driven model can be time and resource consuming or for the case predicting the impact of a future development then the data is unlikely to exist.

The main objective of the study was to develop an integrated approach for rapid prediction of bathing water quality in estuarine and coastal waters. Faecal Coliforms (FC) were used as a water quality indicator and two of the most popular data mining techniques, namely, Genetic Programming (GP) and Artificial Neural Networks (ANNs) were used to predict the FC levels in a pilot basin. In order to provide enough data for training and testing the neural networks, a calibrated hydrodynamic and water quality model was used to generate input data for the neural networks. A novel non-linear data analysis technique, called the Gamma Test, was used to determine the data noise level and the number of data points required for developing smooth neural network models. Details are given of the data driven models, numerical models and the Gamma Test. Details are also given of a series experiments being undertaken to test data driven model performance for a different number of input parameters and time lags. The response time of the receiving water quality to the input boundary conditions obtained from the hydrodynamic model has been shown to be a useful knowledge for developing accurate and efficient neural networks.

It is known that a natural phenomenon like bacterial decay is affected by a whole host of parameters which can not be captured accurately using solely the deterministic models. Therefore, the data-driven approach has been investigated using field survey data collected in Cardiff Bay to investigate the relationship between bacterial decay and other parameters. Both of the GP and ANN models gave similar, if not better, predictions of the field data in comparison with the deterministic model, with the added benefit of almost instant prediction of the bacterial levels for this recreational water body.

The models have also been investigated using idealised and controlled laboratory data for the velocity distributions along compound channel reaches with idealised rods have located on the floodplain to replicate large vegetation (such as mangrove trees).

Keywords: Data-driven Model, Numerical models, Genetic Programming, Artificial neural networks, recreational water, vegetation, winGamma

TABLE OF CONTENTS

LIST OF FIGURES

CHAPTER 2

CHAPTER 3

CHAPTER 4

CHAPTER 5

CHAPTER 6

CHAPTER 7

LIST OF TABLES

CHAPTER 7

NOTATIONS

C	Chezy's bed roughness coefficient
C	E. Coli concentration at time t
C_0	E. Coli concentration at time t = 0
C_f	Courant number
C_D	drag coefficient
C_e	eddy viscosity coefficient
C_w	air/ fluid resistance coefficient
D_{xx}	depth averaged dispersion coefficient in x-direction
D_{xy}	depth averaged turbulent diffusion coefficient in x direction
D_{yx}	depth averaged turbulent diffusion coefficient in y direction
D_{yy}	depth averaged dispersion coefficient in x-direction
f	Coriolis parameter
f	Darcy-Weisbach resistance coefficient
g	gravitational acceleration
H	total water depth
H	water depth below datum
I_0	irradiance at the surface
I_z	irradiance at depth z
k	rate of bacterial decay
k_a	after growth rate of bacteria
k_d	rate of bacterial decay in darkness
k_e	vertical light extinction coefficient
k_i	rate of bacterial decay due to irradiance

k_l	depth averaged longitudinal dispersion coefficient
k_n	bacterial growth due to nutrient presence
k_s	Nikuradse equivalent sand grain roughness
k_s	rate of bacterial decay due to sunlight
k_s	net loss (or gain) due to settling (or resuspension) of bacteria
k_T	rate of bacterial decay at local water temperature
k_{20}	rate of bacterial decay at 20°c temperature
k_t	depth averaged lateral turbulent diffusion co-efficient
n	number of row or column
p	discharge per unit width in x-direction
q	discharge per unit width in y-direction
q_m	source or sink discharge per unit horizontal area
R_e	Reynolds number
S	depth averaged solute concentration
S_s	sources and sinks of the solute
T_E	duration of the maintenance phase in bacteria die-off
T_{90}	time required for 90% die-off of the bacteria
U	depth average velocity components in x-direction
U_*	bed shear velocity
u	instantaneous velocity component in x-direction
V	depth average velocity components in y-direction
v	instantaneous velocity component in y-direction
v_s	net loss rate of particulate bacterial forms
W_x	wind velocity component in x-direction
W_y	wind velocity component in y-direction

Z	vertical water depth
$‰$	solute parts per thousand (ppt)
α	proportionality constant
β	momentum correction factor
β	solpe of the log10 plot of die-off against irradiance
Δx	grid spacing in x direction
Δy	grid spacing in y direction
Δt	time interval
ϕ	depth averaged solute concentration
ε	depth averaged turbulent eddy viscosity
η	water surface elevation above datum
κ	von karman constant
ν	kinematic viscosity of fluid
θ	empirical constant for bacterial decay
ρ	density of fluid
ρ_a	density of air
τ_b	bed shear stress
τ_{xw}	surface shear stress due to wind action in x-direction
τ_o	mean shear stress
ω	speed of earth's rotation

ABBREVIATIONS

ADI	Alternating Direction Implicit
ADV	Acoustic Doppler Velocimeter
ANNs	Artificial Neural Networks
AODC	Acridine Orange Direct Counting
APHA	American Public Health Association
ART	Adaptive Resonance Theory
BOD	Biological oxygen demand
CHA	Cardiff Harbour Authority
CFU	Colony Forming Unit
CoD	Coefficient of Determination
CSO	Combined Sewer Overflows
DIVAST	Depth Integrated Velocities and Solute Transport
DWAF	Department of Water Affairs and Forestry
EU	European Union
EU BWD	European Union Bathing Water Directive
FC	Faecal Coliform
FDE	Finite Difference Equation
FS	Faecal Streptococci
GP	Genetic Programming
GRNN	Generalised Regression Neural Network
LVQ	Learning Vector Quantization
MF	Membrane Filter
MLP	Multi Layer Perceptron
MPN	Most probable Number

NTU	Nephelometric Turbidity Units
PDE	Partial Differential Equations
RBF	Radial Basis Function
RMSE	Root Mean Squared Error
SOD	Sediment Oxygen Demand
TS	Threat Score
USEPA	United States Environmental Protection Agency
WHO	World Health Organisation

CHAPTER 1

INTRODUCTION

1.1 Introduction

Today's major challenges for hydraulics engineers and water managers include: securing water for a varying degree of usage, protecting vital aquatic ecosystems and dealing with variability and uncertainty of water in space and time. The importance of clean, safe recreational bathing waters can not be any more overstated, with increasing numbers of people participating in recreational activities. Whether the water bodies are used for sport or relaxation, health or pleasure, there is something about the enjoyment, relaxation and sense of well-being derived from the experience of recreational water activities. Maintaining safe recreational waters requires a concerted effort from all stakeholders. From government regulatory authorities at all levels, to local businesses and industry, to beach managers, community members and recreational water users - all of these stakeholders have a role to play in helping keep the beaches clean and bathing waters safe.

Under the risk management approach to safe recreational water quality, an inspection of the bathing water area is first used to identify all of the potential sources of risk to human health and safety. Appropriate procedures or management actions are then introduced as barriers to reduce these risks. This concept is similar to the Multiple Barrier Approach (USEPA 2003) used in the management of safe drinking water supplies. Using this approach, compliance with the Guidelines becomes but one key piece of a larger picture of preventative risk management. For example, if an inspection of the bathing area has determined that the water quality results are poor following rainstorms, one action might be to restrict bather access immediately following periods of heavy rainfall. Knowing what the risks are, and how to manage them, is an important way to help ensure that recreational waters remain open for everyone to enjoy.

In recent years the health aspects of the recreational use of the aquatic environment has attracted increasing attention from members of the public, concerned professionals and regulatory agencies, and there has been increasing pressure to update legislation using improved epidemiological knowledge and more sophisticated managerial methods. The World Health Organisation (WHO) recognised the limitations of regulatory regimes for the microbiological quality of recreational water, based mainly on a percentage compliance with faecal indicator counts. For such a regulatory practice, management systems can only be retrospective, with actions taking place after humans have been exposed to the hazard. Additionally, waters are classified as safe or unsafe, where, in reality, there is a gradient of increasing severity, variety and frequency of health effects with increasing sewage pollution. As a result the WHO proposed new guidelines, known as the 'Annapolis Protocol' (WHO 1999), which looks towards an improved approach for the control of recreational water environments that better reflect health risks and provide enhanced scope for effective management intervention. The European Union have decided that bathing water quality should be monitored and tested in order to protect bathers from health risks and to preserve the aquatic environment from pollution. As needed by the guidelines, microbiological tests take place throughout the bathing season. However, there will always be time lapses between the sample being taken and the microbiological quality being known. At sites where water quality is known to be variable, or affected by short term water deterioration, this does little to inform the situation on any given day. Hence the current 'predict and protect' approach is being promoted as a means of recreational bathing water management. This 'predict and protect' approach is an important development from the historically more reactive approach for EU designated bathing waters, where results are posted retrospectively and as they become available.

1.2 Current practices of Hydro-environmental Modelling

Traditionally, Scientists and engineers build mathematical/numerical models in order to analyse and better understand the behaviour of real world systems. A mathematical model of a natural system can be derived in a deductive manner, using the laws of physics which describe the conservation of mass, momentum and energy. The behaviour of a more general aquatic system can be simulated to a high degree of accuracy in terms of other kinds of mathematical representation, involving the most important biological and chemical processes occurring in the natural system.

Establishing an acceptable model of an observed system is a challenging task that occupies a major portion of the mathematical modeller's time. It involves observations and measurements of the system's behaviour under various conditions, selecting a set of variables that are important for modelling, and formulating the model itself. The first milestone in the process of modelling a real-world system is the choice of the modelling formalism. Differential equations are one of the most widely accepted formalisms for modelling of dynamic systems, i.e., systems that change their state over time (Gershenfeld, 1999).

The next stage of the problem is concerned with transforming this point or interval representation into a distributed representation over an entire solution domain and for all time. Even for the simple mathematical representation of natural systems, bounded by a set of assumptions, the resulting expressions generally become almost impossible to be solved analytically, especially when the domain and the boundary conditions become complicated. Hence, the mathematical equations are represented in algebraic form to be solved using various numerical techniques. This has led to the extensive, and now almost universal, use of numerical methods to solve the governing equation, where arbitrary points and integral descriptions are extended to finite spatial descriptions. In doing so the differential terms in the governing equation are generally replaced by finite difference, finite element or finite volume representations of the various terms.

In managing bacterial water quality, individual pathogens are generally difficult and expensive to measure, hence for water quality studies it is therefore common practice to measure and/or model the levels of related indicator organisms. Historically deterministic numerical models have been used for predicting flow and water quality processes in coastal aquatic basins, with these models solving numerically the equations of mass (fluid and solute constituents) and momentum conservation. One such model is the DIVAST (Depth Integrated Velocities And Solute Transport) model as developed by Falconer (1976) and Lin and Falconer (1997). This model has been used in the current study.

1.3 Problems Associated with Current Practice

The practice of numerical simulation of flows and other processes occurring in water has now matured into an established and efficient part of hydraulics. Numerical models are increasingly reliable in simulating the natural aquatic environment. Such models assist water managers in understanding the flow regime under certain condition to work out the affects of any future development, such as how the water quality downstream would be affected due to the development of a new housing estate etc. Therefore the need for numerical modelling in water resources management is unquestionable. At the same time, however, the models themselves often become very much extended. In many situations, given the divergence between the response-time requirements and the computational-time requirements of numerical models, these models cannot be used as on-line decision support tools. As mentioned in previous chapters, following the 'predict and protect' philosophy recreational water managers are in need of warning citizens in advance of any potential short term water quality deterioration.

The main motivation for this study therefore lies in overcoming the long computational times usually required for physically based computational or numerical models, particularly for long term management and/or planning. In most practical case studies the model requires a very large number of grid points to represent the problem domain with sufficient resolution to output better and more accurate simulations. Even if current day increased processing computer speeds are taken into account, such numerical models still require too much computational time with respect to real time forecasting. As a result such models in themselves are not ideal for providing assistance in making real time beach management decisions, such as in response to occurring rainfall radar information, or increased pollution loads upstream, and thereby aiding in predicting imminent potential threats to public health.

An alternative modelling approach, so called 'data driven' modelling, can be easily put in place with very little human intervention would be able to flag up any unusual event associated with expected high level of bacterial loads, posing a potential threat of health risks. Once trained, the model becomes a parametric description of the function being approximated, which can then be used for future predictions within a significantly

lower time limit. As a result these techniques can be used for online decision making for practical scenarios, such as recreational water management.

1.4 Data Driven Modelling: an Alternative Approach

As opposed to the current practice of modelling based on a description of the physical or biochemical behaviour, the alternative approach is based on the analysis of all the data characterising the system under study. A model can be defined on the basis of connections between the system state variables (e.g. input, output and internal variables) with a limited knowledge of the details of the physical behaviour of the system. Such models are generally called data driven models. The emergence of new data sources and data analysis methods is allowing a new approach to the development and use of models for decision support and various management practices. Data driven modelling describes a process of model-building wherein models are created that fit the dynamics of the data rather than assuming a priori relationships among variables and their influence. Although more complex than their predecessors, the capabilities of these new data driven decision-support models make them potentially very powerful tools for prediction, function approximation and improving understanding of complex real world dynamics, while suggesting improved and more rapid decision alternatives.

During the last decade, due to the increasing availability of data, such models have become quite popular. The most popular technique by far is Artificial Neural Networks (ANNs) (Solomanite, 2002), but they are not the only technique. There are a wide range of machine learning techniques, such as decision trees, Bayesian methods fuzzy-rule based systems, support vector machines (SVM) and evolutionary algorithms, with all been successfully applied to model different civil engineering systems. In recent years data-driven models have been increasingly used for various types of water management studies. In this study, details are presented of the application of two popular data-driven modelling types, namely Artificial Neural Networks (ANNs) and Genetic Programming (GP), to predict the Faecal Coliform levels in estuarine and coastal waters and velocity distribution of an idealised flood channel.

1.5 Integrated Modelling Approach

Data driven models are data intensive and the performance of these models very much depends on the quality of the data. Sufficient quality data are often a stumbling block for this approach, particularly as high frequency data collection over a prolonged period is a rarity in water quality monitoring as most the survey campaigns are either short term, with a high frequency, or carried out for longer terms with a low frequency. A well calibrated numerical model can provide as much data as needed for data driven model development. Another limitation of the data driven approach is the extrapolation ability, which is thought to be not common for the data driven models (Hettiarachchi et al. 2005; Yin et al. 2003, Drécourt and Madsen, 2001, Minns and Hall, 1996). Therefore, if an extreme event is not presented in the training process of data driven models, then the model prediction for such a case would be less than reliable. It is not possible to make sure that events like 1 in 100 years occurrences can be included in the field data unless these situations really take place. The numerical models can simulate the affect of all extremities for a given water body, hence the model can be trained for all potential incidents. More importantly, the data on which the data driven models should be developed might not be sufficient or even exist as in cases when a future development is proposed such as construction of a new housing estate or land reclamation in coastal areas. Numerical models are flexible enough to easily accommodate all changes that might take place in the future. Therefore, it is desirable to integrate the benefits and positive aspects of both modelling approaches in any decision support system. The main objective of the study was therefore to develop an integrated approach for rapid prediction of bathing water quality in a large estuary, with the main water quality indicator being enteric bacteria.

1.6 Water Quality Indicators – Enteric Bacteria

The measurement of the abundance of enteric bacteria (i.e. total coliform, faecal coliform and faecal streptococci) is one of the most commonly used methods to establish the quality of natural coastal and estuarine waters. These measured data very much form a part of coastal water quality management, as monitoring the enteric bacteria counts in the coastal environment is one of the key aspects of the EU Bathing Water Directive, and is especially important as a standard parameter for the usage of water against human pathogens in recreational waters.

Research on the survival of enteric bacteria in the water environment has received considerable attention in recent years, primarily as the main indicator of water quality. Furthermore, the use of such data can be used to discover new relationships from the measured data, which would enable an on-line predictive model to be provided which would enable water managers to make day to day decisions leading to more effective control of water quality in coastal zones.

In past studies most effort has been concentrated on field investigations, laboratory studies or deterministic numerical modelling, but only limited studies have focused on deploying and developing data driven modelling techniques. Hence this study dedicated particular emphasis on applying data driven techniques for two different sets of problems, namely the water quality predictions in estuarine waters and velocity predictions in a vegetated compound channel.

1.7 Flow in Compound Channels with Vegetated Floodplains

Many rivers consist of a channel with adjacent floodplains. The bottom of the floodplain is generally higher and rougher than the bottom of the main channel, so that during flood events the river consists of a relatively deep channel and shallow floodplains, giving a so-called compound channel. Understanding the hydraulics of flow in a compound channel with vegetated floodplains is very important for determining the stage-discharge curve and for supporting the management of fluvial processes. Flow in a compound channel differs from that in a simple channel because at high discharges the water in a compound channel flows in an out-of-bank manner onto the adjoining floodplain. Because the shape of the cross section varies and the roughness of the main channel and the floodplains is very often different, the flow structure of a compound channel is usually very complex. Momentum transfer between the main channel and the floodplain generally decreases the discharge in the main channel, increases the discharge on the floodplain, and decreases the channel's total discharge capacity. This has been called the "kinematic effect" (Yang et al, 2007).

Vegetation generally increases the flow resistance, changes the velocity distribution, and affects the discharge capacity and sediment transport rate in any riverine system. The experimental results of Huang et al. (1999, 2002) showed that the velocity in the

main channel increased significantly after the floodplains were covered with vegetation. There are numerous research studies reported in the literature that are dedicated to the better understanding of the complex flow regime in such a complex ecosystem and research in this field is still very much ongoing. In this study, these new data driven modelling techniques have been applied to the complex and poorly understood phenomena of flow through idealised vegetation. The ability to predict, with improved accuracy, velocities within wetlands and other vegetated areas would be advantageous as these regions are increasingly being recognised for their natural flood alleviation properties. In this study, laboratory data collected in a flume, with steady flows over a deep channel and with relatively shallow vegetated floodplains were used to induce the formulation of expressions using a data driven discovery technique, namely genetic programming (GP). The Artificial Neural Network was also used in the same context, albeit purely as a velocity prediction tool.

1.8 Aim and Objectives of the Study

The primary aim of this study and thesis has been to investigate the use of conventional numerical models with the new paradigm of data driven modelling as part of a proposed integrated modelling practice. GP and ANNs were used to develop alternative and rapid simulation tools for predicting Faecal Coliform levels, which can offer advantages over traditional numerical models by providing decision support tools for day-to-day recreational water management. However, in order to investigate the capability of data driven models, prediction performances based purely on natural or laboratory data were undertaken to compare the scopes of GP and ANNs for water quality predictions. Such models were also studied using the deterministic models to provide data for training and analysing the data driven models.

In order to achieve these aims the following objectives have been paramount:

- Development of data driven models using synthetic data, field data and laboratory data and assessing the range of variability in performance.
- Use of existing numerical models as data generators for scenarios where insufficient or no field data are available to develop a data driven modelling tool. The numerical model needed to be refined in order to incorporate decay

parameters, to enable these parameters to be included in the development of the data driven models.

- Use of numerical models to analyse the flow field and other hydrodynamic characteristics in order to facilitate better performance of the data driven models.
- Integrating conventional numerical modelling techniques with advanced hydro-informatics techniques to offer a more practical approach to online decision making, especially when there are not enough data to develop a data driven model and not enough time to run a deterministic numerical model.
- Use of advanced nonlinear data analysis tools for pre-processing and finding the sensitivity of results to various input parameters.
- Application of GP and ANN modelling tools to predict the water quality indicator levels at two different sites, namely the Ribble Estuary and Cardiff Bay.
- Application of GP and ANN modelling tools for predicting velocity distributions in an idealised compound channel with vegetated flood plains.
- Comparisons of the performance of GP and ANN predictive tools for practical problems relating to water quality management.

1.9 Outline of the Thesis

The details of the study reported herein can be summarised as follows:

Chapter 1 presents a brief description of the needs of an alternative predictive tool in order to develop and apply the 'predict and protect' concept to bathing and recreational water quality management.

Chapter 2 describes the occurrence of pathogens in the environment, the concept of indicator organisms and discusses current legislation on water quality in further detail. A comprehensive literature review is then given, primarily focusing on investigating bacterial die-off factors in surface waters. The survey of peer reviewed papers and technical reports has been analysed to investigate the relationship between these variables and faecal indicator organism decay rates. This has led to the development of a series of functions which were applied to assist in the determination of suitable T_{90} values for use in hydro-environmental deterministic models.

Chapter 3 explains the governing equations of fluid flow and contaminant transport. These include the mass and momentum conservation equations in one-, two- and three-dimensional form. The advective-diffusion equation for solute transport is also explained. The important terms of the respective equations are briefly discussed.

Chapter 4 introduces two data driven modelling paradigms, namely: Genetic Programming (GP), based on the principles of evolutionary computing, and Artificial Neural Networks (ANNs), based on learning techniques similar to the human brain. The fundamentals of genetic programming have then been discussed, along with some currently popular artificial neural network variants.

Chapter 5 describes the model development using ANNs and GP to predict the faecal coliform levels in the Ribble Estuary. In order to provide sufficient data for training and testing the neural networks, a calibrated hydrodynamic and water quality model was used to generate input data for the neural networks. The hydrodynamic model has been refined to incorporate more parameters for bacterial decay. A novel non-linear data analysis technique, called the Gamma Test, was used to determine the data noise level and the number of data points required for developing smooth model results using the aforementioned techniques.

Chapter 6 provides details of the application of ANNs and GPs to the prediction of faecal coliform level predictions in Cardiff Bay, a freshwater body. The data used for the model development in this study were collected from routine maintenance surveys, undertaken by Cardiff Harbour Authority. The data for this study was also analysed using the Gamma Test.

Chapter 7 provides details of the velocity predictions in a compound channel with vegetated floodplains, using both GP and ANN techniques to analyse the data, acquired in the Hyder Hydraulics Laboratory. The models were then tested to check their accuracy of predicting velocity distributions across the channel for a range of hydrodynamic conditions.

Chapter 8 provides a summary of the studies and the main conclusions of the findings from this research programme, followed by recommendations for further study.

CHAPTER 2

WATER QUALITY AND HEALTH EFFECTS

2.1 Pathogens

Micro-organisms can be found in both aquatic and terrestrial environments, and most perform important functions within their respective environments. Ecosystems rely on micro-organism decomposers, which convert organic matter to nutrients that can then be used by plants and animals higher in the food chain. Humans and animals have micro-organisms resident in their digestive tracts and rely on them for digestion. These micro-organisms are then excreted in large numbers in faecal matter. A small percentage of these micro-organisms have been linked to disease and often death. These disease causing micro-organisms are known as pathogen.

Waterborne pathogens are disease-causing organisms, micro-organisms, viruses or protozoans that can be transmitted to people when they consume or come into contact with untreated or inadequately treated water. Generally these pathogens are present in human and animal faeces, and are deposited directly into water bodies by surface water flow and/or sub-surface water flow.

Urban pathogens are transported by storm water runoff, combined sewer overflows in many parts of the world and directly from wastewater treatment plants. Pathogenic micro-organisms also originate from many animal species left on watersheds including wildlife, pets and agricultural animals. Rosen (2000) identified the following characteristics of waterborne pathogens of concern:

a) The organisms are shed into the environment in high numbers, or they are highly infectious to humans at low doses.

b) The organisms can survive and remain infectious in the environment for long periods, or they are highly resistant to water treatment

c) Some kinds of bacterial pathogens can multiply outside of a host under favourable environmental conditions.

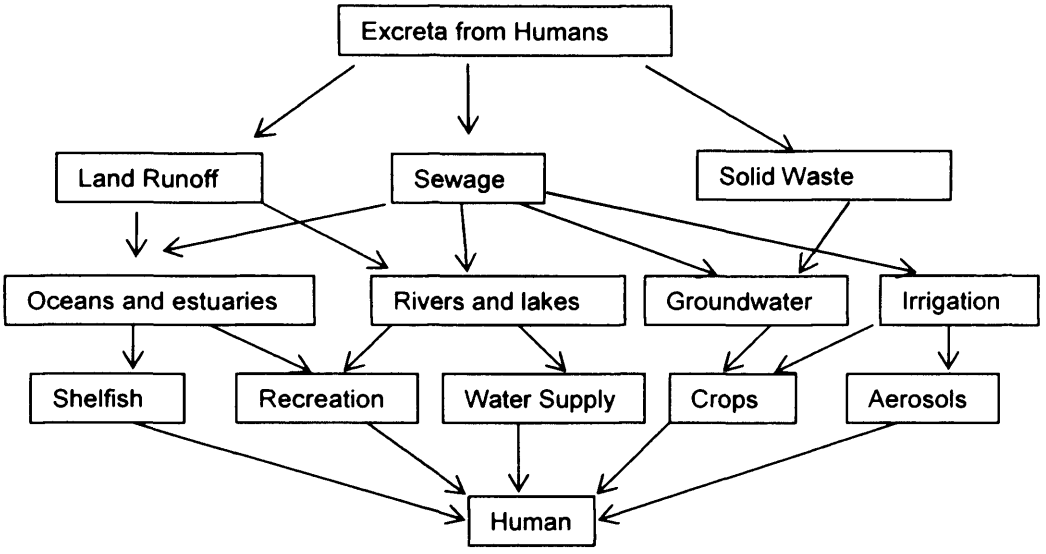


Figure 2.1: Routes of pathogens transmission

Once in a water body, pathogens infect humans through contaminated fish, selfish, skin contact, or ingestion of water. Figure 2.1, adapted from Bosch (1998) shows the pathogenic pathways to human.

A pathogen may be a bacteria protozoa, virus or fungi. The pathogens that are of most interest to aquatic related illnesses can be grouped into three subcategories – bacteria, protozoa, and viruses.

Bacteria are microscopic, unicellular organisms that reproduce by binary fission. They exist either as free living organisms or as parasites. They play a fundamental role in the decomposition and stabilisation of organic mater and in biological sewage treatment processes. Not all bacteria are pathogenic, but pathogenic bacteria that can be found in surface waters are often classified as coming from warm-blooded animals. The USEPA (2002) assessed bacteria as one of the leading causes of impairment of surface waters. With increasing demands on water resources, the potential for contamination of water by

pathogenic enteric bacteria is likely to rise world-wide resulting in an increase in the outbreaks of waterborne disease.

Protozoa are also unicellular organisms that reproduce by binary fission. Pathogenic protozoans exist in the environment as cysts, protecting themselves from harsh conditions such as temperature and salinity. Once the cysts are ingested they hatch, grow and multiply, infecting the host with the associated disease. Ingestion of only a few Protozoa by human causes disease, as they reproduce rapidly inside a host organism.

Viruses are sub-microscopic infectious agents that require a host to survive. The virus has a nucleic acid core that is protected by a protein or lipoprotein shell that can determine what surface to which it will attach itself. Once inside the host the virus reproduces, manifesting the associated diseases. Viruses are excreted in the faeces of infected individuals. Enteric viruses can cause major threats to human health.

2.2 Health Effects

Rapid population growth and urban development have resulted in regional domestic sewage and urban runoff problems and beach contamination has become the focus of public safety concerns. Recreational waters generally contain a mixture of pathogenic and non-pathogenic micro-organisms. These micro-organisms may be derived from sewage effluents, the recreational population using the water (from defecation and/or shedding), livestock (cattle, sheep, etc.), industrial processes, farming activities, domestic animals (such as dogs) and wildlife. In addition, recreational waters may also contain free-living pathogenic micro-organisms. These sources can include pathogenic organisms that cause gastrointestinal infections following ingestion or infections of the upper respiratory tract, ears, eyes, nasal cavity and skin.

Since the 1950s epidemiological studies have investigated the relationship between health risk and swimming. The risk of the health problems associated with swimming is related to the micro-biological quality of the water and it increases with increasing pollution levels. Fleisher et al. (1996) reviewed 11 previously published major studies and found an increased risk of gastroenteritis among bathers relative to non bathers. Corbett et al. (1993) reported that swimmers are almost twice as likely as non-swimmers to report symptoms.

Pruss (1998) presented Figure 2.2 which is adapted and updated from Pike et al. (1991) to show the predicted risk of illness to swimmers from adverse water quality. Swimming in contaminated marine and fresh recreational waters may result in a broad spectrum of illnesses, including: infections of the eyes, ears, skins, gastro-enteritis and upper respiratory tract diseases, although the definitions of these ailments and the associated risks have varied widely among studies.

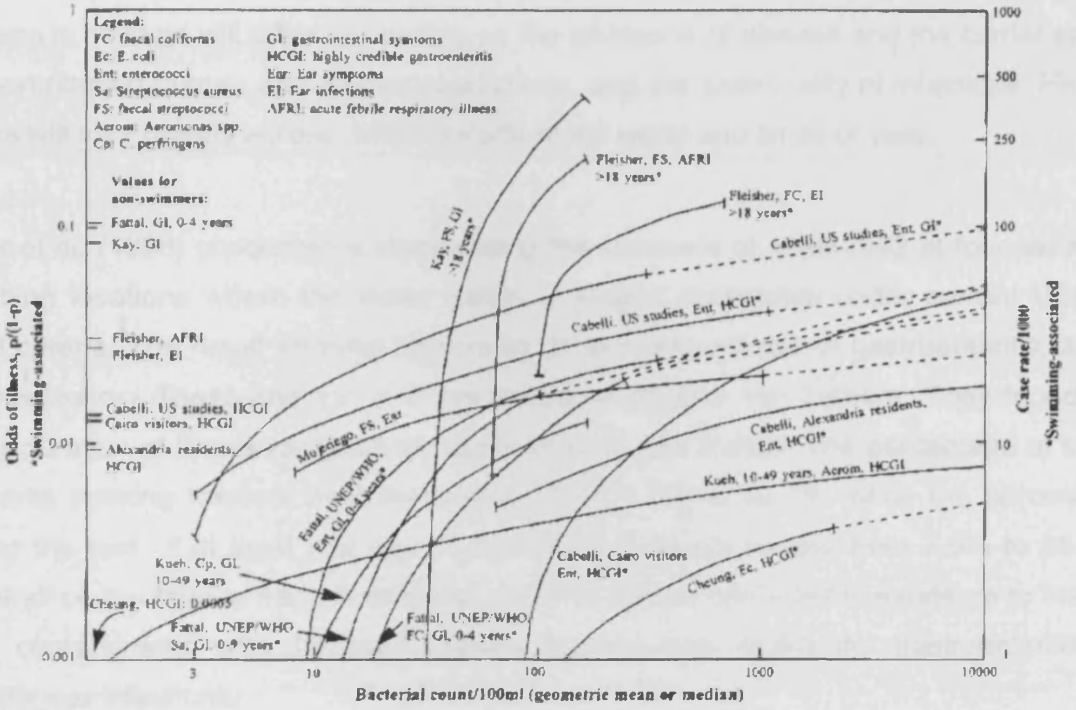


Figure 2.2: Predicted risks of illness in swimmers against bacterial count in marine water (Pruss, 1998)

Available evidence suggests that the most frequent adverse health outcome associated with exposure to faecally contaminated recreational water is enteric illness, such as self-limiting gastroenteritis, which may often be of short duration and may not be formally recorded. Thus the infections and illnesses due to recreational water contact are difficult to detect through routine surveillance systems. Even where the illness may be more severe, it may still be difficult to attribute the illness to water exposure. Targeted epidemiological studies, however, have shown a number of adverse health outcomes (including gastrointestinal and respiratory infections) to be associated with faecally polluted recreational waters. The transmission of pathogens that can cause gastroenteritis is biologically plausible and is analogous to waterborne disease transmission in drinking-water, which is well documented.

This can result in a significant burden of disease and economic loss. The number of micro-organisms (dose) that may cause infection or disease depends upon the specific pathogen, the form in which it is encountered, the conditions of exposure and the host's susceptibility and immune status. For viral and parasitic protozoan illnesses, this dose might be very few viable infectious units (Okhuysen et al., 1999). In reality, the body rarely experiences a single isolated encounter with a pathogen, and the effects of multiple and simultaneous pathogenic exposure is poorly understood (Esrey et al., 1985). The types and numbers of pathogens in sewage will differ depending on the incidence of disease and the carrier states in the contributing human and animal populations, and the seasonality of infections. Hence, numbers will vary greatly across different parts of the world and times of year.

Fleisher et al. (1998) conducted a study during the summers of 1989-1992 at four separate UK bathing locations where the water quality is judged acceptable under current USEPA and EU criteria. The result showed bathers to be at increased risk of gastroenteritis, acute febrile respiratory illness and ear and eye infections relative non bathers. They found the average duration of illness ranged from approximately 4 to 8 days. The percentage of study participants seeking medical treatment ranged from 4.2% to 22.2%, while the percentage reporting the loss of at least one day of normal daily activity ranged from 7.0% to 25.9%. The overall percentage of each illness that could be directly attributed to exposure to marine waters, contaminated with domestic sewage, ranged from 34.5% (for gastroenteritis) to 65.8% (for ear infections).

Some of these studies have suggested that the health risk also exists in those bathing waters meeting the bacteriological criteria of the EU 76/160 Directive and other guidelines. Fleisher et al. (1998) also reported a large burden of illness occurring in marine waters meeting both the current USEPA and EU criterion governing marine bathing waters. Pruss (1998) reviewed 22 studies and found that increased risk of gastro-intestinal symptoms were reported in water quality values ranging from only a few counts/100ml to about 30 indicator counts/100ml. These values are low compared to the water quality frequently encountered in coastal recreational waters. These observations question the appropriateness of such criteria and if these results are confirmed in future studies; they should be taken into account in establishing recommended levels for bathing water compliance. However, the verification of accuracy of current guidelines is beyond the scope of the current study.

2.3 Indicators Organisms

Indicators organisms are micro-organisms that denote faecal pollution and are detected at high numbers in polluted natural waters whenever pathogens are also present. The routine assessment of the sanitary quality of recreational waters is based on the analysis of these organisms.

The detection of pathogenic micro-organisms from a water sample is desirable to test the microbiological criterion for water quality, although many waterborne pathogens are difficult to detect in water samples, due to the fact that their presence is intermittent and at low levels. Also an optimal methodology for specific pathogen recovery may not yet have been developed.

In 1983, Cabelli listed several reasons why the use of indicators was a sound practice. These reasons remain sound today:

a) A large number of pathogenic bacteria and viruses are potentially present in municipal sewage, and each has its own probability of illness associated with a given dose;

b) Routine monitoring for each of the pathogens would be a Herculean task;

c) Enumeration methods for some of the more important pathogens are unavailable (e.g., hepatitis, rotaviruses and parvo-like viruses), and for the rest are difficult;

d) Pathogen density data are difficult to interpret because the methodology generally is imprecise and inaccurate and there are too little data available of dose-response;

e) On theoretical grounds, the intent is not to index the presence of the pathogen but rather its potential to be there in sufficient numbers to cause unacceptable health effects.

Snedecor (2003) described the criteria for an acceptable indicator organism as below

a) it must be part of the faecal-oral route

b) be present in the same or higher numbers as the target organism,

c) exhibit similar survival characteristics, and

d) be easily detectable

Classically, the indicators most widely used are total coliforms, faecal coliforms and faecal streptococci.

The coliform bacteria group is found in the intestines of warm-blooded animals. The presence of these bacteria is an indication that pathogens from untreated or partially treated sewage or contaminated runoff may be found in the relevant aquatic system. The most common types of microbial indicators are described below:-

2.3.1 Total Coliforms

The total coliform group is a general group encompassing all coliform bacteria. The group is easier to test for, but does not make the distinction between coliforms coming from faecal matter of warm-blooded animal vis-à-vis those naturally present in the environment.

2.3.2 Faecal Coliforms

Faecal coliforms bacteria are a sub-group of total coliform bacteria. They are more closely related to faecal matter and do not readily replicate in the water environment. (DWAF, 1995) The presence of faecal pollution by warm-blooded animals indicates the possible presence of pathogens responsible for infections diseases.

A set volume of the sampled water is cultured on an m-FC agar at 44.5°C. Faecal coliform bacteria will produce blue colonies within 20 – 24 hours of incubation. The colonies are then counted and the results are given as colony counts per 100ml or colony forming units (CFU) per 100ml. (DWAF, 1996)

2.3.3 Escherichia coli (E. coli)

Escherichia coli (E. coli) is a member of the faecal coliform bacteria group. E.coli is used as an indicator because it is highly specific to faecal contamination from humans and warm-blooded animals and because these bacteria cannot normally replicate in any natural water

environment. (DWAF, 1995) The presence of faecal pollution by warm-blooded animals indicates the possible presence of pathogens responsible for infectious diseases.

As with faecal coliform bacteria, E. coli will produce blue colonies on an m-FC agar within 20 – 24 hours of incubation at 44.5°C however only E.coli bacteria will test indole-positive at 44.5°C, (DWAF, 1996) E. coli are enumerated as colony counts per 100 ml or CFU per 100ml.

A few examples of bacterial pathogens whose presence are indicated by E. coli are: Salmonella spp., Shigella spp., Vibrio cholerae and pathogenic E.coli. These bacteria can cause diseases such as gastroenteritis, dysentery, cholera and typhoid fever. (DWAF, 1996).

2.3.4 Enterococci: Faecal Streptococci

Enterococci (faecal streptococci) bacteria are used to indicate the presence of faecal pollution by warm-blooded animals, which could contain pathogens responsible for infectious diseases. They are the preferred indicators of faecal pollution in the marine environment, as they survive longer than coliform bacteria in the water columns and the sediments (DWAF, 1995).

The bacteria produce typical reddish colonies on an m-enterococcus agar after 48 hours incubation at 35°C (DWAF, 1996). These colonies are counted and the results are given as the number of colony counts per 100ml or CFU per 100ml.

A few examples of bacterial pathogens for which streptococci is an indicator of are: Salmonella spp., Shigella spp., Vibrio cholerae and pathogenic E. coli. These bacteria can cause diseases such as: gastroenteritis, dysentery, cholera and typhoid fever. (DWAF, 1996).

2.4 Which is the Best Indicator?

In both marine and freshwater studies of the impact of faecal pollution on the health of recreational water users, several faecal index bacteria, including faecal streptococci and

intestinal enterococci, have been used for describing water quality. These bacteria are not suggested as the causative agents of illnesses in swimmers, but appear to behave similarly to the actual faecally derived pathogens (Prüss, 1998).

From time to time it is suggested that the pathogens in the water column should be measured directly rather than measuring indicators which may not directly indicate or quantify the presence of pathogenic bacteria and viruses. This issue has probably existed since the time when the use of indicator organisms was first introduced.

Prieto et al. (2001) reported that the count of total coliforms is the best predictor as they observed a relationship between gastrointestinal and skin symptoms and the degree of pollution, by total pollution while faecal coliform and staphylococci were reported to be more suitable by some other investigators. Fleisher et al. (1996) found that faecal streptococci exposure was predictive of acute febrile respiratory illness, while faecal coliform exposure was more relevant for ear ailments. Pruss (1998) suggested that enterococci/faecal streptococci for both marine and freshwater and E. Coli for freshwater correlates best with the health outcomes. However, Davis et al (1977) suggested that the total coliform group did not constitute a reliable source of information as to the pollutant content or condition of a water source. Reliance on the coliform group created serious problems both in measuring environmental quality and in calculating risks to public health (Saylor et al 1975). Borrego et al (1987) observed that total coliform displays different relationships with salmonella in marine, fresh and estuarine zones and hence concluded that the absence of total coliform in fresh and estuarine water indicates that Salmonella is also absent, but this correlation is not valid in marine waters. In fact detection of Salmonella in waters where there is no total coliform is also reported by Dutka (1973), Fugate et al. (1975) and Mack (1977).

Faecal coliforms have been considered indicators of faecal pollution of waters due to their presence in faeces, their relation with the presence of enteric pathogens such as Salmonella, and because they are present in higher concentrations in polluted waters than pathogens (Dufour, 1977). Several authors have reported loss of recovery of faecal coliforms in seawater due to microbial die-off and to entry of the micro organisms into a viable, but non-culturable, state (Roszak and Colwell, 1987; Barcina et al., 1997). However, it has been suggested that these processes do not invalidate the use of faecal coliforms as indicators of a recent pollution in such environments (Elliot and Colwell, 1985).

Faecal streptococci are considered by many authors to be a good indicator of Faecal pollution because they are more resistant than coliforms to environmental stress (Borrego et al., 1983; Philipp, 1991; Rees, 1993). Despite the opinion that faecal streptococci are rarely found in unpolluted environments (Rees, 1993), several studies seem to demonstrate that they can be found in other habitats (Geldreich and Kenner, 1969)

Dioniso et al (2002) found that for moderately-high levels of faecal pollution the best and most reliable indicators were faecal coliforms and E. coli, while for high concentration of faecal pollution the most appropriate indicators to detect the presence of pathogenic micro-organisms were faecal streptococci and coliphages. The unreliability of traditional bacterial indicators led some authors to suggest that coliphages might serve as indicators for faecal pollution, or for water potability testing. Coliphages have also been included in the new promulgation of the EU guidelines (1984). Several authors have suggested that coliphages may be good indicators for both the viral and faecal pollution of the waters (Borrego et al., 1987, 1990; Moriñigo et al., 1992), on the basis of their viral nature and on their higher resistance to physicochemical factors in water.

However, there are still many questions concerning the effectiveness of the way in which water quality is measured and monitored; a number of environmental and physical factors may influence the usefulness of faecal bacteria as indicators.

Borrego et al. (1987) remained convinced that the absence of indicator species does not guarantee absolutely clean water. There are a host of possible reasons for indicator presence and pathogen absence or vice versa (Ashbolt et al., 2001). No single indicator or approach is likely to represent all the facets and issues associated with contamination of waterways with faecal matter.

2.5 Measurement of Faecal Coliform

Bacteria are single-celled organisms that can only be seen with the aid of a very powerful microscope. However, coliform bacteria form colonies as they multiply, which may grow large enough to be seen. By growing and counting colonies of coliform bacteria from a sample of water, it is possible to determine approximately how many bacteria were present originally.

There are several ways coliform bacteria are grown and measured. Methods commonly used include the most probable number (MPN) method and the membrane filter (MF) method.

2.5.1 Most Probable Number Method (MPN)

In the MPN method, a "presumptive test" is performed first. A series of fermentation tubes that contain lauryl tryptose broth are inoculated with the water sample and incubated for 24 hours at 35°C. Fermentation tubes are arranged in 3 or more rows, with 5 or 10 tubes per row, and with varying dilutions of the samples in the tubes. The fermentation tube contains an inverted tube to trap gases that are produced by the coliform bacteria. After 24 hours, the fermentation tube is examined for gas production. If there is no gas production, the samples are incubated for another 24 hours and re-examined. If gas production is observed at the end of 48 hours, then the presumptive test is positive and coliform bacteria are present in the sample. A "confirmed test" is then performed to determine whether or not faecal coliform bacteria are present. For the confirmed test, some of the contents of the fermentation tube are transferred with a sterile loop to a fermentation tube containing another broth. The sample is incubated in a water bath at 44.5°C for 24 hours. Gas production in the fermentation tube after 24 hours is considered a positive reaction, indicating faecal coliform. Based on which dilutions showed positive for coliform and/or faecal coliform, a table of most probable numbers is used to estimate the coliform content of the sample. The results are reported as most probable number (MPN) of coliform per 100 ml (APHA, 1998).

2.5.2 Membrane Filter Method (MF).

The MF method is more rapid than the MPN method, but the results are not as reliable for samples that contain many non-coliform bacteria, high turbidity, and/or toxic substances such as metals or phenols. The water sample is filtered through a sterile membrane filter. The filter is transferred to a sterile petri dish and placed on a nutrient pad saturated with broth. The plates are inverted, placed in watertight plastic bags, and incubated in a water bath at 44.5°C for 24 hours. Colonies produced by faecal coliform bacteria are blue, and are counted using a microscope or magnifying lens. The faecal coliform density is recorded as the number of organisms per 100 ml. Sometimes the unit of colony producing units per 100 millilitres of water (CPU/100 ml) is used; this is equal to the number of organisms per 100 ml.

2.6 Classification of Recreational Water Use

The WHO (2000) classified the degree of water contacts as the following groups: Whole-body contact, Incidental contact and No contact recreation. The overall basis of risk reduction strategy depends on broad classification of recreational activities. The degree of water contact directly influences the degree of contact with infectious and toxic agents and physical hazards found in water and therefore the likelihood of being injured or contracting illness.

2.6.1 Whole-body Contact Recreation

Whole-body contact recreation is determined by the fact that the full body is likely to come into contact with and ingest water during the activity. Such activities include: swimming, diving, water skiing, surfing, paddle skiing and wind surfing. The people that participate in these activities span a wide range of ages, from infants to the elderly. The health status of users may also vary. Citizens that are not completely healthy are still inclined to swim while citizens taking part in more strenuous activities, such as surfing, are more likely to be fit and healthy.

2.6.2 Incident Contact recreation

Intermediate-contact recreation occurs when only limbs are regularly wetted and in which greater contact, including swallowing water, is unusual. Such activities include boating, wading and angling. The age groups that participate in such activities vary from children to the elderly and the health status of these individuals may also vary.

2.6.3 No Contact Recreation

Non-contact recreation involves recreation with no direct contact with the water and includes sightseeing, walking, horse riding, etc. These activities are predominantly concerned with the aesthetic appreciation of the water.

2.7 Current Guidelines for Water Quality Monitoring

In many fields of environmental health, guideline values are set at a level of exposure at which no adverse health effects are expected to occur. National and international institutions

have established several directives or guidelines to protect the environment and public health by reducing the pollution of bathing waters and by protecting such waters from further deterioration. These guidelines establish limits for the sanitary quality of recreational waters on the basis of the levels of indicator micro-organisms and of the absence of several pathogenic micro-organisms. Present regulatory schemes for the microbiological quality of recreational waters are primarily, or exclusively, based on the percentage compliance with faecal indicator counts.

However it should be noted that that there is no universally applicable risk management formula. "Acceptable" or "tolerable" excess disease rates are especially controversial because of the voluntary nature of recreational water exposure and the generally self-limiting nature of the most studied health outcomes (i.e. gastroenteritis, respiratory illness). Table 2.1, adopted from WHO (2000), shows a detailed list of guideline values used across the world.

Among the most widely used guidelines are the WHO guideline and the Bathing Water Directive (76/160/EEC) of the European Communities (EC). In 1998 the World Health Organization (WHO) organised an expert consultation to look into the adequacy and effectiveness of present approaches to monitoring and assessment, linked to effective management of microbiological hazards in coastal and freshwater recreational waters. The output of the meeting was the development of such an approach, which has become known as the 'Annapolis Protocol'. The European Communities (EC) adopted the Bathing Water Directive (76/160/EEC) which sets the mandatory and guideline microbiological standards for total coliforms and faecal coliforms, and guideline standards only for faecal streptococci.

The Environment agency in the UK monitors the quality of designated bathing waters in England and Wales against the regulations from the EC Bathing water Directive (76/160/EEC) There are two main sets of standards used for measuring bathing water quality: the minimum standard and the stricter guideline standard. All bathing waters must meet the minimum standard.

The mandatory (or imperative) standard, which should not be exceeded are

- 10,000 total coliforms per 100 ml of water
- 2,000 faecal coliforms per 100 ml of water

Table 2.1: Examples of guidelines and standards for microbiological quality of water (number of organisms per 100 ml) (WHO, 2000)

Country	Shellfish Harvesting		Primary Contact Recreation			Protection of Indigenous Organism		Reference
	TC*	FC**	TC*	FC**	Other	TC*	FC**	
Brazil		100%<100	80%< 5000[m]	80% <1000[m]				BrazilMinisterio del Interior, 1976
Colombia			1000	200				Colombia, Ministerio de Salud, 1979
Cuba			1000[a]	200[a] 90% < 400				Cuba, Ministerio de Salud, 1986
EC,[b] Europe			80%<500[c] 95%<10000[d]	80%<100[c] 95%< 2000[d]	Faecal streptococci 100[c] Salmonella 0/litre[d] Enteroviruses 0 PFU/ litre[d] Enterococci 90% <100			EEC, 1976 CEPOL, 1991
Ecuador			1000	200				Ecuador, Ministerio de Salud Publica, 1987
France, Israel			<2000 80%<1000[g]	<500	Faecal Streptococci< 100			CEPOL/UNEP, 1991
Japan	70		1000					Japan Environment Agency, 1981
Mexico	70[e] 80%<230		80% < 1000[f] 100%<10000[k]			10000[e] 80%<10000 100%<20000		Mexico, SEDUE, 1983
Peru	80%<1000	80%<200 100%<1000	80%< 5000[f]	80%<1000[f]		80%<20000	80%< 4000	Peru, Ministerio de Salud, 1983
Poland					*E. coli* <1000			WHO, 1975
Puerto Rico	70[h] 80%<230			200[h] 80%<400				Puerto Rico, JCA 1983
United Sates, California	70[c]		80%<1000 [i,j] 100%<10000[k]	200[a,j] 90%<400[l]				California State Water Resources Board, undated
United States, USEPA		14[a] 90%<43			Enterococci 35[a](Marine) 33[a](fresh) *E.coli* 126[a] (fresh)			US EPA, 1986; Dufour and Ballentine, 1986
Former USSR					*E.coli* < 100			WHO, 1977
UNEP/WHO		80%<10 100%<100		50%<100[n] 90%<1000[n]				WHO/UNEP, 1978
Uruguay				< 500[n] <1000[o]				WHO/UNEP, 1977
Venezuela	70[a] 90%<230	14[a] 90%<43	90%<1000 100%<5000	90%<200 100%<400				Venezuela, 1978
Yugoslavia			2000					DINAMA, 1998

* Total coliforms
** Faecal or thermotolerant coliforms
a. Logarithmic average for a period of 30 days of at least 5 samples
b. Minimum sampling frequency - fortnightly
c. Guide
d. Mandatory
e. Monthly Average
f. At laest 5 samples per month
g. Mini,u, 10 samples per month
h. At least 5 samples taken sequentially from the waters in a given instance
i. Period of 30 days
j. Within a zone bounded by the shoreline and a distance of 1,000 feet from the shoreline or the 30 foot depth contour, whichever is further from the shoreline
k. No sample taken during the verification period of 48 hours should exceed 10,000 per 100 ml
l. Period of 60 days
m. "Satisfactory" waters, samples obtained in each of the preceding 5 weeks
n. Geometric mean of at least 5 samples
o. Not to be exceeded in at elast 5 samples

Source: Adepted from Salas (1998)

In order for a bathing water to comply with the directive, 95% of the samples must meet these standards, plus other criteria.

The stricter guideline standards are those which should be achieved wherever possible, require:

- No more than 500 total coliforms per 100 ml of water for at least 80% of the samples (i.e. 16 or more out of 20 samples)
- No more than 100 faecal coliforms per 100 ml of water in at least 80% of the samples (i.e. 16 or more out of 20 samples)
- No more then 100 faecal streptococci per 100 ml of water in at least 90% of the samples (i.e. 18 or more out of 20 samples)

The European Commission uses the stricter guideline standard to assess bathing waters across all member states.

As there is no universally applicable risk management formula available, the assessment of recreational water quality should be interpreted or modified in the light of regional and/or local factors. Such factors include the nature and seriousness of local endemic illness, population behaviour, exposure patterns, and socio-cultural, economic, environmental and technical aspects, as well as competing health risk from other diseases, including those that are not associated with recreational water.

Recreational water standards have had some success in driving cleanups, increasing public awareness, contributing to informed personal choice and contributing to public health benefits. These successes are difficult to quantify, but the need to control and minimise adverse health effects has been the principal concern of regulation. Present regulatory schemes for the microbiological quality of recreational waters are primarily or exclusively based on percentage compliance with faecal indicator counts. WHO (2000) identified a number of constraints in the current standards and guidelines as given below:

- Management actions are retrospective and can only be deployed after human exposure to the hazard.

- The risk to health is primarily from human excreta, the traditional indicators of which may also derive from other sources.
- There is poor inter-laboratory and international comparability of microbiological analytical data.

While beaches are classified as safe or unsafe, there is a gradient of increasing severity, variety and frequency of health effects with increasing sewage pollution and it is desirable to promote incremental improvements, prioritising "worst failures".

2.8 Die-off of Indicator Bacteria

Enteric bacteria, and specifically the faecal indicator bacteria, are typically used to measure the sanitary quality of water for recreational, industrial, agricultural and water supply purposes. They are released into the environment with faeces, and are then exposed to a variety of environmental conditions that eventually cause their death. The results of past studies suggested that the survival of bacteria may be effected by any one, or a combination of, various inter related environmental, physical, physio-chemical and biological factors such as: solar radiation , adsorption to particulate matter and sedimentation, temperature, pH, salinity, specific ion toxicity (e.g. NaCl, iodate, heavy metals), lack of nutrients, utilisation of bacteria food by protozoa and other predators, the competitive and antagonistic affect of other micro-organisms and algal toxins with varying magnitudes of importance . However, construction of the quantitative relationships between the decline in bacterial population and these factors has not always been successful, mostly because of unsuitable experimental arrangements, with incomplete reporting of important variables. It is noteworthy that from the literature it appears that most researchers have only looked at the bactericidal processes in sea water. Understandably coastal areas, being the main recreational bathing areas, are subjected to pollution due to discharge of sewage in marine waters, and hence raising concerns to public health.

2.8.1 Sunlight

Sunlight is a major factor for bacteria survival in sea water (Gameson and Saxon, 1967; Fujioka et al., 1981; Bellair et al. ,1977). Fujioka et al (1981) found that in the absence of sunlight bacteria survived for days while in presence of sunlight 90% of faecal coliform and faecal streptococci were inactivated within 30 to 90 minutes and 60 to 80 minutes

respectively. They also found that the bactericidal effect of sunlight can penetrate into at least 3.3 m of clear sea water. A survey of the literature has indicated that solar radiation exerts a great influence on bacteria, bringing about their mortality at much higher rates compared to that of other relevant environmental factors.

Alkan et al (1995) suggested solar radiation is an important bactericidal factor, even for low levels of intensity of radiation, and turbid waters. They suggested that the influence of light is minimised when the least desired combination of the environmental conditions prevail, i.e. high turbidity preventing the penetration of light, high sewage content supporting bacterial life as well as contributing to turbidity, and the minimum degree of vertical mixing resulting in poor transportation of the bacteria to the upper layers of the water, where light penetration is more pronounced.

Solic and Krstulovic (1992) found that the response of the survival of FC was inversely proportional to solar radiation (Fig. 2.3) and the value of T_{90} decreased by about 40% for each 100 Wm^{-2} solar radiation increment. Figure 2.4 shows the relative decrease of solar radiation with depth, resulting in an increase of the T_{90} value. In their study solar radiation strongly affected the survival of FC in the first 30m below the surface. Below 30m, the effect of solar radiation was very weak because at that depth solar radiation accounted for only 10% of the intensity at the surface.

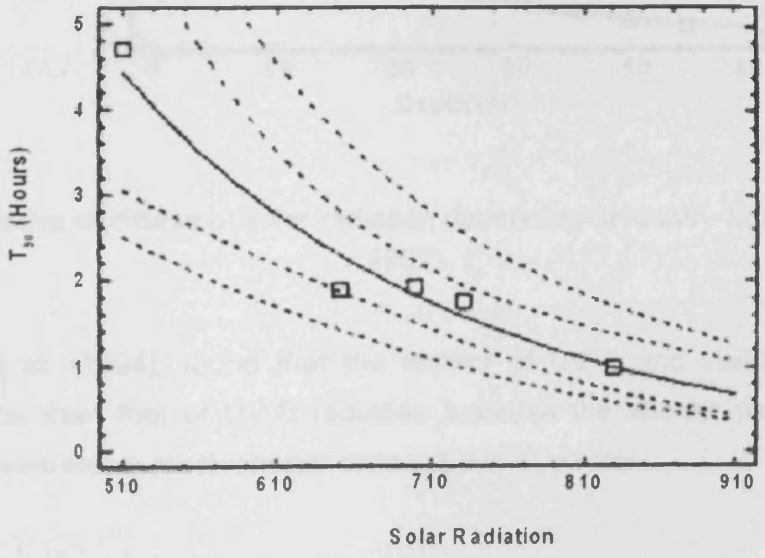


Figure 2.3: Effect of solar radiation on the survival of FC, Solic and Krstulovic (1992)

A significant number of researches have also looked at which part of the solar spectrum is mainly responsible for bacterial decay. Gameson and Gould (1975), Chamberlin and Mitchell (1978), Fujioka et al (1981) and Solic and Krstulovic (1992) found that it is the visible, rather than the ultraviolet, light spectrum of sunlight which is primarily responsible for the bactericidal effect. In fact, visible wavelengths may take on particular significance in natural systems firstly because u.v. wavelengths represent a small (< 3%) fraction of total incident radiation (Jassby and Powell, 1975; Kirk, 1983) and secondly, u.v. radiation is rapidly attenuated in the water column, especially when dissolved organic matter is present (Jerlov, 1968, Kirk, 1983, Wetzel, 1983). Barcina et al. (1989) suggested that the inability of E.Coli cells to take up glucose due to the action of visible light could result in the loss of culterability in freshwater.

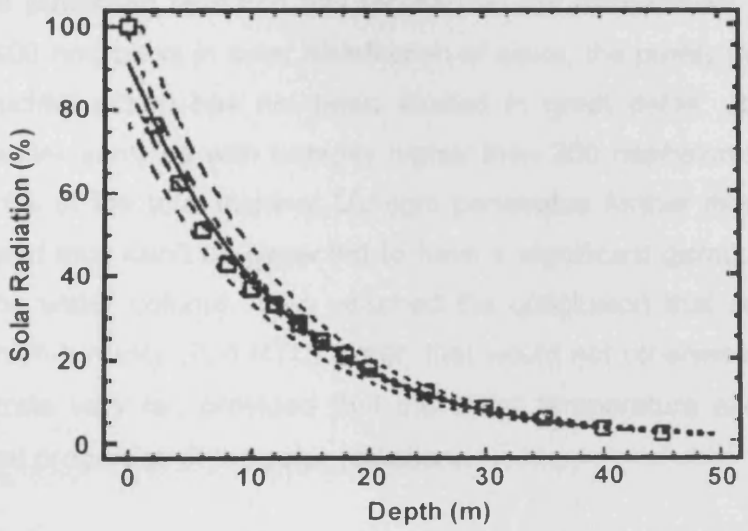


Figure 2.4: Relative decrease of solar radiation depending on depth, Solic and Krstulovic (1992)

Davies-Colley et al. (1994), found that the impact of UV-A and visible radiation in the sunlight is greater than that of UV-B radiation because the shorter wavelength of UV-B (290–320 nm) penetrates a much shorter distance into seawater.

However, Sinton et al. (1999) identified that all three components of the solar spectrum, namely UV-B (290–320 nm), UV-A (320–400 nm) and blue to green visible light (400– 550

nm) are responsible for the bactericidal effect, although they also reported that at wavelengths above 329 nm, when photochemical mechanisms become more important, the overall mechanisms for inactivation of coliforms is not very clear in seawater.

Gameson and Gould (1985) estimated that wavelengths lower than 370nm is responsible for about half the lethal effect of the sunlight. Davies and Evison (1991) found that the UV component of sunlight and high salinity act synergistically and result in a decrease in the number of culturable bacteria. They also carried out experiments on freshwater and observed that the effect of salinity on bacterial mortality is not evident unless exposed to light with a considerable UV component. They concluded that in fresh water the presence of UV absorbing substances (such as humic acid) protect cells from the possible damage to DNA by UV radiation and thereby extend survival.

While much of the published research has focused on the antibacterial role that solar UV radiation (200 to 400 nm) plays in solar disinfection of water, the purely thermal contribution of the solar germicidal action has not been studied in great detail. Joyce et al. (1996) observed that in water samples with turbidity higher than 200 nephelometric turbidity units (NTU), less than 1% of the total incident UV light penetrates further than a depth of 2 cm from the surface and thus can't be expected to have a significant germicidal effect beyond this distance in the water column. They reached the conclusion that solar disinfection is feasible even for high-turbidity (200 NTU) water, that would not otherwise allow incident UV radiation to penetrate very far, provided that the water temperature exceeds 55°C giving credit to the thermal properties of the solar radiation.

Reed (1997) observed different rates of inactivation in aerobic and anaerobic water concluded that solar disinfection of water is only fully effective under aerobic conditions. A decrease of the toxic effect was observed when E. coli was exposed to visible light under anaerobic conditions (Gourmelon et al, 1994).

2.8.2 Temperature

Early observations showed that the rate of disappearance of coliform bacteria in rivers was greater in summer than in winter, which led investigators to pursue a relationship between the death rate coefficient and temperature. An inverse relationship between the survival of

coliform bacteria and temperature has been reported by many researchers (McFeters and Stuart, 1972; Faust et al., 1975; Ayres and 1977, Solic and Krstulovic ,1992).

Wegelin et al. (1994) reported that the synergism of water temperatures above 55°C enhances the solar germicidal effect by a factor of approximately 2 for *Streptococcus faecalis* and E. coli. Flint (1987) observed a temperature-decay correlation based on laboratory studies with E Coli.

It has been found from some research studies that the effects of temperature become negligible in the presence of light, as the effects of sunlight override the effects of temperature (Alkan et al, 1995). Solic and Krstulovic (1992) found that the effect of temperature was obscured by the effect of sunlight up to a depth of 30 m, but below this depth temperature becomes more important as a factor controlling the survival of FC.

