UNIVERSITY OF CALIFORNIA

Los Angeles

An Evaluation of ARIMA (Box-Jenkins) Models for Forecasting Wastewater Treatment Process Variables

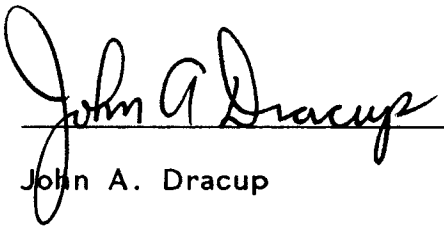
A thesis submitted in partial satisfaction of the requirements for the degree Master of Science in Engineering

by

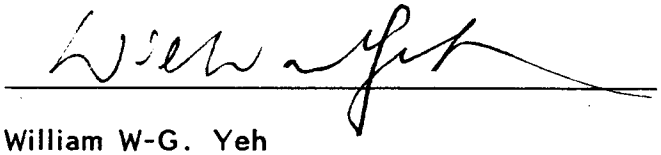
Kevin Michael Reagan

1984

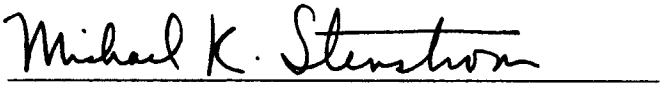
The thesis of Kevin Michael Reagan is approved.



John A. Dracup



William W-G. Yeh



Michael K. Stenstrom, Committee Chair

University of California, Los Angeles

1984

DEDICATION

This thesis is dedicated to Joan and Simone Fujita, whose devotion and understanding were instrumental in its completion.

CONTENTS

DEDICATION iii

ACKNOWLEDGEMENTS vii

ABSTRACT OF THE THESIS viii

Chapter page

I. INTRODUCTION 1

II. BACKGROUND AND OVERVIEW OF THE METHODOLOGY 6

Introduction 6
Quantitative Forecasting 7
General References for Box-Jenkins Modeling 8
Stochastic Processes and Time Series 8
Stochastic vs. Deterministic Models 10
Univariate Box-Jenkins Models: The Basic Idea 10
Autoregressive (AR) Models 14
Moving Average (MA) Models 14
Mixed Autoregressive and Moving Average (ARMA) Models . . 15
ARIMA Models and Stationary Stochastic Processes 15
Multivariate ARIMA Models 18
The Iterative Box-Jenkins Modeling Strategy 19
The Identification Stage 21
The Estimation Stage 25
The Diagnostic Checking Stage 27
Forecasting 29
One-Step-Ahead Within-Sample Forecasts 30
Multi-Step-Ahead Beyond-Sample Forecasts 32
One-Step-Ahead Beyond-Sample Forecasts 33
Summary of Forecast Classifications 35
Seasonality and Seasonal ARIMA Models 36
ARIMA Algebra and Notation 39
The Principle of Parsimony 41

III. LITERATURE REVIEW 42

Background For the Present Research 42
Scope of the Review 43
Previous Applications of Univariate ARIMA Models 44

IV. ARIMA MODELING AND FORECASTING OF WASTEWATER TREATMENT DATA . 56

Introduction 56
Methods of Analysis and Evaluation Criteria 57
Madison Nine Springs Treatment Plant Influent BOD 59
Nonseasonal ARIMA(1,0,0) Model 61
Fitted Values 61
Multi-Step-Ahead Beyond-Sample Forecasts 65
One-Step-Ahead Beyond-Sample Forecasts 68
Fitting and Forecasting Accuracy 71
Seasonal ARIMA(2,0,0)(0,1,1) Model 73
Fitted Values 75
Multi-Step-Ahead Beyond-Sample Forecasts 78
One-Step-Ahead Beyond-Sample Forecasts 80
Fitting and Forecasting Accuracy 80
Comments About the Nine Springs Models 84
Minneapolis-Saint Paul Sewer Station 004 COD 84
Nonseasonal ARIMA(0,1,1) Model 84
Fitted Values 85
Multi-Step-Ahead Beyond-Sample Forecasts 85
One-Step-Ahead Beyond-Sample Forecasts 88
Fitting and Forecasting Accuracy 88
Atlanta R. M. Clayton Treatment Plant Influent BOD 91
Seasonal ARIMA(3,0,0)(0,1,1) Model 91
Model Identification, Estimation, and Diagnostic Checking 91
Fitted Values 92
Multi-Step-Ahead Beyond-Sample Forecasts 95
One-Step-Ahead Beyond-Sample Forecasts 95
Fitting and Forecasting Accuracy 95
Atlanta R. M. Clayton Treatment Plant Total Suspended Solids 99
Seasonal ARIMA(2,0,0)(0,1,1) Model 99
Model Identification, Estimation, and Diagnostic Checking 99
Fitted Values 100
Multi-Step-Ahead Beyond-Sample Forecasts 103
One-Step-Ahead Beyond-Sample Forecasts 103
Fitting and Forecasting Accuracy 106
Minneapolis-Saint Paul Interceptor Sewer Flow Rate . . . 107
Seasonal ARIMA(0,1,0)(0,1,1) Model 107
Model Identification, Estimation, and Diagnostic Checking 107
Fitted Values 108
Multi-Step-Ahead Beyond-Sample Forecasts 111
One-Step-Ahead Beyond-Sample Forecasts 111
Fitting and Forecasting Accuracy 114

V. SIMULATION OF REAL-TIME CONTROL WITH ARIMA FLOW FORECASTING 116

Introduction 116

Methodology 117
Description of the Simulation Runs 120
Results 121
Discussion of Results 123

VI. CONCLUSIONS 126

VII. RELATED TOPICS FOR FUTURE RESEARCH 132

Introduction 132
Effect of the Number of Observations 132
Practical Significance of Forecast Errors and Error Criteria 133
Deseasonalization and ARIMA Models With Seasonal Parameters 134
Influent-Influent Transfer Function Models 135
Automated Box-Jenkins Modeling 135
Comparison of Currently Used Forecasting Techniques . . 136

REFERENCES 137

Appendix page

A. SUMMARY OF ARIMA MODELS PRESENTED 144

B. DEFINITIONS OF ERROR CRITERIA USED 149

LIST OF FIGURES

Figure		page
1.	Time Series Viewed as Result of Random Shocks	11
2.	The Iterative Box-Jenkins Modeling Strategy	20
3.	Example of Computer-Plotted Sample ACF and PACF	22
4.	Nine Springs Influent BOD Data	60
5.	Fitted Values from ARIMA(1,0,0) Model	62
6.	Fitted Values and Residuals from ARIMA(1,0,0) Model	63
7.	Detail of Fitting Error from ARIMA(1,0,0) Model	64
8.	Multi-Step-Ahead Forecasts from ARIMA(1,0,0) Model	66
9.	One-Step-Ahead Forecasts from ARIMA(1,0,0) Model	69
10.	Fitted Values and Residuals from Seasonal Model	76
11.	Detail of Fitting Error from Seasonal Model	77
12.	Multi-Step-Ahead Forecasts from Seasonal Model	79
13.	One-Step-Ahead Forecasts from Seasonal Model	81
14.	Fitted Values and Residuals from ARIMA(0,1,1) Model	86
15.	Multi-Step-Ahead Forecasts from ARIMA(0,1,1) Model	87
16.	One-Step-Ahead Forecasts from ARIMA(0,1,1) Model	89
17.	Fitted Values and Residuals from Atlanta BOD Model	93
18.	Detail of Fitting Error from Atlanta BOD Model	94
19.	Multi-Step-Ahead Forecasts from Atlanta BOD Model	96
20.	One-Step-Ahead Forecasts from Atlanta BOD Model	97

21. Fitted Values and Residuals from Atlanta TSS Model 101

22. Detail of Fitting Error from Atlanta TSS Model 102

23. Multi-Step-Ahead Forecasts from Atlanta TSS Model 104

24. One-Step-Ahead Forecasts from Atlanta TSS Model 105

25. Fitted Values and Residuals from Flow Rate Model 109

26. Detail of Fitting Error from Flow Rate Model 110

27. Multi-Step-Ahead Forecasts from Flow Rate Model 112

28. One-Step-Ahead Forecasts from Flow Rate Model 113

29. Scour Profiles from the Simulation Runs 122

LIST OF TABLES

Table		page
1.	Error Criteria for ARIMA(1,0,0) Model	71
2.	Error Criteria for ARIMA(2,0,0)(0,1,1) Model	82
3.	Error Criteria for ARIMA(0,1,1) Model	90
4.	Error Criteria for ARIMA(3,0,0)(0,1,1) Model	98
5.	Error Criteria for ARIMA(2,0,0)(0,1,1) Model	106
6.	Error Criteria for ARIMA(0,1,0)(0,1,1) Model	114
7.	Scour Variances and Percentage Variance Reductions	123

ACKNOWLEDGEMENTS

There are a number of people who had a significant impact on the course of this thesis, as well as on my general development as a water resources systems analyst. I would like to thank Professor John Dracup, who taught me many practical aspects of the field of water resources engineering. I also owe a great deal to Professor William Yeh, who strengthened my analytical background. The greatest thanks are of course due to my advisor and chairman, Professor Michael Stenstrom, who inspired my thesis topic, helped me perform the simulations, and provided a role model for what a true water pollution "systems engineer" should be.

I would be remiss if I did not give significant credit to two other persons. Dr. Dominique Hanssens of the UCLA Graduate School of Management was gracious enough to meet with me several times to discuss my work; his comments led to substantial gains in my understanding of Box-Jenkins modeling. I would also like to thank my fellow student Prasanta Bhunia, who spent many late hours patiently discussing my approach, and never found any question too naive.

Finally, I mention that I greatly admire the work of Drs. P. M. Berthouex and W. G. Hunter of the University of Wisconsin at Madison; it is in the spirit of their approach to water pollution data analysis that this thesis is presented.

ABSTRACT OF THE THESIS

An Evaluation of ARIMA (Box-Jenkins) Models

for

Forecasting Wastewater Treatment Process Variables

by

Kevin Michael Reagan

Master of Science in Engineering

University of California, Los Angeles, 1984

Professor Michael K. Stenstrom, Chair

In the water pollution control field, ARIMA (Box-Jenkins) models have been applied to model the flow rate and composition of wastewater treatment plant influent and effluent. ARIMA forecasting has also been proposed, with the objective of maintaining real-time control based on current measurements and short-term predictions of the important treatment process variables. To achieve successful control, reasonably accurate predictions of future values are required. However, few published results have reported or even addressed the forecasting accuracy of ARIMA models when applied to wastewater treatment data.

This thesis presents results obtained when ARIMA models were developed for field-measured wastewater data sets and then employed

for forecasting. The forecasts are compared to the values which subsequently occurred, and various measures of forecast error are presented and discussed. A uniform approach to classifying forecasts and presenting results is proposed for such studies. The forecasting accuracy of the models is found to vary for each data set, with average forecast errors ranging from 4% to 24% of the observed values.

In addition, ARIMA forecasting is performed in conjunction with a dynamic wastewater treatment plant simulation model to evaluate the benefits of a real-time control strategy incorporating hourly predictions of influent flow rate. The ARIMA forecasts are found to improve the control strategy, providing an additional reduction in process variability of 6% beyond the reduction achieved by the corresponding control strategy without flow prediction.

Chapter I

INTRODUCTION

> "When the Lord created the world and people to live in it--an enterprise which, according to modern science, took a very long time--I could well imagine that He reasoned with Himself as follows: 'If I make everything predictable, these human beings, whom I have endowed with pretty good brains, will undoubtedly learn to predict everything, and they will thereupon have no motive to do anything at all, because they will recognize that the future is totally determined and cannot be influenced by any human action. On the other hand, if I make everything unpredictable, they will gradually discover that there is no rational basis for any decision whatsoever and, as in the first case, they will thereupon have no motive to do anything at all. Neither scheme would make sense. I must therefore create a mixture of the two. Let some things be predictable and let others be unpredictable. They will then, amongst many other things, have the very important task of finding out which is which.'"
>
> From the book Small is Beautiful
> by E. F. Schumacher

Interest in forecasting the future has captivated the imagination of mankind throughout history. From tarot cards and horoscopes to highly sophisticated computer models for weather prediction, forecasting methods have been developed and are being used routinely in various aspects of daily life. In many scientific and managerial applications, short-term predictions for the next few values in a series of numbers or measurements can be extremely useful for planning, preparing, or controlling the system under study. In 1970,

following a series of articles, George E. P. Box and Gwilym M. Jenkins published a book entitled Time Series Analysis, Forecasting and Control, in which they set forth a comprehensive methodology for modeling and forecasting time series data. Their work has had a far-reaching impact on the entire practice of time series analysis, and has found broad application in numerous fields--particularly economics, management science, and the physical sciences. In the context of the present thesis, it can be stated that the Box-Jenkins approach is one of the most well-known and "fashionable" stochastic modeling techniques in water resources systems engineering today.

As with all mathematical models, much has been written about the theoretical assumptions, implications, and limitations of Box-Jenkins models, also known as ARIMA models (see Chapter 2). The orientation of the present research is, by contrast, quite practical and empirical.

In the field of wastewater treatment, ARIMA models have been used to analyze the flow rate and composition of treatment plant influent and effluent. Forecasting has also been proposed, with the objective of maintaining real-time control based on current measurements and anticipated future values of important process variables. To achieve successful control, reasonably accurate predictions of future values are required. However, few published results have reported or even addressed the forecasting accuracy of ARIMA models when applied to wastewater treatment data. This is in surprising contrast to other fields which utilize quantitative forecasting methods, where various

measures of forecast error are routinely reported in the presentation of results (Carbone and Armstrong, 1982). Numerous quantitative forecasting methods exist; almost all are easier, less costly, and less labor-intensive than ARIMA models (Makridakis et al., 1982). Hence, it is important for the wastewater treatment field to gain experience concerning the accuracy of ARIMA models, so that they may be compared to simpler forecasting procedures.

This thesis is intended to promote better understanding than presently exists regarding the practical utility and limitations of ARIMA models in the wastewater treatment field. From reviewing the water pollution literature it is evident that, thus far, ARIMA models have remained within the realm of the theoreticians. This is because, for the most part, theoreticians have not attempted to make their results understandable to the practitioners. Admittedly, the Box-Jenkins methodology requires some background in probability theory and mathematical statistics that cannot be imparted within a single research paper. However, it should be possible to convey the basic idea of forecasting, and to present practical results accompanied by a discussion of their significance for engineering applications. That is the basic goal of this thesis.

The specific objectives of the present research are to:

1. Perform Box-Jenkins time series analysis and forecasting for some illustrative wastewater treatment data sets, and present the results in a way which provides a much more revealing picture of the forecasting performance of ARIMA models than has previously been reported,

2. apply an ARIMA model in conjunction with a dynamic wastewater treatment plant model to simulate real-time control of an operating treatment plant, where forecasts of the future influent flow rate are used in the control algorithm, and
3. compare the simulated treatment performance obtained using a control strategy incorporating Box-Jenkins forecasting with results from the corresponding control strategy without forecasting.

It is helpful to briefly describe the organization of the thesis. In Chapter 2, an introductory overview of Box-Jenkins modeling and forecasting is given. This provides the background and terminology necessary for presenting the subsequent literature review, Chapter 3. The review chapter in turn describes the forecasting results obtained by previous researchers, and indicates the need for improved, standardized terminology and evaluation criteria when reporting such results.

Chapter 4 presents results obtained when ARIMA models were developed for field-measured wastewater data sets and then employed for forecasting. The forecasts are compared to the values which subsequently occurred, and various aspects of forecast error are discussed. A uniform approach to classifying forecasts and presenting results is proposed for such studies.

Chapter 5 represents the portion of the thesis where an ARIMA model is applied in a way which, to the author's knowledge, has not been previously reported. Box-Jenkins forecasting is performed in

conjunction with a dynamic wastewater treatment plant simulation model to evaluate the benefits of a real-time control strategy incorporating predictions of hourly influent flow rate as inputs to the controller. The results of this simulation are compared to the results from the corresponding control strategy without flow prediction.

Chapter 6 presents the conclusions from the literature review and the modeling studies of Chapters 4 and 5. Chapter 7 provides comments and recommendations concerning possible areas for future additional research.

Chapter II

BACKGROUND AND OVERVIEW OF THE METHODOLOGY

"The true logic of this world is in the calculus of probabilities."

James Clerk Maxwell

2.1 INTRODUCTION

This chapter introduces the general subject of quantitative forecasting and presents a qualitative overview of the basic concepts and procedures involved in Box-Jenkins modeling. Several references of interest to those unfamiliar with the Box-Jenkins approach are cited.

The most important concept presented in this chapter is the classification of Box-Jenkins forecasts into several distinct categories. This distinction is necessary for understanding the results obtained by previous researchers as discussed in the subsequent literature review chapter.

2.2 QUANTITATIVE FORECASTING

The Box-Jenkins approach to modeling and forecasting time series data is but one of a large family of quantitative forecasting methods which have been developed in the fields of operations research, statistics, and management science. Box-Jenkins models are also known as "ARIMA" models, the acronym standing for Autoregressive Integrated Moving Average. This terminology will be made clear in the following sections. Exponential smoothing, linear regression, Bayesian forecasting, and generalized adaptive filtering are some of the other techniques which are termed "extrapolative" forecasting (Makridakis et al., 1982).

Many of these methods have a common element; they utilize only the previous values of a series of numbers to forecast the future values of interest. Hence, they are referred to as univariate models, since the values from a single variable are used to predict the future values of the same variable. This is in contrast to multivariate models, where the variable of interest is also considered to depend on other variables.

Several introductory texts on forecasting are available (Pyndick and Rubinfeld, 1976; Montgomery and Johnson, 1976; Makridakis and Wheelwright, 1978). The Journal of Forecasting and Journal of Time Series Analysis should be consulted for the latest developments and applications. Makridakis et al. (1982) give summary descriptions of 24 important extrapolative forecasting methods in current use.

2.3 GENERAL REFERENCES FOR BOX-JENKINS MODELING

Newbold (1983) has provided an authoritative review of the current state of the art in Box-Jenkins modeling. Box and Jenkins (1968, 1974, 1976) remain the classic references; however, they assume a mathematical and statistical background which may be too advanced for those mainly interested in applications. A number of introductory texts present the Box-Jenkins method at a more applied level (Nelson, 1973; Pyndick and Rubinfeld, 1976; Montgomery and Johnson, 1976; Makridakis and Wheelwright, 1978; McCleary and Hay, Jr., 1980). Perhaps the best single reference in this category is Pankratz (1983), who provides fifteen detailed case studies illustrating the Box-Jenkins approach.

2.4 STOCHASTIC PROCESSES AND TIME SERIES

Here the term time series will refer to a set of N ordered observations or measurements

$$z_1, z_2, \cdots, z_t, \cdots, z_N$$

separated in time. By this definition, a time series is a discrete set of numbers. A discrete series may arise from instantaneous sampling of a continuous process (e.g., flow rate measurements) or by obtaining cumulative values over specified periods (e.g., total hourly flows). In the Box-Jenkins methodology, successive values of the series under study must be separated by equal time intervals. The discrete series is often a measure of some underlying continuous

process. If the observations are recorded appropriately, the discrete series will convey sufficient information about the continuous process. Motivation for discrete analyses arises from the difficulty of processing continuous data on digital computers.

It is important to make the distinction beween an observed time series and the stochastic process from which the series emanates. A stochastic process is represented by a set of N random variables, each of which corresponds to one value of the time series. This collection of random variables is governed by a joint probability density function. A time series should be visualized as one possible realization of the stochastic process. That is, if the process was monitored repeatedly, the random variables would take on different values, resulting in different time series, each being a member of an infinite ensemble of realizations. The basic tenet of the Box-Jenkins methodology is that any actual observed time series bears the signature of the underlying process, and hence may be used to identify an appropriate mathematical model for the process.

There is a fundamental difference between most time series and the familiar "random sample" used for standard statistical analyses. For a time series, the order of observations is important, and they generally cannot be considered independent. In fact, it is the form of dependence which is of interest. The correlation between successive values of the same series, i.e., autocorrelation, is the tool used to identify a model for the underlying stochastic generating process.

2.5 STOCHASTIC VS. DETERMINISTIC MODELS

As opposed to deterministic models, in which a time-dependent quantity evolves exactly in accordance with certain known physical/mathematical laws (e.g., Newton's laws of motion), Box-Jenkins models fall under the category of stochastic models, for which unknown intervening influences are viewed as prohibiting exact prediction of future outcomes. The key difference lies in the probabilistic description associated with stochastic models. A stochastic model can be used to calculate the probability that a future value will fall between certain upper and lower limits, in contrast to an exact prediction. While Box-Jenkins models are used to generate point forecasts (single numerical values), they also provide an estimated confidence interval for each forecast value.

2.6 UNIVARIATE BOX-JENKINS MODELS: THE BASIC IDEA

In the Box-Jenkins approach, the values z_t of an observed time series are considered to be the outputs of a black box (unobservable) process whose inputs a_t are called independent random shocks. This description is represented schematically in Figure 1.

The independent random shocks often cause confusion for the uninitiated, and as such need to be carefully defined. These inputs are regarded as disturbances of randomly varying magnitude, each independent of the preceding one, which enter the black box and are transformed and combined into observed output values. The inputs occur

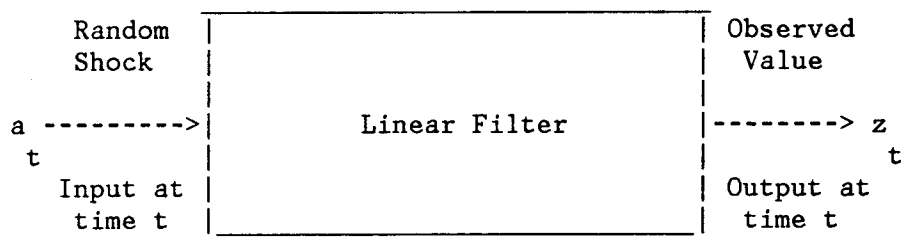


Figure 1: Time Series Viewed as Result of Random Shocks

at equally spaced intervals with the same time index t as the outputs. They do not represent any physical inputs which may be present in the system being studied; they are statistical constructs.

For statistical purposes, time series analysts assume that the random shocks are independent and normally distributed with mean zero and constant variance. A sequence possessing these properties will be referred to here as white noise, although white noise need not be normally distributed. Thus, the Box-Jenkins approach views a time series as the result of a transformation of a white noise process. The black box which transforms the noise into an observed series is called a linear filter. Checking the adequacy of the model used to describe the stochastic process consists of verifying that the model's residuals (observed minus model-calculated values) are white noise. This may be thought of as "running the filter backwards"--the result must be white noise.

The essence of univariate Box-Jenkins models will now be stated. It seems reasonable that the outputs (observed time series values) may depend on

1. the previous and current inputs. The most important factor influencing the current output z_t is the current input a_t. In addition, lingering effects from previous inputs a_{t-1}, a_{t-2}, ... may also play a role, but probably to a lesser extent.

2. the previous output values z_{t-1}, z_{t-2}, ... For example, a particularly large recent output may somehow affect the nature of the process. This change could evidence itself in the next output.

Specifically, the Box-Jenkins approach proposes a simple linear form for the above relationships:

$$z_t = \phi_1 z_{t-1} + \phi_2 z_{t-2} + \ldots + \phi_p z_{t-p} + a_t - \theta_1 a_{t-1} - \theta_2 a_{t-2} - \ldots - \theta_q a_{t-q} \tag{2.1}$$

which shows the current output consisting of a linear weighted sum of previous outputs and inputs (the negative signs on the θ parameters are chosen for a notational convenience to be introduced later).

Note that only "p" nonzero output terms and "q" nonzero input terms are included; this reflects the fact that only a finite number of recent inputs and outputs will have a statistically significant effect on the current output. All earlier influences need not be included in the model. Note also that the current input a_t is always assigned a weight of unity.

2.7 AUTOREGRESSIVE (AR) MODELS

As stated above, an observed value of a time series can be viewed as an output which may depend on previous outputs as well as inputs. When the value of the current output (observation) z_t depends solely on "p" prior outputs and the current input (random shock) a_t, the model

$$z_t = \phi_1 z_{t-1} + \phi_2 z_{t-2} + \ldots + \phi_p z_{t-p} + a_t \tag{2.2}$$

is called an Autoregressive model of order p. The common notation is AR(p). The name is appropriate, as the model involves regressing a variable on previous values of itself (cf. ordinary multiple linear regression), plus an error or random term.

2.8 MOVING AVERAGE (MA) MODELS

When the current output depends solely on the current input and "q" prior inputs, the model

$$z_t = a_t - \theta_1 a_{t-1} - \theta_2 a_{t-2} - \ldots - \theta_q a_{t-q} \tag{2.3}$$

is called a Moving Average model of order q. The notation is MA(q). This is a misnomer, since the model is not the familiar moving average consisting of the arithmetic mean of past observations. However, the term has become traditional.

The idea that an observation could be modeled as a linear weighted sum of random numbers is unfamiliar to most nonstatisticians. It is important to realize that a series composed of such linear sums of white noise elements is not itself white noise, but rather has a definite autocorrelation structure.

2.9 MIXED AUTOREGRESSIVE AND MOVING AVERAGE (ARMA) MODELS

A process which involves elements of both AR and MA processes is modeled by

$$z_t = \phi_1 z_{t-1} + \phi_2 z_{t-2} + \ldots + \phi_p z_{t-p} + a_t - \theta_1 a_{t-1} - \theta_2 a_{t-2} - \ldots - \theta_q a_{t-q} \tag{2.4}$$

and referred to as an Autoregressive Moving Average model of order (p,q), or ARMA(p,q).

2.10 ARIMA MODELS AND STATIONARY STOCHASTIC PROCESSES

The basic elements of univariate Box-Jenkins models have been presented. An additional question remains: What is the "I" (integrated) in ARIMA? It will be seen shortly that the "integrated" portion of Box-Jenkins modeling is concerned with transformations of the original raw time series data which may be necessary before the above AR, MA, or ARMA models may be applied. It should be noted that the ARIMA labels (AR, MA, ARMA) refer either to the stochastic process or its model; the proper model form is sought to describe the process under study.

The Box-Jenkins methodology requires that the time series to be analyzed be stationary. Strictly speaking, stationarity is a mathematical property of the collection of random variables which constitute a stochastic process, as defined in advanced probability theory. However, it is common practice to refer to the data themselves as being stationary or nonstationary. Rigorous discussions of stationarity may be found in the stochastic process literature; for the purposes of this thesis, the operational "definition" used by practitioners will be followed.

A time series is considered stationary if its sample mean and variance are not significantly different, in the statistical sense, for any major subsets of the series. If this is not the case, then the series mean and series variance lose their meaning and cannot be estimated using the familiar formulas. Note that a series can be nonstationary in the mean or the variance. However, a variance-stationary series is, by definition, also mean-stationary. Practically speaking, if a series displays shifts in level or increasing variability over time, it is nonstationary and must be transformed prior to analysis.

Simple transformations exist which will induce mean stationarity in many observed time series. These take the form of simple differences performed on the raw series values. A difference of order one means that each value of the series is subtracted from the next neighboring value:

$$w_t = z_t - z_{t-1} \tag{2.5}$$

This results in a new time series w_t, having one less observation than the original series. A difference of order two means that the order-one differenced series is differenced again

$$y_t = w_t - w_{t-1} = (z_t - z_{t-1}) - (z_{t-1} - z_{t-2}) = z_t - 2z_{t-1} + z_{t-2} \tag{2.6}$$

leaving a new series y_t with two fewer observations than the original series. This may be generalized to d^{th} order differencing, where "d" is the order of differencing required to achieve mean stationarity.

After modeling the differenced series with an appropriate ARMA model, to reclaim the the modeled values corresponding to the original undifferenced series it is necessary to reverse the differencing transformation and "integrate" (sum) "d" times. This is the reason for the "I" (integrated) in the acronym ARIMA.

To achieve stationarity in variance, it may be necessary to perform other types of transformations such as taking logarithms or square roots of the raw series values. For details, the introductory references provided may be consulted. The central point is that, prior to analysis, the series must be made stationary in both mean and variance by suitable transformations. Then the proper AR, MA, or ARMA model is sought for the transformed series.

A process which requires d^{th} order differencing is called an Integrated process of order d, or in notation, I(d). A model which

incorporates aspects of AR, MA, and I models is called an ARIMA(p,d,q) model of order (p,d,q). Note that any AR, MA, I, or combined process may be expressed in this notation by setting p, d, or q to zero separately or in combinations.

2.11 MULTIVARIATE ARIMA MODELS

Thus far, the discussion has been limited to univariate Box-Jenkins models. Of equal importance are the multivariate ARIMA models, for which the variable to be forecast depends not only on its own previous values, but also on current and previous values of related time series variables measured simultaneously. If two time series are involved, the model is bivariate and is referred to as a "transfer function" model by analogy to linear systems theory. The variable to be forecast (the "output") is considered to depend on its own previous values as well as the values of the related ("input") variable.

As might be expected, the modeling process for multivariate ARIMA approaches is considerably more difficult and less developed than the univariate methodology. This thesis will be confined to univariate applications to wastewater treatment data. At present, univariate models stand a much better chance of being applied for real-time forecasting in the wastewater treatment field.

2.12 THE ITERATIVE BOX-JENKINS MODELING STRATEGY

This section outlines the procedures that Box and Jenkins recommend for constructing a univariate ARIMA model from a given time series. The Box-Jenkins approach to model building is represented schematically in Figure 2. The model may then be used to forecast future values. Three different categories of forecasts are described, and a terminology for discriminating between the three types is introduced.

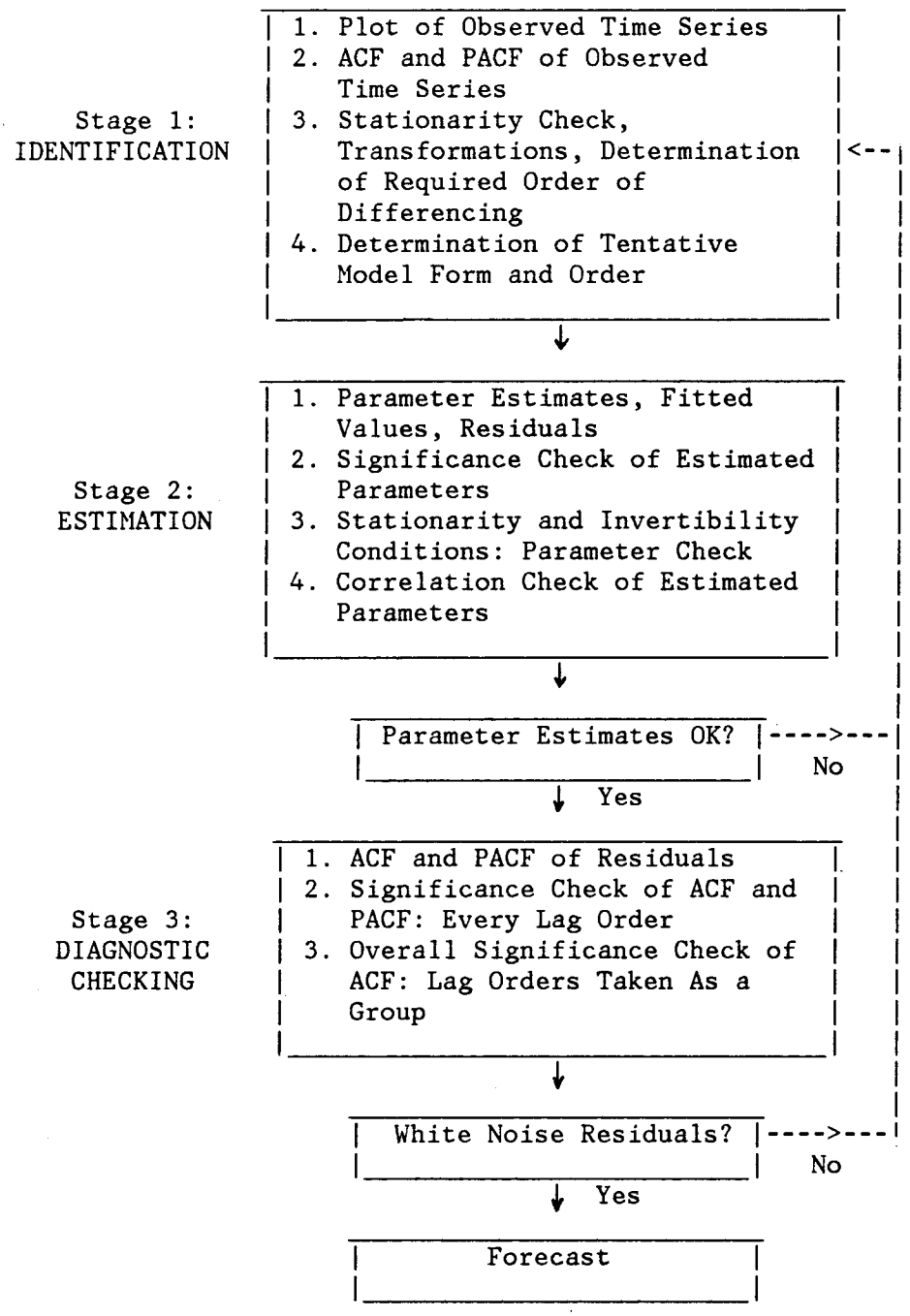


Figure 2: The Iterative Box-Jenkins Modeling Strategy

2.12.1 The Identification Stage

Identification is the stage at which a tentative model for the series is selected from the large family of candidate ARIMA(p,d,q) models. Clearly there are many possible combinations of the orders p, d, and q. Thus, the identification stage consists of specifying the AR, I, and MA orders (p,d,q). Fortunately, it has been found that, in practice, adequate models rarely have values of p, d, and q greater than two. That is, the autoregressive order, moving average order, or degree of differencing required to induce stationarity rarely exceed two. This empirical fact is also related to the "principle of parsimony" to be discussed shortly.

