AD-A254 170

MENTATION PAGE

Form Approved
OMB No. 0704-0188

REPORT DATE	3. REPORT TYPE AND DATES COVERED
December 1991	THESIS

4. TITLE AND SUBTITLE
Adequate Sampling of a Chaotic Time Series

5. FUNDING NUMBERS

6. AUTHOR(S)
Jeffrey A. Doran, Captain

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
AFIT Student Attending: Pennsylvania State University

8. PERFORMING ORGANIZATION REPORT NUMBER
AFIT/CI/CIA-92-054

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)
AFIT/CI
Wright-Patterson AFB OH 45433-6583

10. SPONSORING/MONITORING AGENCY REPORT NUMBER

11. SUPPLEMENTARY NOTES

12a. DISTRIBUTION/AVAILABILITY STATEMENT
Approved for Public Release IAW 190-1
Distributed Unlimited
ERNEST A. HAYGOOD, Captain, USAF
Executive Officer

12b. DISTRIBUTION CODE

13. ABSTRACT *(Maximum 200 words)*

14. SUBJECT TERMS

15. NUMBER OF PAGES
172

16. PRICE CODE

17. SECURITY CLASSIFICATION OF REPORT	18. SECURITY CLASSIFICATION OF THIS PAGE	19. SECURITY CLASSIFICATION OF ABSTRACT	20. LIMITATION OF ABSTRACT

The Pennsylvania State University

The Graduate School

Department of Meteorology

ADEQUATE SAMPLING OF A CHAOTIC TIME SERIES

A Thesis in

Meteorology

by

Jeffrey Alan Doran

Submitted in Partial Fulfillment
of the Requirements
for the Degree of

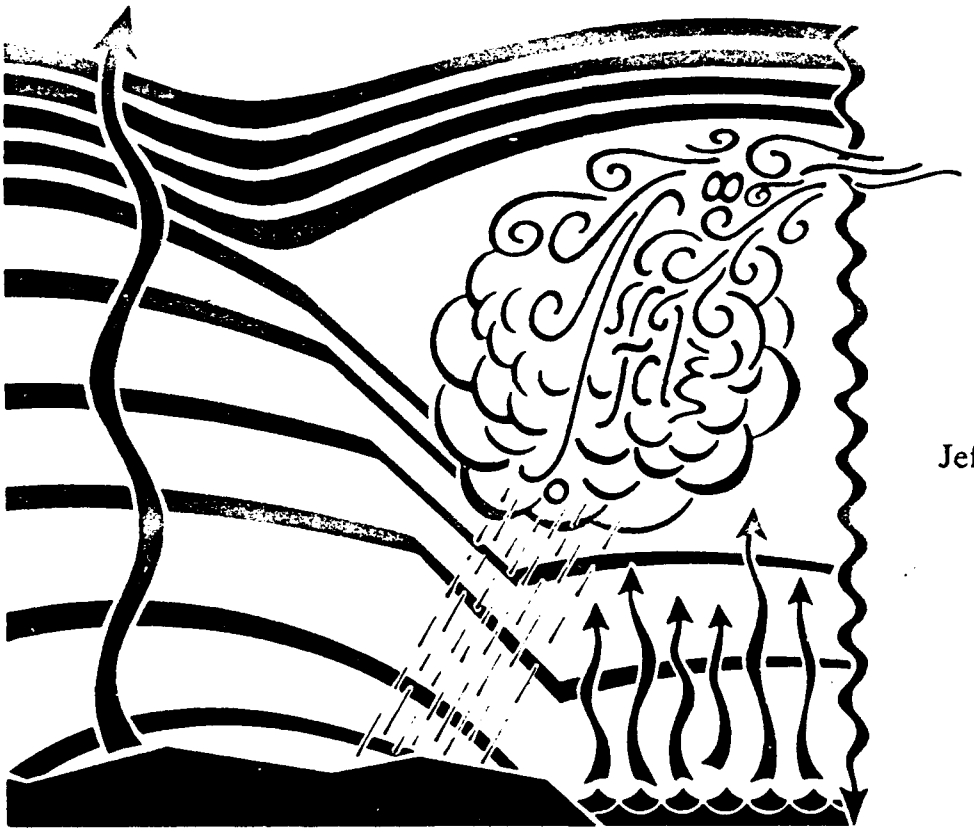
Master of Science

December 1991

PENNSTATE

DEPARTMENT OF METEOROLOGY

ADEQUATE SAMPLING OF A CHAOTIC TIME SERIES



A Thesis

in

Meteorology

by

Jeffrey Alan Doran

Master of Science

December 1991

ABSTRACT

Currently there is some disagreement about what constitutes an adequate sample of a time series with which chaos measures may be quantified. In this thesis, a method for objectively determining such a sample is presented. This method is based on a new, relatively efficient measure, the Histogram Measure, which allows large amounts of data to be considered. This measure also may be used to distinguish the chaotic from the transient, or nonchaotic, portions of the solution that are inherent in any chaotic time series. This is a crucial consideration, since transients contaminate the chaotic characteristics of any time series, be it from observations or models. This measure also leads to a predictability estimate--that of loss of information gain--as functions of sample length and elapsed time.

The Histogram Measure is tested with time series generated by the Lorenz (1963) three-component model of Rayleigh-Bénard convection. It is shown that the determination of criteria for quantifying adequate samples of data yields a definitive cost/benefit result. In effect, there is a balance between obtaining the greatest possible accuracy and spending the fewest resources; beyond a particular time or number of data points, only a minimal benefit is realized for the increased cost. It is also shown that the

results are extremely sensitive to the manner in which the data are sampled; the greatest of these sensitivities is in the time step size that is used to create the data set. Data sets generated with two different time steps are considered; when viewed in terms of series length, the ones created with the larger time step are shown to more efficiently produce convergent results. A similar conclusion is also reached for estimating the Correlation Dimension Measure (Grassberger and Procaccia 1983b). These results give validity to the notion that optimal sampling strategies can be found that best lead to acceptable values for chaos quantities, at least to within suitably small tolerances. This finding has important consequences for chaotic time series that are longer, more complicated, and more operational than those for the Lorenz model.

The Pennsylvania State University

The Graduate School

Department of Meteorology

ADEQUATE SAMPLING OF A CHAOTIC TIME SERIES

A Thesis in

Meteorology

by

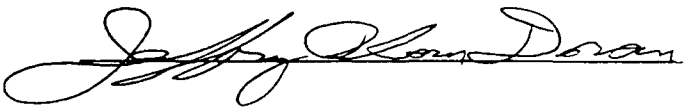
Jeffrey Alan Doran

Submitted in Partial Fulfillment
of the Requirements
for the Degree of

Master of Science

December 1991

I grant The Pennsylvania State University the nonexclusive right to use this work for the University's own purposes and to make single copies of the work available to the public on a not-for-profit basis if copies are not otherwise available.



Jeffrey Alan Doran

We approve the thesis of Jeffrey Alan Doran.

Date of Signature

Hampton N. Shirer — 21 August 1991

Hampton N. Shirer
Associate Professor of Meteorology
Thesis Adviser

Robert Wells — 21 August 1991

Robert Wells
Professor of Mathematics

William M. Frank — 21 Aug 1991

William M. Frank
Professor of Meteorology
Head of the Department of Meteorology

ABSTRACT

Currently there is some disagreement about what constitutes an adequate sample of a time series with which chaos measures may be quantified. In this thesis, a method for objectively determining such a sample is presented. This method is based on a new, relatively efficient measure, the Histogram Measure, which allows large amounts of data to be considered. This measure also may be used to distinguish the chaotic from the transient, or nonchaotic, portions of the solution that are inherent in any chaotic time series. This is a crucial consideration, since transients contaminate the chaotic characteristics of any time series, be it from observations or models. This measure also leads to a predictability estimate--that of loss of information gain--as functions of sample length and elapsed time.

The Histogram Measure is tested with time series generated by the Lorenz (1963) three-component model of Rayleigh-Bénard convection. It is shown that the determination of criteria for quantifying adequate samples of data yields a definitive cost/benefit result. In effect, there is a balance between obtaining the greatest possible accuracy and spending the fewest resources; beyond a particular time or number of data points, only a minimal benefit is realized for the increased cost. It is also shown that the

results are extremely sensitive to the manner in which the data are sampled; the greatest of these sensitivities is in the time step size that is used to create the data set. Data sets generated with two different time steps are considered; when viewed in terms of series length, the ones created with the larger time step are shown to more efficiently produce convergent results. A similar conclusion is also reached for estimating the Correlation Dimension Measure (Grassberger and Procaccia 1983b). These results give validity to the notion that optimal sampling strategies can be found that best lead to acceptable values for chaos quantities, at least to within suitably small tolerances. This finding has important consequences for chaotic time series that are longer, more complicated, and more operational than those for the Lorenz model.

TABLE OF CONTENTS

TABLE OF CONTENTS (continued)

LIST OF FIGURES

LIST OF FIGURES (continued)

LIST OF FIGURES (continued)

LIST OF FIGURES (continued)

LIST OF FIGURES (continued)

LIST OF FIGURES (continued)

LIST OF FIGURES (continued)

LIST OF FIGURES (continued)

LIST OF FIGURES (continued)

LIST OF TABLES

LIST OF TABLES (continued)

ACKNOWLEDGMENTS

There are so many to thank for the existence of this thesis: First, my faculty advisor, Professor Hampton N. (Nels) Shirer, whose brilliant guidance and thoughtful patience with my every concern were a blessing; second, Professor Robert Wells, the meteorology department's resident mathematician and the significant other in the weekly Nels and Wells "meeting of the minds."

Many of my fellow graduate students also lent me a hand. Thanks go to Julie Schramm and Bob Tomas, who helped with the Runge-Kutta and Correlation Dimension codes, respectively. I also thank Tom Salem and particularly Mark Laufersweiler for their expertise in the nuances of the computer system.

Most of all, I want to thank my wonderful wife, Debra, without whose endless love and support this thesis would not have been completed--she was always there for me, good days and bad. I could not end without thanking my two kitties--the little, striped Spike and the big, black Boo-Bear. On numerous occasions, they took time out of their busy schedules to jump on my lap in order to proofread the manuscript and to verify the accuracy of the information written on the computer screen.

A final word of thanks goes to the Air Force and Colonel Gary Zeigler (Ret.), who gave me the opportunity to pursue this work.

CHAPTER 1

INTRODUCTION

In recent years, numerous researchers within the atmospheric science community have recognized the importance of quantifying the behavior of the atmosphere and its climate via application of the principles of chaotic dynamical systems, whether with simplified models (e.g., Lorenz 1963, 1982, 1984) or with observed data (e.g., Fraedrich 1986; Hense 1987). Despite the existence of numerous successful methods for finding chaotic attractors of dynamical systems and for quantifying their characteristic fractal structures from time series (e.g., Grassberger and Procaccia 1983a, b; Lorenz 1963; Mandlebrot 1977; Takens 1981), these measures are relatively expensive to calculate with the amount of data that is most often used. As a result, many investigators have published results claiming success with limited data sets (e.g., Brandstäter *et al.* 1983; Nicolis and Nicolis 1986). Therefore, in this study, we primarily determine the amount of data necessary for adequately sampling chaotic time series, using the Lorenz (1963) simplified model of atmospheric convection and

utilizing a new, inexpensive measure that allows us to work with relatively large data sets. In finding these required data sets, we obtain relevant predictability estimates (comparable to loss of information gain) and comment on numerous sampling issues inherent to the model characteristics.

1.1. Background and Motivation

Chaos is the term used to describe a particular type of behavior inherent to dynamical systems and whose principles were first recognized by Edward Lorenz, a meteorologist at the Massachusetts Institute of Technology. In his simplified model of Rayleigh-Bénard convection (Lorenz 1963), he found that trajectories through two points initially very close to one another eventually diverge and evolve in totally different manners. After a certain time, there is no similarity in their evolutionary behavior; they may as well have been randomly picked. This characteristic is more commonly known as *sensitive dependence on initial conditions.* Even more fascinating is that, above a particular value of the system forcing, the time-dependent variables, when displayed together, outline a unique geometric structure on which all of the points

lie; this structure is more commonly known as the *attractor* of the system. All attractors of this type, which are called strange or chaotic, usually exhibit a noninteger or *fractal* dimension value, which must be less than the degrees of freedom of the system (Mandlebrot 1977).

Throughout the 1970s and 1980s, the study of chaotic dynamical systems blossomed within the scientific community, spanning numerous disciplines ranging from mathematics (e.g., Mandlebrot 1977; Takens 1981) to astronomy (Hénon 1976) to physics (e.g., Feigenbaum 1978; Gollub and Swinney 1975; Grassberger and Procaccia 1983a, b; Ruelle 1979) to chemistry (Rössler 1981), to name a few. *Order within disorder* became the buzzwords linking the scientific community together (Gleick 1987).

Generally speaking, it was not until the early 1980s that researchers began using the principles of chaotic dynamical systems to describe the behavior of certain physical quantities in the natural sciences. Guckenheimer and Buzyna (1983) found, using a rotating annulus whose flow was within a geostrophic turbulence regime, that there exists a fractal dimension value between seven and 12 for an attractor. Brandstäter *et al.* (1983) found a stable dimension value between four and five for an attractor within the turbulent regime exhibited by Couette-Taylor flow. Nicolis and Nicolis (1986) obtained

a definitive dimension value (~ 3.1) for a time series of deep-sea oxygen isotropes, while Henderson and Wells (1988) determined a definitive range of dimension values for 500 millibar height indices and vertical wind velocities. Many others have also reported similar successes in quantifying the dimensions of climatic and atmospheric data (e.g., Fraedrich 1987, 1988; Krishna Mohan *et al.* 1989; Thomson and Henderson 1989); some limited success has even been achieved in predicting the future behavior of these and related quantities over time scales of approximately one to seven days (Abarbanel *et al.* 1989).

Despite the advantages of characterizing these physical quantities in terms of their chaotic dynamics, there are limitations and tradeoffs that are encountered when conducting these types of studies. Most notable is the lengthy time needed to calculate these measures and, thus, expense becomes a key issue in determining the extent of the study. As a result, the bulk of these investigations have been done with relatively small data sets (< 3,000 points). For example, Nicolis and Nicolis (1986) claim that they obtain an attactor using the equivalent of only 184 points spanning over 1,000,000 years. Grassberger (1986) counters that their results are not valid, based on his own calculations using a larger sample of the same data series (~ 500 points). He concludes that with these few data points, it is nearly impossible to differentiate between the

deterministic behavior desired and the noisy behavior that contaminates the results.

Despite these limitations, some success has been achieved using methods for optimizing the results given by limited data. Ben-Mizrachi *et al.* (1984) successfully characterized noise in the Lorenz attractor using only 600 points by noting the correlation dimension behavior as a function of the ratio of a range of time scales (small, noise-plagued values versus larger, more deterministic values). Abraham *et al.* (1986) have successfully reconstructed, to within a reasonable tolerance, the values of the correlation dimension for the Hénon attractor using subsets of a larger data set. Ellner (1988) has developed a dimension measure, the Maximum-Likelihood (ML) method, specifically designed for use with small data sets (< 250 points). He claims that this method is superior to the Grassberger and Procaccia (1983b) Correlation Dimension Measure because it is less subject to distortions at small distances owing to finite-sample effects; moreover, the dimension estimate is accompanied by confidence intervals that quantify the uncertainties owing to finite sample size.

Thus, it is apparent that there remain conflicts between those investigators who question the accuracy of results based upon quantifying relatively small data sets and those who maintain that these small data set results are indeed valid. In this thesis we determine criteria for datasets that can be used to estimate unambiguously the chaotic

behavior of a sampled attractor.

1.2. Objectives of Study

Despite the above-mentioned successes obtained with limited data sets, there is little written within the scientific community as to what constitutes unambiguously adequate data samples for quantifying chaotic systems. Because of the expense of these measures, the pervading arguments center on how to optimize chaos estimates using the minimum number of data points. Thus, undersampling of data sets is a key issue that many argue invalidates the results of many of these studies (cf. Grassberger 1986 vs. Nicolis and Nicolis 1986).

In this study, we seek to determine objectively the length of data sets that is necessary for adequately quantifying chaos measures. Because we intend to work with relatively large data sets, we use a chaotic dynamical system, the Lorenz (1963) model of convection. To avoid the large expense that would be required when using such large data sets to estimate the conventional chaotic measures, we develop a new, relatively inexpensive statistical measure, called the Histogram Measure, that quantifies essential

information about the chaotic solutions of the model.

We show that the utility of the Histogram Measure is indeed far-reaching. To find adequate samples of the data, we simply observe the intervals at which the histogram structures converge to within a reasonable tolerance, whether subjectively through superimposing them or more objectively, through mean difference calculations. In doing so, we find that the lengths of these optimum samples are dependent upon numerous factors, but are most sensitive to the time step that we use.