However, some researches drew different conclusions. Auer and Niehaus (1993) observed no significant relationship between coliform mortality and temperature. A number of investigators have reported a similar lack of dependence. Mitchell and Chamberlin (1978) cited work in the Ohio River by Frost and Streeter (1924), which demonstrated that total coliform death rates were virtually identical at 5 and 20°C. Moeller and Calkins (1980) also supported the fact that temperature has no significant effect on the decay of bacteria. Figure 2.5 shows the different findings on dependence of survival of FC on temperature.

Some investigators have suggested that a relationship between temperature and nutrients may facade the temperature effect on death rates in laboratory experiments. Auer and Niehaus (1993) suggested that the rates of biochemical reactions, and thus microbial growth rates, tend to increase as temperatures rise. High growth rates place added demands on nutrient reserves, which may not be renewed in dilute and natural systems, leading to an increase in the death rate. This effect on the nutrient utilisation and variability of nutrient availability in natural systems may explain the observed discrepancy in temperature-death relationship.

2.8.3 Salinity

Many studies have shown the inactivation of E. coli is more rapid in saline waters than inactivation in fresh water. A review of the published literature (Mitchell and Morris, 1969,

Faust et al., 1975; Fujioka et al., 1981) reveals that in most studies coliform bacteria were reported to survive for days in seawater. However, in some in situ studies these bacteria have been reported to be effectively inactivated within a few hours. The bactericidal effect of high salinity, probably caused by osmotic effects or by specific ion toxicity (Carlucci and Pramer, 1960), is reported in several other studies, such as Anderson et al. (1979); Ayres (1977), Carlucci and Pramer (1960) and Dutka (1984).

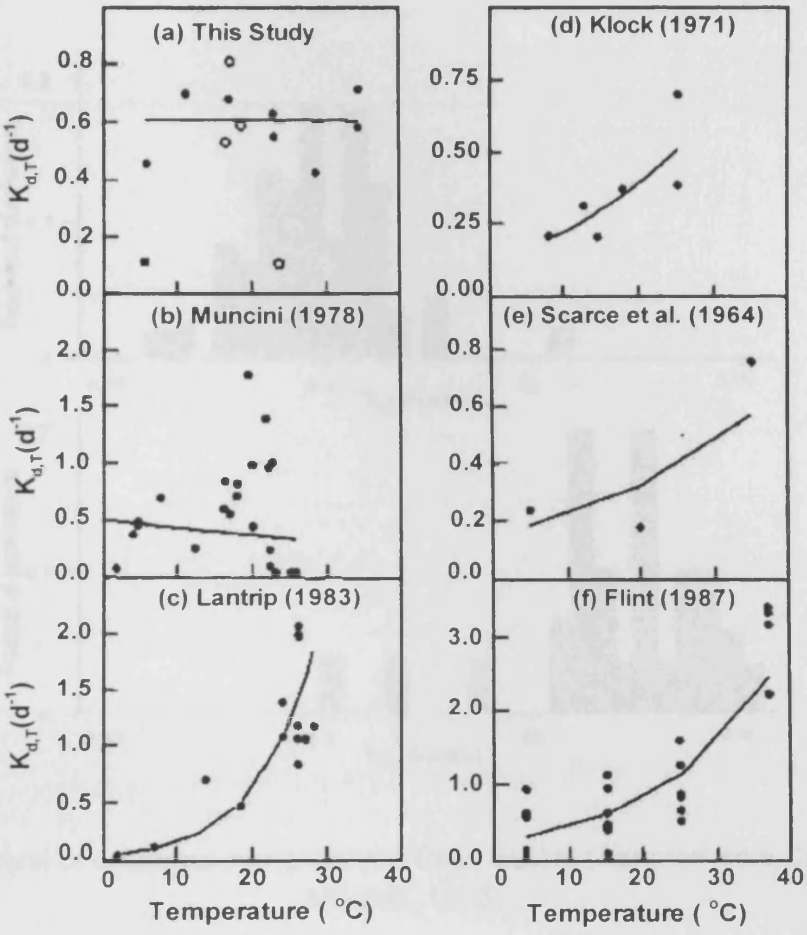


Figure 2.5: Effect of temperature on decay of bacteria in onondaga lake (Auer et al. 1993)

Solic and Krstulovic (1992) observed an inverse relationship between FC survival and salinity concentrations in water (Fig. 2.7). They found that increasing salinity was more detrimental to FC survival at lower salinity (in the range of 7-15‰) than at higher salinity levels (in the range of 15-40‰). In the range of salinity from 7-15‰ the value of T_{90}

decreased by about 55‰ for each 5‰ salinity increment, while in the range of salinity from 15-40‰ the value of T_{90} decreased only by 15‰ salinity increment.

Some authors observed the greatest survival of E. coil at salinity levels between 5 and 15‰, and rapid mortality in fresh water and at upper salinity above 25‰ (Ayres, 1977, Carlucci and Pramer, 1960).

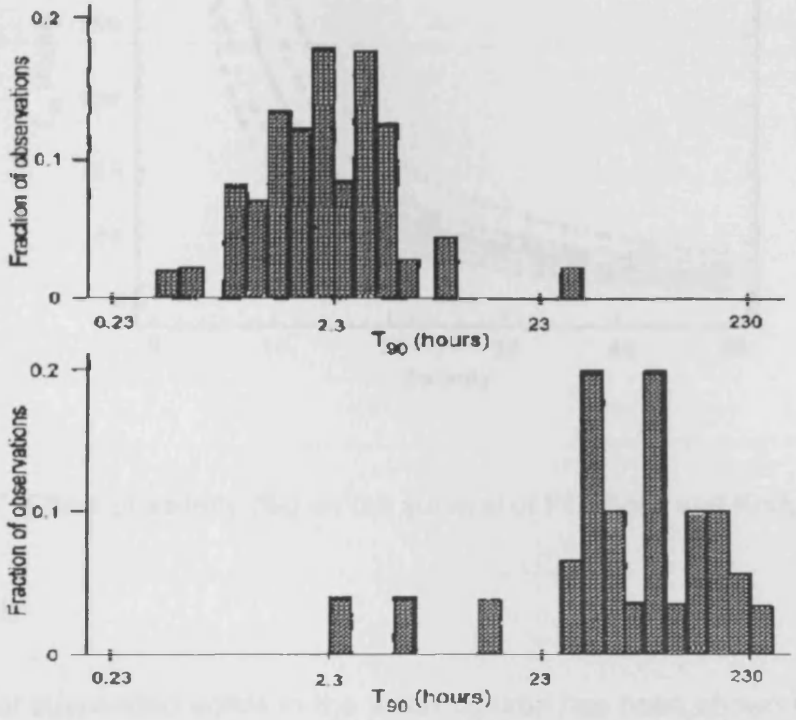


Figure 2.6: Survival of coliforms in marine and fresh waters (Adapted from Chamberlain and Mitchell, 1978)

Dutka (1984) found that the bactericidal effect of salinity is greater in the presence of sunlight. Sunlight and salinity work together to produce complimentary results are greater than either can produce individually. Fujioka et al (1981) noted that sunlight, both direct and indirect, is more lethal in saline waters. Davies and Evison (1991) achieved similar results indicating that the effects of salinity are more pronounced in the presence of UV radiation and that the UV component of sunlight and high salinity levels act synergistically to cause a decrease in the number of culturable bacteria. Bordalo et al. (2002) observed overall survival rates were higher in low salinity water. Light had a further deleterious effect, since it

accelerated the decay of faecal indicators, particularly in highly saline waters. They carried out their experiments in tropical estuarine water. Some authors stated that cells may be injured by solar light or salinity, but remains viable even though they are unable to form colonies.

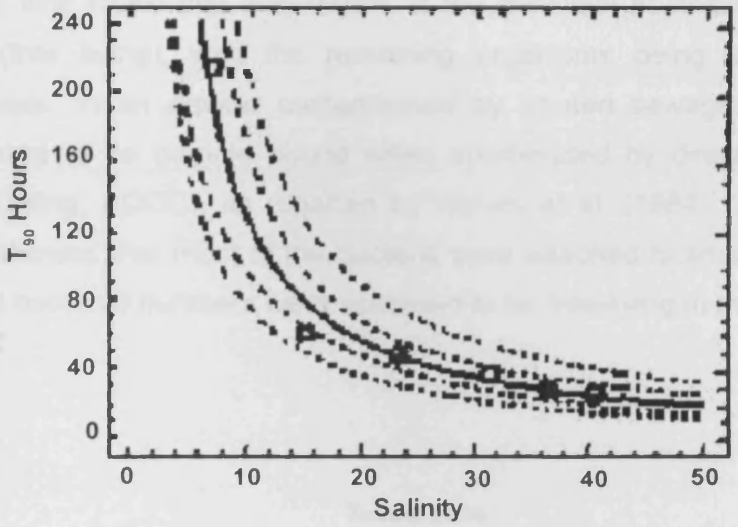


Figure 2.7: Effect of salinity (‰) on the survival of FC (Solic and Krstulovic, 1992)

2.8.4 Turbidity

The presence of suspended solids in the water column has been shown to increase the E. coli survival rates by limiting the affects of sunlight. Alkan et al (1995) documented the significant influence of sewage concentration, turbidity and vertical mixing have on E. coli inactivation in the presence of sunlight. Greater turbidity allows less light penetration, thus E. coli survive longer in turbid conditions; this suggests that the influence of light is minimized when the least desired combination of the environmental conditions prevail, i.e. high turbidity, high sewage content, and a minimal degree of vertical mixing (Alkan et al, 1995). Kay et al (2005) identified Turbidity as a dominant factor influencing decay when light and temperature remains constant. They observed that there is little difference between dark and irradiated T_{90} values above approximately 200 NTU. In fact work done by Joyce et al., 1996 (Key et al. 2005) suggests that at turbidity > 200NTU, around 90% of the incident radiation is absorbed in the first centimetre of the optical path through the water column.

Turbid water with a high suspended solid concentration provides a significant amount of suspended sediment particles with enough living surfaces for bacterial organisms. Past research has indicated that in natural turbid waters most of the bacterial organisms are found to be attached to suspended solids. Marshall (1978) indicated that bacteria readily absorbed to different kinds of interfaces, such as: liquid-solid, liquid-liquid, liquid-gas etc and most of the bcteria are attached to these surfaces. He also quoted the results from Jannasch (1956), who found that only 0.02% of the microbial population in the Nile River was planktonic (free living), with the remaining organisms being attached to mineral particulate materials. In an aquifer contaminated by treated sewage, 96.8-100% of the bacteria were found to be particle bound when enumerated by direct counting (acridine orange direct counting, AODC), as reported by Harvey et al. (1984). Also, Albrechtsen's (1994) research showed that most of the bacteria were attached to small particles and only 0.01% of the total bacterial numbers were assessed to be free-living in the pore-water.

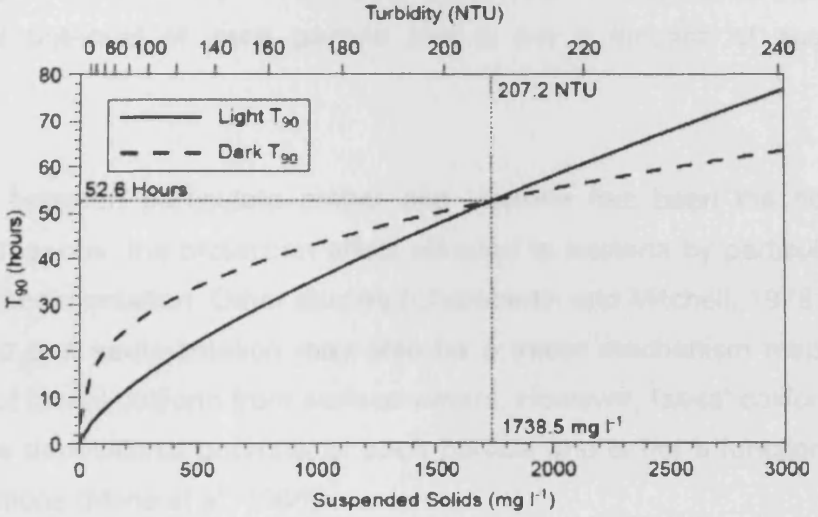


Figure 2.8: Relationship of irradiated and dark T_{90} with suspended solids/turbidity (Kay et al. 2005)

2.8.5 Nutrient concentration

Several authors have mentioned the nutrient concentration, coupled with competition for nutrients, have a considerable influence on bacterial survival rates. Lim and Flint (1989) showed that increased survival times are dependent upon higher nutrient concentrations. In the presence of adequate nutrients and no competition, as water temperature increases

there is a corresponding increase in the E. coli growth rates. High growth rates place added demands on nutrient reserves, leading to an increase in the death rate. It is believed that increased death rates at higher temperatures are a function of increased competition for nutrients by other organisms and increased predation. Rates of biochemical reactions, and thus microbial growth rates, tend to increase as temperatures rise. High growth rates place added demands on nutrient reserves, which may not be renewed in dilute, natural systems, leading to an increase in the death rate.

2.8.6 Sediment

Studies suggest sedimentation is a major mechanism responsible for the disappearance of faecal coliform from surface waters. Cells settle from the water column as discrete entities and as part of larger aggregates of faecal material, storm water debris and other suspended solids (Schillinger and Gannon, 1982). Sedimentation was considered to be one of the most important factors in E. coli removal and inactivation in a study undertaken by Auer and Niehaus (1993). They also noticed that 90.5% of the faecal coliform is associated with particles ranging from 0.45-10 μm. However, faecal coliform deposition was due solely to the depositional potential of each particle and is not a function of suspended solids concentration.

The interaction between particulate matter and bacteria has been the subject of much interest for two reasons: the protection affect afforded to bacteria by particulate matter and the potential for sedimentation. Other studies (Chamberlin and Mitchell, 1978: Gannon et al., 1983) suggested that sedimentation may also be a major mechanism responsible for the disappearance of faecal coliform from surface waters. However, faecal coliform deposition is due solely to the depositional potential of each particle and is not a function of suspended solids concentrations (Milne et al, 1986).

While dealing with the sediment part of the decay equation care is needed in the analysis as bacteria in natural waters exist in two forms in terms of their interaction with sediments (Stapleton et al., 2007). Some of the bacteria exist as free-living bacteria that stay within the water column, while others may attach to the suspended particles. The free-living bacteria move with the flow, while the attached bacteria move with the suspended particles, which could settle out onto the bed sediment surface when the suspended particles deposit and

also the turbulent flow can carry the particles with the attached bacteria to re-suspend into the overlying water. The free-living bacteria are also called free swimming bacteria.

Stapleton et al. (2007) found that, for natural waters 70%-99% bacteria exist as attached bacteria and the remaining 30%-1% bacteria as free swimming bacteria. It is thus understandable that due to the continuous bacterial deposition of bacteria organisms to the bed sediment, the populations of faecal bacteria on the bed sediment are, on average, 100-2000 times greater than the corresponding populations within the water columns.

2.8.7 pH

The affects of pH and predation were investigated in the literature. Faecal coliform, and specifically E. coli, were found to prefer a slightly more acidic environment, i.e. 5.0-7.0, while a pH of 8.0 or greater was found to have a negative impact on E.coli survival (Carlucci and Pramer, 1960). While pH does play a role in bacterial inactivation, the affects were found to be minor in comparison to other more dominant factors affecting bacterial survival.

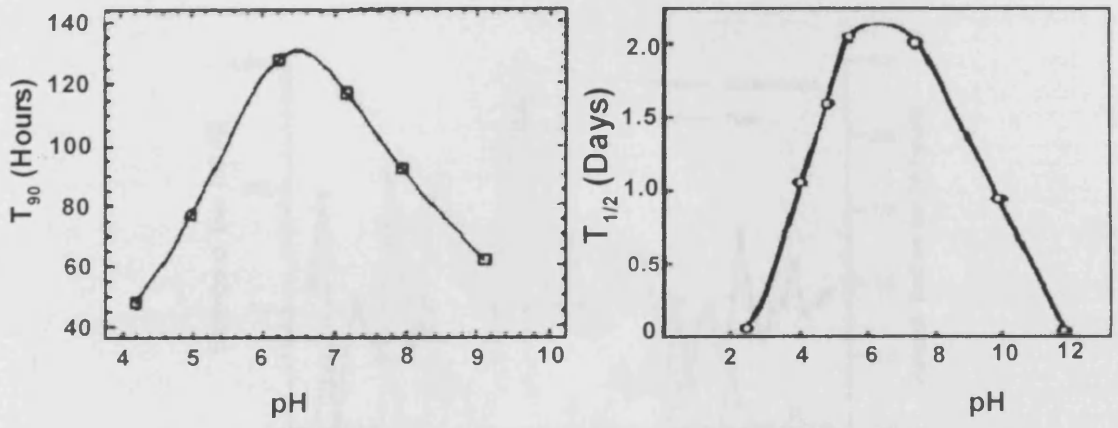


Figure 2.9: Effect of pH on decay rate ; left - Solic 1992, right -McFeters 1972

A review of published reports shows that the optimum pH for the survival of coliform bacteria ranges from 5.5-7.5 (McFeters and Stuart, 1972), and from 7-8 (Ayres, 1977). Šolić and krustulović (1992) had reduced the range to 6-7; although a pH value of 5 (Carlucci and Pramer, 1959) was also reported to be optimal. Šolić and krstulović (1992) observed that there was a slightly higher decrease in the T_{90} value for acid reactions (by about 40% for each value of pH) as compared with alkaline reactions (by about 30% for each value of pH),

even though they conceded that in natural conditions pH was a less important factor in controlling the persistence of FC than other studied factors, due to the small pH variability of seawater.

2.8.8 Predation

Predation, either by viruses, bacteria or protozoans, is another ecological factor that may contribute to the removal of non-indigenous bacteria from the environment.

2.8.9 Rainfall

Rainfall can have a significant effect on indicator densities in recreational waters increasing the densities to high levels, because animal wastes are washed from forest land, pasture land and urban settlings, or because treatment plants are overwhelmed causing sewage to by-pass the treatment process. In either case, the effect of rainfall on beach water quality can be quite dramatic (Figure 2.10) (Calderon, 1990, from WHO 2000). The effect, illustrated in Figure 2.10, on a beach surrounded by forests, was very rapid and usually persisted for 1-2 days.

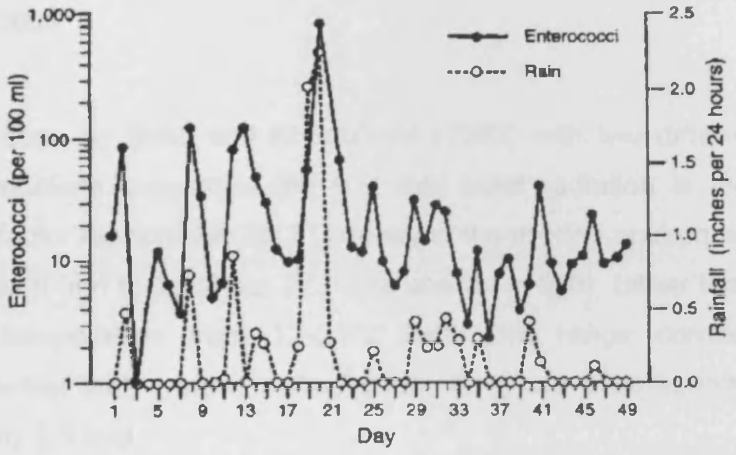


Figure 2.10: Effect of rainfall on enterococci densities in bathing beach waters (Calderon, 1990)

The highly variable effect of rainfall on water quality can result in the frequent closing of beaches. The important question is whether high indicator levels that result from animal wastes carried to surface waters by rain water run-off, indicate the same level of risk to swimmers as would exist if the source of the indicators was a sewage treatment plant. There

are conflicting reports in the literature with regard to risk associated with exposure to recreational water contaminated by animals.

2.9 Combined Effects of Different Decay Parameters

The literature review established the fact that the presence of sunlight was the major factor controlling the survival of faecal coliforms (FC) in seawater. The responses of FC survival to intensity of solar radiation, temperature and salinity were inversely proportional. The optimum pH for FC survival was between pH 6 and pH 7, with rapid decline both above and below these values. However, the interaction between these environmental factors was not considered in the majority of studies. Šolić and Krustulović (1992) investigated the combined effect of these factors, particularly the solar radiation-temperature and the solar radiation-salinity effect. They suggested as the increase of temperature and salinity is more detrimental to FC survival in the presence of sunlight, it may act synergistically with temperature or salinity. Many studies have shown that inactivation of E. coli is more rapid in saline waters than inactivation in fresh water (Carlucci and Pramer, 1960; Anderson et al, 1979; Fujioka et al, 1981; Milne et al, 1989; Davies and Evison, 1991). Sunlight and salinity work together to produce complimentary results, which are stronger than either can produce alone (Darakas, 2002).

The experiments done by Šolić and Krustulović (1992) with two different temperatures in light and dark conditions supported the fact that solar radiation is more important than temperature as a factor responsible for FC decay in the marine environment (Fig. 2.11). The T_{90} value of FC was found to be about 27 times shorter in light, rather than in the dark, while the increase in temperature from 12-24°C (with this range corresponding to yearly oscillations of seawater temperature in the Adriatic Sea) was accompanied by a decrease in the T_{90} value of only 2.5 fold.

An increase of salinity from 10-35% caused relatively more rapid mortality of FC in sunlight than in dark conditions, suggesting that sunlight and salinity may have acted synergistically. This result was confirmed by the significant salinity-solar radiation interaction shown in Figure 2.12. The authors were convinced that the bactericidal effect of salinity is enhanced in the presence of sunlight. They suggested that salinity should be taken into consideration

in determining the FC survival rate, particularly in areas with marked oscillations in salinity (e.g. estuarine waters).

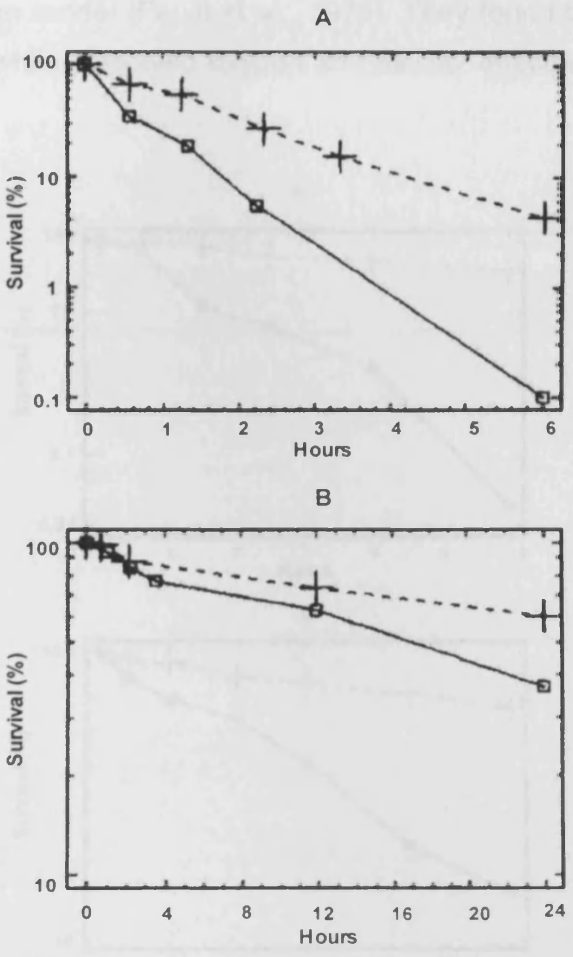


Figure 2.11: Survival of FC in the sunlight (A) and dark (B) at two different temperatures (23.8°C solid lines and 12.1°C dashed lines), (Solic and Krustulvoc 1992)

It is apparent from these results that solar radiation, temperature, and salinity interact to produce the most significant observed decline of FC in seawater. There are few studies reporting the combined effects of these different factors on the survival of coliform bacteria in seawater, in comparison to the number of such studies dealing with their separate (or individual) effects.

McCambridge and McMeekln (1981) found that naturally occurring microbial predators and solar radiation interact to produce part of the observed decline of sewage bacteria in estuarine water samples. That is, the decline in numbers of E. coli cells was found to be

significantly greater in the presence of both naturally occurring microbial predators and solar radiation than when each of these factors was acting independently. The combined effect of water temperature, dissolved oxygen, and salinity on the survival of E. coli was studied using a multiple regression model (Faust et al., 1975). They found that temperature mostly affected E. coli survival, while dissolved oxygen and salinity affected bacterial survival to a much lesser degree.

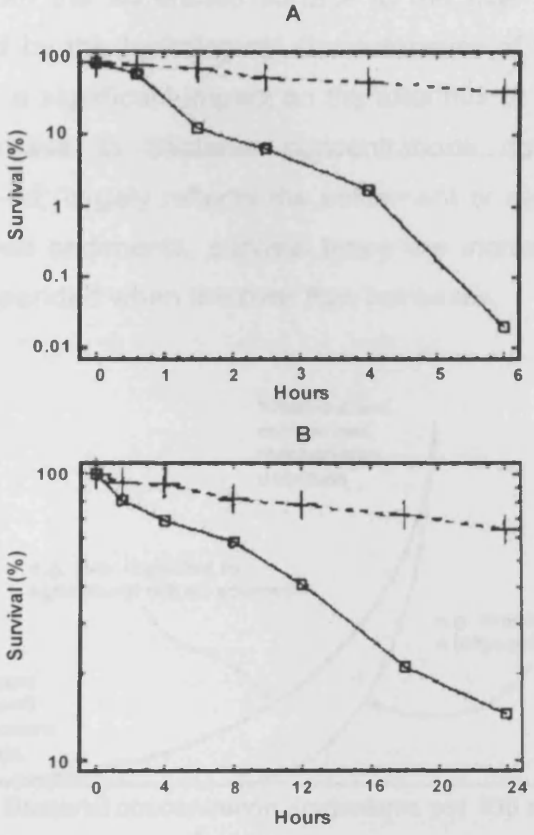


Figure 2.12: Survival of FC in sunlight (A) and in darkness (B) for two different salinities (35‰ - solid lines and 10‰ dashed lines), (Solic and Krustulvoc 1992)

2.10 Hydrological Considerations

Rivers contribute a significant proportion of the bacterial load to coastal bathing waters. In some regions, significant numbers of freshwater beaches are directly affected by river water quality. The bacterial concentration in river water is determined by faecal pollution from point sources and non-point or diffuse sources. Major point sources include sewage effluents, CSOs, industrial effluents and confined animal sources, such as feedlots. Non-point sources relate directly to agricultural activity within the watershed, and are influenced primarily by the

type of livestock and its density. A significant contribution is also derived from urban surfaces.

The transport of microbial contaminants through the watershed to the river, and subsequently through the river system to the marine environment is controlled by the flow of water. Rainfall is a key influence on the concentrations of coliform in bathing waters. Faecal material is transported from the watershed surface to the river and changes in flow are determined by rainfall and by the hydrological characteristics of the basin (soils, bedrock, etc.) which therefore have a significant impact on the total flux of the transported microbes. In river water the decrease in bacterial concentrations downstream of a source, conventionally termed *die-off*, largely reflects the settlement or sedimentation of organisms to the river bed. In riverbed sediments, survival times are increased significantly and the bacteria are readily re-suspended when the river flow increases.

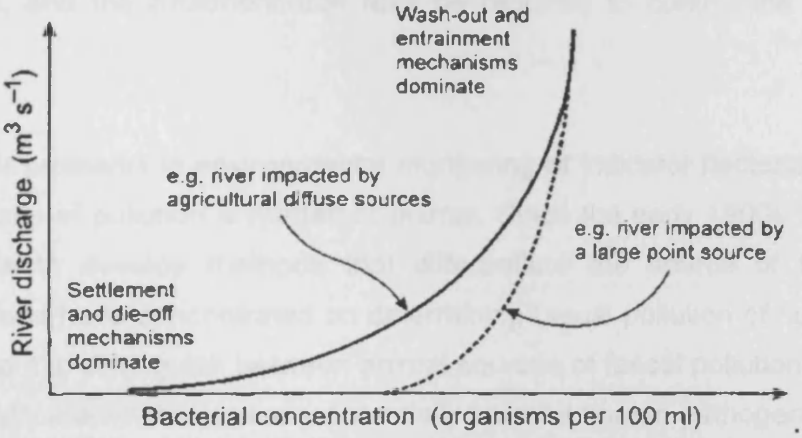


Figure 2.13: The relationship between river discharge and bacterial concentration (WHO, 2000)

All rivers demonstrate a close correlation between flow and bacterial concentration due to the increased supply of bacteria from watershed surfaces and some point sources (e.g. CSOs) during rainfall events (Figure 2.13).

The two curves represent hypothetical examples. In reality, all rivers will exhibit individual relationships depending on their hydrological characteristics and bacterial sources. The shape of the flow relationship will vary between different catchments and may also break

down during prolonged high flows, if the store of organisms in the bed-sediments (or the catchments surface) is exhausted. This phenomenon, however, has only been documented for small streams dominated by diffuse inputs and is less likely to occur for major rivers with multiple point and non-point sources. The processes controlling the transport and fate of bacteria in watersheds are now well understood and river water bacteria concentrations can be modelled and predicted.

2.11 Source of Bacteria: Human or Animal?

Bacterial indicator organisms such as faecal coliforms have been used to test water samples for faecal pollution, but such indicators do not provide specific information on the particular source of pollution. These bacteria may be found in a variety of warm-blooded animals and are not unique to the human intestinal flora. Information on the human or animal origin of faecal pollution gives an indication of the types of pathogens that may be expected, or the risk of infection, and the treatment that may be required to control the transmission of disease.

One of the major problems in environmental monitoring of indicator bacteria is to determine whether the source of pollution is human or animal. Since the early 1900s there have been various attempts to develop methods that differentiate the source of faecal pollution. Traditionally, efforts have concentrated on determining faecal pollution of human origin. It is now also important to distinguish between animal sources of faecal pollution compared to as human source, since animals can carry potentially harmful human pathogens. If animals are the source of indicator organisms, then the control measures and management practices will be different.

There are microbiological and chemical approaches for identifying sources of faecal contamination. Microbiological approaches cover bacterial and viral indicators found in the intestines of warm-blooded animals. Chemical approaches cover natural by-products of human metabolism or human activity. Microbiological approaches include the measurement of the ratio of faecal coliforms to faecal streptococci, or total coliforms, the detection of bacteriophages of bacteroides fragilis HSP40 and some serotypes of specific RNA coliphages, antibiotic resistance analysis, ribotype analysis, rep-PCR DNA technique, and use of human enteric viruses. Chemical approaches include faecal sterol fingerprinting

techniques and the presence of contaminants normally associated with sewage, such as detergents.

2.11.1 The Ratio of Faecal Coliforms (FC) to Faecal Streptococci (FS)

Human faecal material may be distinguishable from animal faecal material using an old method, the ratio of faecal coliforms to faecal streptococci (FC/FS). Faecal streptococci have received widespread acceptance as useful indicators of faecal pollution in natural aquatic ecosystem. Faecal streptococci are more abundant in animal faeces than in humans; in contrast, faecal coliforms are more abundant in human faeces than in animals (see Table 2.2 which is adapted and simplified from Gledrich 1978 and Pitt 1998). Therefore, the faecal coliform to faecal streptococci ratio has been used to differentiate human faecal contamination from that of other warm-blooded animals.

Table 2.2: Bacterial densities in warm-blooded animal faeces (Gledrich 1978 and Pitt 1998)

Source	Faecal Coliform (Avg. no. × 10^6 /gm wet weight)	Faecal Streptococci (Avg. no. × 10^6 /gm wet weight)	FC/FS Ratio
Human	13	3	4.33
Cats	7.9	27	0.29
Dogs	23	980	0.02
Cows	2.3	13	0.02
Sheep	16	38	0.42
Pig	3.3	84	0.03
Horse	0.126	6.3	0.02

A ratio of faecal coliform (FC) to faecal streptococci (FS) concentrations of four or greater is considered human faecal contamination, whereas a ratio of less than 0.7 suggests non-human sources (Edwards et al. 1997). Feachem (1975) suggested that if a series of FC and FS concentrations are obtained through time, an improved estimate of the pollution sources could be obtained. A predominantly human source should exhibit an initially high (>4) ratio which should then fall, whereas a non-human source should exhibit an initially low ratio (0.7) which should subsequently rise (Table 2.3).

Many attempts have been made to use this ratio (i.e. FC:FS) to determine the source of faecal bacteria. For example, Jagals et al. (1995) showed that the ratio of faecal coliforms to

faecal streptococci was close to unity in streams and rivers, which were upstream of urban settlements, and were exposed to faecal pollution predominantly of animal origin. However, downstream of urban settlements, which were exposed predominantly to human faecal pollution, the ratio increased to 3.5 to 4.7. Coyne and Howell (1994) measured the FC/FS ration for two watersheds typical of agricultural land use in Kentucky. They concluded that the FC/FS ratio suggested the probable source of faecal contamination, but they considered their conclusions to be tentative.

Table 2.3: Summary of faecal source related to FC:FS ratios (Feachem, 1975)

Initial FC\FS ratio	**Change through time of FC\FS**	**Probable faecal Source**
> 4	Rise	Uncertain
	Fall	Human
< 0.7	Rise	Non-human
	Fall	Uncertain

The application of this method is now considered unreliable due to the variable survival rates of faecal streptococci species. Fujioka et al (1981) suggested that the significance of the FC/FS ratio established under freshwater conditions should not be extrapolated to include the marine environment (i.e. seawater) where the decay rate for FC and FS varies significantly. Furthermore, the ratio is affected by the methods for enumerating faecal streptococci and by disinfection of wastewater. This method is an inexpensive and moderately complicated laboratory procedure. However, the result of this method taken alone must be quite carefully evaluated. If the method were used with some other methods, such as the detection of bacteriophages, the result will be more reliable.

If this ratio were reliable it would an inexpensive and practical method. Therefore, to use this method to provide information on possible faecal pollution source we have to consider its limits:

a) Sampling needs to occur soon after waste contamination (within 24 hours if possible) because the faecal bacteria may die off at different rates;

b) It becomes difficult to distinguish faecal streptococci in waters from faecal streptococci that are naturally present in soil and water when fewer than 100 faecal streptococci/100ml are present, and

c) The water pH needs to be between 4 and 9 because faecal coliforms die off quicker than faecal streptococci in acid or alkaline water (Geldreich and Kenner,1969; Coyne and Howell,1994).

2.11.2 The Ratio of Faecal Coliforms (FC) to Total Coliforms (TC)

Faecal (thermotolerant) coliforms constitute a subset of total coliforms. These bacteria conform to all the criteria used to define total coliforms, but in addition they grow and ferment lactose with the production of gas and acid at 44.5 ± 0.2°C within the first 48 hours of incubation. The ratio of faecal coliforms to total coliforms (FC/TC) is used to show the percentage of the total coliforms comprising faecal coliforms, i.e., coming from the guts of warm-blooded animals. If the faecal to total coliforms ratio exceeds 0.1 (i.e. faecal coliforms comprise 10% or more of the total coliform group) then this suggests the presence of human faecal contamination.

Hiraishi et al., (1984) measured TC, FC, and BOD from the Tamagawa River and its tributaries in Tokyo. Geometric means of the faecal coliforms to total coliforms ratios ranged from 0.007 to 0.069 in streams, which were located on the upstream of human contamination sources, but downstream of human sources, the ratio ranged from 0.21 to 0.26. Noble et. al., (2000) measured the FC/TC in a regional survey of the microbiological water quality along the shoreline of the Southern California coast. This method illustrates the possibility of faecal pollution but this method is not suitable for distinguishing human from animal–derived faecal matter. One of the shortcomings of this method is the potential growth of faecal coliforms in soils in tropical areas. As a result, its application in tropical areas is questionable. However, the method should not be discarded for topical areas, since it may be useful in conjunction with other methods.

Among others methds, the most frequently used and a well-tested method is genetic fingerprinting. Promising methods on the horizon include techniques using PCR, multiple antibiotic resistance, and bacteriophages. However, it should be noted that there is no easy, low cost, method for differentiating between human and non-human sources of bacterial contamination. No single indicator or approach is likely to represent all the facets and issues associated with contamination of waterways with faecal matter. At present, the best hope of distinguishing faecal pollution of human and animal origin is an appropriate combination of indicators. Statistical analyses of appropriate groups of methods offer the best possibility of

identifying human sources. Unfortunately, relying on a combination of methods will probably require a longer period of analysis than relying on a single method

2.12 Modelling the Decay of Bacteria

The prediction of the evolution of microbiological pollution of seas, rivers and lakes, following the construction of sewage disposal outfalls, is one of the most pressing problems that environmental water engineers need to confront during the planning stage of such constructions. Enteric bacteria, and specifically faecal indicator bacteria, are typically used to measure the sanitary quality of water for recreational, industrial, agricultural and water supply purposes. The knowledge of enteric bacteria survival kinetics is very important for environmental scientists. This biological–biochemical phenomenon is affected by a large number of factors, as mentioned earlier. Extensive laboratory experiments and field investigations have been undertaken to discover the process of bacterial decay in water. These efforts have produced a body of general qualitative knowledge (which cannot, however, easily be used by planning engineers) and a number of empirical formulae for the quantitative description of decay kinetics.

Bacterial decay rates were assumed to follow a first order decay model according to Chick's Law, 1910 (Key et al., 2005):

$$\frac{\partial C}{\partial t} = -k_t C \tag{2.1}$$

where

C = the enterococci concentration (cfu /100 ml),

t = the time (minutes), and

k_t = the die off rate over time (/min).

This allows k_t to be represented by the slope of the line of best-fit from least-squares regression of $\log_{10}$ transformed enterococci concentration ($\log_{10}C$) plotted against time (t). Thus, T_{90} is calculated as (Key et al., 2005):

$$T_{90} = \frac{1}{k_t} \tag{2.2}$$

Equation 2.1 is the longest established one for bacterial die off. However workers in the field recognize that Equation 2.1 does not in fact represent the actual process of decay of enteric bacteria disposed in natural waters. Orlob (1956) provided a general view of the various curves and the corresponding equation, which can be used for the calculation of enteric bacteria survival in natural waters. The disadvantage shared by these equations is that each of them has different empirical constants, so in each actual case the determination of the mathematical values of the constants demands expensive field investigations. This is why Equation 2.1, which contains only one constant, has been widely used for the last thirty years. In their review of modelling enteric bacterial die-off, Crane and Moore (1985) stated that first-order decay has been used with "moderate success" to describe bacterial die-off.

In order to avoid the expense of field investigations, there has been an attempt to define the coefficient k as a function of temperature, salinity, etc (Lantrip, 1983; Mancini, 1978; Mitchell and Chamberlin, 1978). However, this research has not produced such conclusive results that can be recommended for use by planning engineers (Grace, 1978).

Mancini (1978) and Crane and Moore (1985) described three commonly observed patterns of coliform die-off: first-order decay; bacterial growth followed by first-order die-off; and a die-off rate that changes with time.

Darakas and Hadjianghelou (1997) (Darakas, 2000) showed that the survival curves of E. coli in double logarithmic representation consisted of two phases, the maintenance phase and the decay phase. The duration of the maintenance phase provided a time scale for the decay phase and was contained as a parameter in the equation, which described the decay phase. The curve of the decay phase in a double logarithmic representation is a simple exponential curve. It is accurately described by Equation 2.3.

$$\left.\begin{aligned} C &= C_o \quad for\ t < t_E \\ \log\frac{C}{C_o} &= -\left[\log\frac{t+1}{t_E+1}\right]^5 \quad for\ t \geq t_E \end{aligned}\right\} \tag{2.3}$$

where

C = E.Coli concentration for $t \geq t_E$

C_0 = E.Coli concentration for $t < t_E$

t = time

t_E = duration of the maintenance phase.

They have shown that this equation is applicable for a wide range of temperatures between 4 °C and 37°C. It can be applied in the field of the quantitative description of the decay kinetics if the temperature is known. The duration of the maintenance phase t_E, which is a parameter in Equation 2.3, also provide a time scale for the decay phase. The authors have presented the mean values of t_E for different temperatures in the Table 2.4.

Canale et al. (1993) proposed a linear relationship between temperature and the death rate coefficient for modelling total coliform bacteria in Grand Traverse Bay, Lake Michigan. Laboratory studies of coliform death rates in Grand Traverse Bay (Gannon and Meier, unpublished data), used in developing this relationship, showed no difference in the death rate coefficient at 5, 10 and 15°C. The death rate coefficient increased in field measurements for temperature of 19°C, but this may have been due to the bactericidal effects of light (Auer and Niehaus, 1993).

Table 2.4: Maintenance phase duration (t_E mean values)

	Temperature (°C)				
	4	10	20	30	37
t_E (mean, days)	4.4	13.6	7.0	2.9	0.5

Work done by Thomann and Mueller, 1987 (Auer and Niehaus, 1993) provided the basis for the development of an overall kinetic coefficient. They proposed the following equation for an overall kinetic coefficient

$$k = k_d + k_i + k_s \tag{2.4}$$

Where

k_d = rate coefficient for death for dark conditions; including the effects of temperature, salinity, predation, etc (d^{-1})

k_i = rate coefficient for death as mediated by irradiance (d^{-1}), and

k_s = rate coefficient for sedimentation loss (d^{-1}).

Snedecor (2003) expanded and modified the terms used herein to include a nutrient component and a salinity component. They divided the dark death coefficient (k_d) proposed by Thomann and Mueller (1987) into two parts; k_{dt} was defined as the rate of coefficient in the dark (includes the effects of temperature) and a new coefficient term k_{ds} was added to reflect the rate coefficient for death in the dark and saline condition. An additional term, k_n, was included to reflect the impacts of potential growth due to water nutrient concentration. Thus the overall kinetic equation becomes:

$$k = k_{dt} + k_{ds} + k_i + k_s + k_n \tag{2.5}$$

Harris et. al (2002) used the constant decay rates (T_{90}) for day (6am to 6pm) and night (6pm to 6am) as 30 hours and 100 hours respectively. Bellair (1977) undertook a series of experiments that commenced at 6.30 am. Initially the rate of die-off was small, but it was found to increase rapidly, reaching a maximum around noon. Recorded values ranged from 19h to 40h. This implies that inactivation rates also vary greatly, particularly over a diurnal cycle, with the rate of die-off at any period of the year being approximately proportional to the intensity of irradiance received by the sample (Gameson and Saxon 1967). Their experience revealed that the effect of sunlight on the mortality of faecal coliform die-off was related to the solar irradiance by a power law as given by

$$k_s = \alpha I^{\beta} \tag{2.6}$$

Where

k_s = the die-off or decay rate due to sunlight (day^{-1});

I = the irradiance (W/m^2), a is a constant of proportionality, and

β = the slope of the $\log_{10}$ plot of die-off against irradiance, I.

α = the proportionality constant

The degree of penetration of sunlight into the water column has a significant effect on the bacterial die-off beneath the water surface. The turbidity of the water interferes with the light penetration through the water column and thus affects the bactericidal effectiveness of sunlight. Therefore, in more turbid waters, the bacteria survival time is increased, mainly because of the decreased effect of UV light, which is partially adsorbed by the suspended matter. The penetration, or conversely the extinction, of incoming solar radiation is usually described by introducing the extinction coefficient. This is proportional to the water depth, and may be calculated from solar radiation measurements taken at a range of water depths. Thus is represented by the Lambert (or Beer–Lambert) law, given as:

$$I_z = I_o e^{-k_e z} \tag{2.7}$$

Where

I_0 = the irradiance (solar intensity) at the surface (W/m^2),

I_z = the irradiance at depth z (W/m^2),

z = the depth (m), and

k_e = the vertical light extinction or attenuation coefficient (m^{-1}).

Gameson and Gould (1975) (Harris 2002) reported that the effect of sunlight on coliform die-off was additive and independent of temperature; die-off was expressed as the sum of die-off for darkness, k_d, and die-off due to sunlight, k_s. Assuming that the total faecal coliform mortality rate, k, can be defined by a simple relationship, taking into account disappearance for dark and light mortality conditions, then

$$k = k_d + k_s \tag{2.8}$$

Where

k = the total die-off rate (d^{-1}),

k_d = the die-off rate in darkness (d^{-1}), and

k_s = the die-off rate due to sunlight (d^{-1}).

Mitchel and Chemberlin (1975) suggested that the die-off rate is related to depth, with an effective attenuation coefficient of about $0.22m^{-1}$; the rate proportional to light intensity; and die-off is essentially first order with respect to coliform concentration. The relationship could be formulated as:

$$\frac{dC}{dt} = -kI_o e^{-\alpha z} C \qquad (2.9)$$

Where

C = the concentration of coliform bacteria at time t and depth z,

k = proportionality co-efficient,

I_0 = is the light intensity just below the water surface which is generally a function of time and latitude, and

α = the effective attenuation coefficient.

Thomann and Mueller (1987) suggested that the principal components of the net decay rate can be written as

$$k = k_b + k_l + k_s - k_a \qquad (2.10)$$

Where

k_b= basic rate as a function of temperature, salinity, predation,

k_l=death rate due to sunlight,

k_s= net loss (or gain) due to the settling (or resuspension) and

k_a= after growth rate.

They have also compiled a table of reported overall decay rates for bacteria and viruses, based on data and results reported in the literature.

They suggested that the temperature effects are corrected according to the Streeter-Phelps formulation:

$$k_T = k_{20}\theta^{(T-20)} \tag{2.11}$$

Where

k_T = value of rate constant at local water temperature

k_{20} = value of rate constant at standard temperature (i.e. 20°C)

T = local water temperature

θ = empirical constant for bacterial decay, which is 1.07 according to them.

For the effect of sunlight they referred to the Gameson and Gould (1975) relationship, giving

$$k_I(t) = \alpha I_o(t) \tag{2.12}$$

Where

$k_I(t)$ =decay rate at surface

α= proportionality constant (from the data of Gameson and Gold (1975) , α= 1) and

$I_o(t)$ = surface solar radiation cal/cm^2 hr

They also showed that the depth averaged sunlight decay rate was given as

$$K = \frac{\alpha I_o(t)}{HK_e}\left[1 - \exp(-K_e H)\right] \tag{2.13}$$

Where H is the depth in metre over which the average is taken and K_e (m^{-1}) is the vertical light extinction co-efficient.

Mancini,1978 (Thomann and Mueller, 1987) has evaluated the available data to incorporate salinity, temperature and solar radiation. On the basis of Mancini's work, and the depth averaged solar effects, Thomann and Mueller (1987), have deduced that:

$$K = \left[0.8 + 0.006(\%seawater)\right]1.07^{(T-20)} + \frac{\alpha I_o(t)}{HK_e}\left[1 - \exp(-K_e H)\right] + \frac{v_s}{H} \qquad (2.14)$$

where v_s is the net loss rate of the particulate bacterial forms (in m/day), which can be positive, zero or negative depending on the degree of resuspension.

From the above discussion, two representations can be utilized to model bacteria die-off. The first and most simple model, uses the overall net loss rate K as the measure of bacterial kinetics and no attempt is made to describe the individual mechanism or kinetic structure. At most K is considered as a function of temperature. This simple model recognises that there may be considerable uncertainty in the input loads in certain problem contexts and that it is really not practical or meaningful to describe the decay kinetics in greater detail. The second level incorporates some principal kinetics discussed above. The increasing complexity of the formulation for K is worthwhile for situations where the input loads are known with some degree of confidence.

2.13 Summary

This chapter has outlined the occurrence of pathogens in the environment, the concept of the indicator organism and current and revised legislation. The findings of a comprehensive literature review to investigate bacterial die-off in surface waters are summarised, with important relationships being quoted to develop the link between environmental variables and faecal indicator organism decay rates. This has led to the development of a series of functions, which were found to assist in the determination of suitable T_{90} values for use in hydro-environmental models. Studies were reviewed of health risk assessment and the transmission potential for human populations. Overall, this chapter establishes the degree of complexity and uncertainty related to the modelling of bacterial decay, which justifies the introduction of data driven modelling approaches in this key field of water management.

CHAPTER 3

HYDRO-ENVIRONMENTAL MODELLING

3.1 Introduction

During the past few decades, significant progress has been made in research with regard to the environmental impact on the biosphere and related anthropogenic activities. The cause and effect relationships between pollutant sources and degrading quality of the environment (both air and water) are better understood through research with the use of mathematical models.

In order to investigate the fate of pollutants once discharged into a water body it is necessary to understand the physical processes that control the movement of the solute within the receiving waters. In particular, how it becomes diluted, dispersed and advected from the point of discharge.

This chapter examines the hydrodynamic processes which cause pollutants and natural substances to be mixed in the receiving waters and transported with the flow. The numerical modelling process proceeds by describing the physical system with a set of equations and conservation laws acting upon it. The set of numerical operations transform the description of the system at one time to a description at a later time, providing a prognosis of the projected variable or property. The solute dispersion model is able to assist in the management strategy, including for example the design, location of discharges and operational options of a water body, for a range of bathymetric, hydrodynamic and meteorological conditions. The same numerical model may be used for different estuaries by altering the bathymetry, boundary conditions and other parameters.

3.2 Governing Equations for Hydrodynamic Process

Numerical modelling of flow is based on the principal of the conservation of laws mass and momentum within the body of fluid to be studied. In many cases, the flow is defined by the Reynolds equations, which describe the three dimensional turbulent motion of an incompressible fluid. For flows which show little variation in the vertical direction, it is appropriate to integrate these equations over depth of water, resulting in the simplified 'two-dimensional depth-averaged' equations of motion. When integrated over the depth, the equations governing fluid motion are as follows (Falconer, 1994, Falconer et al.,1999):

Conservation of mass:

$$\frac{\partial \eta}{\partial t} + \frac{\partial p}{\partial x} + \frac{\partial q}{\partial y} = q_m \tag{3.1}$$

Conservation of momentum:

$$\frac{\partial p}{\partial t} + \frac{\partial \beta pU}{\partial x} + \frac{\partial \beta pV}{\partial y} = fq - gH\frac{\partial \eta}{\partial x} + \frac{\rho_a}{\rho} C_w W_x \sqrt{W_x^2 + W_y^2} - \frac{gp\sqrt{p^2+q^2}}{H^2C^2} + \varepsilon\left[2\frac{\partial^2 p}{\partial x^2} + \frac{\partial^2 p}{\partial y^2} + \frac{\partial^2 q}{\partial x \partial y}\right] \tag{3.2}$$

$$\frac{\partial q}{\partial t} + \frac{\partial \beta qU}{\partial x} + \frac{\partial \beta qV}{\partial y} = -fp - gH\frac{\partial \eta}{\partial y} + \frac{\rho_a}{\rho} C_w W_y \sqrt{W_x^2 + W_y^2} - \frac{gq\sqrt{p^2+q^2}}{H^2C^2} + \varepsilon\left[\frac{\partial^2 q}{\partial x^2} + 2\frac{\partial^2 q}{\partial y^2} + \frac{\partial^2 p}{\partial x \partial y}\right] \tag{3.3}$$

where,

p (=UH), q(=VH) discharges per unit width in the x and y directions

respectively ($m^3/s/m$)

q_m source discharge per unit horizontal area ($m^3/s/ m^2$)

U,V depth averaged velocity components in the x and y directions respectively (m/s) defined as:

$$U = \frac{1}{H}\int_{-h}^{\eta} u dz, \quad V = \frac{1}{H}\int_{-h}^{\eta} v dz \tag{3.4}$$

β momentum correction factor for a non-uniform vertical velocity profile.

f Coriolis parameter due to earth's rotation $f = 2\omega Sin\varphi$, with ω= angular rotation speed of the earth and φ= geographical angle of latitude; $\omega = 2\pi/(24 \times 3600)$ $= 7.27 \times 10^{-5}$ radians/s, see Martin and McCutcheon(1999) and Kundu (1990),

g gravitational acceleration (=9.806 m/s^2)

H total water depth = (η+h) - see Figure (3.1),

η water surface elevation above datum see Figure (3.1),

ρ_a density of air ($\cong$ 1.292 kg/m^3),

ρ density of fluid (kg/m^3)

C Chezy roughness coefficient ($m^{1/2}/s$)

C_w air/fluid resistance coefficient (assumed to be 2.6×10^{-3} Falconer, 1991)

ε depth averaged turbulent eddy viscosity (m^2/s)

Further details of the derivation of continuity and momentum equations can be found in Kundu (1990), Versteeg and Malalasekara (1995) and Falconer (1994).

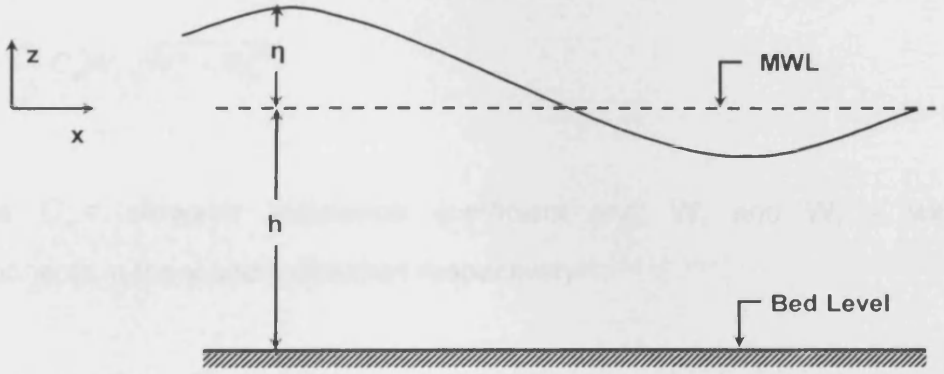


Figure 3.1: Co-ordinate system for depth integrated equations

3.3 Momentum Correction Factor

For an assumed logarithmic vertical velocity profile, the momentum correction factor can be calculated using:

$$\beta = 1 + \frac{g}{C^2 \kappa^2} \tag{3.5}$$

Where κ = von Karman constant = 0.4. For a velocity profile defined by the seventh power law, the value of β=1.016 and it is 1.20 for a quadratic velocity profile (Falconer and Chen, 1991).

3.3.1 Wind Effects

Wind exerts a drag force as it blows over the water surface. The shear stress at the air-water interface is calculated by assuming that it is proportional to the square of the wind speed at a particular height above the water surface. Various empirical formulae have been proposed to calculate the surface-water resistance coefficient, similar to that used to estimate for the drag coefficient in a turbulent flow field.

For the surface shear stress due to wind action, resolving forces horizontally for steady uniform flow gives for the x-direction:

$$\frac{\tau_{xw}}{\rho} = \frac{\rho_a}{\rho} C_w W_x \sqrt{W_x^2 + W_y^2} \tag{3.6}$$

where C_w = air-water resistance coefficient and, W_x and W_y = wind velocity components in the x and y direction respectively.

For water bodies with a strong current, such as occurring in estuaries and rivers, then the wind stress is often small compared to the bottom shear stress. In contrast wind generally plays a prominent role in the open sea and in lakes (Falconer et al., 2001).

3.3.2 Bottom Friction

In most coastal, estuarine and river flow studies the bed shear stress is represented in the form of a quadratic friction law, based on a relationship derived for steady uniform open channel (Henderson, 1966). Thus, in the x-direction the bed shear stress can be written as:

$$\frac{\tau_b}{\rho} = \frac{gp\sqrt{p^2+q^2}}{H^2C^2} \tag{3.7}$$

Bottom friction has a non-linear, retarding effect on the flow. The Chezy coefficient is a semi-empirical bottom friction coefficient, which was originally derived from a uniform flow condition in open channels. Under a rough turbulent flow condition and a logarithmic velocity profile, the Chezy bottom friction coefficient is assumed to be independent of the Reynolds number and varies only with the relative roughness of the bed and can be defined as follows (Henderson, 1966):

$$C = \sqrt{\frac{8g}{f}} = -2\sqrt{8g}\log_{10}\left(\frac{k_s}{12.0H}\right) \tag{3.8}$$

Where k_s = Nikuradse equivalent sand grain roughness, f = Darcy-Weisbach resistance coefficient.

Under transitional flow conditions, i.e. the Chezy coefficient varies with the Reynolds number. The corresponding Chezy coefficient can be obtained by using a slightly modified Colebrook-White equation of the form:

$$C = -2\sqrt{8g}\log_{10}\left(\frac{k_s}{12.0H} + \frac{2.5}{\sqrt{8g}\,\mathrm{Re}}C\right) \tag{3.9}$$

Where, R_e = Reynolds number $\left(R_e = \frac{4\left(\sqrt{U^2+V^2}\right)H}{\nu}\right)$, and ν = kinematic viscosity of fluid.

3.3.3 Turbulence

The turbulent shear stress refers to the flow resistance associated with random fluctuations in the fluid with regard to space and time. The momentum exchange brought about by turbulence causes the vertical velocity distribution to be more uniform

than under laminar flow conditions. The turbulence model in this study applies Boussinesq's approximation for the mean shear stress τ_o in turbulent flow:

$$\tau_o = \varepsilon \frac{dv}{dy} \tag{3.10}$$

Where ε = eddy viscosity, which is dependent on the turbulent characteristics of the flow and may be several orders of magnitude greater than the molecular viscosity (Falconer et al., 1999).

If the turbulent shear stress is dominated by bottom friction, a relationship between the Chezy coefficient and eddy viscosity exists. The depth averaged eddy viscosity may be calculated by using Fischer's approximation to give:

$$\varepsilon = C_e.U_*.H \tag{3.11}$$

Where U_* = bed shear velocity, as given as by:

$$U_* = \frac{\sqrt{g\left(U^2 + V^2\right)}}{C} \tag{3.12}$$

Substituting Equation (3.12) into Equation (3.11) gives the eddy viscosity as:

$$\varepsilon = C_e \frac{H}{C} \sqrt{g\left(U^2 + V^2\right)} \tag{3.13}$$

Where C_e = eddy viscosity coefficient, with Fischer's (1979) suggestion of the eddy viscosity co-efficient $C_e \approx 0.15$ based on laboratory data. Values of C_e are frequently found to be much larger for actual tidal flows in estuaries and coastal waters; in this study $C_e \approx 1.00$ has been used (Falconer et al., 2001)

3.4 Governing Equation for Solute Transport Processes

When a cloud of dissolved or suspended material is released into receiving waters, the cloud propagates, dilutes and spreads as it moves with the flow due to the effects of advection, diffusion and dispersive transport processes. The advection refers to the transport of the material by the flow current, such as the tidal current in estuarine and coastal waters. Diffusion includes the scattering of particles by molecular and turbulent motion. The dispersion, as distinct from diffusion, is the dilution process associated with the stretching out and distortion of a cloud of solute in a non-uniform flow by the effect of the velocity shear and the consequential averaging of the flow distribution over the depth for the two-dimensional models (Smith, 1992).

For a horizontal or quasi-horizontal flow, the three dimensional solute mass balance equations can be integrated over the water depth to give the two-`dimensional depth integrated advective-diffusion equation (see Bedford, 1994)

$$\frac{\partial H\phi}{\partial t}+\frac{\partial HU\phi}{\partial x}+\frac{\partial HV\phi}{\partial y}=\frac{\partial}{\partial x}\left[D_{xx}H\frac{\partial\phi}{\partial x}+D_{xy}H\frac{\partial\phi}{\partial y}\right]+\frac{\partial}{\partial y}\left[D_{yx}H\frac{\partial\phi}{\partial x}+D_{yy}H\frac{\partial\phi}{\partial y}\right]+HS_s \tag{3.14}$$

Where ϕ = depth averaged solute concentration (weight/volume) or temperature (°C), D_{xx}, D_{xy}, D_{yx}, D_{yy} = depth averaged dispersion-diffusion co-efficient in the x and y direction, respectively (m^2/s), which was shown (Preston, 1985; Holly, 1984) to be of the form:

$$D_{xx}=\frac{\left(k_l p^2+k_t q^2\right)\sqrt{g}}{C\sqrt{p^2+q^2}}=\frac{\left(k_l U^2+k_t V^2\right)H\sqrt{g}}{C\sqrt{U^2+V^2}} \tag{3.15}$$

$$D_{yx}=D_{yx}=\frac{\left(k_l-k_t\right)pq\sqrt{g}}{C\sqrt{p^2+q^2}} \tag{3.16}$$

$$D_{yy} = \frac{\left(k_l q^2 + k_t p^2\right)\sqrt{g}}{C\sqrt{p^2+q^2}} = \frac{\left(k_l V^2 + k_t U^2\right)H\sqrt{g}}{C\sqrt{U^2+V^2}} \tag{3.17}$$

In which k_l and k_t are the depth averaged longitudinal dispersion and lateral turbulent diffusion co-efficient respectively and where k_l= *5.93* and k_t= *0.23* according to Elder (1959).

S_s summarises all of the other sources and sinks of the solute. Sources and sinks include discharges from outfalls and rivers as well as chemical and biological transformations.

In coastal and estuarine flows the water depth H may vary rapidly, thus the monotonicity of the depth integrated concentration (ϕH) may be different from the monotonicity of the solute concentration (ϕ).Therefore, advective-diffusion equation (3.14) needs to be rearranged as given by Wu and Falconer (1998).

3.5 Numerical Methods

The most widely used numerical method to solve the governing equations, which have been widely applied to fluid flow and solute transport, are the finite difference method, the finite element method and the finite volume method. The finite difference approximation is the oldest method applied to obtain the numerical solution of the differential equations, and the first application was developed by Euler in 1768 (Hirsch, 1988). The finite difference method has been applied to two- and three-dimensional hydrodynamic and solute transport studies in estuarine and coastal waters by many researchers such as Yin et al. (2000), Zoppou et al. (2000), Shu and Chew (1998), Huang and Li (1997), Lin and Falconer (1997a, 1997b, 1995), Owens and Falconer (1987), Abbott and Basco (1989), Stelling et al. (1985) etc.