The basic tools for model identification are the graphs of the sample autocorrelation function (ACF) and sample partial autocorrelation function (PACF) obtained from the series. The ACF (sometimes called the correlogram) indicates the degree of correlation within the series for lags 1, 2, 3,... etc. In a similar fashion, the PACF indicates the degree of correlation at a given lag after accounting for the correlation from the intervening lags. Pankratz (1983) gives a lucid explanation of the PACF. The ACF and PACF are plotted as spikes occurring at each lag order. Examples are shown in Figure 3. Estimated 95% confidence intervals are also plotted at each lag order in order to check whether each lag's individual autocorrelation coefficient is significantly different from zero or not, indicating the absence or presence of correlation for that lag order. If a spike lies outside the confidence limit lines, the correlation at that lag is significant.

AUTOCORRELATIONS

LAG	CORRELATION	-1 9 8 7 6 5 4 3 2 1 0 1 2 3 4 5 6 7 8 9 1
0	1.00000	|**********|
1	0.34722	|********.
2	0.35003	|********.
3	0.27501	|*******.
4	0.16682	|****.
5	0.22072	|*****.
6	0.13557	|****.
7	0.11648	|**.
8	0.06858	|*.
9	0.03628	|*.
10	0.07236	|*.
11	0.02038	|.
12	0.08353	|**.
13	0.23953	|******.
14	0.23950	|******.
15	0.16561	|****.
16	0.18672	|*****.
17	0.15348	|****.
18	0.09650	|**.
19	0.11413	|**.
20	0.11816	|**.

'.' MARKS TWO STANDARD ERRORS

PARTIAL AUTOCORRELATIONS

LAG	CORRELATION	-1 9 8 7 6 5 4 3 2 1 0 1 2 3 4 5 6 7 8 9 1
1	0.34722	|********.
2	0.26092	|******.
3	0.11541	|**.
4	-0.02041	|.
5	0.10232	|**.
6	0.00089	|.
7	-0.00517	|.
8	-0.03737	|.
9	-0.02269	|.
10	0.04273	|*.
11	-0.01998	|.
12	0.06499	|*.
13	0.24715	|******.
14	0.15199	|****.
15	-0.05751	|.
16	0.00865	|.
17	0.01608	|.
18	-0.08279	|.
19	-0.03234	|.
20	0.04862	|*.

Figure 3: Example of Computer-Plotted Sample ACF and PACF

To determine the order of differencing d, the time series must be checked for nonstationarity. If nonstationarity is indicated, differencing or other transformations must be performed prior to further analysis. There are basically two methods currently in use by practitioners (Ali and Thalheimer, 1983). One is to simply inspect the plotted time series for shifts in level or increasing variability. The other involves examination of the ACF. If the ACF spikes fail to die out rapidly (i.e., remain statistically significant at high lag orders), differencing may be required. The required order of differencing determines "d". Ali and Thalheimer (1983) recently proposed more formal statistical tests for determining nonstationarity.

To determine the AR and MA orders p and q, inspection of the ACF and PACF of the series (or differenced series, if called for) is performed. It can be shown that, in theory, the number of successive ACF spikes at lags greater than zero equals the order of the moving average component, q. In a similar fashion, the number of significant PACF spikes at lag orders greater than zero indicates the order of the autoregressive component, p. In addition, other patterns in the ACF and PACF help validate these tentative indications.

The identification techniques just described are based on the mathematical fact that the theoretical ACF and PACF for a particular ARIMA(p,d,q) process are unique. That is, for any specified set of integer values (p,d,q), the theoretical ACF and PACF from such a stochastic process provide unique "fingerprints" for identifying the

process. However, in practice the idealized procedure of counting significant spikes is confounded by sampling error in the estimated ACF and PACF, and proper identification can become quite difficult (some analysts call them "smudged fingerprints"). It should be borne in mind that the time series under study is only one realization from which one attempts to estimate the ACF and PACF. The appearances of the ACF and PACF also depend on the signs of the AR and MA parameters ϕ and θ. It is useful to compare the the sample ACF and PACF to plots of the theoretical ACF's and PACF's from various AR, MA, and ARMA processes of orders (p,q) less than or equal to two. Pankratz (1983) contains a good collection of plotted ACF's and PACF's for comparison.

Because of the described identification procedure, proponents of Box-Jenkins models proclaim that the methodology is superior to other modeling techniques because it "lets the data speak for themselves," rather than imposing a specific model form onto the data *a priori*. However, for the sake of presenting a balanced discussion, it should be noted that Klemes (1982) has provided arguments (his own as well as those of noted statisticians) against this philosophy. It is feared that the analyst will not bother trying to understand the physical basis for the data. It is also felt that the assumption of system linearity implicit in ARIMA models is itself a type of *a priori* model specification.

2.12.2 The Estimation Stage

After a tentative model has been identified, the AR and/or MA parameters ϕ and θ are estimated from the time series data using an efficient nonlinear least-squares algorithm. The residuals, i.e., the differences between the observed time series values and the model-calculated or "fitted" values, are also obtained at this stage. The least-squares estimates of ϕ and θ are those values which minimize the sum of the squared residuals.

It is extremely important to understand how the fitted values are obtained, and what the residuals represent in the context of ARIMA models. The model-calculated values are found by inserting initial estimates $\bar{\phi}$ and $\bar{\theta}$ for the AR and MA parameters, setting the current random shock term a_t to its expected value of zero, and using the resulting estimated model together with the observed data to sequentially generate a series of fitted values. This procedure is repeated, adjusting the parameter estimates at every iteration, until the least-squares fit is obtained. It then follows that the differences between the observed and fitted values, the residuals, will be estimates of the random shocks $\bar{a}_t$.

This is how statistical estimates for the unobservable random shocks, which constitute the conceptual driving force in ARIMA models, are obtained. Notice that in the context of ARIMA models, the residuals can never all be zero (corresponding to a "perfect fit" to

the data) because they are estimates of shocks which have zero mean but randomly varying magnitudes. Instead, the residuals from a properly identified ARIMA model indicate the magnitude of the "noise" present in the observations which is not accounted for by the model. Pankratz (1983) presents a detailed example of how fitted values and residuals are calculated.

In addition to the above considerations, at the estimation stage there are certain conditions which the parameter estimates must meet in order for the proposed model to be deemed acceptable. First, all AR and MA parameter estimates must be statistically significant (i.e., significantly different from zero). This is verified by simple t-tests for the estimates. If any parameter is not significantly different from zero, it should be dropped from the tentative model. Second, the parameters must satisfy certain inequality relations known as the stationarity and invertibility conditions.

A detailed discussion of the stationarity and invertibility conditions is not central to this presentation. The stationarity conditions apply only to the AR parameters, whereas the invertibility conditions pertain to the MA parameters. These inequalities reflect the fact that the influence of time series and random shock values in the past must diminish with time, in accordance with common sense.

Finally, it is necessary to ensure that the parameter estimates are not too highly correlated. Pankratz (1983) suggests 0.9 correlation as a rule-of-thumb cutoff level. Checking is performed by inspecting a correlation matrix, which shows the correlations between all pairs

of AR and/or MA parameter estimates. High correlation often indicates that the estimates are of poor quality. A model based on such estimates will not be robust to changes in the pattern of the data, and hence will perform poorly for forecasting.

If the tentative model has significant parameters, whose values lie within the bounds of stationarity and invertibility and are not highly correlated, then the analyst may proceed to the last stage: diagnostic checking. If not, the analyst must return to the identification stage and formulate an alternate model based on the information gained at the estimation stage.

2.12.3 The Diagnostic Checking Stage

During the estimation stage, the residuals (which constitute estimates of the random shocks) are obtained. Now, the basic idea of Box-Jenkins modeling is to explore the autocorrelation structure of the data and use that information to identify an adequate model for the series of interest. When an adequate model has been fitted to the data, it should have "filtered out" (accounted for) the autocorrelation structure, leaving uncorrelated residuals. Hence, the diagnostic checking stage consists of verifying that the residuals obtained at the estimation stage are white noise.

By definition, the ACF of a white noise series will show no significant autocorrelation at any lag order. Thus, an ACF plot of the residuals is inspected to verify that they have no remaining autocorrelation pattern. The autocorrelation at each lag may be

tested for significance using a t-test, which is equivalent to inspecting the ACF for spikes which lie above the confidence limit lines. In practice, there may be a few ACF spikes which are close to significance; one might expect approximately 5% to be statistically non-zero by chance alone for a 95% confidence limit test. It is generally more important to consider at which lag-orders such "borderline" cases occur, because they may indicate an improperly identified model.

In addition to testing the significance of the individual ACF spikes at each lag, the residual autocorrelations are also tested as a set using a chi-square statistic computed from their values. If this statistic is too large (because some autocorrelations are still significant), then the hypothesis of white noise residuals must be rejected.

It may also be useful to check whether the residuals can be accepted as normally distributed with mean zero and constant variance. The extent to which these assumptions are violated indicates how much faith one can put in the proposed model and its statistically derived results.

If the ACF of the residuals displays a pattern of autocorrelation not accounted for by the tentative model, then it is necessary to return to the identification stage and reformulate the model. This may be done simply by reexamining the original series ACF and PACF for another interpretation, or the model reformulation may be based on the pattern remaining in the residual ACF. Often the significant spikes

remaining in the residual ACF or PACF will provide clues as to how the initial model should be modified.

In theory, the three-stage iterative process of model identification, estimation, and diagnostic checking is repeated until an adequate model yielding white noise residuals is arrived at. In practice, however, the analyst may have to settle for the model which has the "whitest" residuals, i.e., the model which least seriously violates any of the Box-Jenkins screening criteria for model adequacy outlined in this section and summarized in Figure 2.

2.12.4 Forecasting

Once an adequate model has been identified and its parameters estimated, it may be used for forecasting future values of the series. To see how this is done, consider the general ARMA(p,q) model

$$z_t = \phi_1 z_{t-1} + \phi_2 z_{t-2} + \ldots + \phi_p z_{t-p} + a_t - \theta_1 a_{t-1} - \theta_2 a_{t-2} - \ldots - \theta_q a_{t-q} \tag{2.7}$$

and replace the time index t by the time one step ahead, t+1:

$$z_{t+1} = \phi_1 z_t + \phi_2 z_{t-1} + \ldots + \phi_p z_{t-p+1} + a_{t+1} - \theta_1 a_t - \theta_2 a_{t-1} - \ldots - \theta_q a_{t-q+1} \tag{2.8}$$

This equation states that, if the current time is t, then the next value z_{t+1} will be made up of the "p" previous observations and "q" previous random shocks, plus the shock at time t+1. Now, at the estimation stage, estimates $\bar{\phi}$ and $\bar{\theta}$ are obtained for the parameters,

so their values are determined. In addition, the fitted residuals $\bar{a}_t$ are obtained, and hence estimates for the shock terms are also available, except for the next shock a_{t+1} which has not yet "occurred." This is simply set to its expected value of zero. The previous values of the time series have all been observed, so they are known as well. Substituting in the estimates and known quantities and setting the next shock to zero yields the forecast

$$\bar{z}_{t+1} = \bar{\phi}_1 z_t + \bar{\phi}_2 z_{t-1} + \ldots + \bar{\phi}_p z_{t-p+1} - \bar{\theta}_1 \bar{a}_t - \bar{\theta}_2 \bar{a}_{t-1} - \ldots - \bar{\theta}_q \bar{a}_{t-q+1} \tag{2.9}$$

for the next value in the time series.

The expression obtained by replacing the parameters with their estimated values and setting the future shock term to zero is called the forecast function associated with the model. It may be applied to forecast future values in several ways. The time t beyond which forecasting is to be performed is called the forecast origin. The following subsections explain three different types of forecasting which may be undertaken and the important differences between them.

2.12.4.1 One-Step-Ahead Within-Sample Forecasts

The first type of forecasts to be discussed are not really forecasts at all. It is common practice for the fitted or model-calculated values obtained at the estimation stage to be called the "one-step-ahead forecasts" by some Box-Jenkins analysts. The reason

for this becomes clear if one recalls that the fitted values are calculated by inserting initial estimates for the model parameters, setting the random shock term to zero, and using the resulting model together with the data values to sequentially generate a new series corresponding to the observations. This process is repeated, adjusting the parameters at each iteration, until the least-squares fit is found. In light of the previous discussion of the forecast function, it should be apparent that, upon convergence, the fitted values obtained in this manner represent a series of one-step-ahead forecasts using a model with parameters estimated from the entire time series.

These forecasts might be called "forecasts of hindsight," because the parameter values used to generate the forecasts are obtained from a least-squares fitting of the entire time series. Such "forecasts" cannot be determined until all the observations in the series have occurred. They are not true forecasts because they do not extend past the last observation in the series. This type of "forecast" will be called a one-step-ahead within-sample forecast. The one-step-ahead within-sample forecasts constitute a check on model "calibration," since the data used to estimate the parameters are also used to generate the model-calculated values. These forecasts do not provide model "verification," however.

2.12.4.2 Multi-Step-Ahead Beyond-Sample Forecasts

The second type of forecasts that may be produced using the forecast function are forecasts that extend beyond the last observation in the series, and for which it is assumed that future observations have not occurred (become available as new information). If the current time is t, and the forecast function containing parameters and residuals estimated from the observations up to and including time t is

$$\bar{z}_{t+1} = \bar{\phi}_1 z_t + \bar{\phi}_2 z_{t-1} + \ldots + \bar{\phi}_p z_{t-p+1} - \bar{\theta}_1 \bar{a}_t - \bar{\theta}_2 \bar{a}_{t-1} - \ldots - \bar{\theta}_q \bar{a}_{t-q+1} \tag{2.10}$$

then it is possible to make a one-step-ahead prediction for the next observation z_{t+1}, because all quantities in the forecast function have been observed or estimated. However, if it was desired to forecast two steps ahead, it would be necessary to have a value for the observation z_{t+1}, and also an estimate $\bar{a}_{t+1}$ for the shock a_{t+1} to put in the forecast function

$$\bar{z}_{t+2} = \bar{\phi}_1 z_{t+1} + \bar{\phi}_2 z_t + \ldots + \bar{\phi}_p z_{t-p+2} - \bar{\theta}_1 \bar{a}_{t+1} - \bar{\theta}_2 \bar{a}_t - \ldots - \bar{\theta}_q \bar{a}_{t-q+2} \tag{2.11}$$

In practice, what is done is to use the one-step-ahead forecast $\bar{z}_{t+1}$ as the best estimate for the unknown z_{t+1}, and to set the unknown random shock a_{t+1} to its expected value of zero:

$$\bar{z}_{t+2} = \bar{\phi}_1 \bar{z}_{t+1} + \bar{\phi}_2 z_t + \ldots + \bar{\phi}_p z_{t-p+2} - \bar{\theta}_2 \bar{a}_t - \ldots - \bar{\theta}_q \bar{a}_{t-q+2} \quad (2.12)$$

By repeating this process recursively, it is possible to generate forecasts as far beyond the last observation in the collected series as desired. Note that the farther out one forecasts, the more random shock terms will have to be set to zero. Thus, if the forecasts extend far enough, the forecast function becomes purely AR in nature and each new forecast is based solely on previous forecasts, and not on any previous observations. Pankratz (1983) has called this type of forecasting "bootstrap" forecasting, because it constitutes "pulling oneself forward by one's bootstraps," with no new information being used to produce the forecasts. Clearly, the forecast error would be expected to increase rapidly a few steps beyond the last observation, and this is indeed what occurs. This type of forecast is, however, a true forecast and will be called a multi-step-ahead beyond-sample forecast. The number of steps ahead to be forecast is known as the forecast lead time.

2.12.4.3 One-Step-Ahead Beyond-Sample Forecasts

In light of the drawbacks associated with multi-step-ahead beyond-sample forecasts as just described, it is preferable to perform a similar type of forecasting which incorporates the observations occurring after the original sample used for identification and estimation.

Suppose that the current time is t, so that the most recent observed value is z_t. A one-step ahead forecast $\bar{z}_{t+1}$ is then made for the next value. The analyst then waits until time t+1 and observes the true value z_{t+1}, which will differ somewhat from the forecast. The amount of difference, or error, is an estimate of the random shock a_{t+1}:

$$\bar{a}_{t+1} = z_{t+1} - \bar{z}_{t+1} \tag{2.13}$$

Given the new observation z_{t+1} and the estimated shock $\bar{a}_{t+1}$, it becomes possible to make another one-step-ahead forecast $\bar{z}_{t+2}$ for the value at time t+2. This forecast will depend on observations and estimated shocks, and should therefore be an improvement over a multi-step-ahead forecast which does not utilize new available information. The process may be repeated indefinitely.

This type of true forecasting, with continual updating of the observations and random shocks (forecast errors), will be called one-step-ahead beyond-sample forecasting. Within this classification, there are actually two types of forecasting which may be performed, depending on whether the parameters in the forecast function are updated as well. It is possible to reestimate the parameters after every new data point becomes available, but this is usually

unnecessary because the estimates generally do not change significantly over only one period. Pankratz (1983) suggests that a change in parameter value greater than 0.1 be considered significant.

It is often sufficient to forecast a number of steps using fixed parameter estimates obtained at the most recent estimation. After a significant number of periods (perhaps ten or twenty) have passed, it will be necessary to reestimate the parameters using the new, most recent observations. Some of the data at the beginning of the data set may be dropped off after new observations are included. The updated model may then be used to forecast again, using new fixed parameter estimates. The decision to utilize continual or intermittent parameter updating involves making a tradeoff between computational time and forecasting accuracy. If forecasting takes place over an extended period, it may even become necessary to completely reidentify the model, which may change its functional form (orders p, d, and q) over time as the data pattern shifts.

Clearly, such updated forecasting is ideally suited for real-time control applications. However, it is equally clear that the amount of work required for continual updating and periodic checks on the model identification is much greater than for multi-step-ahead (non-updated) forecasting.

2.12.4.4 Summary of Forecast Classifications

In summary, three types of forecasts have been discussed:

1. One-step-ahead within-sample: These are the values obtained from "fitting" a model to past observations. They are not forecasts of future observations.
2. Multi-step-ahead beyond-sample: These are true forecasts which can be generated without waiting for new observations occurring after the initial sample. Each successive forecast is based on preceding forecasts, a procedure referred to as bootstrapping. The forecast error increases rapidly for longer lead times.
3. One-step-ahead beyond-sample: These are also true forecasts, generated one at a time as each new observation becomes available. They do not depend on previous forecasts; instead they are based on the most recent observations and forecast errors (shock estimates). The lead time for each forecast is only one interval, thereby improving forecasting accuracy.

2.13 SEASONALITY AND SEASONAL ARIMA MODELS

Thus far, the discussion has been limited to the so-called nonseasonal ARIMA(p,d,q) models, which apply to time series which contain no "seasonal" component. However, many time series display periodic behavior. The term "seasonal" has become traditional because time series models are frequently used to analyze monthly or quarterly cyclic data. The period of the cycle is denoted by the letter "s". Depending on the data being modeled, a "season" can be a day, a week, a year, or a century. For example, in wastewater treatment plants there are often diurnal cycles present in the influent flow and

composition. Thus, for hourly measurements the data display a periodicity of s = 24 hours.

It can be shown that the entire procedure for identification, estimation, and diagnostic checking of seasonal ARIMA models is identical to that for nonseasonal models, except that attention is focused on the autocorrelation patterns in the ACF and PACF occurring at the seasonal lags. The seasonal lags are those separated by the period of interest "s"; i.e., s, 2s, 3s,... etc..

First, there may be seasonal nonstationarity. This is indicated by ACF spikes at the seasonal lags which die out slowly. It may be necessary to perform seasonal differencing, which consists of subtracting observations "s" intervals apart. The order of seasonal differencing required is given the letter "D."

Second, the ACF and PACF may indicate that the series is autoregressive at the seasonal lags. The model for a seasonal AR process of order P is written as

$$z_t = \Phi_1 z_{t-s} + \Phi_2 z_{t-2s} + \ldots + \Phi_P z_{t-Ps} + a_t \qquad (2.14)$$

which shows the current observation represented by a linear combination of "P" previous observations occurring every period, plus the usual random shock.

Similarly, the ACF and PACF may suggest a seasonal moving average component. The model for a seasonal MA process of order Q is written as

$$z_t = a_t - \Theta_1 a_{t-s} - \Theta_2 a_{t-2s} + \ldots + \Theta_Q a_{t-Qs} \qquad (2.15)$$

which shows the current observation as a linear sum of "Q" previous periodic shocks, plus the current shock.

In general, a time series may contain both seasonal and nonseasonal components. A model which has been found useful for representing such series is the mixed seasonal-nonseasonal model denoted by ARIMA(p,d,q)(P,D,Q)$_s$. The lower-case letters in parentheses refer to the non-seasonal orders, and the upper-case letters refer to the seasonal orders. The subscript for the period s will be omitted; its value will be specified when describing a particular model. The identification of ARIMA(p,d,q)(P,D,Q) models is more complicated than identification of the simpler ARIMA(p,d,q) models because the superimposed seasonal and nonseasonal patterns complicate interpretation of the ACF and PACF.

When written out explicitly, the general ARIMA(p,d,q)(P,D,Q) model will be a complicated linear combination of previous time series values and random shock values, occurring at both seasonal and nonseasonal lag orders. It is still possible to write a linear forecast function based on this explicit form. Just as for nonseasonal models, this is determined by substituting in estimates for the parameters and random shocks (residuals) obtained by fitting the seasonal model to the observed series at the estimation stage, and setting the current random shock term to zero. Any of the three types of forecasting discussed previously may then be carried out.

2.14 ARIMA ALGEBRA AND NOTATION

As seen in the preceding sections, it can be cumbersome to explicitly write out models of higher orders. Box and Jenkins analysts use a shorthand notation for succinctly representing their models. While simplifying the writing of models, the notation may appear strange or difficult when first encountered. This section provides a brief explanation.

Consider the general AR model of order p, or AR(p):

$$z_t = \phi_1 z_{t-1} + \phi_2 z_{t-2} + \ldots + \phi_p z_{t-p} + a_t \tag{2.16}$$

Bring all the z-terms to the left hand side, yielding

$$z_t - \phi_1 z_{t-1} - \phi_2 z_{t-2} - \ldots - \phi_p z_{t-p} = a_t \tag{2.17}$$

Define an operator B called the backshift operator such that

$$B^k z_t = z_{t-k} \tag{2.18}$$

i.e., B operating "k" times shifts the time index of the variable back "k" periods. Then the AR(p) model may be written compactly as

$$(1 - \phi_1 B - \phi_2 B^2 - \ldots - \phi_p B^p) z_t = a_t \tag{2.19}$$

Similarly the general MA model of order q, MA(q)

$$z_t = a_t - \theta_1 a_{t-1} - \theta_2 a_{t-2} - \ldots - \theta_q a_{t-q} + a_t$$

(2.20)

may be written as

$$z_t = (1 - \theta_1 B - \theta_2 B^2 - \ldots - \theta_q B^q) a_t$$

(2.21)

where

$$B^k a_t = a_{t-k}$$

(2.22)

The process of differencing may also be expressed in the backshift notation. Define the differencing operator (1-B) such that

$$(1-B) z_t = z_t - B z_t = z_t - z_{t-1}$$

(2.23)

Similarly, for seasonal differencing of period "s"

$$(1-B^s) z_t = z_t - B^s z_t = z_t - z_{t-s}$$

(2.24)

Combining all the notation, for differencing of order "d", the general ARIMA(p,d,q) model is then

$$(1 - \phi_1 B - \phi_2 B^2 - \ldots - \phi_p B^p)(1-B)^d z_t = (1 - \theta_1 B - \theta_2 B^2 - \ldots - \theta_q B^q) a_t$$

(2.25)

Corresponding expressions may be written for the seasonal ARIMA(p,d,q)(P,D,Q) models.

2.15 THE PRINCIPLE OF PARSIMONY

The principle of parsimony is a statistical concept credited to J. W. Tukey (1961), which states that the best model for a given set of data is the very simplest model which can account for the observed properties of the data. In the context of ARIMA models, a parsimonious model is a model which contains the minimum number of parameters necessary to yield white-noise residuals. If two candidate models are comparable in terms of fit to the data and whiteness of residuals, then the analyst will always prefer the model having lower parameter-order. As stated earlier, it has been found that adequate models for many observed data sets require only nonseasonal and seasonal orders p, d, q, P, D, and Q less than or equal to about two.

Chapter III

LITERATURE REVIEW

"A prudent man foresees the difficulties ahead and prepares for them; the simpleton goes blindly on and suffers the consequences."

Proverbs 22:3
Tynedale Translation

3.1 BACKGROUND FOR THE PRESENT RESEARCH

This research was motivated by the work of Stenstrom (1976) and Stenstrom and Andrews (1979). Stenstrom and Andrews used a Fourier series model for hourly flow prediction in conjunction with a wastewater treatment plant simulation model to test various plant control strategies. A finite Fourier series model was fitted to hourly influent flows recorded in a field survey (Anderson, 1973). A random noise component was then superimposed on the Fourier series; this combination was used as the simulated input to the treatment plant model. Tanthapanichakoon and Himmelblau (1980) have subsequently utilized this form of stochastic input sequence also. The smooth, time-dependent function expressed by the pure Fourier series model was evaluated at discrete intervals into the "future" to simulate flow prediction in advance of occurrence. The variance of

the random noise sequence was set equal to the variance of the residuals obtained from fitting the Fourier series model to the observed data. In this manner, the error that would occur in forecasting was simulated.

Maintaining constant organism growth rate in the activated sludge treatment process is known to improve performance, but this can be quite difficult to achieve in light of the dynamically varying nature of the influent. A useful measure of growth rate is the specific oxygen uptake rate (SCOUR). Stenstrom and Andrews (1979) found that utilizing the additional information of predicted flow for 1-, 2-, and 3-hour lead times decreased the variability of SCOUR by as much as 48% over a corresponding control strategy without flow prediction. They recommended that future research be directed at improved forecasting, and indicated that Box-Jenkins models could be utilized. The present thesis was undertaken in direct response to these comments.

3.2 SCOPE OF THE REVIEW

This review will primarily be concerned with previous applications of univariate ARIMA models to wastewater treatment flow and composition data. A number of applications of multivariate ARIMA models have also been reported (Shih, 1976; Berthouex et al., 1976, 1978a, 1979, 1983; Berthouex and Hunter, 1982; Labadie et al., 1976; Murphy et al., 1977; Adeyemi et al., 1979; Filion et al., 1979; Debelak and Sims, 1981), but their level of complexity relative to univariate models appears to have hindered their application to real-

time forecasting. Other closely related references which provide perspective on time series applications in the water quality field include Fuller and Tsokos (1971), Young and Whitehead (1975), Beck (1976, 1977), Olsson (1976), Berthouex et al. (1978b), D'Astous and Hipel, 1979; Hansen et al., 1980; and McLeod et al., 1983. Far more extensive use of ARIMA models has been made in the field of stochastic hydrology (see for example Carlson et al., 1970; McMichael and Hunter, 1972; McKerchar and Delleur, 1974; McLeod et al., 1977; Kottegoda, 1980; Salas et al., 1980).

3.3 PREVIOUS APPLICATIONS OF UNIVARIATE ARIMA MODELS

The earliest published application of Box-Jenkins models to wastewater treatment data appears to be the work of McMichael and Vigani (1972). In a discussion paper extending the work of Wallace and Zollman (1971), McMichael and Vigani introduced Box-Jenkins models for representing hourly grab samples of chemical oxygen demand (COD) from a municipal combined sewer system. No mention was made of a specific practical use for the models; rather, the paper was intended to introduce the Box-Jenkins approach to water quality data analysts. For six data sets, the authors showed that low parameter-order, nonseasonal ARIMA(p,d,q) models could explain from 25% to 50% (R-squared) of the data variance. Superimposed plots were presented showing the observed COD data and the fitted model values, described as "one interval ahead forecasts." These results were presented in a section entitled Forecasting. As described in the previous chapter,

one-step-ahead within-sample forecasts are not true post-observation forecasts, and do not necessarily reflect the future forecasting performance of a proposed model.

In the same year, Goel and LaGrega (1972) presented a paper in which they analyzed hourly averages of the influent flow rate to a conventional activated sludge wastewater treatment plant. They developed a seasonal ARIMA(2,1,0)(0,1,1) model (s = 24) for forecasting hourly average flow values 24 hours in advance of occurrence. An example was presented in which the proposed model was used to perform multi-step-ahead beyond-sample forecasting 48 hours in advance (i.e., for lead times = 1, 2, ... , 48). The next 48 hourly flows were then measured and found to be in close agreement with the predicted values (Goel, 1984; LaGrega, 1984). This modeling was carried out as one component of a broader study of the beneficial effects of flow equalization on treatment plant unit processes (LaGrega and Keenan, 1974). It appears to be the only reported study where ARIMA forecasts were actually used to influence the operation of a wastewater treatment system.

Huck and Farquhar (1974) modeled hourly chloride and dissolved oxygen (DO) levels in the St. Clair River, Canada. Although their study did not involve wastewater treatment systems per se, it is included here because the authors presented many important concepts pertinent to water quality forecasting not found elsewhere. The authors described in detail the actual steps taken in the identification, estimation, diagnostic checking, and forecasting

stages. These are features frequently lacking in other reported results.

Huck and Farquhar (1974) set forth important contrasts between the "time domain" approach (of which Box-Jenkins modeling is an example) and the "frequency domain" (spectral analysis) approach to time series analysis of water quality data. Many earlier studies utilized the variance spectrum and Fourier series frequency component decomposition method (Gunnerson, 1966; Thomann, 1967; Wastler and Walter, 1968). Thomann (1970) applied spectral analysis to characterize the variability of daily effluent biochemical oxygen demand (BOD) from eight wastewater treatment plants. As Huck and Farquhar pointed out, these approaches result in complex spectral models, whereas "a single parameter in the Box-Jenkins model could replace the contribution of a score of amplitude and phase parameters of a response Fourier series model." Furthermore, they noted that the frequency domain approach requires that the data be made stationary, while ARIMA models accomodate typical forms of nonstationarity directly.

Upon applying Box-Jenkins models to four river quality data sets, Huck and Farquhar (1974) found that low parameter-order, nonseasonal ARIMA(p,d,q) models could account for 60% to 70% of the variance in the series under study. However, the authors made the important observation that the remaining (unexplained) variance, which for a properly identified ARIMA model measures the the magnitude of the purely random disturbances in the system, could pose significant problems in applications.

Using the models developed, Huck and Farquhar (1974) presented "forecasts" for lead times of 1 hour and 24 hours, and compared the results to observed values. However, the authors' figures show that the time origin of the forecasts was chosen within the period during which observations were taken. Hence the results are not true post-observation forecasts, since the model parameters were estimated using data beyond the forecast origin. In their literature review, Huck and Farquhar also asserted that McMichael and Vigani (1972) "fitted models to <Wallace and Zollman's> data and then employed the models for forecasting." As discussed previously, McMichael and Vigani only presented one-step-ahead within-sample forecasts, which are simply fitted values.