This measure also allows us to distinguish transient, or nonchaotic, subsets of the data that are inherent in chaotic time series. This is an important finding, as transients will contaminate the chaotic characteristics of any time series; eliminating the transient behavior in any time series is thus crucial for obtaining valid results. We find that the duration of the transient behavior is highly dependent upon the initial conditions that we use.

We also quantify the predictability characteristics of the adequate data samples, namely their information loss as functions of series length and elapsed time. We believe that these quantities may provide an inexpensive analog to the widely used Lyapunov exponents (Osledec 1968) or local divergence rates (Nese 1989). These rates of

information loss show exponential behaviors that are remarkably independent of sample size and initial condition.

To aid us in attempts at finding adequate samples with the Histogram Measure, we also use the more conventional Correlation Dimension Measure (Grassberger and Procaccia 1983b) because its results are well known. Although not able to use nearly the same amount of data with this measure that we can with the Histogram Measure, we do find fascinating quantitative links between the two. As a result, we theorize that the Histogram Measure may have far-reaching benefits in the quantification of chaotic time series.

By finding adequate samples of chaotic data, we overcome the problems involved with chaos estimates used on potentially undersampled data sets. In addition, we have a control case upon which to base *optimal sampling strategies* of the data. Based on sampling issues uncovered when finding adequate data sets (particularly that of the benefits of using a larger time step to sample the data over that of a smaller one), we theorize that these adequate data sets can indeed be sampled in ways that most likely preserve their chaotic characteristics, at least to within suitably small tolerances. If these strategies indeed prove viable, then chaos estimates can be obtained by using fewer points, which is particularly important when working with relatively long time

series.

In Chapter 2, we first describe in detail the characteristics of the Lorenz (1963) Rayleigh-Bénard convection model with which we have chosen to work. Having done this, we then explain the procedures that we use to obtain both the Histogram Measure and the standard Correlation Dimension Measure, providing introductory examples upon which to base further computation.

CHAPTER 2

A SIMPLE QUANTITATIVE MEASURE OF ATTRACTOR STRUCTURE

We wish to work with a nonlinear, deterministic system that exhibits chaotic behavior, is well studied, and is an archetypical model of the atmosphere. Edward Lorenz's three-component convection model (Lorenz 1963) fits these criteria nicely. During the past 30 years, this simple model has been used to forge a basic understanding of predictability characterisitics and their possible links to the chaotic behavior seen in the atmosphere. However, many of the conventional measures presently used to quantify the characteristics of this chaotic attractor are time-consuming, and thus expensive, to use. In this chapter, we introduce a new, simple measure, the Histogram Measure, that not only provides us with essential quantitative information about this chaotic attractor, but at a much lower cost.

2.1. Lorenz Rayleigh-Bénard Convection Model

Before discussing how the Histogram Measure works, we first need to understand fully the motivations for and intricacies of the Lorenz model and its solutions. Classical Rayleigh-Bénard convection describes the evolution of fluid motions as functions of heating from below and cooling from above. As the fluid is initially warmed, heat is transported vertically by conduction, or molecular processes, until a critical temperature gradient across the fluid domain is reached. Above this critical temperature gradient, heat is transported instead by a convective, or overturning, process that causes roll-like motions to occur. The Lorenz (1963) three-component model of classical Rayleigh-Bénard convection provides the simplest representation of some of these convective states (Shirer 1987b).

The set of three ordinary differential equations of the Lorenz model describes a shallow, two-dimensional, incompressible Boussinesq flow. Using trigonometric functions to describe the velocity and temperature characteristics of the flow, restricting the choice of vertical and horizontal wavenumbers to n=1, m=1 and n=2, m=0 respectively, and integrating over the cyclic, dimensionless domain $0 \leq x \leq 2\pi$ and $0 \leq$

$z \leq \pi$ yields the spectral system (Shirer 1987b)

$$\dot{X} = \frac{dX}{dt} = -PX + PY , \tag{2.1}$$

$$\dot{Y} = \frac{dY}{dt} = -XZ + RX - Y , \tag{2.2}$$

$$\dot{Z} = \frac{dZ}{dt} = XY - BZ , \tag{2.3}$$

in which X describes the temporal behavior of the velocity field and both Y and Z describe the temporal behavior of the temperature field. The variable P is the Prandtl number that measures the relative efficiency of the dissipation of momentum to that of heat. The normalized Rayleigh number R is the ratio of the vertical temperature difference driving the flow to the viscous effects that retard it and is the forcing parameter for this system. The variable $B = 4/(1+a^2)$ is related to the aspect ratio a of the flow field; this ratio is defined to be the height divided by the half-width of the convective domain. The time-dependent solutions X(t), Y(t), and Z(t) are confined to a three-dimensional region in phase space. In contrast to physical space, which we most commonly use to describe fluid motions via the temporal variations of certain scalar and vector quantities at fixed spatial positions, phase space solutions are those in which fluid motions are represented by the temporal variations of the amplitudes of specified

waveforms (Higgins 1987). Indeed, physical and phase space comparisons are easily illustrated graphically as seen in Figure 2.1.

The three-component Lorenz model (2.1-2.3), as with most similar systems, can only be solved analytically in certain special cases. Here, we are forced to determine its chaotic solutions using a numerical method. Over the years, many methods have been employed successfully to find the temporal solutions to this nonlinear hydrodynamical system. We decided to use one of the more efficient and accurate differential equation solvers known as the IMSL subroutine DVERK that uses a fifth-order Runge-Kutta scheme. All of the results in this thesis are based on time series produced by this numerical method. After specifying some initial condition for each of the phase space variables X, Y, and Z in (2.1-2.3) and after fixing the values of B, P, and R, we use this routine to provide a value for all three variables at each time step. The values of B, P, and R that we use are the standard ones originally chosen by Lorenz and by most others studying this system: B=8/3 , P=10, and R=28. For these particular values of B and P, this value of R is greater than the critical value $R_H \sim 24.74$ at which the steady convective solutions become unstable and the only stable solution is chaotic. For now, we choose a relatively small time step of $t_S = 0.005$ to ensure an accurate representation of the model solutions. The chaotic, or Lorenz, attractor is visualized most readily via

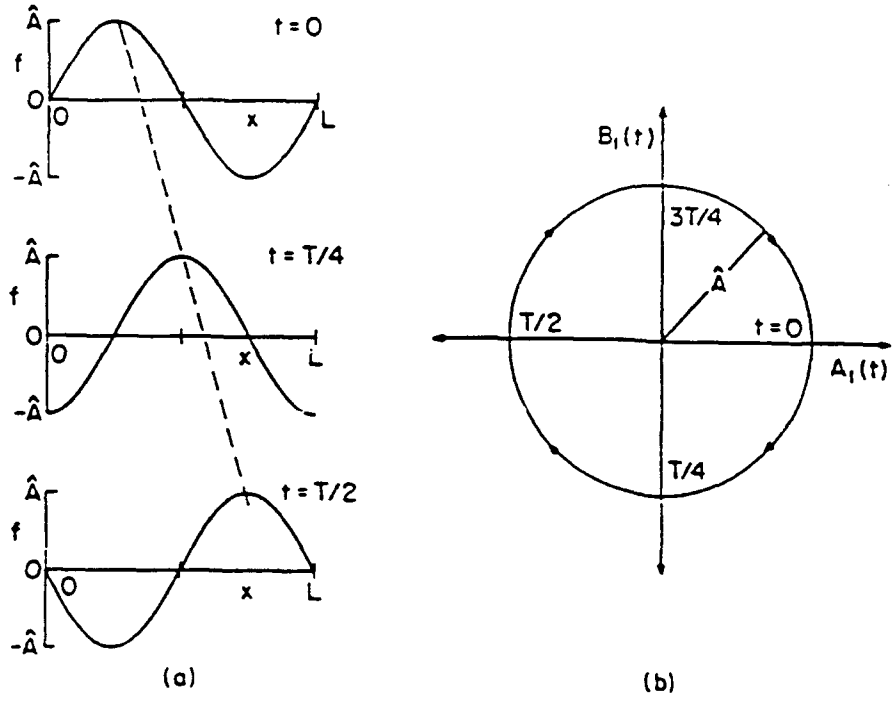


Figure 2.1: A representation of linear wave translation in physical space (a) and the corresponding trajectory in phase space (b). In (b) the arrows point in the direction of increasing values of t (from Higgins, 1987).

plotting one of the variables X, Y, or Z against any or both of the remaining two (Lorenz 1963).

The existence of this chaotic attractor needs explanation. In the chaotic regime, the phase space solutions are characterized by three Lyapunov exponents--one positive, one zero, and one negative (Nese 1987). Together, these exponents give a quantitative description of the average stability properties of orbits on the phase space attractor. Negative Lyapunov exponents measure the average rate of exponential convergence of trajectories onto and within the attractor, while positive exponents measure the average rate of exponential divergence. For the Lorenz attractor, the positive Lyapunov exponent is associated with the growth of solutions away from the unstable convective solutions and accounts for the stretching of the attractor along the unstable manifold. The negative Lyapunov exponent is associated with dissipation in the system and accounts for shrinking along the stable manifold. However, this argument alone does not explain the unique geometry exhibited by this attractor. For this and other chaotic dynamical systems, it is possible to define the existence of a trapping or bounded region in phase space. The existence of such a trapping region ensures that all orbits in the neighborhood of any fixed point are bounded, regardless of their local stability properties (Dutton and Wells 1984). These effects are illustrated in Figure 2.2, the

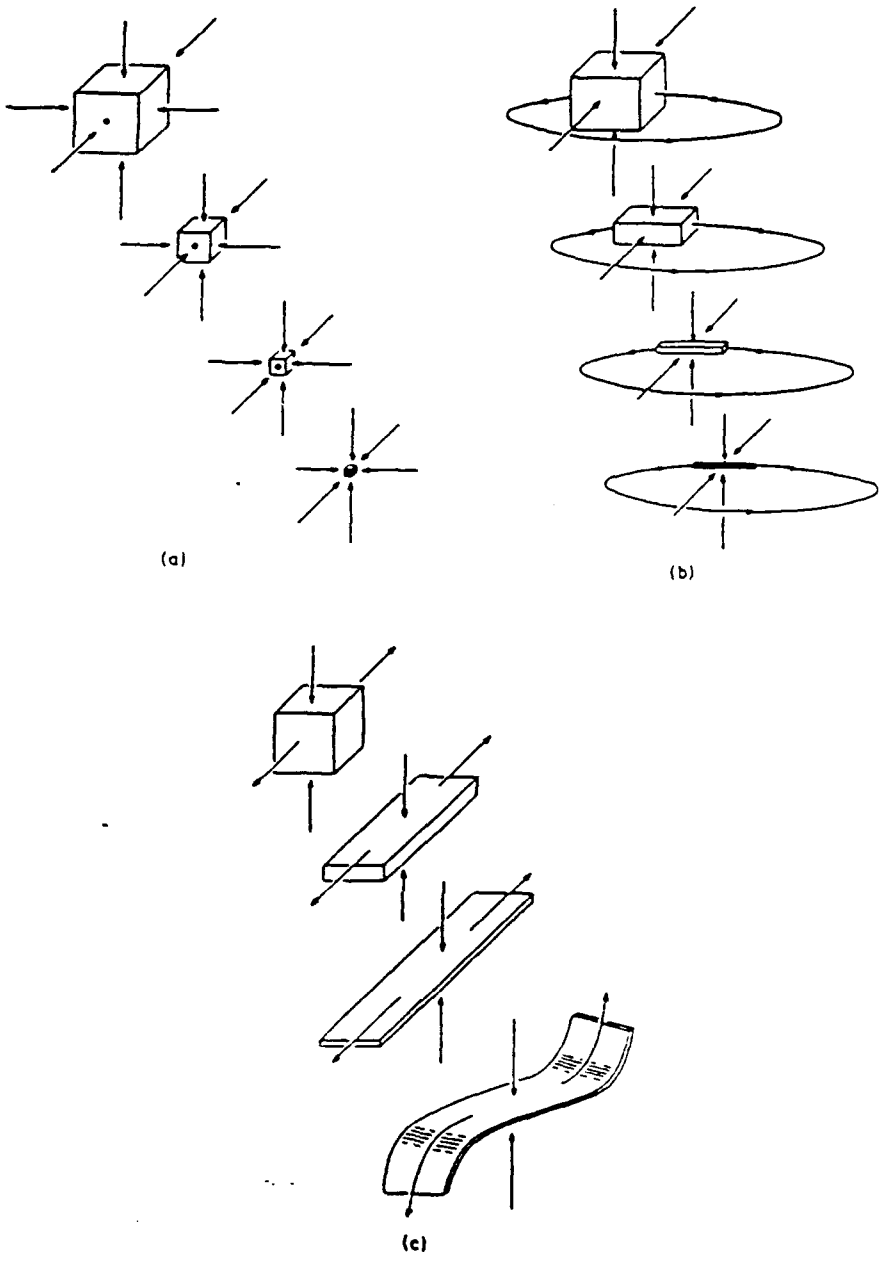


Figure 2.2: The evolution of a three-dimensional phase space volume for three types of attractors: (a) For a stable fixed point, all three Lyapunov exponents are negative and the volume contracts in all three directions. (b) For a stable periodic attractor, two exponents are negative and one zero, so the volume contracts in only two of the three directions. (c) For a strange attractor, one exponent is positive, so the volume evolves into a sheet that, because of its boundedness in phase space, is infinitely folded by the flow (from Nese, 1987).

combination of which yield the Lorenz attractor in which sheets of divergent trajectories are infinitely folded by the bounded flow (Nese 1987). Figures 2.3-2.5 show the two-dimensional projections of the Lorenz attractor that we produced with the standard parameter values; shown are the values from the last 5,000 time steps of a 15,000-step integration for which the initial conditions are X=0, Y=1, Z=0. Confident that the integration scheme is producing the correct model representation, we now describe the new measure that we have developed, one that quantifies the chaotic structure and characteristics of the model solutions.

2.2. The Histogram Measure

We seek a measure for accurately quantifying the structure of a chaotic attractor in a computationally inexpensive manner. Now as we run the model integration, instead of providing only the value of each phase-space variable, we also calculate a three-dimensional Euclidian distance d_U given by

$$d_U = \left[(X - X_0)^2 + (Y - Y_0)^2 + (Z - Z_0)^2 \right]^{1/2}, \qquad (2.4)$$

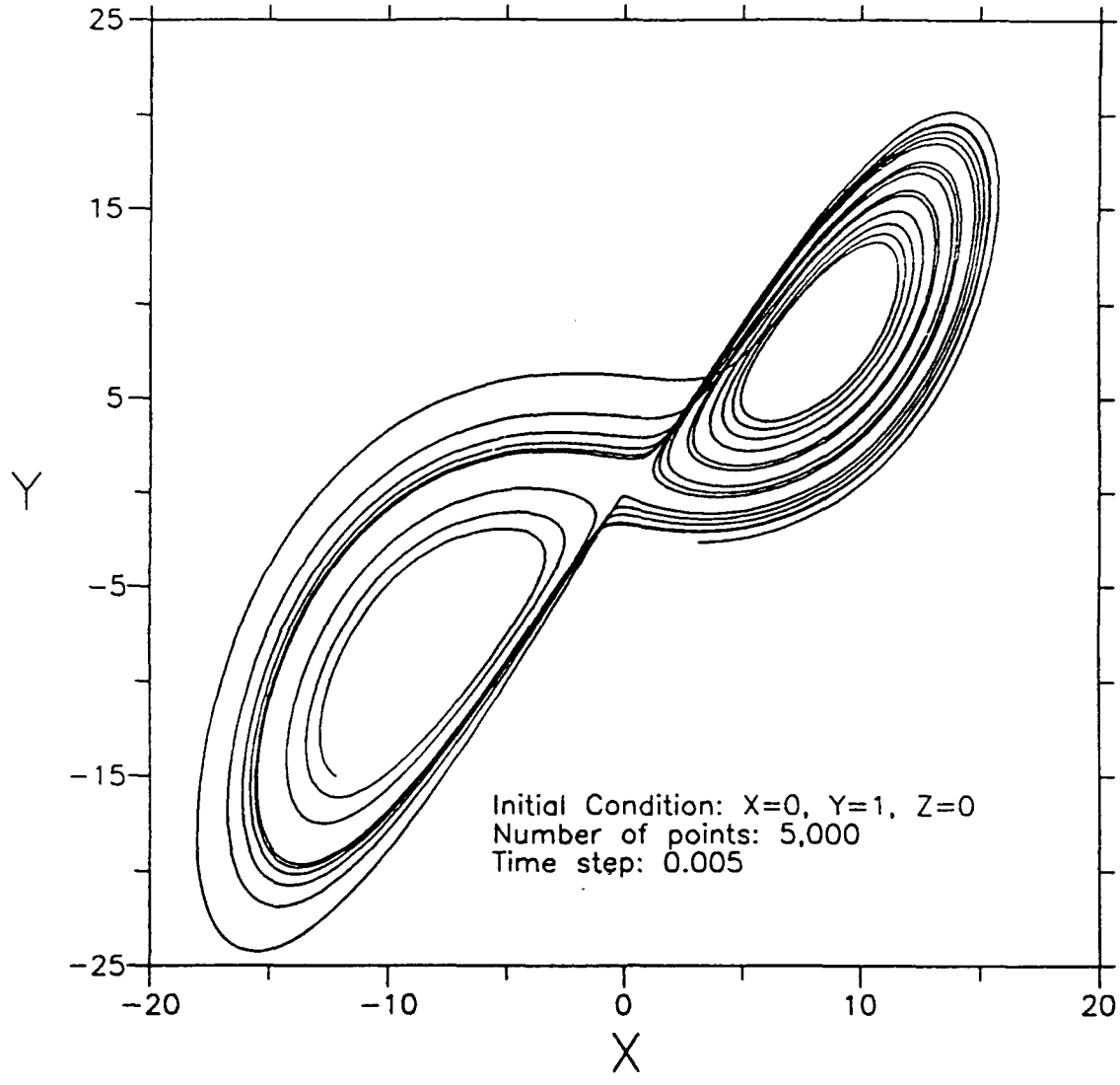


Figure 2.3: The Lorenz attractor in X−Y phase space representing the last 5,000 points of a 15,000−point data set.