3.6 Finite Difference Method

Although the conservation of mass and momentum within any physical domain may be represented by the corresponding governing partial differential equations, these

equations can be written in many different finite difference forms. Consequentially, for any given set of differential equation it is desirable to be able to compare and contrast these finite difference schemes so that the most appropriate representation of the governing differential equations can be applied. Approximations to the derivatives are obtained to replace the individual terms in governing differential equations. The following figure provides a schematisation of the steps required, and some of the key terms used to ensure that the results obtained are in fact a solution of the original partial differential equations.

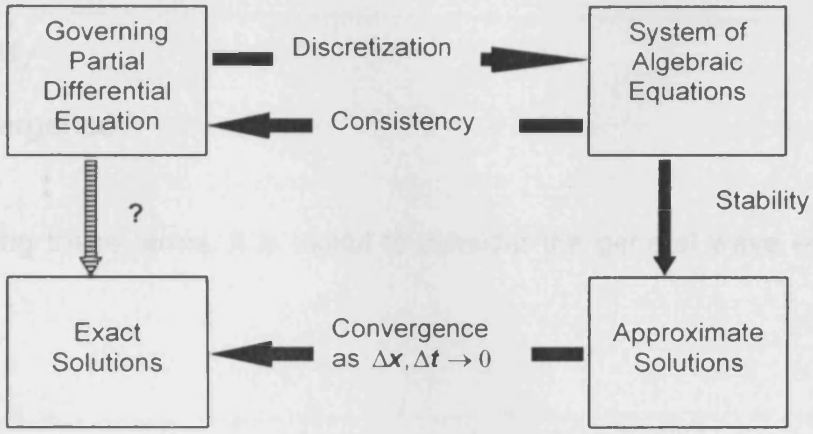


Figure 3.2: Overall procedure used to develop a CFD solution procedure

For the Finite difference representation of the Partial Differential Equations (PDEs) the approximations to the derivatives obtained above can be used to replace the individual terms in the partial differential equations. Figure 3.2 provides a schematic of the steps required, and some of the key terms used to ensure that the results obtained are in fact the solution of the original partial differential equation. Each of these new terms is defined below.

Accurate numerical methods for the partial differential equations require that the physical features of the PDE be reflected in the numerical solution algorithm. The selection of a particular finite difference approximation depends upon the physics of the problem being studied. In general, the type of PDE is crucial, and thus a determination

of the type, i.e. elliptic, hyperbolic, or parabolic, is extremely important. The mathematical type of the PDE must be used to construct the numerical scheme for approximating partial derivatives. Some advanced methods obscure the relationship, but it must still exist.

In Figure 3.2, several important terms have been introduced which require definition and discussion including:

- discretisation
- consistency
- stability
- convergence

Before defining these terms, it is useful to consider the general wave equation given by:

$$\frac{\partial u}{\partial t} = \alpha \frac{\partial^2 u}{\partial x^2} \tag{3.18}$$

This equation can be discretised using a forward difference in time and central difference in space formulations following the grid schematisation shown in Figure 3.3.

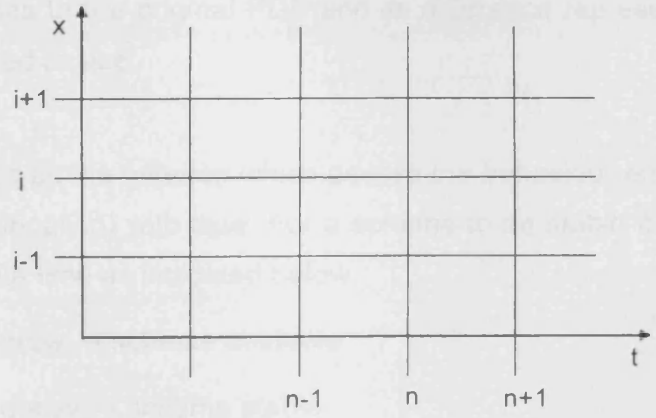


Figure 3.3: Grid nomenclatures for discritisation of wave equation

The differential and finite difference form of the wave equation can be written as:

$$\underbrace{\frac{\partial u}{\partial t} - a\frac{\partial^2 u}{\partial x^2}}_{PDE} = \underbrace{\frac{u_i^{n+1} - u_i^n}{\Delta t} - \frac{a}{(\Delta x)^2}\left(u_{i-1}^n - 2u_i^n + u_{i-1}^n\right)}_{FDE} + \underbrace{\left[-\frac{\partial^2 u}{\partial t^2}\bigg|_i^n \frac{\Delta t}{2} + a\frac{\partial^4 u}{\partial t^4}\bigg|_i^n \frac{(\Delta x)^2}{12} + \ldots\right]}_{Truncation\ Error} = 0 \tag{3.19}$$

Where the superscript denotes time and the subscript denotes spatial location. In Equation (3.19) the partial differential equation is converted to the related finite difference equation (FDE), giving a truncation of the form $O\left(\Delta t, \Delta x^2\right)$

Discretisation is the process by which finite difference approximations are used to replace derivatives with an approximation at a discrete set of points (i.e. the mesh). This introduces an error, due to the truncation error arising from the finite difference approximation and any errors due to the treatment of the boundary conditions.

Consistency is defined as the substantiation that the finite-difference representation of a PDE is converges to the original PDE and its difference representation vanishes as the mesh is reduced in size.

Stability is defined as the criterion which assess the behaviour errors from any source (e.g. round-off, truncation) with time. For a scheme to be stable it is necessary for the errors to decay with time as indicated below:

errors grow → scheme unstable

errors decay → scheme stable

and

- Stability is normally thought of as being associated with time marching problems.
- Stability requirements often dictate the allowable step sizes, such as for explicit schemes where the governing are that $\frac{\Delta t}{\Delta x}\sqrt{gh} \leq 1$.
- In many cases a stability analysis can be made to define the stability requirements.

Convergence is defined as where the solution of the FDEs approach the solution of the PDEs as the mesh is refined. In the case of a linear equation there is a theorem which proves that the numerical solution to the FDE is in fact the solution of the original partial differential equation provided that the scheme is both stable and consistent.

In practice, numerical experiments must be conducted to determine if the solution appears to converge with respect to mesh size. Machine capability and computing budget (time as well as money) dictate limits to the mesh size. Many results presented in the literature are not completely converged with respect to the mesh and grid independence is a key concern relating to many published papers.

For any finite difference scheme only a limited number of terms in the Taylor's series expansion can be included, therefore an approximation error, known as truncation error, will be introduced when a differential equation is approximated by a Taylor's series expansion. Since the hydrodynamic governing equations involve only first and second order derivatives, first and second order difference approximations of these derivatives will be shown here. For different approximations of the first and second order derivatives, various finite difference schemes can be found in Tannehill et al. (1997) and Abbott and Basco (1989).

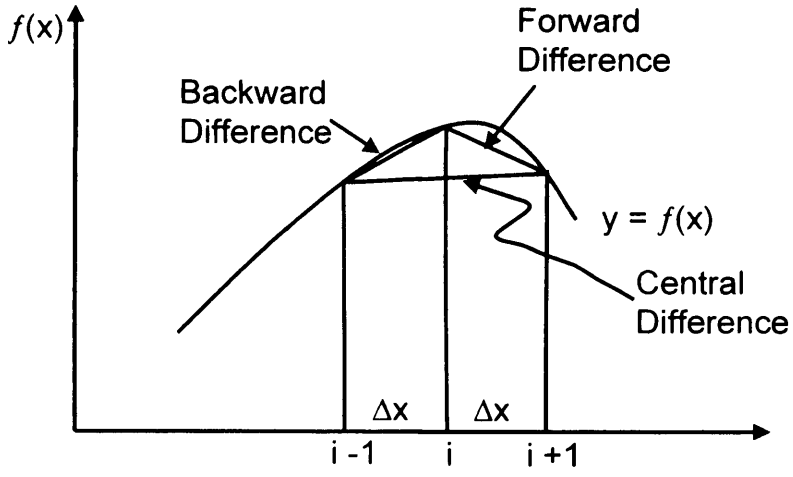


Figure 3.4 : Geometrical interpretation of difference formulae for first order derivatives (Hirsh, 1988)

3.6.1 Discritisation of the Governing Equations

Discritisation of the Continuity Equation

In the hydrodynamic model used in this study, an alternating direction implicit finite difference scheme is used to solve the governing equations, where for the first half time step the terms in the x-direction are treated implicitly and the terms in the y direction are treated explicitly. The continuity equation (Eq. 3.1) can therefore be discretised as follows:

$$\frac{\left(\eta_{i,j}^{n+\frac{1}{2}} - \eta_{i,j}^{n}\right)}{\frac{\Delta t}{2}} + \frac{\left(p_{i+\frac{1}{2},j}^{n+\frac{1}{2}} - p_{i-\frac{1}{2},j}^{n+\frac{1}{2}}\right)}{\Delta x} + \frac{\left(q_{i,j+\frac{1}{2}}^{n} - q_{i,j-\frac{1}{2}}^{n}\right)}{\Delta y} = q_m \qquad (3.20)$$

Where i, j = grid point location in the x- and y-directions respectively, subscripts n, n+1/2, n+1 represent variables evaluated at time levels $t = n\Delta t$, $t = (n+1/2)\Delta t$ and $t = (n+1)\Delta t$, respectively, where Δt represents the time-step for computations and n is the time step number. Using a square grid then, i.e. $\Delta x = \Delta y$, equation (3.20) can be rewritten as :

$$\eta_{i,j}^{n+\frac{1}{2}} = \eta_{i,j}^{n} - \frac{\Delta t}{2\Delta x}\left(p_{i+\frac{1}{2},j}^{n+\frac{1}{2}} - p_{i-\frac{1}{2},j}^{n+\frac{1}{2}} + q_{i,j+\frac{1}{2}}^{n} - q_{i,j-\frac{1}{2}}^{n} \right) + \frac{\Delta t.q_m}{2} \tag{3.21}$$

For the second half-time step, the terms in the y-direction are treated implicitly and the terms in the x-direction are treated explicitly. In this case, the continuity equation is discritised as:

$$\frac{\left(\eta_{i,j}^{n+1} - \eta_{i,j}^{n+\frac{1}{2}} \right)}{\frac{\Delta t}{2}} + \frac{\left(p_{i+\frac{1}{2},j}^{n+\frac{1}{2}} - p_{i-\frac{1}{2},j}^{n+\frac{1}{2}} \right)}{\Delta x} + \frac{\left(q_{i,j+\frac{1}{2}}^{n+1} - q_{i,j-\frac{1}{2}}^{n+1} \right)}{\Delta y} = q_m \tag{3.22}$$

and for a square grid cell Equation (3.22) reads:

$$\eta_{i,j}^{n+\frac{1}{2}} = \eta_{i,j}^{n} - \frac{\Delta t}{2\Delta x}\left(p_{i+\frac{1}{2},j}^{n+\frac{1}{2}} - p_{i-\frac{1}{2},j}^{n+\frac{1}{2}} + q_{i,j+\frac{1}{2}}^{n+1} - q_{i,j-\frac{1}{2}}^{n+1} \right) + \frac{\Delta t.q_m}{2} \tag{3.23}$$

Equations (3.20) to (3.23) are fully centred in both time and space over a whole time step, giving a second order accurate solution.

Discritisation of the momentum conservation equations

The momentum equation in the x-direction, i.e. Equation (3.2), can be written in the same manner as the continuity equation. Therefore, for the first half time step:

$$\frac{\left(p^{n+\frac{1}{2}}_{i+\frac{1}{2},J}-p^{n-\frac{1}{2}}_{i+\frac{1}{2},J}\right)}{\Delta t}+\beta\left[\frac{\left(\hat{U}\hat{p}\right)^{n}_{i+\frac{3}{2},j}-\left(\hat{U}\hat{p}\right)^{n}_{i-\frac{1}{2},j}}{2\Delta x}+\frac{\left(\overline{\hat{V}}\hat{\hat{p}}\right)^{n}_{i+\frac{1}{2},j+\frac{1}{2}}-\left(\hat{V}\hat{\hat{p}}\right)^{n}_{i+\frac{1}{2},j-\frac{1}{2}}}{\Delta y}\right]$$

$$=\bar{f}q^{n}_{i+\frac{1}{2},J}-\frac{g.H^{n}_{i+\frac{1}{2},j}}{2\Delta x}\left(\eta^{n+\frac{1}{2}}_{i+1,j}+\eta^{n-\frac{1}{2}}_{i+1,j}-\eta^{n+\frac{1}{2}}_{i,j}-\eta^{n-\frac{1}{2}}_{i,j}\right)$$

$$+\frac{\rho_a}{\rho}C_w W_x\sqrt{W_x^2+W_y^2}-\frac{g\left(p^{n+\frac{1}{2}}_{i+\frac{1}{2},J}-p^{n-\frac{1}{2}}_{i+\frac{1}{2},J}\right)\sqrt{\left(\hat{p}^{n}_{i+\frac{1}{2},j}\right)^2-\left(\bar{q}^{n}_{i+\frac{1}{2},j}\right)^2}}{2\left(H^{n}_{i+\frac{1}{2},j}.C^{n}_{i+\frac{1}{2},j}\right)^2} \qquad (3.24)$$

$$+\frac{\varepsilon.H^{n}_{i+\frac{1}{2},J}}{\Delta x^2}\left[\begin{array}{l}2\left(\hat{U}^{n}_{i+\frac{3}{2}}+\hat{U}^{n}_{i-\frac{1}{2}}\right)+\hat{U}^{n}_{i+\frac{1}{2},j+1}+\hat{U}^{n}_{i+\frac{1}{2},j-1}-6\hat{U}^{n}_{i+\frac{1}{2},j}\\+V^{n}_{i,j-\frac{1}{2}}-V^{n}_{i,j+\frac{1}{2}}-V^{n}_{i+1,j-\frac{1}{2}}+V^{n}_{i+1,j-\frac{1}{2}}\end{array}\right]$$

Where $\hat{U}$ denotes a term that is updated by iteration as:

$$\hat{U}^{n}=\begin{cases}U^{n-\frac{1}{2}} & \textit{for the first iteration,}\\ \frac{1}{2}\left(U^{n-\frac{1}{2}}+U^{n+\frac{1}{2}}\right) & \textit{for the } 2^{nd} \textit{ and remaining iteration}\end{cases} \qquad (3.25)$$

In Equation (3.24), $\overline{V}$ denotes a velocity value obtained by averaging the corresponding values at surrounding grid points giving:

$$\overline{V}^{n}_{i+\frac{1}{2},j+\frac{1}{2}}=\frac{1}{2}\left(V^{n}_{i,j+\frac{1}{2}}+V^{n}_{i+1,j+\frac{1}{2}}\right) \qquad (3.26)$$

and $\vec{P}$ denotes a value obtained from the upwind algorithm where

$$\dot{P}^{n}_{i+\frac{1}{2},j} = \begin{cases} P^{n}_{i+\frac{1}{2},j+1} & \text{if} \quad V^{n}_{i+\frac{1}{2},j} > 0, \\ P^{n}_{i+\frac{1}{2},j+1} & \text{if} \quad V^{n}_{i+\frac{1}{2},j} < 0 \end{cases} \tag{3.27}$$

Likewise, for the momentum equation in the y-direction, Equation (3.3) can be written for the second half time step as:

$$\begin{aligned} &\frac{\left(q^{n+1}_{i,j+\frac{1}{2}} - q^{n}_{i,j+\frac{1}{2}}\right)}{\Delta t} + \beta\left[\frac{\left(\hat{V}\hat{q}\right)^{n+\frac{1}{2}}_{i,j+\frac{3}{2}} - \left(\hat{V}\hat{q}\right)^{n+\frac{1}{2}}_{i,j-\frac{1}{2}}}{2\Delta y} + \frac{\left(\overline{U}\hat{\vec{q}}\right)^{n+\frac{1}{2}}_{i+\frac{1}{2},j+\frac{1}{2}} - \left(\overline{U}\hat{\vec{q}}\right)^{n+\frac{1}{2}}_{i-\frac{1}{2},j+\frac{1}{2}}}{\Delta x}\right] \\ &= -f\overline{p}^{n+\frac{1}{2}}_{i,j+\frac{1}{2}} - \frac{g.H^{n+\frac{1}{2}}_{i,j+\frac{1}{2}}}{2\Delta y}\left(\eta^{n+1}_{i,j+1} + \eta^{n}_{i,j+1} - \eta^{n+1}_{i,j} - \eta^{n}_{i,j}\right) \\ &+ \frac{\varepsilon.H^{n+\frac{1}{2}}_{i,j+\frac{1}{2}}}{\Delta x^2}\left[\begin{array}{l} 2\left(\hat{V}^{n+\frac{1}{2}}_{i,j+\frac{3}{2}} + \hat{V}^{n+\frac{1}{2}}_{i,j-\frac{1}{2}}\right) + \hat{V}^{n+\frac{1}{2}}_{i+1,j+\frac{1}{2}} + \hat{V}^{n+\frac{1}{2}}_{i-1,j+\frac{1}{2}} - 6\hat{V}^{n+\frac{1}{2}}_{i,j+\frac{1}{2}} \\ + U^{n+\frac{1}{2}}_{i-\frac{1}{2},j} - U^{n+\frac{1}{2}}_{i+\frac{1}{2},j} - U^{n+\frac{1}{2}}_{i-\frac{1}{2},j+1} + U^{n+\frac{1}{2}}_{i+\frac{1}{2},j+1} \end{array}\right] \end{aligned} \tag{3.28}$$

Where $\hat{V}, \overline{U}$ and $\hat{\vec{q}}$ have similar expression to those given in Equations (3.25) to (3.27) except that the current time level is $n+\frac{1}{2}$ instead of n.

To obtain the water elevation gradient for the first half-time step, the depth integrated continuity equation (3.21) and the momentum equation in the x-direction (3.24) are written for all grid nodes across the domain. For instance, Equation (3.24) is written at point $\left(i+\frac{1}{2}, j\right)$ while Equation (3.23) is centred at point (i, j). These two equations can be simplified to give:

$$a_{2i-1}U^{n+\frac{1}{2}}_{i-\frac{1}{2},j} + b_{2i-1}\eta^{n+\frac{1}{2}}_{i,j} + c_{2i-1}U^{n+\frac{1}{2}}_{i+\frac{1}{2},j} = d_{2i-1}$$
$$a_{2i}\eta^{n+\frac{1}{2}}_{i,j} + b_{2i}U^{n+\frac{1}{2}}_{i+\frac{1}{2},j} + C_{2i}\eta^{n+\frac{1}{2}}_{i+1,j} = d_{2i} \qquad (3.29)$$

Where $U^{n+\frac{1}{2}}$ and $\eta^{n+\frac{1}{2}}$ are the unknown velocity and water elevation variables while a, b, c and d are coefficients obtained through arrangement of equations. If there is velocity boundary at two ends, and there are total 'I' grid squares in the x-direction for the j^{th} row, then the number of unknown is (2I-1) in the whole domain for the same number of equations. The system of equations obtained can be expressed in a tri diagonal matrix as follows:

$$\begin{bmatrix} b_1 & c_1 & & & & & 0 \\ a_2 & b_2 & c_2 & & & & \\ & \ddots & \ddots & \ddots & & & \\ & & a_i & b_i & c_i & & \\ & & & \ddots & \ddots & \ddots & \\ & & & & a_{2I-2} & b_{2I-1} & c_{2I-2} \\ 0 & & & & & a_{2I-1} & b_{2I-1} \end{bmatrix} \cdot \begin{bmatrix} \eta_{1,j} \\ U_{\frac{1}{2},j} \\ \eta_{2,j} \\ \vdots \\ U_{I-\frac{1}{2},j} \\ \eta_{I,j} \end{bmatrix}^{n+\frac{1}{2}} = \begin{bmatrix} d_1 \\ d_2 \\ \vdots \\ d_i \\ \vdots \\ d_{2I-2} \\ d_{2I-1} \end{bmatrix} \qquad (3.30)$$

The system of equations is solved by the Thomas algorithm.

By solving the system of equations given in Equation (3.30), the velocity component $U^{n+\frac{1}{2}}_{i+\frac{1}{2},j}$ and the water elevation $\eta^{n+\frac{1}{2}}_{i,j}$ can be determined across the domain. Likewise, a similar system of equations is formed for the seconds half time step by discretising Equations (3.23) and (3.28), to solve for $\eta^{n+1}_{i,j}$ and $V^{n+1}_{i,j+\frac{1}{2}}$.

3.6.2 Alternating Direction Implicit (ADI)

In this study, a particular discretisation of the governing hydrodynamic equations which is based upon the Alternating Direction Implicit (ADI) technique. This is the best example of a splitting technique that was first applied by Peaceman and Rachford (1955) and Fletcher (1991). The technique was then generalised by Douglas and Gunn (1964).

For the ADI technique, each time step is split into two half time steps (see Figure 3.5). Thus a two dimensional problem can be solved by considering only one dimension implicitly for each half time step, without solving the two dimensional matrix. On the first half time step the water elevation (η) and the V velocity component (or the unit width discharge q) are solved implicitly in the y-direction, with the other variables being represented explicitly. With the boundary conditions included, the resulting finite difference equations for each half time steps are solved using the method of Gauss elimination and back substitution (or the Thomas Algorithm).

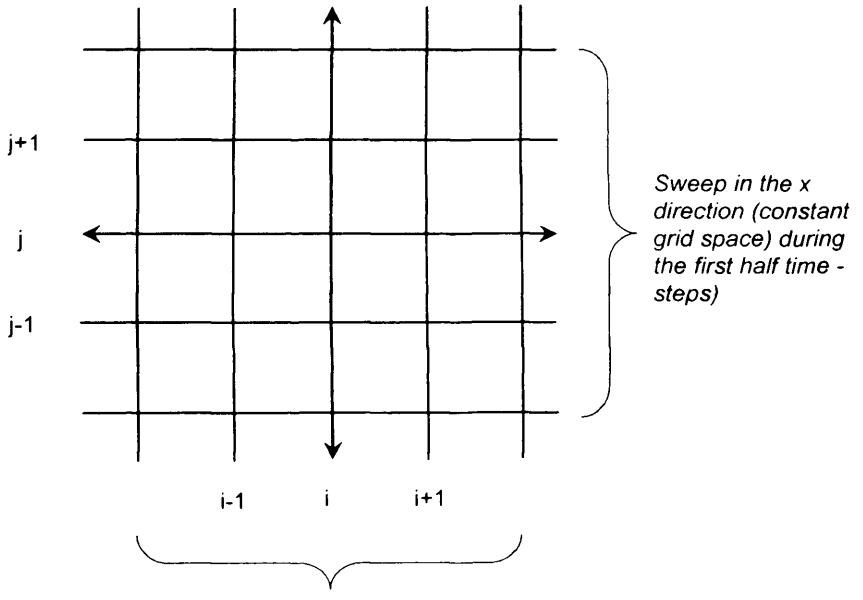


Figure 3.5: ADI implementation (Fletcher, 1991)

There are two families of solution techniques for linear algebraic equations: direct methods and indirect or iterative methods. Simple examples of direct methods are Cramer's matrix inversion method and Gaussian elimination. Jacobi and Gauss Seidel point by point methods are well known examples of iterative methods. These iterative methods are easy to implement in simple computer programs, but they can be slow to converge when the equation system is large. Hence they are not considered suitable for Computational Fluid Dynamic procedures. Thomas (1949) developed a technique for rapidly solving a tri-diagonal system that is now called the Thomas algorithm or the tri-diagonal matrix algorithm. It is computationally relatively efficient and has the advantage that it requires a minimum amount of storage (Versteeg and Malalasekara, 1995). In this study, the Thomas algorithm is used to solve the algebraic equations.

3.6.3 Staggered Grid System

In applying the finite difference method to solve the equations of mass, momentum and convective-diffusion for estuarine studies there are a number of advantages in not representing all the of variables η, U, V and S at the same grid points. The use of a space staggered system prevents the appearance of oscillatory solutions, which tends to arise for a collocated grid for space centred differences (Fletcher, 1991).

In the space staggered grid system, which is used in this study, the variables η (elevation) and ϕ (concentration) are located at the grid centre with the velocity components U, V or p, q (discharge per unit width in X and Y direction respectively) being located at the midpoint of the sides, as shown in Figure 3.6. The depths are specified directly at the velocity points so that twice as much bathymetric detail can be included in comparison with the traditional methods, further by locating the velocities and the depths at the same point then the fluxes of flow and concentration can be evaluated more accurately. Thus, this method allows bed topography to be represented more accurately, particularly for non linear bed variations and complicated bed profiles. However, in practice specification of bed topography is often restricted by the available data (Falconer et al., 2001).

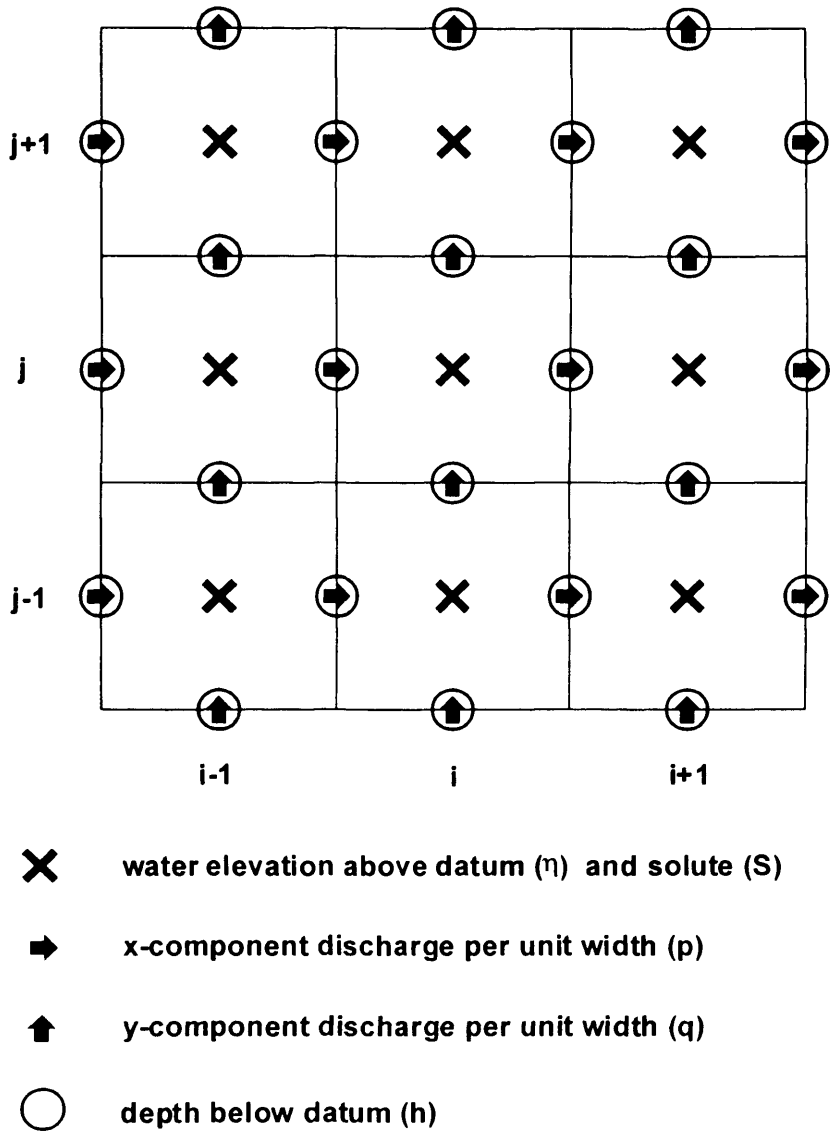


Figure 3.6: Computational Space Staggered Grid System

The ADI scheme used in this study is basically second order accurate, both in time and space and with no stability constraints due to the time centred implicit structure of the technique. However, it has been recognised that the time step needs to be restricted so that reasonable computational accuracy can be achieved (Chen, 1992). A maximum Courant number (C_f) was suggested by Stelling et al. (1985) as:

$$C_r = 2\Delta t\sqrt{gH\left(\frac{1}{\Delta x^2}+\frac{1}{\Delta y^2}\right)} \le 4\sqrt{2} \tag{3.31}$$

With the average depth being adopted for H. When 2-D solute transport equation is also solved for each half time step, then the choice of the time step should also consider the stability requirements for the solute transport equation.

3.7 Summary

The Numerical solution methods used in the model for this research study are reviewed briefly in this chapter. The governing equations were discretised using an appropriate numerical scheme for the hydrodynamic differential equations (i.e. continuity and momentum conservation). Boundary conditions for each case, solution procedures for the discreitsed equations and the interpolation technique for the velocity and sediment concentrations have been outlined.

Chapter 4

DATA DRIVEN MODELLING: GENETIC PROGRAMMING AND ARTIFICIAL NEURAL NETWORK

4.1 Data, Information and Knowledge

Data have been commonly seen a set of as simple facts that can be structured to become information. There are several variations of this widely adopted concept. The common idea is that data are something less than information and information is less than knowledge. Data are assumed to be simple isolated facts. When such facts are put into a context, and combined within a structure, then information emerges. When information is given a meaning by interpretation it, then information becomes knowledge.

Davenport and Prusak (1998) provided a comprehensive discussion of the differences between data, information and knowledge. They suggest that data are simply facts, records or transactions about some kind of event that has occurred; information adds value to the data by providing context and interpretation and encoding it into some kind of message.

Davenport and Prusak (1998) stated that:

> *Data is a set of discrete, objective facts about events...Data describes only a part of what happened; it provides no judgment or interpretation and no sustainable basis of action...Data says nothing about its own importance or relevance.*

According to the authors, however, data turn into information as soon as it is given meaning. Information must inform, as outlined:

> *it's data that makes a difference...Unlike data, information has meaning ...Data becomes information when its creator adds meaning*

Wurman (1989) viewed the hierarchy as

> *Data are facts and figures that have no inherent meaning, Information is data to which people assign meaning and Knowledge is data that people can apply in their lives.*

Information is "data endowed with relevance and purpose" (Druker, 1995), or data that make a difference (King, 1993). Bourdeau and Couillard (1999) see information as the result of analysing and interpreting data-phrases or images that carry a meaning. Thus information is normally associated with meaning. Knowledge is information made actionable in a way that adds value to the enterprise (Vail, 1999).

Spiegler (2000) views the chain as "yesterdays data are today's information and tomorrow's knowledge..... If data becomes information when they add *value* in some way, then information becomes knowledge when it adds *insight*, abstractive value, better understanding."

Knowledge Discovery denotes the overall process of extracting high-level knowledge from low-level data, however the terms Data Mining and Knowledge Discovery are often used interchangeably. The rapidly emerging field of knowledge discovery has grown significantly in the past few years. This growth is driven by a mix of daunting practical needs and strong research interests. The technology for computing and storage has enabled people to collect and store information from a wide range of sources at rates that were, only a few years ago, considered unimaginable.

This chapter establishes the notion of data based modelling in context and then describes briefly two such types of data-based models namely: Genetic Programming (GP) models and Artificial Neural Networks (ANNs), which have been applied in various water management studies in this research project.

4.2 Modelling: Knowledge of Processes and Data

A *model* is a theoretical framework that represents a 'reality' with a set of variables and a set of logical and quantitative relationships between them. The objective of modelling is to explain or to predict that 'reality'. Models in this sense are constructed to enable reasoning within an idealized logical framework about these processes and are an important component of scientific theories. Modelling includes: studying the system, formulating/establishing the problem, experimentation/data collection, analysis of the experimental results/data, building the model, verifying the model with real life data. Traditionally , the term model is used in one of two senses (Solomatine 2002),

- A mathematical model based on the description of behaviour of a system or phenomenon under study.
- A physical or scaled model based on material components or objects.

Behavioural models based on mathematical descriptions are used widely. Traditional modelling of physical processes is often named *physically-based modelling* (or knowledge-driven modelling) since the model tries to explain the underlying processes. For this case the emphasis is on a theory, which demands that appropriate data be obtained through observation or experiment. In such an approach, the discovery process may be referred to as theory driven. Especially when a theory is expressed in mathematical form, theory-driven discovery may make extensive use of strong methods associated with mathematics and with the subject matter of the theory itself. An example of such a model is the hydrodynamic model DIVAST used in this research and based on the solution of the Navier-Stokes partial differential equations, solved numerically using finite-difference scheme, and used in this study.

Another approach is based on the analysis of data characterising the system under study. A model can then be developed on the basis of interactions between the system state variables. These models are referred to as *data-driven* models. Data-driven models possess the attractive property that they can be built up generically in the sense that no underlying physical, chemical laws etc. about the system variables need to be known. This concept relies on the data describing input and output characteristics, primarily employs Artificial Intelligence (AI) techniques and is based on a limited knowledge of the modelling process. Statistical models, such as linear

regression, follow the same approach. These methods, take a body of data as its starting point and search for a set of generalisations, or a theory, to describe the data parsimoniously or even to explain it. Usually such a theory takes the form of a precise mathematical statement of the relations existing among the data. Thus they are able to make abstractions and generalizations of the process and often play a complementary role in physically-based models (Keijzer and Babovic 2000).

Data-driven modelling uses results from such overlapping fields as data mining, artificial neural networks (ANNs), machine learning, evolutionary computations, rule-based type approaches, such as expert systems, fuzzy logic concepts, rule-induction and machine learning systems. Sometimes "hybrid models" are built which combine both types of these models. In this research study particular attention has been given to examining suitability of Genetic programming and Artificial Neural Networks approaches, as a supplement to the conventional behavioural models.

4.3 Model Induction from Data

One particular mode of data mining is that of model induction. Inferring models from data is a method of deducing a closed-form explanation based on observations. These observations, however, more often than not represent a limited source of information. The question emerges how such a limited flow of information from a physical system to the observer can result in the formation of a model that is complete, in the sense that it can account for the entire range of phenomena encountered within the physical system in question. The confidence in model performance can not be based on data alone, but might be achieved by grounding models in the domain so that appropriate semantic content is obtainable. This should be the ultimate goal of knowledge discovery.

The rapid advance in information processing systems in recent decades had directed engineering research towards the development of intelligent systems that can evolve models of natural phenomena automatically without any human intervention. In this respect, a wide range of machine learning techniques, like Decision Trees, Artificial Neural Networks, Bayesian methods, Fuzzy-Rule based systems and Evolutionary Algorithms, have been successfully applied to solve various problems, including a significant number of civil engineering and water resources modelling and management

issues. These techniques have also shown their potential as an alternative approach to conventional modelling.

The goal of learning from examples is to find the general rule that created the specific examples, and this is achieved by trying out different model topology and related parameters. Of the various possible methods for model induction from data Genetic Programming (GP) and Artificial Neural Networks (ANNs) are reviewed in the rest of this chapter and the implementation of these new data mining and knowledge discovery processes are then shown to produce a valid management tool in the following chapters.

4.4 Genetic Programming

Genetic programming became a popular branch of Evolutionary Algorithms in the early 1990s due primarily to the work by Koza (Koza 1992). Genetic Programming, as envisioned by Koza, does not process computer programs in the same way that human programmers would. There are no ASCII files, no countless data types that can be mixed, or no repulsive syntax with various special symbols that can be misplaced to produce a synthetically meaningless result. The standard genetic programming system operates using abstractions of computer programs. In the short time, since the publication of Koza's 1992 book, over eight hundred GP papers have been published (Banzhaf et al., 1998). Researchers have devised many different systems that may fairly be called genetic programming – system that use tree, linear and graph genomes; systems that use high crossover rates and mutation rates.

No exposition on Genetic Programming would be complete without some reference to the ideas which inspired them. The next section describes the very notion of evolutionary computation, before focusing on the details of Genetic Programming in the subsequent sections.

4.4.1 Evolutionary Computation

The principle of evolution is the primary unifying concept of biology, linking every organism together in a historical chain of events. Over many generations, random

variation and natural selection have shaped the behaviour of individuals and species to fit the demands of their surroundings. Whilst evolution itself has no intrinsic purpose, it is capable of engineering solutions to the problem of survival that are unique to the circumstance of each individual. Harnessing the evolutionary process within a computer provides means for solving complex engineering problems that traditional algorithms have been unable to solve. Indeed, the field of evolutionary computation is one of the fastest growing areas of computer science and engineering, enabling the solution to be found for many problems that were previously unsolvable.

Evolutionary algorithms mimic the process of natural evolution, the driving process for the emergence of complex and well-adapted organic structures. Evolutionary Algorithms (EAs) are engines simulating grossly simplified processes occurring in nature and implemented in artificial media – such as the computer.

Charles Darwin (1859) described a unifying view of the origin and further evolution of organisms in nature in his book 'The origin of species' based on the principal of natural selection. He stated (Banzhaf et al., 1998):

> *....if variation useful to any organic being do occur, assuredly individuals thus characterized will have the best chance of being preserved in the struggle for life; and from the strong principal of in heritance they will tend to produce offspring similarly characterized. This principle of preservation, I have called, for the sake of brevity, Natural Selection (C. Darwin, 1859)*

Using the same principle, Evolutionary Computation tackles difficult problems by evolving approximate solutions. Starting with a primordial diversity of random solutions, repeated variation and selection are applied to improve the accuracy of the solutions. The basic criterion for evolution to take place – either in biology or through computers – have been summarized by Maynerd-Smith (Maynerd-Smith 1975) as-

Criterions of Heredity The copying process is highly dependable and offspring are similar to their parents.

Criterion of Variability The copying process is not perfect and offspring are not exactly the same as their parents

Criterion of Fecundity A different number of off-spring resulted from different variants. Any specific variation has an effect on behaviour, and this behaviour has an effect on reproductive success.

Darwinian evolution uses the principles of competition, inheritance and variation within a population. These concepts are used to define a class of iterative improvement meta heuristic search methods. These methods, evolutionary algorithms, use a population of solutions and genetic operators to carry out searches. Specifically, the evolutionary algorithm employs the following items:

- A population of candidate solutions called individuals,
- A *fitness function* that evaluates and assigns each individual a score, or *fitness value*,
- Transformation operators that produce *offspring* individuals from *parent* individuals, implementing the concept of inheritance through stochastic variation, and
- A stochastic selection method for selecting individuals with better fitness to produce offspring.

The definition of the basic evolutionary algorithm is representation-free. It does not mention what form of solutions should be considered and, in effect, many representations are used in the field of evolutionary computation. Evolutionary algorithms are often categorised into four main branches (Babovic and Abbot 1997) that are mainly distinguishable by their commonly used representation and operators: *Genetic Algorithms* that use a bit-string and two-parent crossover, *Evolutionary Strategies* which use a real-valued vector and Gaussian mutation, *Evolutionary Programming* which employs a finite-state machine and mutation operators, and *Genetic Programming* which uses a computer program or executable structure and two-parent crossovers. Despite their differences, they share the same main ingredients as population of solutions, innovation operations, conservation operations, quality

differentials and selection. These classifications represent common, or initial, implementations. Many implementations use components from different branches and make the classifications less accurate.

4.4.2 Fundamentals of Genetic Programming

Langdon and Poli (2002) summarised the theoretical foundations of genetic programming. Theories of evolutionary algorithms use abstract representations of the solution space, called schemata, to describe various components and behaviour of the algorithm. Holland's (1975) notion of schema for genetic algorithms was extended by Koza (1992) to include syntax trees. Syntax trees are the common representation in genetic programming, and these schemas were intended to represent trees that have common subtrees. In this case, subtrees represent functional codes that are combined over time to create better programs (Altenberg, 1994; Rosca and Ballard, 1996). Rosca (1997) defined a similar schema based on rooted trees, also using wildcards to allow for schema to represent templates.

The use of this flexible coding system allows the algorithm to perform structural optimisation. This can be useful for the solution of many engineering problems. For example, GP may be used to perform symbolic regression. While conventional regression seeks to optimise the parameters for a pre-specified model structure, with symbolic regression, the model structure and parameters are determined simultaneously. Similarly, the evolution of control algorithms, scheduling programs, structural design and signal processing algorithms can be viewed as structural optimisation problems suitable for GP. The basic features of GP are described in this section, with more details given in Banzhaf et al. (1998).

4.4.2.1 Terminals and Function

A parse tree is composed of terminals and functions, together known as nodes. Functions are inner nodes, while terminals are leafs of the tree. The terminals and functions play different roles. In general terms, terminals provide a value to the system and functions process a value already in the system.

Input, constants and other zero-argument nodes are called terminals, or leafs, because they terminate a branch of a tree in tree based GP. In fact a terminal lies at the end of every branch of the tree structure. The notion is to use these terminals as inputs to the program, constant or function without argument. In either case, a terminal returns a numerical value without any need for take an input for itself. Another way of outlining this notion is that terminal nodes have an *arity* of zero. The arity of a node is the number of arguments it expects to receive.

The terminal set also includes constants. In typical tree based GP, a set of real numbered constants is chosen for the entire population at the beginning of the run. These constants do not change their value during the run. They are called *random ephemeral constant* and are represented by the symbol $\Re$. Other constants may be constructed within programs by combining random ephemeral constants using arithmetic functions. On the other hand, in linear GP systems, the constant portion of the terminal set is consists of a number chosen randomly out of a range of real, or integers, constants and these constants may also experience mutation.

The function comprises statements, operators, and functions available to the GP system (Banzhaf et al., 1998) The function set may be application specific and may be selected to fit the problem domain. The range of available functions may be broad, including boolean functions, arithmetic functions, transcendental functions and variable assignment functions. In fact, it may use any construct that is available to any programming language.

The functions and terminal sets used in a GP run should be powerful enough to represent the solution of a problem. A function set consists of too few operators (say, addition only) will not solve many difficult problems. On the other hand too large a function set makes search for a solution harder. So a parsimonious approach to choosing a function is often advised in the literature. A similar approach is also effective in choosing the constant.

4.4.2.2 Representation of Solutions

The original formulations of genetic programming considered Lisp S-expressions as candidate solutions. The Lisp programming language is popular in artificial intelligence

research, as it was designed for symbolic processing. S-expressions, or symbolic expressions, are the basic objects in Lisp and are naturally represented as syntax trees, where leaves represent terminals (variables or constants) and nodes represent functions. In genetic programming, to overcome typing related to passing arguments and function values to functions, it is standard practice to use a strongly-typed system (Montana, 1995), where variables and functions have the same type. Grammar-based genetic programming systems can easily use multiple types, but require a defined grammar and specialised operators (Whigham, 1995; Ryan et al., 1998). The evaluation of a syntax tree is performed as a depth-first walk of the tree. Before evaluating a node, each of its arguments must be evaluated first. Thus, one may think of the syntax tree evaluation algorithm as a recursive call on the root node of the tree, which in turn evaluates each of its children, typically from left to right. Function and terminal nodes return their values up the tree to their parent, where terminals can only return their value. In the domain of mathematical functions, the expression $((3/(1+2))+(-1))$ can be represented by an S-expression in prefix as $(+(/(3(+(12))))-1)$. Which can be represented as a tree as-

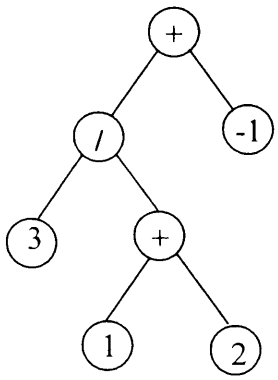


Figure 4.1: Tree representation

The representation of candidate solutions in genetic programming is not limited to single syntax trees. Auxiliary data structures, such as memory arrays, are also used (Spector and Luke, 1996), as are data structures containing several syntax trees to represent one solution (Luke, 1998). The encapsulation of functions has been accomplished with automatically defined functions (Koza, 1994) or automatically

defined macros (Spector, 1996). More recent advances have seen "architectural-altering" operators, loops and recursion (Koza et al., 1999).

4.4.2.3 Population and Initialisation

The first step of performing a GP run is to initialise the *population*. This means creating a variety of program structures for later evolution. The role of the population is to hold (the representation of) possible solutions. A population is a multi set of genotypes. The population forms the basic units of evolution. Individuals are static objects, not changing or adapting, it is the population that changes. Defining a population can be as simple as specifying how many individuals are present, provided representation exists. The *diversity* of a population is a measure of the number of different solutions present. No single measure of diversity exists. The number of different fitness values present, the number of different phenotypes present, or the number of different genotypes are among the popular measures for diversity.

Initialisation can be done in several ways. There are two different methods of initialising tree structures commonly in use. They are called *full* and *grow* methods (Koza, 1992). In the grow method, a primitive - be it a function or a terminal – is selected at random, and as long as there are unresolved subtrees, then the process is repeated. When a predefined depth or size limit is reached, then only terminals are chosen.

If the terminal and functions allowable to the program tree is selected as $T = \{a, b, c, d, e\}$ and $F = \{+, -, \times, /\}$, then one of the numerous possible trees could be as Figure 4.2.

In Figure 4.2 the branch that ends with input 'a' has a depth of only three as choosing terminals is random throughout initialisation. Tree initialised methods therefore are more likely to be of an irregular shape.

On the other hand in the Full method, functions are chosen only until a node is at maximum depth, and then it chooses the terminals only. As a result every branch of the tree goes to its maximum depth. Figure 4.3 has been initialised with the full method with a maximum depth of three.

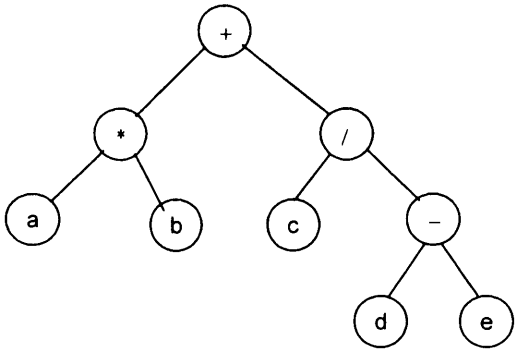


Figure 4.2: Tree of maximum depth four initialised with the grow method

The above methods have their drawbacks as they could result in a uniform set of structure in initial populations as the routine is same for all individuals. But diversity is valuable for GP populations. To prevent this Koza (1992) devised the *ramped half-n-half* method. This method probabilistically selects between two recursive tree generating methods: *Grow* and *Full*. It is intended to enhance the population diversity from outset. An overview of alternative tree initialization routines and an empirical composition between those can be found in Luke and Panait (2001).

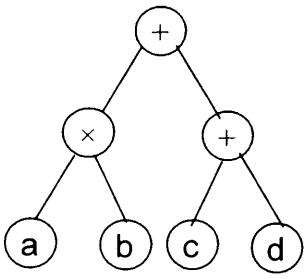


Figure 4.3: Tree of maximum depth three initialised with the full method

4.4.2.4 Genetic Operators

An initialised population usually has a very low fitness. Evolution proceeds by transforming the initial population by the use of genetic operators. The three principal genetic operators are:

a) Crossover;

b) Mutation; and

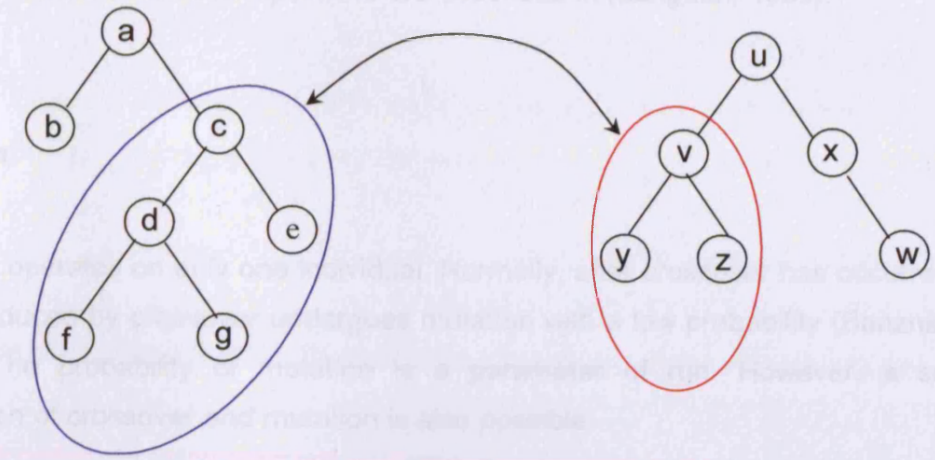
c) Reproduction

Two basic operators are described here.

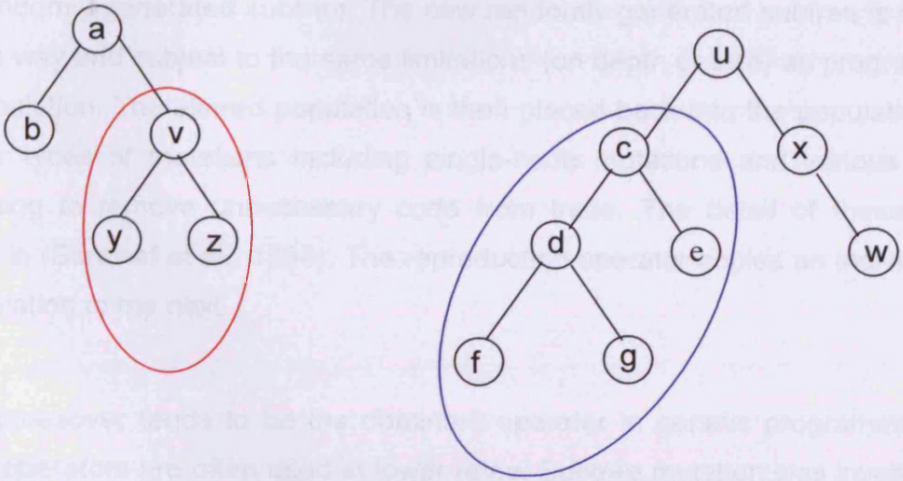
Crossover

The crossover operator combines the genetic material of two parents by swapping a part of one parent with a part of other. The mechanics of the subtree crossover method were initially described by Cramer (1985) and Koza (1992) and have been examined in detail in a variety of studies, e.g. (D'haeseleer, 1994; Luke and Spector,1998; Langdon, 2000; Gathercole and Ross, 1996). The algorithm can be described as follows: two trees are selected from the population, a subtree in each tree is selected and the two subtrees are exchanged between the trees. Either one or both children are considered for the new population.

Subtree selection is done by assigning a uniform probability to all internal nodes and leaf nodes separately. Then, an internal node selection probability, usually set to 0.9, defines the frequency of leaves or subtrees selected for recombination. It may be noticed that, as trees grow in size, the probability of selecting subtrees grows near the leaves. This is because there are more nodes in these locations, giving them a higher cumulative probability of being selected.



Before crossover



After crossover

Figure 4.4: The crossover operator acting on two parse trees (the branches to be copied from each are circled)

Since in canonical genetic programming all functions and terminals return and expect the same type, any exchange of subtrees between two trees will be valid. Many possible variations of recombination exist. For example, Homologous and Size Fair

crossover (Langdon, 2000) attempt to preserve tree structures and the size of exchanged subtrees. Other operators are described in (Langdon, 1998).

Mutation

Mutation operates on only one individual. Normally, after crossover has occurred, each child produced by crossover undergoes mutation with a low probability (Banzhaf et al., 1998). The probability of mutation is a parameter of run. However, a separate application of crossover and mutation is also possible.

When an individual is selected for mutation, one type of mutation operator in tree GP selects a point in the tree randomly and replaces the existing subtree at that point with a new randomly generated subtree. The new randomly generated subtree is created in the same way and subject to the same limitations (on depth or size) as programs in the initial population. The altered population is then placed back into the population. There are other types of mutations including single-node mutations and various forms of code-editing to remove unnecessary code from trees. The detail of these types is available in (Banzhaf et al., 1998). The reproduction operator copies an individual from one population to the next.

Subtree crossover tends to be the dominant operator in genetic programming, while mutation operators are often used at lower rates. Subtree mutation was investigated in comparison with subtree crossover in (Luke and Spector 1998).

4.4.2.5 Fitness and Selection

Selection is an essential process in EAs that removes individuals with a low fitness and drives the population towards better solutions. After the quality of an individual has been determined by applying a fitness function, a decision has to be made whether to apply genetic operators to that individual and whether to keep it in the population or allow it to be replaced. This task is called selection. The fitness function used in this study is the Root Mean Squared Error (RMSE) and Coefficient of Determination (CoD).

The mathematical expressions for these functions are given in Chapter 5 (see Equations 5.1 and 5.2)

Selection defines how the algorithm updates the population from one iteration to the next and is responsible for the speed of evolution. In general, selection either replaces the entire population or only a fraction of it. The former approach is used in generational EAs whereas the latter is employed in steady-state EAs. There are a few major differences between the two approaches, with the most common selection algorithms being described below.

Tournament selection

Tournament selection is not based on competition within the full generation but a subset of population. In each tournament, the process picks a number of individuals, called the tournament size and is selected randomly. A selective competition then takes place and the individuals with the better fitness are then allowed to replace those of the worse individuals. In the smallest possible tournament two individuals compete. The better of the two is allowed to reproduce with mutation. The result of this reproduction is returned to the population, replacing the loser of the tournament.

The selection pressure in tournament selection is dictated by the tournament size. A small tournament causes a low selection pressure, with a large tournament size causing a high pressure. The selection pressure can be lowered by introducing stochastic winners in tournaments with two individuals. Hence, the fittest individual wins with probability $p > 0.5$. Typical values are $p = 0.75$ or $p = 0.8$. Setting $p = 0.5$ is equivalent to random selection.

Tournament selection is easy to implement, produces good results within short time, requires very little computing time, is controlled by only a few parameters and above all does not require a centralised fitness comparison between all individuals. For these reasons, tournament selection is probably nowadays the most commonly used selection operator.

Proportional selection

Proportional selection assigns the probability of an individual's survival according to the fitness of the individual. The probability is calculated by dividing the fitness of the individual by the fitness sum of the whole population, i.e., an individual's chance of survival depends on its relative fitness to the other individuals. Each individual is assigned to a "slot" of the interval [0; 1] according to the individual's probability of survival. An individual is selected if a random number of the interval [0; 1] is within its slot. This selection method is often illustrated as a biased roulette wheel, where the interval slots correspond to the slots of a roulette wheel and the "winners" are copied to the next generation. The drawback of proportional selection is that the selection pressure depends on the relative fitness of the individuals, instead of a parameter such as tournament size. In proportional selection, a few very good individuals can quickly take over the entire population, because they dominate a large part of the roulette wheel and are therefore frequently copied when the next generation is formed. For this reason, proportional selection is not as popular as it used to be.

Ranking selection

Ranking selection is a variant of proportional selection that deals with the uncontrolled selection pressure. It is based on the fitness order, into which the individuals can be sorted. In ranking selection, the selective superiority of an individual is determined by a fixed probability of survival according to its fitness rank. The ranking is obtained by sorting the individuals according to their fitness. Each individual is then assigned a probability of survival, which is determined by the used ranking scheme. The selection is performed using the roulette wheel approach.

The difficult part of applying ranking selection is to determine a good probability of survival for each rank. A scheme that is too generous towards low-fit solutions might slow down the convergence, while a scheme favouring the best individuals might lead to a premature loss of genetic diversity.

Steady-state selection

Evolutionary algorithms that are based on steady-state selection, also known as steady-state EAs, update only a small fraction of the population in every iteration. The evolutionary operators create λ potential solutions from the parent population with size μ. The $(\lambda + \mu.)$ individuals are then sorted and λ individuals with the lowest fitness are

discarded. Common values are $\mu = 100$ and $\lambda = 15$. This approach is fundamentally different from tournament, proportional and ranking selection. In steady-state selection the populations are overlapping and all the surviving individuals are deterministically selected, which is only the case for the elite individuals in the other three selection techniques. The steady-state selection method is that used for the GP adopted in this thesis in most of the cases.

Manual selection

In some applications the quality of a solution is based on a subjective evaluation of issues that are hard or impossible to capture mathematically; for instance, the beauty of a design. Instead, the selection process can be handled by a human operator. The algorithm displays the current solutions and asks the operator to select a subset of the presented solution. The selected solutions are then used to create a new population and the process is repeated. Examples of manual selection include evolution of robot controllers, mixing of food-colours and more experimental applications in evolutionary art.

Stopping Criterion

The termination criterion is based upon the problem that is being solved. Normally, an exact solution cannot be obtained and so the search for a solution is complete after a certain number of generations have been performed. In the case of the time series model that is developed in this paper, an exact solution to the training data is neither desirable (since this would imply that the solution has been overly specialised) nor is it likely to occur. Hence, the solution that is accepted is the one with the best fitness after a fixed number of generations.

4.4.3 Dimensionally Aware Genetic Program

The standard GP is ignorant of the dimensionality of its terminals and as such can only produce dimensionally correct formulations if it is applied to problems composed of dimensionless numbers. However, given the symbolic nature of GP and its ability to manipulate the structure of functional relationships, the inclusion of units of measurement, and the information contained within them, is thought to result in improved search efficiency. The inclusion of dimensionality is termed as Dimensionally Aware Genetic Programming (DA GP) (Keijzer and Babovic, 1999) and differs from

generally used evolutionary computing approaches, in that the raw observations are used together with their units of measurement. In the DA GP every node in the parse trees maintains a description of the units of the measurement associated with each terminal, with randomly generated constants allowed only as dimensionless quantities. The application of arithmetic functions on dimension-augmented terminals should be undertaken without violating the dimensional constraints. For example, adding metres to seconds renders a dimensionally incorrect result, whereas dividing metres by seconds gives the dimensionally correct expression for linear motion. In cases where dimensional correctness is not maintained additional functions are introduced in order to repair trees and guarantee closure. Further details of dimensionally aware genetic programming may be found in Keijzer and Babovic (1999).

4.5 Artificial Neural Networks

Artificial Neural Networks (ANNs), also referred to as Neural Networks, are a class of artificial intelligence algorithm that operate analogously to the biological process of a brain. Artificial Neural Networks are composed of a number of interconnected simple processing elements called neurons or nodes. Each node receives an input signal which is the total "information" from other nodes or external stimuli, processes it locally through an activation or transfer function, and produces a transformed output signal to other nodes or external outputs. The mathematical models are much simpler than their biological counterparts. It should be highlighted that not all ANNs are models of biological neurons. However the inspiration for the field of ANNs stems from the notion of producing a system of an artificial brain.

Although each individual neuron implements its function rather slowly and imperfectly, collectively a network can produce a surprising performance (Reilly and Cooper, 1990). This information processing characteristic makes ANNs a powerful computational device, able to learn from examples, and then to generalize to examples never before seen. In recent years ANNs have become extremely popular for prediction and forecasting in a number of areas, including finance, power generation, medicine, water resources and environmental science. Data classification (or grouping) and function approximation (or mapping) are among other common applications of ANNs.

Although the concept of artificial neurons was first introduced in 1943 (McCulloch and Pitts, 1943), research into the application of ANNs has blossomed since the introduction of the back propagation training algorithm for Feedforward ANNs in 1986 (Rumelhart et al., 1986). ANNs may thus be considered a fairly new tool in the field of prediction and forecasting.

4.5.1 Biological Inspiration of Artificial Neurons

The long course of evolution has given the human brain many desirable characteristics, such as massive parallelism, distributed representation and computation, learning ability, adaptivity, inherent contextual information processing, and fault tolerance. Thus the human brain has the ability to perform difficult operations and to recognise complex patterns, even if these patterns are distorted by 'noise'. The particular ability of the brain to learn from experience without a predefined knowledge of the underlying physical relationship makes it an exceptionally flexible and powerful calculating device that researchers have long tried to mimic.

ANNs are not an exact computational representation of the human brain, but are merely inspired by the limited understanding of the activities that take place in the brain. The human brain consists of approximately 10 to 100 billion (10^{11}) brain cells, known as neurons. These neurons are massively connected, with each neuron being connected to 10^3 to 10^4 other neurons. In total, the human brain contains approximately 10^{14} to 10^{15} interconnections. When the brain performs a task, like recognising a pattern, a number of processing steps are undertaken, some delays are caused by the travel time of information between neurons, but the brain still has the ability to process steps within a second. On the other hand a single switching operation in a digital computer is of the order of magnitude of 1 nanosecond (10^{-9} sec) which is about 1 million times faster than the response time of a biological neuron. Even with this capability most modern computers are still incapable of performing better then human brain for most cases. The brain compensates for the relatively slow switching speed by massively parallel configurations. The inspiration for ANNs therefore lies in the desire to mimic this functionality of the human brain on a computer which lacks the parallelism.