A discussion of the forecasts presented by Huck and Farquhar (1974) was provided by Litwin and Joeres (1975), who admonished Huck and Farquhar for drawing sweeping conclusions about the forecasting potential of Box-Jenkins models from their limited results. Litwin and Joeres made the important point that Box-Jenkins models are best suited for forecasting with continual updating (one-step-ahead beyond-sample) as new observations become available. In their response, Huck and Farquhar (1975) conceded that updated forecasting could have readily been performed.

Goel and LaGrega (1972), Huck and Farquhar (1974), and Litwin and Joeres (1975) all indicated the potential application of Box-Jenkins forecasts to real-time control of water quality systems. Berthouex et al. (1975) took the first step by modeling hourly influent BOD to an

activated sludge plant, with the anticipated goal of developing a bivariate ARIMA model of effluent behavior based on previous effluent observations, influent observations, and stochastic disturbances. Berthouex's research team performed a massive data collection effort, obtaining hourly grab samples (with two replicates) and flow measurements at a municipal activated sludge plant over a two-week period (Shih et al., 1974).

The starting point for the transfer function approach to forecasting and control is the identification of a univariate ARIMA model for the input time series (Box and Jenkins, 1976). Following a brief outline of the steps performed in model identification, estimation, and diagnostic checking, Berthouex et al. (1975) proposed that a simple, nonseasonal AR(1) model adequately represented the influent BOD data and could be used for forecasting or "many engineering purposes." Evidence for this conclusion was presented in a superimposed plot of the 332 actual data and corresponding "forecasts" from the AR(1) model (Berthouex et al., 1975, Figure 1, p. 129). Elsewhere in the paper, however, the same figure was described as "the fitted and observed values." This is the more correct description, since the forecasts shown were simply one-step-ahead within-sample forecasts. The AR(1) model accounted for 65% of the original series variance.

In a discussion paper reviewing the work of Berthouex et al. (1975), Adams (1975) criticized their lack of distinction between fitting and true forecasting, and pointed out that a plot of one-step-

ahead within-sample forecasts from an AR(1) model will of course closely resemble the plot of actual data. However, it should be noted that Berthouex et al. (1975) included a proviso stating that the AR(1) model identified might not be the best-fitting or most refined model possible. They presented the (slightly lower) residual variance from a second order AR(2) model as an example of alternative, higher parameter-order models that could be explored.

In a comment with direct bearing on this thesis, Adams (1975) further related that he had "encountered some instances in the analysis of time-dependent treatment plant performance data where the Box and Jenkins methods were of marginal value for these purposes." Adams also noted with irony that, "If BOD forecasts are made on a 1-hour lead time, it may prove most difficult to update the forecast function with 5-day BOD measurements of the previous hour's input provided for the next hour." Adams expressed doubt about the operational capabilities of Box-Jenkins forecasting for real-time control of treatment plants if the time interval necessary for process control is on the order of an hour. This statement was based on a presumed "remote likelihood" of being able to obtain hourly measurements for the "many variables" affecting treatment plant performance. However, simple univariate and bivariate (transfer function) ARIMA models require measuring only one or two variables simultaneously.

Berthouex et al. (1975) did not report any seasonal ARIMA(p,d,q)(P,D,Q) models for their BOD series, although mention was

made of the "diurnal and seasonal trends" present in wastewater flows. In a subsequent discussion paper, Shahane (1975) stated that the sample autocorrelation function (ACF) for the BOD data indicated periodicity, and questioned the simple autoregressive identification arrived at by Berthouex et al.

The paper by Berthouex et al. (1975) was based on a portion of a Ph.D. dissertation by Shih (1976), which incorporated and resolved the criticisms by Adams (1975) and Shahane (1975). Shih developed transfer function models relating effluent BOD to influent BOD and flow rate using data collected in field surveys at three activated sludge plants. In order to determine the (bivariate) transfer function models, Shih modeled the influent series using univariate ARIMA models. This was necessary because ARIMA models for the input series were used for "prewhitening" (filtering) the output series as part of the transfer function model construction procedure (Box and Jenkins, 1976).

Shih (1976) modeled twelve different input time series consisting of hourly or bi-hourly influent BOD concentrations, flow rates, and BOD mass loading rates. He obtained values of R-squared ranging from 42% to 82%. It is interesting to note that for all twelve series, five were represented by AR(1) models, and the remaining seven were represented by the same ARIMA(1,0,0)(0,1,1) model form (with s = 24 or s = 12). This is surprising, since the data sets involve different variables, different times of observation, and different treatment plants.

For the influent BOD series previously reported on by Berthouex et al. (1975), Shih compared 25 different seasonal and nonseasonal ARIMA models of varying complexity. He concluded that a seasonal ARIMA(1,0,0)(0,1,1) model (s = 24) best satisfied all the diagnostic checks for model adequacy. However, he chose to report and utilize the simpler, nonseasonal AR(1) model for prewhitening because it had a higher R-squared, even though it yielded correlated residuals. This appears to clear up the the issue of improper identification raised by Shahane (1975), but raises new questions about the use of models which are not developed in full accordance with the Box-Jenkins methodology. This could be viewed as a retreat to ad hoc models in place of models constructed according to a structured, rational methodology. Shih stated that the AR(1) model was adequate for prewhitening, but that the the seasonal model would be necessary if forecasting was the modeling objective. Shih did not perform any forecasting. Apparently in response to the comments by Adams (1975), Shih is to be commended for describing his model-calculated results as fitted values, avoiding the term "one-step-ahead forecasts" entirely. Shih's work formed the basis for a series of papers by Berthouex et al. (1978a, 1979, 1983).

Barnes and Rowe (1978) brought the wastewater treatment field up to date with stochastic hydrology by applying univariate ARIMA models to generate synthetic sewer flow sequences. This method had been applied much earlier for synthetic streamflow generation (Carlson et al., 1970). Barnes and Rowe modeled time series consisting of flows averaged over 4-hour periods, with six values per day over three

months time. The flows were measured at the headworks of two different treatment plants. Interestingly, it was found that the the same model form, seasonal ARIMA(1,1,2)(0,1,1) (s = 6), was adequate for both treatment plants. No forecasting was performed, although synthetic generation is essentially multi-step-ahead forecasting with random shocks being provided by a random number generator. This method may be used to generate many different hypothetical design flow sequences, each preserving the statistical properties of the historically observed series. Such synthetic sewer inflow series should be useful as simulation inputs for treatment plant modeling studies.

A unique application of ARIMA models for forecasting was presented by MacInnes et al. (1978), who modeled 3-hour flow rate averages measured at a primary treatment plant. Eight observations per day were measured over a 2-year period. After removing the periodicity in the data by spectral analysis, the first year's data was used to identify and estimate a nonseasonal ARIMA(9,0,1) model for the stochastic component. The inclusion of a ninth order AR component is quite unusual. The second year of data was then used in conjunction with the model to simulate real-time, in-line flow equalization with forecasting to remove diurnal flow fluctuations.

MacInnes et al. simulated the operation that would have taken place if forecasts and corresponding responsive adjustments in pumping controls had been made every three hours. Each forecast was made for the succeeding 24-hour period, i.e., eight forecasts of 3-hour

averages. The simulated pumping rate was then adjusted so as to maintain the mean forecasted flow rate of the 24-hour period. The model form and parameters were assumed to be constant throughout the second year. The main objective was to develop flow equalization basin sizing curves, showing the annual percentage diurnal variance reduction as a function of basin design storage volume. No measures of forecasting performance or accuracy were reported.

Since MacInnes et al. subtracted out the periodic component from the data and modeled the residuals with an ARIMA model, their approach cannot be easily compared with a direct Box-Jenkins modeling of the raw data. The periodic function accounted for 61% of the variance in the first year's data, and thus should contribute a similar amount to the forecasts for the second year. Therefore, the forecasts were largely deterministic and not indicative of direct Box-Jenkins forecasting.

As part of their work developing a transfer function model relating influent and effluent COD measurements from an industrial activated sludge process, Debelak and Sims (1981) found univariate ARIMA models for both series. However, in place of the iterative Box-Jenkins strategy of identification, estimation, and diagnostic checking, Debelak and Sims relied on the "Akaike information criterion" (Akaike, 1974) to select the best models. They modeled daily influent and effluent COD measurements taken over a 14-month span. In both cases they found that the "best" (minimum information criterion) model was the simple first-difference ARIMA(0,1,0). They also found that upon

attempting to improve the effluent COD model by including influent COD as a predictor variable, the information criterion approach indicated that the univariate first-difference model was actually "better" than the bivariate model. Debelak and Sims thus concluded that influent COD had no explanatory power as a predictor of effluent COD.

It is important to understand what the ARIMA(0,1,0) or first-difference model implies. This model states that each value of the time series is equal to the preceding value plus a random shock. In terms of forecasting, the model implies that the best forecast for tomorrow's COD measurement is simply today's measured value (plus the expected value of the random shock, which is zero). Debelak and Sims reported:

> "A comparison of this model with the actual data for a one step ahead forecast is shown in Figure 14. The model does predict the trends in the actual data but for the purposes of control is probably not precise enough."

This was certainly true, since the plot of one-step-ahead forecasts was simply the plot of the original data shifted over by one interval. Makridakis and Wheelwright (1978) have called the first-difference model "The Naive 1 Model", and use it as the "worst case" for comparing to all other proposed forecasting models.

In a subsequent discussion paper, Berthouex and Hunter (1982) strongly criticized Debelak and Sims (1981) for "letting the Akaike criterion mechanically identify the 'best' model." Berthouex and Hunter pointed out that the Akaike criterion should not even be applied to models which do not pass the basic diagnostic checks for

model adequacy, and that the first-difference model did not pass, as evidenced by the residual ACF. However, it seems ironic that Berthouex and Hunter should criticize an inadequate first-difference model, when Shih (1976) and Berthouex et al. (1978) have themselves reported and utilized an admittedly inadequate (Shih, 1976, p. 74) AR(1) model for prewhitening in transfer function development. A first-difference model may be thought of as an AR(1) model with an AR parameter of unity.

Chapter IV

ARIMA MODELING AND FORECASTING OF WASTEWATER TREATMENT DATA

"Nothing is easier than to 'prove' that a hypothesis is true by testing it by an experiment which is sufficiently inaccurate."

F. Yates

4.1 INTRODUCTION

This chapter presents illustrative results from ARIMA modeling and forecasting of selected wastewater treatment data sets. Two of the data sets and models are taken directly from the existing literature; however, the results obtained from fitting the models and employing them for forecasting are reexamined using previously unreported graphical and quantitative evaluation criteria. These additional measures provide a more revealing evaluation of the forecasting performance of ARIMA models applied to wastewater treatment data. For one of the previously reported data sets, an alternative model is identified and presented because it was found that the goodness-of-fit and whiteness of the residuals could be improved. The remaining data sets presented here have not, to the author's knowledge, been previously analyzed using ARIMA models.

4.2 METHODS OF ANALYSIS AND EVALUATION CRITERIA

The data sets were chosen solely on the basis of availability to the author and the sampling interval of the observations. For real-time control applications, frequent measurements are required. The more readily available daily average measurements from treatment plants were not considered useful for this study. As noted by previous researchers (e.g., Wallace and Zollman, 1971; Shih, 1976), more frequent data are seldom available. The required sampling frequency will depend on the intended application; as a point of interest it is noted that Berthouex et al. (1979) reported essentially the same model form and information content (explanatory power) using either a two-hour or a one-hour sampling interval, for the particular data sets they studied.

All analyses were performed on the University of California, Los Angeles campus IBM 3033 computer using the commercial statistical software package "SAS"; specifically, the "ARIMA" procedure (SAS Institute Inc., 1982) and "SAS/GRAPH" graphics system (SAS Institute Inc., 1981).

For each data set, an ARIMA model was developed using the Box-Jenkins strategy of identification, estimation, and diagnostic checking. For a data set and model which had already been reported in the literature, the model's adequacy was checked, and if found to be acceptable, the previously proposed model was utilized. If the previously reported model did not pass diagnostic checking (according to present-day diagnostic test statistics which had not been developed

at the time the model was first reported), an improved model was sought and reported for comparison. The identified model was fitted to a smaller subset of the original time series, and used to perform both multi-step-ahead and one-step-ahead beyond-sample forecasting for the remaining period at the end of the data set, chosen to be 24 hours in order to simulate a typical day's forecasting results. Appendix A contains a summary of all the models studied, together with their estimated parameter values from the SAS ARIMA conditional least-squares estimation routine (SAS Institute, 1982).

The fitting and forecasting results were plotted and then evaluated by examining the structure and magnitudes of the errors. The following error criteria were also calculated: R-squared (fitting only), the range of absolute errors (RAE), the mean absolute error (MAE), the range of absolute percent errors (RAPE), the mean absolute percent error (MAPE), and the root mean square error (RMSE). Their definitions are given in Appendix B. The term "absolute" refers to absolute values of the errors, which may be positive or negative.

In the following section, the first data set is analyzed and discussed at some length in order to illustrate the important concepts and limitations associated with ARIMA forecasting of wastewater treatment data. The remaining data sets reinforce these initial indications.

4.3 MADISON NINE SPRINGS TREATMENT PLANT INFLUENT BOD

The first data set consists of N = 332 hourly observations of influent BOD from the Nine Springs Sewage Treatment Plant at Madison, Wisconsin. Figure 4 shows a plot of the data, with straight lines connecting each hourly value. The data were taken from Shih et al. (1974). This data set has also been analyzed and discussed in Berthouex et al. (1975, 1978a, 1979, 1983), Berthouex and Hunter (1982), and Shih (1976).

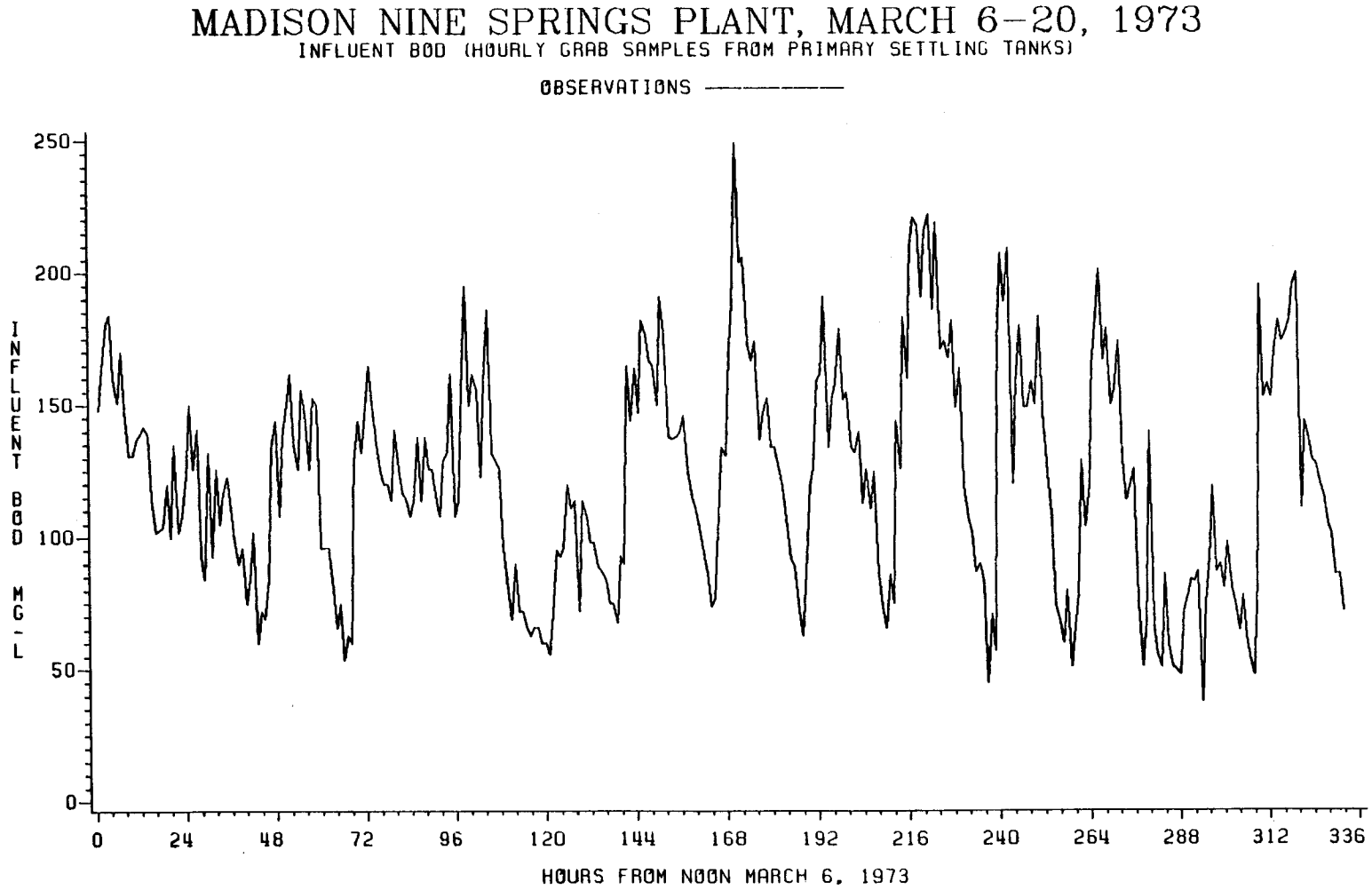


Figure 4: Nine Springs Influent BOD Data

4.3.1 Nonseasonal ARIMA(1,0,0) Model

As mentioned previously in the literature review, Berthouex et al. (1975) proposed an AR(1) or ARIMA(1,0,0) model for the time series. This model represents an observed value's deviation from the series mean by a lag-one autoregression on the previous hour's deviation.

4.3.1.1 Fitted Values

Figure 5 shows the fitted values obtained from the AR(1) model (fitted to all N = 332 observations) as they appeared in Berthouex et al. (1975, Figure 1, p. 129), where they were called "forecasts." Plotted on this scale, the fitted values appear to closely follow the observed data with perhaps minor fitting errors. However, Figure 6 presents the same fitted and observed values together with the absolute values of the residuals (differences between observed and fitted values) plotted as vertical deviations. This allows the absolute error in mg/l BOD to be read directly from the same ordinate scale used for the observed and fitted values. It can be seen that many of the residuals are quite large, exceeding 50 to 75 mg/l, with some errors extending into the same magnitude range as the data themselves. This is particularly true at the "turning points" where the time series takes large jumps upward or downward.

Figure 7 provides a magnified view of the last 45 data points together with their corresponding fitted values and absolute errors. In addition to the sizeable magnitudes of some residuals, a more troubling problem can be seen--the fitted values or one-step-ahead

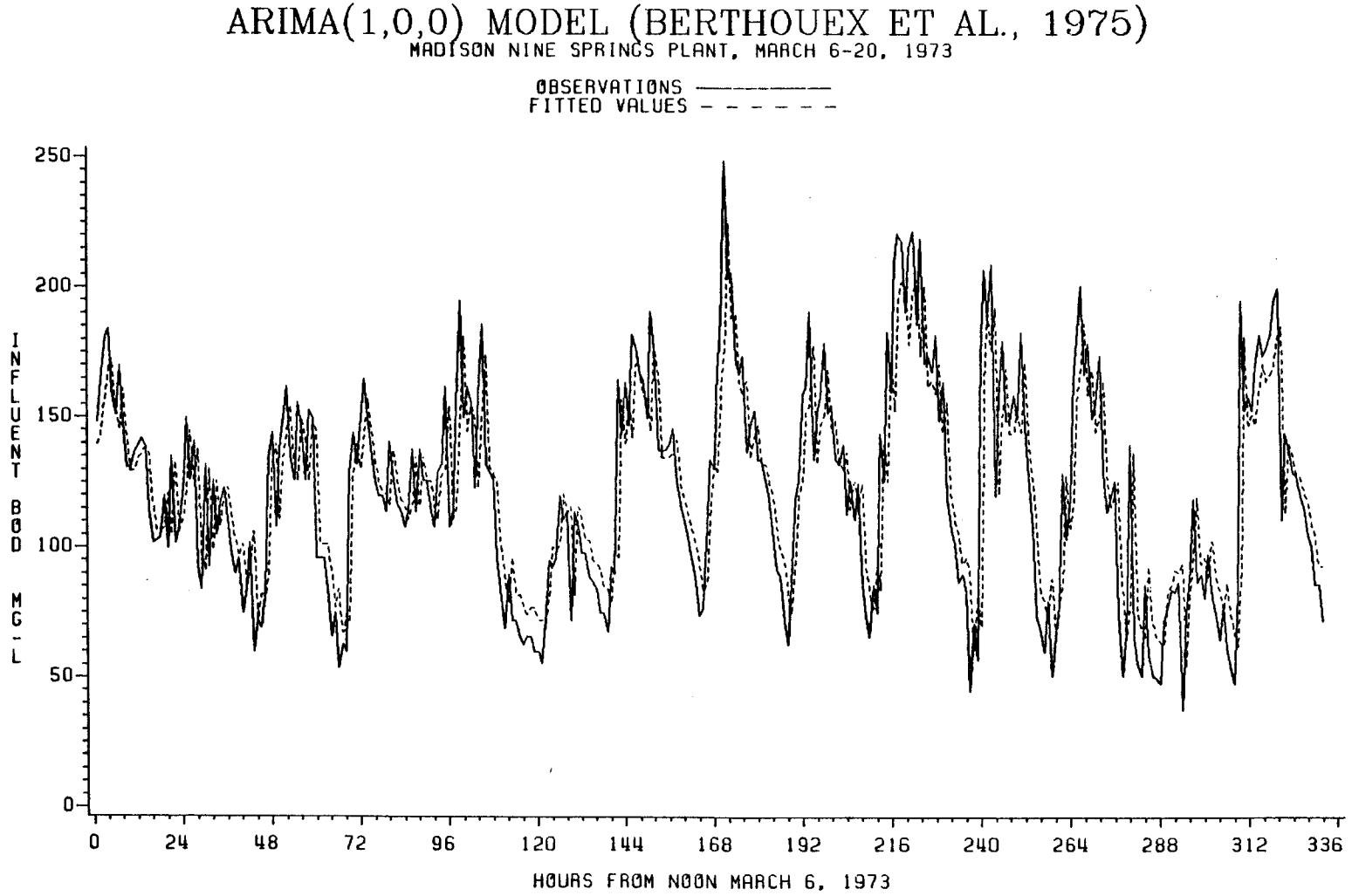


Figure 5: Fitted Values from ARIMA(1,0,0) Model

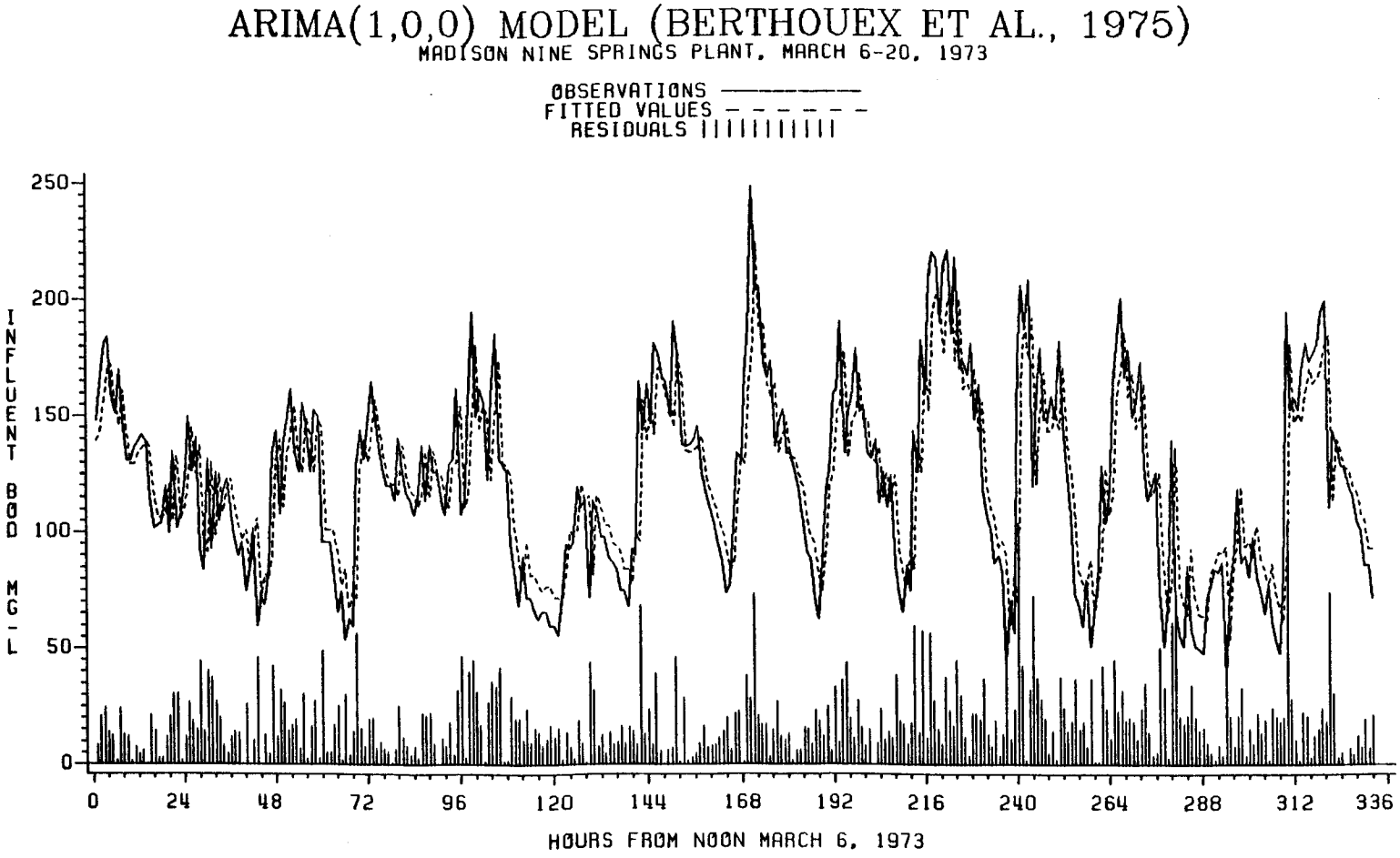


Figure 6: Fitted Values and Residuals from ARIMA(1,0,0) Model

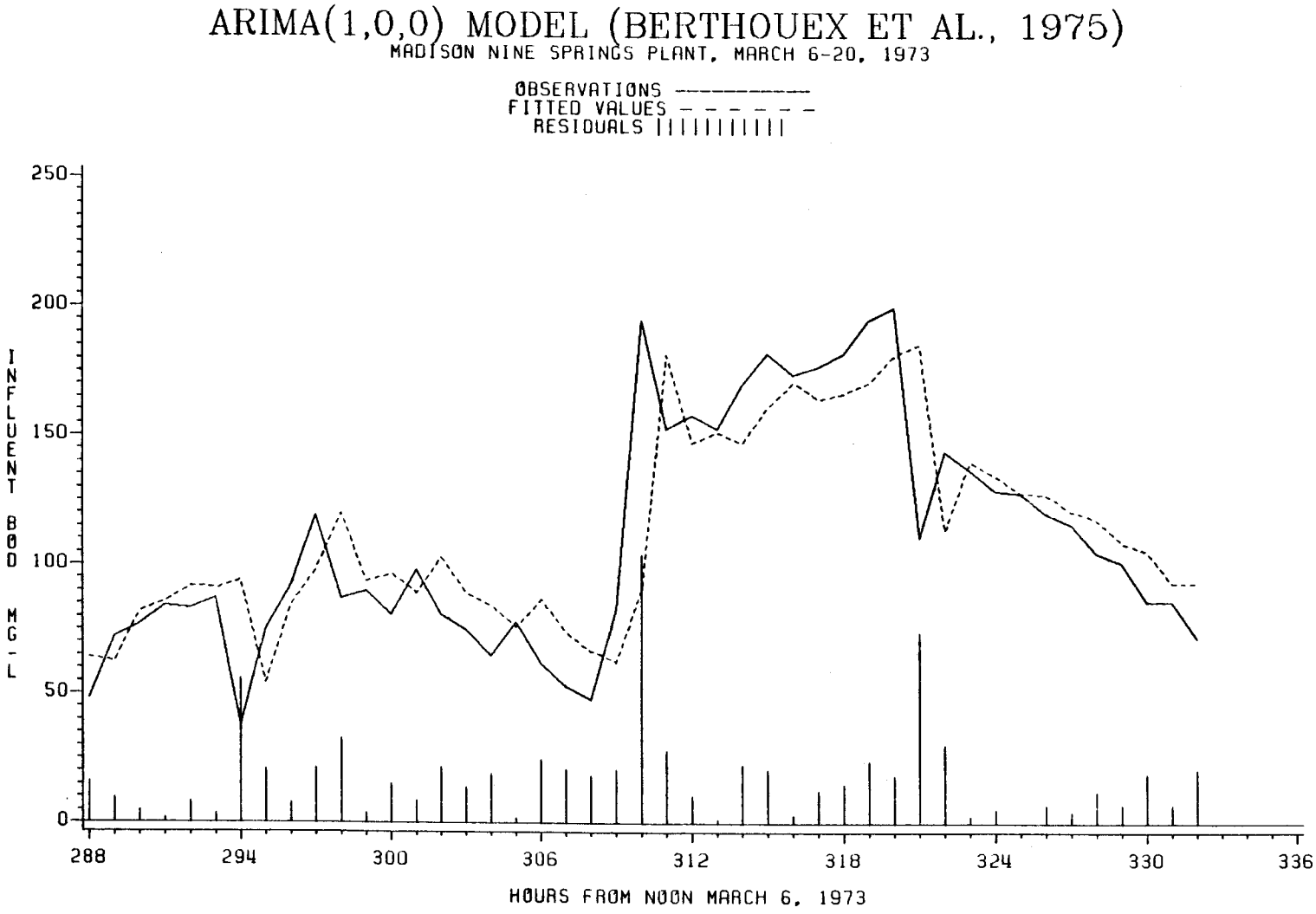


Figure 7: Detail of Fitting Error from ARIMA(1,0,0) Model

within-sample forecasts appear to lag behind the observed BOD measurements by one hour. That is, the predicted values achieve their "peaks and valleys" (maxima and minima) one hour later than the actual data.

Shih (1976, p. 74) also noted the failure of the AR(1) model to reproduce jumps in the data and made mention of the lag effect. He attributed the problems to the fact that "since no seasonal terms were built in <the model> it cannot respond to the strong diurnal variation fast enough." However, the present author has observed the same one-interval lag effect in every plot of observed and fitted values from univariate ARIMA models reviewed in the references for this research. This was true of both nonseasonal and seasonal models, for a broad range of water resources quality and quantity data. This point will be returned to in later subsections and in the conclusions.