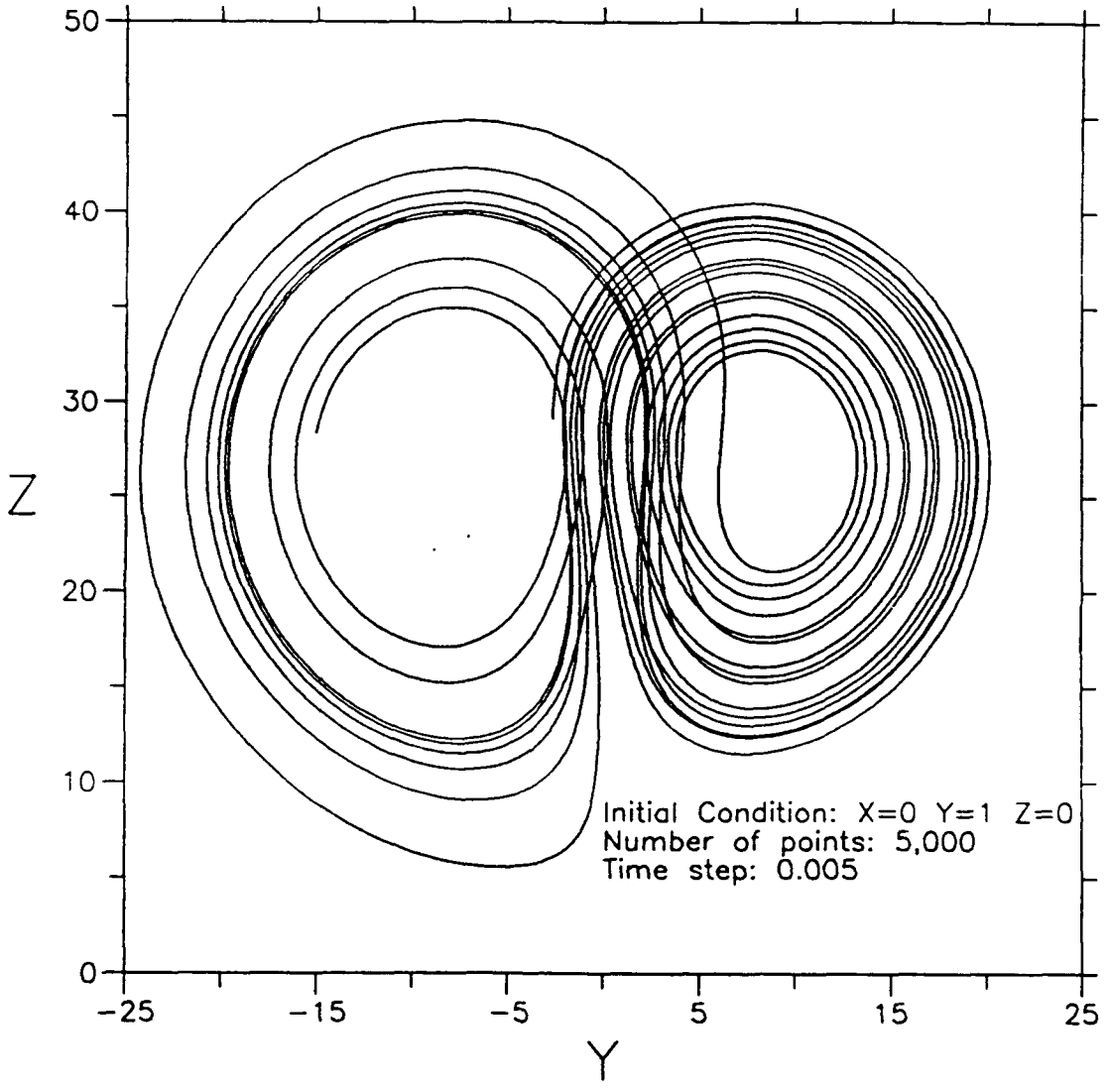


Figure 2.4: The Lorenz attractor in Y−Z phase space representing the last 5,000 points of a 15,000−point data set.

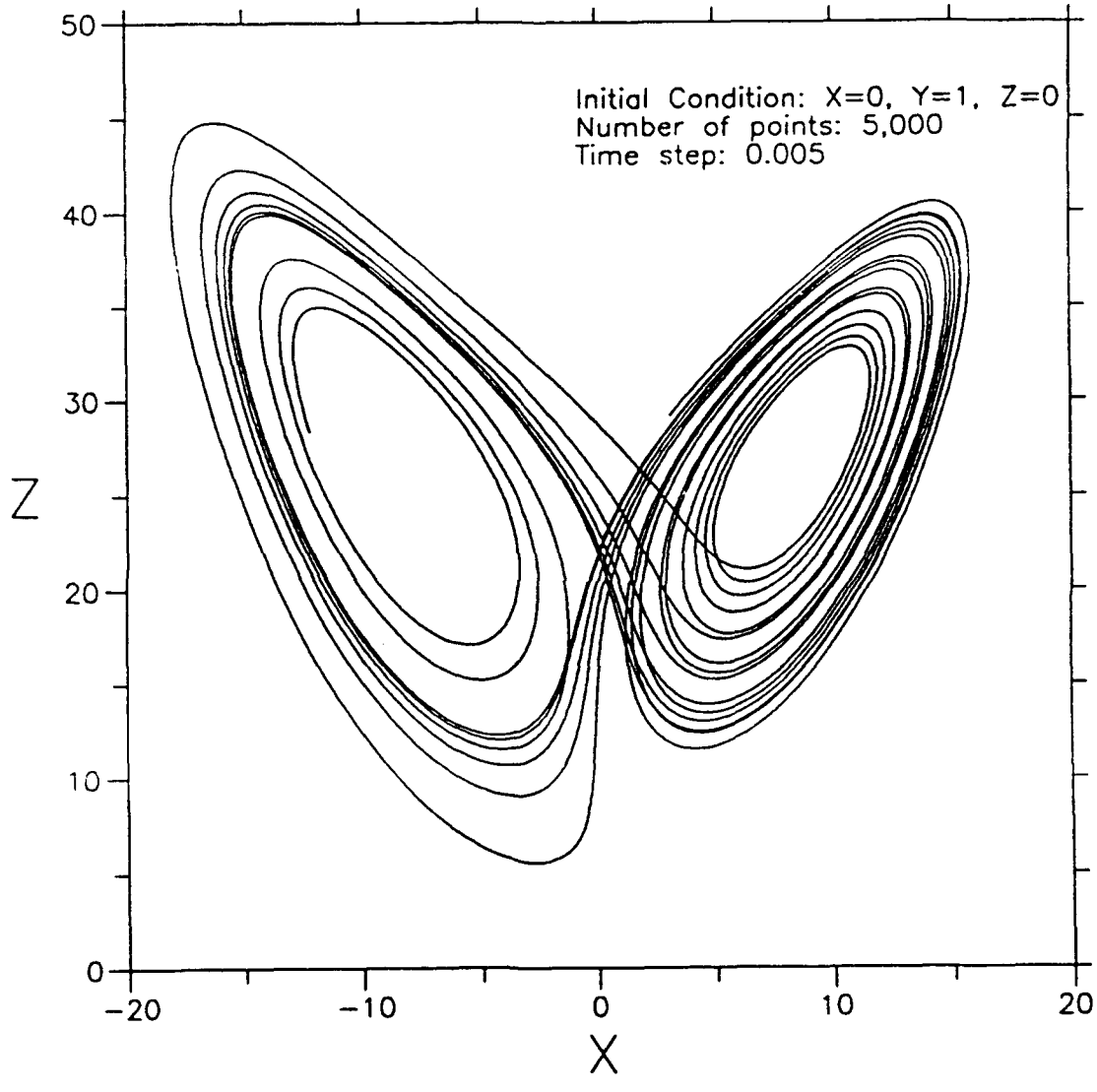


Figure 2.5: The Lorenz attractor in X–Z phase space representing the last 5,000 points of a 15,000–point data set. Note the classical butterfly shape.

in which X_0, Y_0, and Z_0 represent the coordinates of some specified reference point. Once this distance has been calculated, it is placed into a specified *bin* or distance interval. When the integration is complete, the number of points in each bin is normalized by the total number of points produced by the integration. Normalizing the results ensures that we can easily compare results obtained when using different numbers of points to represent the solution.

To fully understand the intricacies of this measure, we first need to explain the motivations for the particular values that we use to generate it. First, we must choose a reference point (X_0, Y_0, Z_0). For simplicity, we picked the phase space origin (0,0,0). This is not, however, the only reason for our choice. In the Lorenz model, it is this unstable stationary point that represents the conductive solution and plays a unique role in both defining the stability characteristics of the system and in creating the chaotic attractor itself. As the forcing rate R is increased to one, this solution changes from a stable and globally attracting one to an unstable one from which trajectories tend to move toward the two convective solutions (Sparrow 1982). In the chaotic regime of this model, the origin and nearby points in phase space on the attractor, although rarely visited, help define the unstable manifold along which trajectories diverge. With the

origin as our reference point, we seek the rate of occurrence of these rare events.

Second, we need to determine how to bin the distances. To do this, we must ensure that we capture all of the possible distances from the origin in phase space; that is, we first need to know the maximum Euclidean distance d_M. Once this value is found, we must define a constant bin width value b_w for each bin. Specifying the number of bins N_b yields the following relation:

$$b_w = \frac{d_M}{N_b}. \tag{2.5}$$

The maximum Euclidean distance is estimated by simply examining the attractor plots shown in Figures 2.3-2.5 and noting the associated phase space distances. Upon doing this, we choose $d_M = 50$. Using 128 distance bins N_b, we obtain a bin width value b_w of 0.4. For now, we reserve further comment on the bin width.

Finally, we must specify the initial conditions. This specification is most important, as we want eventually to discover a *control histogram* that is independent of the initial condition used. To do this, we choose three separate sets of initial conditions. These values are based upon radial distances from one of the two unstable convective solutions. These two solutions are given by

$$X_C = Y_C = \pm \left(B\,(R\text{-}1) \right)^{1/2} ;\; Z_C = R\text{-}1 \,, \qquad (2.6)$$

We arbitrarily choose the solution with *positive* values, although it does not matter which we choose because the attractor is symmetric. The three sets of initial conditions used are 1.0, 0.10, and 0.01 percent perturbations from this solution. Specifically, these sets correspond to $(X_0, Y_0, Z_0) = (1.01\,X_C\,,\, 1.01\,Y_C\,,\, 1.01\,Z_C)$, $(1.001\,X_C\,,\, 1.001\,Y_C\,,\, 1.001\,Z_C)$, and $(1.0001\,X_C\,,\, 1.0001\,Y_C\,,\, 1.0001\,Z_C)$ respectively.

To illustrate the information provided by this measure, we introduce examples to analyze briefly. In Figures 2.6-2.8 are shown the histograms that we produced for each of the three initial conditions by using the last 80,000 points of a 100,000-point series. On examining these figures, we observe a number of similarities and differences between them. The similarities involve the general shape of the histogram structure. Each has a bimodal appearance with closely corresponding primary and secondary maxima and a relative minimum between them. On either side of these maxima, the curve drops off quite dramatically toward zero. The primary maximum in the percentage of distance values h occurs between bins 45 and 55, with the secondary maximum occurring between bins 85 and 95; these bins correspond to phase space

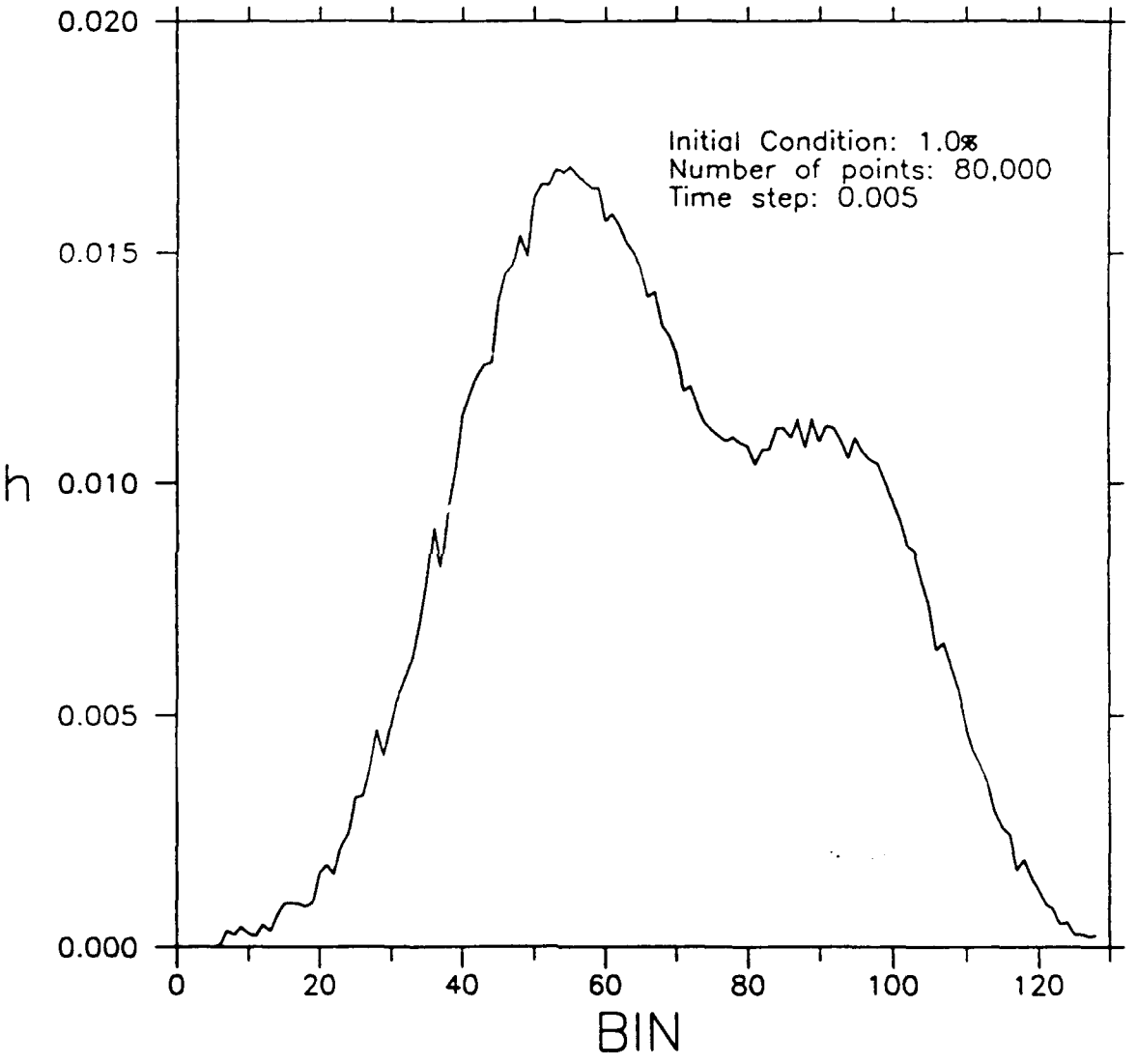


Figure 2.6: The normalized histogram for the 1.0% initial condition representing the last 80,000 points of a 100,000-point data set. Note the spiky appearance.