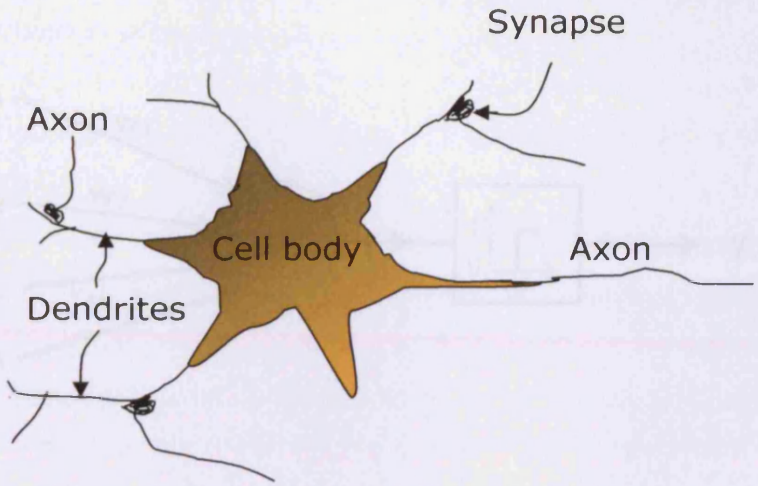


Figure 4.5: A sketch of biological neuron

Neurons (or nerve cells) are special biological cell that process information (see Figure 4.5). They are composed of a cell body, or soma, and two types of out-reaching tree-like branches: the axon and the dendrites. The cell body has a nucleus that contains information about hereditary traits and plasma that holds the molecular equipment for producing material needed by the neuron. A neuron receives signals (impulses) from other neurons through its dendrites (receivers) and transmits signals generated by its cell body along the axon (transmitter), which eventually branches into strands and sub strands. At the terminals of these strands are the synapses. A synapse is an elementary structure and functional unit between two neurons (an axon strand of one neuron and a dendrite of the other). When the impulse reaches the synapse's terminal, certain chemicals called neurotransmitters are releases. The neurotransmitters diffuse across the synaptic gap, to enhance or inhibit, depending on the type of synapse and the receptor neuron's own tendency to emit electrical impulses. Neurons communicate though short series of these pulses. The synapses' effectiveness can be adjusted by the signals passing through it so that the synapses can learn from the activities in which they participate. For a more complete description on how biological neurons actually perform computations, reference is made to the work of Hopfield (1994).

McCulloch and Pitts (1943) proposed a binary threshold unit as a computational model for an artificial neuron (see Figure 4.6), which stems from an extremely oversimplified description of the operation of biological neurons.

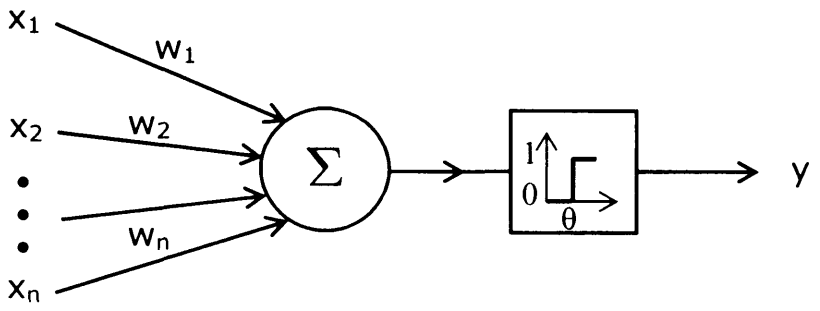


Figure 4.6: McCulloch and Pitts model of a neuron

This mathematical neuron computes a weighted sum of its n input signals, x_i where i = 1, 2... n, and generates an output of one or zero, depending upon whether this sum is above or below a given threshold value, u. Mathematically,

$$y = \theta\left(\sum_{i=1}^{n} w_i x_i - u\right) \tag{4.1}$$

where θ (η) is a unit step function and w_i is the synapse weight associated with the i^{th} input, when $\eta \leq 0$, $\theta(\eta) = 0$ and $\theta(\eta) = 1$ otherwise. McCulloch and Pitts (1943) showed how a synchronous assembly of these model neurons could compute any logical function for a suitable selection of weights.

Positive weights correspond to excitatory synapses, while negative weights model inhibitory ones. McCulloch and Pitts proved that in principle suitably chosen weights let a synchronous arrangement of such neurons perform universal computations. So as a crude analogy to biological neurons, connection weights represent synapses and the threshold function approximates the activity in a cell body or soma.

This demonstrated the ability of these devices to be used for numerical computations also. In fact, the systems of model neurons provide a complete computational model capable, in principle, of performing the same computations that can be performed by any digital computer. The ability of these artificial neurons, however, is only a first step towards emulating the functionality of the human brain. One of the most fundamental properties of the human brain is its ability to learn from example. In case of artificial neuron, it learns through iterative adjustments of the weights, w_i in order to perform a desired computation.

A learning process in the neural network context can be considered as the problem of updating network architecture and connection weights so that a network can efficiently perform a given task. The network usually must learn the connection weights from available training patterns, with performance being improved over time by iteratively updating the weights in the network.

4.5.2 Types of Neural Network

Neural networks generally consist of a number of interconnected processing elements (PEs) or neurons. How the inter neuron connections are made, and the way information flows through the network, determine the network architecture. How the strengths of the connections are adjusted or trained to minimize prediction errors is governed by its learning algorithm. Many different ANN models have been proposed since the 1980s, with Neural networks being classified according to their structure and learning algorithm (Pham and Liu, 1995).

4.5.2.1 Based on Network Structure

Feedforward network: In a Feedforward network the neurons are arranged in distinct layers. A signal flows from the input layer to the output layer through unidirectional connections. The neurons of one layer are connected only to the neurons of the next layers. Perhaps the most influential models are the multi-layer perceptrons (MLP) (Rumelhart and McClelland, 1986), radial-basis function networks (Park and Sandberg, 1991), the learning vector quantisation (LVQ) network (Kohonen, 1989) and the group method of data handling (GMDH) network (Hectch-Nielson, 1990). Feedforward network can most naturally perform static mapping between the input and output

space, i.e. the out put of a given instant is a function of the input of that instant (Pham and Liu, 1995). For dynamic systems mapping it needs to be treated explicitly (Krishnapura and Jutan, 1997; Gencay and Liu, 1996), which can be achieved by introducing lagged inputs.

Recurrent networks: In recurrent networks neurons of one layer can connect to the neurons of the next layer, the previous layer, same layer and even to themselves. Examples of recurrent networks include the Hopfield network (Hopfield, 1982), Elman network (Elman, 1990) and Jordan network (Jordan, 1986). Recurrent network have a dynamic memory and their output at a given instant reflect the current inputs, as well as previous inputs and outputs. Recurrent networks can model dynamic properties implicitly (Krishnapura and Jutan, 1997; Gencay and Li, 1996).

Based on Learning Algorithm

Networks are trained mainly by using two types of learning algorithms namely supervised and unsupervised learning algorithms. There is also a third type of learning algorithm, called reinforcement learning, however this can be regarded as a special form of supervised learning (Pham and Liu, 1995).

Supervised learning: In supervised learning training data contains examples of inputs along with the corresponding outputs and the network adjusts the strengths or weights of the interneuron connections according to the difference between the desired and actual network output corresponding to a given input. Examples of a supervised learning algorithm include: delta rule (Widro and Hoff, 1960), the generalised delta rule or back propagation algorithm (Rumelhart and McClelland, 1986) and the LVQ algorithm (Kohonen, 1989).

Unsupervised learning: Unsupervised learning is used in cases when learning (fitting of models) in this case cannot be guided by previously known classifications. It does not require the desired outputs to be known. The training dataset in this case contains input data only and the network adjusts the weights in reference to the input patterns. Unsupervised learning algorithms attempt to locate clusters in the input data based on similar features. Examples of unsupervised learning include the Kohonen Network (Kohonen, 1989), also known as Self-Organizing Feature Maps (SOFM) and

Carpenter-Grossberg Adaptive Resonance Theory (ART) (Carpenter and Grossberg, 1988) competitive learning algorithms.

MLP is perhaps the most common type of artificial neural network. However Radial Basis Networks, Hopfield networks and Kohonen's self organizing networks offer some advantages over MLP in some cases and have been widely used. The description of some common types of networks is given below.

4.5.3 Feedforward Network

An MLP is typically composed of several layers of nodes, where the nodes in one layer can be connected to nodes in the next layer, the previous layer, the same layer and even to themselves. This variety of MLP is called recurrent networks. Feedforward network is a type of MLP where the nodes in one layer are only connected to nodes in the next layer.

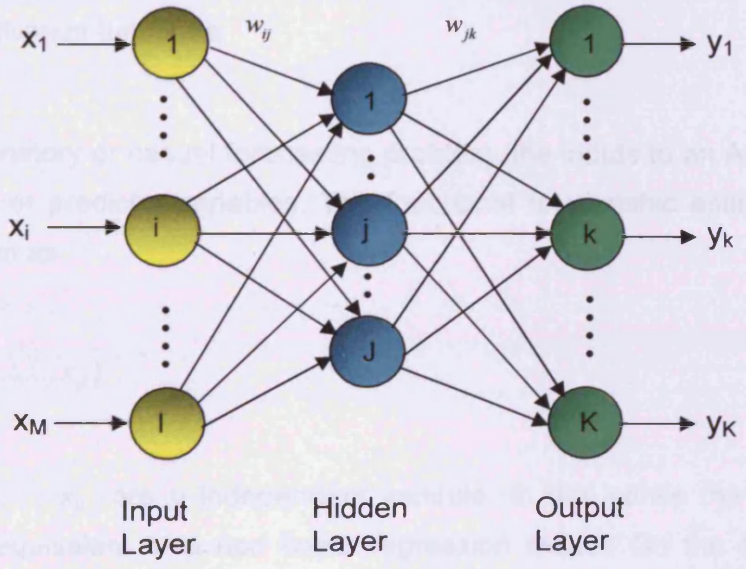


Figure 4.7: A typical Feedforward Network (MLP)

In a Feedforward network the first layer is an input layer where external information is received. The last layer is an output layer where the problem solution is obtained. The input layer and output layer are separated by one or more layers, called the hidden layers. Figure 4.7 gives an example of the general structure of a Feedforward network with one hidden layer.

In this network there are M input nodes, N hidden nodes in the single hidden layer, and P output nodes. This can be expressed mathematically as:

$$y_k = f_1\left(\sum_{j=1}^{J} w_{jk} f_2\left(\sum_{i=1}^{I} w_{ij} x_i + \theta_1^j\right) + \theta_2^k\right), \qquad \forall k \in 1,2,.....K \tag{4.2}$$

where y_k is the output from the k^{th} node of the output layer; x_i is the input at the i^{th} node of the input layer; w_{jk} is the connection weight between j^{th} node of the hidden layer and k^{th} node of the out put layer; w^{ij} is the connection weight between i^{th} node of the input layer and j^{th} node of the hidden layer. θ_1^j and θ_2^k are the bias terms, and $f_1(\bullet)$ and $f_2(\bullet)$ are activation functions.

For an explanatory or casual forecasting problem, the inputs to an ANN are usually the independent or predictor variables. The functional relationship estimated by the ANN can be written as

$$y = f(x_1, x_2, ..., x_p) \tag{4.3}$$

where $x_1, x_2, \ldots, x_p$ are p independent variable. In this sense the neural network is functionally equivalent of a non linear regression model. On the other hand, for an extrapolative and time series forecasting problem, the inputs are typically past observations of the data series and the output is a future value. The ANN performs the following function mapping

$$y_{t+1} = f(y_t, y_{t-1}, ..., y_{t-p}) \tag{4.4}$$

where y_t is the observation at time t. Thus the ANN is equivalent to the nonlinear autoregressive model for time series forecasting problems. It is also easy to incorporate both predictor variables and time-lagged observations into one ANN model.

4.5.4 Activation Function

Each hidden or output layer in a neural network receives values from input or adjacent layer. Each non-input unit in a neural network combines values that are fed into it via the connections from the units of a previous layer and they are combined to produce a single value called the net input. There is no standard term in the ANN literature for the function that combines these values. Generally linear vector to scalar combination functions are used in Feedforward networks. After the combination the scalar value is passed through a transfer function also known as an activation function. Activation functions for the hidden units are needed to introduce nonlinearity into the network that makes multilayer networks so powerful. Without nonlinearity, hidden units would not make nets any more powerful than just plain perceptrons. However, activation functions can be a linear function as well, in which case the net input is not transformed. The choice of transfer function may strongly influence the performance and complexity of a network.

This function typically falls into one of three categories:

- linear (or ramp)
- threshold
- sigmoid

For linear units, the output activity is proportional to the total weighted output. For threshold units, the outputs are set at one of two levels, depending on whether the total input is greater than or less than some threshold value. For sigmoid units, the output varies continuously but not linearly as the input changes. Sigmoid units bear a greater resemblance to real neurones than do linear or threshold units. The details of the most common transfer functions are given in Table 4.1.

Table 4.1: Characteristics of common activation functions

Name	Function	Derivative	Output Range
Linear function	$y = x$	$\frac{dy}{dx} = 1$	No Limit
Logistic Sigmoid	$y = \frac{1}{1+e^{-x}}$	$\frac{dy}{dx} = y(1-y)$	0 to 1
Tanh Sigmoid	$y = \frac{e^{x} - e^{-x}}{e^{x} + e^{-x}}$	$\frac{dy}{dx} = 1 - y^2$	-1 to 1
Percepton or Hard Limit	$y = \begin{cases} 0, & \text{if less then threshold} \\ 1, & \text{if not less then threshold} \end{cases}$	Not Applicable	0 or 1

4.5.5 Back-Propagation Training Algorithm

The most popular algorithm for training Feedforward networks is 'error back-propagation algorithm'. It is often referred to as back propagation training algorithm. The purpose of training is to adjust the network weights such that the network produces the desired output in response to the input patterns. During the training phase, any difference between network output and the target is treated as an error, with the purpose of the network being to minimise the error.

Consider a network with I number of input nodes, J number of hidden nodes and K number of output nodes as shown in Figure 4.7. Back propagation can be applied to any Feedforward network with differentiable transfer functions. Let us consider an input vector $\boldsymbol{x} = (x_1, x_2, x_3, \ldots, x_I)$ applied to the input layer, noting that the input layer does not perform any operation upon the input signal but simply sends the input signal x_i to

the units of a hidden layer. The net input to the j^{th} hidden unit for a given training pattern, p is

$$s_j = \sum_I w_{ij} x_i + \theta_j \tag{4.5}$$

where w_{ij} is the weight of the connection from the i^{th} input unit to the j^{th} hidden unit and θ_j is a bias term. The output of this node can be written as

$$y_j = f(s_j) \tag{4.6}$$

where the activation function f is of any form as shown in Table 4.1.

Similarly the output from a unit k in the output layer is

$$y_k = f(s_k) \tag{4.7}$$

where

$$s_k = \sum_J w_{jk} y_j + \theta_k \tag{4.8}$$

Updating the output layer units

The error measure E is defined as the total quadratic error for pattern p at the output units

$$E = \frac{1}{2} \sum_{k=1}^{K} (d_k - y_k)^2 \tag{4.9}$$

where d_k is the desired output. It can now be written –

$$\frac{\partial E}{\partial w_{jk}} = \frac{\partial E}{\partial s_k}\frac{\partial s_k}{\partial w_{jk}} \tag{4.10}$$

From Eq. (4.8) we see that the second factor is

$$\frac{\partial s_k}{\partial w_{jk}} = \frac{\partial}{\partial w_{jk}}\left(\sum_j w_{jk} y_j + \theta_k\right) = y_j \tag{4.11}$$

If we define

$$\delta_k = -\frac{\partial E}{\partial s_k} \tag{4.12}$$

then,Eq. (4.10) becomes,

$$\frac{\partial E}{\partial w_{jk}} = -\delta_k y_j \tag{4.13}$$

By applying the principle gradient descent method, weights must be changed in proportion to the amount given by Eq. (4.13), i.e.

$$\Delta_p w_{jk} = \eta \delta_k y_j \tag{4.14}$$

The factor η is called the learning rate parameter.

To calculate δ_k, for each unit we can write Eq. (4.12) as

$$\delta_k = -\frac{\partial E}{\partial s_k} = -\frac{\partial E}{\partial y_k}\frac{\partial y_k}{\partial s_k} \tag{4.15}$$

where it follows from the definition of E^p from Eq. (4.9) that

$$\frac{\partial E}{\partial y_k} = -(d_k - y_k) \tag{4.16}$$

and from Eq. (4.7)

$$\frac{\partial y_k}{\partial s_k} = \frac{\partial}{\partial s_k} f(s_k) = f'(s_k) \tag{4.17}$$

Hence Eq. (4.10) becomes

$$\delta_k = -(d_k - y_k) f'(s_k) \tag{4.18}$$

Updating the hidden layer units

Similarly for the hidden unit

$$\frac{\partial E}{\partial w_{ij}} = \frac{\partial E}{\partial s_j} \frac{\partial s_j}{\partial w_{ij}} \tag{4.19}$$

From Eq. (4.5) we see that the second factor is

$$\frac{\partial s_j}{\partial w_{ij}} = \frac{\partial}{\partial w_{ij}} \left(\sum_i w_{ij} x_i + \theta_j \right) = x_j \tag{4.20}$$

Hence, if we define

$$\delta_j = -\frac{\partial E}{\partial s_j} \tag{4.21}$$

then Eq. (4.19) becomes

$$\frac{\partial E}{\partial w_{ij}} = -\delta_j x_j \tag{4.22}$$

and the change of weight should be

$$\Delta_p w_{ij} = \eta \delta_j x_i \tag{4.23}$$

To calculate δ_j, for each unit we can write (4.21) as

$$\delta_j = -\frac{\partial E}{\partial s_j} = -\frac{\partial E}{\partial y_j}\frac{\partial y_j}{\partial s_j} \tag{4.24}$$

and from Eq. (4.6) we get

$$\frac{\partial y_j}{\partial s_j} = f'\left(s_j\right) \tag{4.25}$$

And then we can write

$$\frac{\partial E}{\partial y_j} = \sum_k \frac{\partial E}{\partial s_k}\frac{\partial s_k}{\partial y_j}$$

$$= \sum_k \frac{\partial E}{\partial s_k}\frac{\partial}{\partial y_j}\left(\sum_j w_{jk} y_j + \theta_k\right)$$

$$= \sum_k \frac{\partial E}{\partial s_k} w_{jk}$$

$$= -\sum_k \delta_k w_{jk} \tag{4.26}$$

Substituting Equations (4.20) and (4.21) to Eq. (4.19), we get:

$$\delta_j = f'(s_j)\sum_k \delta_k w_{jk} \tag{4.27}$$

Equations (4.15) and (4.27) gives a recursive procedure for computing δ for all units in the network, which are then used to compute the weight changes according to Eq. (4.16). This procedure is also called the generalised delta rule.

The application of generalised delta rule thus involves two phases. During the first phase, the input x_i, is presented to the network and propagated forwards through the network to the output layer. Here the desired output, d_k and the computed output y_k are compared with each other and the error signal δ_k from Eq. (4.15) and the corresponding weight adjustment from Eq. (4.14). In the second phase, this error is propagated backwards through the network to each intermediate layer. At each intermediate layer the error signal δ_j is computed from Eq. (4.27) and the associated weight adjustment from Eq. (4.23). For a network having more then a hidden layer, this process is repeated for every layers until an input layer is reached.

Momentum parameter

The learning procedure requires the weight change to be proportional to $\partial E / \partial w$. True gradient descent requires this to be infinitesimal. The constant of proportionality is the learning rate η. For practical purposes the learning rate is chosen to be as large as possible without leading to oscillation. One way to avoid oscillations at large values of η is to add a momentum term. The idea is to stabilise the weight trajectory by making the weight change a combination of a gradient-decreasing term plus a fraction of the previous weight change. With momentum the current weight change is a combination of a step down from negative gradient, plus a fraction $0<\alpha<1$ of the previous weight change. The weight update rule can be written as:

$$\Delta w(t+1) = \delta w + \alpha \Delta w(t) \tag{4.28}$$

where t indexes the presentation number and α is a constant which determines the effect of the previous weight change.

In order to optimise the performance of the Feedforward network trained with back propagation algorithm, it is essential to have a good understanding the impact of step size on training (Dai and MacBeth, 1997 and Maier and Dandy 1998)

4.5.6 Radial Basis Function Network

Radial basis function networks (RBF) are another popular Feedforward network. The fundamental difference between MLP and radial basis functions is the way in which the hidden nodes combine with signals from preceding layers in the network. MLPs have one or more hidden layers, for which a combination function is the inner product of inputs and weights, with a bias being added. The activation function is usually *logistic* or *tanh* function. On the other hand RBF, have one hidden layer for which the combination function is dependent upon the position of data relative to some centre point, i.e.

$$\Phi(\|x - w\|) \tag{4.29}$$

where $\|.....\|$ denotes some distance measure of x from the centre point w. The distance measure is often chosen to be Euclidean, so that:

$$\delta_j = \|x - w\| = \sqrt{\sum (x_i - w_i)^2} \tag{4.30}$$

and the most commonly used combination function is a Gaussian function, i.e.

$$\Phi(\delta_j) = y_j = e^{-\lambda \delta_j^2} \tag{4.31}$$

where λ is some constant.

Thus, unit *j* gives a maximum response to the input vectors near *w*. As a result each hidden unit occupies a region in the input space centred upon w_j. The idea is then to pave the input space with these receptive fields. If an input vector *x* lies in the middle of the receptive field for unit *j*, then only unit *j* will be activated. If the input vector lies between two receptive field centres, then the network will make a smooth interpolation between the two units.

The output of the RBF network is a linear combination of the response function (4.31). This approach is guaranteed to produce a function that fits all data points as long as there is a basis function for each input. As the radial units of the hidden layer uses a *Gaussian* response surface which is nonlinear in nature, only one hidden layer is sufficient to model any shape of function.

Having one hidden unit for each input means that noisy data will also be a part of the model, which will affect the generalisation ability of the network. This can be improved by reducing the number of hidden units (Minns, 1998). The selection of the coefficient for the linear combination of basis function outputs is then a simple problem of linear optimisation, which will readily find a global optimum solution.

There are advantages of RBF networks over MLPs. Firstly, RBF networks can be trained faster, with the simple linear transfer in the output layer can be optimised by traditional linear modelling technique, which are fast and do not suffer from the problem of local minima. Secondly, as it can model any non linear function with a single hidden layer, there is no decision making needed to select the number of hidden layers. Minns (1998) found that RBFs offer a superior performance over MLPs while dealing with small input data. However, as the data set increased the generalisation property of RBFs deteriorated and they were subsequently out performed by the MLPs.

On the other hand according to Dibike's (2002) observation RBF networks achieve a similar performance as Feedforward network in a lesser time period although they require more data to reach same level of accuracy.

4.5.7 Kohonen Network

Kohonen Networks also known as Self Organising Feature Maps were introduced by Von der Malsburg (1973), and in their present form by Kohonen (1982). This network differs considerably from the Feedforward back propagation neural network. The main difference is that the Kohonen network is trained in an unsupervised mode. This means that the Kohonen network is presented with data, but the correct output that corresponds to those data are not specified. However it does not only differs in how it is trained but also how it recalls a pattern. The Kohonen neural network does not use any sort of activation function or any sort of a bias weight. Output from the Kohonen neural network does not consist of the output of several neurons. When a pattern is presented to a Kohonen network one of the output neurons are selected as *winner*. Often these winning neurons represent groups in the data that are presented to the Kohonen network.

The structure of a Kohonen network consists of an input layer and an output layer. A Kohonen network or self-organising feature map has two layers, an input buffer layer to receive the input pattern and an output layer (see Figure 4.8)

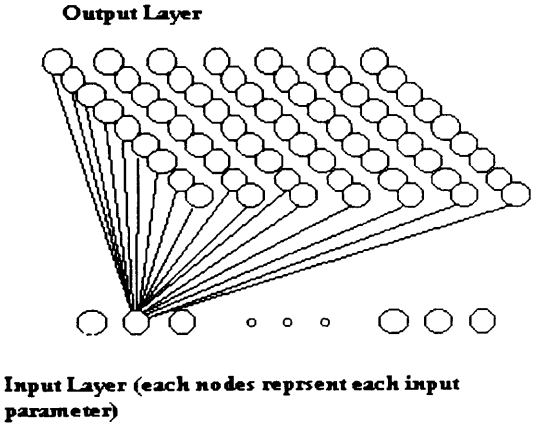


Figure 4.8: Kohonen Network

The neurons in the output layers are usually arranged into a regular two dimensional array. Each neuron in the output layer stores a weight vector (an array of weights), each of which corresponds to one of the inputs in the data.

Each input neuron is connected to all output neurons, with the weights of the connections from the components of the reference vector being associated with the given output neuron. The learning process corresponds to repeatedly modifying the synaptic weights of all the connections in the system, in response to the input patterns and according to a prescribed rule until a steady configuration is achieved.

The basic components of the learning of a Kohonen network involve the following steps:

- Initialise the reference vectors of all out put neurons to small random values,
- Present a training input pattern,
- Determine a winning output pattern, i.e. the neurons whose reference vector is closest to the input pattern,
- Update the reference vector of the winning neuron and those of its neighbours. These reference vectors are brought closer to the input vector. The adjustment is greatest for the reference vector of the winning neuron and decreases for the reference vectors of the neurons further away.

When presented with a new input pattern, each neuron calculates its activation level based on the following

$$\sqrt{\sum_{i=0}^{n}\left(w_i - p_i\right)^2} \tag{4.32}$$

where w_i is the i^{th} element of the weight vector and p_i is the p^{th} element of the input pattern. The neuron which is closest in Euclidian space to the new input pattern has the lowest activation level and is allowed to adjust its weights so that it is closer to the input pattern. Some of the nodes nearer to it also adjust their weight, the number of those neighbouring nodes is determined as the algorithm runs, beginning at all the nodes and decreasing linearly throughout the training process.

The amount by which each neuron changes its weight vector is determined by the definition

$$\delta w_i = -\alpha(w_i - p_i) \tag{4.33}$$

where α is the learning rate, which begins as specified by the user and decreases to 0 as the algorithm runs, and δw_i is the change in w_i. This change is carried out for each element in the weight vector.

Kohonen network is a relatively simple network to construct, which can be trained very rapidly. However it also has its own limitations, as having only two layers, it can only be applied to linearly separable problems.

4.5.8 Hopfield Network

The Hopfield network consists of a set of N single layer neurons (Figure 4.9) in which each of the neurons are connected to all other neurons, which results in a recurrent network. The updates of the activation values of the neurons are asynchronous and independent of other neurons. All neurons are both input and output neurons and normally accept binary (0 or 1) and bipolar inputs (+1 or -1).

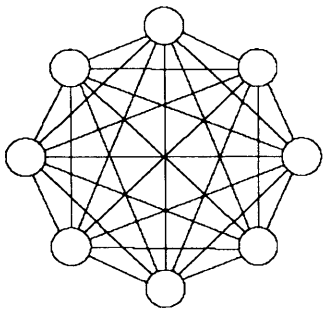


Figure 4.9 : Hopfield Network

The training of Hopfield network takes only one step, the weights w_{ij} of the network being assigned directly as follows:

$$w_{ij} = \begin{cases} \frac{1}{N}\sum_{c=1}^{P} x_i^c x_j^c & where\ i \neq j \\ 0 & where\ i = j \end{cases} \tag{4.34}$$

In this expression w_{ij} is the connection weight from neuron i to neuron j, and x_i^c (which is either +1 or -1) is the i^{th} component of the training input pattern for class *c*, *P* is the number of classes and *N* is the number of neurons (or the number of components in the input pattern). It should be noted that in equation (4.34) $w_{ij} = w_{ji}$ and $w_{ii} = 0$, a set of conditions that guarantee the stability of the network. Removing the restriction of bidirectional connections (i.e., $w_{ij} = w_{ji}$) results in a system that is not guaranteed to settle to a stable state.

When the network experiences an unknown input pattern, the outputs are initially set equal to the components of the unknown pattern, i.e.

$$y_i(0) = x_i \qquad 1 \leq i \leq N \tag{4.35}$$

Starting with initial values, the network iterates according to the following equation until it reaches a minimum energy state, i.e. its output stabilise to constant values:-

$$y_i(k+1) = f\left[\sum_{j=1}^{N} w_{ij} y_i(k)\right] \qquad 1 < i \leq N \tag{4.36}$$

where f is a hard limit function, defined as

$$f(x) = \begin{cases} -1 & x < 0 \\ 1 & x > 0 \end{cases} \tag{4.37}$$

4.5.9 Modelling Issues

The idea behind ANN modelling is to identify the underlying relationship between the inputs and outputs. Ideally the function has to be smooth and continuous, so that a small change in input will give rise to a small change input. It is important the inputs have adequate information related to the target, so that a mathematical function can be achieved to relate outputs (with a desired degree of accuracy) to inputs, ANNs will not learn with a non existent function. Finding good inputs for ANN modelling and collecting sufficient training data take more time and effort then training the network. The major issues related to modelling using ANNs are discussed in the flowing sections.

4.5.9.1 Division of data

In ANN methodology, the available data set is generally subdivided into two or three parts, namely training, test and validation sets. The training sample is used for learning, that is, to fit the parameter weights of classifiers. Test samples are used only to assess the generalisation performance. If the test results are not good, then training should not be continued. If the testing confirms that the data are acceptable, the model can be further checked with the validation set. A Validation sample is also used to avoid the over fitting problem or to determine the stopping point of the training process (Weigned et. al. 1992).

There is no general rule for dividing data into training, test and validation sets. Several factors, such as the problem characteristics, the data type and the size of the available data, should be considered during making a decision. It is important to have both the training and test sets (also validation sets, if present) representing the population or the underlying mechanism. An inappropriate division of datasets will adversely affect the ANN performance adversely. ANNs are, generally, unable to extrapolate beyond the range of data used for training (Minns and Hall, 1996). Consequently, poor forecasts/predictions can be expected when the validation data contain values outside of the range of those used for training.

Literature offers little guidance in selecting the training and test datasets. Most authors select the sets based on the rule 90% - 10%, 80% - 20% or 70% -30% etc. Some chose the sets based on particular problems. Another closely related issue is the size

of the dataset. No definite rule exists for the requirement of the size for a given problem. In general, for in any data driven method the larger the data set, the more accurate the results. Nam and Schaefer (1995) test the effect of different training sample size and found that as training sample size increases, then the performance of the ANN gets better.

4.5.9.2 Data pre-processing

In any modelling problem different input variables have different ranges. The contribution of an input will be heavily dependent on its variability, relative to other inputs. If one input has a range of 0 to 1, while another input has a range of 0 to 1 million, then the contribution of the first input is more likely to be dwarfed by the second input. Data standardisation ensures that all variables receive equal attention during the training process. There is another issue related to the nature of transfer functions. The non linear transfer functions typically restrict possible outputs from a node to (0,1) or (-1,1). Hence, the variables should be scaled in order to be commensurate with the limits of the activation functions used in the output layer (Minns and Hall, 1996). However, if the transfer functions in the output layer are linear then scaling is not strictly required (Karunanithi et al., 1994).

There are other benefits of scaling suggested in literature, such as meeting algorithm requirements (Sharda and Patil, 1992), facilitation of network learning (Srinivasan et al., 1994) and avoiding computational problems (Lapedes and Farber, 1988). It is important to note that the available data need to be divided into training, testing and validation subsets, before any data pre-processing is carried out (Burden et al., 1997). It is common to standardise each output to the same range, or the same standard deviation if there is a lack of better prior information. If some inputs are more important then others, then it may be better to scale the inputs such that the more important ones have larger variances and or/ranges (Neelakantan, 2005).

4.5.9.3 Determination of model inputs

The number of nodes in the input layer is fixed by the number of model inputs, whereas the number of nodes in the output layer equals the number of model outputs. The number of input variables corresponds to the number of input nodes used in the network. Hence, the selection of input is a part model construction process. There is no

guideline to determine this number at present. Ideally, it would be intended that a minimal number of input parameters will unveil the features embedded in the data. Too few or too many numbers will affect either the learning or the prediction capability of the network. Selection of appropriate model inputs is extremely important for any prediction or forecasting problem. However, as the data driven approaches have the ability to determine which model inputs are critical there is no need for a '*priori rationalisation about relationships between variables*' (Lachtermacher and Fuller, 1994). Presenting a large number of inputs to ANN models, and relying on the network to determine the critical model inputs usually increases network size. This has a number of disadvantages, such as decreasing processing speed and increasing the amount of data required to estimate the connection weights efficiently (Lachtermacher and Fuller, 1994). The problem is exacerbated in time series applications, where appropriate lags have to be chosen for each of the input variables. Consequently, there are distinct advantages in using analytical techniques to help determine the inputs for multivariate ANN models. Many authors design experiments for selection of the number of input nodes while others adopt some empirical relationship. For example, Sharda and Patil (1992) used 12 inputs for monthly data and 4 for quarterly data. Recently, Genetic Algorithms have been increasingly in use for the optimal design of neural networks (Koza and Rice ,1991, Schiffmann et al., 1993).

The number of output nodes is relatively easy to specify as it is directly related to the problem being studied. For a time series forecasting problem, the number of output nodes often corresponds to the forecasting horizon.

4.5.9.4 Determination of network architecture

Network architecture determines how the information flows through the network, as well as the number of connection weights. Determination of appropriate network architecture is one of the most difficult, albeit important, tasks in the model building process.

Type of connection and degree of connectivity

Feedforward networks have traditionally been used for prediction and forecasting applications. For most of the forecasting, as well as other problems, networks are fully

connected in that all nodes in one layer are only connected to all nodes in the next layer. However, it is possible to use partial connectivity (Chen et al., 1992).

Recurrent networks have also been recently proposed as an alternative, for example, Warner and Misra (1996), Khotanzad et al (1997) described Feedforward networks to perform well in comparison with recurrent networks in many practical applications. However it follows that Feedforward networks are special cases of recurrent networks.

Feedforward networks require dynamic systems to be treated explicitly while recurrent networks can model dynamical properties implicitly (Krishnapura and Jutan, 1997). In case of a Feedforward network, a dynamic system can be achieved by including lagged inputs. In current research legged inputs have been used to represent the dynamic system of bacterial decay.

Lin et al (1996) found that recurrent networks have difficulties in capturing long-term dependencies when inputs at high lags have a significant effect on network outputs. They also suggested that the inclusion of inputs at explicit time lags can improve the performance considerably in such cases. According to Masters (1993) the processing speeds of a Feedforward network *'is among the fastest of all models currently in use'* (Masters, 1993). As for the advantage of using recurrent networks, they can handle moving average components (Connor et al., 1994), whereas Feedforward networks are unable to do so. Examples of different types of recurrent networks that have been used for time series applications include those proposed by Elman (1990); Williams and Zipser (1989); Krishnapura and Jutan (1997).

Network Geometry

Network geometry determines the number of hidden layers and chooses the number of nodes in each of these layers. However, it has also been suggested that during training it might be best to fix the number of nodes, rather than the number of hidden layers, and to optimise the connections between the nodes, as well as the connection weights associated with each connection (Kumar, 1993).

The choice of the number of nodes in the hidden layers is a critical factor and hence the number of connection weights. If the balance between having sufficient free parameters (weights) to enable representation of the function to be approximated and having too many free parameters is not struck, then this may result in overtraining, with this issue having been discussed widely in the literature (Maren and Harston, 1990; Rojas, 1996). Smaller networks with a few hidden layers usually have better generalisation ability (Castellano et al., 1997, Neelakantan, 2005), require fewer physical resources (e.g. require less storage space), have higher processing speeds (during training and testing) and make rule extraction simpler (Towell et al., 1991). Bebis and Georgiopoulos (1994) reported that despite the fact that smaller network can be implemented on hardware more easily and economically, the error surface is more complicated and contains more local minima. Moreover the smaller networks generally need a large number of training samples to deliver good generalisation property.

Number of hidden layers plays an important role in capturing the pattern of data and performing non linear mapping. It has been shown that ANNs with one hidden layer is sufficient to approximate any complex nonlinear function (Hornik et al., 1989, Cybenko, 1989). However, in practice many functions are difficult to approximate with one hidden layer (Flood and Kartam, 1994). Barron (1994) suggested two hidden layers may provide more benefits to some of problems. There is quite a high variability in the number of nodes suggested by the various rules, however guidelines do not ensure optimal network geometry, where optimality is defined as the smallest network that adequately captures the relationship in the training data. Traditionally, optimal network geometries have been found by trial and error. More recently, a number of systematic approaches for determining optimal network geometries have been proposed, including pruning and constructive algorithms (see Bebis and Georgiopoulos, 1994). However, it must be stressed that the optimal network geometry is highly problem dependent.

4.6 Summery

The Management and control of water resources systems is traditionally based on mathematical models describing the behaviour of the natural process, with these models requiring a good understanding of the underlying processes. Recent developments in the field of data mining and knowledge discovery have introduced a

whole new branch of data driven-modelling techniques. These techniques have also shown their potential as an alternative approach to conventional deterministic modelling. This chapter has briefly introduced two such types of data-based models namely: Genetic Programming (GP) models and Artificial Neural Networks (ANN), with a view to using these models as efficient alternatives to solve some of the various problems of water resources modelling and management discussed in subsequent chapters.

Data-driven modelling methods are increasingly being developed and used to gain information from data. They are especially useful when the data sets are large, and where it becomes impractical for any human to sort through the data to obtain further insights into the processes that have led to the data. Many data-rich problems can be solved by using novel data-driven models together with other techniques. Data mining methods are also being increasingly used to gain a greater understanding and knowledge from large data sets.

While physically based or process-based models are appealing to those who wish to better understand these natural processes, they clearly remain approximations of reality. In some water resources problem applications, the complexities of the real system are considerably greater than the complexities of the models built to simulate them. Hence, it should not be surprising that in some cases data-based models, which convert input variable values to output variable values in ways, which are not directly related to the processes, may produce more accurate results than physically-based models. When these models are used, they often produce results faster than their physical counterparts and as accurately or more so, but only within the range of values observed in the data used to build these models.

Chapter 5

MODEL DEVELOPMENT AND APPLICATION TO RIBBLE ESTUARY

5.1 General Description

The Ribble Estuary discharges along the north-west coast of England, in Lancashire. At the mouth of the estuary there are two popular seaside resorts, namely Lytham St Anne's and Southport, both designated bathing waters. The Fylde Coast, which is bounded between Fleetwood in the north and the Ribble Estuary in the south, includes Blackpool, one of the most famous beaches in England for tourism, and with an average of more than 17 million visitors per annum.

In order to improve the bathing water quality along the Fylde coast about £600 million was invested during the 1990s. New sewerage works and treatment plants were constructed along the Fylde coast and in the estuary. The improvements included upgrading the waste water treatment works at Clifton Marsh from primary treatment to include UV disinfection; reducing storm water discharges from the wastewater network by constructing 260 Ml of additional storage. These waste water treatment works contributed to a significant reduction in the input bacterial loads, and as a result the concentration of bacteria in the receiving coaster waters has been reduced. However, occasional high levels of Faecal Coliform (FC) counts have still been measured and, subsequently, the bathing waters under these conditions have failed to comply with the EU mandatory water quality standards. As a popular tourist attraction the bacteria concentration in the coastal area has to be monitored continuously in order to comply with the EC mandatory water quality standards.

A numerical modelling study was undertaken to establish the water quality of the EU designated bathing waters located at the mouth of the Ribble Estuary. A hydrodynamic and water quality model was used in that study. In order to reduce the possible

inaccuracies caused by setting up the boundary conditions required by the numerical models, the upstream boundaries were set up at the tidal limits of the rivers Ribble, Darwen and Douglas (see Figure 5.1) and the downstream boundary was located around the 25 m depth contour in the Irish Sea. The model was verified using six sets of hydrodynamic and water quality data were collected during the winter period of 1998 and the summer period of 1999 by the UK Environment Agency. The survey data included water depths, current speed and directions, salinity levels and concentrations of suspended solids, Faecal Coliforms and Total Coliforms, and Faecal Streptococci at all discharge sites, upstream river boundaries and four calibration sites (see Figure 5.1). At each site, a survey typically provided 25 data points for calibration. More details regarding this study may be found in Kashefipour et al (2002).

Figure 5.1: Fylde Coast and Ribble Estuary with its tributaries

5.2 Methodology

All data mining activities are data intensive. They are truly useful only in situations where a considerable body of data exists. Unfortunately, in water quality monitoring an intensive survey of bacterial load for a water body is rarely available. Most of the survey campaign is carried out either as long term operations, collected at a low frequency (e.g. monthly) or at high frequencies but with the period of data collection being very short (e.g. a few days).

High frequency observations spanning longer terms are almost non existent. Even though the concentration of bacteria is measurable in a natural river and estuary, considering the amount of data needed to run a data driven method, such as Genetic Programming (GP) or Artificial Neural Networks (ANNs), is at the very least prohibitive in terms of effort and cost. Hence in this study the existing 2-dimensional hydrodynamic and water quality model has been used to generate the data. Such a numerical model employs all available knowledge about the decay of bacteria, the dispersion and diffusion of solutes, the influence of the tidal cycle and riverine inflows, all of which have an important role on the bacterial transport and die-off. Moreover the data generated from such a model can be used as a noise free approximation of the phenomenon under study. The main objective of this application was therefore to verify the employability of appropriate data mining technology rather than providing a deterministic solution to a particular problem. Taking the raw data from a deterministic numerical model allowed the study to focus on correctly predicting these data, and verifying the applicability of data mining technology in water quality model predictions for the basin. If the input data were collected from the field, then significant resources would be required in obtaining these data, which in this case would not add much to the research objective.

5.3 Numerical Model

In this study a depth integrated two-dimensional numerical model, namely, Depth Integrated Velocities and Solute Transport (DIVAST) was used to predict the bacterial indicator concentration distributions. DIVAST was developed for simulating the hydrodynamic, solute and sediment transport processes in estuarine and coastal

waters. It has been calibrated and validated against many laboratory and practical field studies over the past 25 years. The hydrodynamic module of the model is based on the solution of the depth integrated Navier-Stokes equations and it includes the effects of: local acceleration, advective acceleration, earth's rotation, pressure gradient, wind stress, bed resistance and turbulent shear stresses.

For the water quality module, the advective-diffusion equation (ADE) is solved for a range of water quality indicators, including:- salinity, total and faecal coliforms, biochemical oxygen demand, dissolved oxygen, the nitrogen and phosphorous cycles and algal growth. The ADE defines the dynamic distributions of the bacterial indicators due to the flow characteristics, diffusion processes and die-off rates. The Faecal coliform decay rates are expressed as a first order decay model according to Chick's Law.

In current study DIVAST was modified to incorporate the effects of salinity and temperature and was verified against the field data collected during the study of Kashefipour et al. (2002). In general, good agreement was obtained between the model predictions and the measured data for both the hydrodynamic and water quality calibration studies. Predicted Faecal Coliform (FC) levels at 2 selected locations, e.g., 7 Milepost and 11 Milepost were compared with the corresponding measured values for the field survey. The corresponding results are illustrated in Fig. 5.2.

The modified DIVAST model, used as a data generator, provided the convenience of fast data generation and an ability to produce results for any combination of boundary conditions. A number of DIVAST runs were performed to generate data for the model development with GP and ANN. The flow and concentration varied through the input files of DIVAST. The model itself calculated the salinity and concentration at the points previously mentioned at every time step.

The flow and concentration inputs from the rivers Ribble, Darwen and Douglas, at a given time were acquired from the input files and the corresponding salinity and water depths and FC concentrations at a particular location (e.g.11 Milepost) were acquired from the model output. The flow and concentrations at the rivers, salinity and depth

data at those particular locations have been used as input data, while the FC concentrations at those locations were used as target data.

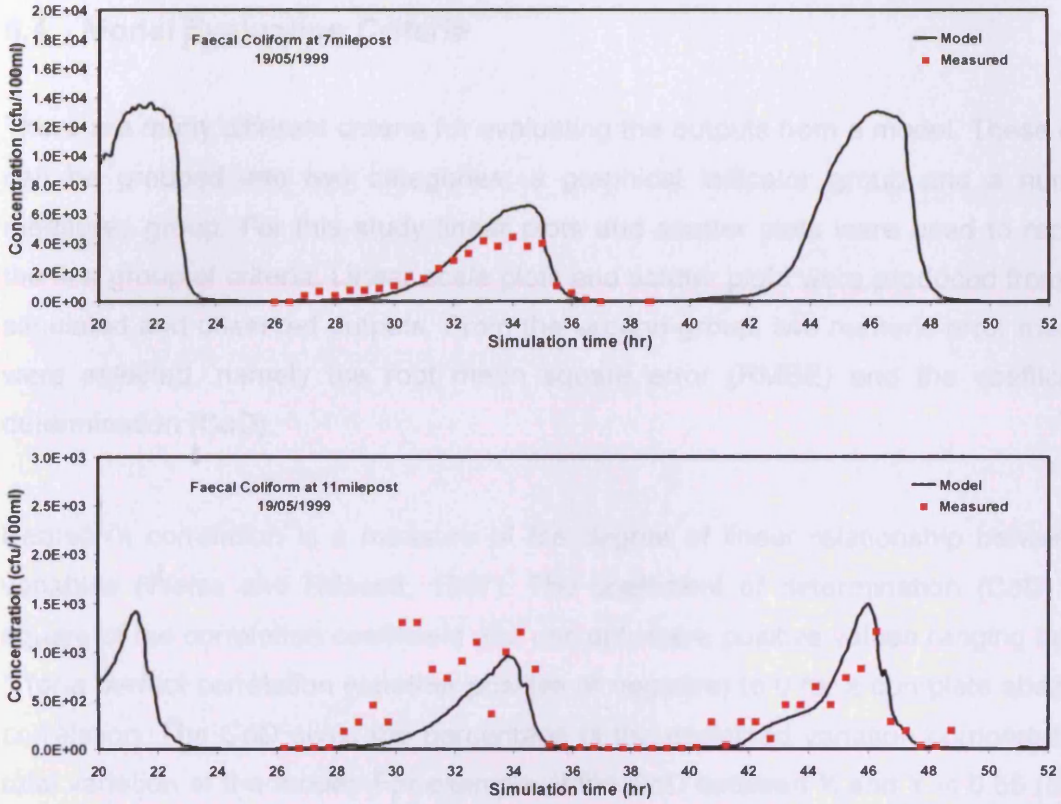


Figure 5.2: Comparison between predicted and measured faecal coliform concentrations for the survey on 19 May, 1999, in Ribble Estuary

It would have been ideal to use sunlight data as a parameter, especially since sunlight is an important factor for bacterial decay. However, no sunlight data were recorded during the original surveys and this parameter not be incorporated in the hydro dynamic model. As a result sunlight was also ignored for the current model development process. For a given boundary condition the model was run for 50 hours and the data were collected every 15 minutes. The dataset comprised the result from a total of 8 DIVAST runs. This data set will be referred to as the observed value for the remainder of this chapter.

In this study, the Faecal Coliform levels in the Ribble estuary have been modelled by using two data mining technique, namely:

- Genetic Programming (GP) and
- Artificial Neural Networks (ANNs).

5.4 Model Evaluation Criteria

There are many different criteria for evaluating the outputs from a model. These criteria can be grouped into two categories: a graphical indicator group and a numerical measures group. For this study linear plots and scatter plots were used to represent the first group of criteria. Linear scale plots and scatter plots were produced from of the simulated and observed outputs. From the second group, two numeric error measures were selected, namely the root mean square error (RMSE) and the coefficient of determination (CoD).

Pearson's correlation is a measure of the degree of linear relationship between two variables (Weiss and Hassett, 1987). The coefficient of determination (CoD) is the square of the correlation coefficient and can only have positive values ranging between 1 for a perfect correlation (whether positive or negative) to 0 for a complete absence of correlation. The CoD gives the percentage of the explained variation compared to the total variation of the model. For example, if the CoD between X and Y is 0.55 (55%), it can be said that 55% of the variability of X is explained by the variability in Y (Weiss and Hassett, 1987). These two statistical measures is widely used for model evaluation. (see Iliadis and Maris 2007, Cigizoglu 2005a, Choi et al. 2004, Linne et al. 2000)

The definition of RMSE and CoD are given below:

$$RMSE = \sqrt{\frac{\sum_{i=1}^{N}(O_i - S_i)^2}{N}} \qquad (5.1)$$

$$CoD = \left(\frac{\sum_{i=1}^{N} (O_i - \overline{O})(S_i - \overline{S})}{\sqrt{\sum_{i=1}^{N} (O_i - \overline{O})^2 \sum_{i=1}^{N} (S_i - \overline{S})^2}} \right)^2 \quad (5.2)$$

where:

O_i – Observed or measured data;

S_i – Predicted or simulated data;

$\overline{O}$ - Mean value of the observed data;

$\overline{S}$ - Mean value of the simulated data;

N – Size of the data set

The RMSE value gives a quantitative indication of the model error; it measures the deviation of the forecasted value from the actual observed value. The CoD values represent the proportion of the variation in the data that has been explained by the regression line. The ideal value for RMSE is 0, for CoD is 1. Thus the all the model results in this study were analysed and assessed by visual inspection of the observed and modelled FC concentrations, as well as by their absolute RMSE and correlation coefficients.

As the model developed is intended to be used for day to day monitoring of recreational waters, another additional criterion was also included this particular study. In order to use GP and ANN type models for monitoring purposes it is more important to detect when the water quality fails to comply with the guideline compliance values which for the EU Bathing Water Directive is 2000 cfu/100ml. Lin et al (2003) employed this type of evaluation criterion by reporting the number of days when the observed and predicted FC concentrations exceeded the regulatory water quality standard. In this work the number of failed sample was reported for both the observed and predicted samples, as well as the number of occurrences when the failed sample was correctly predicted as *failed.*

Kim and Barros (2001) used the *Threat Score* (TS) to perform similar verification for flood forecasting. This is a dimensionless value which evaluates the performance of a

model relative to some threshold value. The value of the coefficient ranges from 0 to 1. When the model performs perfectly then the TS will be 1. The TS is defined as –

$$TS = \frac{CP_n}{Obs_n + Pre_n - CP_n} \tag{5.3}$$

where

TS = Threat Score

Obs_n = Number of observed sample more then 2000 cfu/100ml

Pre_n = Number of predicted sample more then 2000 cfu/100ml

CP_n = Number of correct prediction, i.e. when both observed and predicted values are more then 2000 cfu/100ml

The threat score of the selected models are reported here only.

5.5 Data Analysis

Table 5.1 shows the statistical analysis of the data used for building the hydroinformatics models. Although varying numbers of input parameters were used for different experiments, only the important data statistics are presented in the table. It is interesting to note that FC levels at the target sites, i.e. 7 and 11 Mileposts ranges from 10^1 to 10^5. This large variation is very much expected in such a complex natural domain; however, on the other hand it shows the degree of difficulty to model such phenomenon by a data driven model. The standard deviation is also in the range of 10^4. This indicates how widely the data are spread particularly since there are no outliers in the dataset.

As the GP and ANNs are data driven, the quality of a model depends primarily on the quality of the data and hence, data analysis is very important prior to any model building operation. In the current study the nonlinear data analysis technique called the Gamma Test was used perform the data analysis.

Table 5.1: Statistical analysis of the important data used

	Ribble Flow	Darwen Flow	Douglas Flow	FC in Ribble	FC in Darwen	FC in Douglas	FC in 11 Milepost	FC in 7 Milepost
Minimum	12.78	1.47	1.18	318.00	1100.00	5224.70	34.90	195.20
Maximum	108.88	17.70	10.20	4.49×10^{5}	2.57×10^{5}	3.4×10^{7}	1.04×10^{5}	2.26×10^{5}
Mean	40.14	4.30	2.98	32597.11	28915.16	9.91×10^{5}	9950.00	2.68×10^{4}
Median	33.05	3.79	2.59	4689.95	7660.00	5.6×10^{4}	1744.40	2904.00
Std. Dev.*	25.71	2.88	1.52	76288.85	48155.86	3.21×10^{6}	1.8×10^{4}	4.7×10^{4}

Gamma test examines the relationships between input and output datasets. Suppose there is a set of input–output observations of the form

$$\{(x_i, y_i) | 1 \le i \le M\} \tag{5.4}$$

where the inputs $x \in ¡^{m}$ are vectors confined to some closed bounded set $C \in ¡^{m}$ and, without loss of generality, the corresponding outputs $y \in ¡$ are scalars. Rather than pre-suppose some particular parametric form for the underlying non-linear model it is considered that it belongs to some *general class of functions*. In general terms the underlying relationship can be assumed of the form

$$y = f(x_1, \ldots\ldots, x_m) + r \tag{5.5}$$

where f is a suitably smooth and unknown function that maps the components of the input vector $\boldsymbol{x}$ to the output y and r is a stochastic variable which represents noise. The mean of the distribution of r is assumed to be zero.

* Standard Deviation

Even though the underlying function f is unknown, the Gamma test can estimate the variance of r, $\text{var}(r)$ directly from the data. This estimate is called the Gamma statistic and denoted by Γ. As the number of data samples increases, the Gamma statistic approaches an asymptotic value, which is the variance of noise on the particular output. For more details on the theory of Gamma statistic, reference is made to Evan and Jones (2002).

In Gamma Test the critical graph to look at first is the scatter plots and $(\delta(p), \gamma(p))$ regression line. The scatter plot shows point pairs (δ, γ) where δ is the squared distance of an input (**x**) from one of its near neighbours. Figures 5.3 and 5.4 shows two such scatter plots. It can be seen from Figure 5.3 that an empty wedge appears at the top left corner of the graph. This indicates that the input data and output data are closely related, thus an underlying smooth model exist. On the other hand, there is no wedge at the top left corner in Figure 5.4 indicating a high level of noise in the data or there is no smooth underlying model.

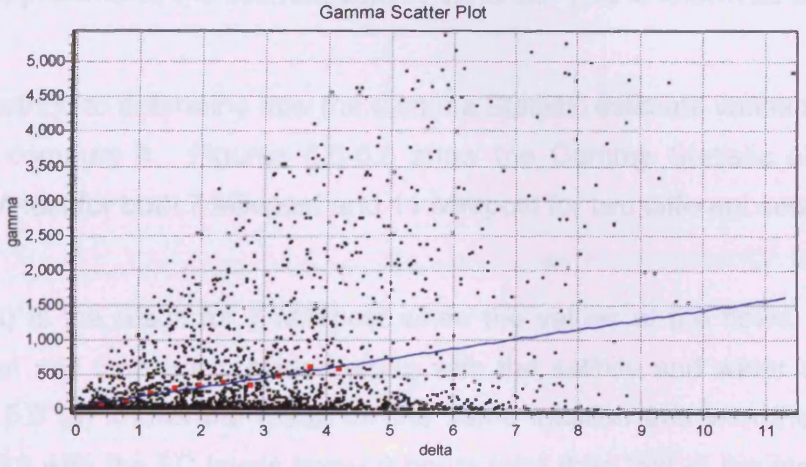


Figure 5.3: Scatter plot of polynomial equation $y = x + x^2 + x^3$ (empty 'wedge' at the top left corner)

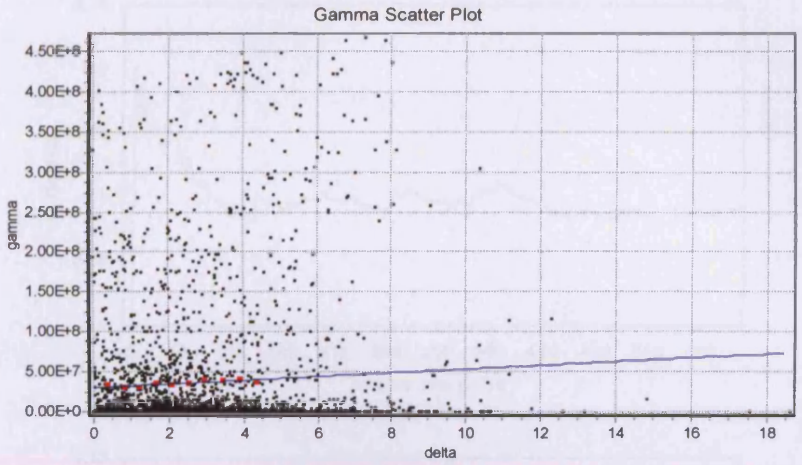


Figure 5.4: Scatter plot of polynomial equation $y = x + x^2 + x^3$ with random noise (no 'wedge' at the top left corner, indicates difficulty or impossibility of finding a smooth model)

The reliability of the Γ statistic is determined by running a series of Gamma test for increasing M, to establish the size of data set required to produce a stable asymptote and thus indicates how much data is required to build a model with a mean squared error which approximates the estimated noise variance. This is known as an M-test.

M-test is a method to determine how the Gamma Statistic estimate varies as more data are used to compute it. Figures 5.5-5.6 show the Gamma Statistic obtained from running the M-test for both 7 Milepost and 11 Milepost for two different sets of data.

Figure 5.5 (a) is the result for 7 Milepost when the values of the flows, FC levels of three rivers at any time are provided along with the salinity and water depth of that time. Figure 5.5 (b) shows the result for the same location and with the same input parameters but with the FC levels being 9 hours later than that of the input data. The red line shows the possible asymptote.

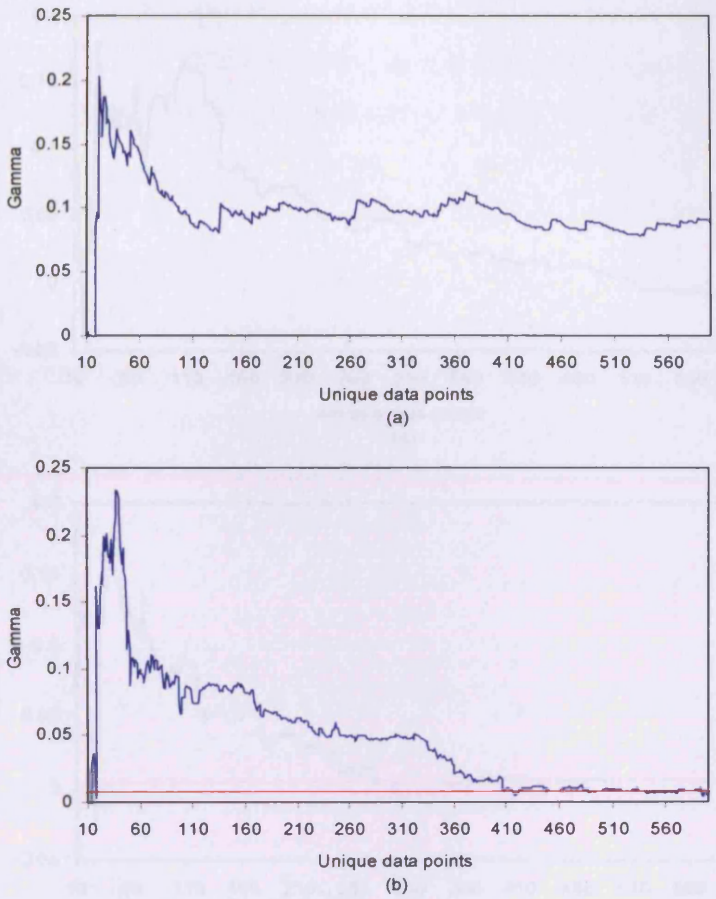


Figure 5.5: M Test performed on randomised scaled data at 7 Milepost: (a) with no time lag information and (b) with 9 hr time lag

It can be seen from these 2 figures that the value of Gamma Statistic reduced significantly with the addition of the time lag information. This is due to the fact that that the travel time of contaminates from the upstream boundaries, particularly the Ribble river, to 7 Milepost is around 9 hours, so boundary information 9 hours in advance will have the most significant impact on the FC level at 7 Milepost. This can also be confirmed by plotting the numerical model results. From Figure 5.5 it can be seen that after 420 points the Gamma statistic is fairly stable, this means that an adequate model can be built using 420 or more data points. A very small value of Gamma Statistic indicates that the noise level in the data is low, indicating a smooth model can be developed.

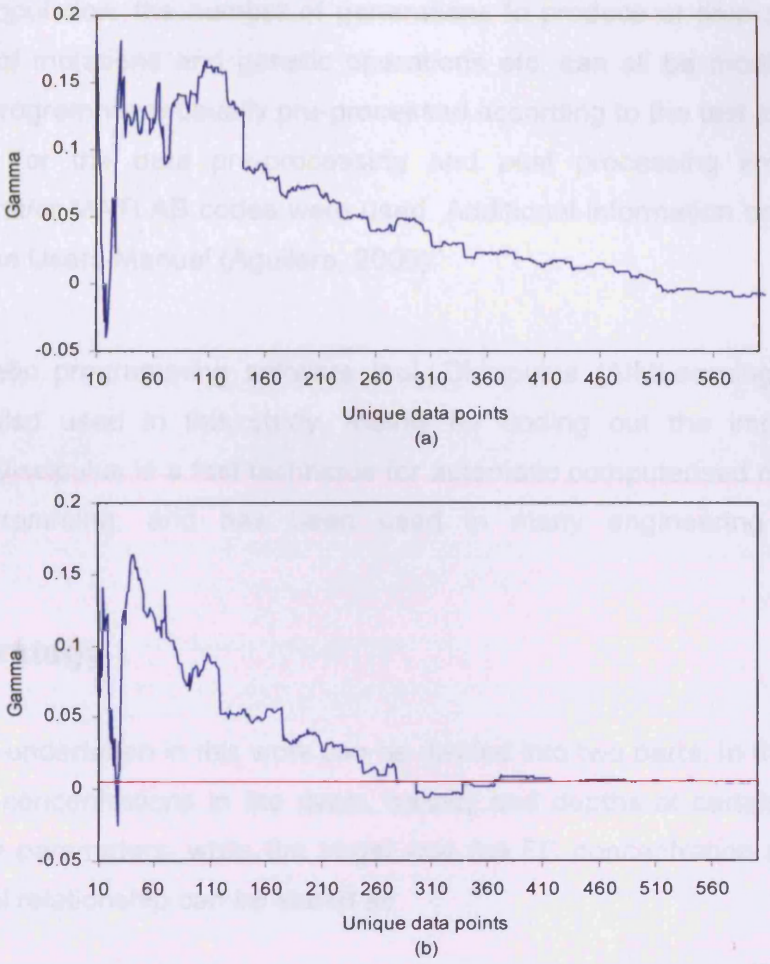


Figure 5.6: M Test performed on randomised scaled data at 11 Milepost: (a) with no time lag information and (b) with 11 hr time lag

A similar trend can be seen for 11 Milepost from Figure 5.6, in which a time lag of 11 hours was used as found from the numerical model analysis. Again, the Gamma statistic becomes stable after about 420 data points.