4.3.1.2 Multi-Step-Ahead Beyond-Sample Forecasts

Figure 8 shows a sequence of 24 multi-step-ahead beyond-sample forecasts (lead times = 1, 2 ,..., 24) that were made using the AR(1) model. For this stage of the modeling, only the first N = 300 observations were used for estimation of the model. The resulting estimated model (see Appendix A), which was still assumed to be AR(1), was then used to forecast the next 24 values without using any information (observations) occurring after hour 300. Since hour 300 corresponds to midnight of Sunday, March 18, this sequence of "bootstrap" forecasts simulates what would have happened if a

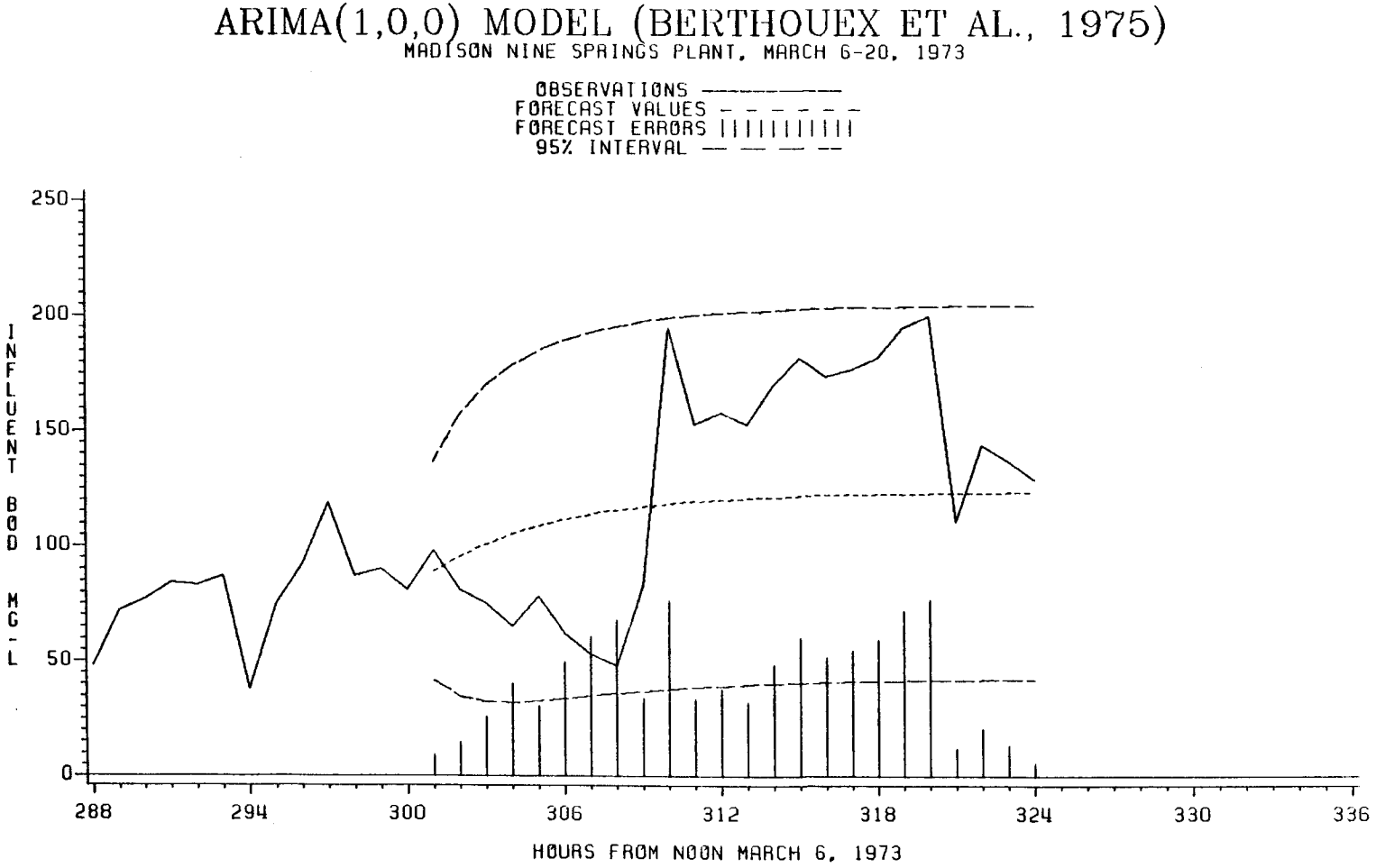


Figure 8: Multi-Step-Ahead Forecasts from ARIMA(1,0,0) Model

treatment plant engineer had used the AR(1) model to forecast Monday's entire hourly BOD profile in advance. The short-dashed line shows the forecast values, which are bracketed above and below by the long-dashed lines representing the upper and lower 95% confidence limits for each forecast. It might be argued that the simulated forecasts are not physically realistic, since it is not possible in practice to obtain hourly 5-day BOD measurements in real time. However, the results would be comparable if some surrogate measure of BOD such as total organic carbon (TOC) were used instead. TOC can be readily determined on an hourly or even more frequent basis.

It is apparent from the figure that the AR(1) model is of little use in predicting future BOD values via multi-step-ahead forecasts, i.e., for long lead times. This is to be expected, since it can be shown that multi-step-ahead forecasts from such an AR(1) model simply converge to the mean of the series (Box and Jenkins, 1976). The forecasts display no pattern or structure such as would be needed to anticipate the observed diurnal variation. This is where a seasonal model might be of significant benefit.

Of equal importance are the extremely wide confidence intervals for the forecasts from the AR(1) model. It should be understood that although stochastic ARIMA models are used to generate point forecasts (single predicted values), a point forecast is simply a conditional expected value based on the previous history of the series. The proper interpretation of the forecast is actually as follows: the future value is "expected" to be the model-generated value, but the

true value will fall between the upper and lower confidence limits about 95% of the time. Since the confidence limits in Figure 8 encompass almost the entire range of observed BOD values (approximately 50 to 200 mg/l BOD), such probabilistic information is of little value.

The width of the confidence interval depends on the estimated variance of the forecast errors, which is in turn proportional to the estimated variance of the random shocks (Box and Jenkins, 1976). Since the variance of the random shocks is estimated by the sample variance of the fitted residuals from the estimation stage, which is seen to be large (Figure 6), the confidence intervals will therefore be relatively wide. Practically speaking, this means that if the data contain large unexplained variation (e.g., experimental error from BOD tests) so that the ARIMA model developed has a low R-squared, then the forecast errors will also display large variation and the confidence intervals will be wide.

4.3.1.3 One-Step-Ahead Beyond-Sample Forecasts

Figure 9 shows a sequence of one-step-ahead beyond-sample forecasts that were generated from the AR(1) model. Again, the model was estimated using only the first N = 300 observations. The resulting model was then used to generate a series of one-step-ahead forecasts, this time utilizing one additional observation to make each successive forecast. That is, the forecast for hour 301 was based on the observation from hour 300, the forecast for hour 302 was based on the

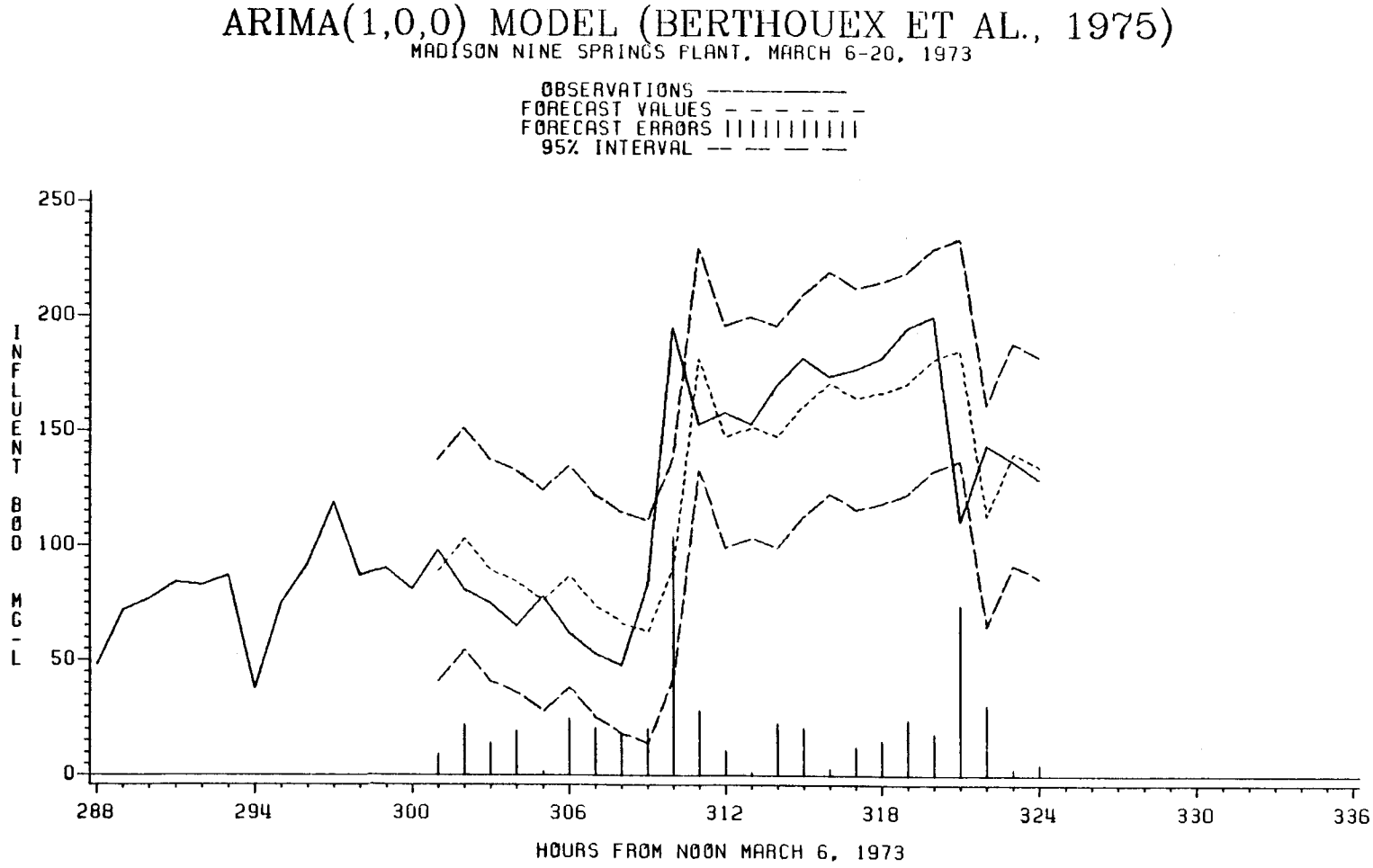


Figure 9: One-Step-Ahead Forecasts from ARIMA(1,0,0) Model

observed value from hour 301, and so on, up to hour 324. This is in contrast to the multi-step-ahead forecasts in the preceding subsection, where each new forecast relied solely on the previous hour's forecast. The parameter values were held fixed throughout the 24-hour forecast period. This is a good assumption, since Appendix A shows that the parameter estimates obtained using N = 300 or the full N = 332 observations are not significantly different. This was verified by a t-test at the 5% level (Pankratz, 1983).

Careful study of Figure 9 and comparison with Figure 7 indicates that the one-step-ahead beyond-sample forecasts are almost identical to the fitted values; displaying the same one-hour lag effect and hence failing to predict rapid changes in level of the series. This is to be expected, since it can be shown that if the parameter values do not change from hour 300 to hour 332, the one-step-ahead within-sample forecasts (fitted values) from hour 301 to hour 324 are the same as one-step-ahead beyond-sample forecasts from hour 301 to hour 324.

The one-step-ahead beyond-sample forecasts simulate what would have happened if forecasting had been performed hourly, as new observations became available. Note the increase in effort required compared to multi-step-ahead forecasting, where the entire 24 hours can be forecast in advance (e.g., at midnight of each day). Note also the improved forecasting accuracy as indicated by the reduced magnitudes of the absolute errors and the decreased width of the 95% confidence intervals. This illustrates an important point concerning the

tradeoff between computational labor and accuracy for one-step-ahead versus multi-step-ahead beyond-sample forecasting.

4.3.1.4 Fitting and Forecasting Accuracy

Criterion	Fitted Values (N = 332)	Multi-Step-Ahead Forecasts (N = 24)	One-Step-Ahead Forecasts (N = 24)
R-squared	0.65	---	---
RAE	0 - 104 mg/l	6 - 77 mg/l	2 - 104 mg/l
MAE	19 mg/l	41 mg/l	22 mg/l
RAPE	0 - 147 %	4 - 141 %	1 - 67 %
MAPE	17 %	38 %	20 %
RMSE	25 mg/l	46 mg/l	31 mg/l

TABLE 1

Error Criteria for ARIMA(1,0,0) Model

Figures 6 and 7 provide a qualitative, intuitive view of the overall accuracy with which the AR(1) fitted values match the observed values. While such plots are extremely useful, the graphical description can be made quantitative by computing the error criteria shown in the first column of Table 1. These criteria indicate several important aspects of the fitted values, which are enumerated explicitly in order to illustrate the utility of the quantitative criteria:

1. The model explains only 65% of the variance in the data (R-squared).
2. Some residuals (fitting errors) are as large as 104 mg/l (see RAE).
3. On the average, the absolute fitting error (see MAE) is about 19 mg/l.
4. Some absolute percent errors are as large as 147% (see RAPE).
5. On the average, the absolute percent error (see MAPE) is about 17%.
6. The overall fitting error, as measured by RMSE, is 25 mg/l.

By themselves, these numbers provide practical information about the amount of error incurred in fitting the AR(1) model to the observed data. More importantly, they may be used for comparison to alternative models as a basis for selection.

Figures 8 and 9 indicate visually the relative accuracy of the AR(1) model for forecasting 24 hourly values in the observed influent BOD series. The accuracy of the forecasts can be evaluated by reviewing the error criteria computed in the second and third columns of Table 1.

Several important points can be seen from consideration of these values:

1. Although the maximum absolute forecast error (see RAE) is larger for the one-step-ahead forecasts, on the average the absolute errors (see MAE) are nearly halved by performing one-step-ahead forecasting.

2. The range of absolute percent errors (RAPE) is also greatly reduced for one-step-ahead forecasting.
3. On the average, one-step-ahead forecasting reduces the percentage error (MAPE).
4. Overall, the RMSE was reduced by one-step-ahead forecasting.

Consider the one-step-ahead beyond-sample forecasts, which would clearly be preferable if real-time forecasting were to be performed. It should be noted that, in practice, an error as large as 104 mg/l or 67% when forecasting influent BOD which ranges from 50 to 250 mg/l could have important consequences. This is particularly disturbing because the largest errors occur when forecasting the critical maxima and minima, which may affect process stability.

The preceding subsections indicate that there are many more considerations inherent in forecasting with an ARIMA model than are conveyed in a single plot of fitted values such as Figure 5.

4.3.2 Seasonal ARIMA(2,0,0)(0,1,1) Model

As discussed in the literature review, Shih (1976) tested 25 nonseasonal and seasonal ARIMA models for the Nine Springs influent BOD series and found that the model which best passed the diagnostic checking stage was a seasonal ARIMA(1,0,0)(0,1,1) model with s = 24. However, Shih found that this seasonal model had a lower R-squared

value (0.63) than the simpler AR(1) originally reported by Berthouex et al. (1975). For this reason, as well as parsimony of parameters, Shih chose to use the AR(1) model for prewhitening in transfer function development. However, he mentioned that the seasonal model should be used for forecasting. The following sections illustrate the utility of this suggestion.

Shih (1976) reported that the AR(1) model did not pass diagnostic checking, but found that the ARIMA(1,0,0)(0,1,1) model did so. In order to apply Shih's seasonal model for forecasting, the present author first ran diagnostic checks to verify the model's adequacy. It was found that the ARIMA(1,0,0)(0,1,1) model (see Appendix A) showed small but significant spikes at lags one and three in the residual ACF, and at lag one in the residual PACF. The author also observed that the residual ACF did not pass the chi-square test; the chi-square value was significant at the 1% level for all lag orders up to lag 42 (see comment below). On the basis of these findings, an alternative model was sought.

Prior to reviewing Shih's dissertation (Shih, 1976), the present author had independently analyzed the Nine Springs influent BOD series using the data given in Shih et al. (1974). As expected, the present author arrived at almost the same seasonal model form as Shih; however, the author found that including one additional nonseasonal AR term at lag two, i.e., ARIMA(2,0,0)(0,1,1), yielded white noise residuals. Appendix A gives the parameter estimates for the proposed ARIMA(2,0,0)(0,1,1) model. It was found that the ACF spikes at lags

one and three were eliminated, as was the lag-one PACF spike. In addition, the chi-square statistic was not significant at the 5% level (ranging from significance at the 15% level to over the 40% level for all lag orders through lag 42). It should be pointed out, however, that Shih (1976) used the older "Box-Pierce" chi-square statistic and included 50 lag orders, whereas the present author used the more recently developed "Ljung-Box" statistic for all analyses and included 42 lag orders (see for example Pankratz, 1983 for a discussion of the differences). Lastly, the R-squared value (0.65) for the proposed ARIMA(2,0,0)(0,1,1) model was slightly higher than for Shih's seasonal model. The proposed model is a better model in terms of present-day formal Box-Jenkins diagnostic checking procedures. Notice, however, that the seasonal model provides no improvement in fit over the AR(1) model, which also had an R-squared of 0.65.

4.3.2.1 Fitted Values

Figure 10 shows the fitted values and residuals obtained from the proposed seasonal ARIMA(2,0,0)(0,1,1) model. Twenty four fitted values at the beginning of the series are not calculated due to the seasonal differencing with s = 24. It is immediately seen that the residuals do not appear any smaller overall than those from the AR(1) model, as expected from the equal R-squared values for the models.

Figure 11 shows a magnified view of the last 45 data points, fitted values, and residuals obtained from the proposed model. It is apparent that the seasonal model also displays the one-interval lag

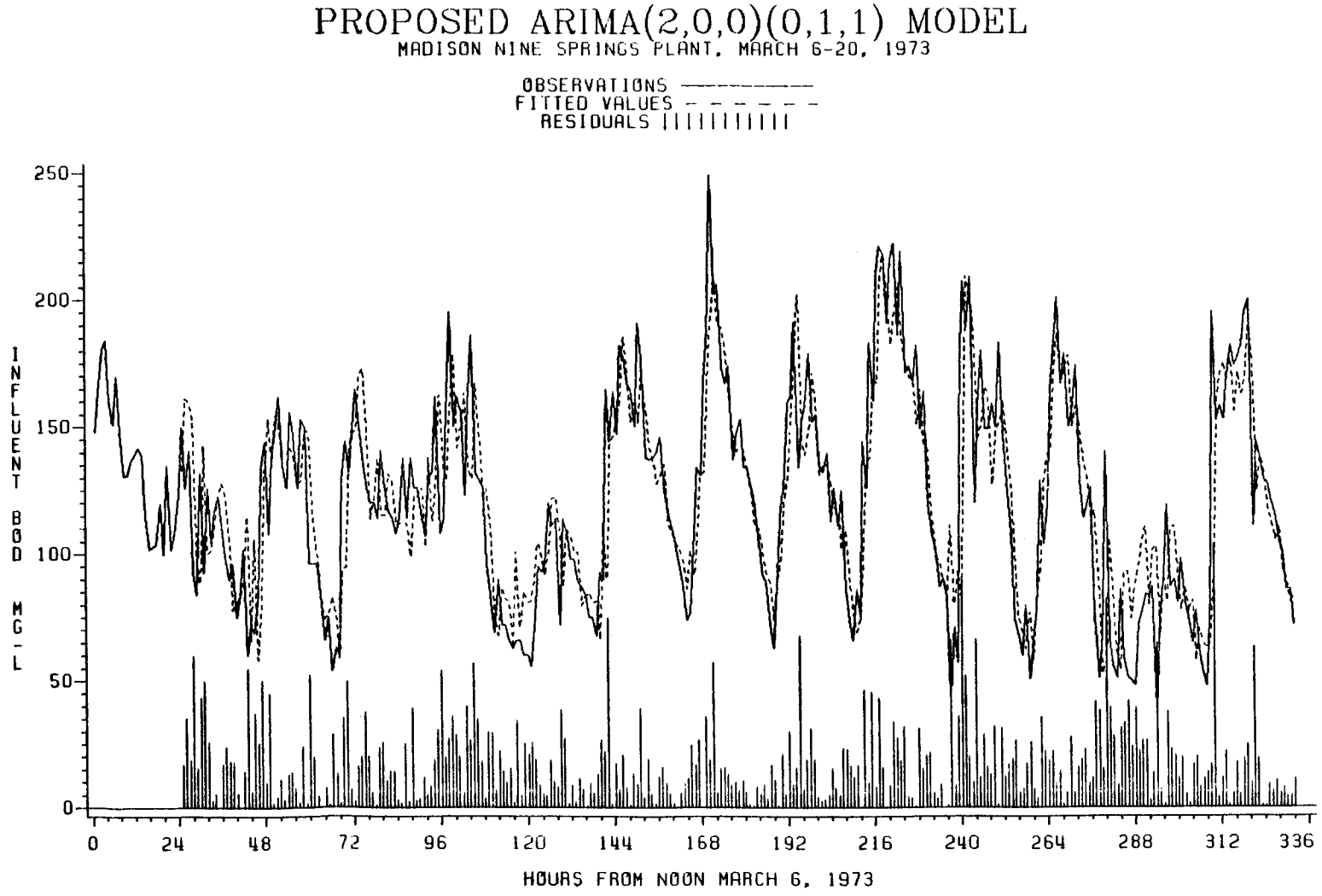


Figure 10: Fitted Values and Residuals from Seasonal Model

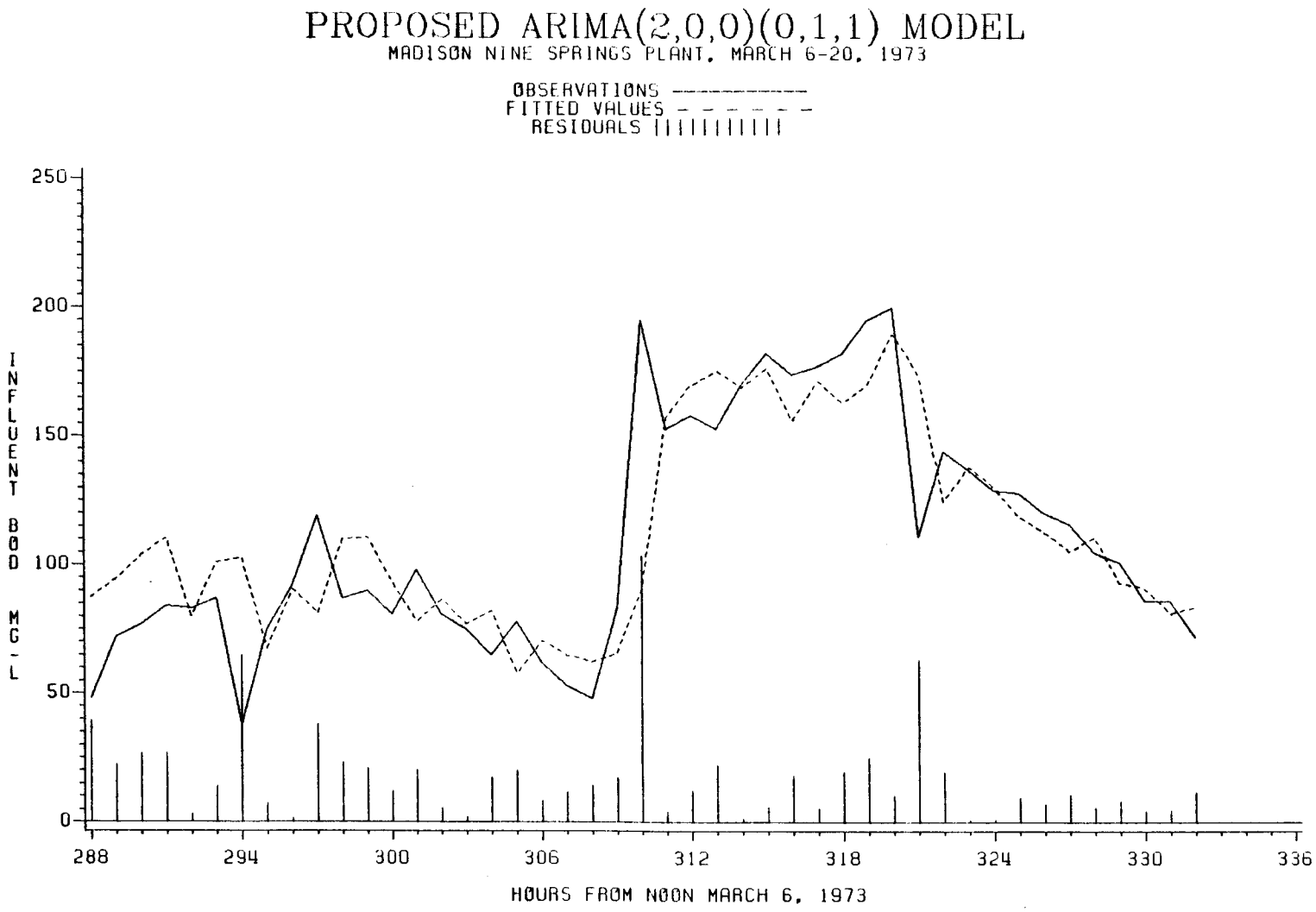


Figure 11: Detail of Fitting Error from Seasonal Model

effect noted previously for the AR(1) model. Again, the largest residuals occur at the turning points of the series. Thus, for this particular data set, a seasonal model does not alleviate these problems.

4.3.2.2 Multi-Step-Ahead Beyond-Sample Forecasts

Figure 12 shows 24 multi-step-ahead beyond-sample forecasts (lead times = 1, 2, ... , 24) generated using the seasonal ARIMA(2,0,0)(0,1,1) model fitted to the first N = 300 observations (see Appendix A). No observations occurring past hour 300 were utilized in producing the forecasts. It is seen that a marked improvement in multi-step-ahead beyond-sample forecasting accuracy is gained from the seasonal model as compared with the AR(1) model (cf. Figure 8). The forecasts follow a structured pattern which anticipates a diurnal cycle, based upon the previous cycles in the data. This occurs because the seasonal model's forecast function incorporates information from not only the two preceding lag intervals, but also from the corresponding lags 24, 25, and 26 hours before. By contrast, the forecast function for the AR(1) model only utilizes information from the preceding hour. These statements may be verified by expanding the AR and MA polynomials in the backshift-operator forms for the models, as shown in Appendix A.

It is particularly interesting to note that the 24 forecasts from hour 301 to hour 324 correspond to Monday, March 19 and are partially based on observations from Sunday, March 18 (hours 277 to 300). The

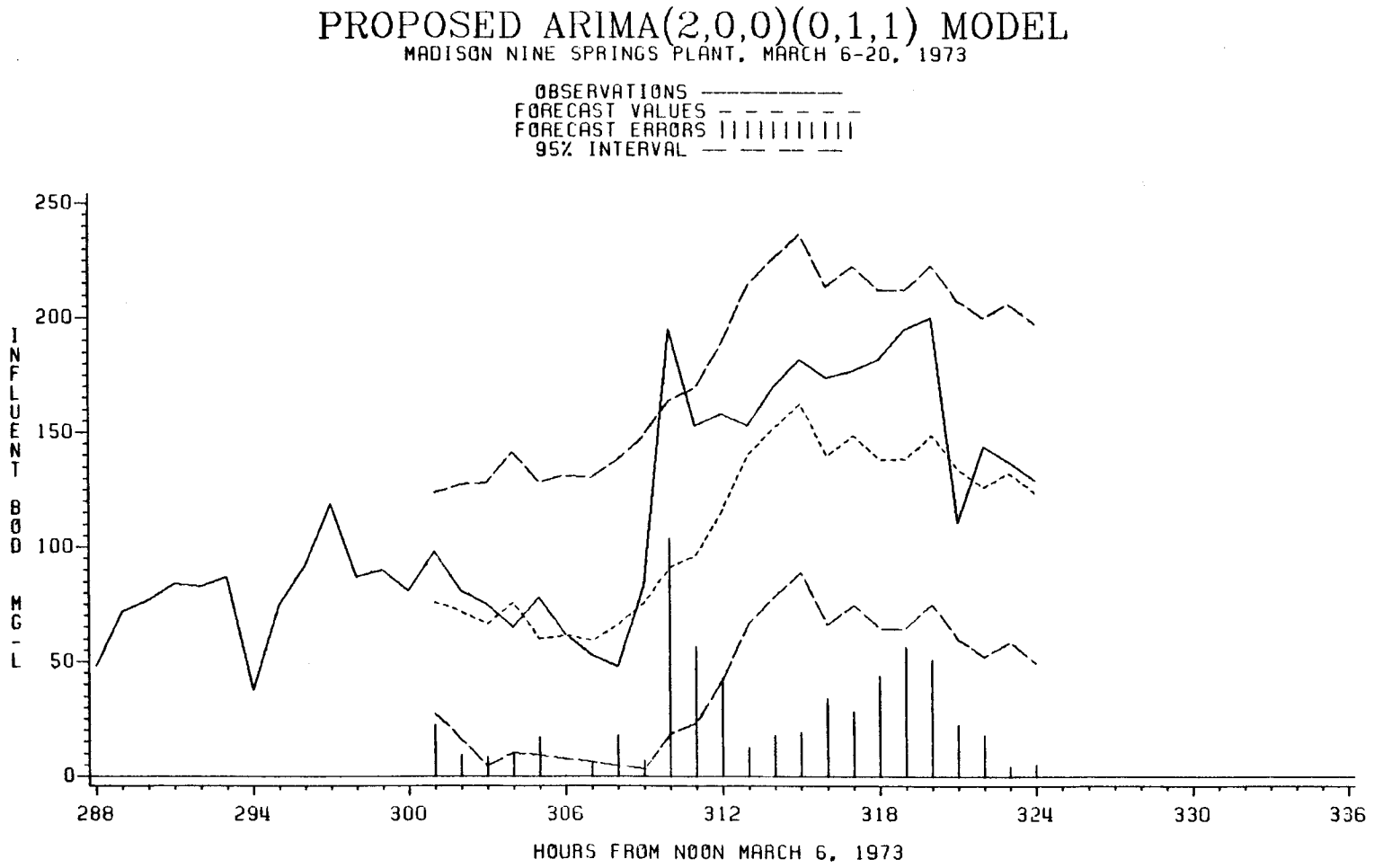


Figure 12: Multi-Step-Ahead Forecasts from Seasonal Model

diurnal variation for Sunday's BOD is quite different from the other days of the week, yet the forecasts anticipate Monday's profile reasonably well. The relative utility of the seasonal model over the nonseasonal AR(1) model for multi-step-ahead beyond-sample forecasting is apparent. Recall that this could not be inferred on the basis of the fitted results from the two models alone.

4.3.2.3 One-Step-Ahead Beyond-Sample Forecasts

Figure 13 shows a sequence of 24 one-step-ahead beyond-sample forecasts (all lead times = 1) from the seasonal model fitted to the first N = 300 observations. While the one-step-ahead beyond-sample forecast errors are somewhat lower than the multi-step-ahead errors (cf. Figure 12) as expected, the forecasts do not appear to be much of an improvement over those from the simpler AR(1) model (cf. Figure 9). This is a potentially significant result, because it implies that in some cases a much simpler model may suffice if one-step-ahead forecasting is the objective. Examination of the largest forecast errors (hours 310 and 321) indicates that for this particular data set, the seasonal model performs no better than the AR(1) model for predicting rapid changes in the concentration of incoming BOD.