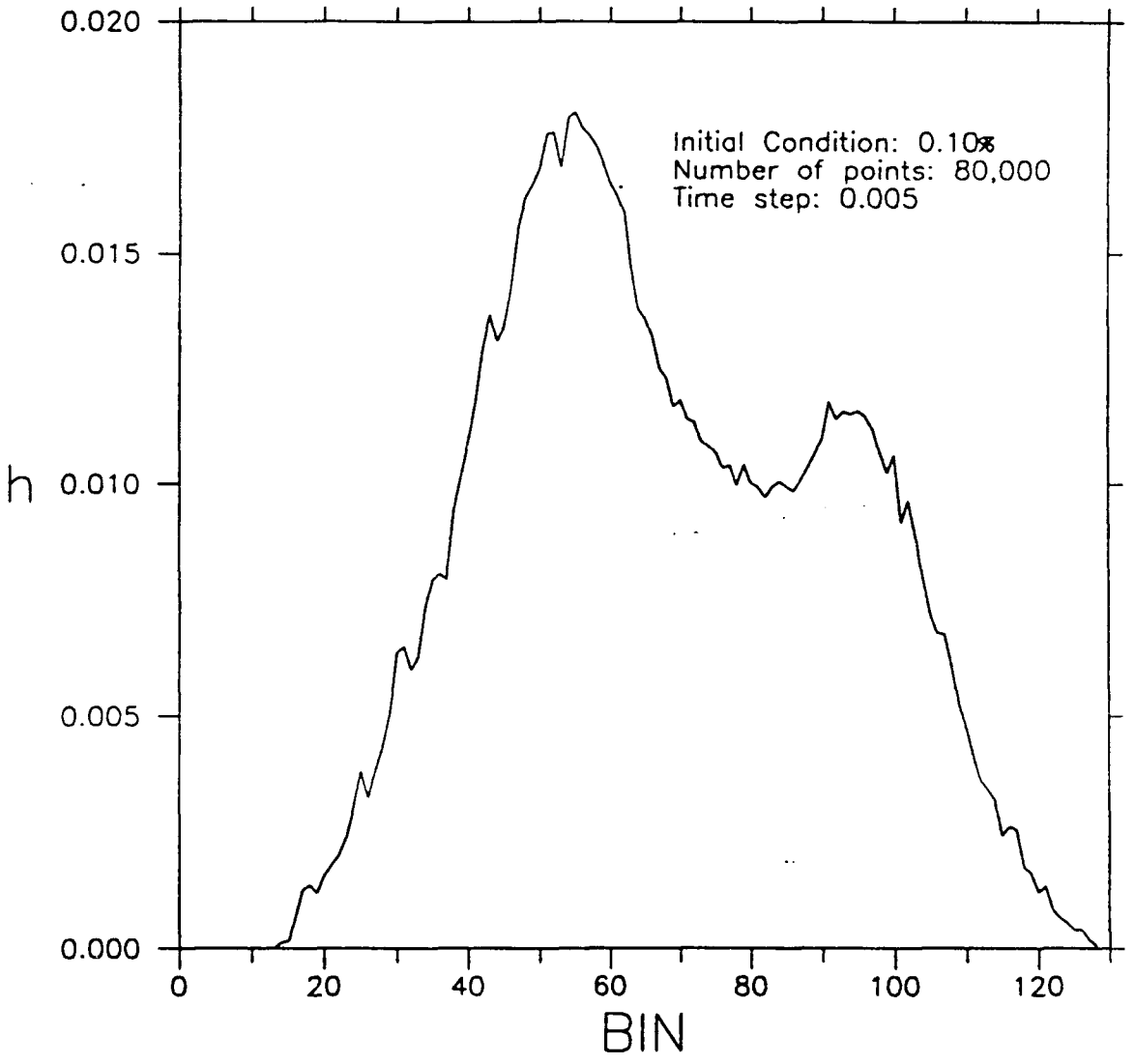


Figure 2.7: The normalized histogram for the 0.10% initial condition representing the last 80,000 points of a 100,000-point data set. Note the spiky appearance.

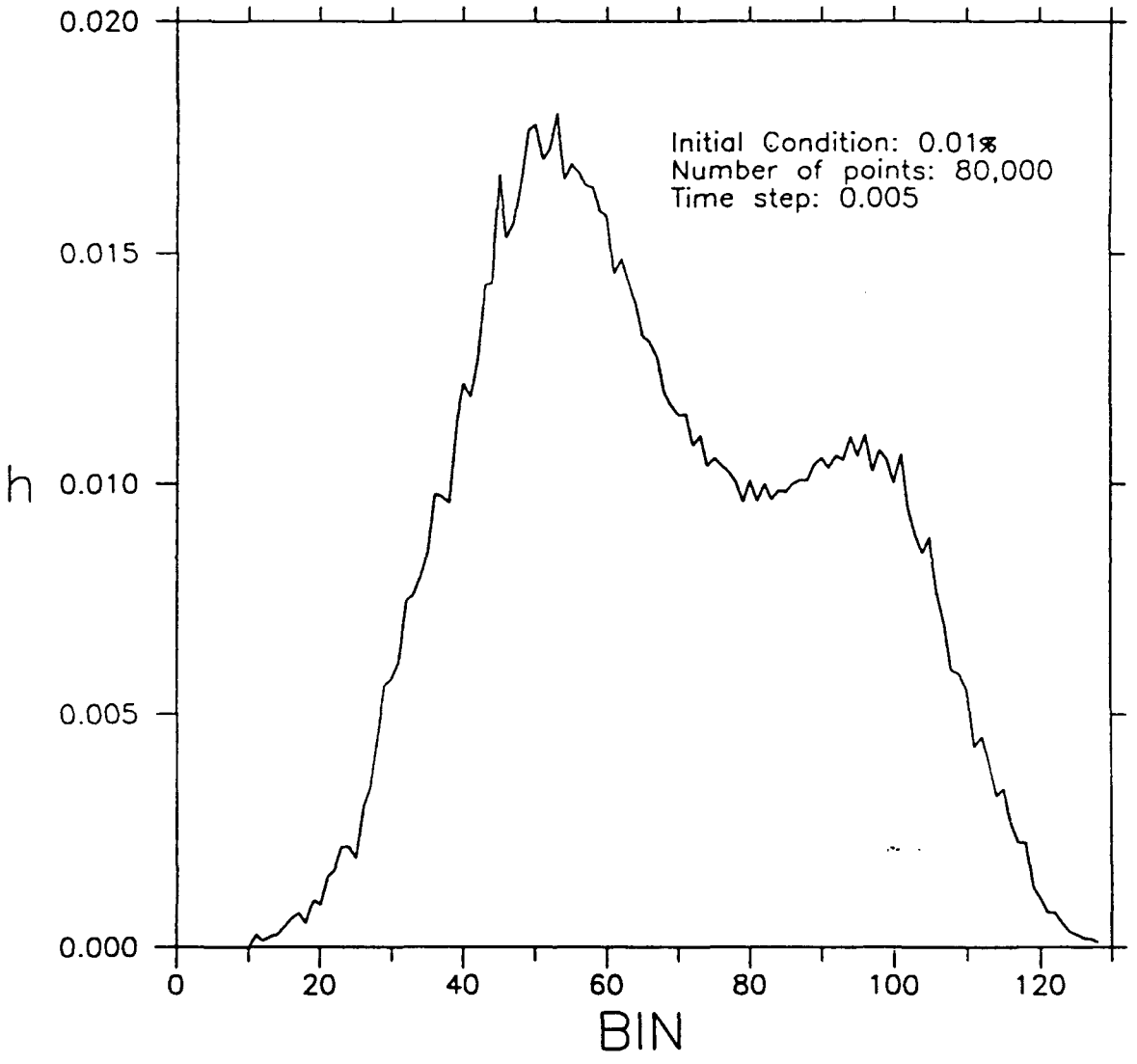


Figure 2.8: The normalized histogram for the 0.01% initial condition representing the last 80,000 points of a 100,000-point data set. Note the spiky appearance.

distances of approximately 18 to 22 and 34 to 38, respectively. The relative minimum exists between bins 75 and 85, or at distances of 30 to 34.

At first glance, the histogram shape is somewhat disturbing since the attractor is spatially distributed symmetrically about the Z-axis. However, the *time* spent in each portion of the attractor is not the same. That is, as the trajectory passes along the attractor, its speed depends upon its location on that attractor (Nese 1989). Thus, we conclude that the distance maxima correspond to those parts of the attractor in which trajectories are moving slowly, with the relative minimum corresponding to those parts in which the trajectories are moving quickly. The sharp decrease in the percentage of bin distances visited is simply related to the bounded nature of the attractor.

By superimposing the three figures as seen in Figure 2.9, we observe that remarkable differences also exist. These substantial differences are intriguing. Each form is somewhat jagged, suggesting a noisy behavior possibly related to insufficient sampling. In Figure 2.10 we show the results of applying a 1-3-1 smoother to the data in Figure 2.9. Although the magnitudes of the structural differences are slightly smaller, they do remain. These differences are troubling, since they imply structure that is *dependent* upon initial conditions and therefore inconsistent with that of other measures presently used to quantify chaotic time series. This result, however, may be just a

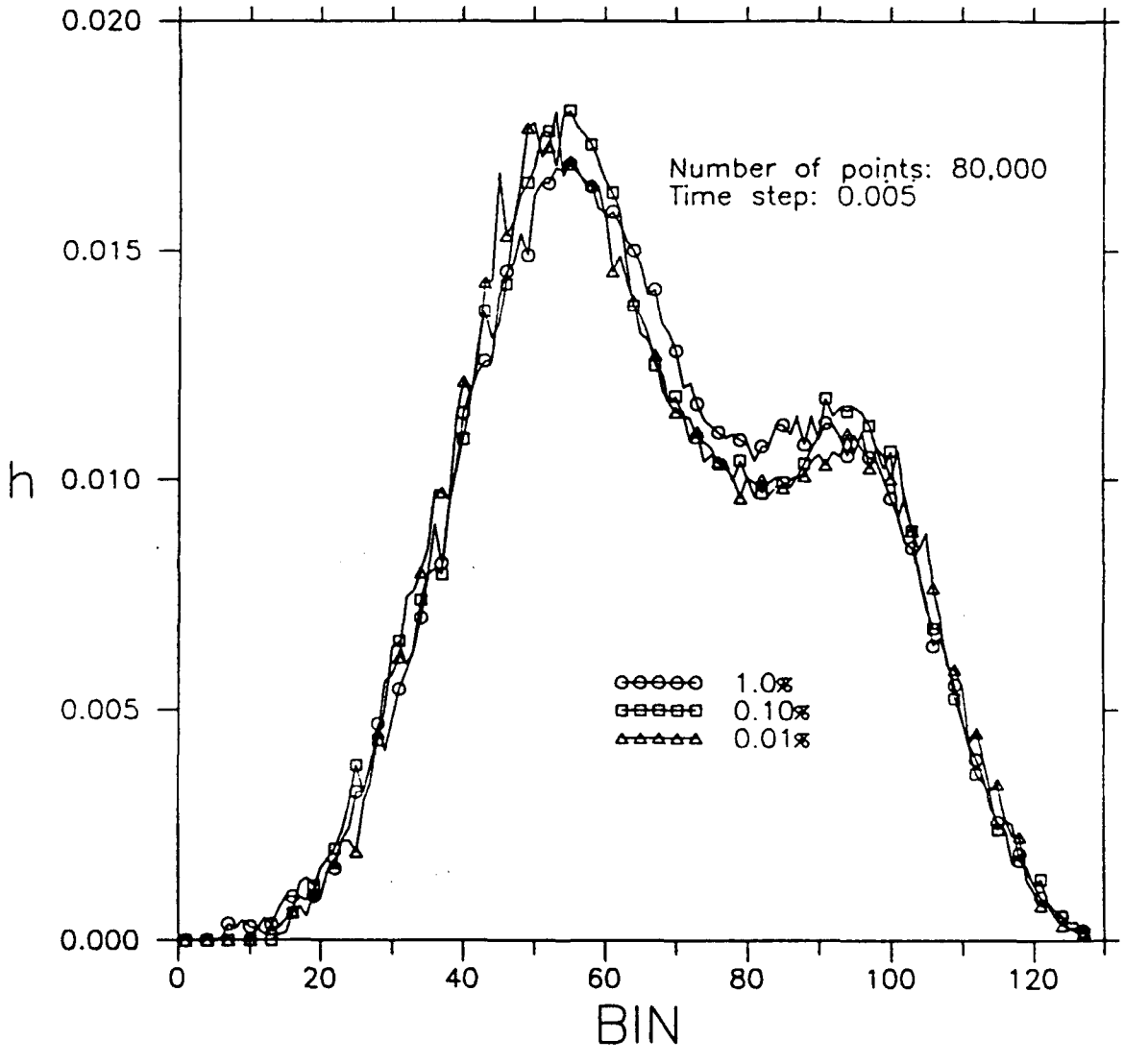


Figure 2.9: The normalized histograms superimposed for all three initial conditions representing the last 80,000 points of a 100,000–point data set.

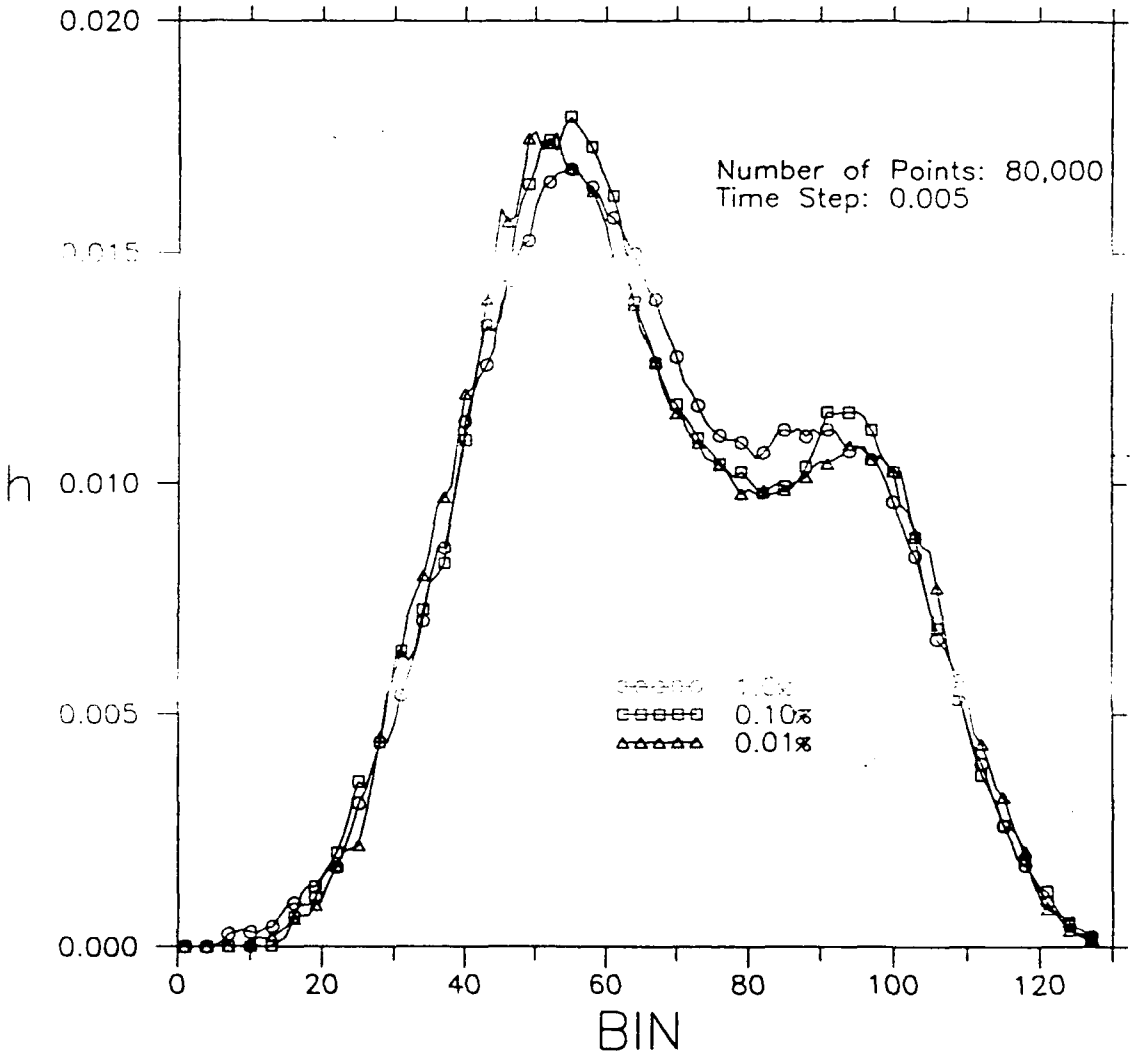


Figure 2.10: The normalized histograms superimposed for all three initial conditions and smoothed with a 1–3–1 filter, representing the last 80,000 points of a 100,000–point data set.

function of the way that we have sampled the data. Before making any firm conclusions, we must test the effects of including more points. It may be that we require more than 80,000 points for the measure to converge. We return to this and related considerations in Chapter 4.