5.6 FC modelling with GP

The software used for the GP simulations is called GPKernel (Genetic Programming Kernel). The code was developed by Maarten Keijzer and Vladan Babovic at the Danish Hydraulic Institute (DHI). GpKernel is a line-command program that looks for mathematical relations based on a set of input parameters, constants, operators, genetic parameters and on user-defined target(s). The genetic parameters, such as the

size of the population, the number of generations to produce or time of run, different probabilities of mutations and genetic operations etc. can all be modified. The input data for the program were usually pre-processed according to the test structure agreed in advance. For the data pre-processing and post processing small self-written FORTRAN and/or MATLAB codes were used. Additional information on GPKernel can be found in the Users Manual (Aguilera, 2000).

Another genetic programming software tool, Discipulus (AIMLearning™ Technology 2000) was also used in this study, mainly for finding out the impact of various parameters. Discipulus is a fast technique for automatic computerised modelling, using genetic programming, and has been used in many engineering and scientific applications.

General Strategy

The GP tests undertaken in this work can be divided into two parts. In the first part only the flow and concentrations in the rivers, salinity and depths at certain location were used as input parameters, while the target was the FC concentration at that location. The functional relationship can be stated as

$$y = f(x_i) \qquad \forall i \in 1, 2, \ldots\ldots \tag{5.6}$$

where y is the target and x_i are inputs

The following experiments were carried out as part of the first GP test on data gathered from 11 Milepost and 7 Milepost respectively:

Experiment 1 – In this experiment the inputs were flow and FC concentrations of 3 rivers and salinity and depth of the location in question. The target was the FC at that location.

Experiment 2 – In this experiment the same inputs were used as for the previous experiment, but change in concentration (dFC) was used as the target. Some

literatures (e.g. Minns 1998) suggested use of the derivatives of targets instead of the absolute values. The FC concentration increment used as a target in the simulations was calculated as follows:

$$dFC(t) = FC(t) - FC(t-1) \tag{5.7}$$

Experiment 3 - In this experiment the logarithmic value of the FC concentrations was used, both for the inputs and target, instead of the absolute value or its derivative. This combination was chosen because the faecal coliform concentrations ranged from less then 35 cfu/100ml to more then 45,000,000. This parameter represented the small changes in the bacterial concentrations, when absolute concentrations were low. Since there were no significant differences between the minimum and maximum value for other parameters, they remained unchanged.

Hence, for experiments 1, 2 and 3 only 8 input parameters were used, which included flow and FC concentrations of the three rivers and the salinity and water depth at 7 and 11 Mileposts respectively.

In the second part of this study the past values of the same parameters were also used for the model development in order to capture the time series nature of the problem. In this case the following experiments were undertaken.

Experiment 4 – The previous simulation values of the flow (i.e. 3, 6, 12, 18 and 24 hours before simulation), concentrations of the rivers and the current values of salinity and water depths at a certain location were used as input parameters, while the output were the values of the FC concentration at the same location. Hence, the functional relationships were represented as -

$$y_{t+1} = f\left(x_t^i, x_{t-j}^i, salinity, depth\right) \qquad \forall i \in 1,2,.....6 \quad \forall j \in 3,6,12,18,24\ hr. \tag{5.8}$$

where x^i are the flow and concentration values and y are the output values. The total number of inputs in this experiment were 48.

Experiment 5- In this experiment previous value of the FC concentrations were also used as inputs in addition to the inputs used in experiment 4. The outputs remained the same. The functional relationship was described as

$$y_{t+1} = f\left(x_t^i, x_{t-j}^i, salinity, depth, y_{t-j}\right) \qquad \forall i \in 1,2,.....6 \quad \forall j \in 3,6,12,18,24\ hr \tag{5.9}$$

where x^i are the inputs and y_i are the outputs. The total number of inputs used was 53.

Experiment 6 – In this experiment an attempt had been made to reduce the number of inputs from 53. For this task the Discipulus is used. In experiment 6, 37 inputs are selected for 7 Milepost and 40 inputs are selected for 11 Milepost based on the impact table produced by Discipulus.

For all the experiments the dataset was randomised for running all the experiment to make sure that there was no influence of one combination of data over the next combination, thus GP had to treat the every set data of data as stand alone combination.

Test Setup

The first three experiments were run for two tests, whereas Experiments 4 and 5 were run for three tests. A summary of all of the experiment run setup parameters for GP Kernel is provided in the Table 5.2.

In Table 5.2 μ is the population size and λ is the number of offspring to be produced. In all runs a possibility for using random constants in the evolutionary process has been utilised. No hypothesis proposed by the GP considered the dimensional equality and hence the experimental runs were not dimensionally constrained. The *language* is the set of mathematical operators that were specified before each of the runs. The number of records used in the training was 75% and the Evolution strategy was tournament (size 3) for first run of experiment 1, 2 and 3 and the first two runs of experiment 3 and 4 while Elitist strategy was used for the last run for each experiment.

Table 5.2: GPKernel set up parameters

No	ID	Run	Experiment		Objectives	GP Parameters		Language
			No	Time (min)		μ	λ	
1	Exp. 1	1	3	720	CoD,RMSE	1000	1300	+,-,×,/, sqrt
2		2	3	845	CoD,RMSE	1200	1000	+,-,×,/,sqrt, pow
3	Exp. 2	1	3	953	CoD,RMSE	1000	1300	+,-,×,/, sqrt
4		2	3	886	CoD,RMSE	1200	1000	+,-,×,/,sqrt, pow
5	Exp. 3	1	3	1035	CoD,RMSE	1000	1300	+,-,×,/, sqrt
6		2	3	1168	CoD,RMSE	1200	1000	+,-,×,/,sqrt, pow
7	Exp. 4	1	2	1148	CoD,RMSE	600	900	+,-,×,/, sqrt
8		2	2	1243	CoD,RMSE	650	850	+,-,×,/,sqrt, pow
9		3	2	1056	CoD,RMSE	620	400	+,-,×,/,sqrt, pow,exp
10	Exp. 5	1	2	1232	CoD,RMSE	600	900	+,-,×,/, sqrt
11		2	2	1130	CoD,RMSE	650	850	+,-,×,/,sqrt, pow
12		3	2	1272	CoD,RMSE	620	400	+,-,×,/,sqrt, pow,exp

Test Result and Analysis

The output of GPKernel is typically a file that contains the best expressions, along with the values of the objective functions, which the program has found during the evolutionary process. The objective values are calculated based on the 75% (as uses for training purpose) of the input data. The raw expressions from GPKernel are usually of the following form:

```
FC = (((((((sqrt(((FC24 * FC3) * FC3)) + ((Qrib6 / Qdar18) *
(FCrib * Qrib18))) / sqrt(FCrib)) + (((Qrib18 + sqrt((FCdoug12 *
(FCrib3 * Qrib18)))) / Qdar12) + Qdoug18)) + ((((Qrib6 + FC24) /
Qdar12) / Qdar18) + (Qdar18 / FC12))) + Qdoug12) + (sqrt((Qdar12
+ FCrib)) / sal)) + (((FC24 + Qrib18) + Qrib6) / Qrib18)) / sal)
```

with the description of the variables being presented later in this chapter.

These expressions can be inserted directly into a MATLAB file and used for any calculations, but it has to be simplified for easier analysis, with the simplification being

done mainly manually. The following steps have been undertaken to evaluate the test results produced by GPKernel:

i. Choose the best expression from GPKernel based on CoD and RMSE.

ii. Insert into a self-written MATLAB file and calculate the predicted FC concentrations, RMSE and CoD using the entire data set.

iii. Save the observed and calculated (predicted) FC concentrations into a file

iv. Generate plots using this result, and finally

v. Simplify the expression

For the first set of experiments the result are presented in Table 5.3. To obtain the RMSE value for Experiment 3, the result is converted from a logarithmic value. The result shows that the prediction of FC was not efficient when no information about its previous time steps was included in the GP. It was very difficult to fit the variation over such a large range without including the time series data during the model build up. Hence, it was decided to undertake the second part of the experiments, which contain experiments 5 and 6 respectively.

Table 5.3: Statistical analysis of results obtained using GP models for Part 1

Experiment	11 Milepost		7 Milepost	
	CoD	RMSE	CoD	RMSE
Exp.1	0.4405	97535	0.4905	107735
Exp. 2	0.4930	82315	0.4347	115125
Exp. 3	0.4750	85964	0.4150	119964

In these the experiments the data for all the input parameters was included from 3 to 24 hours before a certain time, with the data being provided in 3 to 6 hr interval alongside the current values. Hence, the GP now has some information about how the values were reached to a certain level over an interval of time. However, in the first experiment (Experiment 4), only the time series values of inputs were presented to GP. A total of 48 input parameters were used for this experiment. The results of two best GP solutions are presented in Table 5.4.

Table 5.4: Statistical analysis of the results obtained from GP models for Experiment 4 (Part 2)

		Statistical measures		Number of failed sample		
Milepost	Run	CoD	RMSE	Obs.	Model	SE
7	1	0.8941	15413	339	509	326
	2	0.8782	16532	339	339	285
11	1	0.8968	5821	346	561	346
	2	0.8990	8757	346	561	346

In this table the number of samples, which are greater then 2000cfu/100ml have been identified as failed samples as this is the imperative guideline level in EU Directive. SE in the leftmost column of the table stands for the same event. When some event is spotted as 'failed' in both observed and modelled data then this is termed as same event failure.

As can be seen from Table 5.4, the statistical measures had been improved significantly and the model was capable of picking up the failed samples with a good level of precision. At 11 Milepost all failed sample had been successfully detected, although it had identified more then 150 samples to be failed. These cases are termed as *false positive*. Clearly the model was over predicting and hence a little too conservative. However, the model is less conservative for 7 Milepost (run 2) and it had failed to identify a number of failed samples. The cases where observed samples are more than the guideline value, but the model predicts as less the guideline can be termed as *false negative*. Overall, there had been a significant improvement but there was still scope for further improvements.

A subsequent trial (Experiment 5) was carried out to improve further on the result obtained above. In this experiment the past observations of the FC was also included

as part of input data. Total number of input in this case was 53, which are shown in Table 5.5.

Table 5.5: Statistical analysis of the results obtained from GP models for Experiment 5

		Statistical measures		Number of failed sample		
Milepost	Run	CoD	RMSE	Obs	Model	SE
7	1	0.9277	12759	339	409	301
	2	0.9204	13386	339	480	333
11	1	0.9190	5158	346	484	329
	2	0.9299	4797	346	462	323

As can be seen from Table 5.5, the statistical measures had been further improved and the models in general were over predicting, but to a lower extent then before. As a result the number of false positives predicted by the model had been reduced. However, the model failed to identify a number of observed failed. Overall, this result is encouraging, although the the number of parameters used in this experiment would generally be regarded as too many. In Experiment 6 an attempt had been made to reduce the number of parameters.

As described in section 5.4.1, the GP software Discipulus was used to reduce the number of inputs. The GP had been run for 150 generations and had evaluated around 10^9 possible solutions. When runs were completed, Discipulus produced a detailed report on the importance of the various inputs, called the Input Impacts. The input impacts are shown in Table 5.6.

Form this table the less important input parameters were determined. The less frequent inputs were then eliminated to reduce the number of input parameters. 40 inputs were selected for the 11 Milepost after eliminating any parameters that were less than 20% frequent in the best 30 programs. For the 7 Milepost, inputs that were less than 15% frequent were eliminated with 36 parameters remaining for further use.

Table 5.6: Impact and frequency of inputs in best GP programs

		Milepost 11			Milepost 7		
V	**Variable Description**	**F**	**Av**	**Max**	**F**	**Av**	**Max**
v[00]	Ribble Flow	0.20	0.01	0.01	0.13	0.00	0.00
v[01]	Darwen Flow	0.07	0.00	0.00	0.17	0.00	0.01
v[02]	Douglas Flow	0.20	0.01	0.01	0.20	0.01	0.01
v[03]	FC at Ribble	0.23	0.20	0.32	0.23	0.88	0.95
v[04]	FC at Darwen	0.27	0.57	0.80	0.03	0.00	0.00
v[05]	FC at Douglas	0.03	0.00	0.00	0.30	0.02	0.10
v[06]	Salinity	0.80	0.24	0.43	0.90	0.30	0.68
v[07]	Depth at location	0.53	0.11	0.35	0.47	0.19	0.28
v[08]	Ribble Flow preceding 3 hours	0.07	0.00	0.00	0.10	0.00	0.00
v[09]	Darwen Flow preceding 3 hours	0.10	0.00	0.00	0.40	0.13	0.45
v[10]	Douglas Flow preceding 3 hours	0.10	0.00	0.00	0.20	0.01	0.01
v[11]	FC at Ribble preceding 3 hours	0.13	0.00	0.00	0.20	0.43	0.53
v[12]	FC at Darwen preceding 3 hours	0.27	0.67	0.69	0.43	0.08	0.15
v[13]	FC at Douglas preceding 3 hours	0.07	0.00	0.00	0.17	0.01	0.02
v[14]	Salinity preceding 3 hours	0.00	0.00	0.00	0.17	0.00	0.00
v[15]	Water Depth preceding 3 hours	0.17	0.01	0.01	0.00	0.00	0.00
v[16]	Ribble Flow preceding 6 hours	0.63	0.04	0.09	0.30	0.00	0.01
v[17]	Darwen Flow preceding 6 hours	0.27	0.00	0.01	0.40	0.02	0.08
v[18]	Douglas Flow preceding 6 hours	0.30	0.08	0.11	0.23	0.00	0.00
v[19]	FC at Ribble preceding 6 hours	0.20	0.05	0.05	0.40	0.29	0.62
v[20]	FC at Darwen preceding 6 hours	0.30	0.46	0.89	0.70	0.24	0.46
v[21]	FC at Douglas preceding 6 hours	0.03	0.00	0.00	0.33	0.05	0.09
v[22]	Salinity preceding 6 hours	0.60	0.04	0.06	0.37	0.02	0.04
v[23]	Water Depth preceding 6 hours	0.03	0.00	0.00	0.17	0.02	0.03
v[24]	Ribble Flow preceding 12 hours	0.07	0.00	0.00	0.33	0.09	0.19
v[25]	Darwen Flow preceding 12 hours	0.60	0.03	0.07	0.23	0.00	0.00
v[26]	Douglas Flow preceding 12 hours	0.37	0.01	0.09	0.27	0.00	0.00
v[27]	FC at Ribble preceding 12 hours	0.60	0.26	0.34	0.30	0.09	0.18
v[28]	FC at Darwen preceding 12 hours	0.57	0.13	0.26	0.93	0.32	0.52
v[29]	FC at Douglas preceding 12 hours	0.70	0.05	0.13	0.17	0.02	0.02
v[30]	Salinity preceding 12 hours	0.17	0.01	0.02	0.30	0.00	0.00
v[31]	Water Depth preceding 12 hours	0.30	0.07	0.09	0.10	0.00	0.01
v[32]	Ribble Flow preceding 18 hours	1.00	0.25	0.45	0.37	0.06	0.20
v[33]	Darwen Flow preceding 18 hours	0.37	0.02	0.02	0.27	0.00	0.00
v[34]	Douglas Flow preceding 18 hours	0.20	0.00	0.00	0.37	0.00	0.01
v[35]	FC at Ribble preceding 18 hours	0.07	0.00	0.00	0.13	0.03	0.04
v[36]	FC at Darwen preceding 18 hours	0.27	0.05	0.08	0.00	0.00	0.00
v[37]	FC at Douglas preceding 18 hours	0.00	0.00	0.00	0.10	0.00	0.00
v[38]	Salinity preceding 18 hours	0.27	0.00	0.00	0.13	0.00	0.01
v[39]	Water Depth preceding 18 hours	0.17	0.01	0.01	0.20	0.11	0.34
v[40]	Ribble Flow preceding 24 hours	0.57	0.01	0.16	0.40	0.01	0.03
v[41]	Darwen Flow preceding 24 hours	0.20	0.14	0.51	0.20	0.11	0.17
v[42]	Douglas Flow preceding 24 hours	0.33	0.03	0.09	0.17	0.00	0.00
v[43]	FC at Ribble preceding 24 hours	0.43	0.03	0.06	0.33	0.21	0.43
v[44]	FC at Darwen preceding 24 hours	0.23	0.04	0.18	0.00	0.00	0.00

		Milepost 11			Milepost 7		
V	Variable Description	F	Av	Max	F	Av	Max
v[45]	FC at Douglas preceding 24 hours	0.03	0.00	0.00	0.20	0.02	0.03
v[46]	Salinity preceding 24 hours	0.00	0.00	0.00	0.13	0.00	0.00
v[47]	Water Depth preceding 24 hours	0.00	0.00	0.00	0.23	0.11	0.17
v[48]	FC at milepost preceding 3 Hours	0.73	0.08	0.12	0.43	0.05	0.08
v[49]	FC at milepost preceding 6 Hours	0.07	0.00	0.00	0.80	0.02	0.06
v[50]	FC at milepost Preceding 12 Hours	0.37	0.02	0.09	0.30	0.00	0.01
v[51]	FC at milepost Preceding 18 Hours	0.13	0.04	0.17	0.77	0.03	0.10
v[52]	FC at milepost Preceding 24 Hours	0.27	0.01	0.01	0.33	0.04	0.17

where V = Variable input

F = Frequency, percentage of best 30 programs containing inputs

Av = Average effect of removing all instances of inputs from best 30 programs

Table 5.7 gives the results of Experiment 6 and shows that the statistical analysis for both the mileposts had been improved significantly.

Table 5.7: Statistical analysis of the results obtained from GP models for Experiment 6

		Statistical measures			Number of failed sample		
Milepost	Run	CoD	RMSE	R^2	Obs.	Model	SE
7	1	0.9295	12762	0.9399	339	406	301
	2	0.9395	11667	0.9401	339	390	298
11	1	0.9398	4445	0.9422	346	433	329
	2	0.9365	4521	0.9399	346	495	340

The GP model also detected even less failed samples in comparison with that of experiment 5. However, the problem of failing to spot some of the failed sample was still present. These comparisons indicated at this stage that there is a trade off between avoiding over prediction and identifying the maximum number of failed samples. It had been evaluated that if the output obtained from experiment 6 was increased by 10% then it was possible to pick up 99% of the failed sample. If these models are applied to a real life situation then it is the environmental managers' decision to decide how to approach the problem.

GP produces an expression for every model it proposes. As these expressions are very complex in nature only one the best performing equations are given here for each of the locations. For 7 Milepost the expression produced from Experiment 6 run 2 is as below:

$$\varphi_{7mp} = \frac{1}{1.988s} \times \left(\varphi_{7mp24} + \frac{\varphi_{dg9}}{\sqrt{\frac{s^2 \times \varphi_{7mp24} + \sqrt{\varphi_{dg24}} \times e^{Q_{dr6}}}{(Q_{rb24})^5} + \frac{\varphi_{dr9}}{Q_{rb24}}}} \right) + \frac{1}{s} \times \left(\frac{(\varphi_{dr9} + \varphi_{rb24}) \times Q_{rb18}}{(Q_{rb24})^2} + \varphi_{dr9} \right) \quad (5.10)$$

where:

φ_{7mp} = FC concentration at 7 Milepost at certain time t

φ_{7mp24} = FC concentration at 7 Milepost at time t-24 hr

φ_{rb24} = FC concentration at Ribble boundary at time t-24 hr

φ_{dr9} = FC concentration at Darwen boundary at time t-9 hr

φ_{dg9} = FC concentration at Douglas boundary at time t -9 hr

φ_{dg24} = FC concentration at Douglas boundary at time t-24 hr

Q_{rb18} = Flow at Ribble boundary at time t-18 hr

Q_{rb24} = Flow at Ribble boundary at time t-24 hr

s = salinity at 7 Milepost at time t

This expression can be supported theoretically. It shows that if the FC levels in the rivers increases then the FC levels at 7 Milepost are also increased. Furthermore, the flow in the rivers has a positive correlation with the concentration at 7 Milepost, with increasing salinity values leading to reduction in the FC concentrations.

The expression produced for 7 Milepost shows that a reduction in the water depth leads to increased FC concentrations. These are supported in the literature, with the result obtained using this expression being shown in Figure 5.7 and 5.8.

The equation for 11 Milepost produced by GP in Experiment 6, Run 1 is as follows:

$$\varphi_{11mp} = 1 + \frac{1}{s}\left(1 + \varphi_{dr11} + \varphi_{11mp24} + \frac{1}{Q_{dg18}}\left(\varphi_{11mp24} + \varphi_{rb24} + \frac{\varphi_{rb24}}{s + d_{18} + (s_{12})^2} + \frac{\varphi_{dg11}}{d_3(s + d_3 + d_{18})} + \frac{Q_{rb18}}{s + d_{18}}\right)\right) \quad (5.11)$$

where:

φ_{11mp} = FC concentration at 11 Milepost at time t

φ_{11mp24} = FC concentration at 11 Milepost at time t-24 hr

φ_{rb24} = FC concentration at Ribble boundary at time t-24 hr

φ_{dr11} = FC concentration at Darwen boundary at time t-11 hr

φ_{dg11} = FC concentration at Douglas boundary at time t-11 hr

Q_{rb18} = Flow at Ribble boundary at time t-18 hr

Q_{dg18} = Flow at Douglas boundary at time t-18 hr

s = salinity at 11 Milepost at time t

s_{12} = salinity at 11 Milepost at time t-12

d_3 = water depth at 11 Milepost at time t-3

d_{18} = water depth at 11 Milepost at time t-18

The result produced using this equation are shown in Figures 5.9 and 5.10. According to this expression the FC levels and the flow in rivers have a positive correlation with concentration in 11 Milepost while salinity and water depth an inverse relation.

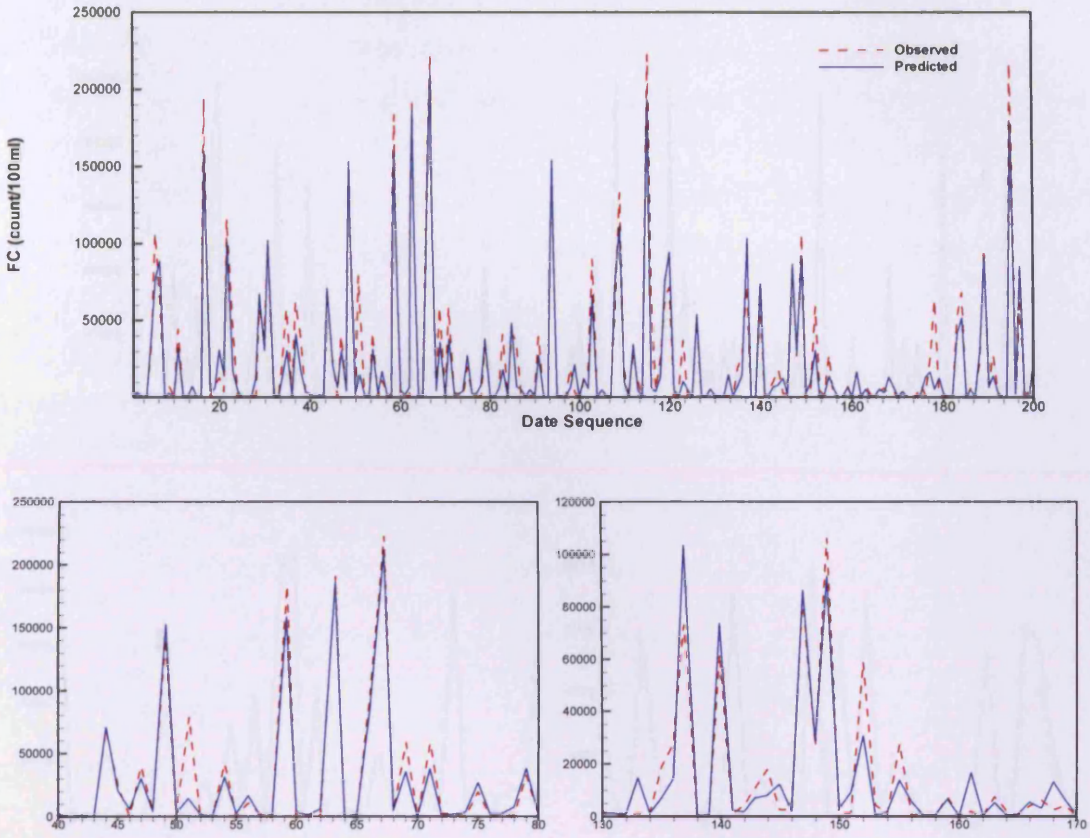


Figure 5.7: Comparison between observed and GP predicted FC levels at 7 Milepost

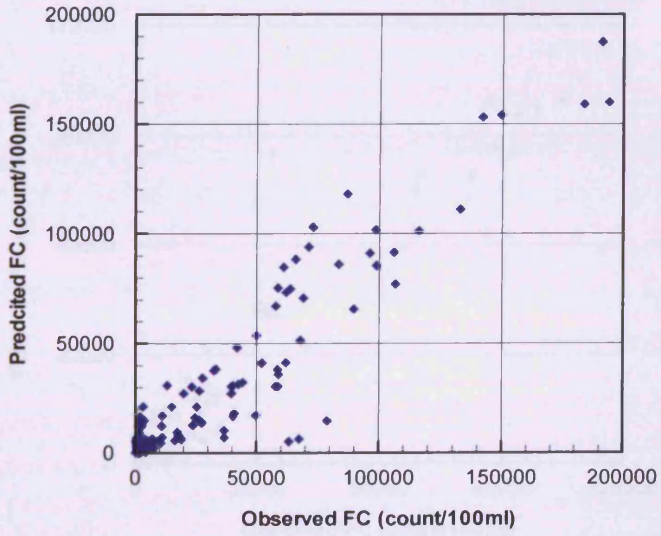


Figure 5.8: Scatter plot for observed and GP predicted FC levels at 7 Milepost

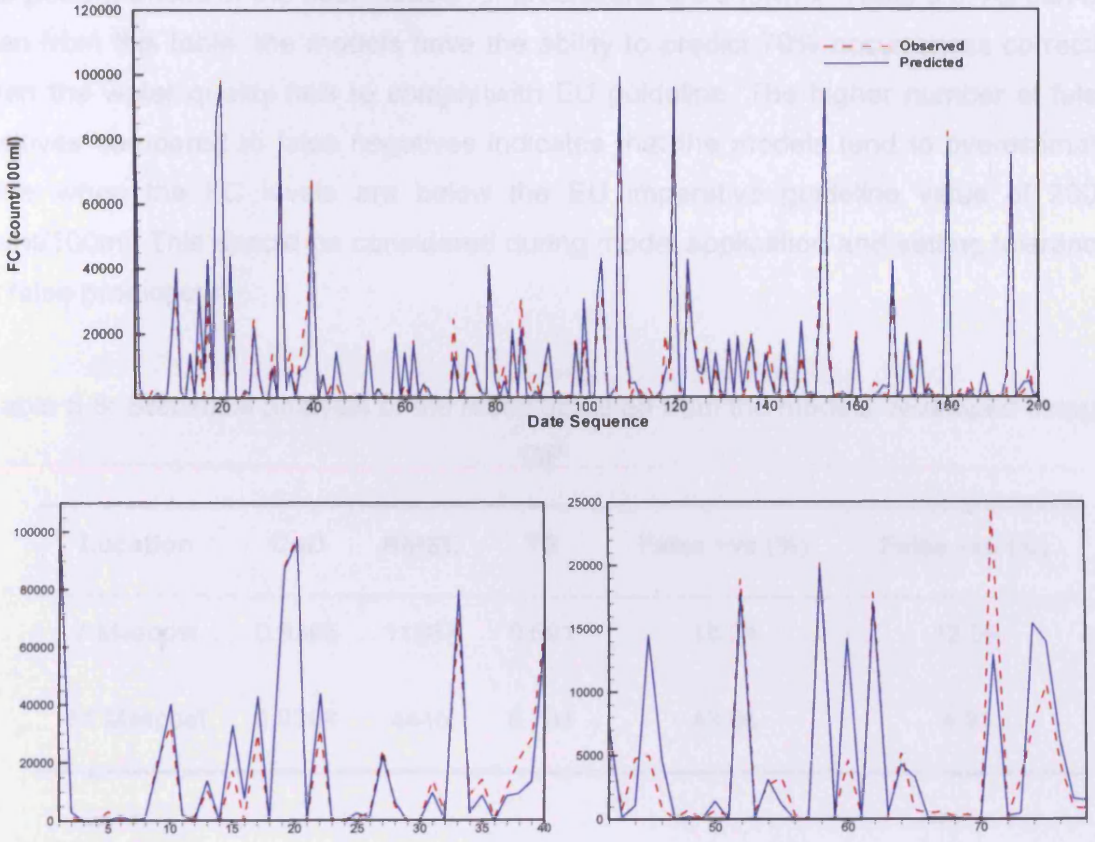


Figure 5.9: Comparison between observed and GP predicted FC levels at 11 Milepost

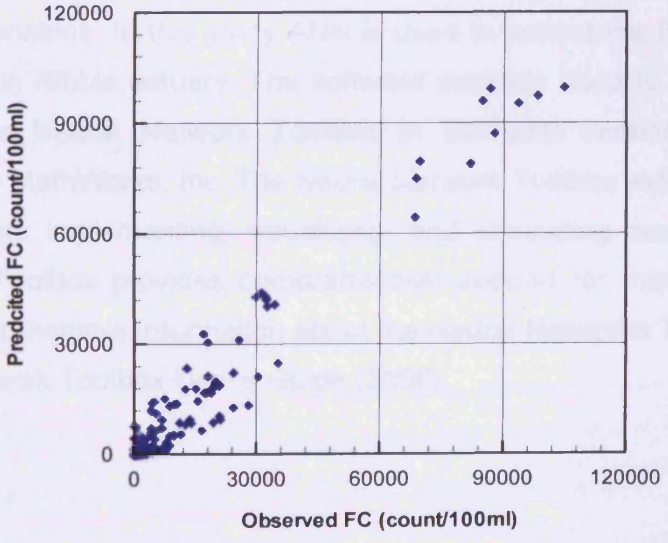


Figure 5.10: Scatter plot for observed and GP predicted FC levels at 11 Milepost

The performances of the best models for predictions are shown in Table 5.8. As can be seen from this table, the models have the ability to predict 70% occurrences correctly when the water quality fails to comply with EU guideline. The higher number of false positives compared to false negatives indicates that the models tend to overestimate value when the FC levels are below the EU imperative guideline value of 2000 count/100ml. This should be considered during model application and setting tolerance for false predictions.

Table 5.8: Statistical analysis of the result obtained from the models developed using GP

Location	CoD	RMSE	TS	False +ve (%)	False –ve (%)
7 Milepost	0.9395	11667	0.691	15.04	12.09
11 Milepost	0.9398	4445	0.731	43.06	4.91

5.7 FC modelling with ANN

Neural networks are increasingly being used for prediction and forecasting various water resource variables. In this study ANN is used to predict the FC concentration at various locations in Ribble estuary. The software package used to develop the neural network model is Neural Network Toolbox in MATLAB version 7.1 release 14, developed by The MathWorks, Inc. The Neural Network Toolbox extends MATLAB with tools for designing, implementing, visualizing, and simulating neural networks. The Neural Network Toolbox provides comprehensive support for many proven network paradigms. Comprehensive information about the Neural Networks Toolbox is available in the Neural Network Toolbox User's Guide (2006).

General Strategy

For the Neural Networks modelling, the flow discharges and concentrations at the rivers, salinity and depth data at boundaries have been used as input data, while the FC concentrations at the 2 selected locations were used as target data. The data were normalised for ANN model construction. The specifications of the rivers and the variables are listed in Table 5.9.

Table 5.9: Variable specifications used for ANNs

Variables	Symbol
Flow at Ribble River	Qrib
Flow at Darween River	Qdar
Flow at Douglas river	Qdoug
FC at Ribble River	FCrib
FC at Darween River	FCdar
FC at Douglas River	FCdoug
Water depth at 7Milepost	Dep7MP
Salinity at 7 Milepost	Sal7MP
Water depth at 11 Milepost	Dep11MP
Salinity at 11 Milepost	Sal11MP

Test Setup

The 3 rivers are the main contributors of FC in the whole estuary while the transport of FC to the estuary depends on, among other factors, velocity or flow of the rivers. Hence the FC levels and the flow of the three rivers were included as input for neural network model building. It was considered that the transport of FC is also heavily affected by the water level at the tidal boundary in the estuary; therefore the water depth on the location in question was also included as an input. Salinity level was also included as an input as salinity has a detrimental effect on the FC survival. It would have been ideal to use sunlight data as a parameter, especially since sunlight is an important factor for bacterial decay. However, as mentioned earlier, no sunlight data

were recoded during the original surveys and this parameter could not be incorporated in the hydrodynamic model. As a result the sunlight was also ignored for the neural networks.

The data were divided into three subgroups, including training, validation and testing. The training data were used to find optimal set of connection weights, the validation data were used to verify the trained network, and the testing data set was used to test the true generalisation capability of the model. However, in the literature testing data set and validation data set is used interchangeably. The way these sub division is used can have significant influence on the model performance. In this study the data set was divided in a way so that all of the patterns were presented in the calibration set. Finally it was ensured that the ranges of all input parameters were roughly the same in the three subsets. To obtain this goal the data order were randomised by sorting the contiguous block of data using a sequence of random numbers.

A total of 840 data points were used. According to the results from Gamma Test, 1/2 of the data were used as training set while 1/4 each used for validation and test dataset. The statistical parameters of the training, validation and test data is shown in Table 5.10.

It can be seen that the training set contains maximum values of most of the parameters. This is important as some literature suggests ANNs have a poor extrapolation ability, which could result in a sub-optimal model.

After dividing the data into 3 subsets the data were then transformed. In the past, it has been commonly been perceived that data standardisation is not necessary for ANN models. However, more recent studies claimed that generally data standardisation improves neural network's performance. The input variables were selected according to the possible relationships between the bacterial concentrations and those variables. In this study 4 different experiments have been carried out based on different combination of parameters. In order to avoid over-training the noise values obtained result from Gamma Test RMS error were used as the stopping criteria.

Feed forward network was used with back propagation learning algorithm. The number of hidden layer in the network had an effect on network performance. A complex network with too many hidden layers may reduce the generalisation ability and it has been shown that ANNs with one hidden layer can approximate any function, given that sufficient degrees of freedom (i.e. connection weights) are provided (e.g. Hornik et al., 1989). The number of nodes in hidden layer was determined through trial and error.

Table 5.10: Statistics of the training, validation and test data sets after data division

	Qrib	Qdar	Qdoug	FCrib	FCdar	FCdoug	Sal	Dep
Training Dataset								
Mean	51.92	5.19	3.28	40,117	29,607	1,401,337	10.75	3.89
Min	14.46	1.63	1.51	364	272	5,228	1.06	1.02
Max	134.24	17.58	9.97	449,000	257,000	34,000,000	23.82	6.75
Std Dev	35.73	3.43	1.59	88,458	49,261	4,344,983	7.53	1.92
Validation Dataset								
Mean	47.70	5.27	3.42	34,069	29,259	724,573	10.60	3.69
Min	14.41	1.62	1.52	348	1,178	5,228	1.06	1.08
Max	133.98	17.70	10.20	405,000	237,000	19,500,000	23.66	6.76
Std Dev	32.40	3.69	1.89	80,444	48,883	2,345,617	7.66	1.95
Test Dataset								
Mean	45.70	5.25	3.24	23,535	25,043	1,166,961	11.68	4.04
Min	14.41	1.63	1.51	364	1,360	5,408	1.14	1.08
Max	132.33	17.70	9.23	300,000	248,000	27,600,000	23.71	6.77
Std Dev	30.85	3.62	1.49	56,085	44,514	3,713,009	7.29	1.90

Four different experiments were carried out for this study. The parameters used in these experiments are listed in Table 5.11. The first experiment was carried out with only 8 parameters as inputs which were the flow and FC concentration levels at the upstream boundaries of the 3 rivers and the salinity level and depth predicted at the sites in question. The target was the FC at that location. The second experiment used the same input parameters but the FC and Flow values used were 9 and 11 hrs before

for 7 Milepost and 11 Milepost respectively. As mentioned earlier these are the receiving water response times, or time lags. It has already been known from the Gamma test that the time lag information improves the performance of the networks. Therefore two sets of experiment were carried out to quantify the impact of including this information on the neural network performance.

In Experiment 3 time series (i.e. 3, 6, 12, 18 and 24 hours before) values of input parameters were used as input parameters, while the output were the values of the FC concentration. The previous value of the FC concentrations in the locations (7 Milepost or 11 Milepost) were also used as inputs, thus the total number of inputs used in this case is 53. In Experiment 4, time lag information was used and the number of input parameters was reduced to 27.

Test Result and Analysis

The accuracy of the model predictions was evaluated using the Root Mean Square (RMS) error and the coefficient of determination (CoD). The RMS error measures the deviation of the predicted FC values from the observed values. The CoD values represent the extent to which the observed and predicted FC concentrations "varying together", i.e., whether a positive correlation exists. This is an analysis tool for examining whether large values of predicted FC tend to be associated with large observed FC values, and vice versa.

Table 5.12 shows the summary of the simulation results for the experimental runs. It can be seen from Table 5.12 that generally the statistical indexes obtained from the testing dataset are very close to those of the validation dataset, which indicates a good generalisation ability of the neural networks used in this study. This is primarily due to the fact that the noise level in the input data, which was predicted from the Gamma Test, was used as the stopping criterion in training the ANNs. In this way, the over-training problem, which often makes the ANN testing results significantly worse than the validation results, had been avoided.

Table 5.11: Parameters used for ANN models

Site	Exp.	No. of para-meters	Description
7MP[+]	1	8	Qrib, Qdar, Qdoug, FCrib, FCdar, FCdoug, Dep7MP, Sal7MP
	2	8	Qrib-9, Qdar-9, Qdoug-9, FCrib-9, FCdar-9, FCdoug-9, Dep7MP-9, Sal7MP-9,
	3	53	Qrib, Qdar, Qdoug, FCrib, FCdar, FCdoug, Dep7MP, Sal7MP, Qrib-3, Qdar-3, Qdoug-3, FCrib-3, FCdar-3, FCdoug-3, Dep7MP-3, Sal7MP-3, Qrib-6, Qdar-6, Qdoug-6, FCrib-6, FCdar-6, FCdoug-6, Dep7MP-6, Sal7MP-6, Qrib-9, Qdar-9, Qdoug-9, FCrib-9, FCdar-9, FCdoug-9, Dep7MP-9, Sal7MP-9, Qrib-12, Qdar-12, Qdoug-12, FCrib-12, FCdar-12, FCdoug-12, Dep7-12, Sal7MP-12, Qrib-18, Qdar-18, Qdoug-18, FCrib-18, FCdar-18, FCdoug-18, Dep7-18, Sal7MP-18, FC7MP-3, FC7MP-6, FC7MP-12, FC7MP-18
	4	27	Qrib-9, Qdar-9, Qdoug-9, FCrib-9, FCdar-9, FCdoug-9, Dep7MP-9, Sal7MP-9, Qrib-10, Qdar-10, Qdoug-10, FCrib-10, FCdar-10, FCdoug-10, Dep7MP-10, Sal7MP-10, Qrib-14, Qdar-14, Qdoug-14, FCrib-14, FCdar-14, FCdoug-14, Dep7MP-14, Sal7MP-14, FC7MP-6, FC7MP-8, FC7MP-10
11MP	1	8	Qrib, Qdar, Qdoug, FCrib, FCdar, FCdoug, Dep11MP, Sal11MP
	2	8	Qrib-11, Qdar-11, Qdoug-11, FCrib-11, FCdar-11, FCdoug-11, Dep11MP-11, Sal11MP-11
	3	53	Qrib, Qdar, Qdoug, FCrib, FCdar, FCdoug, Dep11MP, Sal11MP, Qrib-3, Qdar-3, Qdoug-3, FCrib-3, FCdar-3, FCdoug-3, Dep11MP-3, Sal11MP-3, Qrib-6, Qdar-6, Qdoug-6, FCrib-6, FCdar-6, FCdoug-6, Dep11MP-6, Sal11MP-6, Qrib-9, Qdar-9, Qdoug-9, FCrib-9, FCdar-9, FCdoug-9, Dep11MP-9, Sal11MP-9, Qrib-12, Qdar-12, Qdoug-12, FCrib-12, FCdar-12, FCdoug-12, Dep11MP-12, Sal11MP-12, Qrib-18, Qdar-18, Qdoug-18, FCrib-18, FCdar-18, FCdoug-18, Dep11MP-18, Sal11MP-18, FC11MP-3, FC11MP-6, FC11MP-12, FC11MP-18
	4	27	Qrib-11, Qdar-11, Qdoug-11, FCrib-11, FCdar-11, FCdoug-11, Dep11MP-11, Sal11MP-11, Qrib-12, Qdar-12, Qdoug-12, FCrib-12, FCdar-12, FCdoug-12, Dep11MP-12, Sal11MP-12, Qrib-14, Qdar-14, Qdoug-14, FCrib-14, FCdar-14, FCdoug-14, Dep11MP-14, Sal11MP-14, FC11MP-6, FC11MP-8, FC11MP-10

As the neural networks were intended to be used as a tool for day to day monitoring of bathing water quality, a comparison between the neural networks predicted number of failed samples and the observed (by numerical model) numbers was made.

[+] The numbers '-3', '-6' etc are used to refer to the values at 3 and 6 hours before, respectively.

It can be seen from Table 5.12 that the CoD correlation is relatively high and the RMS error is reasonably low in all cases. For example, for training and validation the correlation coefficient ranges from 82.9% (Experiment 1, 11 Milepost) to 99.6% (Experiment 3, 11 Milepost). For model testing, in which unseen data were used, the correlation coefficient ranges from 82.1% (Experiment 1, 7 Milepost) to 95.3% (Experiment 4, 11 Milepost). The maximum RMS error is 18666 (Experiment 1, 7 Milepost), which is less than 2% of the maximum FC concentration level.

ANN model performance improved with an increasing number of input parameters. The RMS errors obtained from Experiment 3 are smaller than those obtained from Experiment 1 and the number of predicted failed samples resulting from Experiment 3 is closer to the observation than that resulting from Experiment 1. Similarly, the predictions obtained from Experiment 4 are significantly better than those obtained from Experiment 2. Good correlations were obtained between the predictions made by the neural networks and the observed FC values, see Figures 5.12 and 5.14.

From Table 5.12 it can be seen that the predicted FC concentration level varies from 10^3 to 10^5, which shows a similar degree of variation in the FC level as found in observed data. However, the neural networks generally over-predict at low FC concentration levels, while under-predict at very high concentration levels, see Figures 5.11 and 5.13. In particular, the neural networks over-predict when the FC concentration levels are around 2000cfu/100ml. For the day to day water quality management, such predictions will be slightly conservative.

It can also be seen from Table 5.12 that the model results obtained from Experiment 4 are generally similar to those obtained from Experiments 3, even though the number of input parameters used in Experiments 4 was only half of that used in the Experiment 3. The testing CoD values obtained from Experiment 4 are slightly higher than those from Experiment 3, but the average CoD values are both over 90%. On the other hand, the number of currently predicted failed samples from Experimental 4 is slightly lower that that from Experiment 3, with the average error from both experiments being lower than 10%.

Table 5.12: Statistical analysis of the result obtained from models developed by ANNs

	Statistical measures		Number of failed sample		
	CoD	RMSE	observed	ANN	SE
Experiment 1:					
7 Milepost					
Training	0.936	11959	168	186	129
Validation	0.877	18666	89	101	72
Testing	0.821	17072	82	110	75
11 Milepost					
Training	0.957	4031	176	208	162
Validation	0.829	7069	91	164	91
Testing	0.869	5735	79	107	75
Experiment 2:					
7 Milepost					
Training	0.948	10802	168	186	129
Validation	0.914	15608	89	106	77
Testing	0.889	14343	82	91	56
11 Milepost					
Training	0.989	2024	176	180	169
Validation	0.933	4428	91	153	89
Testing	0.881	5474	79	117	77
Experiment 3:					
7 Milepost					
Training	0.992	4043	168	194	162
Validation	0.947	12256	89	92	77
Testing	0.926	10927	82	87	75
11 Milepost					
Training	0.996	1144	176	179	171
Validation	0.942	4122	91	152	90
Testing	0.885	5371	79	102	79

Experiment 4:					
7 Milepost					
Training	0.994	3664	168	195	157
Validation	0.949	12005	89	97	76
Testing	0.928	10790	82	84	72
11 Milepost					
Training	0.994	1439	176	178	173
Validation	0.944	4042	91	135	91
Testing	0.953	3435	79	107	74

Thus in constructing the ANN models if consideration is given to the hydrodynamic process, the number of input parameters can be significantly reduced. Also, all of the input data used in Experimental 4 are collected at least 6 hours before the prediction time. This is very useful for bathing water managers to give forward warning to potential visitors.

5.8 Comparison between GP and ANN Models

Unlike ANNs, GPs are not a purely black box models as they offers some symbolic expression, as a result it is easy to understand how a GP has come up with certain result. However, there are number of research projects ongoing (e.g. Setiono et al, 2002) where the objective is to extract an expression of the models developed by using ANNs. It is widely believed that ANNs are more efficient to deal with data with noise, although it worth noting that there is no noise present in the datasets used for this study, which is contrary to the real life measured datasets.

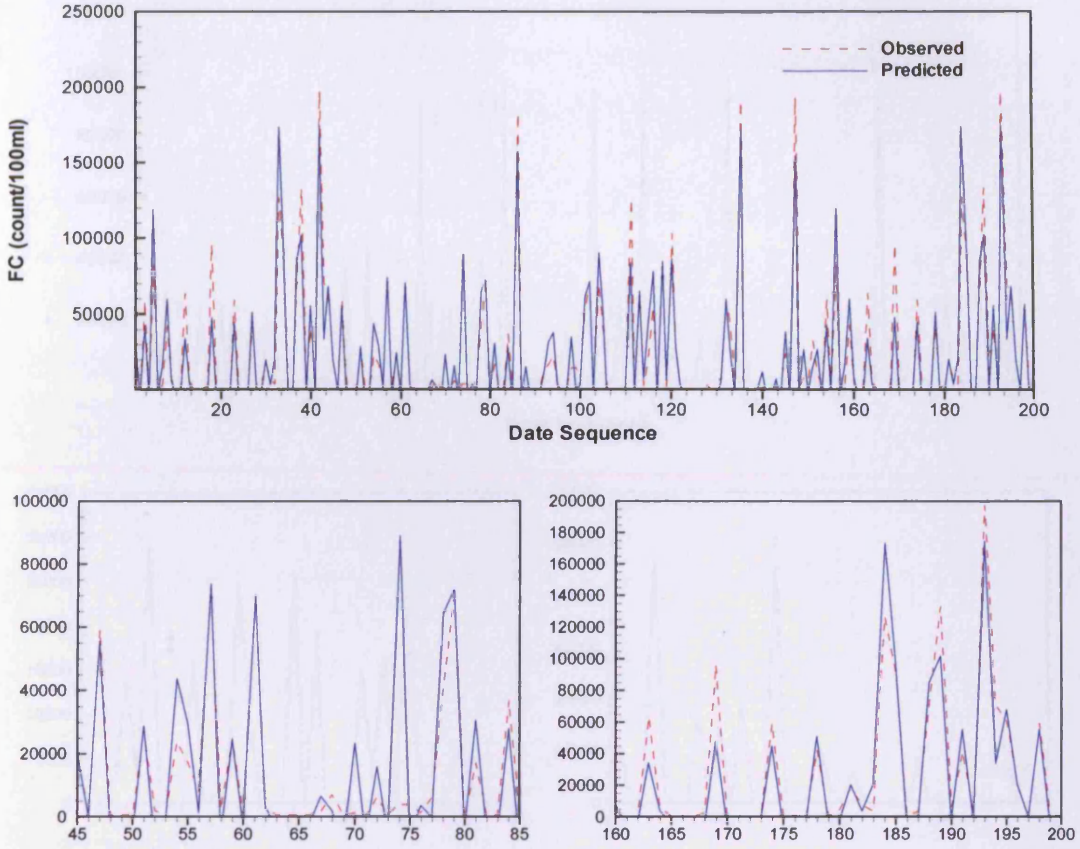


Figure 5.11: Comparison between observed and ANN predicted FC levels at 7 Milepost

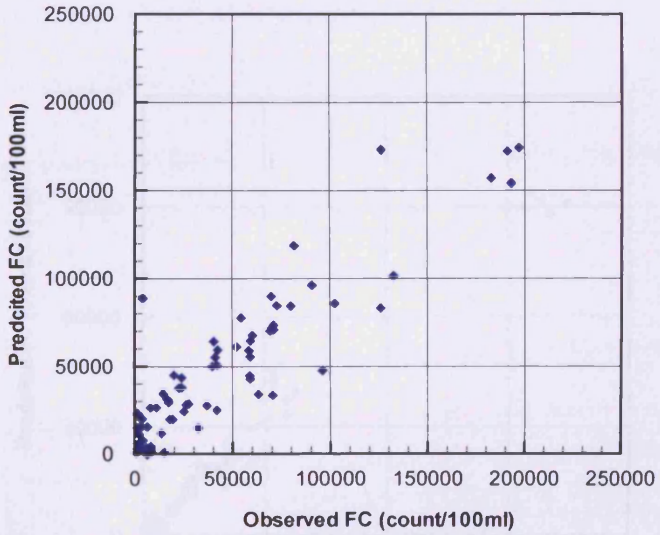


Figure 5.12: Scatter plot for ANN test data set for 7 Milepost

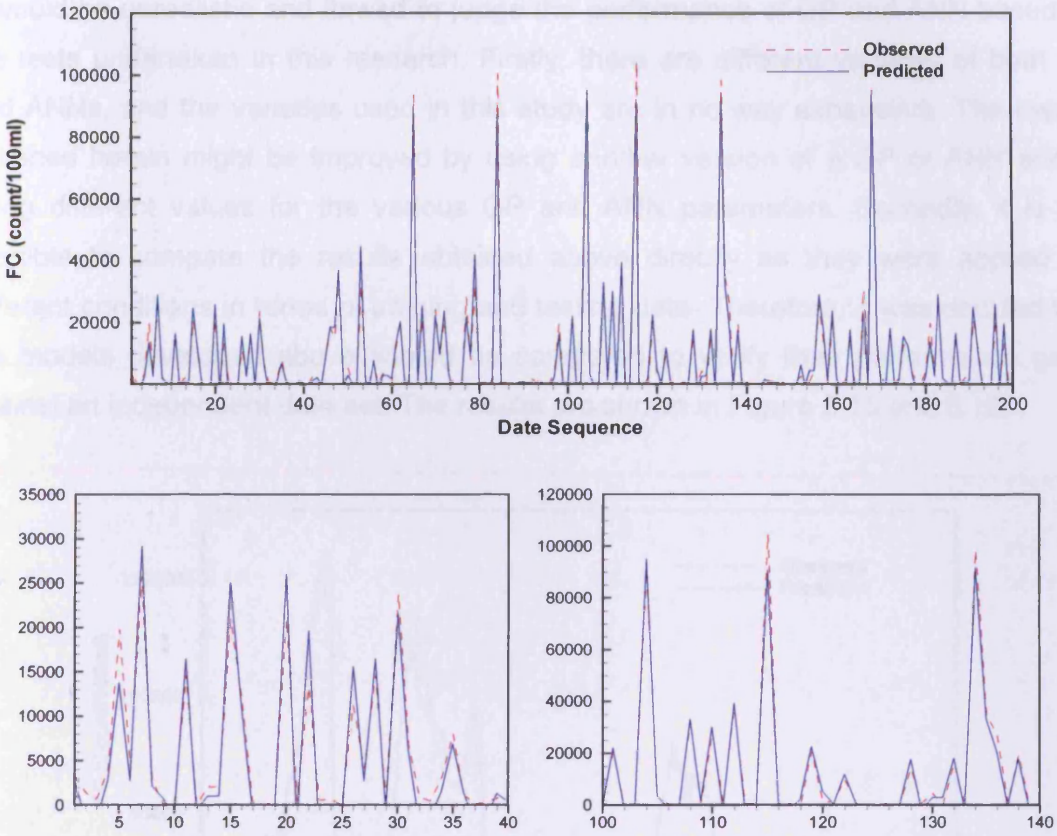


Figure 5.13: Comparison between observed and ANN predicted FC levels at 11 Milepost

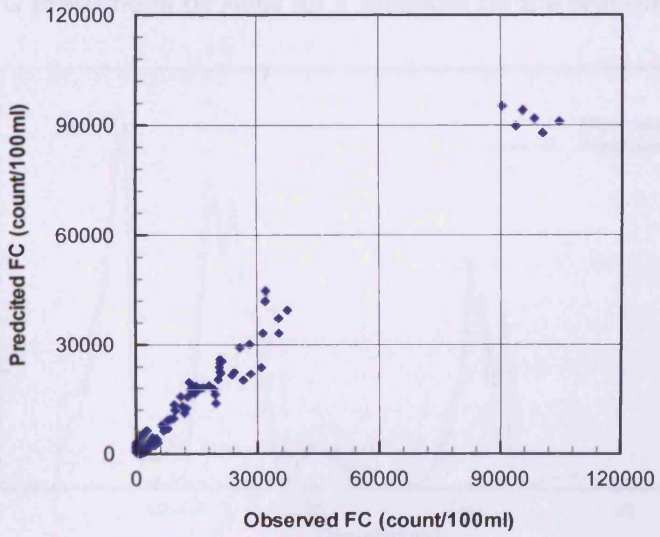


Figure 5.14: Scatter plot for ANN test data set for 11 Milepost

It would be unrealistic and flawed to judge the performance of GP and ANN based on the tests undertaken in this research. Firstly, there are different varieties of both GP and ANNs, and the varieties used in this study are in no way exhaustive. The results obtained herein might be improved by using another version of a GP or ANN and/or using different values for the various GP and ANN parameters. Secondly, it is not possible to compare the results obtained above directly as they were applied for different conditions in terms of training and testing data. Therefore, it was decided that the models developed above should be compared to verify their performance gains against an independent data set. The results are shown in Figure 5.15 and 5.16.

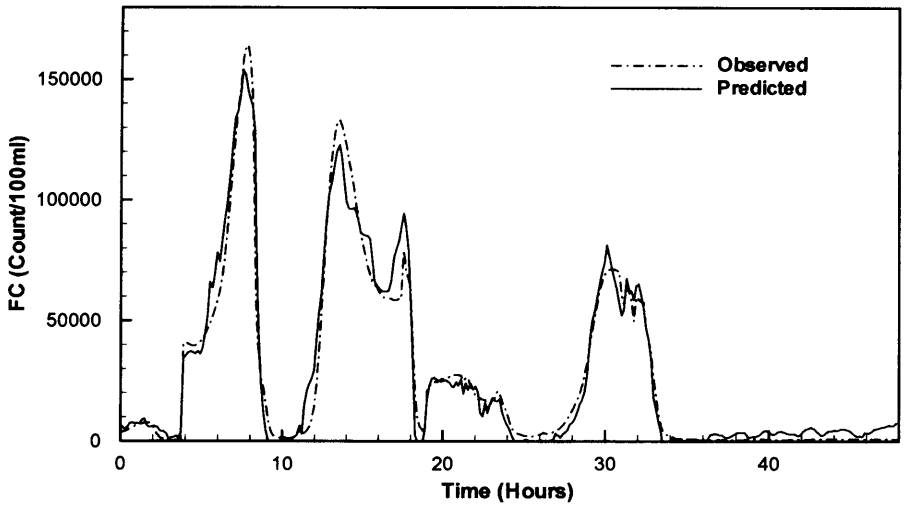


Figure 5.15: FC predictions by ANN for 7 Milepost on the test data time series

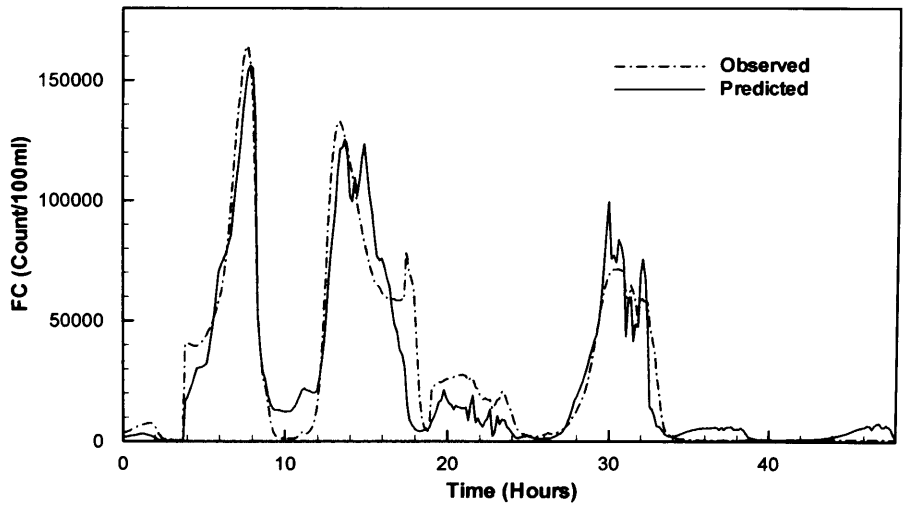


Figure 5.16: FC predictions by GP for 7 Milepost on the test data time series

From the plots it is clear that ANN model produced better prediction of the data although it also offered some negative values while the observed values were very low. Another aspect worth mentioning is the time needed for model development. The GP was run for around 12-21 hours to produce the results obtained while it took only around 10 minutes to obtain the results from the ANN models. However, once the model was developed it was faster to obtain results using the GP expressions compared to obtaining the results using the neural network models.

5.9 Summary

Modelling with ANNs and GPs is a relatively new approach being applied to the difficult problem of predicting FC concentrations in recreational waters. Contrary to conventional hydrodynamic models, GP and ANN models are *black box* models, in the sense that they do not need any physical insight to the problem in question at least to the same extent as a deterministic model. Although in practice some physical insight of the domain is necessary to supply the GP or ANN with all of the information necessary and at the same time avoiding any redundancy in the supplied data. One advantage of GP models are that they propose some symbolic expressions which offer some clues of the physical process, but it is generally very difficult to propose a physical model based on the GP model. It is also important to use real life data for data driven model development. The data used in this study was synthetic and fully noise free data, whilst in real life even a one day dataset would not be perfect. It would be interesting to see how the model reacts with the data including noise.

It is accepted that the black-box models, in general, do not work very well outside the conditions used for their development and calibration. A deterministic model can provide the necessary data to cover the whole range of possible cases. The results from this study have shown that these models are useful operational tools and possibly the only option for forecasting cases where there may be insufficient time available to run a hydrodynamic model. The development and application of such models as decision support tools could be a great benefit to environment managers and engineers involved in managing the safe use of bathing and recreational waters

Chapter 6

MODEL DEVELOPMENT AND APPLICATIONS TO CARDIFF BAY

6.1 Cardiff Bay Study

During the 1920s the Port of Cardiff was one of the largest trading ports in the UK. However with the decline in the coal mining and steel industry the port also experienced a prolonged period of decline. A massive urban regeneration of the docks area was seen as the most appropriate means of reviving the southerly part of Cardiff. A plan to construct a 1.4km long tidal exclusion barrage across the mouth of Cardiff Bay was considered as the focal point of this regeneration and was given Royal Assent in 1993. A freshwater lake, with a plan surface area of about 200 hectares, has been created following the construction of the barrage.

Cardiff Bay encompasses the estuaries of the Taff and Ely, which have contributing catchment areas of 512km^2 and 163km^2, respectively. Before impoundment the mean spring tidal range was 11.1m, with the estuary having the second highest tidal range in the world. The barrage impounds the rivers Taff and Ely at around 4.8m A.O.D. (above Ordnance Datum), which is close to the mean high water level (Edwards, 1997 and Jones, 1994). The creation of the bay has enhanced opportunities for sailing and recreational water use. Although major sewage and other outfalls have been diverted from discharging directly into the impounded waters, there are still inputs of sewage, industrial effluent and land drainage from the river catchments and there are discharges from some combined sewer overflows (CSOs) during high rainfall conditions (Hill et al., 1996).

Cardiff Bay is not currently designated a bathing water and therefore is not required to comply with the standards outlined in the EU Bathing Water Directive (Council of the European Communities, 1976). However, the waters are used significantly for a mix of

recreational uses, such as canoeing and sailing. Cardiff Harbour Authority (CHA), part of Cardiff City Council, has been responsible for monitoring the water quality and managing environmental operations within Cardiff Bay and the Harbour Limits since its formation in April 2000; this task has been undertaken by a specialist Water Quality team. Although the main responsibility of CHA is compliance with the Cardiff Bay Barrage Act 1993, i.e. maintaining dissolved oxygen levels of at least 5 mg/l, data are also collected for a range of water quality indicators such as dissolved oxygen, temperature, pH, conductivity and turbidity. The water quality team also samples the bay and river waters twice per week, to assess the bacteriological content in terms of Escherichia coli, total coliforms and faecal streptococci. River flow, wind speed and air temperature data are made available by the Environment Agency Wales and the Central Climate Unit of the Meteorological (Met) Office.