4.3.2.4 Fitting and Forecasting Accuracy

The first column of Table 2 shows the fitting error criteria calculated for the seasonal model. Comparison with the corresponding values from Table 1 for the AR(1) model reveals the following:

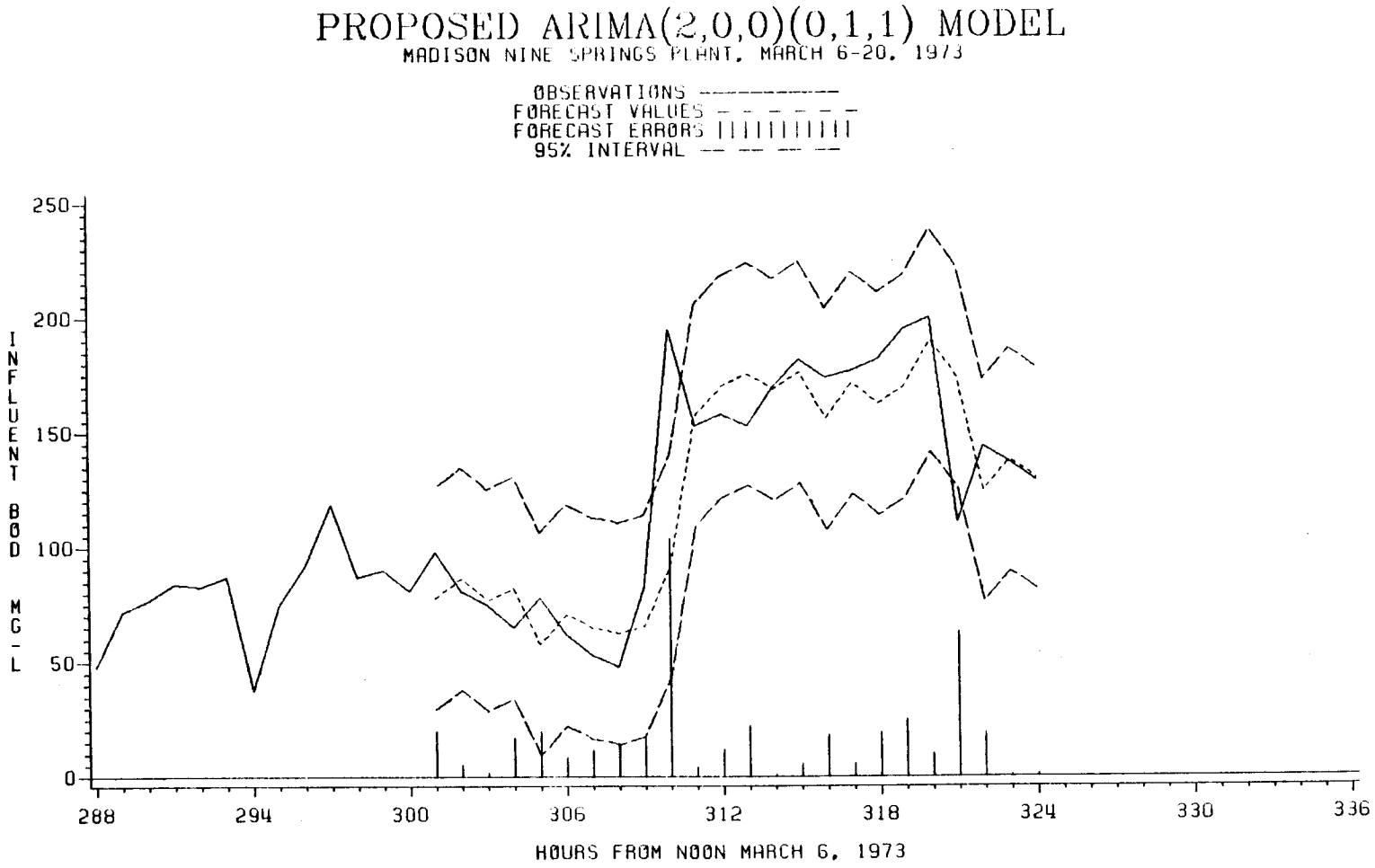


Figure 13: One-Step-Ahead Forecasts from Seasonal Model

Criterion	Fitted Values (N = 308)	Multi-Step-Ahead Forecasts (N = 24)	One-Step-Ahead Forecasts (N = 24)
R-squared	0.65	---	---
RAE	0 - 104 mg/l	1 - 104 mg/l	1 - 104 mg/l
MAE	18 mg/l	26 mg/l	18 mg/l
RAPE	0 - 171 %	1 - 53 %	1 - 57 %
MAPE	17 %	19 %	15 %
RMSE	24 mg/l	35 mg/l	28 mg/l

TABLE 2

Error Criteria for ARIMA(2,0,0)(0,1,1) Model

1. The maximum absolute error is the same as from the AR(1) model.
2. The MAE is only 1 mg/l lower for the seasonal model.
3. The maximum absolute percent error is higher for the seasonal model.
4. The MAPE is the same as found with the AR(1) model.
5. The RMSE is only 1 mg/l lower for the seasonal model.

Overall, from the standpoint of fitting error only, the seasonal model seems to be no better than a simpler, more parsimonious AR(1).

The second and third columns of Table 2 provide the error criteria for both types of forecasts obtained from the seasonal model. Several points arise from comparison of the multi-step-ahead and one-step-ahead beyond-sample forecasts:

1. Neither type of forecast anticipates the large BOD peak at hour 310 (absolute error = 104 mg/l).
2. The overall MAE, MAPE, and RMSE are reduced by performing one-step-ahead beyond-sample forecasting.
3. Both types of forecasts yield a similar range of absolute percent errors.

It is also instructive to compare the forecasts from the seasonal model to those from the AR(1) model (cf. Table 1). The following facts are noted:

1. The overall MAE, MAPE, and RMSE are demonstrably improved by using the seasonal model for multi-step-ahead beyond-sample forecasting. In fact, the multi-step-ahead beyond-sample forecasts are comparable in accuracy to the one-step-ahead beyond-sample forecasts from the AR(1).
2. The overall MAE, MAPE, and RMSE are only slightly improved by using the seasonal model for one-step-ahead beyond-sample forecasting.
3. The accuracy difference between one-step-ahead and multi-step-ahead beyond-sample forecasts is not as dramatic as with the AR(1) model.

4.3.3 Comments About the Nine Springs Models

It should be emphasized that the presentation of the ARIMA(2,0,0)(0,1,1) model is not intended to discount the AR(1) or ARIMA(1,0,0)(0,1,1) proposed by Berthouex et al. (1975) and Shih (1976). A seasonal model was presented in order to highlight the differences in forecasting (particularly multi-step-ahead beyond-sample) performance and character between seasonal and nonseasonal models. The same conclusions would have been reached had Shih's ARIMA(1,0,0)(0,1,1) model been used for forecasting instead. In fact, the author found that the inclusion of the additional AR term at lag two made little difference in terms of fitting or forecasting accuracy. Plotted results from both seasonal models were almost indistinguishable.

4.4 MINNEAPOLIS-SAINT PAUL SEWER STATION 004 COD

The second data set analyzed consists of N = 96 hourly COD grab samples from Sewer Station 004 of the Minneapolis-Saint Paul Sanitary District's combined sewer system. The data were taken from Table 1 of Wallace and Zollman (1971), and cover a four-day period during April, 1967.

4.4.1 Nonseasonal ARIMA(0,1,1) Model

McMichael and Vigani (1972) proposed a nonseasonal ARIMA(0,1,1) model for the "Table 1" COD series. This model represents the first difference of the data by a lag-one moving average component.

4.4.1.1 Fitted Values

Figure 14 shows the fitted values obtained from the ARIMA(0,1,1) model as they appeared in McMichael and Vigani (1972), where they were referred to as "one interval ahead forecasts." The fitted residuals are also shown in order to indicate the magnitudes of the absolute errors. One fitted value is lost due to the nonseasonal differencing of order one.

While the ARIMA(0,1,1) fitted values display a different character than those from an AR(1) model (which always appear as lagged, scaled versions of the original data), it can be seen that a one-interval lag effect is also present in these results. As a consequence, the model fails to predict the turning points, particularly at hours 22 and 42.

4.4.1.2 Multi-Step-Ahead Beyond-Sample Forecasts

Figure 15 shows 24 multi-step-ahead beyond-sample forecasts generated from the ARIMA(0,1,1) model fitted to the first N = 72 observations (see Appendix A). Those unfamiliar with Box-Jenkins models may be surprised to see that the forecasts after hour 73 simply follow a straight line. It can be shown that all forecasts for lead times greater than one from an ARIMA(0,1,1) model are equal to the forecast for lead time = 1 (Box and Jenkins, 1976). Clearly, such a model is entirely unsuitable for multi-step-ahead beyond-sample forecasting of treatment plant influent, which will never remain constant over time. This drawback of the ARIMA(0,1,1) model has also been noted by Litwin and Joeres (1975).

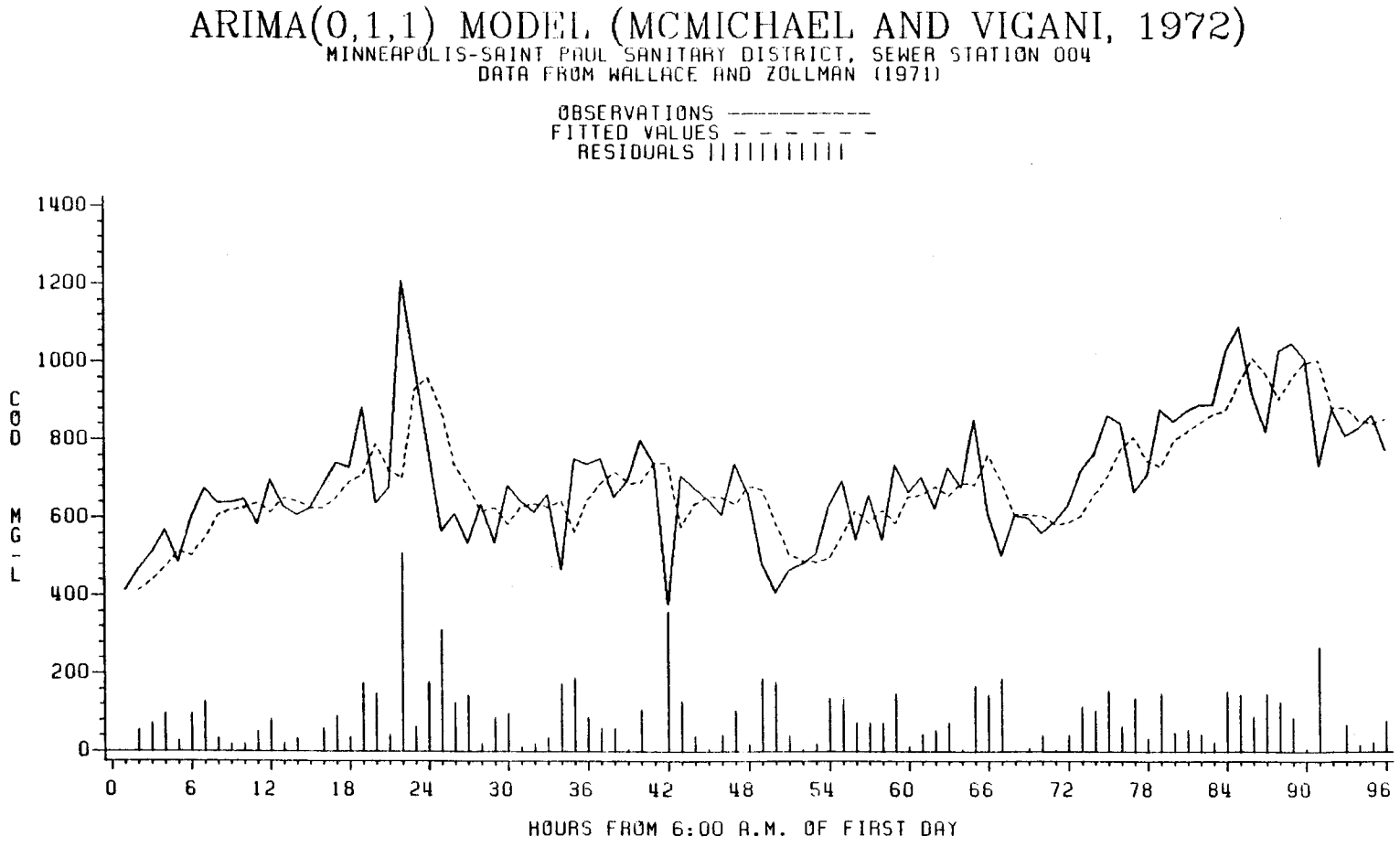


Figure 14: Fitted Values and Residuals from ARIMA(0,1,1) Model

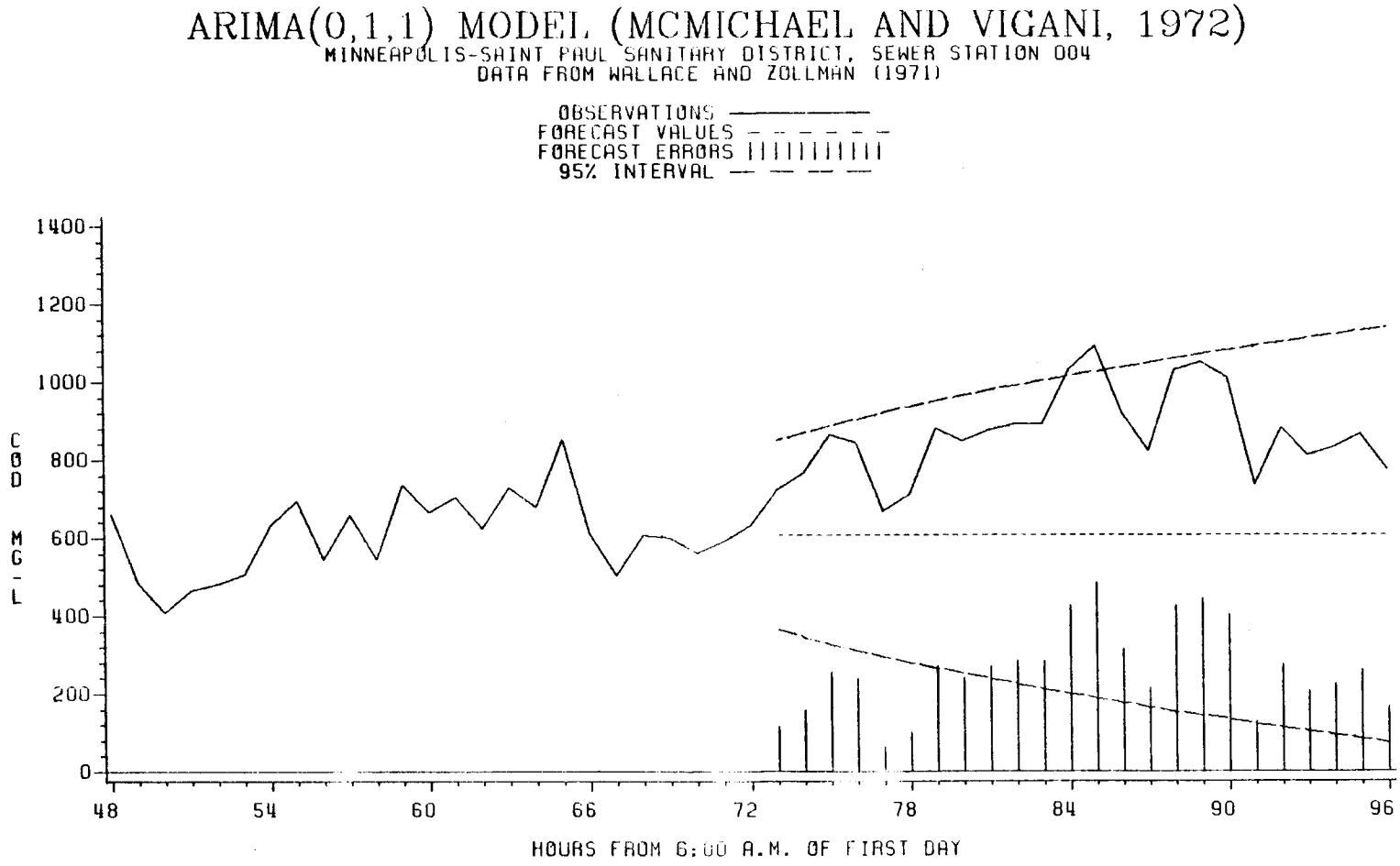


Figure 15: Multi-Step-Ahead Forecasts from ARIMA(0,1,1) Model

4.4.1.3 One-Step-Ahead Beyond-Sample Forecasts

Figure 16 shows 24 one-step-ahead beyond-sample forecasts from the ARIMA(0,1,1) model fitted to the first N = 72 observations. The confidence intervals are markedly reduced. These forecasts generally appear to follow the data fairly well, but this is largely due to the fact that this portion of the series (hours 73 to 96) does not exhibit any large, sudden fluctuations. The choice of scale for the ordinate axis (determined by a figure size limitation) also affects the appearance of "accuracy." The one-interval lag effect can again be clearly seen.

4.4.1.4 Fitting and Forecasting Accuracy

Table 3 shows the fitting and forecasting accuracy results for the ARIMA(0,1,1) model. Several considerations about the results may be noted:

1. The ARIMA(0,1,1) model only accounts for 42% of the variance in the data.
2. The multi-step-ahead beyond-sample forecast errors are rather large as expected, since the straight line forecast provides no information about future COD variability. The absolute error averages 261 mg/l, and the overall RMSE is 283 mg/l over a portion of the series where the COD ranges from 669 to 1090 mg/l.
3. The one-step-ahead beyond-sample forecasts are much more useful, with absolute errors averaging only 93 mg/l and

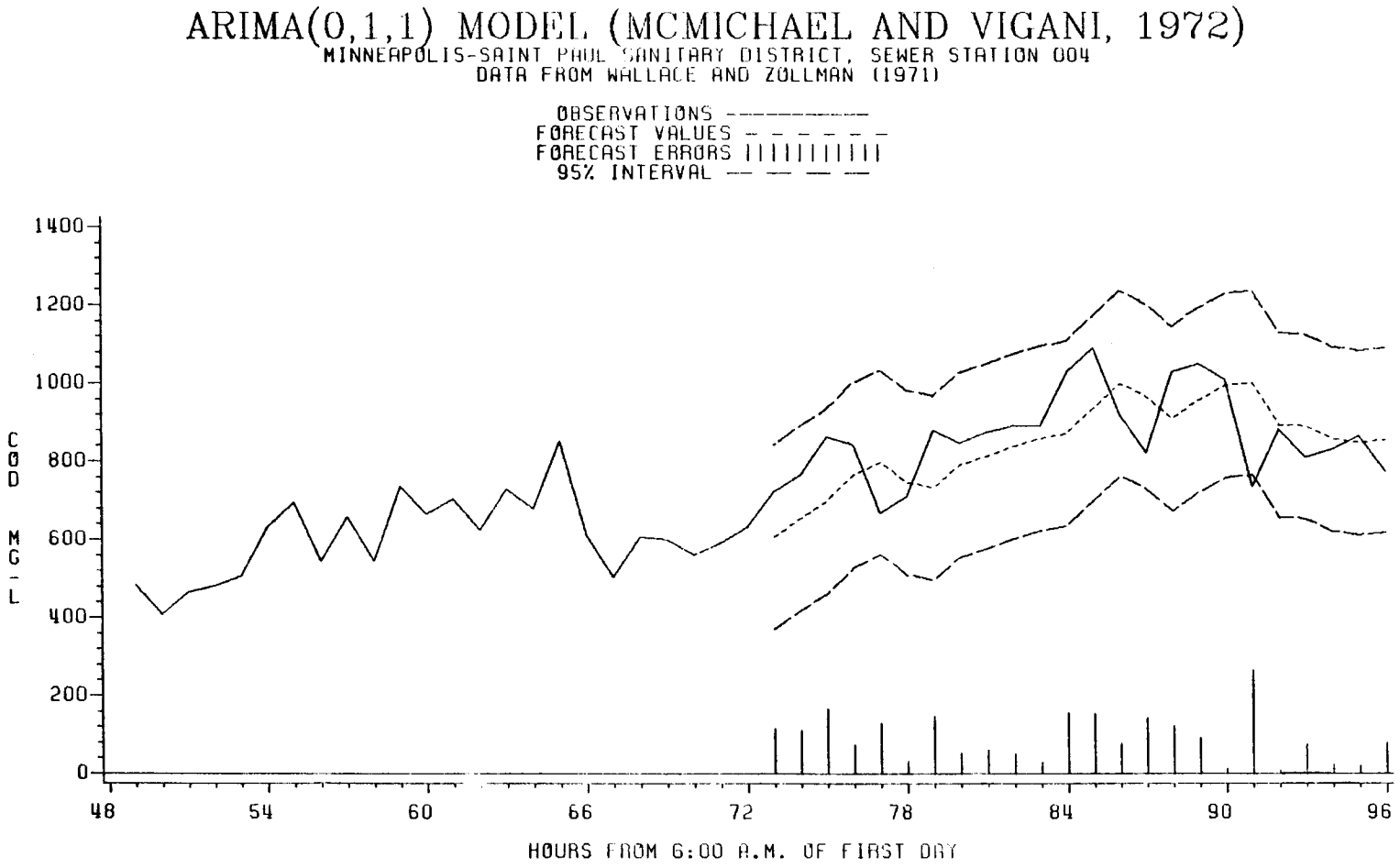


Figure 16: One-Step-Ahead Forecasts from ARIMA(0,1,1) Model

Criterion	Fitted Values (N = 95)	Multi-Step-Ahead Forecasts (N = 24)	One-Step-Ahead Forecasts (N = 24)
R-squared	0.42	---	---
RAE	1 - 509 mg/l	63 - 484 mg/l	11 - 265 mg/l
MAE	88 mg/l	261 mg/l	93 mg/l
RAPE	0 - 94 %	9 - 44 %	1 - 36 %
MAPE	13 %	29 %	11 %
RMSE	120 mg/l	283 mg/l	110 mg/l

TABLE 3

Error Criteria for ARIMA(0,1,1) Model

absolute percent errors averaging only 11%. The maximum absolute error from these forecasts is about as large as the average absolute error for the multi-step-ahead forecasts.

It is clear from the results presented that the ARIMA(0,1,1) model can only be considered useful for one-step-ahead beyond-sample forecasting.

4.5 ATLANTA R. M. CLAYTON TREATMENT PLANT INFLUENT BOD

The third data set analyzed consists of N = 168 hourly measurements of influent BOD from the R. M. Clayton wastewater treatment plant in Atlanta, Georgia. The data were taken from Briggs (1972), and have been previously cited and analyzed using a Fourier series model by Stenstrom (1976). Hour "0" corresponds to midnight, Sunday night and the data cover exactly a one-week period.

4.5.1 Seasonal ARIMA(3,0,0)(0,1,1) Model

The Atlanta BOD series was analyzed using the Box-Jenkins approach. The following subsection describes the modeling steps.

4.5.1.1 Model Identification, Estimation, and Diagnostic Checking

Stationarity of the series was first checked by plotting the data and computing the raw series ACF and PACF. Determination of the total mean square (sample variance) was made by fitting an ARIMA(0,0,0) model to the deviations about the mean (following Berthouex et al., 1975). Large autocorrelations at lags 24, 48, ..., etc. in the ACF indicated the need for a first seasonal differencing of D = 1, s = 24. The seasonal difference was taken, and the ACF and PACF were computed for the differenced series. The ACF showed autoregressive decay at the first few nonseasonal lags, plus a single large, negative spike at lag 24. The PACF showed seasonal moving average decay at lags 24 and 48, plus three significant nonseasonal spikes at lags one, two, and three. This led to a tentative identification of ARIMA(3,0,0)(0,1,1) with s = 24.

When the tentative model was estimated, the nonseasonal AR parameter at lag two was found to be insignificant, and was therefore dropped. The resulting model was then reestimated (see Appendix A), and its parameters were found to be significant, uncorrelated, and within the bounds of stationarity and invertibility.

Proceeding to the diagnostic checking stage, the ACF and PACF of the residuals were found to contain no significant correlations for all lag orders through lag 24. The Ljung-Box chi-square statistic was not significant (5% level) for all lag orders through lag 24, ranging from significance at the 30% to the 46% levels. The ARIMA(3,0,0)(0,1,1) model with zero lag-two AR parameter was therefore deemed acceptable. An attempt was also made to fit an ARIMA(2,0,0)(0,1,1) model because the model containing a nonseasonal AR(3) component with no lag-two term was considered suspect (rare). However, this alternative model did not yield white-noise residuals and was discarded.

4.5.1.2 Fitted Values

Figure 17 shows the BOD values, fitted values, and residuals from the ARIMA(3,0,0)(0,1,1) model fitted to all N = 168 observations. Many of the fitted residuals exceed 50 mg/l BOD. Again, the largest residuals occur at the turning points, and a distinct one-interval lag effect is observed. Figure 18 provides a magnified view of the last 48 hours' fitted results.

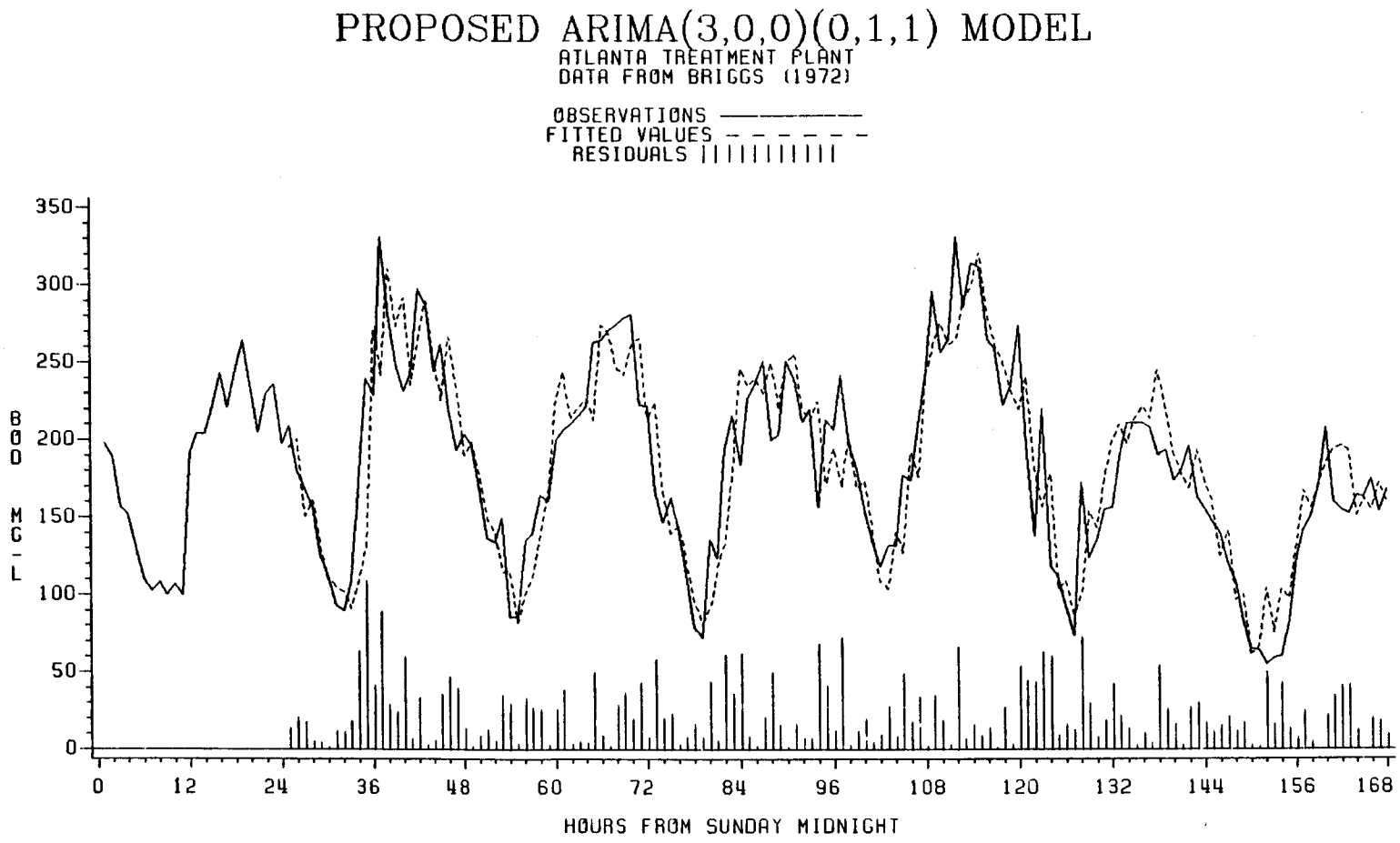


Figure 17: Fitted Values and Residuals from Atlanta BOD Model

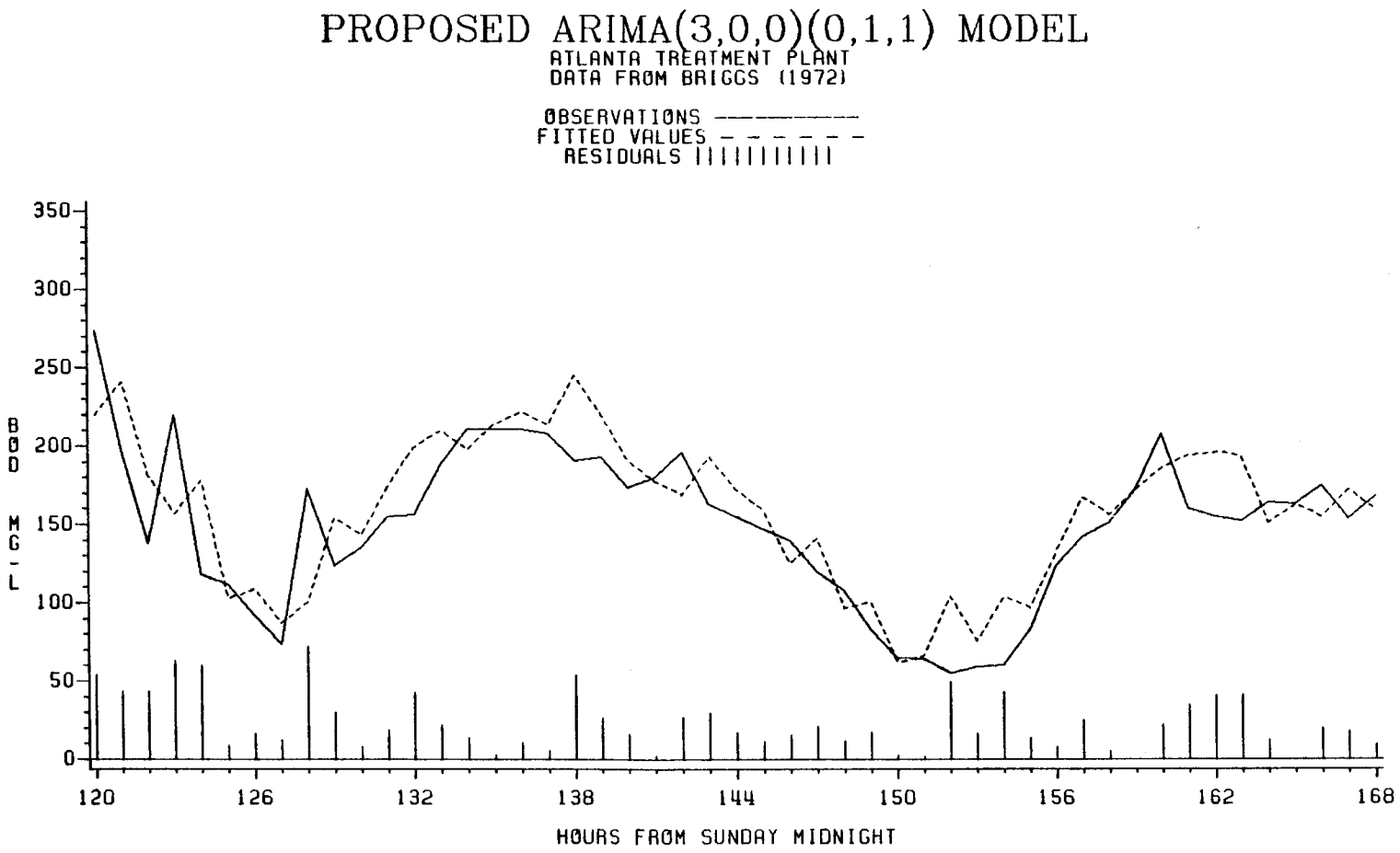


Figure 18: Detail of Fitting Error from Atlanta BOD Model

4.5.1.3 Multi-Step-Ahead Beyond-Sample Forecasts

Figure 19 shows 24 multi-step-ahead beyond-sample forecasts from the ARIMA(3,0,0)(0,1,1) model fitted to the first N = 144 observations (see Appendix A). The 24 hours (hour 145 to hour 168) correspond to Sunday, the last day for which measurements were taken. It can be seen that the forecasts consistently overestimate the actual BOD profile that occurred. Since the behavior of the multi-step-ahead beyond-sample forecasts is based only on information from Monday through Saturday, which all had higher minima than Sunday, the model cannot be expected to anticipate the Sunday cycle. The general diurnal cycle is reproduced by the seasonal model, however.

4.5.1.4 One-Step-Ahead Beyond-Sample Forecasts

Figure 20 shows 24 one-step-ahead beyond-sample forecasts generated by the ARIMA(3,0,0)(0,1,1) model fitted to the first N = 144 data. Each forecast is based on data from one hour, three hours, 24 hours, 25 hours, and 27 hours into the "past" (see Appendix A). It is seen that the forecasts track the observed data fairly well, but still tend to overestimate the BOD concentration at most hours. The forecast errors are greatly reduced compared to the multi-step-ahead beyond-sample forecast errors.