To understand further the benefits that the Histogram Measure provides, we compare it throughout this thesis with a more conventional measure that is used to quantify chaotic attractor structure--the Correlation Dimension. By comparing a standard quantitative measure with the Histogram Measure, we will learn the utility of the measure as well as the data requirements for using it.

2.3. The Correlation Dimension Measure

Many measures have been developed to quantify the complex fractal structure of chaotic attractors. One of these is the Correlation Dimension Measure. Whereas measures such as the fractal and information dimension are based upon estimating the number of cubes of a certain size that are required to cover the attractor, the Correlation Dimension Measure is based upon the average density of a trajectory in neighborhoods

of points, and therefore on the distribution of nearby leaves of the attractor, as well as on certain trajectory recurrence characteristics. Thus, unlike these other dimensions, the correlation dimension captures some of the dynamics of the system (Nese *et al.* 1987).

The correlation dimension is calculated as follows (Henderson and Wells 1988). Let X_i, i=1,..., n be n points on an attractor in an N-dimensional phase space. A point X_j is selected from these data and the distances $\|X_i - X_j\|$ between this point and the remaining n-1 points are calculated using any normalized distance definition. Then, the number of points falling within a specified distance ε of the point X_j is calculated for numerous choices of ε. This process is repeated for all points X_i on the attractor and yields the Correlation Integral $C(\varepsilon)$ given by

$$C(\varepsilon) = \lim_{n \to \infty} \left[\frac{1}{n^2 - n} \sum_{\substack{j,k=1 \\ j \neq k}}^{n} H\left(\varepsilon - \|X_i - X_j\|\right) \right], \qquad (2.7)$$

in which H(y) is the Heaviside Function that has a value of one if $y \geq 0$ and zero otherwise. Finding $C(\varepsilon)$ is equivalent to calculating the density of points on an attractor within a range of distances ε from a point X_j and then finding the average of the density over all n points. Generally if our attractor is ν-dimensional, then we expect that

$$C(\varepsilon) = \varepsilon^{\nu}, \qquad (2.8)$$

where ν is the correlation dimension and measures the size of the subset of the phase space that is continually visited by the trajectory. Since there is no information on contraction or expansion in ν, this parameter can only quantify the geometric structure

The greatest advantage of the Correlation Dimension Measure is its ability to distinguish between deterministic chaos and random noise in an N-component data series, because in a random series, $C(\varepsilon) = \varepsilon^{N}$ (Grassberger and Procaccia 1983b). Its primary disadvantage is the lengthy computational time required to calculate it, and so it is relatively expensive to determine. Since its introduction by Grassberger and Procaccia (1983b), the Correlation Dimension Measure has been used to quantify the dimension of numerous attractors. For the Lorenz attractor, the value of ν is approximately 2.06.

To determine ν from a time series, we must also consider two other fundamental concepts: time lag and embedding dimension. The time lag Δt is related to the notion that knowledge of a sampled trajectory of points $[x(t), x(t+\Delta t)]$ is equivalent to the

knowledge of the actual trajectory of points $[x(t), \dot{x}(t)]$ (Takens 1981). Unfortunately, the optimum choice of Δt is generally a matter of trial and error. If Δt is too small, then the points $x(t)$ and $x(t+\Delta t)$ are not independent and will mimic a one-dimensional system; if Δt is too large, then we undersample the time series. Assuming the Shadowing Lemma is valid (i.e. the Lorenz attractor is sufficiently hyperbolic), then the

sufficiently such that the spatial correlation is maintained.

The embedding dimension d_E is related to the idea that a phase space trajectory $x(t)=[x_1(t), x_2(t), x_3(t), \ldots, x_n(t)]$, given that the above theory holds, can be replaced by a trajectory in some m-dimensional *artificial phase space* where $x(t)=[y(t), y(t+\Delta t), \ldots, y(t+m\Delta t)]$ and m is the number of times that the series is lagged by Δt from $y(t) = x_1(t)$. Thus, a sequence of points representing the attractor can be constructed in this artificial phase space. For any chaotic attractor, the embedding dimension d_E is the mimimum dimensionality m of the artificial phase space that is necessary to capture the attractor. For n-dimensional smooth manifold dynamical systems, it has been shown that $d_E < 2n+1$; for more general attractors, it has been conjectured that $d_E \sim 2r+1$ where r is the attractor dimension (Packard *et al.* 1980; Takens 1981). Given this relation for the Lorenz attractor, we should see convergence of the correlation dimension by an

embedding dimension of approximately seven.

Therefore, we must first specify both a proper time lag and an embedding dimension. The embedding dimension is found by calculating the value of ν for increasing values of m until convergence is seen (e.g., Henderson and Wells 1988). The appropriate time lag, however, is harder to determine. To find this quantity, we rely on another qualitative procedure--that of *model reconstruction* (Takens 1981). The appropriate time lag is the one that provides us with the optimum reconstructed view of our attractor. By simply plotting the value of one variable from a time series with a single time-lagged value of that variable, we seek the two-dimensional view that best simulates the actual two-dimensional attractor. In Figures 2.11-2.13, we show the results of plotting the X(t) data series against the lagged series X(t+Δt), where Δt=5, Δt=20, and Δt=40, respectively; many others values of the time lag were investigated as well. Comparing these three reconstructed attractor plots against the *true* attractor representation shown in Figure 2.3, we conclude that the time lag value Δt = 20 is the optimum one for a time series created with a time step t_S=0.005.

Using the values of d_E and Δt, we can next calculate the correlation dimension for any time series that we create; here we use the standard Euclidean distance in our dimension calculation. The classical way to determine the correlation dimension is to

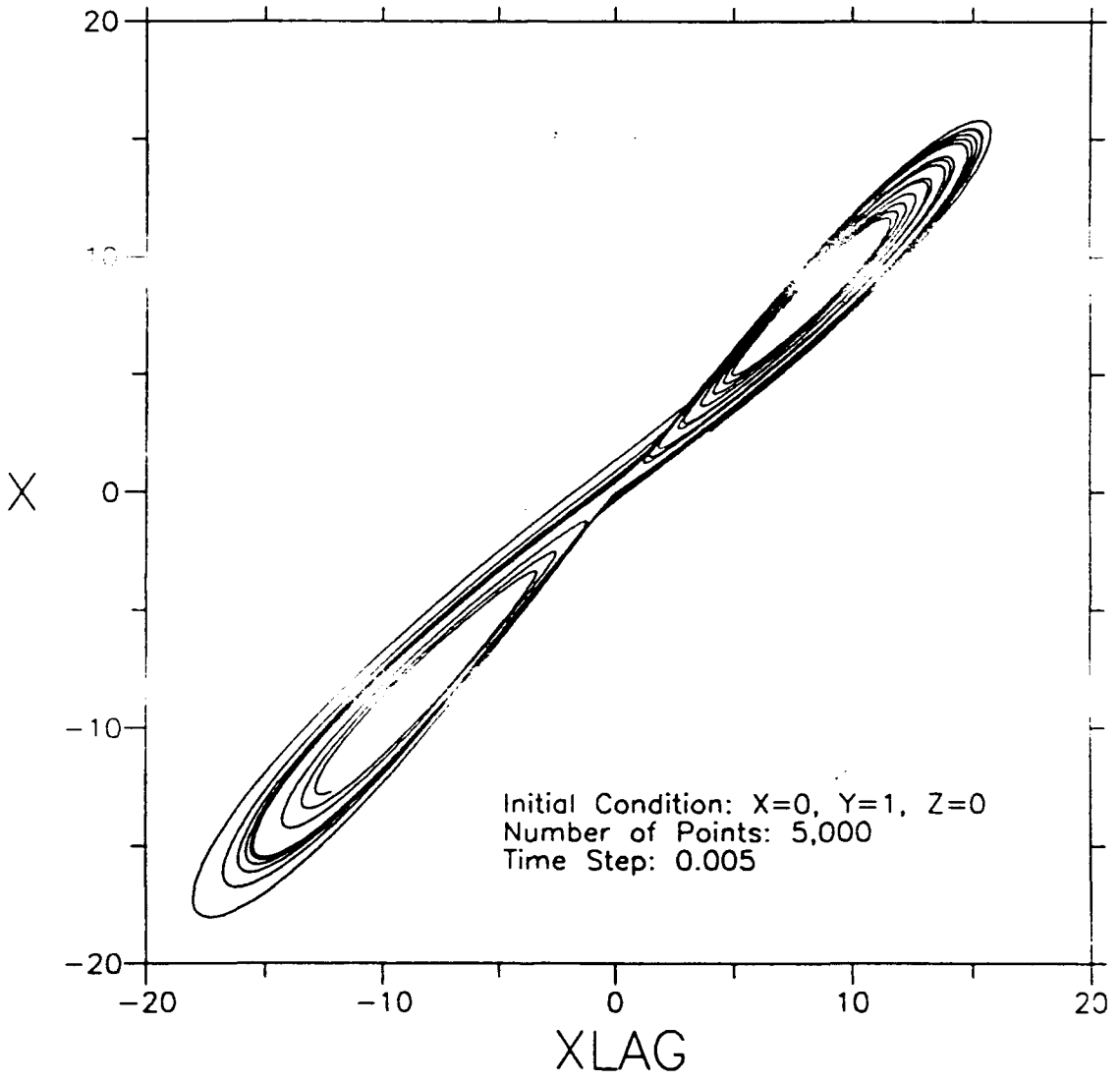


Figure 2.11: The reconstructed Lorenz attractor versus its lagged series XLAG for a time lag value of five. Note that this shape does not best simulate the actual attractor structure.

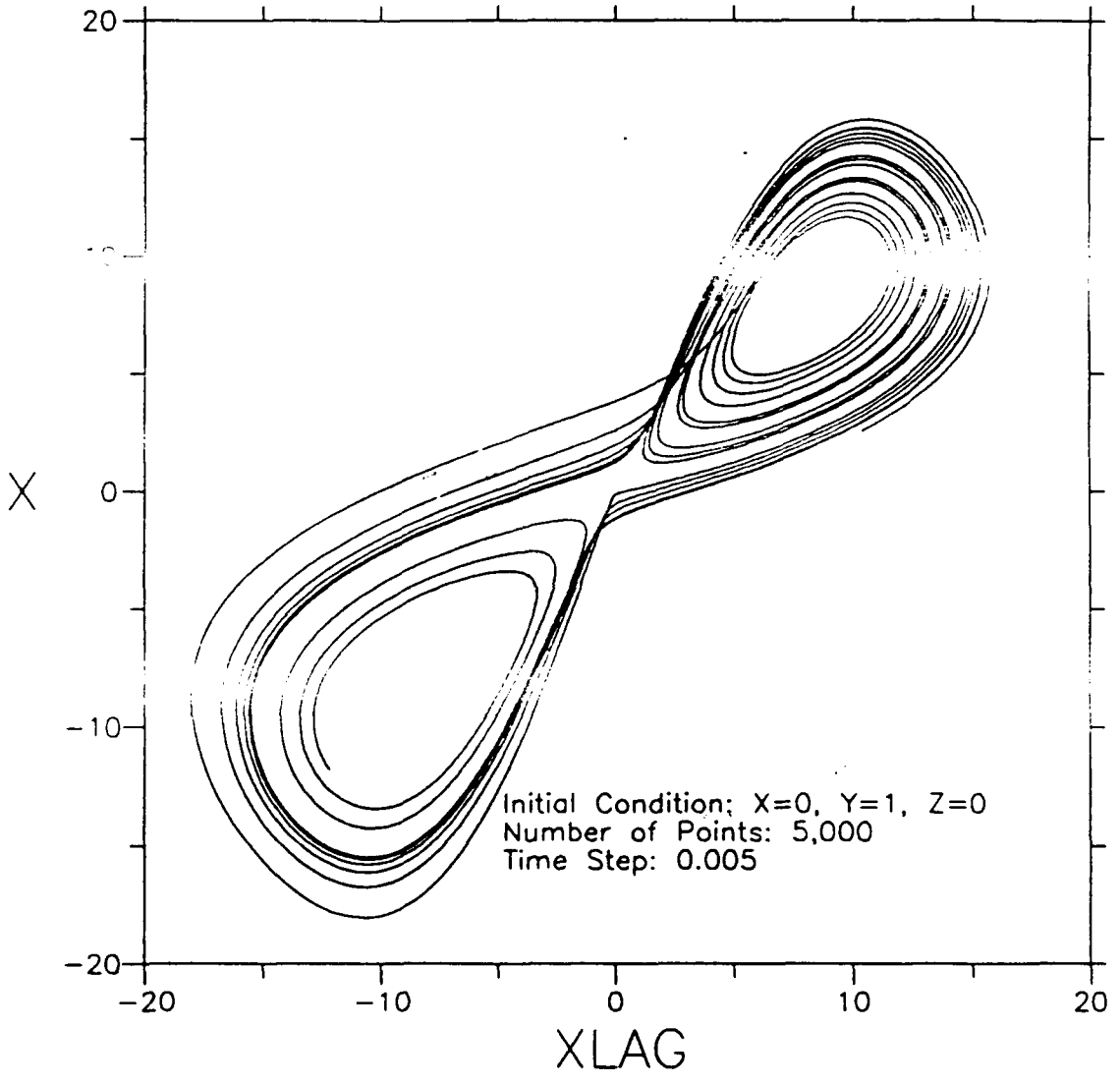


Figure 2.12: The reconstructed Lorenz attractor versus its lagged series XLAG for a time lag value of 20. Note that this shape best simulates the actual attractor structure.

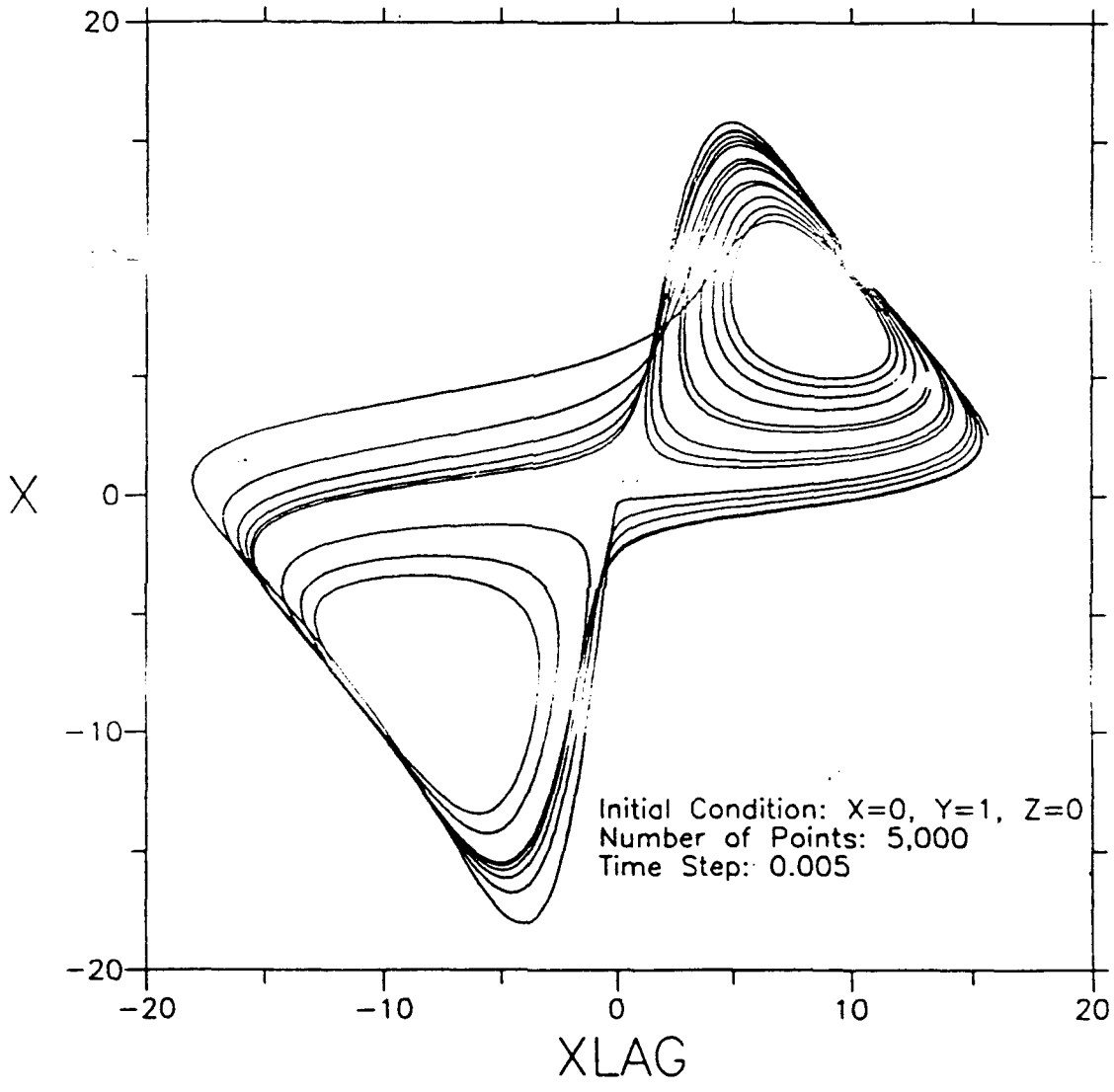


Figure 2.13: The reconstructed Lorenz attractor versus its lagged series XLAG for a time lag value of 40. Note that this shape does not best simulate that of the actual attractor structure.

find the slope of a plot of ln[C(ε)] versus ln(ε) (Grassberger and Procaccia 1983b) since $\nu = \ln C(\varepsilon)/\ln(\varepsilon)$ by (2.8). For this introductory example, we have specified integer values of the critical distance ε. Figure 2.14 shows a correlation dimension plot that we have produced using a time series that represents the last 10,000 points of a 100,000-point data set with the 0.01 percent initial condition. We use a time step t_S of 0.005, a time lag At of 20, and an embedding dimension d_E of seven. The correlation dimension value is calculated by determining the slope of the curve. This can be done with more sophisticated techniques such as finite-differencing, or more simply, by eye. However, regardless of the method we choose, finding the appropriate value of the slope is still a highly subjective process. Although this particular case, Figure 2.14, is relatively well-behaved, upon closer inspection, the curve does not increase with increasing ε in a perfectly linear manner, and so we can not be certain of the appropriate ε interval in which to estimate the dimension; this appropriate interval is known as the *scaling region*. The very small ε range is sampled too infrequently to truly represent the attractor structure, and so these results most likely include noise. Too large a value of ε provides too coarse a resolution and so misses relevant details in the structure. Thus, the resultant value of our correlation dimension ν is highly dependent on the ε range that is examined.