6.2 Previous Modelling Studies

A number of modelling studies have previously been undertaken of Cardiff Bay, pre impoundment, by Delft Hydraulics, Hydraulics Research and by Hyder Consulting Ltd using the Delft Hydraulics, Delft3D modelling software. Hydraulics Research constructed a three-dimensional model with a relatively fine vertical resolution, but a coarse horizontal resolution, to investigate sediment oxygen demand (SOD). Delft Hydraulics constructed a single box model of the impoundment and concentrated on the main processes contributing to oxygen supply and demand (Hyder Consulting Ltd 1997). The study undertaken by Hyder Consulting Ltd, using the Delft Hydraulics Delft3D modelling software, was to examine the potential for compliance with the Cardiff Bay Barrage Act (1993). The Barrage Act requires the dissolved oxygen levels in the Bay to be maintained at/or above 5mg/l in all places and at all times. As a result an aeration system has been installed to ensure this standard is maintained.

As detailed in Chapter 3 individual pathogens are generally difficult and expensive to measure and, therefore, in water quality studies it is common practice to measure and/or model the levels of related indicator organisms. Numerical model based on solving the solute transport and kinetic equations are used by governmental bodies, consultants and water companies for the prediction of the distributions of bacterial concentrations, particularly for assessing compliance with the EU Bathing Water

Directive (EU 1976). However these models can require some time to set-up, particularly when detailed bathymetric and boundary condition data must be obtained.

Data driven models, such as Artificial Neural Networks (ANNs), have been widely used in non-linear time series modelling of multivariate signal processing and controls, and in recent years there have been a number of successful applications in water management, for example Minns (1998) and DiBike et al (1999). Lin et al (2003) have recently applied ANNs to predict the compliance of coastal waters along the coastline of Firth of Clyde, Scotland, with the EU Bathing Water Directive (EU BWD).

Although not an EU designated bathing water, Cardiff Harbour Authority is collecting large quantities of water quality data for the Cardiff Bay impoundment relating to recreational water use. In the current study, Artificial Neural Networks (ANNs) and Genetic Programming (GP) have been used to predict the Faecal Coliform (FC) concentration levels at two specific location using the data collected within Cardiff Bay and the contributing rivers. Genetic Programming is able to generate symbolic non-linear regressions for the input output relationships. The development of a decision making tool that is able to predict the water quality and its variability will be of great benefit to the management and operation of Cardiff Bay, enhancing opportunities for safe recreational water use.

6.3 Data Availability

There are several continuous monitoring stations deployed around the Bay and the Rivers Taff and Ely. These stations operate continuously, measuring parameters such as: dissolved oxygen, temperature, pH, conductivity and turbidity. The water quality data collected is telemetered back to the Harbour Authority office via radio. The water quality team ensure the continual operation and maintenance of these stations. Depth profiling is carried out at a number of sites throughout the bay in parallel with monitoring water quality. The equipment used records general water quality indicators such as: dissolved oxygen (% saturation and mg/l), turbidity, conductivity, salinity, pH, temperature and depth. The Harbour Authority also carries out routine sampling of the bay and river waters for a number of key determinants, such as: biochemical oxygen demand, nitrogen, nitrates, ammonia, phosphorous, suspended solids and some

metals. Figure 6.1 and Table 6.1 detail the location of the sites where this monitoring work is regularly undertaken. The Water Quality team also samples the bay and river waters once a week, to assess the bacteriological content. The water samples are analysed for the presence of the eschericia coli, total coliforms and faecal streptococci. Less frequently, the water is analysed for the presence of enterovirus, cryptosporidium and salmonella. As mentioned earlier, Cardiff Bay is not a designated bathing water; however, the EU BWD standards are the only currently available measure of bacteriological quality for recreational waters and are being used by Cardiff Harbour Authority as a guide for public information.

Table 6.1: Locations and data availability at different sites

Site No	Position		Bacterio-logical Samples	Depth Profiling	Continuous monitoring	General Water quality
	Easting (m)	Northing (m)				
1	317811.32	176170.21	×	×		×
2	318037.16	175415.25		×		
3	318560.85	174610.46	×	×		
4	318238.61	173869.03		×	×	×
5	318930.4	173571.09	×	×	×	
6	319302.12	173830.79		×	×	×
8	319135.53	173350.31	×			
9	318330.9	173128.38	×	×	×	
10	318904.64	173076.21		×	×	×
15	318491.12	172466.15	×	×		×
16	318440.98	172583.18	×	×		×
17	317368.25	173133.96		×	×	
18	316935.41	173860.48		×		
19	316024.05	175106.62	×	×		×

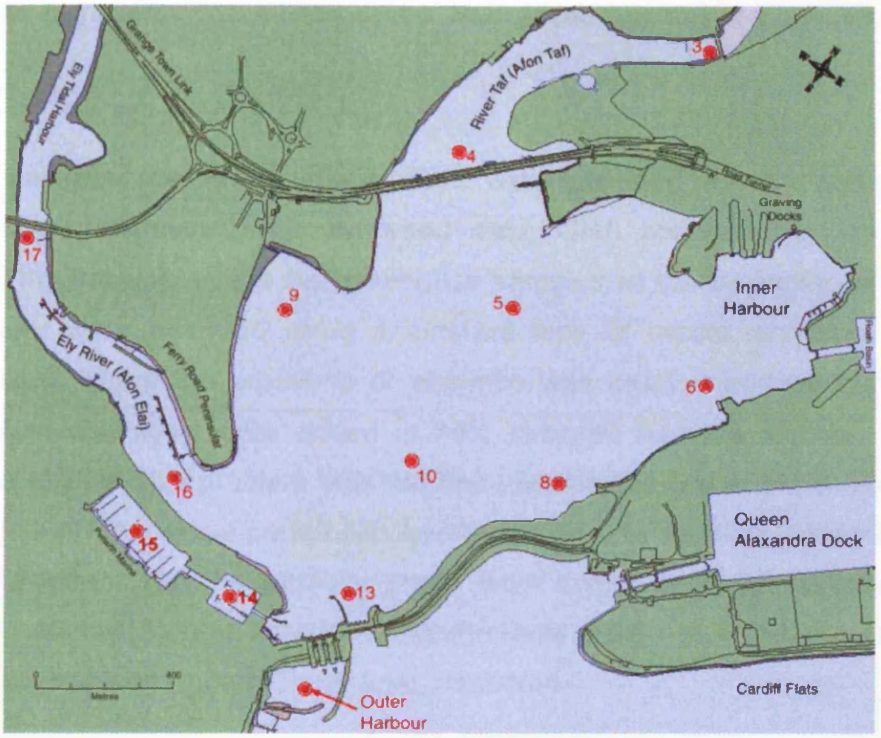


Figure 6.1: Schematic diagram of Cardiff Bay sampling locations

Data from Cardiff Bay Harbour Authority (CHA) was provided from the period of the beginning of summer 2001 (1/5/01) to the end of first quarter 2005 (30/3/05). The data provided was as follows:

- Bacteriological Water Quality - Samples from Cardiff Bay and the rivers Taff and Ely was collected on average twice per week to assess the bacteriological content. The water samples were analysed for the presence of faecal coliform and Escherichia coli (E. coli).
- Water Quality - Measurements were taken using continuous monitors in the bay (former tidal rivers). The measurements included dissolved oxygen (% saturation and mg/l), turbidity, conductivity, salinity, pH, temperature and depth.
- Meteorological Data – Data were collected at the CHA station and included: total daily radiation; maximum radiation over 5 minute intervals; radiation at mid-day; modal wind direction; mean wind speed; and, maximum wind speed
- River Taff and Ely flow data – The rivers were gauged at 15 minute intervals at Black weir for River Taff and former Arjo wiggins weir on River Ely, with the

data being provided by the Water Resources Section of Environment Agency Wales.

The bacteriological results for summer 2002 were excluded from the current study as there was a problem with the analysed data. CHA changed the laboratory that undertook the analysis of the bacteriological samples at the beginning of April 2002. The samples were analysed using a different type of media developed for testing potable water, where the presence or absence was more important than the actual number. Also chemicals were added to help stressed bacteria recover, thus further biasing the results. This problem was rectified towards the end of the summer of 2002, thus the data for the whole period has been excluded. In addition, during the summer months, CHA removed the turbidity probe from some sites and replaced it with a chlorophyll monitor. Where turbidity measurements were not a continuous annual (or summer/winter) record the readings were discarded.

6.4 Model Evaluation Criterion

The model evaluation in this study was carried out using the coefficient of determination (CoD) and the root mean squared error (RMSE). The error measurement consists of an analysis of the error between the observed and predicted values. The overall performance of trained neural networks can be judged with respect to criterion such as CoD. This coefficient is independent of the scale of data used and is useful in assessing goodness of fit of the model (Dawson and Wilby, 1999). CoD ranges from 0 at the worst case to +1 for a perfect correlation. RMSE were used to show a quantitative indication of the model error; which measures the deviation of the forecasted value from the actual observed value.

As the model developed herein is intended to be used for day-to-day monitoring of recreational waters, another additional criterion is also used for this particular study. In order to use these models for monitoring purposes it is more important to detect when the water quality fails to comply with the guideline threshold values, for example, the EU imperative limit of 2000 cfu/100ml. Lin et al. (2003) employed this type of evaluation criterion by reporting the number of days when the observed and predicted FC concentrations exceeded the regulatory water quality standard. In this study the

number of failed samples is reported, for both the observed and predicted sample as well as the number of occurrences when the failed sample was correctly predicted as failed.

6.5 Data Pre-processing

The data were first scanned for sensor malfunctions, missing values or data entry errors. In some cases these data errors continue for significant intervals. Their cause is unknown, but could be due to sensor failure or periods of routine maintenance. It was apparent that the data would have to be *cleaned* to reduce the effect of faulty sensor readings prior to analysis or modelling. The measurements that were significantly different from the previous reading, taken at the same site, were identified as 'incorrect measurement'. A simple threshold algorithm (twice the standard deviation of each time series), designed to operate in real-time when future sensor values would not be known, was used to correct obviously faulty readings by replacing them with their last known reliable value. This routine is effective with time-series values that can be expected to change relatively smoothly over time. It is obviously inappropriate for the FC concentrations, where the values were effectively discontinuous, and no attempt was made to adjust the FC data. For most cases where there were single missing values the data cleaning procedure provided a simple and effective approximation. A disadvantage of this technique occurs when a string of missing values are assigned the last valid measured value. For long strings this algorithm would in all likelihood produce increasingly inaccurate approximations.

6.6 Data Analysis

The non linear data analysis tool winGamma was used for the data analysis. The brief description of winGamma is given in Chapter 5. A simple Gamma test was run with pmax =10 (where pmax is the number of near neighbour, the prescribed value is 10). The result obtained is given in Table 6.2. The Gamma statistic is actually the vertical intercept of the regression line in Figure 6.2. This is the estimated variance of the errors for any smooth model built on the data. Table 6.3 shows that the unscaled values of Γ is 662894.9. This means that any smooth model built on unscaled data will have a standard deviation of the prediction error of approximately 815. In considering

the range of the output (FC concentrations) data [3, 32000], then this corresponds to about 2.54% of the range.

Table 6.2: Result of Gamma test for unscaled and scaled data

	Unscaled	**Scaled**
Gamma	662894.9	0.048079
Gradient	0.065911	0.15184
V-Ratio	0.014876	0.048316
Near Neighbours	7	7
Start Vector	1	1
Unique Points	230	230
Evaluated Output	1	1
Zero Nearest Neighbours	0	0
Lower 95% Confidence	0	0
Upper 95% Confidence	0	0

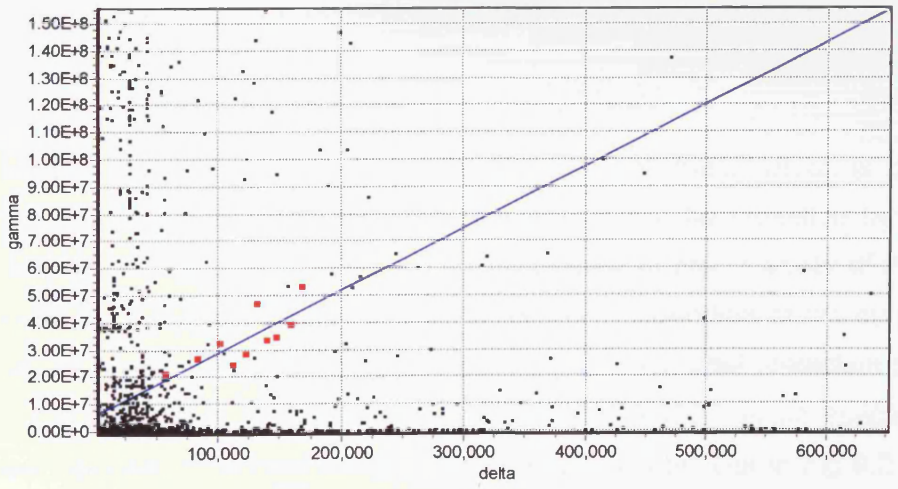


Figure 6.2: Scatter plot and regression line for the dataset used to build the model

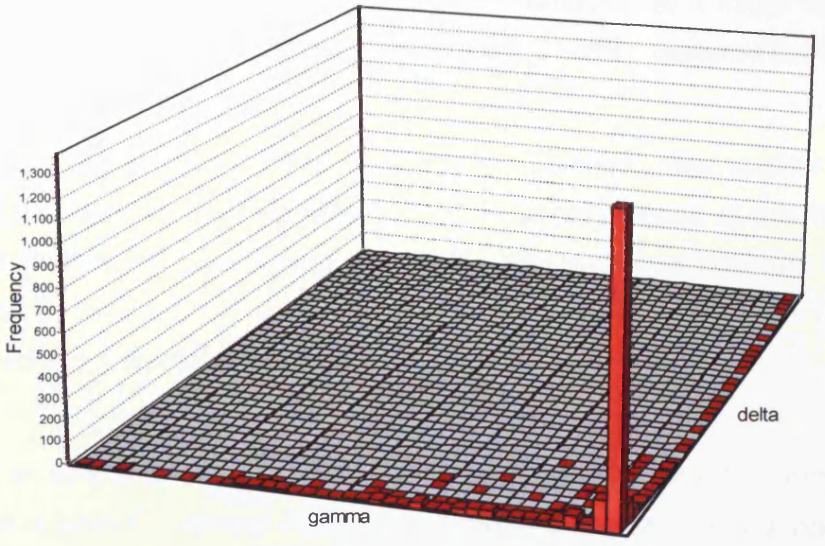


Figure 6.3: 3-D histogram for the dataset used to build the model

In the Gamma Test the critical graph is the scatter plots and the $(\delta(p),\gamma(p))$ regression line. The scatter plot shows point pairs (δ,γ), where δ is the squared distance of an input (**x**) from one of its near neighbours and γ is one half of the squared distance between two corresponding scalar output (y) values. The points to which the regression line is fitted are calculated by finding the mean $\delta(p)$ of δ and the mean $\gamma(p)$ of γ, where p refers to the first nearest neighbour (p=1), second nearest neighbour (p=2) and so on, up to the maximum number of near neighbours (pmax) which is set by the user.

Another important measure of data predictability is the Vratio, which is defined as Gamma/Var (output). It indicates how well the output can be modelled by a smooth function. A Vratio close to zero indicates a high degree of predictability of a particular output. Vratio is a better parameter to look at as it is independent of the output range. In this case the Vratio of 0.0148 indicates difficulty in the model prediction due to the presence of noise. Jones (2001) reported a Vratio of 0.0007 as an indicator of low noise presence. This observation is reinforced by the scatter plot in Fig 6.2. However the 3-D histogram in Fig 6.3 shows that the majority of the data are in good agreement. For model building purposes the noise was not removed as it might eliminate some useful information that would otherwise influence the output captured in the input.

For model building purposes the noise was not removed as it might eliminate some useful information that would otherwise influence the output captured in the input.

For the model building purpose it is important to know how quickly the estimate returned by the algorithm will stabilise to a close approximation of the noise variance. One simple method for quantifying this parameter is to compute the Γ statistic for increasing M. By plotting the Γ values over M it can be seen whether the graph appears to be approaching a stable asymptote.

Performing an M-test prior to model building can establish whether there is sufficient data to get a reliable Γ estimate. The fact that the graph has stabilised indicates that we have enough information (i.e. data) to estimate accurately the noise and so to construct a feasible surface, with the performance corresponding to the measured noise level. The Gamma test itself provides the criterion for ceasing training of a non-parametric model, such as a neural network. This is based on the idea that one criterion of a good model is that when tested on unseen data it can be expected to produce a root mean squared error (RMSE) which is the same as (or close to) the true or estimated noise variance of r associated with the data. Figure 6.4 shows the result of M tests performed for different combination the data at Site 9. It can be seen that Γ value did not reach asymptotic value when no FC data were provided (Figure 6.4(a)). It improves with the addition of FC data (Figure 6.4(b)) and Improves even further when the maximum values of some decay parameters were included (Figure 6.4(c)).

6.7 Model Inputs

It was initially intended to develop a model which would be able to predict the FC concentration at certain points within Cardiff Bay, without having any priori knowledge regarding the concentration levels at any point across the whole modelling domain. However, the M test results obtained during data analysis made it clear that it is not possible to build such a model without offering any FC data to the ANN. As a result the FC concentrations for Taff and Ely, at Black weir for River Taff (Site 1) and former Arjo wiggins weir on River Ely (Site 19) respectively, were also included in the data set. This data did improve M Test result, however, as the velocity in the rivers are generally very low, there is always a possibility that a significant fraction of the FC might decay by the

considered for the model development. For the individual tests different numbers of combinations of inputs were used to find out the most efficient combination. The list of the inputs included:-

- *Flow data:* The average flow for the preceding 24 hr for the Taff and Ely (2 parameters)

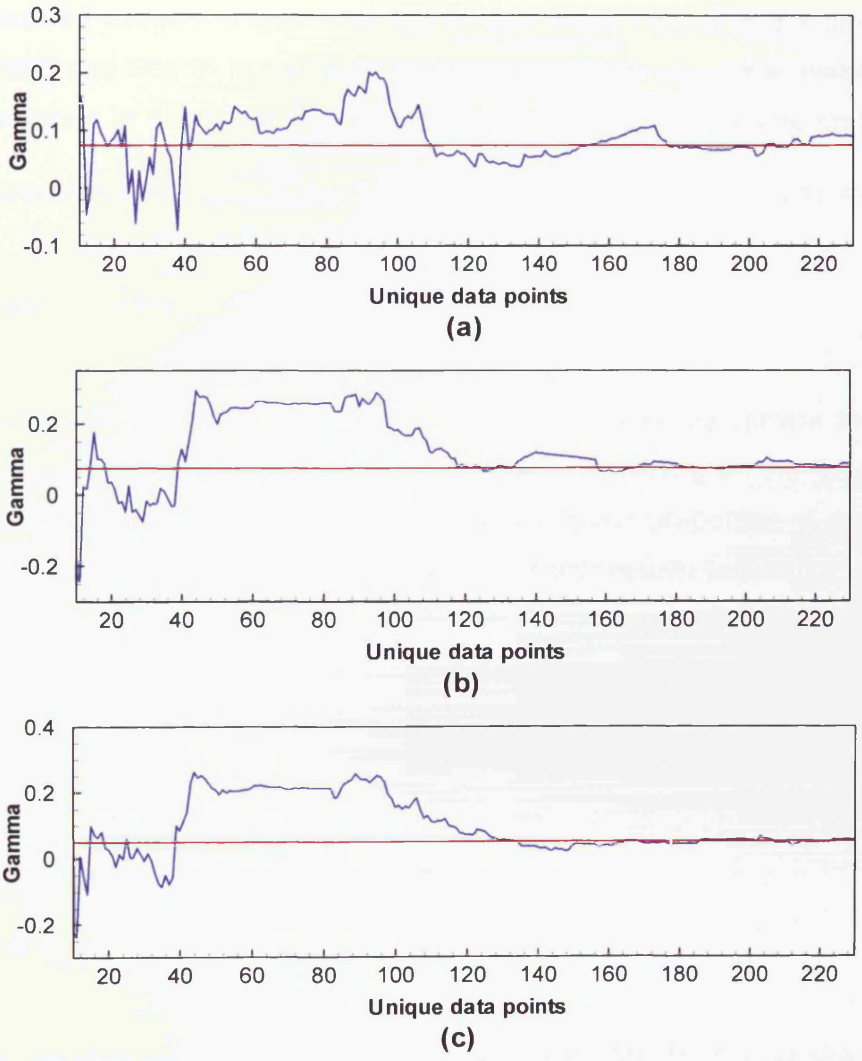


Figure 6.4: M Test performed on randomised scaled data, red line corresponds to the potential Γ value for FC at site 9: (a) when no FC data were provided (b) with FC data and (c) with FC data and maximum values of some decay parameters

- *FC data*: FC concentrations at Sites 1 and 3 (on river Taff) and 16 and 19 (at river Ely) (4 parameters)

- *FC data*: FC concentrations at Sites 1 and 3 (on river Taff) and 16 and 19 (at river Ely) (4 parameters)
- *Meteorological data:* The daily total radiation, maximum radiation and average rainfall. The average and maximum radiation and average rainfall for the previous 6 hr was also provided. (6 parameters)
- *Water quality data:* Average and maximum daily temperature, turbidity, pH and dissolved oxygen. It was intended to use these values at the target locations (Site 5 and Site 9) but all data for these locations were unavailable. Therefore the values at Site 4 and Site 10 were used instead. (16 parameters).
- *Water depth* average for preceding 24 hr at sites 4 and 10 (2 parameters).

6.8 Data Division

Once the input and output variables were defined, it was convenient to classify the complete data set into two categories: dry and wet, depending on the flow volumes for each quarter. All three data subdivisions contained same proportion of data from each period which was ensured for the analysis. The M test results previously indicated that at least 150 data were needed to build a model. To achieve this criterion 4 data points out of every 6 were assigned for training purpose. Finally, 157 data points were included in the training dataset and 38 data points included in validation and test data set.

6.9 Model Development using GP

6.9.1 Test Setup

The model development using Genetic Programming (GP) for the prediction of the FC levels, at Sites 5 and 9 in Cardiff Bay was undertaken using the academic version of a genetic programming software tool, namely Discipulus (AIMLearning™ Technology 2000). Discipulus is a fast technique for automatic computerised modelling using genetic programming and has been used in many engineering and scientific applications. It produces high precision models independently built from the data supplied. It has the advantage of being self-tuning and self-parametising. Once the tool

has been set-up and run, the output produces the details the performance of the best programs and allows the user to edit, optimise and simplify, via the interactive evaluator, the expressions developed through the program.

Two sets of solutions are produced in Discipulus, namely the best team and the best program. Team solutions are a combination of the best programs in the project, which frequently perform better that the individual program solutions. It evolves Pentium machine code for either numerical function fitting or binary classification problems, or runs on all Windows operating systems.

The GP parameters that had been looked at were the crossover rate, mutation rate, population size, instruction set and distribution of the initial program sizes, termination criteria, and parsimony pressure (i.e. fitness advantages for smaller programs). Table 6.3 shows the values used for the various GP parameters. These values were obtained after running a number of tests.

Table 6.3: Values of the GP parameters

Parameters	Value
population size	500
Mutation rate	0.85
crossover rate	0.50
parsimony pressure	0.20
Homologous crossover	0.90
Block mutation rate	0.30
Instruction mutation rate	0.25
Number of demes	10
Cross over between demes	0
Migration rate	1

As can be seen from Table 6.3, a high value for homologus crossover had been used for the GP runs. Homologous crossover is a recombination between equal length

program fragments in the same positions, in each parent. This reduces the tendency of evolved programs to become larger without correlated fitness improvements.

Discipulus has an option of using multiple groups of relatively isolated populations, known as demes (Banzhaf et al. 1998). This feature allows migrations between adjacent demes in a ring topology. Though sometimes demes produce better solutions, their main advantages are in supporting parallel computation or in increasing population diversity. Discipulus, however, does not support multiple processors and does not provide easy access to individuals in different demes. In this study this feature has been tested for several experiments but no convincing evidence was found that it improves the results, therefore this facility was not used for subsequent runs.

The inputs used for GP model building were flows in the two rivers, FC concentrations at 4 sites (namely, Sites 1,3, 16 and 19), temperature, dissolved oxygen, pH, turbidity and water depth at Sites 4 and 10. The maximum and average solar radiation were also included in the inputs. The output or target remained evaluating the FC at Sites 5 and 9 as before.

6.9.2 Test Results

As mentioned earlier, Discipulus produces two solutions for each problem; one relates to the best GP program and the other to the best team of solutions. Figure 6.5 and Table 6.4 show the individual best program and the team selection that have produced best results. It can be seen that the difference of the performance between the best team and the best program was minimal in this case, therefore for the sake of simplicity the best program solution had been accepted as the GP solution.

The evolved program for both the best program and best teams can be seen as machine code. Figure 6.6 shows part of the evolved program for Site 5. It can be seen that the evolved code is rather incomprehensible.

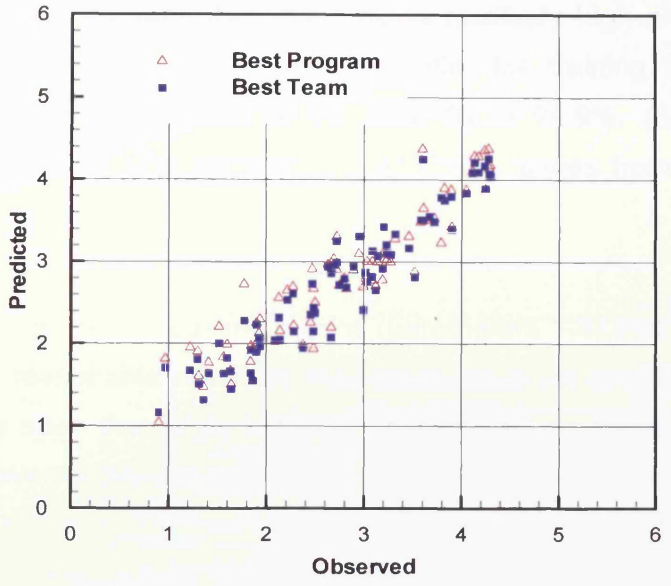


Figure 6.5: Correlation between measured and GP predicted FC Concentration (best program and best team)

However, it does give an indication of which input parameters were used frequently and what operation had been performed with these parameters. A better representation of the relative importance of the input parameters can be found from the impact table, which is provided in Table 6.5 and Figure 6.7.

Table 6.4: Performance of best GP program and best team

		Best Team		Best Program	
Site	Subset	RMSE	CoD	RMSE	CoD
Site 5	Training	0.217	0.949	0.253	0.932
	Validation	0.242	0.935	0.284	0.913
	Test	0.327	0.863	0.389	0.809
Site 9	Training	0.266	0.933	0.326	0.900
	Validation	0.327	0.893	0.342	0.884
	Test	0.427	0.785	0.447	0.756

From Table 6.4 it can be seen that, the CoD is relatively high and RMS errors are reasonably low in all experiments. For example, for training and validation the correlation coefficient ranges from 88.4% (Site 9) to 94.9%. (Site 5). The model performance on test data (the correlation coefficient) ranges from 75.6% (Site 9) to 80.9% (Site 5).

Figure 6.7 shows that the FC concentrations (parameters 1-4) had the most dominant effect, which is a reasonable result as they were the main contributors of FC in the domain. It can be seen that while the other parameters had been used in a similar frequency, their overall impact on the model out was very low. It suggests these parameters have been used to perform the fine tuning of the model and improve the model accuracy.

```
f[0]-=f[1];
cflag=(f[0] < f[1]);
f[1]+=f[0];
f[1]/=f[0];
f[0]/=-1.924433708190918f;
f[0]/=-1.907608032226563f;
f[0]*=v[0];
f[0]/=v[2];
f[1]*=f[0];
cflag=(f[0] < f[0]);
f[0]=sqrt(f[0]);
tmp=f[1]; f[1]=f[0]; f[0]=tmp;
f[0]=sqrt(f[0]);
if (cflag) f[0] = f[1];
f[0]*=f[1];
if (cflag) f[0] = f[0];
f[0]-=f[1];
f[0]-=v[2];
f[1]*=f[0];
```

Figure 6.6: Generated expression for FC prediction at Site 5

Table 6.5 : Impact of input parameters in the best GP programs

	Site 5			Site 9			
V	F	Av	Max	F	Av	Max	Variable Description
1	0.50	0.16	0.81	0.63	0.05	0.21	FC concentrations at site 1
2	0.23	0.13	0.76	0.83	0.22	0.89	FC concentrations at site 3
3	1.00	0.46	0.76	1.00	0.72	0.90	FC concentrations at site 16
4	0.57	0.11	0.42	0.37	0.05	0.17	FC concentrations at site 19
5	0.20	0.07	0.14	0.23	0.02	0.02	Average* Flow at Taff
6	0.23	0.03	0.09	0.60	0.05	0.18	Average Flow at Ely
7	0.83	0.03	0.07	0.17	0.05	0.07	Average Radiation
8	0.47	0.03	0.06	0.17	0.01	0.02	Maximum Radiation
9	0.13	0.03	0.05	0.23	0.05	0.05	Average Temperature at Site 4
10	0.43	0.02	0.05	0.30	0.01	0.01	Average DO at Site 4
11	0.07	0.04	0.04	0.10	0.00	0.00	Average pH at Site 4
12	0.03	0.04	0.04	0.40	0.03	0.07	Average Turbidity at Site 4
13	0.43	0.02	0.04	0.20	0.04	0.07	Average Depth at Site 4
14	0.07	0.01	0.01	0.13	0.00	0.00	Average Temperature at Site 10
15	0.03	0.00	0.00	0.17	0.00	0.00	Average DO at Site 10
16	0.00	0.00	0.00	0.27	0.04	0.11	Average pH at Site 10
17	0.00	0.00	0.00	0.03	0.00	0.00	Average Turbidity at Site 10
18	0.00	0.00	0.00	0.17	0.02	0.02	Average Depth at Site 10

Where V = Variable input

F = Frequency, percentage of best 30 programs containing input

Av = Average effect of removing all instances of input form best 30 programs

Max = Maximum effect of removing all instances of input form best 30 programs

* All average values and maximum radiation are taken for preceding 24 hours

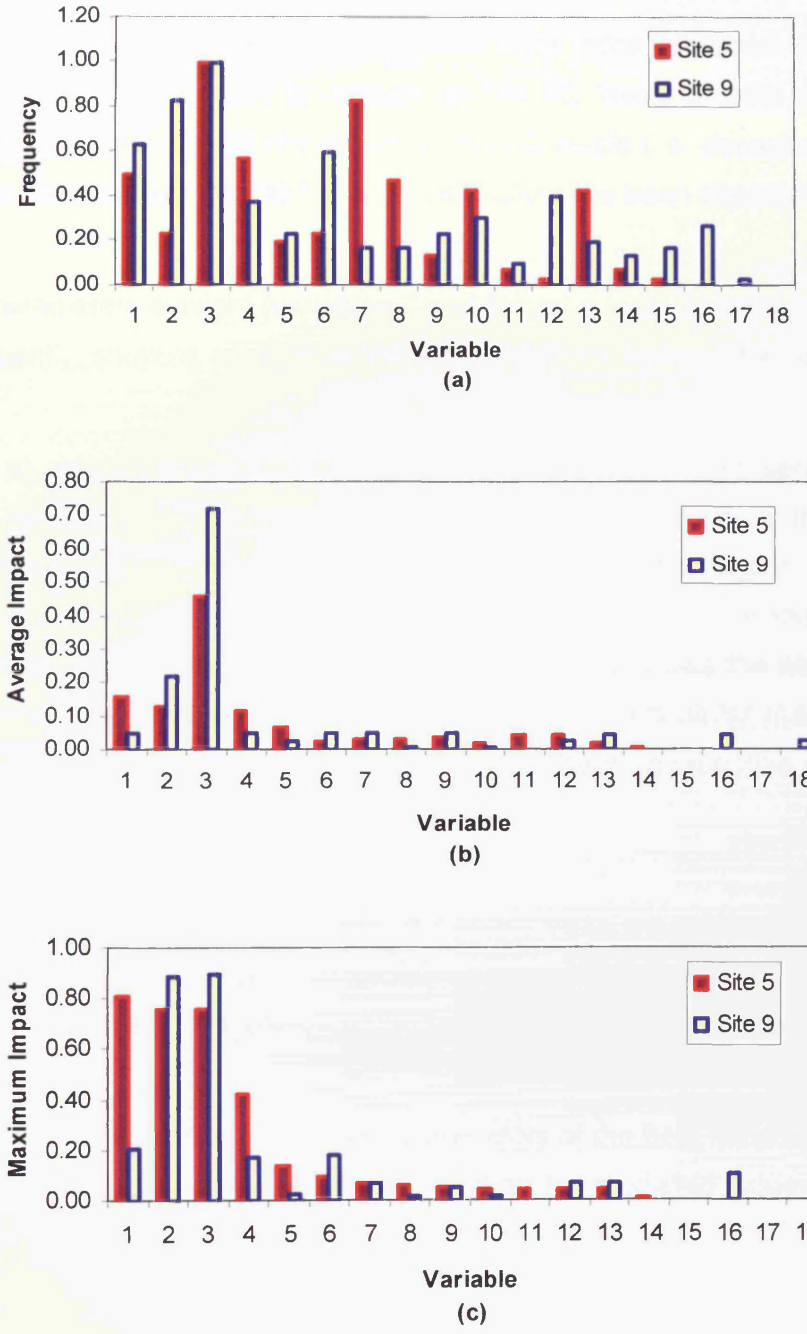


Figure 6.7: Impact of different parameters on the FC levels at Sites 5 and site 9. (a) frequency of the parameter within best 30 program (b) average impact and (c) maximum impact of parameters

As for the water quality parameters, Site 4 values were more influential than those at Site 10. Whilst Site 4 values had an impact on the FC levels of both of the target locations, water quality at Site 10 did not affect the FC levels (i.e. decay) at Site 5. As Site 10 is further downstream of Site 5 this result would have been expected anyway.

Among other parameters sunlight (variables 7 and 8) had a more prominent impact and was used frequently, showing sunlight as the most important factor of bacterial decay.

Figure 6.7 (b, c) also shows that the FC levels at Site 9 were more sensitive to the water depth. This finding can be explained by the fact that Site 9 was in the shallower region, where the water level fluctuations might result in a situation when solar irradiation penetrated most, if not the whole, of the water column. On the other hand Site 5 being in the deeper water (i.e. Figures 6.17 and 6.18 shows the water depth in Cardiff Bay), there was always a region where solar irradiation never reached part of the water column ensuring a more favourable environment, irrespective of the water level in the domain.

It is also clear that the flow in the rivers Taff and Ely were more influential in the FC levels at Sites 5 and 9 respectively. However, while the Taff flow had some affect on the FC at Site 9, flow in Ely had relatively little impact on the FC levels at Site 5.

Figure 6.8 to 6.11 show the line graph and scatter plots of the best results produced by the GP, for both Sites 5 and 9. It can be seen that the predicted values are well in agreement with the observed values.

Table 6.6 shows the performance of the GP models for detecting the samples which were above the EC Bathing Water Directive imperative value of 2000 cfu/100ml. It can be seen that most of the failed samples were detected by the GP models. The percentage of samples identified wrongly ranged from 0 - 30%, however it should be noted that that the number of failed samples was very low for this experiment, which resulted in a high percentage of errors.

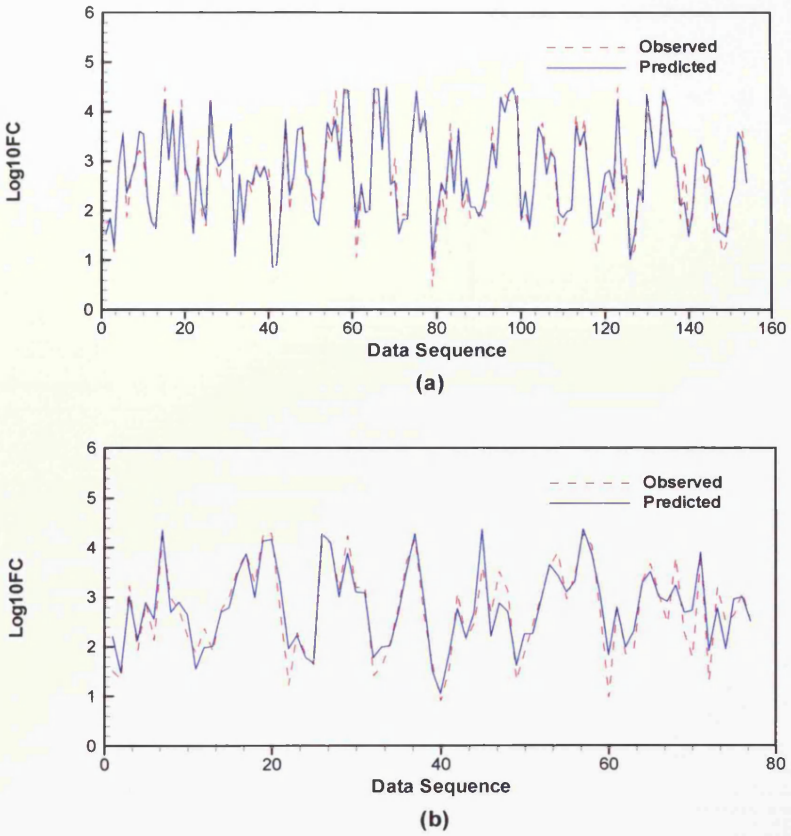


Figure 6.8: Observed and GP predicted FC concentrations (logarithmic value) at Site 5 for : (a) training, and (b) validation (1-38) and test (39-76) dataset

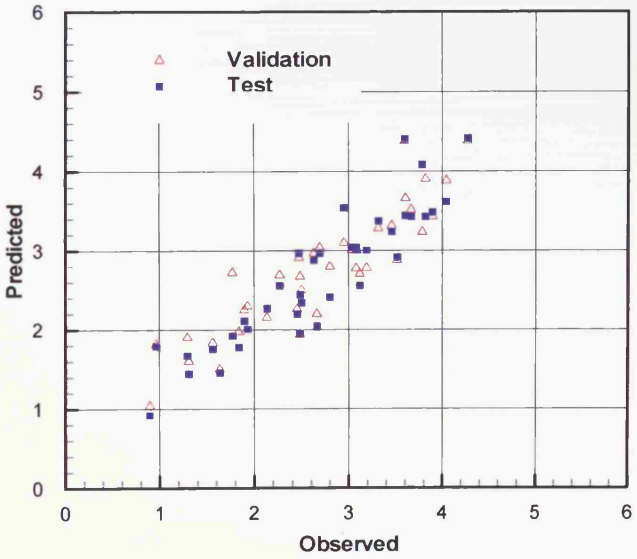


Figure 6.9: Comparison of GP predicted and measured FC concentrations at Site 5

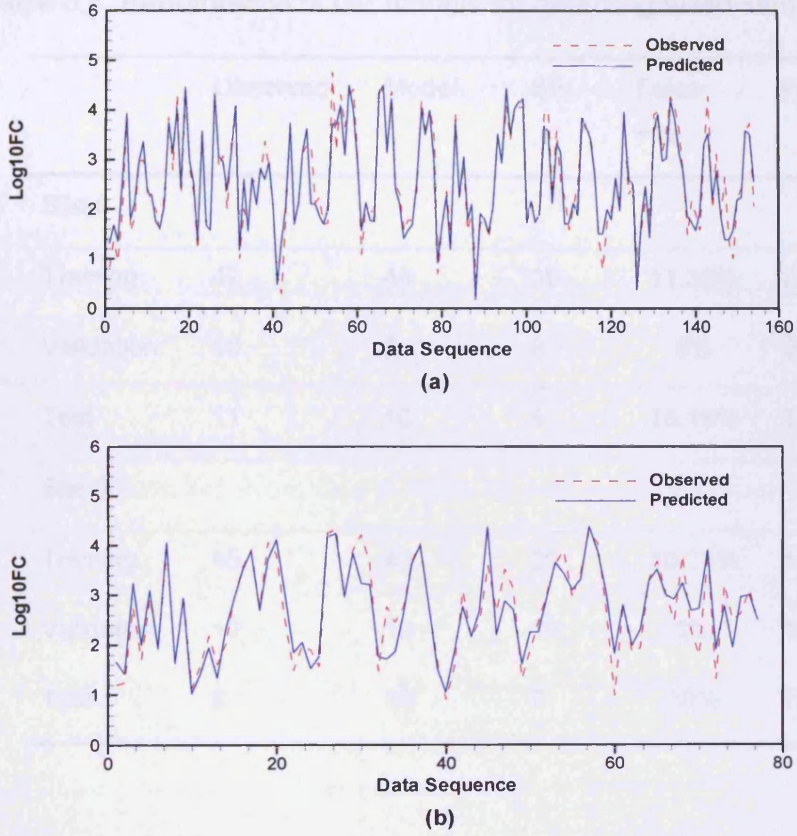


Figure 6.10: Observed and GP predicted FC concentrations (logarithmic value) at Site 9 for: (a) training and (b) validation (1-38) and test (39-76) dataset

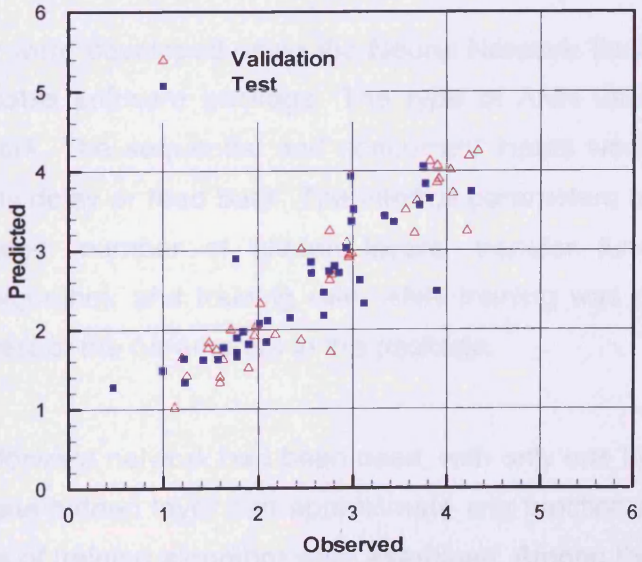


Figure 6.11: Comparison of GP predicted and measured FC concentrations at Site 9

Table 6.6: Performance of GP models for detecting failed sample

	Observed	Model	SE	False +ve	False -ve
Site 5					
Training	42	44	39	11.36%	7.14%
Validation	10	8	8	0%	20%
Test	11	10	9	18.18%	20%
Site 9					
Training	45	43	39	10.25%	13.33%
Validation	12	10	10	0%	16.67%
Test	8	10	7	30%	12.5%

6.10 Model Development using ANN

6.10.1 Test Setup

ANNs in this study were developed using the Neural Network Toolbox in Matlab 7.0, a commercially available software package. The type of ANN used in this study was Feedforward network. The sequential and concurrent inputs were applied to a static network without any delay or feed back. The internal parameters of the ANN that were manipulated included: number of hidden layers, transfer functions, input–output scaling, learning algorithm, and training rate. ANN training was performed using the default values for rest of the parameters in the package.

A multi-layer feed forward network had been used, with only one hidden layer as it had been shown that one hidden layer can approximate any function (Hornik et al. 1989). Four different types of training algorithm were examined. Among the training algorithms that have been tested, the Levenberg-Marquardt Back Propagation algorithm delivered the best results.

In order to avoid overfitting, the early stopping method was used. In this technique the available data are divided into three subsets. The first subset is the training set, which is used for computing the gradient and updating the network weights and biases. The second subset is the validation set. The error for the validation set is monitored during the training process. The validation error will normally decrease during the initial phase of training, as does the training set error. However, when the network begins to overfit the data, the error for the validation set will typically begin to rise. When the validation error increases for a specified number of iterations, the training is stopped, and the weights and biases at the minimum of the validation error are returned. While the early stopping method being applied for Levenberg-Marquardt Back Propagation algorithm, some training parameters were set to a prescribed value (e.g., momentum terms, mu as 1, mu_dec as 0.8 and mu_inc as 1.5) (Demuth and Beale, 2003) so that convergence is relatively slow. The network goal is selected as RMSE = 0.05 as achieved during Gamma analysis. The ANNs were trained starting from 20 different initial networks, randomly initialised, with the best performing network on training data being chosen as the trained network. A tan-sigmoidal unit was chosen for the hidden layer after a series of run using other functions with a linear transfer function always being used in the output layer.

The scaling of the network inputs and targets is done by normalisation, based on the mean and standard deviation of the dataset. The inputs and targets have been normalised in such a way that they will have zero mean and a unit standard deviation. The outputs of the model are then converted back into their original scale. The number of nodes in the hidden layer is determined by trial and error. Figure 6.12 shows that 10 hidden nodes provided the best RMSE for site 5.

It should be noted that 7 hidden layers produces the lowest validation error, however as 10 hidden nodes produces same level of RMSE for both Test and Validation dataset, it offers better generalisation.

Han (2002) (Han et al., 2007) proposed the following relationship for number of hidden layers:

$$\textit{Number of hidden nodes} = (\textit{Number of input neurons} + \textit{Number of input neurons}) \times \frac{2}{3}$$

Following this relationship there should be 13 (considering 19 inputs and 1 target) neurons in this case. It can be seen that 13 hidden nodes does not produce the best result, however further investigation is needed before making a conclusive comment on Han's work.

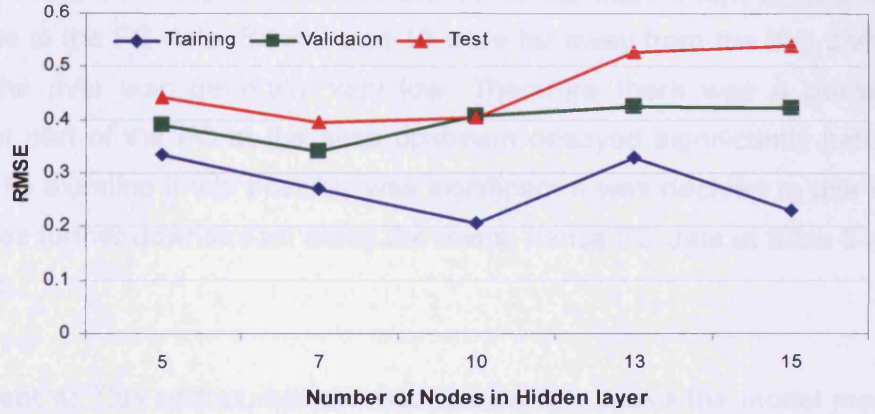


Figure 6.12: The Effect of hidden nodes on network performance

6.10.2 Network Inputs

Selection of the inputs is a very important aspect of building a successful network model. This is to avoid costly data collection, to eliminate unwanted impacts of irrelevant data and to build a simple model. This is referred to as part of the complexity regularisation problem. The primary criterion of complexity regularisation problem involves selection of an appropriate number of inputs and hidden neurons for a network. Among the input parameters stated in section 6.7 different input combinations were tested. These combinations are described as experiments and details are given below:

Experiment 1: FC data for 4 sites (Sites 1, 3,16 and 19), daily average flow of rivers Taff and Ely (2 nos), daily average radiation, maximum radiation (2 nos) and daily average temperature, dissolved oxygen, pH, turbidity and water depth for Site 4 and 10

(2 × 5 nos). All these input values were taken for the preceding 24 hr. Therefore a total of 18 parameters were used.

Experiment 2: This test was designed to minimise the number of FC inputs for model development. Therefore only FC dataset for site 1 and 19 were used, while the other inputs remained the same as for Experiment 1. The total number of inputs was 16.

Experiment 3: This experiment was the same as Experiment 2. The only change made was to the FC data. Sites 1 and 19 were far away from the bay domain and the flow in the river was generally very low. Therefore there was a possibility that a significant part of the FC at the sites upstream decayed significantly before reaching the bay. To examine if this possibly was significant it was decided to use the FC level of two sites further downstream along the rivers, hence FC data at Sites 3 and 16 were also used.

Experiment 4: This test experiment was intended to make the model more compact, thereby offering minimal input data to the neural net. From chapter 2, it can be seen that solar radiation and temperature were the most important factors affecting bacterial decay, particularly in a relatively small and freshwater body such as Cardiff Bay. Therefore apart from the flows in the two rivers and the FC levels at the boundary (Sites 3 and 16) the daily average and maximum solar radiation and daily average temperatures were included in this test. It must be stressed that all inputs values were taken for the preceding 24 hours. The total number of inputs in this test was 7.

6.10.3 Test Results

The results obtained from the ANNs are shown in Table 6.7. It can be seen that, generally the statistical indexes obtained from the testing dataset are very close to those of the validation dataset, which indicates a good generalisation ability of the neural networks used in this study. This is primarily due to the fact that the noise level in the input data, which was predicted from the Gamma Test, was used as the stopping criterion in training the ANNs. In this way, the over-training problem, which often makes the ANN testing results significantly worse than the validation results, has been avoided.

From Table 6.7 it can be seen that, the CoD is relatively high and RMS error is reasonably low in all experiments. For example, for training and validation the correlation coefficient ranges from 74.6% (Experiment 2, Site 9) to 91.3%. (Experiment 4, Site 9). It can be seen that the model worked best on the training data set. However, for the model performance it is important to check how the models work on test data set which had been unseen to the model during model construction process. For model testing, the correlation coefficient ranges from 53.7% (Experiment 2, Site 9) to 84.6% (Experiment 1, Site 5).

For Site 5, Experiment 1 thus produced the best result, while Experiment 4 produced the best result for Site 9. This indicated that the water quality (i.e. turbidity, pH, DO) and water depth played a more significant role on bacterial decay for Site 5 than they did for Site 9.

A performance enhancement in the range of 20 to 45% (of CoD) from Experiment 2 to Experiment 3 can be seen. This shows that a significant portion of bacteria died off before reaching the bay from those upstream sites, in comparison to those from sites 3 and 16. As this decay process can not be captured entirely by the neural net, more decay therefore leads to inferior results.

Figures 6.13 and 6.15 show the best results for both sites, which includes line plot for the training, validation and test datasets. Figures 6.14 and 6.16 show the correlation between the observed and predicted FC concentrations.

Finally, for the best results the number of failed samples was determined for the observed and predicted FC for both sites (Table 6.8). It can be seen that the models can successfully determine the occurrence as to when FC concentrations exceed the EU imperative value.

Table 6.7 Statistical analysis of the results obtained from different ANN models

	Site 5		Site 9	
	RMSE	*CoD*	*RMSE*	*CoD*
Experiment 1				
Training	0.292	0.9127	0.374	0.869
Validation	0.359	0.871	0.368	0.863
Test	0.361	0.846	0.487	0.723
Experiment 2				
Training	0.4619	0.765	0.508	0.767
Validation	0.475	0.747	0.527	0.746
Test	0.532	0.647	0.636	0.537
Experiment 3				
Training	0.374	0.850	0.451	0.828
Validation	0.444	0.788	0.457	0.802
Test	0.425	0.768	0.464	0.741
Experiment 4				
Training	0.311	0.893	0.304	0.913
Validation	0.353	0.859	0.448	0.801
Test	0.452	0.740	0.458	0.750

Table 6.8: Performance of ANN models for detecting failed sample

	Observed	Model	SE	False +ve	False -ve
Site 5					
Training	42	45	41	8.89%	2.38%
Validation	10	12	10	16.67%	0%
Test	11	10	10	0%	9.09%
Site 9					
Training	45	45	42	6.67%	6.67%
Validation	12	14	10	28.57%	16.67%
Test	8	10	7	30%	12.5%

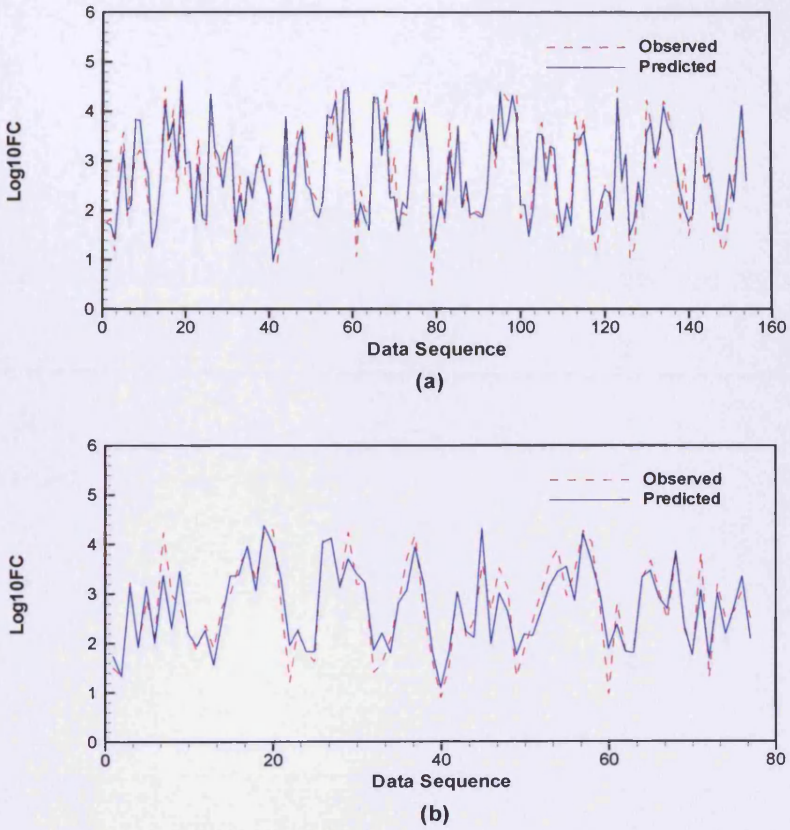


Figure 6.13: Observed and ANN predicted FC concentrations (logarithmic value) at Site 5 for: (a) Training and (b) Validation (1-38) and Test (39-76) dataset

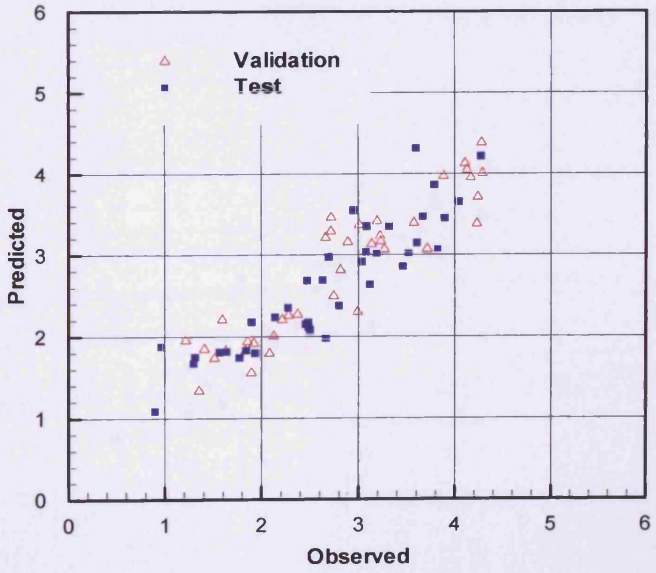


Figure 6.14: Comparison of ANN predicted and measured FC concentration at Site 5

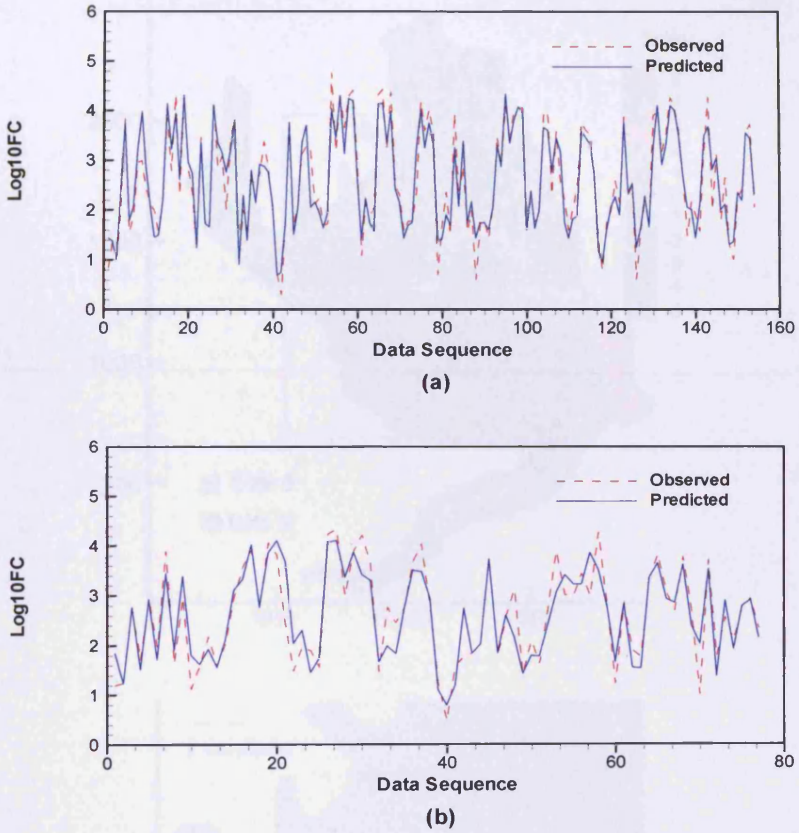


Figure 6.15: Observed and ANN predicted FC concentrations (logarithmic value) at Site 9 for: (a) Training and (b) Validation (1-38) and Test (39-78) dataset

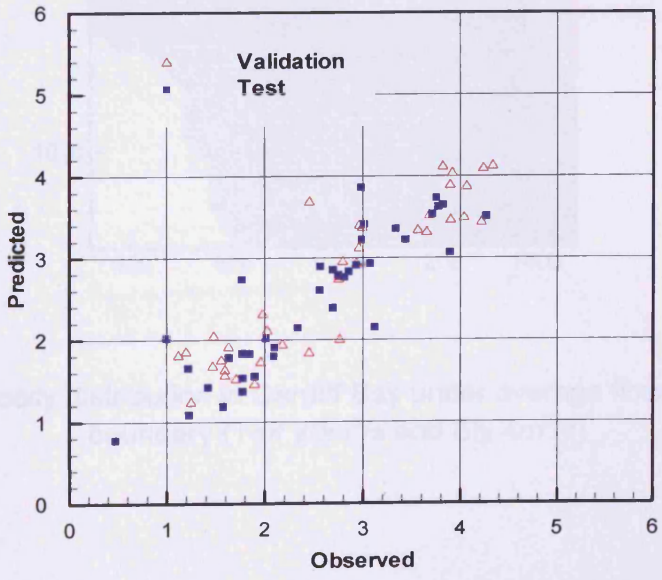


Figure 6.16: Comparison of ANN predicted and measured FC concentration at Site 9

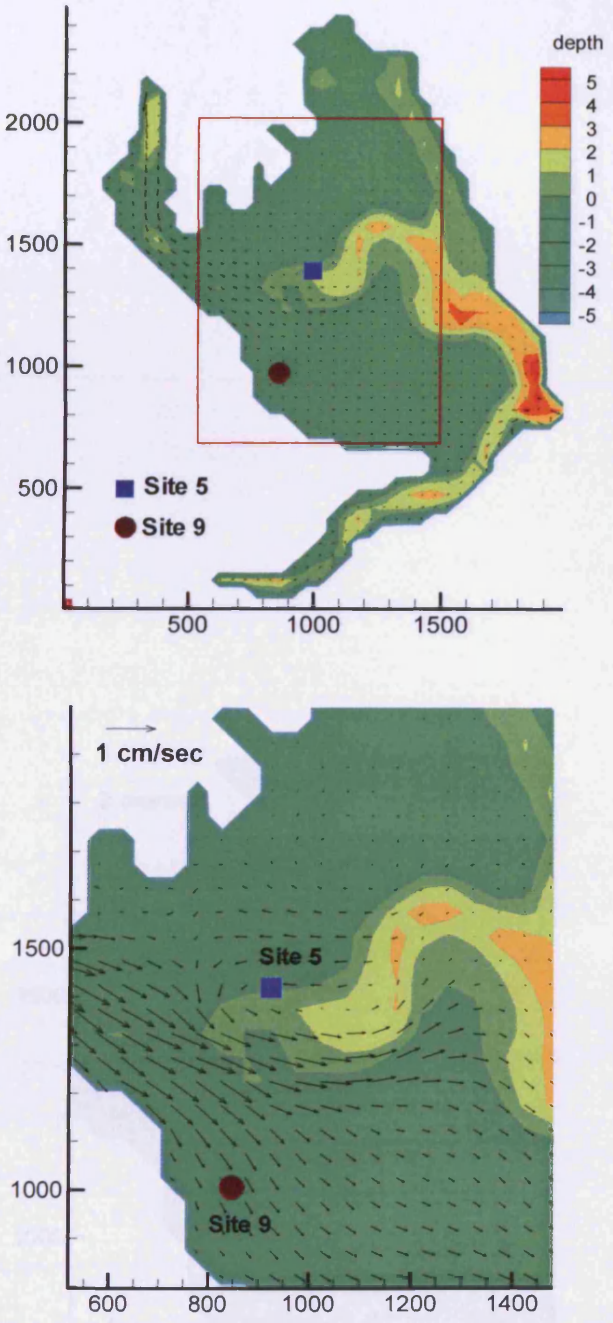
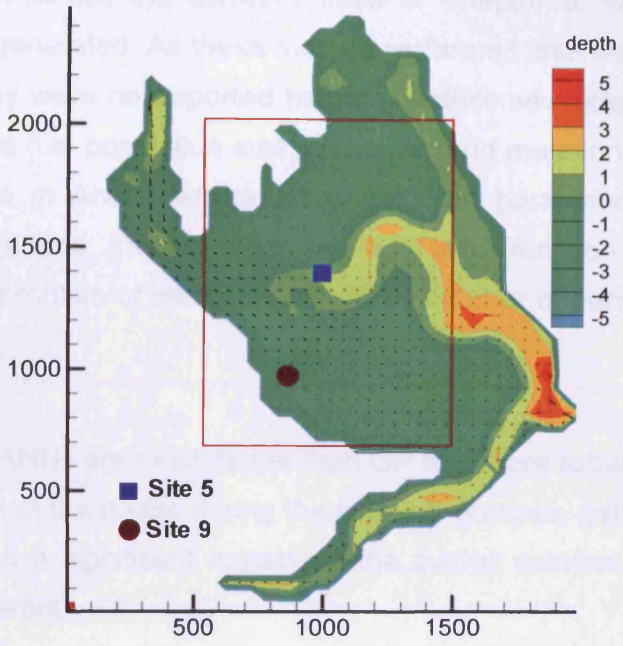


Figure 6.17: Velocity distribution in Cardiff Bay under average flow conditions at the boundary (Taff 20m^3/s and Ely 4m^3/s)



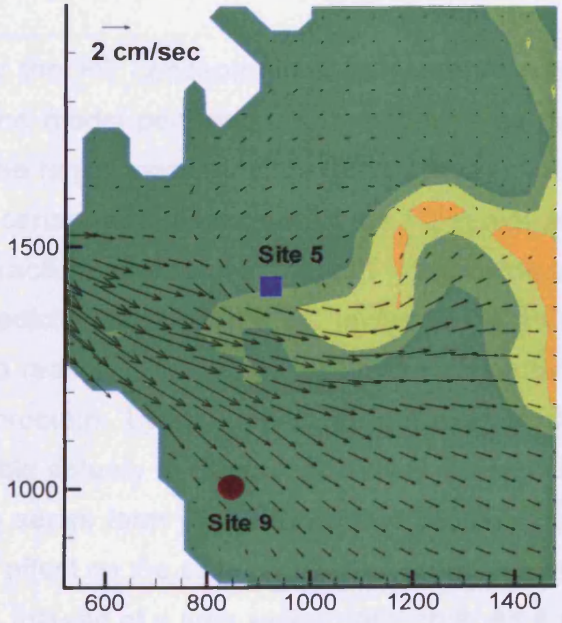


Figure 6.18: Velocity distribution in Cardiff Bay under a higher flow conditions at the boundary (Taff 40m^3/s and Ely 12m^3/sec)

However, when GPkernel, the software used in Chapter 5, was used, a series of expressions were generated. As these models performed less well then those reported in this chapter, they were not reported herein. Another advantage of GP is that there are less parameters (i.e. population size, crossover and mutation probability) in the GP compared to those in ANNs, where many heuristic parameters like network type, number of layers, nodes, training algorithm, activation function etc have included. It thus alleviates the problem of identifying the large number of parameters necessary for model optimisation.