4.5.1.5 Fitting and Forecasting Accuracy

Table 4 shows the error criteria computed for the fitting and forecasting results. The following general observations can be made:

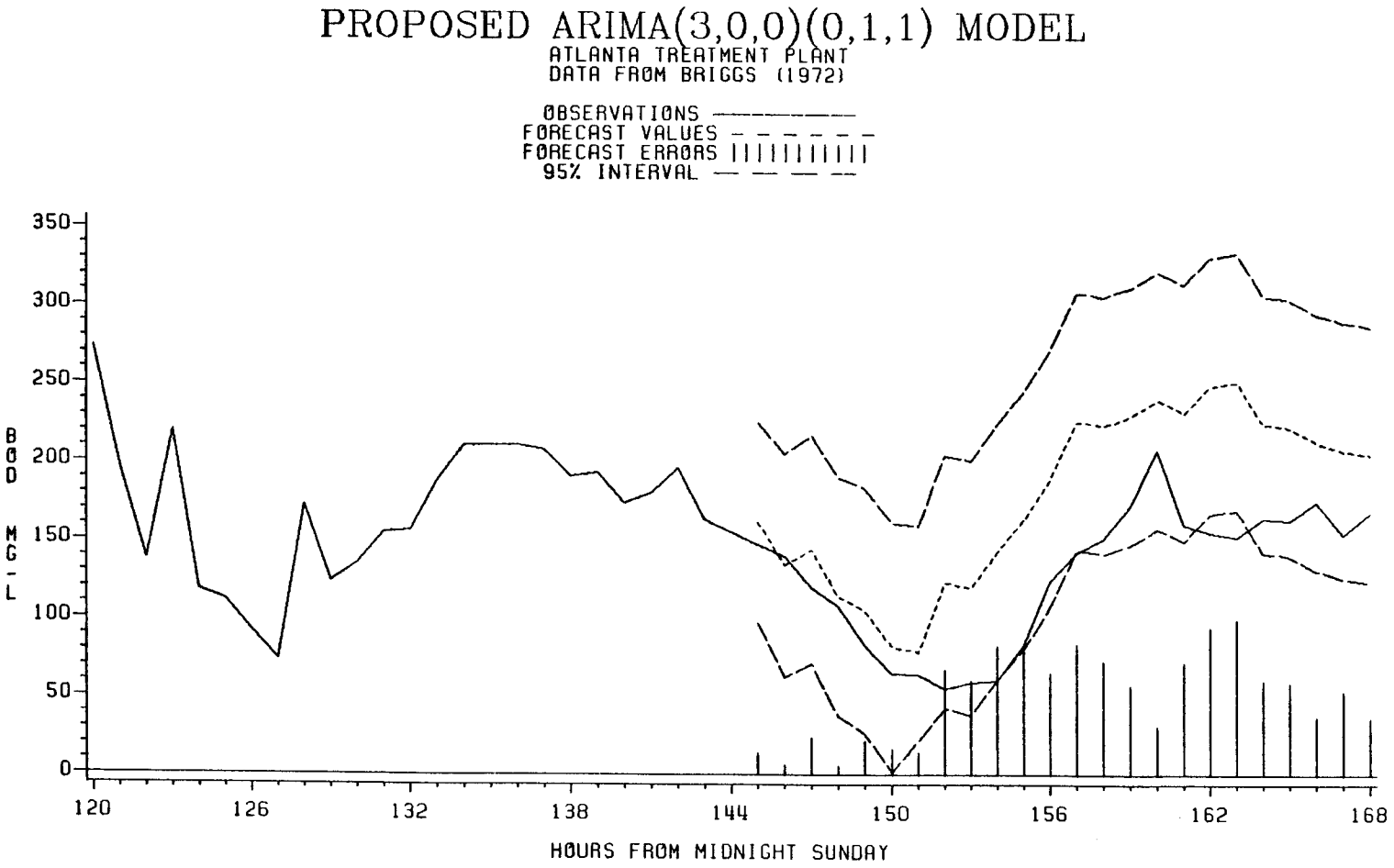


Figure 19: Multi-Step-Ahead Forecasts from Atlanta BOD Model

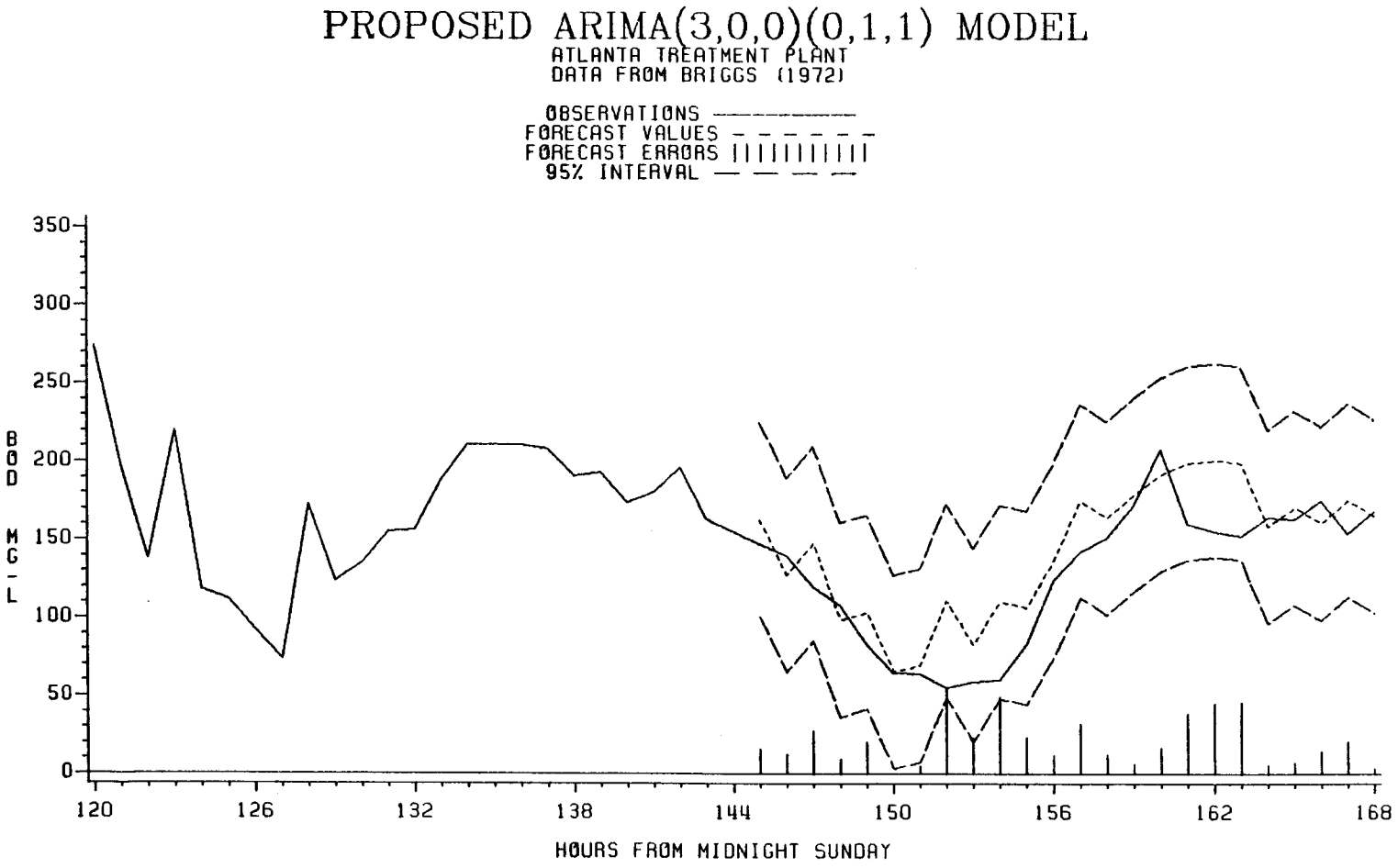


Figure 20: One-Step-Ahead Forecasts from Atlanta BOD Model

Criterion	Fitted Values (N = 144)	Multi-Step-Ahead Forecasts (N = 24)	One-Step-Ahead Forecasts (N = 24)
R-squared	0.74	---	---
RAE	0 - 109 mg/l	5 - 100 mg/l	0 - 56 mg/l
MAE	23 mg/l	51 mg/l	21 mg/l
RAPE	0 - 89 %	4 - 137 %	0 - 101 %
MAPE	14 %	46 %	21 %
RMSE	31 mg/l	58 mg/l	26 mg/l

TABLE 4

Error Criteria for ARIMA(3,0,0)(0,1,1) Model

1. This model has a higher R-squared than previous models, accounting for 74% of the observed variation about the mean.
2. Because the multi-step-ahead beyond-sample forecasts overestimated the observed BOD values, the MAE and the RMSE were greater than 50 mg/l BOD, which are large for data ranging from 55 to 208 mg/l. The MAPE is also larger than was found for previous models.
3. Once again, one-step-ahead beyond-sample forecasting reduced the overall forecast error approximately twofold as measured by MAE, MAPE, or RMSE.

4.6 ATLANTA R. M. CLAYTON TREATMENT PLANT TOTAL SUSPENDED SOLIDS

The fourth data set analyzed consists of N = 168 hourly measurements of influent total suspended solids (TSS) from the R. M. Clayton wastewater treatment plant in Atlanta, Georgia. The data were also taken from Briggs (1972), and were previously cited and analyzed using a Fourier series model by Stenstrom (1976). Hour "0" corresponds to midnight, Sunday night and the data cover exactly a one-week period.

4.6.1 Seasonal ARIMA(2,0,0)(0,1,1) Model

The Atlanta TSS series was analyzed using the Box-Jenkins strategy. The following subsection outlines the modeling steps performed.

4.6.1.1 Model Identification, Estimation, and Diagnostic Checking

Stationarity of the series was investigated by plotting the data and computing the raw series ACF and PACF. Determination of the total mean square (sample variance) was also made. Large autocorrelations at the seasonal lags in the ACF indicated the need for a first seasonal differencing of D = 1, s = 24, after which the ACF and PACF were computed for the resulting differenced series. The ACF suggested autoregressive decay at the lower nonseasonal lags, and had a single large, negative spike at lag 24. The PACF showed seasonal moving average decay at lags 24 and 48, plus two significant nonseasonal spikes at lags one and two. This implied a preliminary identification of ARIMA(2,0,0)(0,1,1) with s = 24.

When this model was estimated (see Appendix A), its parameters were found to be significant, uncorrelated, and within the bounds of stationarity and invertibility.

At the diagnostic checking stage, the ACF of the residuals was found to contain no significant correlations for all lag orders through lag 24. The Ljung-Box chi-square statistic was not significant (5% level) for all lag orders through lag 24, ranging from significance at the 6% to the 19% levels. The PACF showed a single "borderline" significant spike at lag 14; this was considered to be a spurious correlation at a physically meaningless lag order (14 hours). It was felt that one stray spike out of 24 might be expected from sampling error alone for a 95% confidence interval test. The ARIMA(2,0,0)(0,1,1) model was therefore tentatively accepted as adequate.

4.6.1.2 Fitted Values

Figure 21 shows the TSS values, fitted values, and residuals from the ARIMA(2,0,0)(0,1,1) model fitted to all N = 168 observations. This figure displays several interesting features. The observations fluctuate erratically about a general diurnal cycle. The peak values of TSS appear to decrease steadily throughout the week. Of greater interest, however, are the peak values of the forecasts, particularly at hours 37, 61, and 85. These peaks are separated by exactly 24 hours. Due to the nature of the forecast function and the specific parameter estimates for the ARIMA(2,0,0)(0,1,1) model (see Appendix

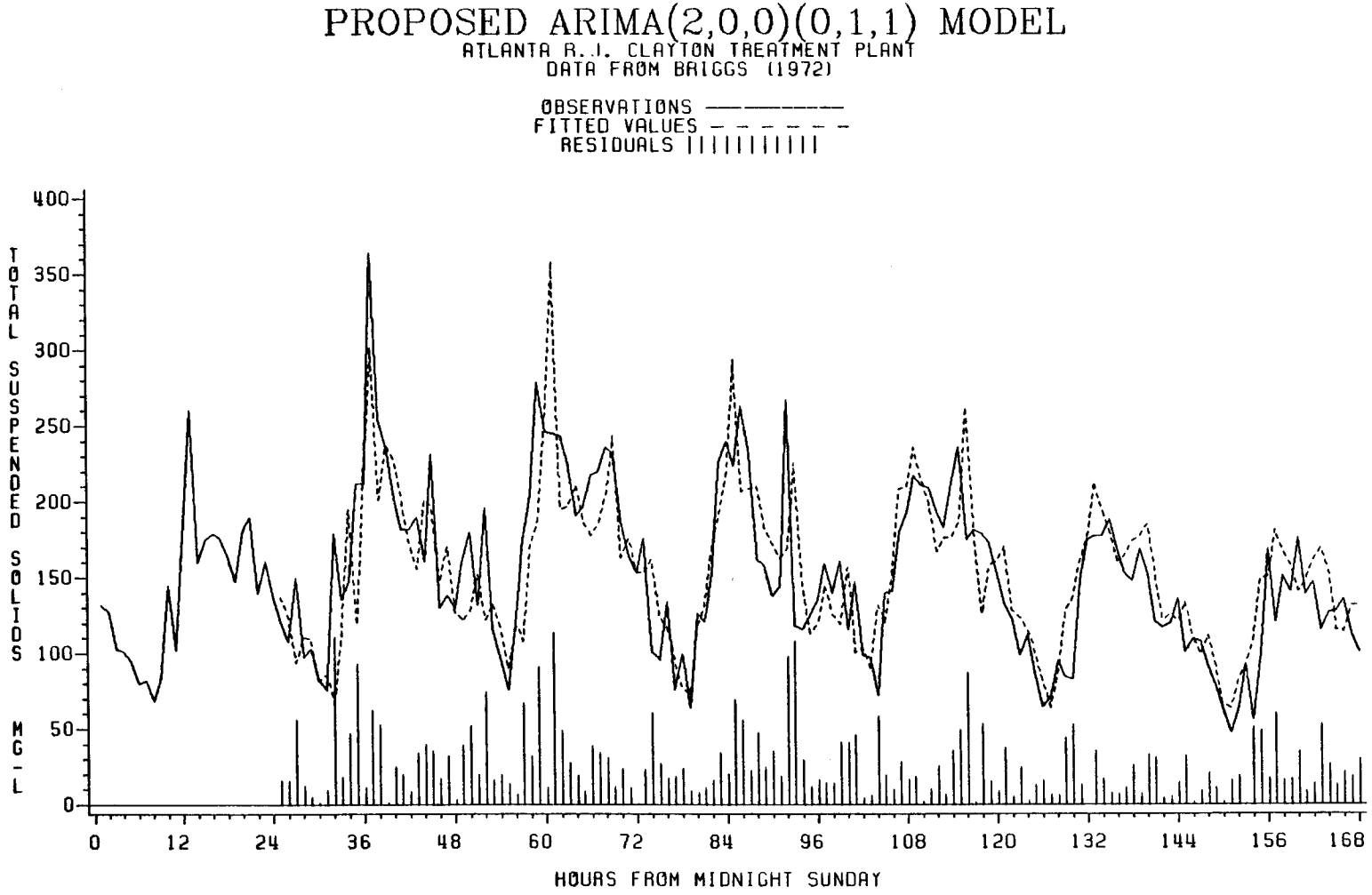


Figure 21: Fitted Values and Residuals from Atlanta TSS Model

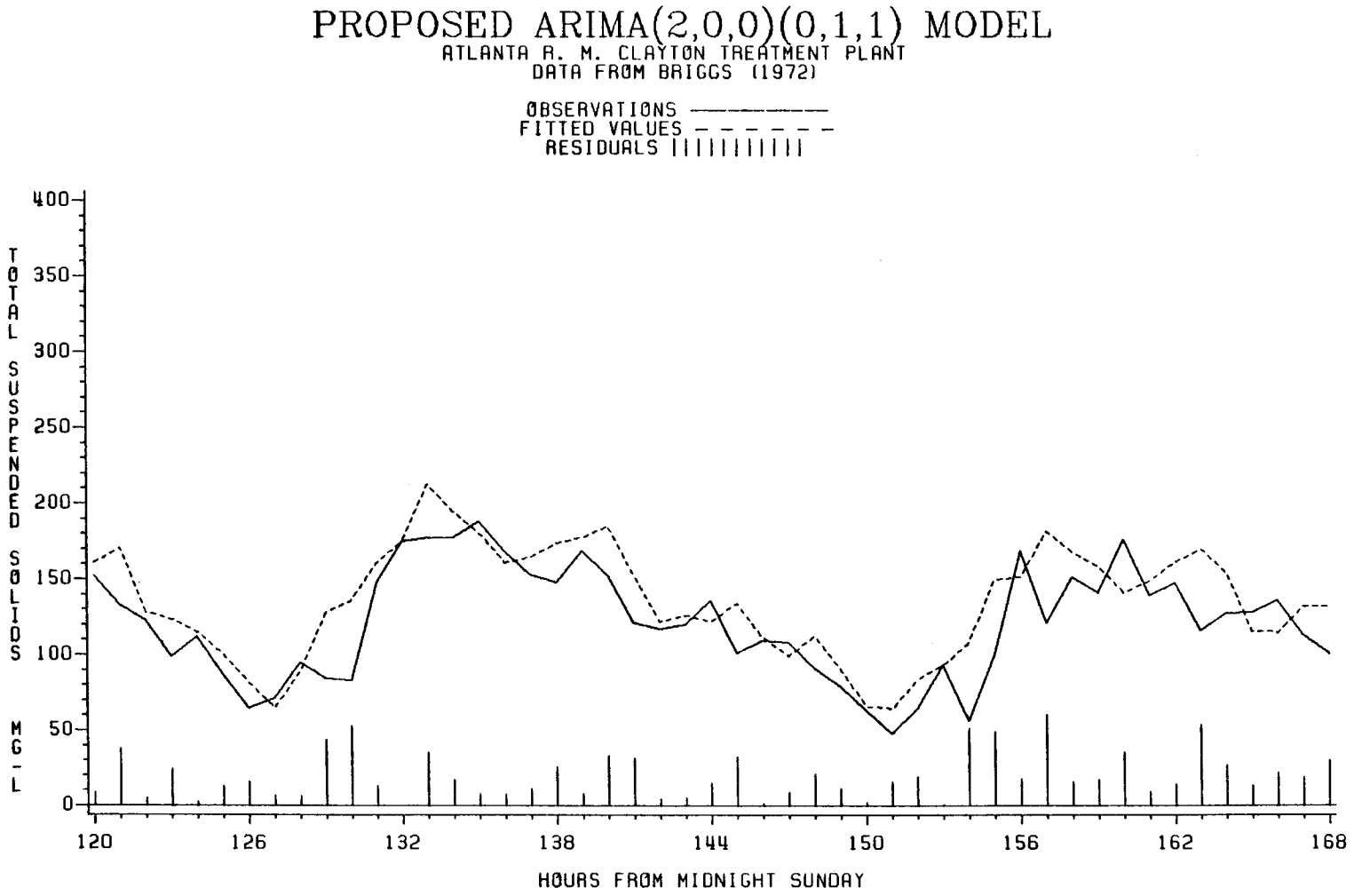


Figure 22: Detail of Fitting Error from Atlanta TSS Model

A), the forecast for a given hour is made up largely of the observed value from 24 hours before. Therefore, the one-step-ahead within-sample forecast (fitted value) occurring 24 hours after an extreme observed peak will also form a large peak. This explains the large fitted value at hour 61. Another large component of each forecast is the fitted residual from 24 hours before. Since the weighting parameter for this fitted residual is negative (see the seasonal moving average parameter in Appendix A), and since the fitted residual at hour 61 is large and negative, this results in a large contribution to the forecast for hour 85.

Figure 22 shows a magnified view of the last 48 hours' fitted results.

4.6.1.3 Multi-Step-Ahead Beyond-Sample Forecasts

Figure 23 shows 24 multi-step-ahead beyond-sample forecasts from the ARIMA(2,0,0)(0,1,1) model fitted to the first N = 144 observations (see Appendix A). The forecasts based on the previous days' cycles overestimate the Sunday cycle, as seen before in the preceding data set.

4.6.1.4 One-Step-Ahead Beyond-Sample Forecasts

Figure 24 shows 24 one-step-ahead beyond-sample forecasts generated by the ARIMA(2,0,0)(0,1,1) model fitted to the first N = 144 data. These forecasts appear to follow the observed data fairly well, but still tend to overestimate the TSS concentration at most hours. The

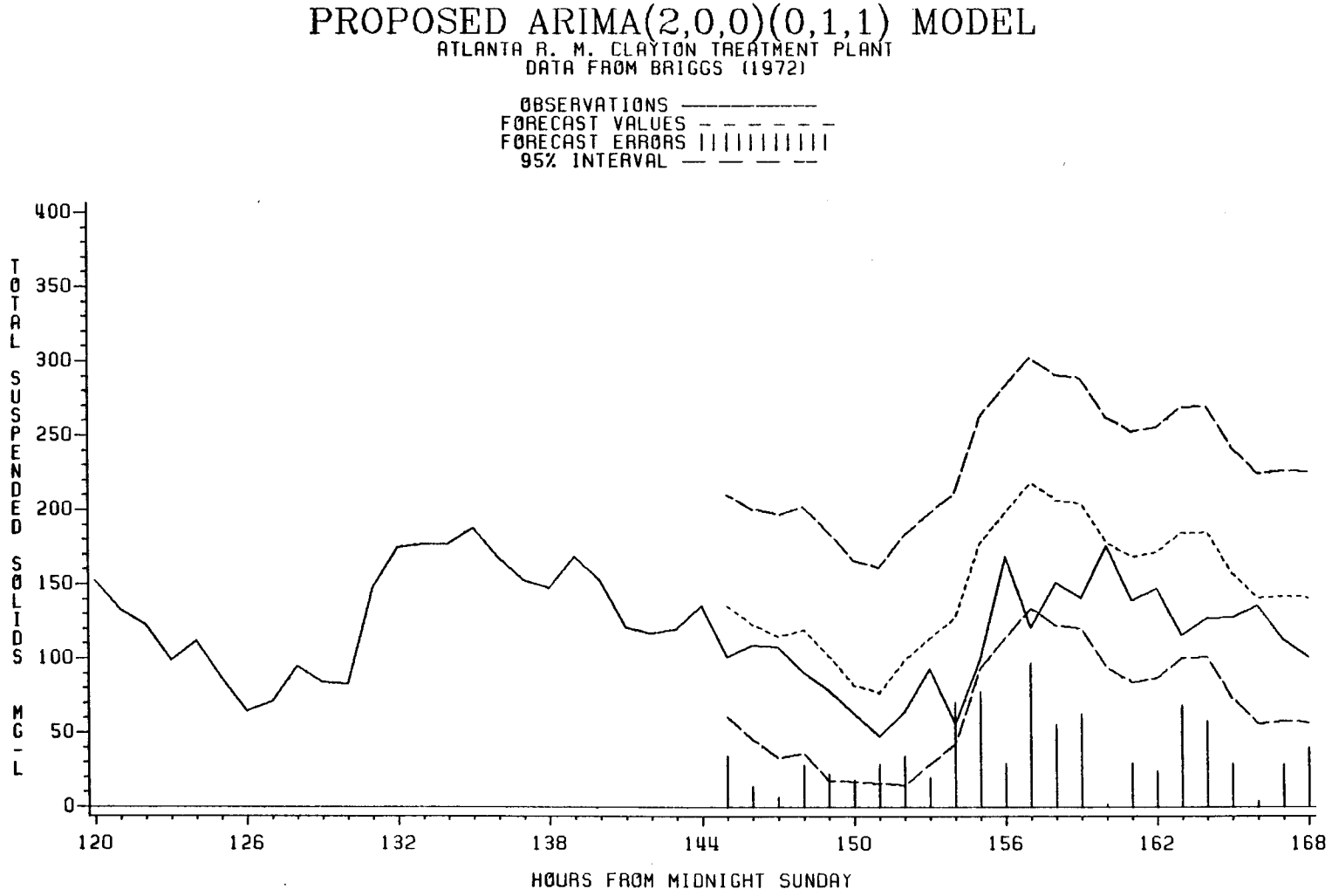


Figure 23: Multi-Step-Ahead Forecasts from Atlanta TSS Model

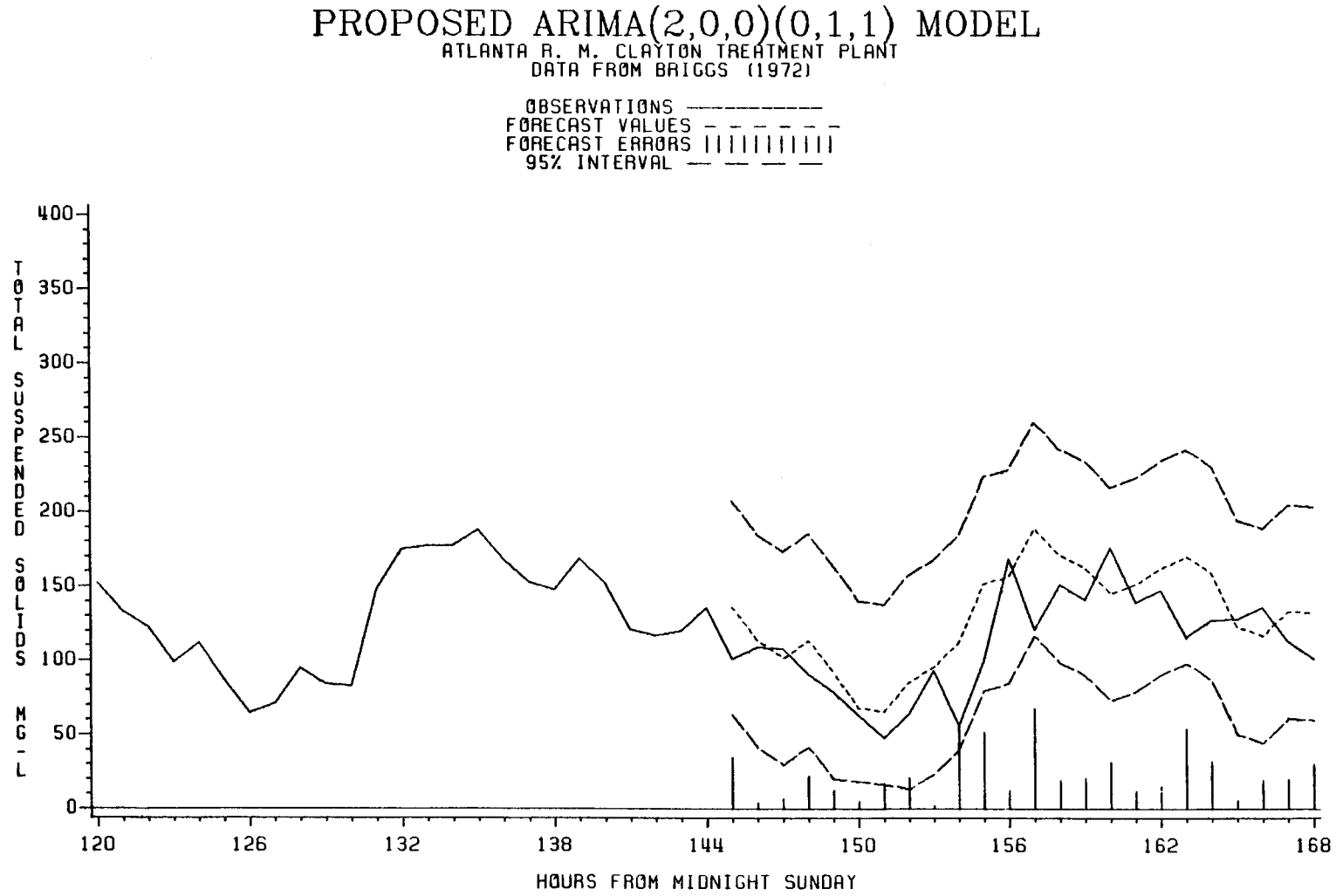


Figure 24: One-Step-Ahead Forecasts from Atlanta TSS Model

forecast errors are only slightly reduced compared to the multi-step-ahead beyond-sample forecast errors. The confidence intervals are also approximately equal in width for both types of forecasts.

4.6.1.5 Fitting and Forecasting Accuracy

Criterion	Fitted Values (N = 144)	Multi-Step-Ahead Forecasts (N = 24)	One-Step-Ahead Forecasts (N = 24)
R-squared	0.53	---	---
RAE	0 - 114 mg/l	2 - 97 mg/l	3 - 67 mg/l
MAE	27 mg/l	37 mg/l	24 mg/l
RAPE	0 - 92 %	1 - 127 %	3 - 100 %
MAPE	19 %	37 %	24 %
RMSE	36 mg/l	44 mg/l	30 mg/l

TABLE 5

Error Criteria for ARIMA(2,0,0)(0,1,1) Model

Table 5 shows the error criteria computed for the fitting and forecasting results. The following points emerge from examining the table:

1. This model has a poor R-squared value (0.53), in accordance with the appearance of the fitted residuals.

2. Although one-step-ahead beyond-sample forecasting reduces the overall forecast error as measured by MAE, MAPE, or RMSE, the improvement is not as notable as it was for the better-fitting models in this study.

4.7 MINNEAPOLIS-SAINT PAUL INTERCEPTOR SEWER FLOW RATE

The fifth data set analyzed consists of N = 168 hourly averages of wastewater flow rate measured in an interceptor sewer of the Minneapolis-Saint Paul Sanitary District's combined sewer system. The data were taken from Anderson (1973), and have been previously cited and analyzed using a Fourier series model by Stenstrom (1976). Hour "0" corresponds to midnight, Sunday night and the data cover exactly a one-week period.

4.7.1 Seasonal ARIMA(0,1,0)(0,1,1) Model

The Minneapolis-Saint Paul flow rate series was analyzed using the procedures of Box and Jenkins. The following subsection summarizes the steps performed in modeling the series.

4.7.1.1 Model Identification, Estimation, and Diagnostic Checking

Stationarity of the series was first checked by plotting the data and computing the raw series ACF and PACF. Determination of the total mean square (sample variance) of the series was also made. Large autocorrelations at lags 24, 48, ..., etc. in the ACF indicated the need for a first seasonal differencing of D = 1, s = 24. The

resulting seasonally differenced series was then used to compute a new ACF and PACF. Slow, positive decay at the nonseasonal lags in the ACF clearly indicated the need for an additional nonseasonal first differencing (d = 1). The resulting twice-differenced (D = 1, d = 1) series was then used to calculate another ACF and PACF. The ACF showed a single large, negative spike at lag 24, and the PACF had a large, negative spike at lag 24 plus another small, negative spike at lag 48. This suggested a seasonal, first-order moving average component; i.e., ARIMA(0,1,0)(0,1,1) with s = 24.

This tentative model was estimated (see Appendix A); the MA parameter estimate was found to be significant and to obey the invertibility condition.

Proceeding to the diagnostic checking stage, the ACF and PACF of the residuals were found to contain no significant correlations for all lag orders through lag 24. The chi-square statistic was not significant (5% level) for all lag orders through lag 24, ranging from significance at the 52% to the 73% levels. The ARIMA(0,1,0)(0,1,1) model was therefore tentatively entertained.