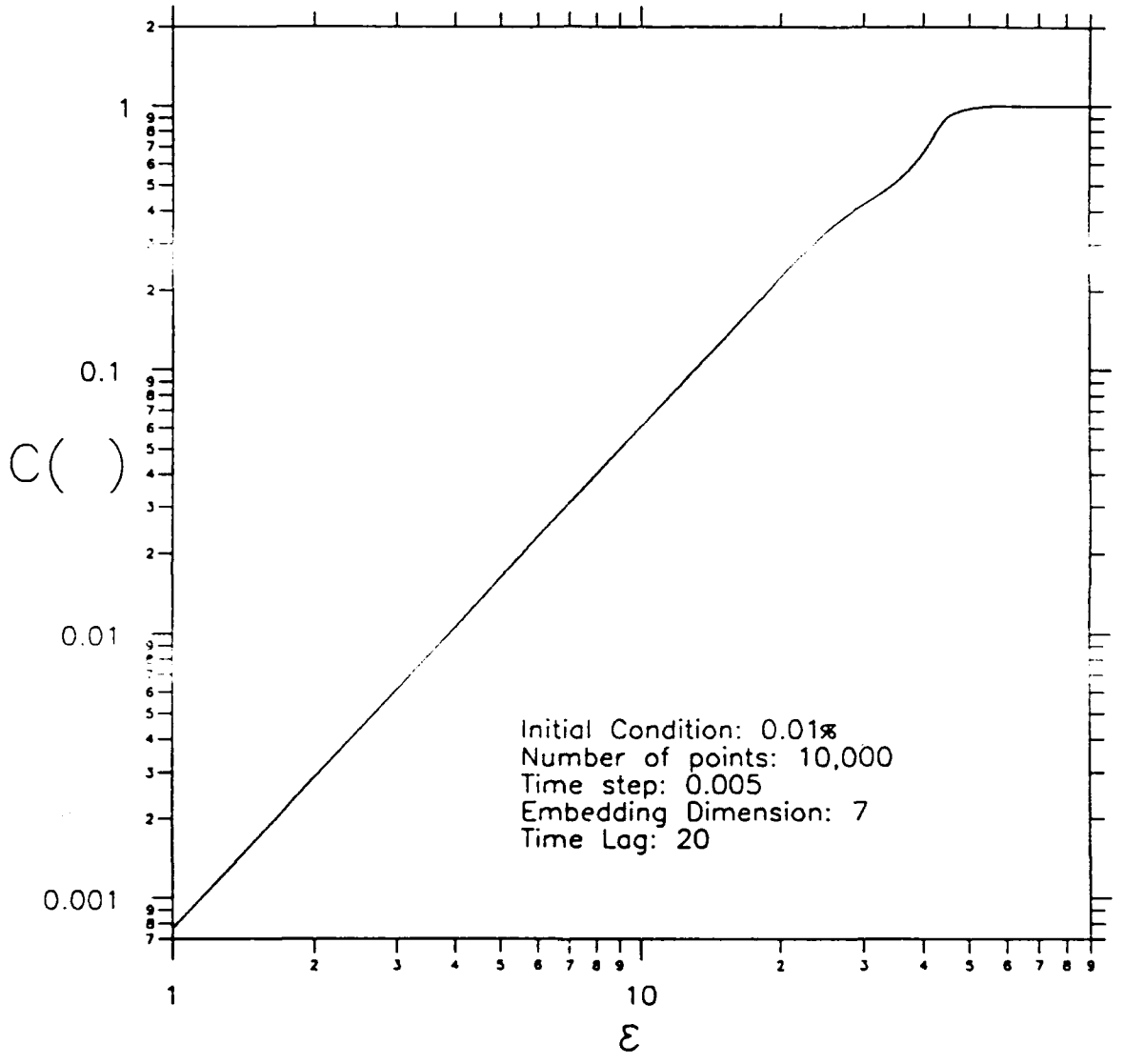


Figure 2.14: The classical way to view the correlation dimension. Its value is determined by calculating the slope of the line.

A more objective way of calculating ν is to employ a slope-finding algorithm to the above curve. Once we plot the slope value as a function of the distance ε, we can determine the scaling region. For simplicity, we employ a scheme based on a Taylor series expansion using three adjacent, unequally spaced points ($\varepsilon - h_2$, ε, $\varepsilon + h_1$) that is given by

$$C'(\varepsilon) \cong \left[\frac{h_2^2 \ln C(\varepsilon + \Delta\varepsilon) + (h_1^2 - h_2^2) \ln C(\varepsilon) - h_1^2 \ln C(\varepsilon - \Delta\varepsilon)}{(h_1 h_2^2 + h_2 h_1^2)} \right], \qquad (2.9)$$

where

$$h_1 = \ln(\varepsilon + \Delta\varepsilon) - \ln(\varepsilon), \qquad (2.10)$$

$$h_2 = \ln(\varepsilon) - \ln(\varepsilon - \Delta\varepsilon). \qquad (2.11)$$

The expressions for h_1 and h_2 involve natural logarithms owing to the logarithmic form of the data. In Figure 2.15, we show the result of using this slope estimate on the same data set used to generate Figure 2.14. We observe that we have slope values for all of the ε distances; these values represent those of the correlation dimension ν. Although we do not obtain *exactly* the accepted value of 2.06 for this particular case, the values of ν that we do obtain are close enough to warrant our confidence in the above

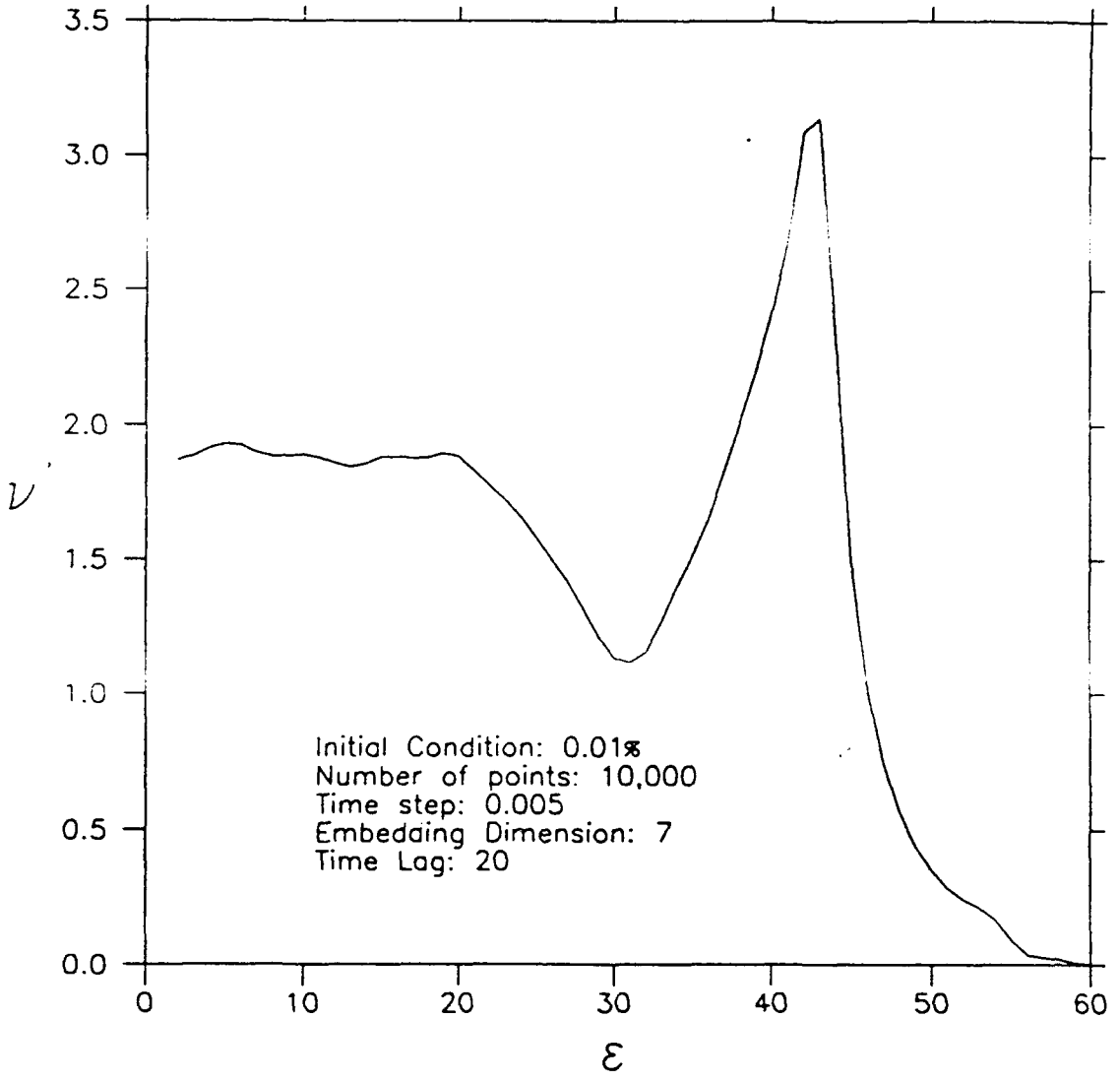


Figure 2.15: The correlation dimension value as a function of bin distance using the slope formula. Note a relatively stable value of correlation dimension (approximately 1.9) through the first 20 distance bins.

slope-finding method. There are many reasons why we did not obtain the expected value; we address this issue further in Chapter 4.

As with the previous section explaining the Histogram Measure, this section is meant strictly to give an introduction to the Correlation Dimension Measure. In Chapter 4, we describe in detail how sensitive this measure is to changes in the parameter values, what these sensitivities tell us about how we sample the attractor, and what its importance is in helping us understand the capabilities of our Histogram Measure. Before we do that, however, we investigate an issue that is important in working with any chaotic time series. We want to ensure that any time series that we use is not contaminated by nonchaotic behavior arising from transients. In the next chapter, we seek to identify these solutions.

CHAPTER 3

IDENTIFYING TRANSIENTS WITH THE HISTOGRAM MEASURE

Given that we normally do not choose an initial condition on the attractor itself, the time-dependent solutions of the Lorenz model must evolve toward the chaotic one. The details of this evolutionary behavior are dependent upon numerous factors, such as integration method, the values of parameters B, P, and R that are chosen, and the initial conditions that are used (Lorenz 1963). These solutions show an amplifying periodic behavior until a nonperiodic, or chaotic, pattern is seen (Figure 3.1). To ensure that we are working with a chaotic time series, it is extremely important that we omit these oscillating or transient portions of the solutions.

Conventional measures successfully quantify the characteristics, properties, and structure of chaotic data, whether model-generated or observed, only when such transients are removed from the data. Figure 3.2 shows the correlation dimension ν calculation for the 0.01 percent initial condition using the *first* 10,000 points of the

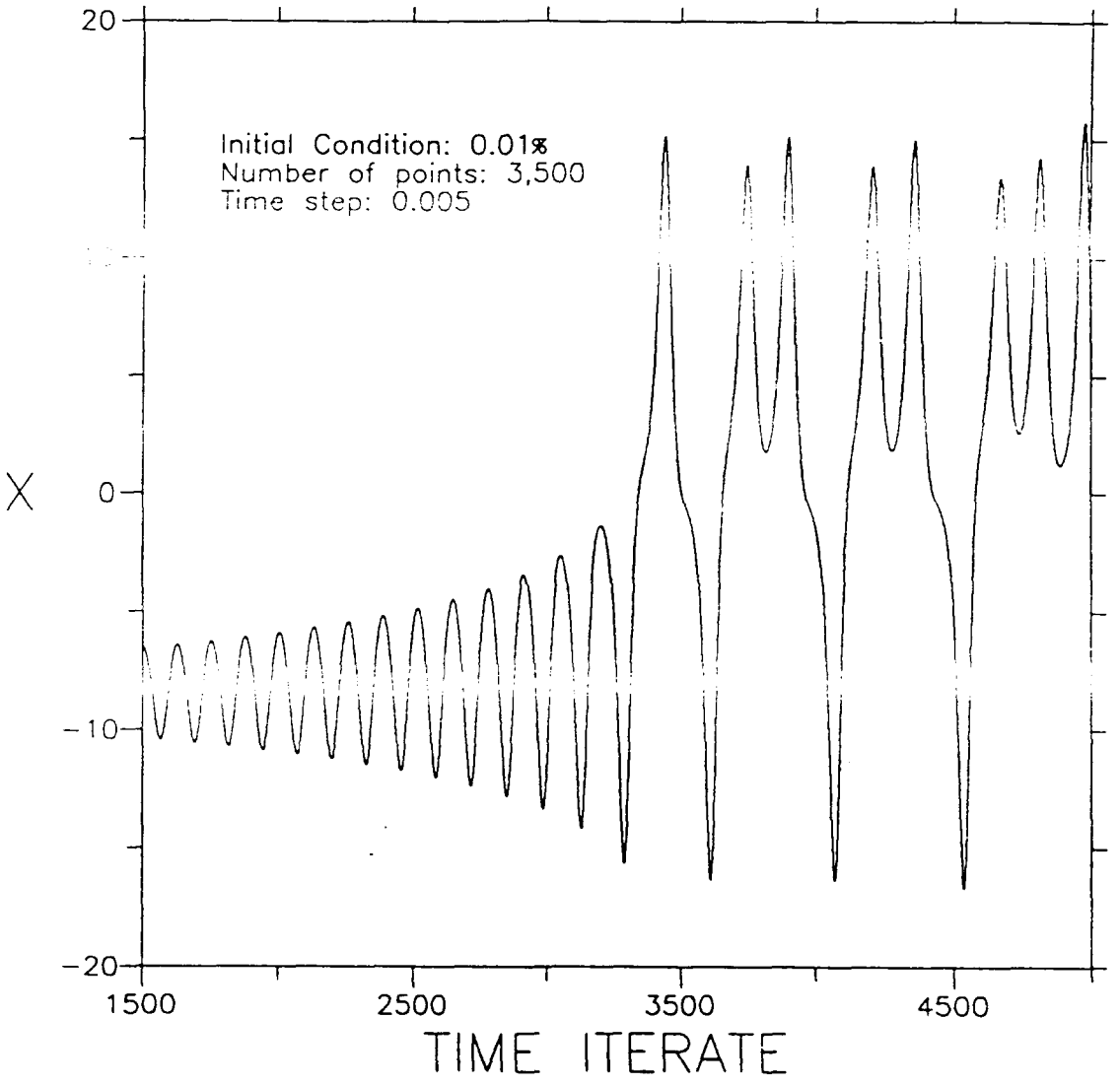


Figure 3.1: The X time series for the Lorenz attractor. Note that the behavior changes from growing periodic to chaotic.