On the other hand ANNs are much faster than GP and more robust. As ANNs gradually changes the weight of the nodes during the learning process, getting few weight wrong does not have such a significant impact on the overall solution, making ANNs more robust and fault tolerant.

6.12 Limitations and Uncertainties

The model predicts the FC concentrations at certain locations, given FC levels at nearby locations. The model performance deteriorates as those known locations are further away from the target location. If the model is supplied with FC levels at some locations, then at a certain time it can predict the FC levels at some target locations at that time. From a practical viewpoint this might not provide a significant advantage in terms of forward prediction since, if the FC levels at the known locations are already available there is no reason to know the FC levels of the target sites, unless there is some accessibility problem. These models do not produce future predictions as was the case for the Ribble estuary in Chapter 5. This is entirely due to the lack of FC data availability in a time series form. The data was collected once a week but it is highly unlikely to have any affect on the data for another week, hence these data were treated as discrete patterns, instead of a time series data input. As a result due to the absence of a time series for FC levels at Sites 5 and 9 it was not possible to build a data driven model to predict FC levels as shown in Chapter 5 for Ribble estuary and using synthetic data. However, the current study shows the possibility and extent to which the data driven model can handle relatively random natural phenomenon such as FC levels. The exercise undertaken in this chapter shows the capability of ANN and GP to handle noisy data from a natural environment and to predict an unknown value. Hence,

it should be possible to develop a model able to predict a time series with a similar level of accuracy.

Carabin et al (2001) reported that bacteriological analysis of environmental samples often provides imprecise estimates of the number of colonies, especially when the number is small. Considering this level of limitations of bacterial count, the accuracy level of the prediction obtained by the data driven models are reasonably good.

Another contentious issue related to such data driven models is that they produce a whole host of solutions for the same problem domain. The reason for the creation of different models is due partially to the stochastic nature of GP and ANN inputs and due to the multiple optimal solutions to the model-fitting problem.

6.13 Summary

The study details the application of ANNs to the problem of FC level predictions in a natural water body. A non linear data analysis tool, namely winGamma was used to assess the data quality prior to model building. It also helped in determining two heuristic aspects of the neural network model building, particularly when to cease the training of neural net and how much data were needed for building a smooth model. This study showed that the neural networks and genetic programming models could be successfully applied to predict FC levels when other approaches cannot succeed, due to the uncertainty and the complicated environmental interaction for bacterial decay. Different test cases were investigated in order to asses their ability and relative performance in encapsulating the site-specific knowledge and data necessary to reproduce the spatial distribution of FC observed in a modelled area.

CHAPTER 7

VELOCITY PREDICTIONS FOR COMPOUND CHANNEL FLOWS WITH VEGETATED FLOODPLAINS

7.1 Introduction

Proper modelling of flow resistance and conveyance capacity of wetlands and vegetated floodplains is very important for river and wetland management. Over the years, wetland areas have been treated as potential sites for agricultural or industrial development. In many developing countries, and particularly in areas of mangrove forestation, the destruction of these wetlands has been undertaken for land reclamation, shrimp farming, timber and chemical production. However, in recent years there have been a number of devastating floods world-wide and engineers and environmental managers are increasingly promoting the restoration or recreation of vegetated zones as a form of natural protection.

It has been found that in general vegetation increases the flow resistance in rivers, changes the back water profile and influence sediment transport and deposition (Yen, 2002). Vegetation induces natural flood alleviation as both the biomass and root systems produce turbulence and reduce the mean flow, thus decreasing the flow energy and attenuating flood flows to provide a naturally integrated flood protection system. Vegetation also acts to stabilise banks, helping to maintain deep channels and thus protecting the coastline and floodplains. However, wetlands remain one of the most complex and poorly understood ecosystems in terms of water management (Harris et al 2003). Current interest in the construction and restoration of wetlands has led to the need to understand the physical, chemical and biological processes that control such ecosystems.

7.2 Vegetation in Channels

Flow vegetation interaction is a complex process, and research in this area has led to significant simplifications in practical applications. Conventional approaches use standard reference publications, such as Chow (1959) and Barnes (1967), to select a roughness coefficient or employ a simple semi-empirical method for their estimation. Recently, attempts have been made to develop physically based models and to relate resistance to the measurable characteristics of vegetation and flow. Though significant advances have been made in the field, the effects of vegetation on flow resistance are still not fully understood (Tsihrintzis 2001).

Previous research in the area of vegetated floodplains has primarily focused on the adaptation of theory driven resistance equations, such as the Manning or Chezy equations. As a result although improvements have been made in the analysis to some degree, the resistance due to form drag has been coupled directly to the bed resistance. This is not the most comprehensive method of dealing with the problem, as vegetation, particularly surface piercing and flexible vegetation, is independent of the bed shear and therefore acts separately. Chow (1959) gave a comprehensive list of typical values for Manning's n values incorporating various factors like surface roughness, vegetation, sediment transport, channel irregularity etc.

A majority of subsequent research on vegetative flow resistance is based on theory and experiments with rigid cylindrical elements. Li and Shen (1973) studied the effects of tall non-submerged vegetation on flow resistance by investigating the wake caused by various cylinder set-ups. Experimental results indicated that different patterns or groupings of cylinders significantly affected flow rates. This wake correction approach was incorporated into the methods of Jordanova and James (2003). Li and Shen (1973) identified four factors that need to be considered in determining the drag coefficient: 1) the effects of open-channel turbulence; 2) the effect of a non uniform velocity profile; 3) the free surface effects; and 4) the effect of blockage. Petryk and Bosmajian (1975) formulated an approach that would calculate the Manning coefficient and also include the vegetation drag coefficient. They calculated the drag force and related this to the shear stress. They estimated Manning's n as a function of hydraulic radius and vegetation density for non-submerged rigid vegetation. Pasche and Rouvè

(1985) and Nepf (1999) investigated the wake caused by various cylinder set-ups resulting in methods to determine drag coefficients for single plants in a group and as a separated friction factor of the vegetation.

Kadlec (1989) developed a coupled equation for the friction slope of a channel based on the Mannings equation modified for laminar flow, thereby employing a Reynolds dependent equation. His work was inspired from the fact that most of the previous research was based on turbulent flow, whereas for the case of overland flows the slopes and depths are frequently not large enough to meet the turbulence criterion.

Naot et al. (1996) investigated the flow in a compound channel with a vegetated floodplain which included the shading effects of multiple cylinder wakes on the velocity distributions. They developed two equations calculating the degree of shading, one for aligned cylinders:

$$D_{SA} = \left(1 - \sqrt{\frac{D_t}{S}}\right)^2 \tag{7.1}$$

and another for a randomly arranged distribution of cylinders:

$$D_{SR} = 1 - \frac{D_t}{S}\left(1 - 0.5\sqrt{\frac{D_t}{S}}\right) \tag{7.2}$$

Where D_t = average vegetation diameter, D_{SA} and D_{sr} are shading factors and S = averaged spacing = $1/\sqrt{\rho_t}$, where ρ_t =averaged vegetation density.

Wu et al. (2001) introduced the term porosity (θ) to take care of the blockage effect on the water flowing through vegetation. They defined porosity as

$$\theta = 1 - \frac{\pi D_t^2 \rho_t}{4} \tag{7.3}$$

A considerable amount of research has also been carried out in developing resistance laws for channels with flexible vegetation (Kouwen and Unny 1973, Temple et al. 1987, Kouwen and Fathi-Moghadam 2000), and various combinations (Sokolov 1980). Recently, several studies have focused on velocity profiles and turbulent characteristics of vegetated channels (Naot et al. 1996, Nepf 1999, López and García 2001). In addition, an increased interest in the application of various bioengineering techniques has prompted several studies covering the hydraulic aspects related to this activity.

Overall, an abundance of studies, however, is based on laboratory experiments with simple artificial roughness (in uniform flow), whereas in reality natural vegetation exhibits a wide variety of forms and flexibility. In hydraulic analysis, non-submerged and submerged conditions are typically distinguished, since flow phenomena become more complicated when the flow depth exceeds the height of plants (Stone and Shen 2002). In addition, two types of vegetation are usually defined namely: rigid (normally woody plants) and flexible (herbaceous plants). The complexity and advances made in the substantial amount research that has been undertaken have prompted various researchers to pursue the data driven route.

Harris et al (2003) carried out for velocity predictions in vegetated channels, while Babovic et al. (2005), Keijzer et al. (2005) and Baptist et al. (2006) described the process of induction of equation for the vegetation induced roughness. However, in all of these research studies they used Genetic programming was used as the Hydro informatics tool. The work presented in this chapter utilises both Genetic Programming and Artificial Neural Network for velocity prediction in a compound channel.

The laboratory data, detailed later in this chapter, were originally collected for the development of an existing numerical model, DIVAST (Falconer et al. 2001) and bears a number of similarities to the research studies undertaken by Naot et al. (1996) and Järvelä (2002).

7.3 Compound Open Channel

During extreme events, flows often overtop the main channel so as to use the wider carrying and storage capacity of the floodplain. Even in the absence of vegetation there

is a significant increase in the complexity of the flow behaviour once overbank flow has occurred. When over the bank flow occurs, special consideration is required in terms of analysing the interaction between the main channel and floodplain flows, the proportion of flow between sub areas, differences in roughness between the main channel and the floodplain, the significant variation in the resistance parameters with depth and flow regimes, the distribution of boundary shear stresses, the use of the hydraulic radius in calculations, the effects of vegetation on retarding the flow, the sediment transport rates and over bank flow in meandering channels (Knight, 2001).

Naot et al (1993) listed three mechanisms that dominate the flow pattern in a compound open-channel. At each intersection between the floodplain and the main channel a pair of longitudinal vortices is formed with intensity similar to that of the vortices formed at the corners of rectangular channels. In addition, an intensive vortex pair was experimentally shown by Tominaga et al. (1989) and Tominaga and Nezu (1991) at the flood-plain threshold, with one vortex on the floodplain and the second one on the main channel. This pair of vortices controls the interactions between the floodplain and the main channel. The third mechanism was noted by Reece (1976) and Naot and Rodi (1982) (Naot et al., 1993) and they suggested that the turbulent eddies do not have sufficient energy to breach the water surface and therefore break down to smaller vortices, redistributing the velocity fluctuations. These two additional interactions, introduced by Naot and Rodi (1982) (Naot et al., 1993) into an algebraic stress model, showed a substantial effect on the longitudinal vortices.

Different hydraulic conditions prevail in the river and on the floodplain, with the mean velocity in the main channel and on the floodplain being very different. Flow in the main channel exerts a pulling or accelerating force on the flow over the floodplain, which naturally generates a dragging or decelerating force on the flow in the main channel, this leads to the transfer of momentum between the channel section and the floodplain. Momentum transfer between the main channel and the floodplain decreases the discharge in the main channel and increases the discharge on the floodplain, ending up in decreasing the total discharge capacity of the channel (Helmio, 2002). The introduction of vegetation onto the floodplain offers additional resistance, this is known to reduce the velocity and increase the turbulence (Kadlec, 1990).

7.4 Experimental Setup

The study makes use of experimental flume data collected to investigate flow conditions over vegetated floodplains. A physical laboratory model was constructed in the Hyder Hydraulics Laboratory at Cardiff University to investigate the flow conditions over vegetated floodplains in a compound channel. This was carried out as part of a research programme undertaken by Westwater (2001). The model consisted of a recirculating flume, with steady flows over a deep channel and with relatively shallow vegetated floodplains on either side. Figure 7.1 shows the upstream end of the laboratory flume looking in a downstream direction. The prototype was scaled up using Froud's scaling law, in order to maintain similar characteristics to those found in mangrove forests. Although the model was developed based around the data for mangrove forests, it focused on a generalised case of vegetated roughness. The vegetation was simulated using non-submerged, rigid, water surface piercing elements. Cylindrical wooden dowels of 8, 12, 18 and 25mm diameters were used as vegetation. Arranging these dowels in different configurations resulted in densities of:- 122.2, 200 and 366.7 dowels /m^2.

Figure 7.1: Laboratory model of compound channel with vegetated floodplain

Only the density or the cylinder size was varied for each experimental run. Data were collected at a several cross-sections along the flume, which were selected to provide the best spread of data. Figure 7.3 shows the location of the sampling sections. Velocity data were collected using an acoustic doppler velocimeter (ADV), Further details of the experimental data collection may be found in Westwater (2001).

7.5 Data Analysis

A total of 960 data values were available for use that had been collected from 5 different sections, for different combinations of vegetation diameters and densities. Figure 7.4 shows two typical velocity profiles measured across different sections. It can be seen that the data collected from section 2 are significantly different from those collected over other sections. Section 2 was at the very beginning of vegetated floodplain, hence the flow was not developed at this section. It was decided to discard the data collected from this section from further consideration. It can be seen from Figure 7.3 that the channel was symmetrical to the central line, which resulted more or less in a symmetrical velocity profile (Figure 7.4). To exploit this fact, the data from the left hand side of the centre line was used for model development and the right hand side data were used as unseen data for model verification.

The data analysis was carried out using the non linear data analysis tool winGamma described in Chapter 5. The result of the Gamma test is given in Table 7.1. An estimated Gamma statistic ≈ 0.00013 indicates a moderate noise level as does Vratio ≈ 0.0208. The Gamma tests also indicated that smooth models built on this data will have standard deviation of prediction error $\sqrt{0.00013} = 0.011$ on unscaled data.

The M charts in Figure 7.2 shows that there are sufficient data available for building a smooth model from the whole data set (i.e. floodplain and main channel together and the main channel). However, the performance of the model generated from the whole data set should be better then that of the main channel, as evident from the higher Γ value for the latter. Figure 7.2 (c) shows that the Γ value merely reached an asymptotic value with the available data which indicates that, an addition of more data would possibly improve the model.

Table 7.1: Result of the Gamma test on the whole data, including the main channel and floodplain

	Unscaled	Scaled
Gamma	0.00013	0.00007
Gradient	0.020827	0.039186
Standard Error	0.000682	0.000688
V-Ratio	0.293394	0.086266
Near Neighbours	10	4
Start Vector	1	1
Unique Points	336	336

7.6 Model Inputs

As mentioned previously, the dimensionless parameters D_{SA}, D_{SR} and θ were included represent the shading factor and blockage effects respectively. The dimensional parameters included were the diameter (D_t), density of the cylinders (ρ_t), the distance from the beginning of the floodplain of a point along the direction of flow (x), the distance from centreline of the channel across a section (y), the width of the main channel and floodplain (W_{mc} and W_{fp}) respectively, the flow in the channel (Q) and Area of flow (A). The target out put was the measured velocity in the laboratory (V).

In order to facilitate the dimensional correctness of the induced equation in the GP experiments, some dimensionless ratios were included in the experiments. The dimensionless ratios were defined as follows:

$$X_r = \frac{x}{L} \tag{7.4}$$

$$Y_{n_1} = \frac{y}{\left(\frac{1}{2} W_{mc}\right)} \tag{7.5}$$

$$Y_{n_2} = \frac{y}{\left(\frac{1}{2}W_{mc} + W_{fp}\right)} \tag{7.6}$$

$$h_r = \frac{h_{fp}}{h_{mc}} \tag{7.7}$$

h_{fp} and h_{mc} are water depth at floodplain main channel respectively.

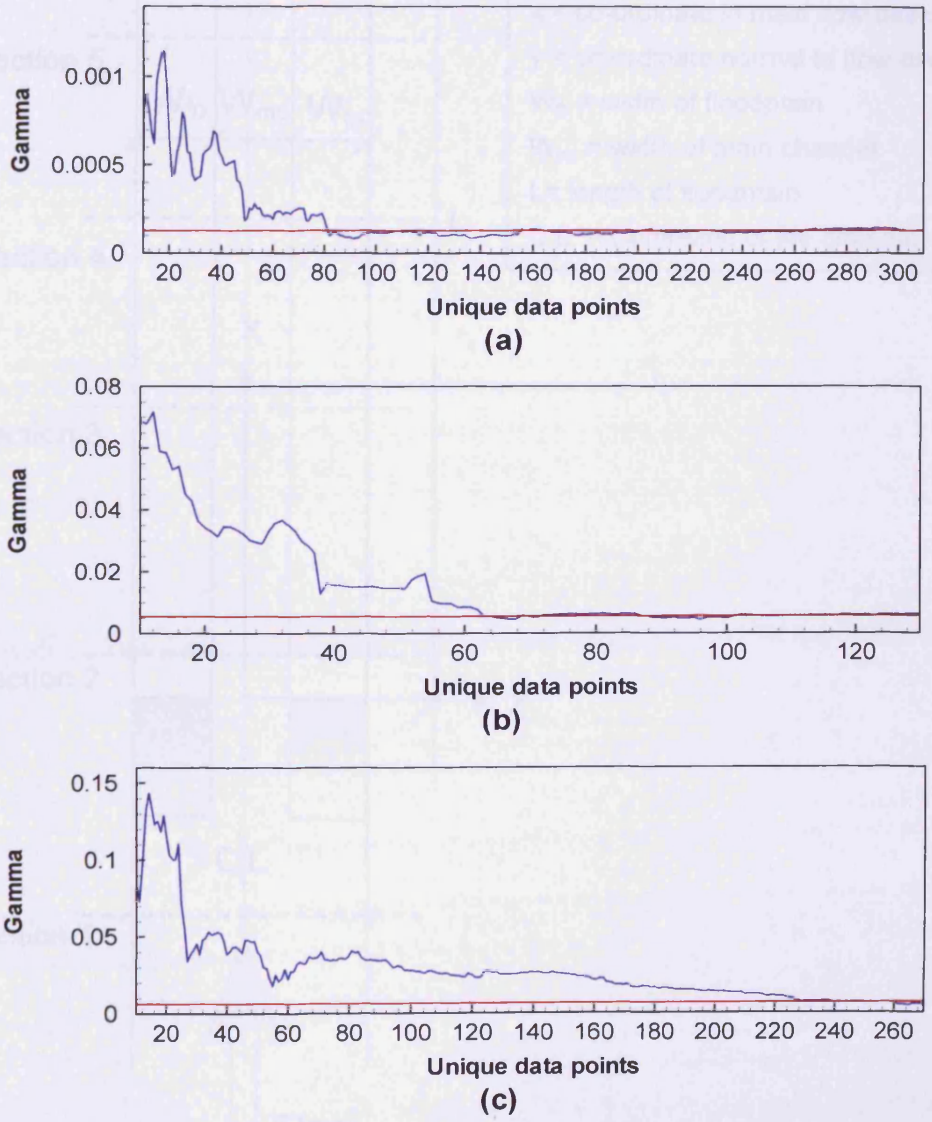


Figure 7.2: M test performed on randomised data, red lines corresponds to the potential Γ value for: (a) main channel and floodplain together (b) main channel (c) floodplain

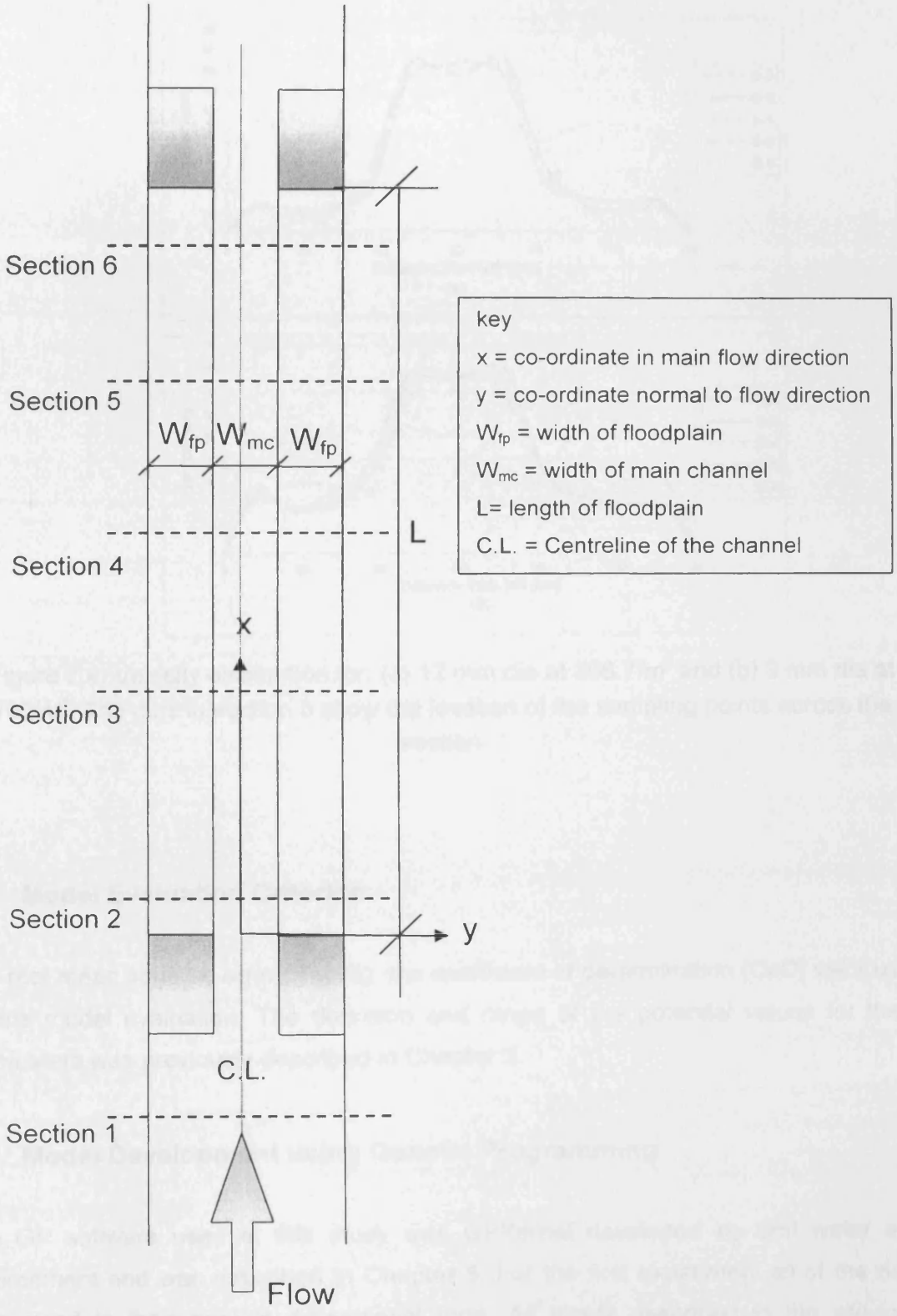


Figure 7.3: Sketch of laboratory flume, showing location of sampling cross-sections

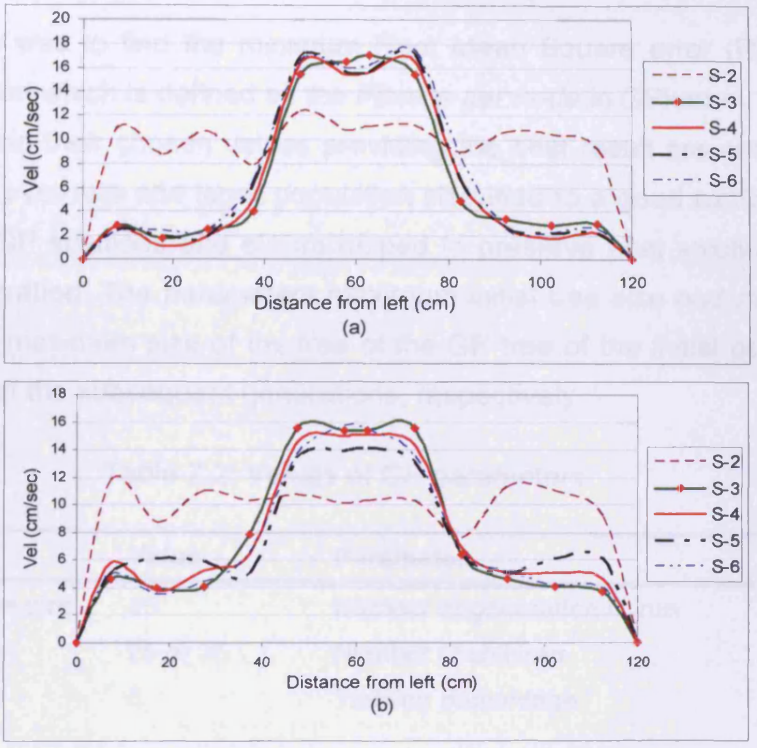


Figure 7.4: Velocity distribution for: (a) 12 mm dia at 366.7/m^2 and (b) 9 mm dia at 100/m^2. The dots in section 3 show the location of the sampling points across the section

7.7 Model Evaluation Criterion

The root mean squared error (RMSE), the coefficient of determination (CoD) were used for the model evaluation. The definition and range of the potential values for these parameters was previously described in Chapter 5.

7.8 Model Development using Genetic Programming

The GP software used in this study was GPKernel developed by DHI water and Environment and was described in Chapter 5. For the first experiment all of the data were used in their original dimensional form. All inputs described in the previous sections were used, other than the dimensionless ratios (i.e. Eq 7.4 to 7.7). Initially the whole dataset was used to train the GP with the target parameter being the velocity

measured at different sections. The function set was $\left(+,-,*,/,\sqrt{},e^x,x^2,\log(x)\right)$. The objective function was to find the minimum Root Mean Square error (RMSE) with a compact expression which is defined as the *Fitness per node* in GPkernel. The relevant GP parameters and their chosen values providing the best result are shown in Table 7.2. A high cross over rate and lager population size lead to a good exploration of the search space of GP solutions and elitism helped to preserve best solution evolved in the previous generation. The parameters maximum initial tree size and maximum tree size stand for the maximum size of the tree of the GP tree of the initial population and of the population of the subsequent generations, respectively.

Table 7.2: Values of GP parameters

Parameter	Value	Parameter	Value
Maximum initial tree size	45	Number of generation to run	1000 or 500
Maximum tree size	25 or 35	Number of children	1000
Tournament size	3	Training percentage	100% or 75%
Population size	1000	Crossover rate	0.9
Number of Experiment	120	Mutation rate	0.08
Breeding method	Tournament	Constant probability	0.3
Elitism used	Yes	Swap mutation rate	0.3

From a series of runs it was observed that the larger initial population size give better results which could be because it leads to a better initial exploration of search space. The values of maximum initial size and maximum size are thus constrained to 45 and 25 respectively. The restriction was necessary as the GP had a tendency to evolve uncontrollably, which deteriorated the compactness of the generated expression. 15 different GP runs were performed with each using a different initial seed.

For the maximum length of the parse tree two techniques were tested. For the first set of experiments the maximum length of parse tree was chosen to 25. This ensured the parsimony of the generated expression, as well as avoid overfitting. In the second approach a cross validation subset was introduced and hence the maximum length of the tree could be increased without the problem of overfitting. However, this increase in tree size could not guarantee parsimony of the expression. Table 7.2 shows the GP parameters related to the first and second approach respectively.

The best produced expression for the first experiment to predict the velocity on both the floodplain and the main channel was as below:

$$V_{GP} = \frac{y}{(y+\theta)\left(y + \frac{y}{e^{x}} + \frac{Q.S}{y^{3}}\right)} \tag{7.8}$$

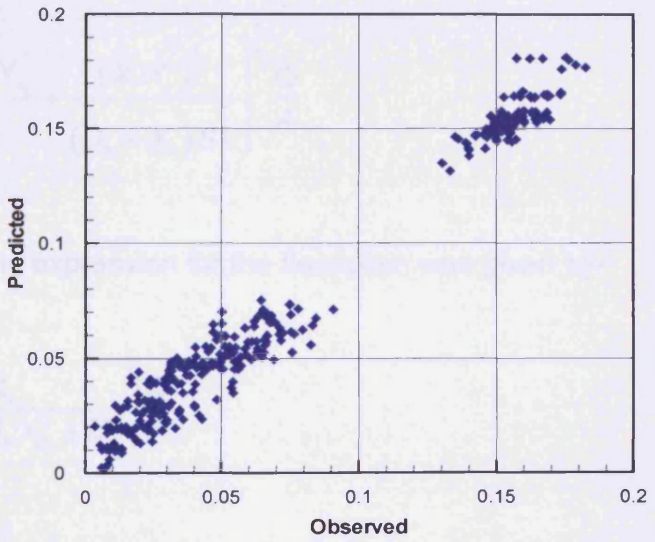


Figure 7.5: Scatter plot for velocity using expression 7.8

This expression showed a very high positive correlation, but it was not dimensionally correct. Moreover, the distribution of the laboratory data showed the presence of two distinct clusters of velocities, one for the higher velocities in the main channel and the other for the lower velocities of the floodplains, respectively. This result was also evident from Figure 7.4.

The next stage of the model development was to separate the dataset into two groups, allowing the GP to generate two expressions for the main channel and the floodplain.

The dimensionless ratios described in Equations 7.4 - 7.7 were included as input parameters in this stage to aid in the generation of dimensionally correct expressions. As the number of inputs increased for this case, the maximum tree size is reduced to 20 to ensure that the number of inputs selected for the generated expression remained more of less same as equation 7.8, which maintained the parsimony of the expressions. To maintain the dimensional correctness an objective function was also included to check the unit error, in addition to the RMSE and fitness per node as used before. The expression that performed best for the main channel was -

$$V_{MC} = \left(\frac{D_s + 3X_r + Y_{n_2}}{D_s + X_r} + \frac{(X.X_r)^{\frac{1}{2}}}{(D_s + X_r)S^{\frac{1}{2}}} \right)^{\frac{1}{4}} \frac{Q}{A} \tag{7.9}$$

The best performing expression for the floodplain was given by-

$$V_{FP} = \left(\frac{D_s X_r}{2Y_{n_1}^2 Y_{n_2}^2 + 2D_{sr} Y_{n_2}^2 + D_s^2} \right) \frac{Q}{A} \tag{7.10}$$

where D_s is the shading factor (either D_{SA} or D_{SR}).

Both the velocity prediction for the floodplain and the main channel showed good correlation (see Figures 7.6 and 7.8), however it can be seen from Figure 7.9 that velocity in the floodplain was under predicted, especially in the region of high velocity. Figure 7.4 shows that the higher velocities in the floodplain were mostly in the region between the floodplain and the main channel. The velocity was influenced by momentum transfer nearer to the main channel. However, no input parameters used in the model development could capture these effects. In order to verify these assumption, the velocities of more then 0.8 m/sec were discarded from the data set, as from the study of the whole dataset it was found that the most of the velocities recorded above 0.8m/s were recorded near the edge of the floodplain which were subjected to a complex flow regime.

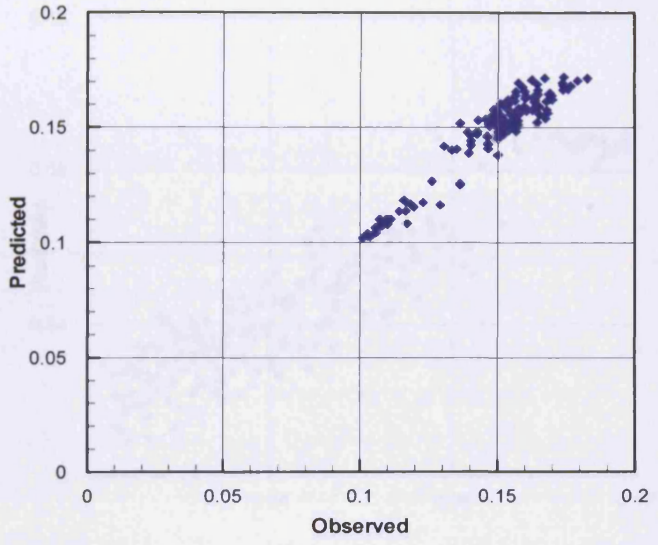


Figure 7.6: Scatter plot for main channel velocity using expression 7.9

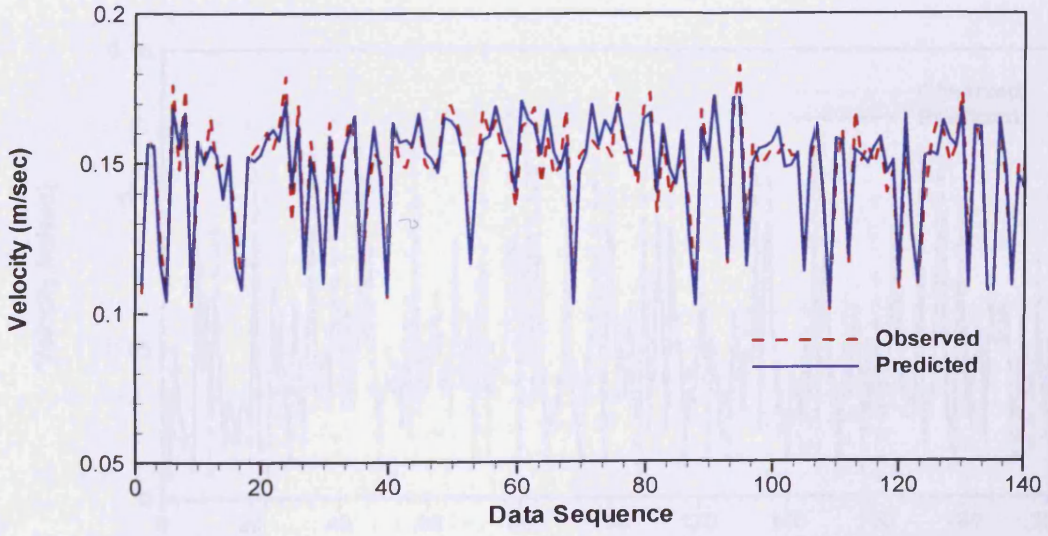


Figure 7.7: Velocity prediction for main channel using expression 7.9

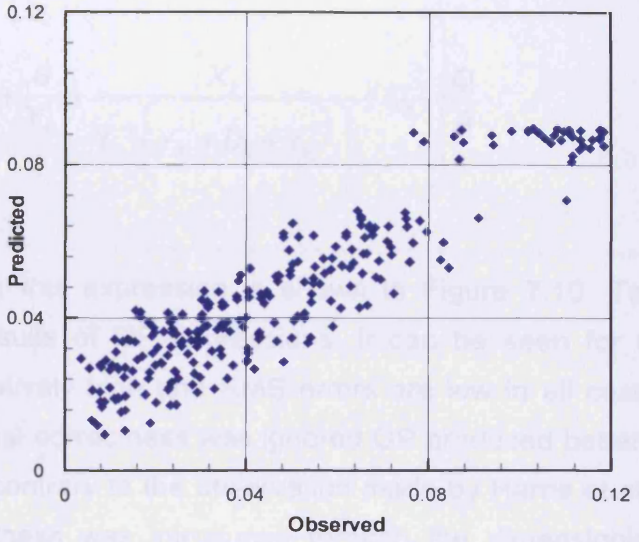


Figure 7.8: Scatter plot for floodplain velocity using expression 7.10

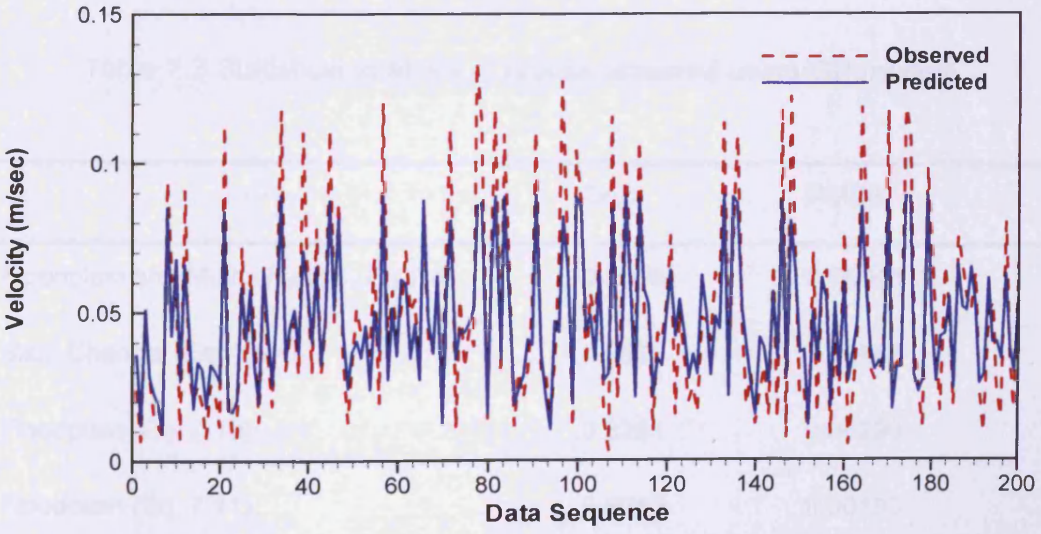


Figure 7.9: Velocity prediction for floodplain using Expression 7.10

Keeping all other parameters the same as before, the best performing expression was found to be:

$$V_{FP'} = \left(\frac{D_s Y_{n_2}^{\ 2} + Y_{n_2}^{\ 4}}{\theta Y_{n_2}^{\ 2} + X_r} + \frac{\theta}{Y_{n_1}} + \frac{X_r}{Y_{n_1}^{\ 3}\left(Y_{n_1} + D_s + Y_{n_2}^{\ \frac{1}{2}}\right)} + D_s^{\ \frac{1}{2}} \right) \frac{Q}{A} \tag{7.11}$$

The scatter plot for this expression is shown in Figure 7.10. Table 7.3 shows the summary of the results of GP expressions. It can be seen for the table, the CoD correlations are relatively high and RMS errors are low in all cases. It is shown that when the dimensional correctness was ignored GP produced better result with 98.16% correlation. This is contrary to the observation made by Harris et al. (2003). When the dimensional correctness was introduced through the dimensionless ratios, the GP result for the main channel was significantly better than the floodplain. As described before, the data of floodplains that were collected from the edge of the flood were left out for the final GP experiment which generated the expression 7.11. It can be seen from Table 7.3 that other statistical measures were also improved in using the expression 7.11. This showed the importance of including some means of incorporating the complex flow field nearer to the main channel.

Table 7.3 Statistical analysis of results obtained using GP models

	CoD	RMSE
Floodplain and Main channel (Eq.7.8)	0.9816	0.00045
Main Channel (Eq. 7.9)	0.9522	0.00087
Floodplain (Eq. 7.10)	0.8251	0.00230
Floodplain (Eq. 7.11)	0.8612	0.00193

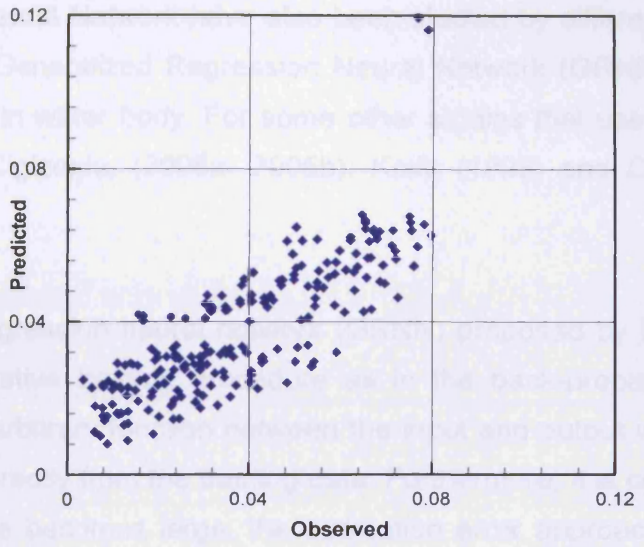


Figure 7.10: Scatter plot for floodplain velocity using expression 7.11

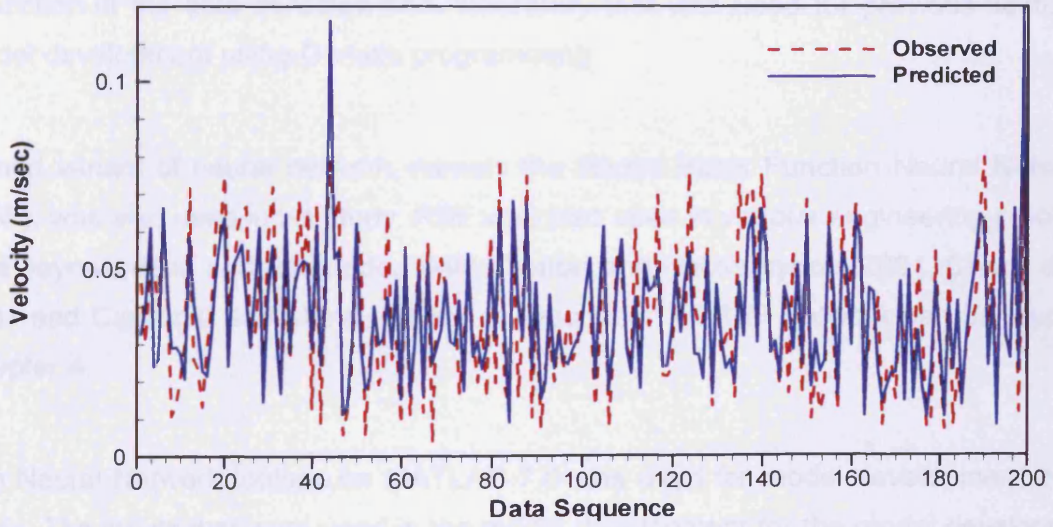


Figure 7.11: Velocity prediction for floodplain using modified Expression 7.11

7.9 Velocity Prediction using ANN

Traditionally Feedforward network is almost exclusively used for prediction and forecasting (Maier and Dandy, 2000). Lin et al. (2003) used this type of network for water quality prediction. Han et al. (2007) reported that Feedforward network with one hidden layer to be the most popular form of neural network. In current research project, the Feedforward network was used exclusively in Chapters 5 and 6. However, the

other variants of Neural Network have also been studied by different authors. Bowden et al (2005) used Generalized Regression Neural Network (GRNN) to forecast Blue-green algae levels in water body. For some other studies that used GRNN, reference can be made to Cigizoglu, (2005a, 2005b), Keifa (1998) and Cigizoglu and Murat (2004).

The generalized regression neural network (GRNN) proposed by Specht (1991) does not require an iterative training procedure as in the back-propagation method, but approximates any arbitrary function between the input and output vectors, drawing the function estimate directly from the training data. Furthermore, it is consistent; that is, as the training set size becomes large, the estimation error approaches zero, with only mild restrictions on the function. Details of the GRNN are presented by Specht (1991),

This study analyzes the performance of Feedforward network and GRNN in velocity prediction of the data collected from laboratory that was used for previous section in model development using Genetic programming.

A third variant of neural network, namely the Radial Basis Function Neural Networks (RBF), was also used in his study. RBF was also used in various engineering problems (see Jayawardena and Fernando, 1998, Gontar and Hatziargyriou, 2001, Chang et al., 2001 and Cigizoglu and Murat, 2004). A description of RBF network can be found in Chapter 4.

The Neural Network toolbox for MATLAB 7.0 was used for model development in this study. The inputs that were used in the model development for the model development for dimensionally floodplain and main channel, floodplain and main channel respectively were used in the all the models described here.

In Feedforward network, a network with one hidden layer was chosen The number of hidden nodes was chosen by trial and error, although in many cases it was found to meet the rule of thumb given in Han (2002) (Han et al. 2007). i.e.

*Number of hidden nodes = (number of inputs + number of outputs) *2/3*

For the training of the networks Gradient descent back propagation or LevenbergMarquardt back propagation was used. Hyperbolic tangent sigmoid transfer function or log sigmoid transfer functions were used in the hidden layer. To ensure good generalisation early stopping was used, along with a limit of the maximum number epoch and goal set in the terms of RMS error. Maximum epoch used was 500 after trial and error. The RMS error was set as the Gamma value found during the data analysis.

As the Feedforward models used the sigmoid function, of which the output values lie in the interval [0,1], all the input values were transformed into the interval [0.05, 0.95], instead of [0,1] because the logistic activation function approaches 0 and 1 asymptotically when the variable approaches negative infinity and positive infinity, respectively. For the case of a tan-sigmoid transfer function being used the range is set to [-0.95, 0.95] to match the range of transfer function which is [-1, 1].

Radial basis networks consist of three layers: an input layer, a hidden radial basis layer an output linear layer. In the model applied in this study a radial basis network was iteratively created one neuron at a time. Neurons were added to the network until the sum-squared error falls beneath an error goal or a maximum number of neurons have been reached. RBFs are determined dynamically and automatically and only one parameter which is the spread has to be assigned for model development.

A GRNN configuration consists of four layers. The input units are in the first layer, the second layer has the pattern units, the outputs of this layer are passed on to the summation units in the third layer, and the final layer covers the output units. Only the biases (spread) are to be given in the model development process. The spread parameters were found simply by trial and error for both the RBF and GRNN models.

The statistical measures of the model produced using all three ANN methods are shown in Table 7.4. It can be seen from the table that the CoD correlation is relatively high and RMS error is reasonably low in all cases. For example, the correlation coefficient ranged from 78.6% (Feedforward network for floodplain) to 93.8% (GRNN for the floodplain and main channel). GRNN produced the worst result of all while

modelling the velocities in floodplain. As experienced with the GP models the best result is produced for the main channel and the floodplain when the dimensional correctness was ignored. The prediction for the floodplain was found to be most difficult with all 3 types of ANNs which was also the case for the GP models. The maximum RMS error was 0.0037 which is less the 4% of the range of velocity data.

Table 7.4: Statistical analysis of result obtained using ANN models

	CoD	RMSE
Floodplain and Main channel		
Feedforward	0.879	0.00099
GRNN	0.938	0.00080
RBF	0.848	0.00121
Main Channel		
Feedforward	0.888	0.00115
GRNN	0.757	0.00273
RBF	0.829	0.00176
Floodplain		
Feedforward	0.786	0.0035
GRNN	0.656	0.0037
RBF	0.806	0.0032

As for the comparison among three types of ANNs none of them consistently outperformed the others for all three cases. Therefore in terms of the performance they all are more or less similar. However, it was found that Feedforward network and RBF resulted in negative velocities for some low velocities. However, this problem was found in GRNN simulations. This observation coincided with that of Cigizoglu and Muart (2004) who indicated that this might be due to the fact that GRNN simulations are bounded by the minimum and maximum of the data value. As no goal was provided in GRNN it did not converge to the poor solutions corresponding to the local minima of

simulations were needed in order to obtain the best Feedforward network as the models performance was very sensitive to the randomly assigned initial weight values. As a result the model development was relatively slower in Feedforward network than the other two types. The scatter plots of the best ANN results are shown in Figures 7.12 to 7.14.

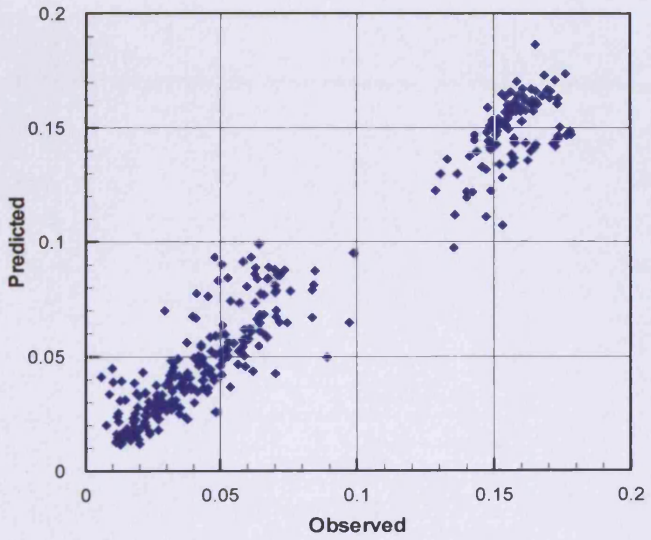


Figure 7.12: Scatter plot for ANN predicted velocity in floodplain and main channel

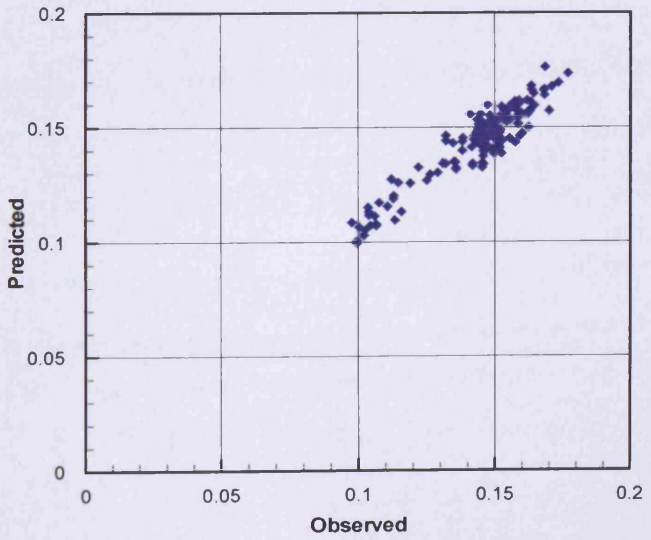


Figure 7.13: Scatter plot for ANN predicted velocity in main channel

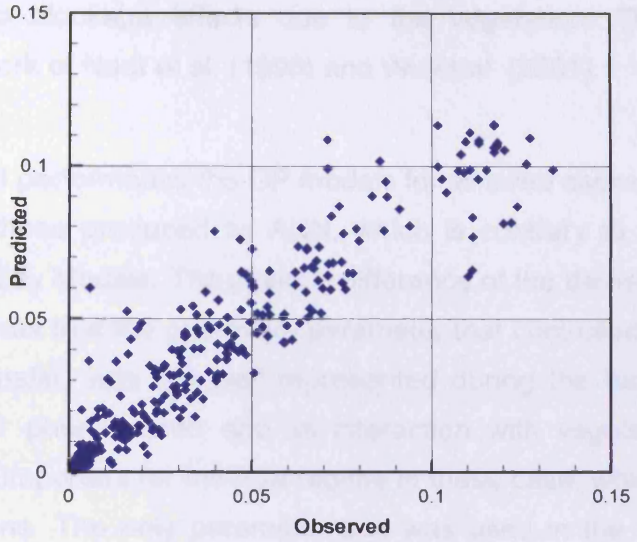


Figure 7.14: Scatter plot for ANN predicted velocity in floodplain

These plots show that the predicted data are in good agreement with the observed data.

7.10 Analysis of Results and Discussion

It is clear from the model results that both GP and ANN perform better for the data collected along the main channel. The results for flow over the floodplain show a higher degree of scattering which is due to the complexity of flow around the vegetation. The presence of vegetation restricted the placing of the ADV probe to some extent, which might have also led to data correction errors as well. The complexity of the flow is due to the complex wake structure in the lee of cylinders, resulting in an increased turbulence. Both GP and ANN tended to under predict the higher velocities along the floodplain which indicates that the accelerating effect of the main channel was missing during the model developments.

The GP expressions (7.10 -7.12) showed that the arrangement of the vegetation (as well as the vegetation density and diameter) was important along with the location of the vegetation in the channel. These expressions showed that hydrodynamic behaviour of the vegetated compound channel largely depended on the parameter describing the

shading factor and blockage effects due to the vegetation. This conclusion was supported by the work of Naot et al. (1996) and Wu et al. (2001).

In comparing the of performance the GP models for all three cases, better results were obtained than for those produced by ANN, which is contrary to the findings for the Ribble and Cardiff Bay Models. The obvious difference of the dataset used in this study from the other two has that the one major parameter that controlled the flow regime i.e. the momentum transfer, was not well represented during the learning process. The momentum transfer phenomenon and its interaction with vegetation presence was reported to be very important for the flow regime in these case, which was discussed in the previous sections. The only parameter that was used in the model development was the distance across the channel which could be weakly related to momentum transfer was the distance across the channel. However, the sharp and irregular changes in the velocity profile across the section indicated that this distance alone could not adequately represent the effect of momentum transfer.

7.11 Summary

It is shown that while experimental science formed a basis for the description of physical phenomena, through the data collection, the knowledge discovery software system, using genetic programming, was able to guide a search for an accurate formulation for the phenomenon under study. Genetic Programming and Artificial Neural Networks have been applied to predict the velocities in a compound channel with vegetated floodplain. Both of the hydro informatics tools produced a reasonably good prediction however it was shown that with improved representation of the underlying system in the model development parameters, the performance of the model could be improved further. It was found that Feedforward, GRNN and RBF neural networks produce more or less similar results. However, GRNN did not provide physically non plausible estimations (negative values). In was also found that both the RBF and GRNN were faster than Feedforward networks.

CHAPTER 8

CONCLUSIONS

8.1 Review and Conclusions

Historically deterministic numerical models have been used for predicting flow and water quality processes in aquatic basins, with these models solving numerically the equations of mass (fluid and constituent) and momentum conservation. Such numerical models have proved to be very useful tools for simulation of various scenarios generally predicting variables with a high level of accuracy. They are very good for long term planning, risk assessment or option assessments. For example, if a land reclamation is planned in a coastal area or a barrage is to be built across a channel, the numerical models can easily simulate the effect of these proposed interventions in the natural environment. However the main problem of this types of models are that they take a long time to run and can not generally be used as online decision support tools. Hence, these models can not be effectively used for predicting imminent threats to public health in recreational waters or an imminent flood risk whereas timely predictions are the essence of the of predict and protect WHO philosophy .

A data driven model can offer an alternative and faster approach for predicting such threats as these cited above. However, data driven models have their own limitations. Firstly, they need a large amount of quality (i.e. noise free) data which might be time and resource consuming. Secondly, as the data driven models lack extrapolation capacity, they can only be used reliably for prediction as long as all the causative parameters are within the range scenarios presented during the model development. Therefore if an extreme event is not present in the training data, then the reliance on data driven models for that extremity would be very risky and resulting some unreliable prediction. Thirdly, data driven models can not predict the effect of a future development, such as a barrage or land reclamation, on water quality or flooding

probability. A numerical model can readily accommodate such aspects as bad level changes due to a land reclamation or possible barrage construction, thereby enough data can be generated with the potential development scenarios through building a data driven model.

The main objective of this thesis was therefore to develop an integrated modelling approach, focused on studying bathing water quality modelling in large estuaries. The ability of data driven techniques to simulate widely varying natural data with noise presence were also examined through the development of models based entirely on field data for study, and on laboratory data in another study. In this research two of the most popular data driven modelling techniques, namely Artificial Neural Networks (ANNs) and Genetic Programming (GP) have been investigated extensively to study their suitability as a prediction tool for water quality and pollution management. These two data mining techniques have been applied for water quality predictions for recreational water quality management and also for predicting the velocity distribution in vegetated compound channels.

The analysis of data for suitability of model building has been carried out using an advanced non-linear data analysis technique, namely the Gamma test. The Gamma test is shown to estimate accurately from the available input/output data the best achievable performance characteristics of a smooth data model. It enabled the model developer to predict the best achievable performance of the model, without the time consuming necessity of estimating this empirically by creating, training and testing a number of networks. The Gamma test also estimates the noise level and establishes whether a smooth model can be built corresponding to the measured noise level. The Gamma test itself provided the criterion for ceasing training of a heuristic model, such as a neural network. This is based on the concept that one criterion of a good model, when tested using unseen data, can be expected to produce a root mean squared error, which is the same (or close to) the true or estimated noise variance associated with the data. The Gamma test used in this study was found to answer two important issues for ANN model construction, namely: (i) specification of the number of data points required for building a smooth model, which in turn helps divide the data set into training, test and validation subsets, and (ii) specification of the stopping criterion in

training a neural network. The latter is shown to be effective in preventing a network from over training.

In the integrated model development exercise, a deterministic numerical model, namely DIVAST, has been used to simulate a whole host of scenarios that included some extreme conditions that can be possibly expected in the study area. The use of a deterministic model as a data generator ensured that the training process of data driven models included all extremes, as a result the question of extrapolation by data driven models can be indirectly addressed. The deterministic models were also used to indicate the response time of the receiving water quality to the upstream boundary conditions (e.g. in Ribble estuary), which were found to be very useful for selection of input parameters for the data driven models. Once a the potential response time was known for the water body then Gamma tests were undertaken to determine the effective input combinations more precisely. Thus a combination of deterministic models and Gamma tests were found to ascertain the input selection for the data driven models. In the present study it was shown that with this information being used for model construction, a network can achieve a similar level of performance to those networks with twice the number of input parameters.

The deterministic model was also used for verifying the input-output relations delivered by the data driven models (e.g. Cardiff Bay Study). As numerical models are flexible they can be run for various conditions to verify the affects and weight of input parameters, identified through the data driven models, and which will enhance confidence in the data driven models.

As for the modelling studies with Genetic Programming (GP) it was found that if dimensional correctness was considered as an objective for model development then the model performance deteriorated. This was due to the fact that adding dimensionality in the evolution of generations reduces the search space. It was therefore desirable not to include dimensionality for purely predicting purposes. However, the dimensional correctness of the generated expression was achieved through the introduction of some dimensionless ratios as shown in Chapter 7. Another issue observed in the generated expressions was that they were generally very complicated and at times contained higher degrees of polynomials. A preferential bias

had been added in the GP program used in this study, so that the search space was reduced but information about parsimony is not lost. It was not very efficient in returning a parsimonious expression which was representative of all important input parameters.

The modelling with Artificial Neural Networks showed that, the use of deterministic model and Gamma analysis prior to model development helped find the models inputs, stopping criterion and data division, all of which were usually determined heuristically. Different variants of the ANNs were not shown to deliver significant and/or consistent, performance benefit in the problems studied in this research.

ANNs were found to perform better than GPs in predicting water quality parameters, whereas in predicting the velocities for the compound channel study then the GP model outperformed the ANN models. However, for a given dataset, whichever method performed best was found to do so consistently for any combination of input and output datasets for that study. By their very nature GPs will supply a symbolic-algebraic relationship between the measured data through the process of evolution and competition between all possible solution expressions. ANNs on the other hand will usually find a relationship between the input and output data, but then the resulting relationships can only be represented sub-symbolically and are therefore essentially 'hidden' from the user. It is worth mentioning that there is a stark difference in the data range used in some of the problems studied. In first case the Faecal Coliform levels in the Ribble were of the order of 10^6, whereas for the velocity predictions in the estuary the range of data were within generally less than 1 m/s. However, in the literature review, there are no indications found to suggest that ANNs perform better than GPs for certain ranges of data or vice versa. In this context, the effect of variability and the range of data would constitute the subject matter for a separate research study.