4.7.1.2 Fitted Values

Figure 25 shows the observed flow, fitted values, and residuals from the ARIMA(0,1,0)(0,1,1) model fitted to all N = 168 observations. The most striking feature of this data set is the lack of erratic variation in the flow rate curve. This is due to the large volume of flow (mean = 180.4 MGD), which masks any disturbances due to sporadic

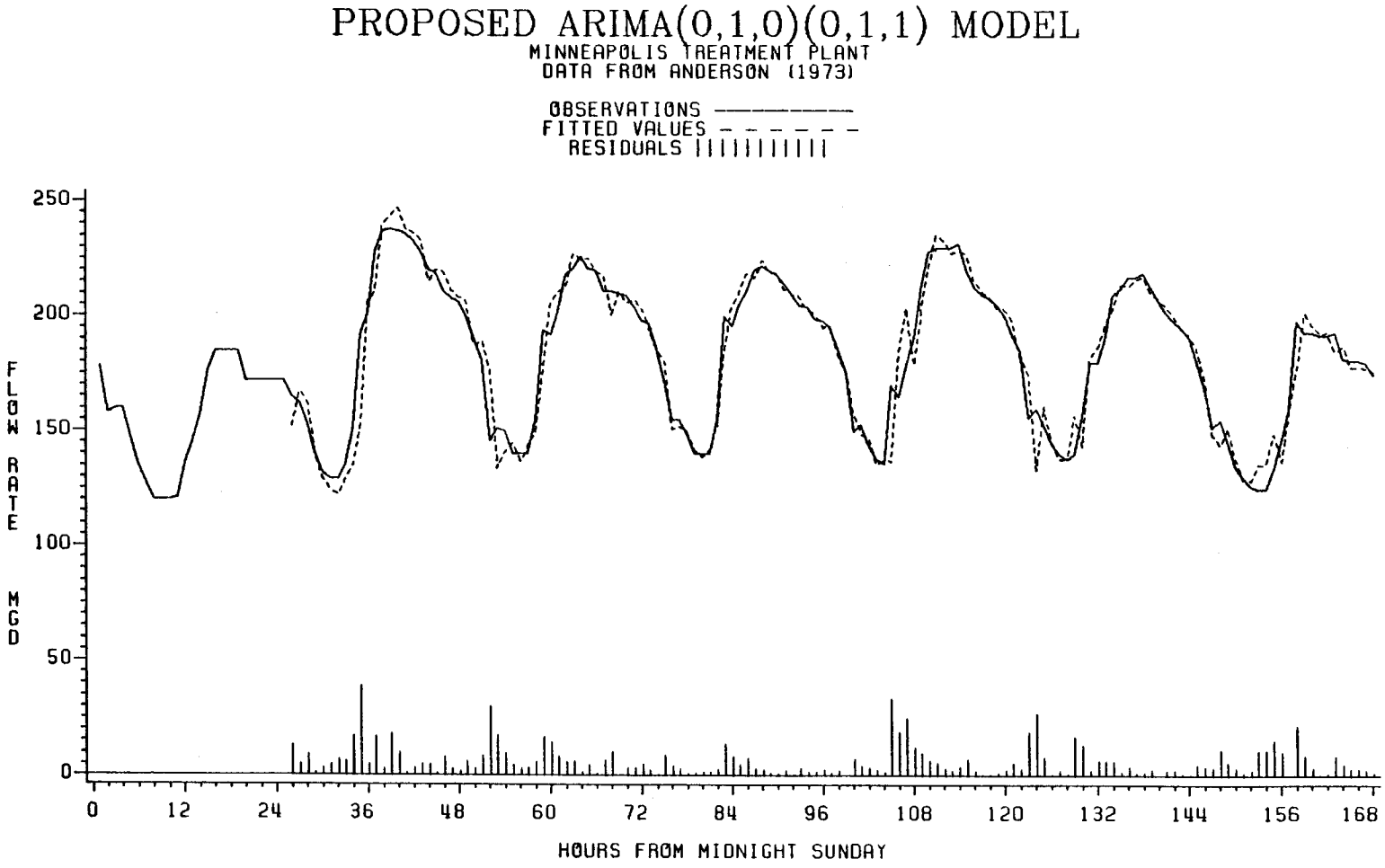


Figure 25: Fitted Values and Residuals from Flow Rate Model

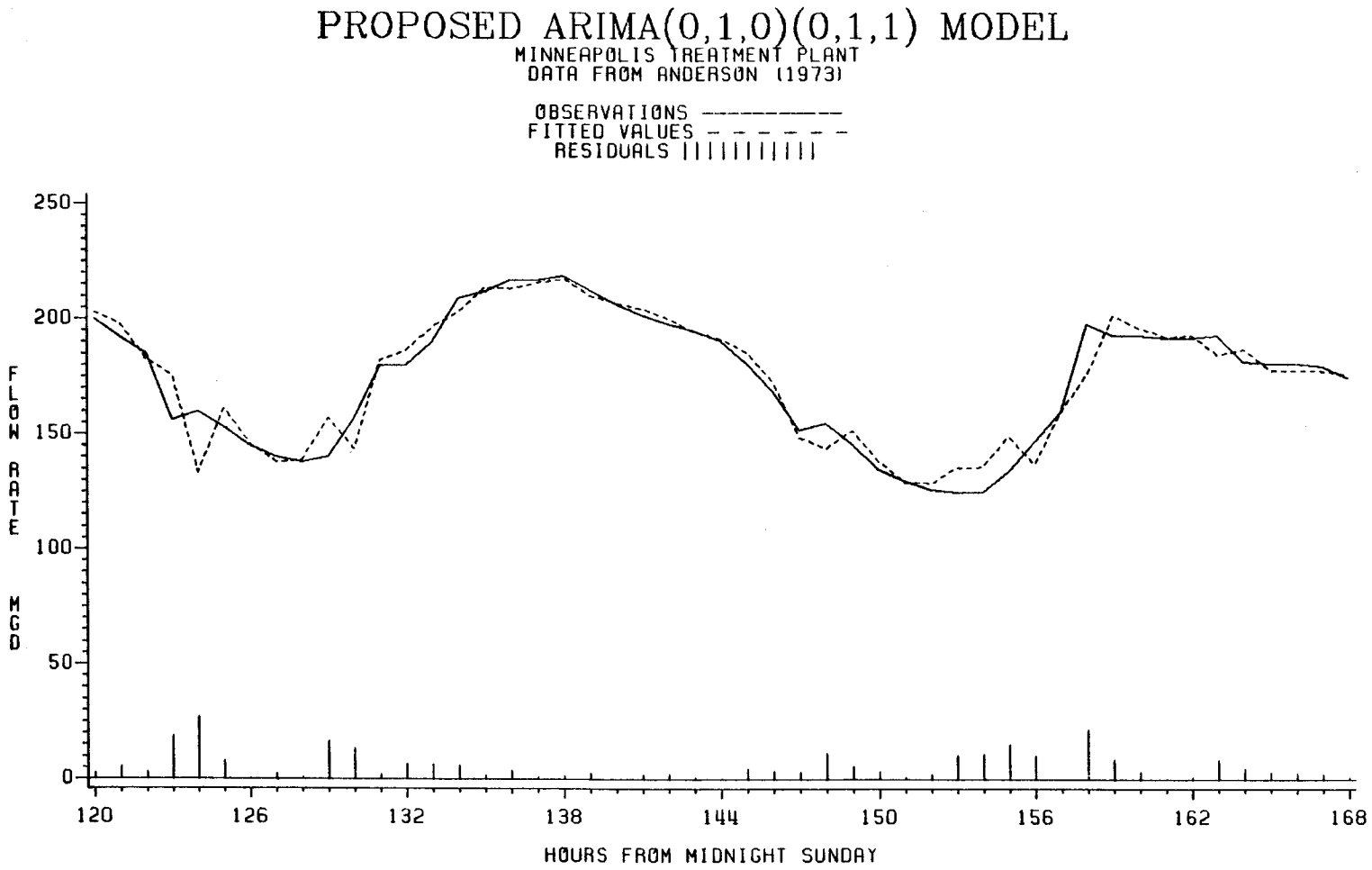


Figure 26: Detail of Fitting Error From Flow Rate Model

municipal or industrial batch discharges. There is also less experimental error in flow rate measurements than in chemical lab analyses such as BOD or COD tests. As a result, the "noise" is reduced and a better model fit is obtained, as indicated by the relatively small absolute residuals.

Figure 26 provides a magnified view of the fitting errors for the last 48 observations in the series. The largest residuals occur at the turning points, and the one-interval lag effect can be discerned at several of the relative maxima.

4.7.1.3 Multi-Step-Ahead Beyond-Sample Forecasts

Figure 27 shows 24 multi-step-ahead beyond-sample forecasts from the ARIMA(0,1,0)(0,1,1) model fitted to the first N = 144 observations (see Appendix A). Although the forecast profile reproduces the general diurnal cycle, the forecasts overestimate the flow rates at most intervals. Since the 24 hours (hour 145 to hour 168) shown fall on a Sunday, the actual flow rate is lower than the forecasts based on previous larger cycles.

4.7.1.4 One-Step-Ahead Beyond-Sample Forecasts

Figure 28 shows 24 one-step-ahead beyond-sample forecasts generated by the ARIMA(0,1,0)(0,1,1) model fitted to the first N = 144 data points. The confidence intervals shown are the smallest ones obtained in this study, reflecting the reduced forecast error variance implied by the quality of fit of the proposed model. The one-interval lag

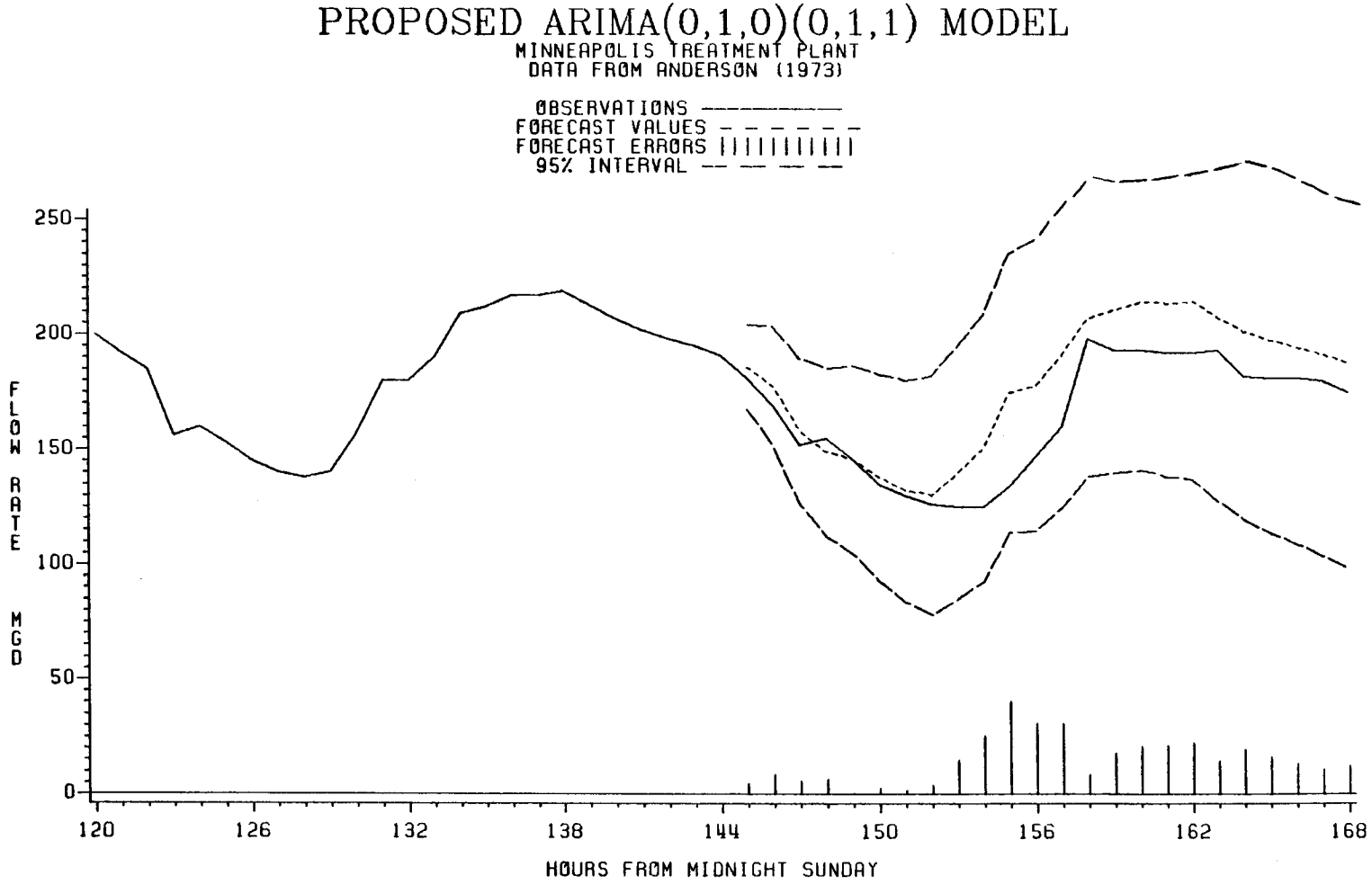


Figure 27: Multi-Step-Ahead Forecasts from Flow Rate Model

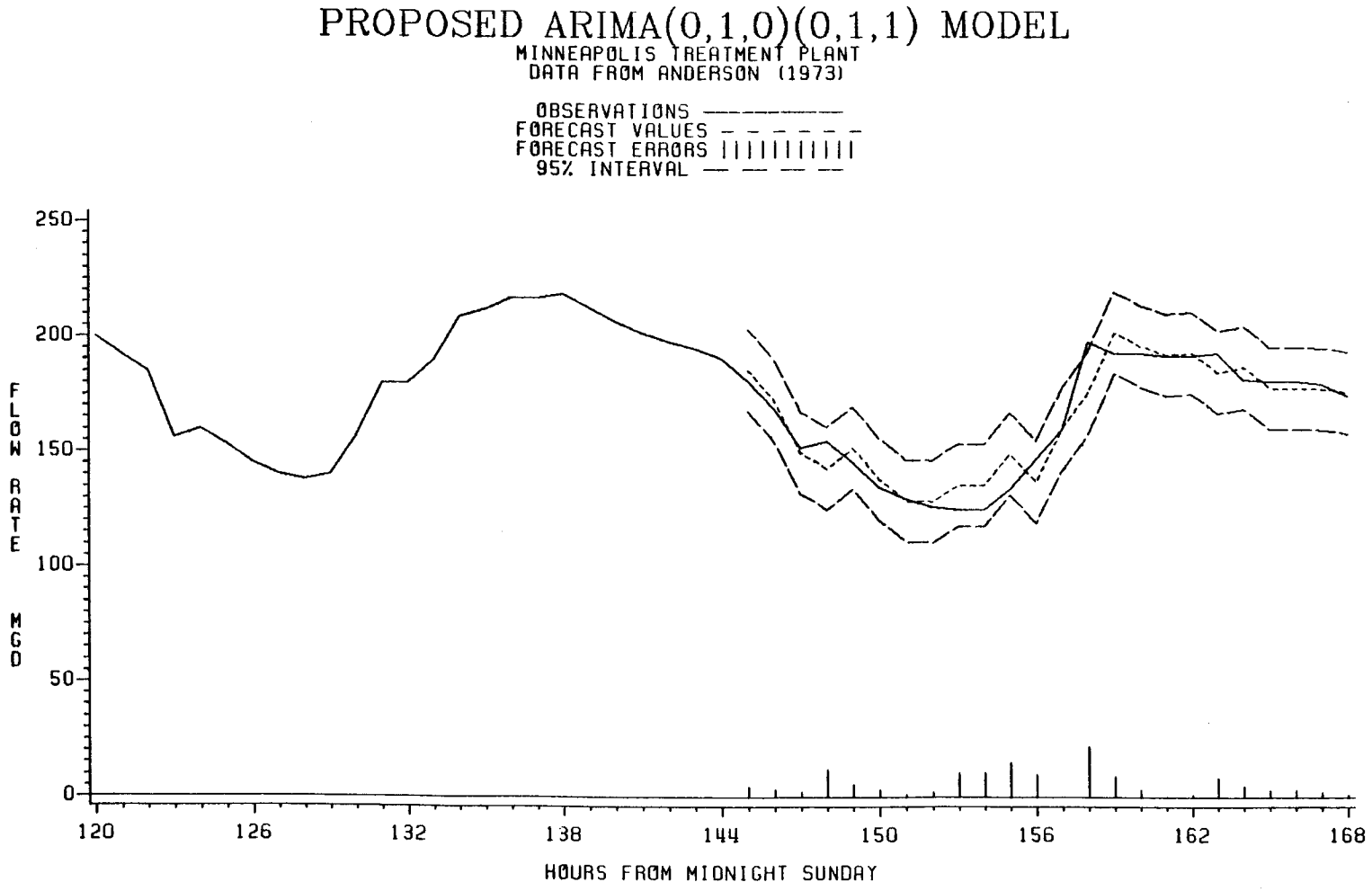


Figure 28: One-Step-Ahead Forecasts from Flow Rate Model

effect is present at the relative maxima and minima, but the resulting forecast errors are not excessive because the variability at these turning points is relatively gradual.

4.7.1.5 Fitting and Forecasting Accuracy

Criterion	Fitted Values (N = 143)	Multi-Step-Ahead Forecasts (N = 24)	One-Step-Ahead Forecasts (N = 24)
R-squared	0.92	---	---
RAE	0.02 - 39.0 MGD	0.52 - 40.6 MGD	0.09 - 22.1 MGD
MAE	6.09 MGD	14.7 MGD	5.83 MGD
RAPE	0 - 21 %	0 - 30 %	0 - 11 %
MAPE	4 %	9 %	4 %
RMSE	9.07 MGD	17.9 MGD	7.88 MGD

TABLE 6

Error Criteria for ARIMA(0,1,0)(0,1,1) Model

Table 6 gives the values of the fitting and forecasting error criteria calculated for the proposed flow rate model. Several interesting points emerge from reviewing this table:

1. This data set and model achieved the highest R-squared value (0.92) in this study.

2. Because the multi-step-ahead beyond-sample forecasts overestimated the true flow rates, one-step-ahead beyond-sample forecasting provided a substantial improvement in accuracy as measured by any of the error criteria.
3. The MAPE's of 9% AND 4% for the multi-step-ahead forecasts and one-step-ahead forecasts respectively are the lowest average percent errors obtained for any model in this study.
4. Considering one-step-ahead beyond-sample forecasting, the maximum absolute error of 22.1 MGD represents a flow volume of 920,833 gallons for a one-hour period. This is a large amount of wastewater to be reckoned with if such a forecasting error should occur in practice. Similarly, for multi-step-ahead forecasting the maximum absolute error of 40.6 MGD constitutes a one-hour volume of approximately 1.7 million gallons.

This last point is raised in order to illustrate the fact that it is not always sufficient to consider only the percent error in forecasting. Another useful measure of forecasting error might be the absolute error divided by the mean of the observations--in this context, the fraction of the average flow by which the forecast was in error.

Chapter V

SIMULATION OF REAL-TIME CONTROL WITH ARIMA FLOW FORECASTING

"The question ultimately is whether we have developed a mildly interesting academic curiosity or a potentially useful management tool."

W. O. Spofford, Jr.

5.1 INTRODUCTION

Over the last decade, there has been increasing interest in dynamic modeling and computer-compatible control strategies for activated sludge wastewater treatment plants. It is not the intention of this discussion to review historical developments or to reiterate the benefits of real-time process control for treatment plants; these have been summarized by Stenstrom (1976) and Stenstrom and Andrews (1979), among others. It will suffice to note that there are two ways in which research on control strategies for treatment plants can be carried out: by direct field experimentation, or by using dynamic mathematical models which simulate the performance and operation of treatment plants. The latter option, mathematical simulation, allows candidate control strategies to be studied without requiring actual plant interruption, costly experiments, and large investments of time and personnel. Ultimately, of course, experimentation must always be

used to verify modeling results. Experimentation and modeling can be utilized in a complementary, iterative way (Stenstrom and Andrews, 1979).

As discussed in the literature review, the motivation for the present research arose from the work of Stenstrom (1976) and Stenstrom and Andrews (1979), who utilized a dynamic, deterministic computer simulation model of a wastewater treatment plant to evaluate various control strategies. They found that control strategies which incorporated short-term predictions of influent flow rate as inputs to the control law greatly reduced treatment process variability as measured by specific oxygen uptake rate (SCOUR). Their predictions were made using a deterministic Fourier series model. Stenstrom and Andrews noted that ARIMA models also showed promise for performing such forecasting. The present thesis was undertaken in direct response to this comment. The following sections describe results from a simulated real-time control strategy using an ARIMA model for hourly flow rate prediction in conjunction with the dynamic simulation model utilized by Stenstrom and Andrews.

5.2 METHODOLOGY

The dynamic mathematical model has been described in detail elsewhere (Stenstrom, 1976); only a brief description is given here. The portions of Stenstrom's model used for this investigation included process models for the primary clarifier, aeration basin, and secondary clarifier, and Fourier series input models for influent BOD,

total suspended solids (TSS), and ammonia nitrogen. The influent flow rate sequence consisted of hourly average flows measured in a field survey (Anderson, 1973). For this analysis, the flow rate was scaled by dividing by 1/10th of the mean of the observed series, yielding a flow series with a mean of 10 MGD. The hydraulic retention time of the aeration basin was 4.3 hours, and a solids-liquid separator with an area of 12,500 square feet was simulated; this allowed mean cell retention times (MCRT) of up to 10 days to be attained. The biological reactor was formulated so as to simulate a "four-pass" aeration basin with step-feed modification.

Stenstrom and Andrews (1979) demonstrated the superiority of SCOUR over the traditional MCRT or food-to-mass ratio (F/M) as a dynamic control variable, pointing out that MCRT and F/M are founded in steady-state assumptions. Only SCOUR retains a direct relationship to the fundamental biological parameter, organism growth rate, under nonsteady-state conditions. Hence, maintaining constant SCOUR implies that growth rate is maintained constant, a desirable objective for decreasing process variability. In this simulation study, the dimensionless variance of SCOUR (Stenstrom, 1976; Stenstrom and Andrews, 1979) was selected as the performance criterion for evaluating the improvement provided by a real-time control strategy with ARIMA flow prediction. The dimensionless SCOUR variance is defined as the variance of SCOUR divided by the squared mean SCOUR over a given period of plant operation.

A two-loop control strategy was adopted for the simulation. SCOUR was used as the controlled variable for the fast loop, and MCRT served as the controlled variable in the slower loop. SCOUR was controlled by manipulating the sludge recycle rate; MCRT was controlled by manipulating the sludge wasting rate. Sludge storage was provided by directing the influent feed to the second segment of the four-pass aeration basin. This control strategy was called "pseudo feedforward-feedback" control by Stenstrom (1976). Further details concerning the control strategy and control law are given in Stenstrom (1976) and Stenstrom and Andrews (1979) and are not central to the present study. The result of interest is the comparison of dimensionless SCOUR variance from three different simulation runs:

1. Operation of the treatment plant with no control strategy ("No Control" case).
2. Operation of the treatment plant with the pseudo feedforward-feedback control strategy ("Control Without Prediction" case).
3. Operation of the treatment plant with the pseudo feedforward-feedback control strategy incorporating hourly one-step-ahead beyond-sample flow rate forecasts as inputs to the controller ("Control With Prediction" case).

5.3 DESCRIPTION OF THE SIMULATION RUNS

For all three simulation runs, the dynamic model was run using Fourier series model values as inputs for the influent BOD, ammonia nitrogen, and TSS, and scaled flow rate measurements as inputs for the influent flow rate. The flow rate series used was a portion of the one analyzed and discussed in Chapter 4 in the section entitled "Minneapolis-Saint Paul Interceptor Sewer Flow Rate." In each case, the dynamic model was run for 168 hours of simulated operation, sufficient time to eliminate the transient effects of the initial conditions. Subsequently, for the next 24-hour period (hour 169 to hour 192), the dimensionless variance of SCOUR was computed.

For the "No Control" case, no control was implemented throughout the entire 192 hours of operation. For the "Control Without Prediction" case, the control strategy was implemented for all 192 hours. Lastly, for the case of greatest interest, "Control With Prediction", control was implemented incorporating hourly one-step-ahead forecasts generated from the ARIMA(0,1,0)(0,1,1) flow rate model described in Chapter 4. For the 24-hour period during which SCOUR variance was computed, the influent flow rate values and corresponding forecasts were exactly those shown in Figure 28 of Chapter 4 (with scaling to obtain a mean of 10 MGD). Computed values of the accuracy criteria for the forecasts are given in Table 6 of Chapter 4.

5.4 RESULTS

The three simulation runs were performed, and three corresponding values of dimensionless SCOUR variance were computed. Each value was computed for the last 24 hours of each simulation run. Figure 29 shows the controlled value of SCOUR over 24 hours for each of the three cases. Table 7 gives the values of dimensionless SCOUR variance obtained from each of the three cases, and the corresponding reduction in variance for the "Control Without Prediction" and "Control With Prediction" cases. It is apparent from the results that implementation of the control strategy greatly reduced the variability of SCOUR (by 77%). It is also observed that the control strategy incorporating ARIMA forecasts reduced SCOUR variability by 83%, an additional 6% reduction beyond the case of control without prediction.

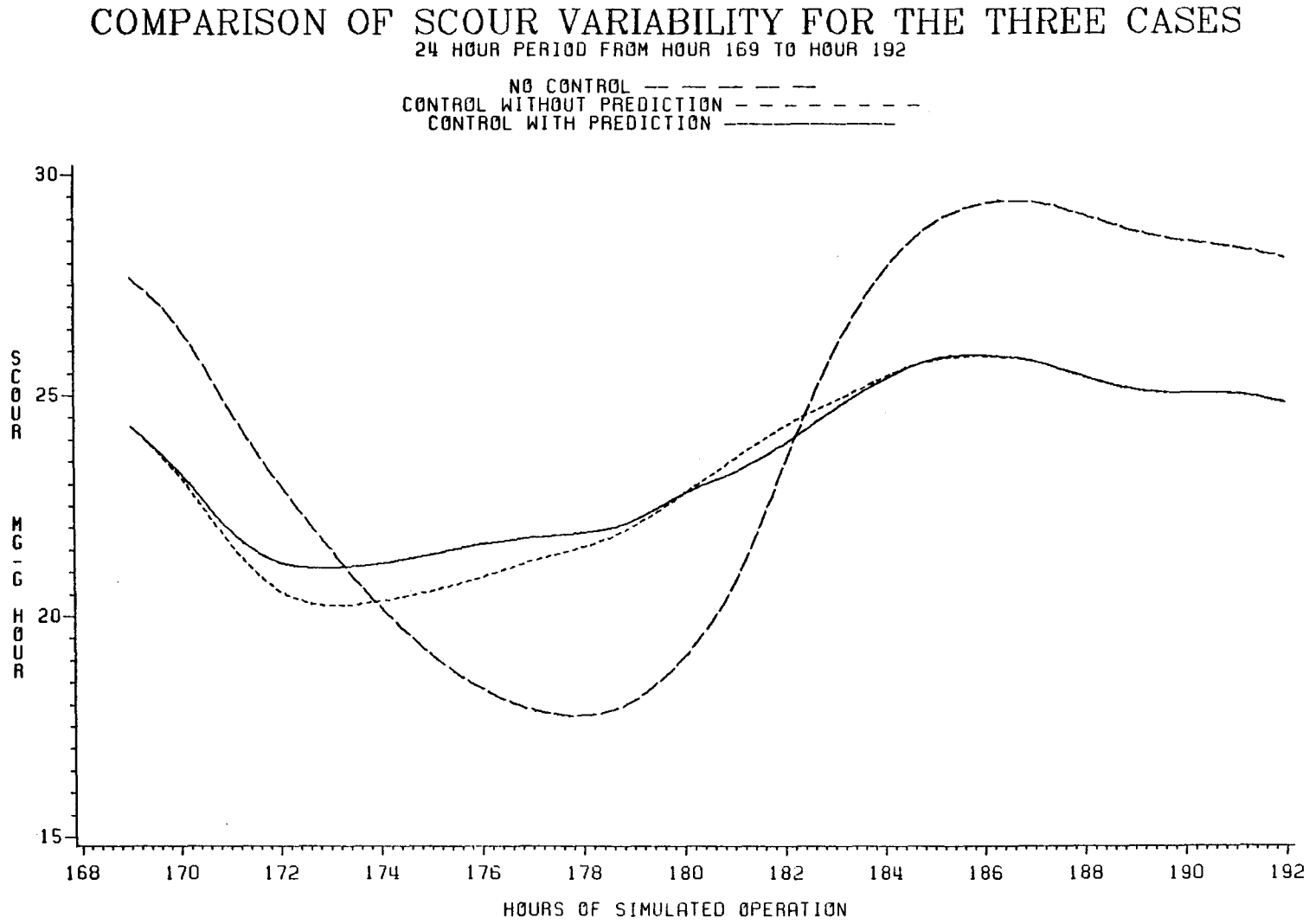


Figure 29: Scour Profiles from the Simulation Runs

Case	Dimensionless SCOUR Variance	Percentage Variance Reduction
No Control	0.03156	---
Control Without Prediction	0.007224	77 %
Control With Prediction	0.005230	83 %

TABLE 7

Scour Variances and Percentage Variance Reductions

5.5 DISCUSSION OF RESULTS

The additional 6% reduction in SCOUR variability achieved by ARIMA flow forecasting is small in appearance, but potentially significant.

Stenstrom (1976) thoroughly evaluated the pseudo feedforward-feedback control strategy, and found that there was little more that could be done to reduce process variability without introducing additional costly measures or some form of additional information. Thus, the control strategy has been utilized to its maximum potential, and further reductions in variability could have high marginal costs (cost per unit of improvement). It is not known at present how beneficial an additional 6% reduction in variability might be for an operating treatment plant. If a process control computer was available and could be programmed to perform the forecasting, the control strategy with forecasting might be an economical way to achieve further improvement. The control strategy without prediction

alone reduced variability by 77%, leaving only 23% additional possible improvement. If the "point of diminishing returns" has been reached, then an economical strategy which provides a 6% reduction out of 23% may prove to be of significant interest.

It should also be emphasized that the results from the single example of ARIMA forecasting presented here cannot be considered an exhaustive evaluation of ARIMA models for forecasting wastewater treatment process variables in general. These results should only be interpreted within the context of the following qualifying factors:

1. The particular single choice of performance criterion.
2. The particular set of input data (BOD, nitrogen, TSS, and flow rate).
3. The arbitrary 24-hour period chosen for analysis.
4. The particular choice of control strategy and control law.
5. The particular variable chosen to forecast.
6. The particular choice of forecast lead time.
7. The particular simulation model utilized.
8. The degree of correspondence between the simulation model and an actual treatment plant operating under the specified conditions.

This case study was presented solely as a preliminary step in investigating ARIMA models for forecasting wastewater treatment process variables. There has been much speculation in the literature

about the forecasting potential of ARIMA models; it is hoped that the particular example presented here will stimulate interest and will be followed by further investigations. It is clear from the list of qualifying factors that there are many interesting variations of the present analysis that could be investigated, and that dynamic simulation models could be utilized for many of them. Of greatest interest will be future experimental field studies and hybrid modeling-experimental studies that utilize complementary features of both modeling and experimentation.

Chapter VI

CONCLUSIONS

From reviewing previous investigations of ARIMA models and the results of the modeling performed for the present study, the following conclusions can be drawn:

1. In the water pollution control literature, four previous investigations (McMichael and Vigani, 1972; Goel and LaGrega, 1972; Berthouex et al., 1975; Debelak and Sims, 1981) involving univariate ARIMA models applied to wastewater treatment data have presented graphical results described as "forecasts." However, only one (Goel and LaGrega, 1972) actually performed and reported true beyond-sample forecasts. The other reported results were simply one-step-ahead within-sample forecasts; i.e., fitted values.
2. Graphs of one-step-ahead within-sample forecasts often appear to closely follow the observed data. Thus, those unfamiliar with the Box-Jenkins methodology and the concept of one-step-ahead within-sample forecasts may have inferred that ARIMA models can forecast the future rather well. However, what is actually portrayed in such graphs is simply the fitting of a model to <u>past</u> observations.

3. The only quantitative measure of goodness-of-fit reported by previous researchers for their ARIMA models was the coefficient of determination (R-squared), i.e., the fraction of variance in the data accounted for by the model. This has been variously reported in terms of R-squared, residual sum of squares, residual mean square, or "standard deviation". All are equivalent when converted to R-squared. Other traditional goodness-of-fit criteria such as mean absolute error, mean absolute percent error, and root mean square error have not been presented.

4. No quantitative measures of forecasting accuracy were presented for the only example of true beyond-sample forecasting reported (Goel and LaGrega, 1972). Generally accepted forecasting accuracy criteria include absolute errors, mean absolute error, percent errors, mean absolute percent error, and mean square or root mean square error (Carbone and Armstrong, 1982).

5. There have been several reported cases (Goel and LaGrega, 1972; Shih, 1976; Barnes and Rowe, 1978) where the same model form adequately described data from different treatment plants, for the same process variable. Similarly, the same model form has been identified for different process variables at the same treatment plant. More studies are needed to determine the potential significance and usefulness of such results. The common denominator may be similar diurnal variations in flow and pollutant concentration at treatment plants. It has been

observed that concentration variation tends to be positively correlated with flow variation (Wallace and Zollman, 1971; Young et al., 1978; Berthouex et al., 1978).