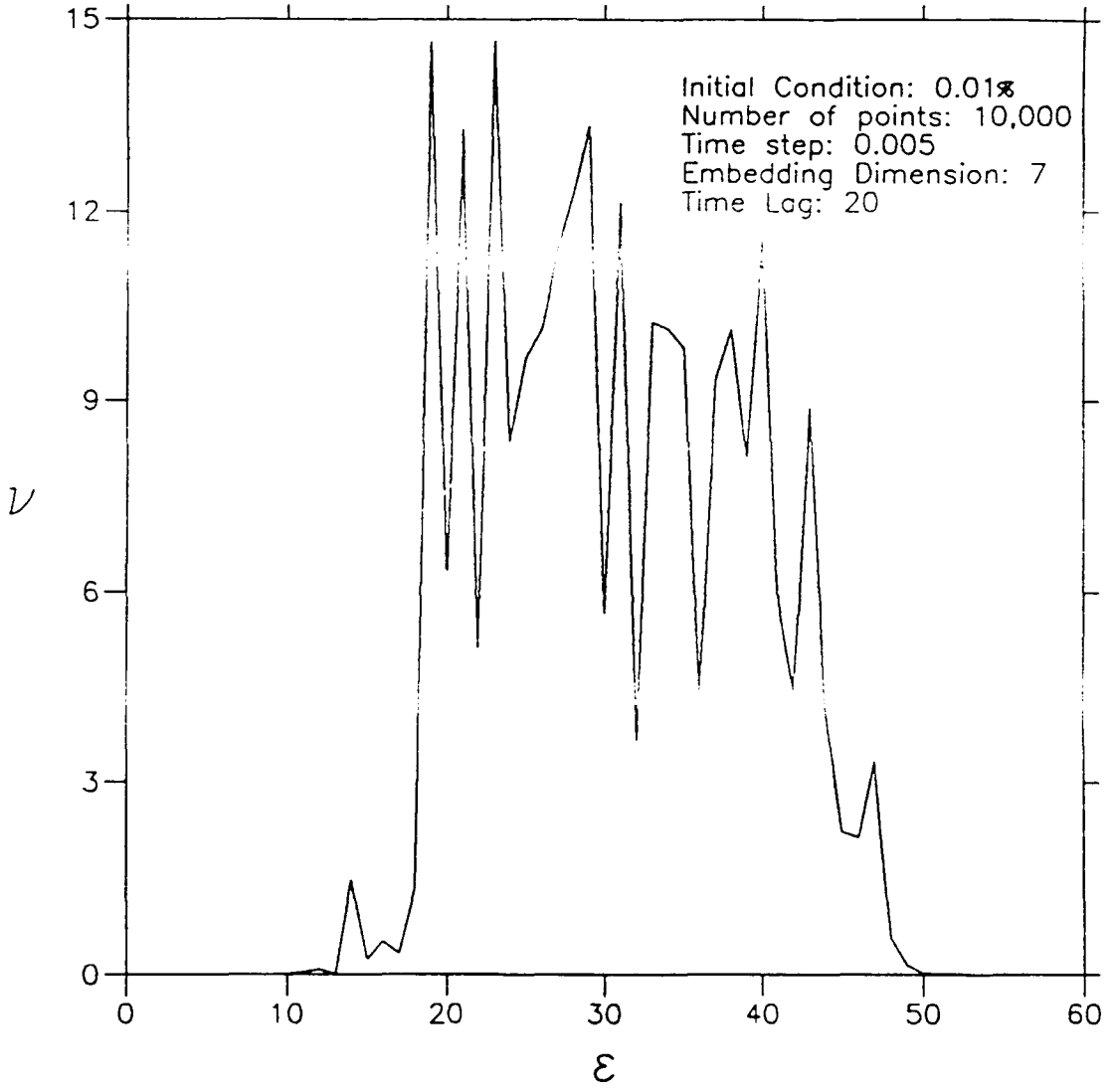


Figure 3.2: The correlation dimension value as a function of bin distance using the slope formula for the first 10,000 points of the 0.01% series, one in which transients are most likely present. Stable behavior at a value near 2.06 is not evident.

model-generated series, a series that is most likely dominated by transients. Comparing these results with those in Figure 2.15, a plot in which we have used the *last* 10,000 points of a 100,000-point series, we note that significant differences exist. Unlike Figure 2.15 that exhibits values of ν rather near its accepted value of 2.06, the data set that contains transients (Figure 3.2) does not produce values for ν anywhere near that value. For a series containing transients, the dimension value obtained from the data is obviously inaccurate and unrepresentative of the attractor.

To ensure an accurate representation of the chaotic structure, we must work strictly with the chaotic portion of the solution. These observations apply to the Histogram Measure as well. In the previous chapter, we described in detail the procedures that we used to develop our Histogram Measure, and we gained some insight into its structure and characteristics. In developing our histogram plots, however, we arbitrarily determined that the duration of the transient portion of the solution was less than the first 20,000 time steps, which we discarded from the data. However, we have not shown that this was a proper assumption to make. In this chapter, we show how to use the Histogram Measure to provide us with simple new ways for identifying transients in time series.

3.1. The Spike Signature Method

Instead of eliminating the first 20,000 time steps from the time series, we use the entire portion. In this example, we choose the same parameter values and conditions that were used to generate the smoothed 0.01 percent histogram plot in Figure 2.10. Figure 3.3 shows the histogram that results. In Figure 3.4, we compare the histograms in Figures 2.10 and 3.3, and we find a remarkable difference between them. In the new histogram, there is a tall, narrow spike that is confined to a small range of bin distances. This range covers the bin distances 74 to 78, corresponding to a phase space distance range of approximately 30 to 32. All values at the other bin distances match nearly identically.

The above results require careful analysis. The range of distance values within which the spike signature occurs is easy to understand. It is related entirely to the X, Y, and Z values that we used to define the initial condition. As described in Chapter 2, these values are combined to form a three-dimensional Euclidean distance from the origin, and they correspond to a Euclidean distance of approximately 31. The narrow range of distances producing the spike is a result of the trajectory slowly spiralling as it

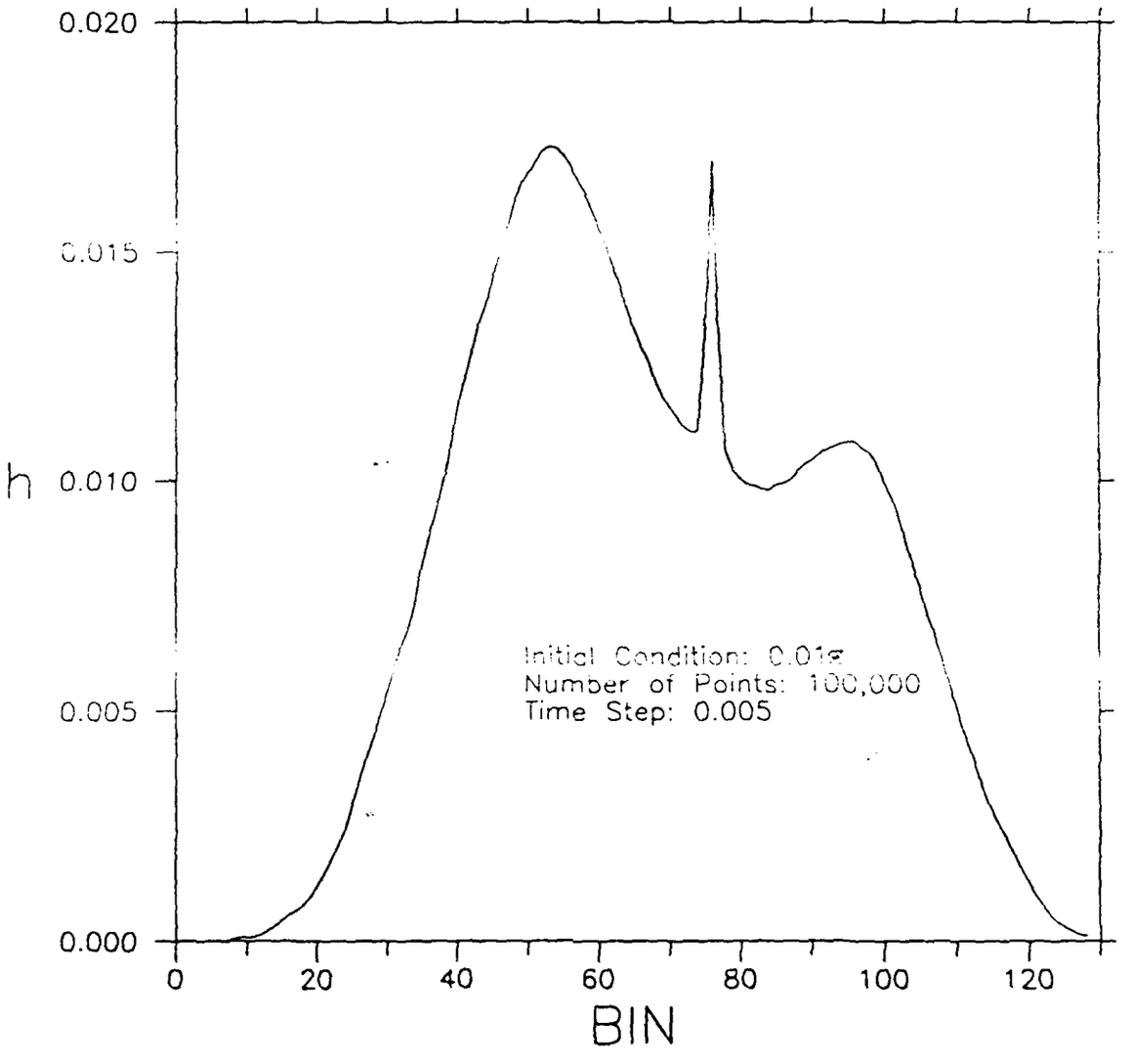


Figure 3.3: The 0.01% initial condition histogram using all 100,000 point of the series. The transient portions of the solution, which are included when using all of the points, are responsible for the large spike seen in bins 75–80.

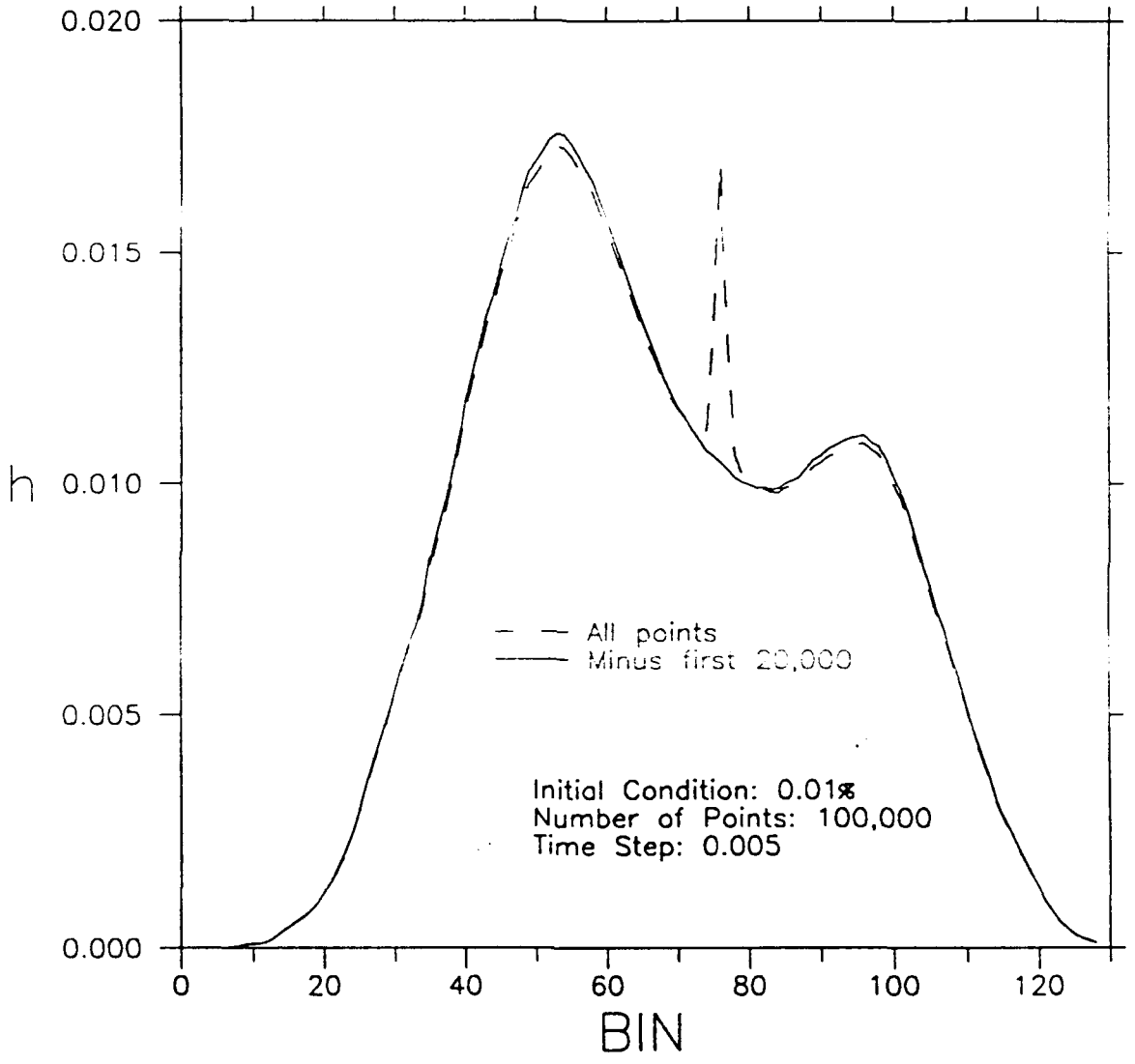


Figure 3.4: Comparison of the 0.01% initial condition histograms between that produced by using all the points of a 100,000-point series and that produced by eliminating the first 20,000 points of the same series. All of their bin values match nearly identically, with the exception of the large spike located in bins 75-80.

evolves inside one of the elliptically-shaped lobes outward from the unstable convective solution and toward the attractor. Not so clear is the reason why the spike is positioned near the primary minimum in the histogram. The existence of this minimum may be linked to the fact that the nontransient portion of the trajectory does not occur at those distances as it winds around the positive and negative attractor lobes. Of course, a portion of the trajectory will exist at these distances, but the motion along the trajectory is probably somewhat rapid, yielding a smaller contribution to the histogram than do other portions of the trajectory along which the motion is slower.

With the discovery that transients produce a spike signature in the histogram, we may use this property to identify the duration of the transient in any model-generated time series. To obtain this duration, we remove intervals of points from the beginning of the data. After we have removed all points contributing to the transient, we should detect the disappearance of the spike. Figure 3.5 shows the result of removing increasing numbers of points in 3,000-point blocks from the beginning of the series for the 0.01 percent case. Since the spike is gone once 12,000 points have been removed, we may be initially misled into believing that the transient is completely gone. However, we must be careful here. Upon closer inspection, we notice that small differences still remain when comparing this histogram with the one produced after

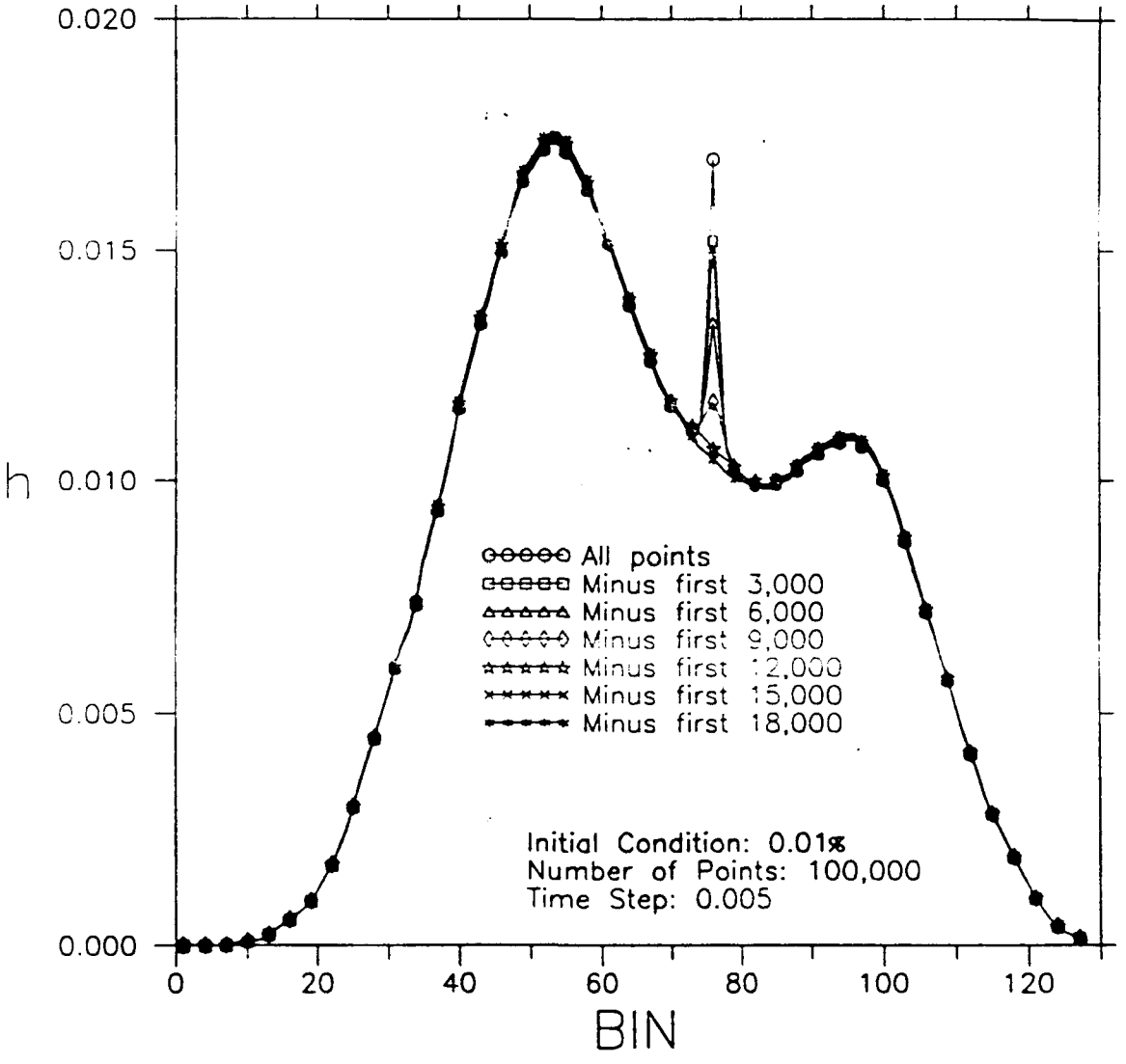


Figure 3.5: The 0.01% initial condition histograms superimposed for increasing blocks of data removed from the initial portion of the series. The histogram structures do not converge until approximately 15,000 points have been removed from the data; this number represents the initial transient portion of the series.