The notion of bacterial level predictions based on data can be argued. However, if the uncertainly and inaccuracy involved in bacterial count of a field sample and the acceptable degree of accuracy of a numerical model are considered as for the case of Ribble then the accuracy levels demonstrated in this study by data driven models can be justifiably accepted without argument.

Data driven models were also developed and tested for on laboratory data which were collected in a flume, for steady flows occurring over a relatively deep channel and with relatively shallow vegetated floodplains. The velocity data were used to test the scope for using both ANNs and GPs to predict the velocity data. However, one important part of this study was to induce the formulation of expressions for the resistance using GPs, which could then be used to improve on the complex frictional representations in 2-D deterministic models. The flow structure and velocity distribution in the vegetated flood plain was very different from that for the in-bank flow condition, due to lateral momentum exchange (van Prooijen et al, 2005). In order to determine the stage discharge relationship for a compound channel the transverse profile of the streamwise velocity in the mixing region needed to be known. A 3-D numerical model can simulate the complicated flow pattern in such flow environments. However, such models are time and resource consuming; therefore any formulation that can be used in a 2-D model would be more practical for current hydro-environmental impact assessment studies. The performance of a GP model is reported for two variations of the GP. The reported results of the experiments were found to be encouraging.

Although ANNs have many attractive features, they suffer from some limitations. The issue of choosing the optimal network architecture is very much subjective. The user has to pre-determine the structure of the ANN network and the training algorithms. On the other hand GPs can optimise both the structure of the model and the parameters.

In Chapter 5 the ability of GPs and ANNs to perform forward predictions of faecal coliform levels in estuarine waters has been demonstrated. This would be very useful for real-time water quality predictions on a day to day water management basis. In this study the numerical model has been used to generate data simulating a wide range of possible scenarios for the Ribble estuary. This integrated approach is particularly helpful as the extrapolation capability of data driven methods is not very reliable. As a result if all possible extreme events are included in the data set, then the model can be more reliable and useful. It can be seen that the expressions generated by GP approaches contain some dimensionally incorrect forms, such as salinity being added to water depth, however these formally incorrect expressions can give some meaningful information. More importantly, imposing dimensionality into the equation

reduces the search options for GPs significantly, which in turn affects the model performance.

For the case of the Ribble Estuary study numerical models were also used to determine the response time, at different locations in the receiving water, to the water quality conditions at the boundary. A knowledge of response time of the receiving water quality relative to the upstream boundary condition is very useful in constructing a data driven model. In Chapter 5 it has been shown that with this information being used for model construction, networks can achieve a similar level of performance to those of networks with twice as much input information. This further strengthens the view that the application of data driven models need proper preparation of the exercises, i.e. an analysis of logical relationships between dependent and independent variables and the choice of variables. Numerical models can play an important role in terms of the selection of input parameters and establishing the relationship between model inputs and outputs. The knowledge response time is also important in specifying the forward prediction window, which is crucial in the provision of data and information for bathing water managers. In particular, for day to day water management, the most important information required is whether the Faecal Coliform levels at certain day are above or below some threshold value. In that respect the data driven models have been shown to be an encouraging development of current management regime.

The main findings of the study can therefore be summarised as given below:

- Data driven models can offer a fast and reliable alternative to the traditional numerical modelling approaches in day to day recreational water management. In the current age of 'predict and protect' regulatory regime concepts the idea of using such models are even more encouraging. These models require very little human intervention, yet can flag up any potential water quality deterioration along designated bathing water beaches. Once such a system is in operation it is possible to retrain the models periodically with the latest data.

- None of the data driven techniques studied herein, namely GP and ANNs, performed constantly better then the other. There may be a role for muti-model systems, using both GP and ANNs, perhaps with related but different sets of

input combinations. The user may have a grater degree of confidence in model predictions if more than one model flags up the same events.

- An integrated approach to modelling that includes both the conventional numerical models and the data driven models should be seriously considered, at least during the model building phase of the data driven models. In order to provide sufficient data points for training and testing of the data driven models a calibrated hydrodynamic and water quality model can be used to generate input data.

- The knowledge on the response time of the receiving water quality to the upstream boundary conditions is very useful in the development of data driven models. In the present study it was shown that with this information being used for model construction, a network can achieve the similar level of performance to those networks with twice the number of input parameters. The numerical models were shown to be very useful for identifying the response time. The knowledge on the response time is also useful for making forward predictions. This is crucial for bathing water managers to give warning to potential visitors.

- Data driven models were also shown to predict the velocities in a compound channel with vegetated flood plains. In a similar manner the water levels can be predicted in flood plains in order to deliver early flood warnings. The faster output of data driven models compared to the numerical models can offer some valuable time gain for flood preparedness. The GP models induce the formulation of expressions for the vegetation resistance which provided valuable insight into the complex flow nature and with an improved model induction could then be used to improve on the complex frictional representations in 2-D deterministic models.

- The data analysis technique used in this study provided answers to address two important issues for ANN model construction, i.e. the specification of a minimum number of data points required for building a smooth model, and the specification of stopping criterion in training a neural network. The stopping criterion thus used was found to be effective in improving the generalisation

capability of ANNs. The data analysis technique was also useful for the data division for both GP and ANNs.

8.2 Recommendations for Further Study

Additional data collection

As seen in Chapters 6 and 7 the data available for model development had limitations. The data available for Cardiff Bay were collected only once a week, hence the data set could not be presented in a time series manner. As a result of this limitation the model could only predict bacterial concentrations at unsampled sites of interest by using known water quality data at other locations and at the same time. An intensive data collection programme, with a higher frequency, would allow the development of a more accurate real-time prediction tool for Cardiff bay.

For the velocity predictions for the flume data, the vertical distribution of the stream wise and transverse velocity directions for various depths could be measured, which would offer more information about the vortex motion and the corresponding form drag and interaction with the bed friction, both on the floodplain and main channel, and also at the interface. The water level should be varied in any future tests, as it has been shown in the literature that changes in the depth ratio (i.e the ratio of the flood plain water depth to the main channel depth) have a differential influence on the velocities along the flood plain and in the main channel. Alternatively, a 3-D numerical model, which would include a higher order turbulence model, such as the $k-\varepsilon$ model, could be used to predict more accurately the complicated shear interface between the channel and flood plain. Such a model would be ideal for generating data. These data could then be used for model induction using a GP model.

Model induction with Genetic Programming

As found in this study GP models can formulate expressions relating the input parameters to the output results which give a reasonable insight into the processes being considered. However, the generated expressions are generally very complicated and at times contain higher degrees of polynomials. A preferential bias has been added

in the GP program to ensure the parsimony of the expression; however in such casea the generated expressions are not representative of all input parameters. The addition of some criterion which forces the GP model to include all of the important parameters, and at the same time maintains the parsimony, would be helpful. In order to exploit the full potential of model induction capability, the dimensional correctness has to be included without constricting the search space.

Modelling the Bacterial decay

In modelling the Faecal Coliform levels in a water body, a constant decay rate is usually used in estimating the natural mortality of pathogens, with a constant being added in the source term of the advective diffusion equation. In the current study a formulation is used to accommodate the salinity and temperature effects. To simulate the effects of solar radiation a constant decay rate is used, but the value is different for day and night. It would be very useful to develop a GP expression containing all of the major decay rate variable parameters. In order to simulate the exact effect of the variables experiments could be carried out, in a controlled laboratory environment, where solar radiation can be varied with a sunlight simulator and turbidity can be varied by mixing different amounts of fine particulate matter. Other parameters can also be varied. Such a formulation would be extremely useful for improving the bacterial predictions using numerical models.

Predicting Gastroenteritis Rates and Waterborne Outbreaks

On a slightly different note data driven models can be used to predict the disease burden in various cases. As for example, waterborne Gastroenteritis might spread even in cities with modern water treatment facilities. Traditional public health surveillance methods rely on detection and reporting of specific pathogens in clinical specimens and have significant limitations in detection the outbreaks of disease rapidly and effectively. Data driven models incorporating weather data, demographic pattern, public and school holidays, disease incidence etc. would be able to flag potential health risks before it takes place.

REFERENCES

Abbott, M. B., and Basco, D. R. (1989). Computational Fluid Dynamics: An Introduction for Engineers, Longman Group UK Ltd.

Aguilera, R. (2000). "Genetic Programming with GPkernel." GPK sofware User Manual.

AIMLearning. (2000). "Discipulus." Fast Genetic Programming based on AIM Learning Technology, Owner's Manual.

Albrechtsen, H. J. (1994). "Distribution of bacteria, estimated by a viable count method, and heterotrophic activity in different size fractions of aquifer sediment." Geomicrobiol Journal, 12, 253-264.

Alkan, U., Elliott, D. J., and Evison, L. M. (1995). "Survival of enteric bacteria in relation to simulated solar radiation and other environmental factors in marine waters." Water Research, 29(9), 2071-2081.

Altenberg, L. (1994). "Emergent phenomena in genetic programming." Evolutionary Programming, V. Sebald and L. J. Fogel, eds., World Scientific Publishing, 233-241.

Anderson, I. C., Rhodes, M., and Kator, H. (1979). "Sub lethal stress in Escherichia coli: a function of salinity."

APHA. (1998). "Standard Methods for the examination of water and wastewater." American Public Health Association, American Water Works Association and Water Environmental Federation.

Ashbolt, N. J., Grabow, W. O. K., and Snozzi, M. (2001). "Indicators of microbial water quality." Water Quality: Guidelines, standards and Health, L. Fewtrell and J. Bartram, eds., IWA Publishing, London, UK.

Auer, M. T., and Niehaus, S. L. (1993). "Modelling faecal coliform bacteria -I. field and laboratory determination of loss kinetics." Water Research, 27(4), 693-701.

Ayres, P. A. (1977). "Coliphages in sewage and the marine environment." Soc. Appl. Bacteriol. Symp. Ser., 6, 275-298.

Babovic, V., and Abbot, M. B. (1997). "The evolution of equations from hydraulic data, Part I:Theory " Journal of Hydraulic Research, 35(3), 1-14.

Babovic, V., Baptist, M., and Mynett., A. (2005). "Man against Machine: Experiments in Vegetation induced resistance." Proceedings of the IAHR Congress, Seoul, Korea.

Banzhaf, W., Nordin, P., Keller, R., and Francone, F. (1998). Genetic Programming- an introduction: on the automatic evolution of computer programs and its applications, Morgan Kaufmann Publishers, Inc.

Baptist, M. J., Babovic, V., Rodríguez, J., Keijzer, M., Uittenbogaard, R. E., Mynett, A., and Verwey, A. (2006). "On inducing equations for vegetation resistance." Journal of Hydraulic Research, 0(0), 1-16.

Barcina, I., Gonzalea, J. M., Iriberri, J., and Egea, L. (1989). "Effect of visible light on progressive dormancy of Escherichia coli cells during the survival process in natural fresh water." Applied and Environmental Microbiology, 55(1), 246-251.

Barcina, I., Lebaron, P., and Vives-Rego, J. (1997). "Survival of allochthonous bacteria in aquatic system: a biological approach." FEMS Microbiol. Ecol., 23, 1-9.

Barnes, H. H. (1967). "Roughness characteristics of natural channels." Geological Survey Water-Supply Paper 1849 US geological Survey, 213 pp.

Barron, A. R. (1994). "A comment on "Neural networks: A review from a statistical perspective'." Statistical Science, 9(1), 33-35.

Bebis, G., and Georgiopoulos, M. (1994). "Feed-forward neural networks: Why network size is so important." IEEE Potentials, October/November, 27-31.

Bedford, K. W. (1994). "Diffusion, dispersion and sub-grid parameterization." Coastal Estuarial and Harbour Engineers Reference Book, M. B. Abbott and W. A. Prince, eds., E&FN Sons Ltd, Chapter 5. pp 61-82.

Bellair, J. T., Parr-Smith, G. A., and Wallis, I. G. (1977). "Significance of diurnal variations in faecal coliform die-off rates in the design of ocean outfalls." J. Water Pollut. Control Fed, 49, 2022-2030.

Bordalo, A. A., Onrassami, R., and Dechsakulwantana, C. (2002). "Survival of faecal indicator bacteria in tropical estuarine waters (Bangpakong river, Thailand)." Journal of Applied Microbiology, 93, 864-871.

Borrego, J. J., Arrabal, F., Vicente, A. d., Gomez, L. F., and Romero, P. (1983). "Study of microbial inactivation in the marine environment." J. Water Pollut. Control Fed., 55, 297-302.

Borrego, J. J., Córnax, R., Moriñigo, M. A., Martínez-Manzanares, E., and Romero, P. (1990). "Coliphages as an indicator of faecal pollution in water. their survival and productive infectivity in natural aquatic environments." Water Research, 24(1), 1-129.

Borrego, J. J., Morinigo, M. A., Devicente, A., Cornax, R., and Romero, P. (1987). "Coliphages as an indicator of faecal pollution in water. Its relationship with indicator and pathogenic microorganisms." Water Research, 21, 1473-1480.

Bosch, A. (1998). "Human enteric viruses in the water environment: a minireview." International Microbiology, 1, 191-196.

Bourdreau, A., and Coulliard, G. (1999). "Systems Integration and Knowledge Management." Information Systems Management, Fall, 24-33.

Bowden, G., Dandy, G., and Maier, H. (2005). "Forecasting Cyanobacteria (Blue-Green Algae) using Artificial Neural Networks." ArtificialNeural Networks in Water Supply Engineering, S. Lingireddy and G. Brion, eds., ASCE, Reston, Virginia.

Burden, F. R., Brereton, R. G., and Walsh, P. T. (1997). "Cross-validatory selection of test and validation sets in multivariate calibration and neural networks as applied to spectroscopy." Analyst 122(10), 1015-1022.

Cabelli, V. J. (1983). "Health effects criteria for marine recreational waters." Research Triangle Park, NCUnited States Environmental Protection Agency: 50.

Calderon, R. (1990). Personal communication, United States Environmental Protection Agency

Canale, R. P., Auer, M. T., Owens, E. M., Heidtke, T. M., and Effler, S. W. (1993). "Modelling fecal coliform bacteria—II. Model development and application." Water Research, 27(4), 703-714.

Carabin, H., Gyorkos, T. W., Joseph, L., Payment, P., and Soto, J. C. (2001). "Comparison of methods to analyse imprecise faecal coliform count data from environmental samples." Epidemiol. Infect., 126( ), 181-190.

CardiffBayBarrageAct. (1993). " Elizabeth II. Chapter 42 ISBN 0105442933."

Carlucci, A. F., and Pramer, D. (1960). "An Evaluation of Factors Affecting the Survival of Escherichia coli in Sea Water - II. Salinity, pH, and Nutrients." Journal of Applied Microbiology, 8(4), 247-250.

Carpenter, G. A., Grossberg, S., and (1988). "The ART of Adaptive Pattern Recognition by a Self-Organizing Neural Network." Computer, 21(3), 77-88.

Castellano, G., Fanelli, A. M., and Pelillo, M. (1997). "An iterative pruning algorithm for feedforward neural networks." IEEE Transactions on Neural Networks, 8(3), 519-531.

Chamberlin, C. E., and Mitchell, R. (1978). "A decay model for enteric bacteria in natural waters." Water Pollution Microbiology, R. Mitchell, ed., John Willey & Sons, Inc, New York, 325-348.

Chang, F.-J., Liang, J.-M., and Chen, Y.-C. (2001). "Flood Forecasting Using Radial Basis Function Neural Networks." IEEE Transactions On Systems, Man, And Cybernetics—Part C: Applications And Reviews, 31(4), 530-535.

Chen, S. T., Yu, D. C., and Moghaddamjo, A. R. (1992). "Weather sensitive short-term load forecasting using non-fully connected artificial neural network." IEEE Transactions on Power Systems, 7(3), 1098-1104.

Chen, Y. (1992). "Numerical Modelling of Solute Transport Processes Using Higher Order Accurate Finite Difference Schemes," Ph.D. Thesis, University of Bradford, England.

Choi, B., Hendtlass, T., and Bluff, K. (2004). "A Comparison of Neural Network Input Vector Selection Techniques." Innovations in Applied Artificial Intelligence Springer Berlin / Heidelberg

Chow, V. T. (1959). Open Channel Hydraulics, McGraw Hill Book Co., New York.

Cigizoglu, H. K. (2005a). "Application of Generalized Regression Neural Networks to Intermittent Flow Forecasting and Estimation " Journal of Hydrologic Engineering, 10(4), 336-341.

Cigizoglu, H. K. (2005b). "Generalized regression neural network in monthly flow forecasting." Civil Engineering and Environmental Systems, 22(2), 71-84.

Cigizoglu, H. K., and Murat, A. (2004). "Rainfall-runoff modelling using three neural network methods." ICAISC 2004, LANI 3070, pp 166-171.

Communities, C. o. t. E. (1976). "Council Directive of 8th December 1975 Concerning Bathing Water Quality (76/160/EEC)." Official Journal of the European Community, No L31/1-7.

Connor, J. T., Martin, R. D., and Atlas, L. E. (1994). "Recurrent neural networks and robust time series prediction." IEEE Transactions on Neural Networks, 5(2), 240-254.

Corbett, S. J., Rubin, G. L., Curry, G. K., and Kleinbaum, D. G. (1993). "The health effects of swimming at Sydney beaches. The Sydney Beach Users Study Advisory Group." American Journal of Public Health, 83(12), 1701-1706.

Coyne, M. S., and Howell, J. M. (1994). "The fecal coliform/fecal sterptococci ratio(FC/FS) and water quality in the bluegrass region of Kentucky." Soil and Science News & Views, 15(9), 1–3.

Cramer, N. L. (1985). "A Representation for the Adaptive Generation of Simple Sequential Programs." Proceedings of the First International Conference on Genetic Algorithms and Their Applications,, In J. J. Grefenstette, ed., pp 183-187, .

Crane, S. R., and Moore., J. A. (1985). Modelling enteric bacterial die-off: a review, Oregon, Corvallis.

Cybenko, G. (1989). "Approximation by superpositions of a sigmoidal function." Mathematical Control Signals Systems 2, 303-314.

Dai, H., and MacBeth, C. (1997). "The application of back-propagation neural network to automatic picking seismic arrivals from single-component recordings." Journal of Geophysical Research, 102(B7), 15115-15113

Darakas, E. (2002). "E. Coli Kinetics - Effect of Temperature on the Maintenance and Respectively the Decay Phase." Environmental Monitoring and Assessment, 78(2), 101-110.

Darakas, E., and Hadjianghelou, A. (1997). "Mathematische Erfassung der Absterbekinetik der Escherichia coli in Gewässern." Wasser-Abwasser, gwf, 138, 86-89.

Darwin, C. (1859). On the origin of the species by means of natural selection, Murray, London, UK.

Davenport, T. H., and Prusak, L. (1998). Working Knowledge: How Organizations Manage What They Know., Harvard Business School Press, Boston, MA.

Davies, C. M., and Evison, L. M. (1991). "Sunlight and the survival of enteric bcteria in natural waters." Journal of Applied Bacteriology, 70, 265-274.

Davies-Colley, R. J., Bell, R. G., and Donnison, A. M. (1994). "Sunlight Inactivation of Enterococci and Fecal Coliforms in Sewage Effluent Diluted in Seawater." Applied and Environmental Microbiology, 60(6), 2049-2058.

Davis, E. M., Casserly, D. M., and Moore, J. D. (1977). "Bacterial relationships in stormwaters." Water. Resource Bulletin, 19, 895-905.

Dawson, C. W., and Wilby, R. L. (2001). "Hydrological modelling using artificial neural networks." Progress in Physical Geography, 25(1), 80-108.

Demuth, H., and Bale, M. (2003). "Neural network toolbox user's guide." The Mathworks. Inc.

D'haeseleer, P. (1994). "Context Preserving Crossover in Genetic Programming." Proceedings of 1st IEEE Conf. on Evolutionary Computation, IEEE Press, 256-261.

Dibike, Y. (2002). "Model induction from data," IHE Delft, PhD Thesis.

Dibike, Y. B., Minns, A. W., and Abbott, M. B. (1999). "Application of Artificial Neural Networks to the Generation of Wave Equations from Hydraulic Data." Journal of Hydraulic Research, 37(1), 81-97.

Dionisio, L. P. C., Garcia-Rosado, E., Lopez-Cortes, L., Castro, D., and Borrego, J. J. (2002). "Microbiological and sanitary quality of recreational seawaters of southern portugal." Water, Air, & Soil Pollution, 138, 319-334.

Douglas, J., and Gunn, J. E. (1964). "A generalized formulation of alternating direction methods. Part I. parabolic and hyperbolic problems." Numer. Math., 6, 428-453.

Drécourt, J.-P., and Madsen, H. (2001). "Role of domain knowledge in data-driven modelling." 4th DHI Software Conference, Scanticon Conference Centre, Helsingør, Denmark.

Drucker, P. E. (1995). "The Post Capitalistic Executive." Management in a Time of Great Change, P. E. Drucker, ed., Penguin New York.

Dufor, A. (1977). "Escherichia coli: the faecal coliform." Bacterial Indicators/Health Hazard Associated with water, A. W. Hoadley and B. J. Dutka, eds., American Society for Testing and Material, Philadelphia, Pa, 48-58.

Dutka, B. J. (1973). "Coliforms are an inadequate index of water quality." Journal of Environmental Health, 36, 39-46.

Dutka, B. J. (1984). "Sensitivity of legionella pneumophila to sunlight in fresh and marine water." Applied and Environmental Microbiology, 48, 970-974.

DWAF. (1995). "South African Water Quality Guidelines for Coastal Marine Waters." Department of Water Affairs and Forestry DWAF.

DWAF. (1996). "South African Water Quality Guidelines, Domestic Water Use (2nd edn.)." Pretoria.

E.U. (1984). "Proposal for a Council Directive Concerning the Quality of Bathing Waters (94/C 112/03)." Brussels.

Edwards, E., Coyne, M., Daniel, T., Vendrell, P., Murdoch, J., and Moore, P. (1997). "Indicator bacteria concentrations of two northwest Arkansas streams in relation to flow and season." American Society of Agricultural Engineers, 40(1), 103-109.

Edwards, R. J. G. (1997). "A Review of the Hydrogeological Studies for the Cardiff Bay Barrage." Quarterly Journal of Engineering Geology, 30, 49-61.

Elder, J. W. (1959). "The Dispersion of Marked Fluid in Turbulent Shear Flow." J. Fluid Mech, 5, 644-650.

Elliot, E. L., and Colwell, R. R. (1985). "Indicator organisms for estuarine and marine waters." FEMS Microbiology Letters, 32(2), 61-79.

Elman, J. L. (1990). " Finding structure in time." Cognitive Science, 14, 179-211.

Esrey, S. A., Feachem, R. G., and etal. (1985). "Interventions for the control of diarrhoeal diseases among young children: Improving water supplies and excreta disposal facilities." Bulletin of the World Health Organizatio, 63(4), 757-772.

Evans, D., and Jones, A. J. (2002). "A proof of gamma test." Proceedings of Royal Society Series A.

Falconer, R. A. (1976). "Mathematical modelling of jet-forced circulation in reservoirs and harbours," PhD Thesis, Imperial College, University of London, London.

Falconer, R. A. (1991). "Review of modelling flow and pollutant transport processes in hydraulic Basins." Proceedings of the first international conference of water pollution, modelling, measuring and prediction, computational mechanics publications, Southampton, UK, September, pp 3-23.

Falconer, R. A. (1994). "An introduction to nearly horizontal flow." Coastal Estuarial and Harbour Engineers Reference Book, M. B. Abbott and W. A. Price, eds., E&FN sons Ltd.

Falconer, R. A., and Chen, Y. (1991). "An improved representation of flooding and drying and wind stress effects in a two-dimensional tidal numerical model." Proc. Inst. Civil Engineers, Part 2, Vol. 91, pp. 3-23.

Falconer, R. A., and Lin, B. (1999). "DIVAST reference manual." School of Engineering, Cardiff University, University of Wales, Cardiff, UK.

Faust, A., Aotaky, A. E., and Hargadon., M. L. (1975). "Effect of physical parameters on the in situ survival of Escherichia coli in an estuarine environment." Appl. Microbiol. , 30, 800-806.

Feachem, R. G. (1975). "An improved role for faecal coliform to faecal streptococci ratios in the differentiation between human and non-human pollution sources (Note)." Water Research, 9, 689-690.

Fischer, H. B. (1979). Mixing and Dispersion in Inland and Coastal Water, Academic Press, Inc., California.

Fleisher, J. M., Kay, D., Salmon, R. L., Jones, F., Wyer, M. D., and Godfree, A. F. (1996). "Marine waters contaminated with domestic sewage: non-enteric illnesses associated with bather exposure in the United Kingdom." American Journal of Public Health, 86(9), 1228-1234.

Fleisher, J. M., Kay, D., Wyer, M. D., and Godfree, A. F. (1998). "Estimates of the severity of illnesses associated with bathing in marine waters contaminated with domestic sewage." International Journal of Epidemiology, 27(4), 722-726.

Fleisher, J. M., Kay, D., Wyera, M. D., and Godfreec, A. F. (1998). "Estimates of the severity of illnesses associated with bathing in marine recreational waters contaminated with domestic sewage." International Journal of Epidemiology 27, 722-726.

Fletcher, C. A. J. (1991). Computational techniques for fluid dynamics Vol II: Specific techniques for different flow categories, 2nd edition, Springer and Verlag, Berlin.

Flint, K. P. (1987). "The long-term survival of Escherichia coli in river water." Journal of Applied Bacteriology, 63(3), 261-270.

Flood, I., and Kartam, N. (1994). "Neural networks in civil engineering. I: Principles and understanding." Journal of Computing in Civil Engineering, 8(2), 131-148.

Forst, W. H., and Streeter, H. W. (1924). "Section 6 Bacteriological Studies." Study of Pollution and Natural Purification of the Ohio River, Public Health Service, Washington, D.C.

Fugate, K. G., Cleaver, D., and Hatch, M. (1975). "Enterovirus and potential bacterial indicators in gulf coast oysters." Journal of Milk and Food Technology, 38(2), 100-104.

Fujioka, R. S., Hashimoto, H. H., Siwak, E. B., and Reginald, H. F. (1981). "Effect of sunlight on survival of indicator bacteria in seawater." Applied and Environmental Microbiology, 41(3), 690-696.

Gameson, A. L. H., and Gould, D. J. (1975). "Effects of solar radiation on the mortality of some terrestrial bacteria in sea water." Discharge of Sewage from Sea Outfalls, A. L. H. Gameson, ed., Pergamon Press, Oxford.

Gameson, A. L. H., and Gould, D. J. (1985). "Investigations of sewage discharges to some British coastal waters." Technical Report TA222, Water Research Centre.

Gameson, A. L. H., and Saxon, J. R. (1967). "Field studies on effect of daylight on mortality of coliform bacteria." Water Research, 1(4), 279-295.

Gannon, J. J., Busse, M. K., and Schillinger, J. E. (1983). "Fecal coliform disappearance in a river impoundment." Water Research, 17(11), 1595-1601.

Gathercole, C., and Ross, P. (1996). "Tackling the boolean even n parity problem with genetic programming and limited-error fitness." Genetic Programming 1997: Proceedings of the Second Annual Conference, J. R. Koza, K. Deb, M. Dorigo, D. B. Fogel, M. Garzon, H. Iba, and R. R. Riolo, eds., Morgan Kaufmann, San Francisco, CA, 119-127.

Gencay, R., and Liu, T. (1996). "Nonlinear Modelling and Prediction with Feedforward and Recurrent Networks " Physica D.

Gershenfeld, N. (1999). The Nature of Mathematical Modelling, Cambridge University Press.

Gledrich, E. E. (1978). "Bacterial populations and indicator concepts in faeces, sewage, storm water and solid wastes." Indicators in Water and Food G. Berg, ed., Ann Arbor Science Pubs, Ann Arbor, MI, 51-97.

Gledrich, E. E., and Kenner, B. (1969). "Concepts of faecal streptococci in stream pollution." Journal of Water Pollution Control Fedaration, 41:R, 336-352.

Gontar, M. Z., and Hatziargyriou, S. M. N. (2001). "Short Term Load Forecasting with Radial Basis Function Network." IEEE Porto Power Tech Conference, Porto, Portugal.

Gourmelon, M., Cillard, J., and Pommepuy, M. (1994). "Visible light damage to Escherichia coli in seawater: oxidative stress hypothesis." Applied Bacteriology, 77(1), 105-112.

Grace, R. A. (1978). Marine Outfall Systems, Planing, Design and Construction, Prentice-Hall, Inc.

Han, D., Kwong, T., and Li, S. (2007). "Uncertainities in real-time flood forecasting with neural networks." Hydrological Processes, 21, 223-228.

Han, J. (2002). "Application artificial neural networks for flood warning sytems," PhD thesis, North Carolina State University.

Harris, E., Falconer, R. A., Kay, D., and Stepleton, C. (2002). "Development of a modelling tool to quantify faecal indicator levels in Cardiff bay." Water & Maritime Engineering, 154(2), 129-135.

Harris, E. L., Babovic, V., and Falconer, R. A. (2003). "Velocity predictions in compound channels with vegetated floodplains using genetic programming." International Journal Of River Basin Management, 1(2), 117-123.

Harvey, R., et al. . (1984). Handbook of occupational hygiene, Kluwer Publishing, Brentford, Middlesex, UK.

Hecht-Nielsen, R. (1990). Neurocomputing, Addison-Wesley, Boston, MA.

Helmio, T. (2002). "Unsteady 1D flow model of compound channel with vegetated floodplains." Journal of Hydrology,, 269(1), 89-99.

Helmio, T. (2004). "Flow resistance due to lateral momentum transfer in partially vegetated rivers." Water Resources Research 40.

Henderson, F. M. (1966). Open channel flow, Macmillan Co. Ltd., 522 p.

Hettiarachchi, P., Hall, M. J., and Minns, A. W. (2005). "The extrapolation of artificial neural networks for the modelling of rainfall; A Titlemash runoff relationships." Journal of Hydroinformatics, 7, 291-296.

Hill, M. I., Carr, O. J., Birch, S. P., and Parker, D. M. (1996). "Cardiff Bay Barrage: Environmental Assessment and Impacts on the Natural Environment." Barrages: Engineering Design and Environmental Impacts, N. Burt and J. W. edt, eds., John Wiley and Sons Ltd, 209-218.

Hiraishi, A., Saheki, K., and Horie, S. (1984). "Relationships of total coliform, fecal coliform, and organic pollution levels in the Tamagawa river." Bulletin of the Japanese Society of Scientific Fisheries, 50(6), 991-997.

Hirsch, C. (1988). Numerical computation of internal and external flows, Volume 1: Fundamentals of Numerical Discrtization, John Willey & Sons Ltd., 515p.

Holland, J. (1975). Adaptation in natural and artificial systems, MIT Press, Cambridge, MA.

Holly, F. M. (1984). "Dispersion in Two-Dimensional Flow." ASCE J. Hydraulic Eng., 110, 905-926.

Hopfield, J. (1982). "Neurons, dynamics and computation " Physics Today, 47(2), 40-46.

Hornik, K., Stinchcombe, M., and White, H. (1989). " Multilayer feedforward networks are universal approximators." Neural Networks, 2, 359–366.

Huang, B., Lai, G., Qin, J., and Lin, S. (1999). "Experimental research on influence of vegetated floodplains upon flood carrying capacity of river." J. Hydrodynamics, 14(4), 468-474.

Huang, B., Lai, G., Qin, J., and Lin, S. (2002). "Hydraulics of compound channel with vegetated floodplains." J. Hydrodynamics, 14(1), 23-28.

Huang, J., and Li, M. (1997). "Finite difference approximation for the velocity- vorocity formulation on staggered and non-staggered grids." Computers and fluids, 25, 59-82.

HyderConsultingLtd. (1998). "Cardiff Bay Development Corporation: Environmental Protection Measures Consultancy. Water Quality Risk Assessment and Intervention Modelling Report. (SH11157/D14/1)."

Iliadis, S. L., and Maris, F. (2007). "An Artificial Neural Network model for mountainous water-resources management: The case of Cyprus mountainous watersheds." Environmental Modelling & Software, 22(7), 1066-1072.

Jagals, P., Grabow, W. K., and de-Villiers, J. C. (1995). "Evaluation of indicators for assessment of human and animal faecal pollution of surface run-off." Water Science and Technology, 31(5-6), 235-241.

Jannasch, H. W. (1956). "Vergleichende Bakterioogische Untersuchung der Adsorptionwirkung des Nil-Treibschalmmes." Ber. Limno. Flusstn Freudenthal, 7, 21-27.

Järvelä, J. (2002). "Determination of flow resistance of vegetated channel banks and floodplains." River Flow 2002, D. Bousmar and Y. Z. (eds), eds., Swets & Zeilinger, Lisse, 31-318.

Jassby, A., and Powell, T. (1975). "Vertical patterns of eddy diffusion during stratification in Castle Lake, California." Limnol. Oceanogr., 20, 530-543.

Jayawardena, A. W., and Fernando, D. A. (1998). "Use of radial basis funtion type artificial neural network for runoff simulation." Computer-aided Civil and Unfrastructure Engineering, 13, 91-99.

Jerlov, N. G. (1968). Optical Oceanography, Elsevier, Amstardam.

Jones, A. A. (2001). "The winGamma user guide."

Jones, F. H. (1994). "Barrage Developments in the Welsh Region: The role of the National Rivers Authority in Protecting the Aquatic Environment." Journal Institution of Water and Environmental Management, 8(4), 432-439.

Jordan, M. I. (1986). "Attractor dynamics and parallelism in a connectionist sequential machine." Proceedings of the Eighth Conference of the Cognitive Science Society, Amherst, MA, 531--546.

Jordanova, A. A., and James, C. S. (2003). "Experimental Study of Bed Load Transport through Emergent Vegetation." Journal of Hydraulic Engineering, 129(6), 474-478.

Joyce, T. M., McGuigan, K. G., Elmore-Meegan, M., and Conroy, R. M. (1996). "Inactivation of faecal bacteria in drinking water by solar heating." Applied and Environmental Microbiology, 62(2), 399-402.

Kadlec, R. H. (1990). "Overland flow in wetlands: vegetation resistance." Journal of Hydraulic Engineering, 116(5), 691-706.

Karunanithi, N., Grenney, W. J., Whitley, D., and Bovee, K. (1994). "Neural networks for river flow prediction." Journal of Computing in Civil Engineeirng 8(2), 201-220.

Kashefipour, S. M., Lin, B., Harris, E., and Falconer, R. A. (2002). "Hydro-environmental modelling for bathing water compliance of an estuarine basin." Water Research, 36, 1854-1868.

Kay, D., Stepleton, C., Wyer, M. D., McDonald, A. T., Crowther, J., Paul, N., Jones, K., Francis, C., Watkins, J., Wilkinson, J., Humphrey, N., Lin, B., Yang, L., Falconer, R. A., and Gardner, S. (2005). "Decay of intestinal enterococci concentrations in high-energy estuarine and coastal waters: towards real-time T90 calues for modelling faecal indicators in recreational waters." Water Research, 39, 655-667.

Keifa, M. A. (1998). "General regression neural network for driven piles in cohesionless soils." Journal of geotechnical and geoenvironmental Engineering, 124(12), 1177-1185.

Keijzer, M., and Babovic, V. (1999). "Dimensionally aware genetic programming." Proceedings of the Genetic and Evolutionary Computation Conference, W. Banzhaf, J. Daida, A. E. Eiben, M. H. Garzon, V. Honavar, M. Jakiela, and R. E. Smith, eds., Morgan Kaufmann, Orlando, Florida, USA. , 1069-1076.

Keijzer, M., and Babovic, V. (2000). "Genetic programming within a framework of computer-aided discovery of scientific knowledge." Proceedings of the Genetic and Evolutionary Computation Conference (GECCO-2000), Morgan Kaufmann, Las Vegas, Nevada, USA.

Keijzer, M., Baptist, M., Babovic, V., and Rodriguez, J. (2005). "Determining equations for vegetation induced resistance using genetic programming." Proceedings of the conference on Genetic and evolutionary computation Washington DC, USA

Khotanzad, A., Afkhami-Rohani, R., Tsun-Liang, L., Abaye, A., Davis, M., and Maratukulam, D. (1997). "ANNSTLF-a neural-network-based electric load forecasting system." IEEE Transactions on Neural Networks, 8, 835-846.

Kim, G., and Barros, A. P. (2001). "Quantitative forecasting using multisensor data and neural networks " Journal of Hydrology, 246, 46-62.

King, J. (1993). "Editorial Notes." Information Systems Research, 291-298.

Kirk, J. T. O. (1983). Light and photosynthesis in aquatic ecosystems, Cambridge University Press, Cambridge, England.

Knight, D. W. (2001). "Scoping study on reducing uncertainty in river flood conveyance—Conveyance in 1-D river models." Sch. of Civ. Eng., Univ. of Birmingham, Birmingham, UK.

Kohonen, T. (1982). "Self-organized formation of topologically correct feature maps." Biol. Cybern, 43.

Kohonen, T. (1989). Self-Organization and Associative Memory, Springer-Verlag, Berlin.

Kouwen, N., and Fathi-Moghadam, M. (2000). "Friction Factors for Coniferous Trees along Rivers." Journal of Hydraulic Engineering, 120(10), 732-740.

Kouwen, N., and Unny, T. E. (1973). "Flexible roughness in open channels." Journal of Hydraulic Division, 99(5), 713-728.

Koza, J. (1992). Genetic programming: on the programming of computers by means of natural selection, MIT Press.

Koza, J. R. (1994). Genetic Programming II: Automatic Discovery of Reusable Programs, MIT Press. , Cambridge MA.

Koza, J. R., Andre, D., Bennett, F. H., and Keane, M. (1999). Genetic Programming 3, Morgan Kaufman.

Koza, J. R., and Rice, J. P. (1991). "Genetic Generation of Both the Weights and Architecture for a Neural Network." International Joint Conference on Neural Networks, IEEE, Washington State Convention and Trade Center, Seattle, WA, USA, 397-404.

Krishnapura, V. G., and Jutan, A. (1997). "ARMA Neuron Networks for Modelling Nonlinear Dynamical Systems." Canadian Journal of Chemical Engineering, 75(3), 574-582.

Kumar, K. K. (1993). "Optimization of the neural net connectivity pattern using a backpropagation algorithm." Neurocomputing 5, 273-286.

Kundu, P. K. (1990). Fluid Mechanics, Academic Press Inc., 638p.

Lachtermacher, G., Fuller, J.D., 1994. . (1994). "Backpropagation in hydrological time series forecasting. ." Stochastic and Statistical Methods in Hydrology and Environmental Engineering. . K. W. Hipel, A. I. McLeod, U. S. Panu, and V. P. Singh, eds., Kluwer Academic, Dordrecht.

Langdon, W. (2000). "Size fair and homologous tree genetic programming crossovers." Genetic Programming and Evolvable Machines, 1, 95-119.

Langdon, W., and Poli, R. (2002). Foundations of Genetic Programming, Springer-Verlag, Berlin, Heidelberg, New York.

Langdon, W. B. (1998). Genetic Programming and Data Structures: Genetic Programming+ Data Structures= Automatic Programming, Kluwer Academic Publishers, Norwell, MA, USA.

Lantrip, B. M. (1983). "The Decay of Enteric Bacteria in an Estuary," Ph.D. Dissertation, School of Hygiene and Public Health. The Johns Hopkins University, Baltimore, Maryland.

Lapedes, A., and Farber, R. (1988). "How Neural Network Works." Neural Information Processing System, D. Z. Anderson, ed., American Institute of Physics, Ed. New York, 442-456.

Li, R. M., and Shen, H. W. (1973). "Effect of Tall Vegetations on Flow and Sediment " Journal of the Hydraulics Division, 99(5), 793-814.

Lim, C. H., and Flint, K. P. (1989). "The effects of nutrients on the survival of Escherichia coli in lake water." Journal of Applied Bacteriology, 66, 559-569.

Lin, B., and Falconer, R. A. (1995). "Modelling sediment fluxes in estuarine waters using a curvilinear co-ordinate system." Estuarine, Coastal and Shelf Science, 41, 413-428.

Lin, B., and Falconer, R. A. (1997b). "Three dimensional layer-integrated Modelling of estuarine flows with flooding and drying." Estuarine, coastal shelf science, 44, 737-751.

Lin, B., and Falconer, R. A. (1997). "Tidal Flow and Transport Modelling Using ULTIMATE QUICKEST Scheme " Journal of Hydraulic Engineering, 123(4), 303-314.

Lin, B., and Falconer, R. A. (1997a). "Tidal flow and transport modelling using ultimate quickest scheme." Journal of Hydraulics Engineering, ASCE, 123, 303-314.

Lin, B., Kashefipour, S. M., and Falconer, R. A. (2003). "Predicting near-shore coliform bacteria concentrations using ANNs." Water Science and technology, 48(10), 225-232.

Lin, T., Horne, B. G., Tin˜o, P., and Giles, C. L. (1996). "Learning long-term dependencies in NARX recurrent neural networks." IEEE Transactions on Neural Networks, 7(6), 1329-1338.

Linne, M., Kane, S., and Dell, G. (2000). A Guide to Appraisal Valuation Modelling, Appraisal Institute, Newburyport, MA.

López, F., and García, M. H. (2001). "Mean flow and turbulence structure of open channel flow through non-emergent vegetation." Journal of Hydraulic Engineering, 127(5), 392-402.

Luke, S. (1998). "Genetic Programming Produced Competitive Soccer Softbot Teams for RoboCup97." Proceedings of the Third Annual Conference, Japan.

Luke, S., and Panait, L. (2001). "A Survey and Comparison of Tree Generation Algorithms." GECCO-2001: Proceedings of the Genetic and Evolutionary Computation Conference, Springer-Verlag

Luke, S., and Spector, L. (1998). "A revised comparison of crossover and mutation in genetic programming." Genetic Programming 1998: Proceedings of the Third Annual Conference, W. B. J. Koza, K. Chellapilla, K. Deb, M. Dorigo, D. Fogel, M. Garzon, D. Goldberg, H. Iba, and R. Riolo, ed., Morgan Kaufmann.

Luke, S., and Spector., L. (1996). "Evolving Teamwork and Coordination with Genetic Programming." Proceedings of the First Annual Conference on Genetic Programming, J. R. Koza, ed., MIT Press1996, Cambridge, MA., 150-156.

Mack, W. N. (1977). "Total coliform bacteria." A. W. Hoadley and B. J. Dutka, eds., American Society for Testing and materials, Philadelphia.

Maier, H. R., and Dandy, G. C. (1998). "The effect of internal parameters and geometry on the performance of back-propagation neural networks: an empirical study " Environmental Modelling and Software,, 13(2), 193-209.

Maier, H. R., and Dandy, G. C. (2000). "Neural network for the prediction and forecasting of water resources variables: A review of modelling issues and applications." Environmental Modelling and Software, 15, 101-124.

Malsburg, C. V. d. (1973). "Self-organization of orientation selective cells in the striate cortex." Kybernetik, 14, 85-100.

Mancini, J. L. (1978). "Numerical estimates of coliform mortality rates under various conditions." Journal of Water Pollut. Control Fed., 50, 2477-2488.

Maren, A., and Harston, C. (1990). Handbook of Neural Computing Applications, Academic Press, San Diego, CA.

Marshall, K. C. (1978). "The Effects of Surfaces on Microbial Activity." Water Pollution Microbiology, 2, 51-70.

Martin, J. L., and McCutcheon, S. C. (1999). Hydrodynamics and transport for water quality modelling, CRC Press Inc.

Masters, T. (1993). Practical Neural Network Recipes in C++, Academic Press, Boston.

Maynerd-Smith, J. (1975). The theory of evolution, Penguin, London.

McCambridge, J., and McMeekin, T. A. (1981). "Effect of solar radiation and predacious microorganisms on survival of fecal and other bacteria." Journal of Applied and Environmental Microbiology, 41(5), 1083 -1087.

McCulloch, W. S., and Pitts, W. H. (1943). "A logical calculus of the ideas immanent in nervous activity." Bulletin of Mathematical Biophysics, 5, 115-133.

McFeters, G. A., and Stuart, D. G. (1972). "Survival of coliform bacteria in natural waters: field and laboratory studies membrane-filter chambers." Applied Microbiology, 24(5), 805-811.

Milne, D. P., Curran, J. C., Findlay, J. S., Crowther, J. M., and Wallis, S. G. (1989). "The effect of estuary type suspended solids on survival of E.Coli in saline water." Water Science and Technology, 21(3), 61-65.

Milne, D. P., Curran, J. C., and Wilson, L. (1986). "Effects of sedimentation on removal of faecal coliform bacteria from effluents in estuarine water." Water Research, 20(12), 1465-1611.

Minns, A. W. (1998). "Artificial neural network as subsymbolic precess descriptors," PhD Thesis, IHE Delft, The Netherlands.

Minns, A. W., and Hall, M. J. (1996). "Artificial neutral networks as rainfall-runoff models." Hydrological Sciences Journal, 41(3), 399-417.

Mitchell, R., and Chamberlin, C. (1978). "Survival of indicator organisms." Indicators of Viruses in Water and Food, G. Berg, ed., Ann Arbor Science Publishers, Inc, Ann Arbor, Michigan, 15-37.

Mitchell, R., and Chamberlin, C. E. (1975). "Factors influencing the survival of enteric microorganisms in the sea: An overview." Discharge of sewage from sea outfalls, A. L. H. Gameson, ed., Pergamon Press, Oxford, 237-251.

Mitchell, R., and Morris, J. (1969). "The Fate Of Intestinal Bacteria In The Sea " Advances in water pollution research, international conference on water pollution research, 811-821.

Moeller, J. R., and Calkins, J. (1980). "Bactericidal agent in wastewater lagoons and lagoon design." Journal of Water Pollut. Control Fed., 52.

Montana, D. J. (1995). "Strongly typed genetic programming." Evolutionary Computation, 3(2), 199-230.

Moriñigo, M. A., Wheeler, D., Berry, C., Jones, C., Muñoz, M. A., Cornax, R., and Borrego, J. J. (1992). "Evaluation of different bacteriophage groups as faecal indicators in contaminated natural waters in southern England." Water Research, 26(3), 267-271.

Nam, K., and Schaefer, T. (1995). "Forecasting international airline passenger traffic using neural network." Logistics and Transportation Review, 31(3), 239–251.

Naot, D., Nezu, I., and Nakagawa, H. (1993). "Calculation of Compound-Open-Channel Flow " Journal of Hydraulic Engineering, 119(12), 1418-1426.

Naot, D., Nezu, I., and Nakagawa, H. (1996). "Hydrodynamic Behaviour of Partly Vegetated Open Channels." Journal of Hydraulic Engineering, 122(11), 625-633.

Naot, D., and Rodi, W. (1982). "Calculation of secondary currents in channel flow." Journal of Hydraulic Division, 108(8), 948-969.

Neelakantan, T. R. (2005). "Artificial Neural Network: An Overview." Artificial Neural Networks in Water Supply Engineering, S. Lingireddy and G. M. Brion, eds., American Society of Civil Engineers, Reston Verginia, 10-22.

Nepf, H. M. (1999). "Drag, Turbulence and Diffusion in Flow through Emergent Vegetation." Water Resources Res, 35(2), 479-489.

Noble, R. T., Dorsey, J. H., Leecaster, M., Orozco-Borbon, V., Reid, D., Schiff, K., and Weisberg, S. B. (2000). "A regional survey of the microbiological water quality along the shoreline of the Southern California bight." Environmental Monitoring and Assessment, 64, 435-447.

Okhuysen, P. C., Chappell, C. L., Crabb, J. H., Sterling, C. R., and DuPont, H. L. (1999). "Virulence of three distinct Cryptosporidium parvum isolates for healthy adults." Journal of Infectious Diseases, 180, 1275-1281.

Orlob, G. T. (1956). "Viability of sewage bacteria in seawater." Sewage Industr.Wastes, 28, 1147-1167.

Owen, P. H., and Falconer, R. A. (1987). "Numerical solution of flooding and drying in a depth averaged tidal flow model." Proceedings of Institute of Civil Engineers, Water Engineering Group, Part 2, Vol 83, pp 161-180.

Park, A. J., and Sandberg, A. I. W. (1991). "Universal approximation using radial-basis-function networks." Journal of Neural Comput. , 3(2), 246-257.

Pasche, E., and Rouve, G. (1985). "Overbank Flow with Vegetatively Roughened Flood Plains." Journal of Hydraulic Engineering, 111(9), 1262-1278.

Peaceman, D., and Rachford, M. (1955). "The numerical solution of parabolic and elliptic differential equations." Journ. Soc. Indust. Appl. Math, 3(28-41).

Petryk, S., and Bosmajian, G. (1975). "Analysis of Flow through Vegetation " Journal of the Hydraulics Division, 101(7), 871-884.

Pham, D. T., and Liu, X. (1995). Neural Networks for Identification, Prediction, and Control, Springer-Verlag, New York.

Philipp, R. (1991). "Risk assessment and microbiological hazards associated with recreational water sports." Reviews in Medical Microbiology, 2, 208-214.

Pike, E. B., Balarajan, R., and Jones, F. (1991). "Health effect of sea bathing (ET 9511) Phase II- studies at Ramsgate and Moreton 1990,1991." DoE 2736-M(P), Department of Environment.

Pitt, R. (1998). Epidemiology and stormwater management, CRC/Lewis Publishers, New York.

Preston, R. W. (1985). "Representation of Dispersion in Two-dimensional Water Flow." Central Electricity Research Laboratories, Leatherhead, England, Report No. TPRD/L/2783/N84, pp 1-13.

Prieto, M. D., Lopez, B., Juanes, J. A., Revilla, J. A., Llorca, J., and Delgado-Rodriguez, M. (2001). "Recreation in coastal waters: health risks associated with bathing in sea water." Journal of Epidemiology & Community Health, 55, 442-447.

Pruss, A. (1998). "Review of epidemiological studies on health effects from exposure to recreational water." International Journal of Epidemiology, 27, 1-9.

Reece, G. R. (1976). "A generalize Reynolds stress model of turbulence," Phd Thesis, University of London, London.

Reed, R. H. (1997). "Solar inactivation of faecal bacteria in water: the critical role of oxygen." Letters in Applied Microbiology, 24, 279-280.

Rees, G. (1993). "Health implications of sewage in coastal waters-The Bristish case'." Marine Pollution Bulletin, 26, 14-19.

Reilly, A., and Cooper, A. (1990). "An overview of neural networks: early models to real world systems." An introduction to neural and electronic networks Academic Press Professional, Inc, 227-248.

Rojas, R. (1996). Neural Networks: A Systematic Introduction, Springer-Verlag, Berlin.

Rosca, J. (1997). " Hierarchical learning with procedural abstraction mechanisms," PhD Thesis, University of Rochester., Rochester, NY.

Rosca, J. P., and Ballard, D. H. (1996). "Complexity drift in evolutionary computation with tree representation." Tech Report, University of Rochester, Computer Science.

Rosen, B. H. (2000). "Waterborne Pathogens in Agricultural Watersheds." U.S. Department of Agricultural, Natural R4esources Conservation Service, Watershed Institute.

Roszak, D. B., and Colwell, R. R. (1987). "Metabolic activity of bacterial cells enumerated by direct viable count." Appl. Environ. Microbiol, 53(2889-2893).

Rumelhart, D., and McClelland, J. (1986). "Parallel Distributed Processing." MIT Press, Cambridge, MA.

Rumelhart, D. E., Hinton, G. E., and Williams, R. J. (1986). "Learning internal representations by error propagation." Parallel Distributed Processing, D. E. Rumelhart and J. L. McClelland, eds., MIT Press, 318-362.

Ryan, C., Collins, J. J., and O'Neill, M. (1998). "Grammatical Evolution: Evolving Programs for an Arbitrary Language." Proceedings of the First European Workshop on Genetic Programming (EuroGP'98), Lecture Notes in Computer Science 1391, MIT Press, Cambridge, MA.

Salas, H. J. (1998). "History and Application of Microbiological Water Quality Standards in the Marine Environment." Centro Panamericano de Ingenieria Sanitaria y Ciencias del Ambiente and Pan American Health Organization, Lima, Peru.

Sayler, G. S., Nelson, J. D., Justice, A., and Colwell, R. R. (1975). "Distribution and significance of faecal indicator organisms in the upper Chesapeake Bay." Journal of Applied Microbiology, 30, 625-638.

Schiffmann, W., Joost, M., and Werner, R. (1993). "Application of Genetic Algorithms to the Construction of Topologies for Multilayer Perceptrons " Proceedings of the International Conference on Artificial Neural Networks and Genetic Algorithms, 975-682.

Schillinger, J., and Gannon, J. (1982). "Coliform attachment to suspended particles in stormwater." US Environmental Protection Agency, NTIS PB 83-108324

Setiono, R., Leow, W. K., and Zurada, J. M. (2002). "Extraction of Rules From Artificial Neural Networks for Nonlinear Regression." IEEE Transactions On Neural Networks, 13(3).

Sharda, R., and Patil, R. (1992). "Connectionist Approach to Time Series Prediction: An Empirical Test." Journal of Intelligent Manufacturing, 3, 317-323.

Shu, C., and Chew, Y. T. (1998). "On the equivalence of generalized differential quadrature and highest finite difference scheme." Computer Methods Applied Mechanics in Engineering, 155, 249-260.

Sinton, L. W., Finlay, R. K., and Lynch, P. A. (1999). "Sunlight Inactivation of Faecal Bacteriophages and Bacteria in Sewage-Polluted Seawater." Applied and Environmental Microbiology, 65(8), 3605-3613.

Smith, R. (1992). "Physics of Dispersion in Coastal and Estuarine Pollution: Methods and Solutions." Scottish Hydraulics Study Group, Glasgow.

Snedecor, J. (2003). "Modelling Bacterial Water Quality with Systems thinking Software." Getting it Done: The role of TMDL Implementation In Watershed restoration, Stevenson, WA.

Sokolov, Y. N. (1980). "Hydraulic resistance of floodplains." Water Resources Res, 5, 563-572.

Solic, M., and Krstulovic, N. (1992). "Separate and combined effects of solar radiation, temperature, salinity and pH on the survival of faecal coliform in seawater." Marine Pollution Bulletin, 24(8), 411-416.

Solomanite, D. P. (2002). "Data driven modelling: paradigm, methods, experience." Hydroinformatics, I. D. Cluckie, D. Han, J. P. Davies, and S. Heslop, eds., IWA Publishing, Cardiff, UK.

Specht, D. F. (1991). "A general regression neural network." EEE Trans. Neural Netw., 2(6), 568–576.

Spector, L. (1996). "Simultaneous Evolution of Programs and their Control Structures." In Advances in Genetic Programming P. Angeline and K. Kinnear, eds., MIT Press., Cambridge, MA.

Spector, L., and Luke., S. (1996). "Cultural Transmission of Information in Genetic Programming." Genetic Programming 1996: Proceedings of the First Annual Conference, Cambridge,MA: The MIT Press.

Spiegler, I. (2000). "Knowledge Management: A New Idea or a Recycled Concept?" Communications of AIS, Volume 3(Article 14).

Srinivasan, D., Liew, A. C., and Chang, C. S. (1994). "A Neural-Network Short-Term Load Forecaster." Electric Power Systems Research, 28(3), 227-234.

Stapleton, C. M., Wyer, M. D., Kay, D., M, M. B., Humphrey, N., and etal. (2007). "Fate and Transport of Particles in Estuaries." Volume IV: Numerical Modelling for Bathing Water Enterococci Estimation in the Severn Estuary, Environment Agency Science Report SC000002/SR4.

Stelling, G. S. (1986). "Practical Aspects of Accurate Tidal Computations." J. Hydr. Engng, 112, 802-817.

Stelling, G. S., Wiersma, A. K., and Willemse, J. (1985). "Practical Aspects of accurate todal computations." Journal of Hydraulic Engineering, ASCE, 112, 802-817.

Stone, B. M., and Shen, H. T. (2002). "Hydraulic Resistance of Flow in Channels with Cylindrical Roughness." Journal of Hydraulic Engineering, 128(5), 500-506.

Tannehill, J. C., Anderson, D. A., and Pletcher, R. H. (1997). Computational fluid mechanics and heat transfer, Taylor and Fransis Series in Computational and Physical Processes in Mechanics and Thermal Sciences, 792p.

Temple, D. M., Robinson, K. M., Ahring, R. M., and Davis, A. G. (1987). "Stability Design of Grass-Lined Open Channels." USDA Agriculture Handbook 667, U S Department of Agriculture, Washington D.C.

Thomann, R. V., and Mueller, J. A. (1987). Principals of surface water quality modelling and control, Harper & Row, New York.

Thomas, L. H. (1949). "Elliptic problems in linear difference equations over a netwo." Watson Sci. Comput. Lab. Rep., Columbia University, New York.

Tominaga, A., and Nezu, I. (1991). "Turbulent structure in compound channel flows." Journal of Hydraulic Engineering, 117(1), 21-41.

Tominaga, A., Nezu, I., Ezaki, K., and Nakagawa, H. (1989). "Three-dimensional turbulent structure in straight open channel flows." Journal of Hydraulic Research, 27(1), 149-173.

Towell, G. G., Craven, M. K., and Shavlik, J. W.(1991) "Constructive induction in knowledge-based neural networks." Proceedings of the 8th International Workshop on Machine Learning, San Mateo, 213-217.

Tsihrintzis, V. A. (2001). "Variation of roughness coefficients for unsubmerged and submerged vegetation." Journal of Hydraulic Engineering, 127(3), 239-245.

USEPA. (2002). "National Water Quality Inventory - 2000 Report." EPA-841-R-02-001, Office of Water, Washington, DC.

USEPA. (2003). "Long Term 2 Enhanced Surface Water treatment (LT2ESWT) - proposed rule." Federal Register, 68 (154) 47640- 47795.

Vail, E. F. (1999). "Knowledge Mapping: Getting Started with Knowledge Management." Information Systems Management, Fall, 16-23.

van-Prooijen, B. C., Battjes, J. A., and Uijttewaal, W. S. J. (2005). "Momentum exchange in straight uniform compound channel flow " Journal of Hydraulic Engineering, 131(3), 175-183.

Versteeg, H. K., and Malalasekara, W. (1995). An introduction to computational fluid dynamics, the finite volume method, Longman Group Ltd., 275p.

Warner, B., and Misra, M. (1996). "Understanding neural networks as statistical tools. ." American Statistician, 50(4), 284-293.

Wegelin, M., S Canonica, Meschner, K., Fleischmann, T., Pesaro, F., and Metzler, A. (1994). "Solar Water disinfection: scope of the process and analysis of radiation experiments." Journal of Water SRT-Aqua, 43, 154-169.

Weigned, A. S., Rummelhart, D. E., and Hubberman, B. A. (1992). "Predicting sunspots and exchange rates with connectionist networks." Nonlinear Modelling and Forecasting., M. Casdagli and S. Eubank, eds., Addison-Wesley, Redwood City, CA,, 395-432.

Weiss, N., and Hassett, M. (1987). Introductory Statistics, Addison-Wesley, USA.

Westwater, D. (2001). "Modelling Hydrodynamic and Shallow Water Processes over Vegetated Floodplains," Cardiff University, UK.

Wetzel, R. G. (1983). Limnology, Saunders College Publishers, Philadelphia, PA.

Whigham, P. A. (1995). "Grammatically-biased Genetic Programming." In: J. Rosca, Edt, Proceedings of the 1995 Workshop on Genetic Programming, Morgan- Kaufmann, pp. 33–41.

WHO. (1999). "Health-based monitoring of recreational waters: the feasibility of a new approach (the 'Annapolis Protocol')." Geneva, World Health Organization.

WHO. (2000). "Monitoring Bathing Waters - A Practical Guide to the Design and Implementation of Assessments and Monitoring Programmes ".

Widrow, B., and Hoff, M. E. (1960). "Adaptive Switching Circuits." IRE WESCON Convention Record, New York, 96 - 104.

Williams, R. J., and Zipser, D. (1989). "A learning algorithm for continually running fully recurrent networks." Neural Computation, 1, 270-280.

Wu, Y., and Falconer, R. A. (1998). "Refined two-dimensional ultimate quickest scheme for conservative solute transport modelling." Proc. of Third International Conference on Hydro-Science and Engineering Cottbus, Germany, Vol 1, No 143 pp 1-13.

Wu, Y., Falconer, R. A., and J, S. (2001). "Mathematical modelling of tidal currents in mangrove forests." Environmental Modelling & Software, 16(1), 19-29.

Wurman, R. S. (1989). Information Anxiety, Doubleday, New York.

Yang, K., Cao, S., and Knight, D. W. (2007). "Flow Patterns in Compound Channels with Vegetated Floodplains." Journal of Hydraulic Engineering, 133(2), 148-159.

Yen, B. C. (2002). "Open Channel Flow Resistance." Journal of Hydraulic Engineering, 128(1), 20-39.

Yin, C., Rosendahla, L., and Luob, Z. (2003). "Methods to improve prediction performance of ANN models " Simulation Modelling Practice and Theory 11(3-4), 211-222.

Yin, J., Falconer, R. A., Chen, Y., and Probert, S. D. (2000). "Water and sediment movements in harbours." Applied Energy, 67, 341-352.

Zoppou, C., Roberts, S., and Renka, R. J. (2000). "Exponential spline interpolation in characteristics based scheme for solving the advective-diffusion equation." International Journal for Neumerical Methods in Fluids, 33, 429-452.