6. Very few investigations of Box-Jenkins forecasting for simulated or actual control of wastewater treatment plants have been conducted. The literature review for this thesis uncovered only two (LaGrega and Keenan, 1974; MacInnes et al., 1978).

7. There is need for a uniform, standardized format and terminology for presenting and discussing forecasting results from ARIMA models. This would promote better understanding of ARIMA forecasting, reduce misconceptions, and allow meaningful comparisons to be made between diverse data sets and models. It is proposed that the format and terminology introduced in this study be used for presenting fitting and forecasting results. The presentation of results would include:
   a) A plot of the observed values, fitted values, and fitted residuals.
   b) A plot of the observed values, forecast values, confidence limits, and forecasts errors.
   c) A table of appropriate computed error criteria for both the fitting and forecasting results.
   d) A statement of the type of forecasts being presented, using the classifications of "one-step-ahead within-sample" forecasts (or "fitted values"), "multi-step-ahead beyond-

sample" forecasts, and "one-step-ahead beyond-sample" forecasts. The forecast origin and lead times of the forecasts should also be specified.

8. In general, goodness-of-fit statistics and graphs of fitted values obtained at the estimation stage convey little information about the future-forecasting performance of ARIMA models. The main information about forecasting gained by fitting a proposed model is an indication of the forecast error variance, as measured by the sample variance of the estimated random shocks (fitted residuals). However, if the parameter estimates for a proposed ARIMA model do not change significantly over the last portion of the time series under study, then the one-step-ahead within-sample forecasts (fitted values) near the end of the data set are valid indicators of the one-step-ahead beyond-sample forecasting capability of the model. Virtually nothing can be inferred about multi-step-ahead beyond-sample forecasting performance from the fitted results alone.
9. Graphical presentations of forecasting results, including 95% confidence intervals and plotted forecast errors, are useful for developing an intuitive feeling for the forecasting accuracy of a proposed ARIMA model. These figures are also helpful for conveying forecasting results to non-technical persons. However, in order to fully evaluate a model for use in forecasting wastewater treatment variables, it is necessary

to consider quantitative measures of forecasting accuracy. In this study, traditional error criteria such as the range of absolute errors, the mean absolute error, the range of absolute percent errors, the mean absolute percent error, and the root mean square error were found to be useful for assessing the forecasting performance of ARIMA models. The proper choice of performance criterion will depend on the particular application for which the model is being considered.

10. The results of this study suggest that nonseasonal, low-order ARIMA(p,d,q) models are not useful for multi-step-ahead beyond-sample forecasting with long lead times, because the forecast profiles from these models cannot anticipate the dynamic variability displayed by wastewater treatment process variables. For multi-step-ahead beyond-sample forecasting, only seasonal ARIMA(p,d,q)(P,D,Q) models were found to provide realistic forecast profiles, because they have longer "memories" (i.e., forecast functions which incorporate higher lag orders) than nonseasonal models. By contrast, the results of this study indicate that in some cases, nonseasonal models may perform just as well as complicated seasonal models for one-step-ahead beyond-sample forecasting. This implies that if one-step-ahead beyond-sample forecasting is the modeling objective, there may be no real justification for seeking a complex model since a simpler one will suffice.

11. All of the one-step-ahead beyond-sample forecasts presented in this study, from both nonseasonal and seasonal models, were found to display a "one-interval lag effect." That is, the forecasted maxima and minima were often predicted to occur one interval after the observed maxima and minima in the actual data. This one-interval lag effect is a direct consequence of the forecast functions for the particular models studied. The forms of the forecast functions (see Appendix A) indicate that one of the largest terms in the forecast for the next interval is the weighted observed value from the preceding interval. Therefore, the one-step-ahead beyond-sample forecasts appear as lagged, scaled versions of the observed values (to a varying degree, depending on the magnitudes of the other terms in the forecast function).
12. The results of the simulations performed for this study suggest that ARIMA forecasting may prove useful for improving real-time control strategies, without the need for additional equipment or redesigning of facilities. Much work remains to be done to fully evaluate the utility of ARIMA models for forecasting wastewater treatment process variables.

Chapter VII

RELATED TOPICS FOR FUTURE RESEARCH

"My interest is in the future, because I'm going to spend the rest of my life there."

Charles F. Kettering

7.1 INTRODUCTION

During the course of this thesis, a number of potentially interesting topics arose which were beyond the scope of the original thesis objectives, but which appear to warrant further consideration. The following sections outline these related topics and provide comments about possible additional research.

7.2 EFFECT OF THE NUMBER OF OBSERVATIONS

In this study, ARIMA models were developed and applied for forecasting using data sets containing as few as N = 96 and as many as N = 332 observations. For reasons involving statistical sampling theory, Box and Jenkins (1976) recommend that at least N = 50 observations be used for model development, since the entire procedure relies on good estimates of the autocorrelation structure as indicated by the sample ACF and PACF. However, is there a useful upper limit for the number of observations needed in ARIMA modeling and

forecasting? Is it always "better" to have more data? What if the underlying model form is different for different subsets of the data? Also, when performing beyond-sample forecasting with updating of the observations, parameter estimates, and model identification, is it better to retain all the available observations including those at the beginning of the collected series, or should the earlier observations be deleted as the newer ones are incorporated? These are questions which merit further investigation.

7.3 PRACTICAL SIGNIFICANCE OF FORECAST ERRORS AND ERROR CRITERIA

This study presented typical ARIMA forecast errors and utilized various accuracy criteria for quantifying the errors. An important area of research is the assessment of the engineering significance of forecasting errors which occur in wastewater treatment plant control strategies incorporating influent forecasts. How much error can be tolerated by the control strategy before performance is seriously impaired?

A related topic is the selection of the appropriate forecast error performance criterion for a particular intended application. For example, is the maximum allowable error for a single forecast the most important consideration, or is the average error over some specified forecasting period more physically meaningful? Answers to questions such as these must be obtained through experimental field studies or mathematical simulation using models which accurately predict observed wastewater treatment plant operation.

7.4 DESEASONALIZATION AND ARIMA MODELS WITH SEASONAL PARAMETERS

Several techniques used in stochastic hydrology or the social sciences might be investigated for use in wastewater treatment time series analysis. The first is known as cyclic standardization (Yevjevich, 1966). Cyclic standardization is a transformation performed on the raw values of a periodic time series; it involves creating a new series by subtracting out periodic means and dividing by periodic standard deviations. For hourly wastewater treatment data displaying diurnal cycles, 24 hourly sample means and 24 hourly sample standard deviations would be calculated, each corresponding to a particular hour of the day. Note that this would require a total of 24M observations, where M is the minimum number of samples (one per day) needed to obtain a valid estimate of the mean and standard deviation for a particular hour. The cyclic transformation would be applied, and the resulting standardized series analyzed with a nonseasonal ARIMA model. The transformation would then be inverted to obtain the fitted values corresponding to the original untransformed series.

Other deseasonalization procedures such as those used by the U. S. Bureau of the Census, Bureau of Labor Statistics, and in the field of economics ("seasonal adjustment") might also prove useful.

Another modification of traditional Box-Jenkins analysis that merits consideration is the class of ARIMA models with periodically varying AR and MA parameters used in stochastic hydrology. Salas et al. (1982) provide a review of these models.

7.5 INFLUENT-INFLUENT TRANSFER FUNCTION MODELS

As discussed in the literature review, a number of investigators have developed transfer function (bivariate) models relating effluent quality to various influent predictor variables. In light of the previously mentioned positive correlation between influent flow rate and influent pollutant concentration, and an observed degree of correspondence between suspended solids and BOD, it seems feasible that transfer function models relating one influent variable to another influent variable might also prove useful. Successful results could have significant implications for real-time forecasting, for it may be possible to forecast a variable which cannot be measured in real time using one that can be rapidly determined.

7.6 AUTOMATED BOX-JENKINS MODELING

A significant limitation of Box-Jenkins forecasting for real-time control is the need for human intervention in the modeling process (e.g., examination and interpretation of ACF's and PACF's). However, several recent developments show promise for fully automating the procedure. Hill and Woodworth (1980) have reported results from an automatic modeling algorithm utilizing pattern recognition and an optimal model-order criterion. Reilly (1981) has described a commercially available, fully automated, interactive Box-Jenkins modeling program called "AUTOBJ", which can currently be implemented on an IBM personal computer possessing expanded options (Reilly, 1984). The program automatically performs identification, estimation,

diagnostics, and forecasting, requiring only initial input and subsequent updating of the time series. It is feasible to envision such data logging and transmittal activities being carried out by an integrated system of sensors communicating with a central process control/modeling computer.

7.7 COMPARISON OF CURRENTLY USED FORECASTING TECHNIQUES

As mentioned in Chapter 2, numerous quantitative forecasting techniques have been developed and are currently in use in the fields of operations research, management science, and statistics. Makridakis et al. (1982) recently conducted an international forecasting-accuracy competition in which 1001 time series were provided to expert proponents of 24 different forecasting techniques, including ARIMA models. Forecasts for the last 6 to 18 observations at the end of each series were made, and the results from the different methods were compared and evaluated. The paper contains summary descriptions of the 24 methods, over 25 tables containing hundreds of numerical measures of performance, and the results are still being debated in the literature (Armstrong and Lusk, 1983). It was reported that the Box-Jenkins methodology was the most costly and time-consuming of the 24 techniques. It was further observed that many simpler, automatic methods performed as well as ARIMA models. Thus, it is important for the wastewater treatment field to gain experience with many different forecasting techniques in order to guide selection of the appropriate model for a given application.

REFERENCES

Adams, B. J. (1975), "Modeling Sewage Treatment Plant Input BOD Data," Discussion, J. Environ. Eng. Div., ASCE, Vol. 101, No. EE5, pp. 877-879.

Adeyemi, S. O., Wu, S. M., and Berthouex, P. M. (1979), "Modeling and Control of a Phosphorous Removal Process by Multivariate Time Series Method," Water Research, Vol. 13, No. 1, pp. 105-112.

Ali, M. M., and Thalheimer, R. (1983), "Stationarity Tests in Time Series Model Building," J. Forecast., Vol. 2, No. 3, pp. 249-257.

Akaike, H. (1974), "A New Look at the Statistical Model Identification," IEEE Trans. Autom. Control, Vol. 19, No. 6, pp. 716-723.

Anderson, J. J. (1973), Personal communication, Cleveland, Ohio (as cited by Stenstrom, 1976).

Armstrong, J. S., and Lusk, E. J. (1983), "Commentary on the Makridakis Time Series Competition (M-Competition)," J. Forecast., Vol. 2, No. 3, pp. 259-311.

Barnes, J. W., and Rowe, F. A. (1978), "Modeling Sewer Flows Using Time Series Analysis," J. Environ. Eng. Div., ASCE, Vol. 104, No. EE4, pp. 639-646.

Beck, M. B. (1976), "Dynamic Modelling and Control Applications in Water Quality Maintenance," Water Research, Vol. 10, No. 7, pp. 575-595.

Beck, M. B. (1977), "The Identification and Adaptive Prediction of Urban Sewer Flows," Int. J. Control, Vol. 25, No. 3, pp. 425-440.

Berthouex, P. M., Hunter, W. G., Pallesen, L. C., and Shih, C. Y. (1975), "Modeling Sewage Treatment Plant Input BOD Data," J. Environ. Eng. Div., ASCE, Vol. 101, No. EE1, pp. 127-138.

Berthouex, P. M., and Hunter, W. G. (1975), "Treatment Plant Monitoring Programs: A Preliminary Analysis," J. Wat. Pollut. Control Fed., Vol. 47, No. 8, pp. 2143-2156.

Berthouex, P. M., Hunter, W. G., Pallesen, L., and Shih, C. Y. (1976), "The Use of Stochastic Models in the Interpretation of Historical Data from Sewage Treatment Plants," Water Research, Vol. 10, No. 8, pp. 689-698.

Berthouex, P. M., Hunter, W. G., and Pallesen, L. (1978a), "Dynamic Behavior of an Activated Sludge Plant," Water Research, Vol. 12, No. 11, pp. 957-972.

Berthouex, P. M., Hunter, W. G., and Pallesen, L. (1978b), "Monitoring Sewage Treatment Plants: Some Quality Control Aspects," J. Quality Tech., Vol. 10, No. 4, pp. 139-149.

Berthouex, P. M., Hunter, W. G., and Pallesen, L. (1979), "Analysis of the Dynamic Behavior of an Activated Sludge Plant--Comparison of Results With Hourly and Bi-Hourly Data," Water Research, Vol. 13, No. 12, pp. 1281-1283.

Berthouex, P. M., and Hunter, W. G. (1982), "Stochastic Modeling of an Industrial Activated Sludge Process," Discussion, Water Research, Vol. 16, No. 7, pp. 1301-1302.

Berthouex, P. M., Hunter, W. G., and Pallesen, L. (1983), "Deriving Components of Variance in a Transfer Function-Noise Model," Water Research, Vol. 17, No. 4, pp. 467-469.

Box, G. E. P., and Jenkins, G. M. (1968), "Some Recent Advances in Forecasting and Control, Part I," Applied Statistics, Vol. 17, No. 2, pp. 91-109.

Box, G. E. P., and Jenkins, G. M. (1974), "Some Recent Advances in Forecasting and Control, Part II," Applied Statistics, Vol. 23, No. 2, pp. 158-179.

Box, G. E. P., and Jenkins, G. M. (1976), Time Series Analysis, Forecasting and Control, Revised Edition, Holden-Day, San Fransisco.

Briggs, J. (1972), Personal communication, Atlanta, Georgia (as cited by Stenstrom, 1976).

Carbone, R., and Armstrong, J. S. (1982), "Note--Evaluation of Extrapolative Forecasting Methods: Results of a Survey of Academicians and Practitioners," J. Forecast., Vol. 1, No. 2, pp. 215-217.

Carlson, R. F., MacCormick, A. J. A., and Watts, D. G. (1970), Application of Linear Random Models to Four Annual Streamflow Series," Water Resources Research, Vol. 6, No. 4, pp. 1070-1078.

D'Astous, F., and Hipel, K. W. (1979), "Analyzing Environmental Time Series," J. Environ. Eng. Div., ASCE, Vol. 105, No. EE5, pp. 979-992.

Debelak, K. A., and Sims, C. A. (1981), "Stochastic Modeling of an Industrial Activated Sludge Process," Water Research, Vol. 15, No. 10, pp. 1173-1183.

Filion, M. P., Murphy, K. L., and Stephenson, J. P. (1979), "Performance of a Rotating Biological Contactor Under Transient Loading Conditions," J. Wat. Pollut. Control Fed., Vol. 51, No. 7, pp. 1925-1933.

Fuller, F. C., and Tsokos, C. P. (1971), "Time Series Analysis of Water Pollution Data," Biometrics, Vol. 27, No. 4, pp. 1017-1034.

Goel, A. L., and LaGrega, M. D. (1972), "Stochastic Models for Forecasting Sewage Flows," presented at the April 26-28, 1972 41st National Meeting, Operations Research Society of America, New Orleans; also presented as "Forecasting Wastewater Flow Rates by Time Series Analysis" at the October 5-10, 1974 47th Annual Water Pollution Control Federation Conference, Denver.

Goel, A. L. (1984), Personal communication.

Gunnerson, C. G. (1966), "Optimizing Sampling Intervals in Tidal Estuaries," J. San. Eng. Div., ASCE, Vol. 92, No. SA2, pp. 103-125.

Hansen, J. L., Fiok, A. E., and Hovious, J. C. (1980), "Dynamic Modeling of Industrial Wastewater Treatment Plant Data," J. Wat. Pollut. Control Fed., Vol. 52, No. 7, pp. 1966-1975.

Hill, G. W., and Woodworth, D. (1980), "Automatic Box-Jenkins Forecasting," J. Opl. Res. Soc., Vol. 31, No. 5, pp. 413-422.

Huck, P. M., and Farquhar, G. J. (1974), "Water Quality Models Using the Box-Jenkins Method," J. Environ. Eng. Div., ASCE, Vol. 100, No. EE3, pp. 733-752.

Huck, P. M., and Farquhar, G. J. (1975), "Water Quality Models Using the Box-Jenkins Method," Closure Discussion, J. Environ. Eng. Div., ASCE, Vol. 101, No. EE6, pp. 1032-1034.

Klemes, V. (1982), "Empirical and Causal Models in Hydrology," in: Scientific Basis of Water-Resource Management, National Academy Press, Washington, D. C., pp. 95-104.

Kottegoda, N. T. (1980), Stochastic Water Resources Technology, John Wiley and Sons, New York.

Labadie, J. W., Grigg, N. S., and Trotta, P. D. (1976), "On-Line Models for Computerized Control of Combined Sewer Systems," in: Proceedings of the Conference on Environmental Modeling and Simulation, April 19-22, 1976, Cincinnati, Ohio. Report EPA 600/9-76-016, U. S. Environmental Protection Agency, Washington, D. C., pp. 755-759, July 1976.

LaGrega, M. D., and Keenan, J. D. (1974), "Effects of Equalizing Wastewater Flows," J. Wat. Pollut. Control Fed., Vol. 46, No. 1, pp. 123-132.

LaGrega, M. D. (1984), Personal communication.

Litwin, Y. J., and Joeres, E. F. (1975), "Water Quality Models Using the Box-Jenkins Method," Discussion, J. Environ. Eng. Div., Vol. 101, No. EE3, pp. 449-451.

MacInnes, C. D., Middleton, A. C., and Adamowski, K. (1978), "Stochastic Design of Flow Equalization Basins," J. Environ. Eng. Div., ASCE, Vol. 104, No. EE6, pp. 1277-1291.

Makridakis, S., and Wheelwright, S. C. (1978), Forecasting: Methods and Applications, John Wiley and Sons, New York.

Makridakis, S., Anderson, A., Carbone, R., Fildes, R., Hibon, M., Lewandowski, R., Newton, J., Parzen, E., and Winkler, R. (1982), "The Accuracy of Extrapolation (Time Series) Methods: Results of a Forecasting Competition," J. Forecast., Vol. 1, No. 2, pp. 111-153.

McCleary, R., and Hay, Jr., R. A. (1980), Applied Time Series Analysis for the Social Sciences, Sage Publications, Beverly Hills, California.

McKerchar, A. I., and Delleur, J. W. (1974), "Application of Seasonal Parametric Linear Stochastic Models to Monthly Flow Data," Water Resources Research, Vol. 10, No. 2, pp. 246-255.

McLeod, A. I., Hipel, K. W., and Lennox, W. C. (1977), "Advances in Box-Jenkins Modeling. 2. Applications," Water Resources Research, Vol. 13, No. 3, pp. 577-586.

McLeod, A. I., Hipel, K. W., and Comancho, F. (1983), "Trend Assessment of Water Quality Time Series," Water Resources Bulletin, AWRA, Vol. 19, No. 4, pp. 537-547.

McMichael, F. C., and Hunter, J. S. (1972), "Stochastic Modeling of Temperature and Flow in Rivers," Water Resources Research, Vol. 8, No. 1, pp. 87-98.

McMichael, F. C., and Vigani, F. C. (1972), "Characterization of Time-Varying Organic Loads," Discussion, J. San. Eng. Div., ASCE, Vol. 98, No. SA2, pp. 444-455.

Montgomery, D. C., and Johnson, L. A. (1976), Forecasting and Time Series Analysis, McGraw-Hill, New York.

Murphy, K. L., Sutton, P. M., and Jank, B. E. (1977), "Dynamic Nature of Nitrifying Biological Suspended Growth Systems," Progress in Water Technology, Vol. 9, Nos. 1-4, pp. 279-290.

Nelson, C. R. (1973), Applied Time Series Analysis for Managerial Forecasting, Holden-Day, San Fransisco.

Newbold, P. (1983), "ARIMA Model Building and the Time Series Analysis Approach to Forecasting," J. Forecast., Vol. 2, No. 1, pp. 23-35.

Olsson, G. (1976), "State of the Art in Sewage Treatment Plant Control," in: Chemical Process Control, A. S. Foss and M. M. Denn, eds., American Institute of Chemical Engineers Symposium Series, Vol. 72, No. 159, pp. 52-76.

Pankratz, A. (1983), Forecasting With Univariate Box-Jenkins Models, John Wiley and Sons, New York.

Pindyck, R. S., and Rubinfeld, D. L. (1976), Econometric Models and Economic Forecasts, McGraw-Hill, New York.

Reilly, D. P. (1981), "Recent Experiences With an Automatic Box-Jenkins Modelling Algorithm," in: Time Series Analysis, O. D. Anderson and M. R. Perryman, eds., North-Holland, New York.

Reilly, D. P. (1984), Personal communication, Automatic Forecasting Systems, Inc., Hatboro, Pennsylvania.

Rowe, F. A. (1977), "Stochastic Models of Wastewater Flows," thesis presented to the University of Texas at Austin, Texas, in 1977, in partial fulfillment of the requirements for the degree of Master of Science.

Salas, J. D., Delleur, J. W., Yevjevich, V. M., and Lane, W. L. (1980), Applied Modeling of Hydrologic Time Series, Water Resources Publications, Fort Collins, Colorado.

Salas, J. D., Boes, D. C., and Smith, R. A. (1982), "Estimation of ARMA Models With Seasonal Parameters," Water Resources Research, Vol. 18, No. 4, pp. 1006-1021.

SAS Institute Inc. (1981), SAS/GRAPH User's Guide, 1981 Edition, SAS Institute Inc., Cary, North Carolina.

SAS Institute Inc. (1982), SAS/ETS User's Guide, 1982 Edition, SAS Institute Inc., Cary, North Carolina.

Schumacher, E. F. (1973), Small is Beautiful, Harper and Row, New York.

Shahane, A. N. (1975), "Modeling Sewage Treatment Plant Input BOD Data," Discussion, J. Environ. Eng. Div., ASCE, Vol. 101, No. EE6, pp. 1036-1038.

Shih, C. Y., Berthouex, P. M., and Hunter, W. G., (1974), "A Field Study of the Dynamic Performance of a Sewage Treatment Plant," unpublished research report, University of Wisconsin at Madison, Wisconsin.

Shih, C. Y. (1976), "Investigation of Dynamic Performance of Wastewater Treatment Plants," thesis submitted to the University of Wisconsin at Madison, Wisconsin, in 1976, in partial fulfillment of the requirements for the degree of Doctor of Philosophy.

Stenstrom, M. K. (1976), "A Dynamic Model and Computer Compatible Control Strategies for Wastewater Treatment Plants," dissertation submitted to Clemson University at Clemson, South Carolina, in 1976, in partial fulfillment of the requirements for the degree of Doctor of Philosophy.

Stenstrom, M. K., and Andrews, J. F. (1979), "Real-Time Control of Activated Sludge Process," J. Environ. Eng. Div., ASCE, Vol. 105, No. EE2, pp. 245-260.

Tanthapanichakoon, W., and Himmelblau, D. M. (1980), "Stochastic Analysis of Process-Control Models," Simulation, Vol. 34, No. 6, pp. 181-193.

Tao, P. C., and Delleur, J. W. (1976), "Seasonal and Nonseasonal ARMA Models in Hydrology," J. Hydr. Div., ASCE, Vol. 102, No. HY10, pp. 1541-1559.

Thomann, R. V. (1967), "Time Series Analyses of Water Quality Data," J. San. Eng. Div., ASCE, Vol. 93, No. SA1, pp. 1-23.

Thomann, R. V. (1970), "Variability of Waste Treatment Plant Performance," J. San. Eng. Div., ASCE, Vol. 96, No. SA3, pp. 819-837.

Tukey, J. W. (1961), "Discussion, Emphasizing the Connection Between Analysis of Variance and Spectrum Analysis," Technometrics, Vol. 3, No. 2, p. 191-219.

Wallace, A. T., and Zollman, D. M. (1971), "Characterization of Time-Varying Organic Loads," J. San. Eng. Div., ASCE, Vol. 97, No. SA3, pp. 257-268.

Wastler, T. A., and Walter, C. M. (1968), "Statistical Approach to Estuarine Behavior," J. San. Eng. Div., ASCE, Vol. 94, No. SA6, pp. 1175-1194.

Yevjevich, V. M. (1966), "Stochastic Problems in the Design of Reservoirs," in: Water Research, A. V. Kneese and S. C. Smith, eds., Johns Hopkins Press, Baltimore, Maryland, pp. 375-411.

Young, J. C., Cleasby, J. L., and Baumann, E. R. (1978), "Flow and Load Variations in Treatment Plant Design," J. Environ. Eng. Div., ASCE, Vol. 104, No. EE2, pp. 289-303.

Young, P., and Whitehead, P. (1975), "A Recursive Approach to Time-Series Analysis for Multivariable Systems," in: Modeling and Simulation of Water Resources Systems, G. C. Vansteenkiste, ed., North-Holland, Amsterdam, pp. 39-52.

Appendix A

SUMMARY OF ARIMA MODELS PRESENTED

Note: The total mean square (TMS) is defined as the sum of the squares of the original data about their estimated mean, divided by the number of observations; i.e., the sample variance. The residual mean square (RMS) is defined as the sum of the squares of the fitted residuals, divided by the degrees of freedom. "Degrees of freedom" is equal to the number of fitted values minus the number of estimated parameters.

MADISON NINE SPRINGS TREATMENT PLANT INFLUENT BOD

Determination of Total Mean Square

Model: $(z_t-\mu) = a_t$

Parameter Estimate: (N = 332) $\bar{\mu}$ = 122.4 mg/l

Total Mean Square : (N = 332) TMS = 1739.4 $(mg/l)^2$

ARIMA(1,0,0)

Model: $(1-\phi_1 B)(z_t-\mu) = a_t$

Forecast Function: $\bar{z}_t = \bar{\phi}_1(z_{t-1}-\bar{\mu}) + \bar{\mu}$

Parameter Estimates: (N = 332) $\bar{\mu}$ = 123.4 mg/l, $\bar{\phi}_1$ = 0.809

(N = 300) $\bar{\mu}$ = 123.7 mg/l, $\bar{\phi}_1$= 0.811

Residual Mean Square: (N = 332) RMS = 607.1 $(mg/l)^2$

ARIMA(2,0,0)(0,1,1) s=24

Model: $(1-\phi_1 B -\phi_2 B^2)(1-B^{24})z_t = (1-\theta_1 B^{24})a_t$

Forecast Function: $\bar{z}_t = \bar{\phi}_1 z_{t-1} + \bar{\phi}_2 z_{t-2} + z_{t-24} - \bar{\phi}_1 z_{t-25} - \bar{\phi}_2 z_{t-26} - \bar{\theta}_1 \bar{a}_{t-24}$

Parameter Estimates: (N = 332) $\bar{\phi}_1$= 0.569, $\bar{\phi}_2$= 0.214, $\bar{\theta}_1$= 0.831

(N = 300) $\bar{\phi}_1$= 0.569, $\bar{\phi}_2$= 0.231, $\bar{\theta}_1$= 0.812

Residual Mean Square: (N = 332) RMS = 609.5 $(mg/l)^2$

MINNEAPOLIS-SAINT PAUL SEWER STATION 004 COD

Determination of Total Mean Square

Model: $(z_t - \mu) = a_t$

Parameter Estimate: (N = 96) $\bar{\mu}$ = 697.6 mg/l

Total Mean Square : (N = 96) TMS = 25,151 $(mg/l)^2$

ARIMA(0,1,1)

Model: $(1-B)z_t = (1-\theta_1 B)a_t$

Forecast Function: $\bar{z}_t = z_{t-1} - \bar{\theta}_1 \bar{a}_{t-1}$

Parameter Estimate: (N = 96) $\bar{\theta}_1 = 0.549$

(N = 72) $\bar{\theta}_1 = 0.593$

Residual Mean Square: (N = 96) RMS = 14,530 $(mg/l)^2$

ATLANTA R. M. CLAYTON TREATMENT PLANT INFLUENT BOD

Determination of Total Mean Square

Model: $(z_t - \mu) = a_t$

Parameter Estimate: (N = 168) $\bar{\mu} = 183.3$ mg/l

Total Mean Square : (N = 168) TMS = 3774.4 $(mg/l)^2$

ARIMA(3,0,0)(0,1,1) s=24

Model: $(1-\phi_1 B - \phi_3 B^3)(1-B^{24})z_t = (1-\theta_1 B^{24})a_t$

Forecast Function: $\bar{z}_t = \bar{\phi}_1 z_{t-1} + \bar{\phi}_3 z_{t-3} + z_{t-24} - \bar{\phi}_1 z_{t-25} - \bar{\phi}_3 z_{t-27} - \bar{\theta}_1 \bar{a}_{t-24}$

Parameter Estimates: (N = 168) $\bar{\phi}_1 = 0.545$, $\bar{\phi}_3 = 0.268$, $\bar{\theta}_1 = 0.768$

(N = 144) $\bar{\phi}_1 = 0.488$, $\bar{\phi}_3 = 0.239$, $\bar{\theta}_1 = 0.733$

Residual Mean Square: (N = 168) RMS = 984.8 $(mg/l)^2$

ATLANTA R. M. CLAYTON TREATMENT PLANT TOTAL SUSPENDED SOLIDS

Determination of Total Mean Square

Model: $(z_t-\mu) = a_t$

Parameter Estimate: (N = 168) $\bar{\mu}$ = 150.8 mg/l

Total Mean Square : (N = 168) TMS = 2857.4 $(mg/l)^2$

ARIMA(2,0,0)(0,1,1) s=24

Model: $(1-\phi_1 B -\phi_2 B^2)(1-B^{24})z_t = (1-\theta_1 B^{24})a_t$

Forecast Function: $\bar{z}_t = \bar{\phi}_1 z_{t-1} + \bar{\phi}_2 z_{t-2} + z_{t-24} - \bar{\phi}_1 z_{t-25} - \bar{\phi}_2 z_{t-26} - \bar{\theta}_1 \bar{a}_{t-24}$

Parameter Estimates: (N = 168) $\bar{\phi}_1$= 0.323, $\bar{\phi}_2$= 0.304, $\bar{\theta}_1$= 0.586

(N = 144) $\bar{\phi}_1$= 0.292, $\bar{\phi}_2$= 0.268, $\bar{\theta}_1$= 0.625

Residual Mean Square: (N = 168) RMS = 1339.5 $(mg/l)^2$

MINNEAPOLIS-SAINT PAUL INTERCEPTOR SEWER FLOW RATE

Determination of Total Mean Square

Model: $(z_t-\mu) = a_t$

Parameter Estimate: (N = 168) $\bar{\mu}$ = 180.4 MGD

Total Mean Square : (N = 168) TMS = 1039.8 $(MGD)^2$

ARIMA(0,1,0)(0,1,1) s=24

Model: $(1-B)(1-B^{24})z_t = (1-\theta_1 B^{24})a_t$

Forecast Function: $\bar{z}_t = z_{t-1} + z_{t-24} - z_{t-25} - \bar{\theta}_1 \bar{a}_{t-24}$

Parameter Estimate: (N = 168) $\bar{\theta}_1 = 0.455$

(N = 144) $\bar{\theta}_1 = 0.497$

Residual Mean Square: (N = 168) RMS = 82.89 $(MGD)^2$

Appendix B

DEFINITIONS OF ERROR CRITERIA USED

Note: The same notation is used here for comparing either fitted values or forecast values to the actual observed values of the time series.

t = time index

N = number of observations and corresponding fitted values (forecasts)

z_t = observed time series value at time t

$\bar{z}_t$ = fitted value (forecast) corresponding to z_t

$\bar{a}_t$ = fitted residual (forecast error) = $z_t - \bar{z}_t$

Absolute Error = AE = $|z_t - \bar{z}_t| = |\bar{a}_t|$

Range of Absolute Errors = RAE = (AE_{min}, AE_{max})

$$\text{Mean Absolute Error} = MAE = \frac{1}{N} \sum_{t=1}^{N} |\bar{a}_t|$$

$$\text{Absolute Percent Error} = APE = 100 \frac{|z_t - \bar{z}_t|}{|z_t|} = 100 \frac{|\bar{a}_t|}{|z_t|}$$

Range of Absolute Percent Errors = RAPE = (APE_{min}, APE_{max})

$$\text{Mean Absolute Percent Error} = \text{MAPE} = \frac{100}{N} \sum_{t=1}^{N} \frac{|\bar{a}_t|}{|z_t|}$$

$$\text{Root Mean Square Error} = \text{RMSE} = \left(\frac{1}{N} \sum_{t=1}^{N} \bar{a}_t^{\,2}\right)^{1/2}$$