3,000 more points have been removed. Only when such consecutive histograms show very few structural differences between them can we be confident that our series contains only the chaotic behavior. Seeing no differences in histogram structure between the 15,000-point and 18,000-point cases, we are fully confident that the transient is gone once we have eliminated the first 15,000 points from the data set. We use the same procedures for each of the two remaining initial conditions, the only difference being in the interval removed; for the 0.10 percent case, we examined 2,000-point intervals and for the 1.0 percent case, we tried 1,000-point intervals. Using the same method for detection, we find that the transient is gone once we have eliminated from the data set the first 10,000 and 5,000 points, respectively, in these two cases, as shown in Figures 3.6 and 3.7.

The fact that the duration of the transient varies with initial condition seems straightforward. What is most interesting, however, is that for each 10-fold magnitude *decrease* in initial distance from one of the two nontrivial unstable convective solutions, there is a 5,000-point *increase* in the length of the transient within a time series. This relation tells us that we can estimate the duration of the oscillating portion of the solution by simply calculating the initial distance from the attractor. However, for initial conditions outside of the two attractor lobes, this relation may not hold. Even so,

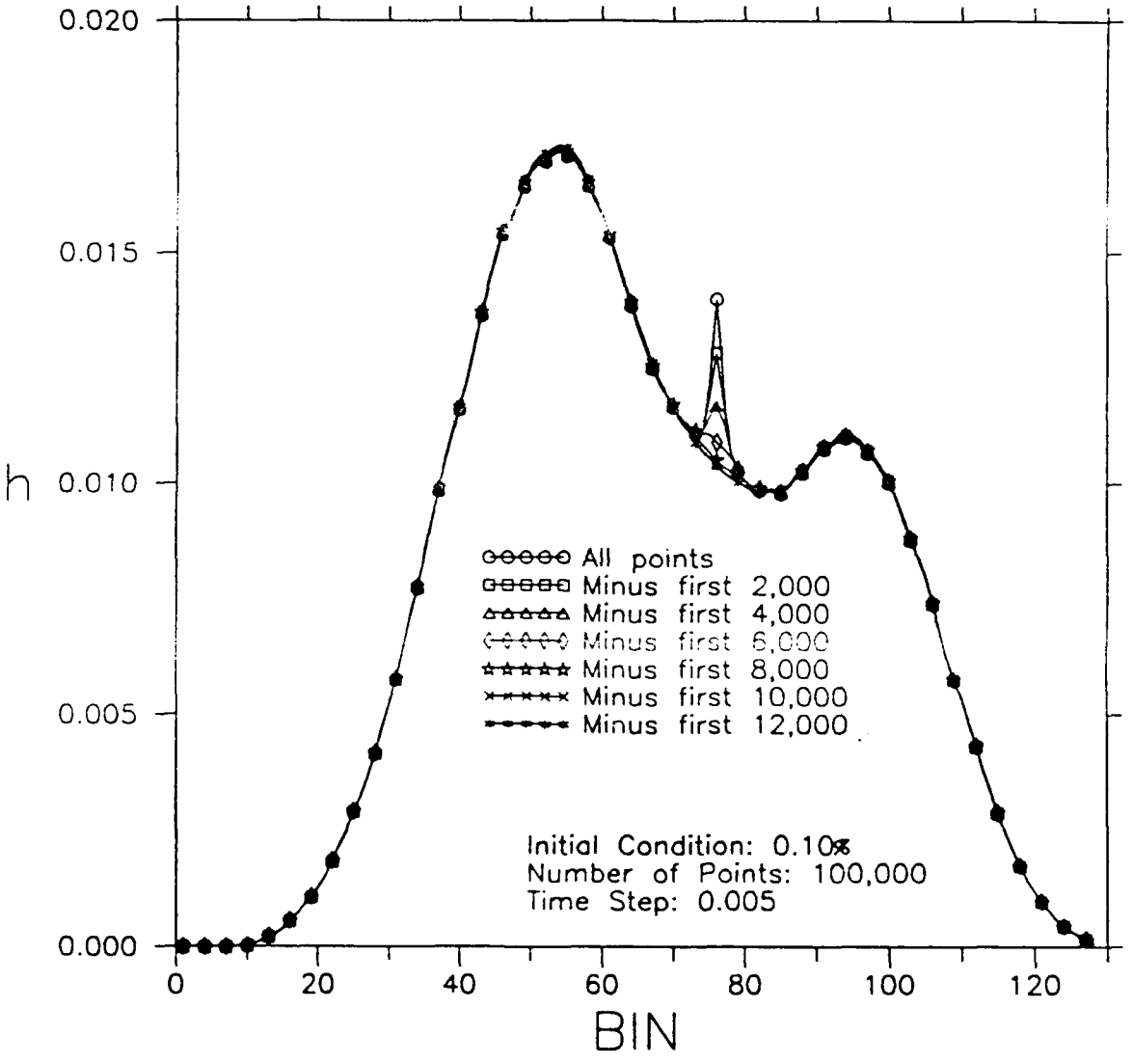


Figure 3.6: The 0.10% initial condition histograms superimposed for increasing blocks of data removed from the initial portion of the series. For this case, the histogram structures do not converge until approximately 10,000 points have been removed from the data; this number represents the initial transient portion for this series.

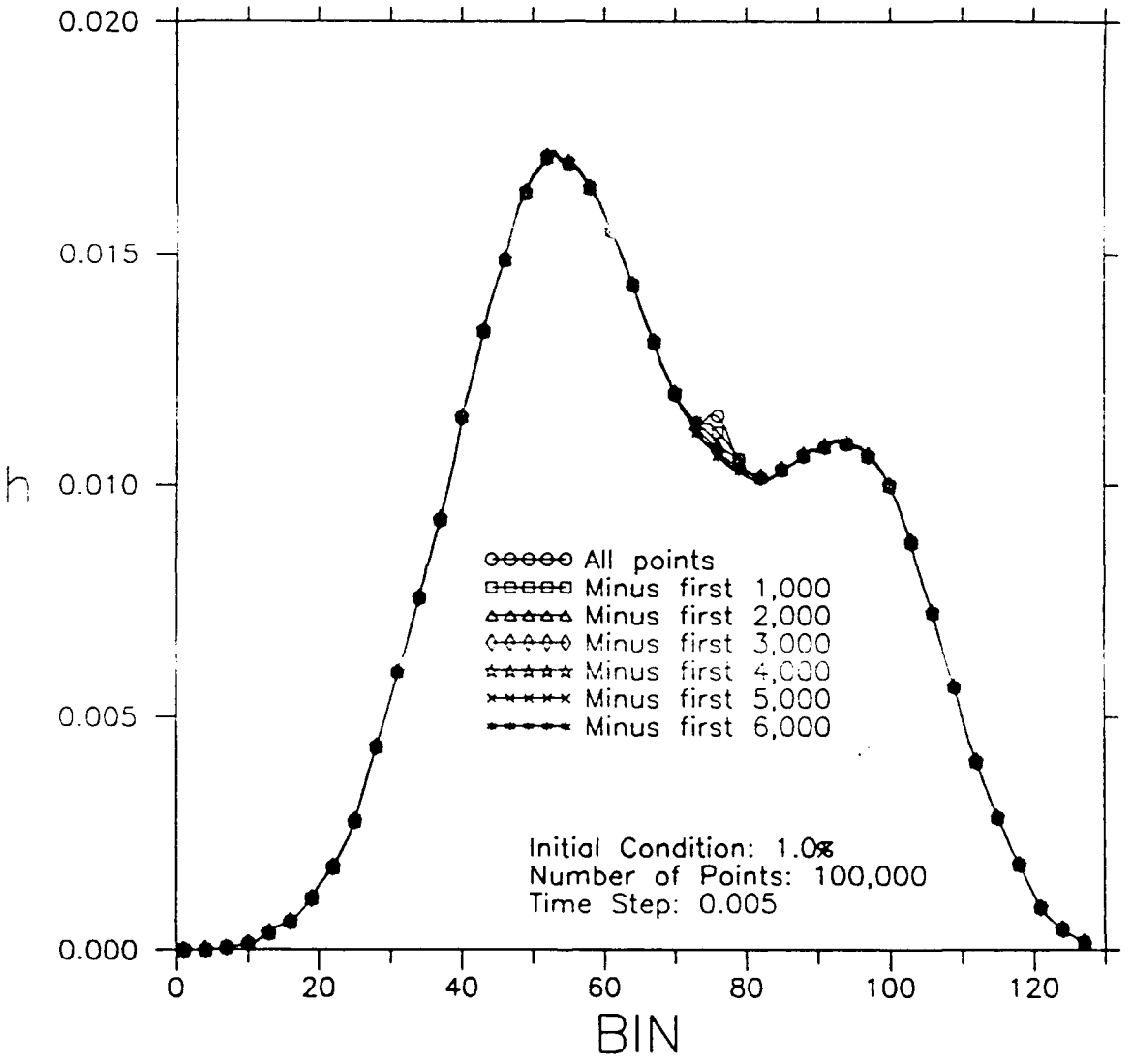


Figure 3.7: The 1.0% initial condition histograms superimposed for increasing blocks of data removed from the initial portion of the series. In this case, the histogram structures do not converge until approximately 5,000 points have been removed from the data; this number represents the initial transient portion for this series.

the transient identification method will still work. Only in this case, the spike is less pronounced because outside of the attractor lobe, the trajectory is not limited to a small range of distances as it evolves toward the attractor; Figure 3.8 shows the results using the initial condition X=0, Y=1, Z=0. We may use the same test for transient duration knowing that the transient is gone once the histogram structures converge. Thus, no matter what initial condition we use, this method provides us with an efficient and inexpensive way to identify a transient-free time series.

3.2. The Offset Mean Absolute Difference Method

Although confident that the spike method can successfully identify transient solutions, we question its accuracy because it involves a rather subjective analysis. In this section, we present a more *quantitative* approach that identifies transient portions of a time series and determines their duration. The results that we obtain with this method are strikingly consistent with those obtained using the spike method. We describe this alternative method here.

Rather than comparing the differences between two histograms by merely

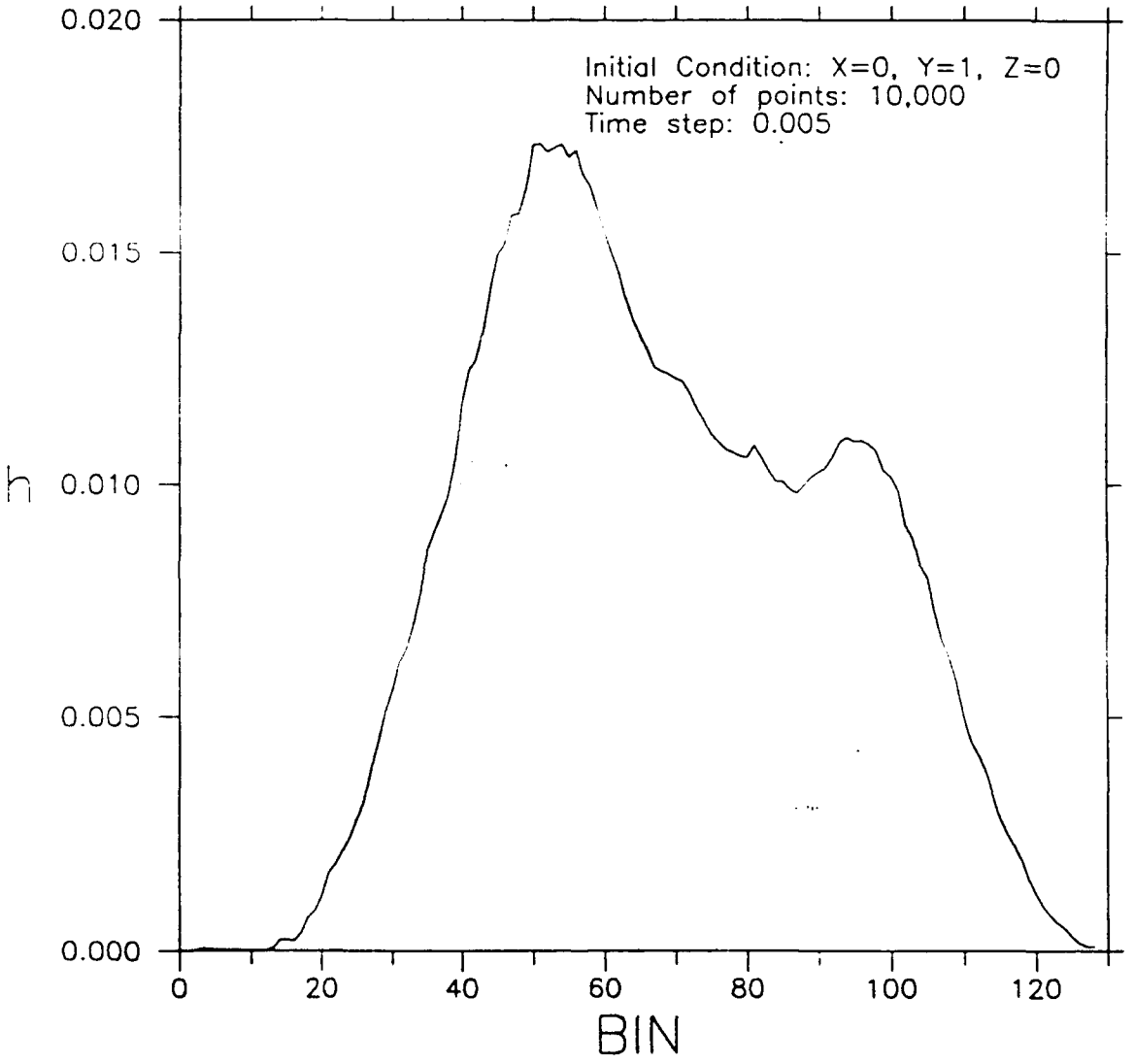


Figure 3.8: The histogram produced using the first 10,000 points for an initial condition outside of the elliptically-shaped lobes. The large spike signature is not evident here because the trajectory is not bound to a small range of distances as it evolves toward the attractor.

superimposing their structures, we instead calculate their mean absolute differences. For each bin j, we compute the differences between two histograms H_a and $H_{a'}$ of fixed length and offset by a predetermined interval, and then square the result to eliminate negative values. We then add these 128 bin differences, take the square root of the result, and finally divide by 128 to obtain a single characteristic offset mean absolute difference value $D_{aa'}$ given by

$$D_{aa'} = \frac{1}{128} \sum_{j=1}^{128} \left[\left(H_a(z_j) - H_{a'}(z_j) \right)^2 \right]^{1/2}. \qquad (3.1)$$

Figure 3.9 shows the results of this calculation for the 0.01 percent case and a sample length of 100,000 points. The ordinate simply shows the $D_{aa'}$ values, while the abscissa identifies the range of points used to create the two histograms H_a and $H_{a'}$. For example at x=0, we compare the first 100,000-point histogram (time steps 0-100,000) with the one produced using a series offset by 5,000 points (time steps 5,000-105,000). At x=5, we compare the 5,000-105,000 time step histogram with the 10,000-110,000 time step one and so on. Our choice of 100,000-point blocks is arbitrary. From the figure, we see a large decrease in difference values as the interval moves away from the initial part of the series. We argue that these large initial differences are a direct result

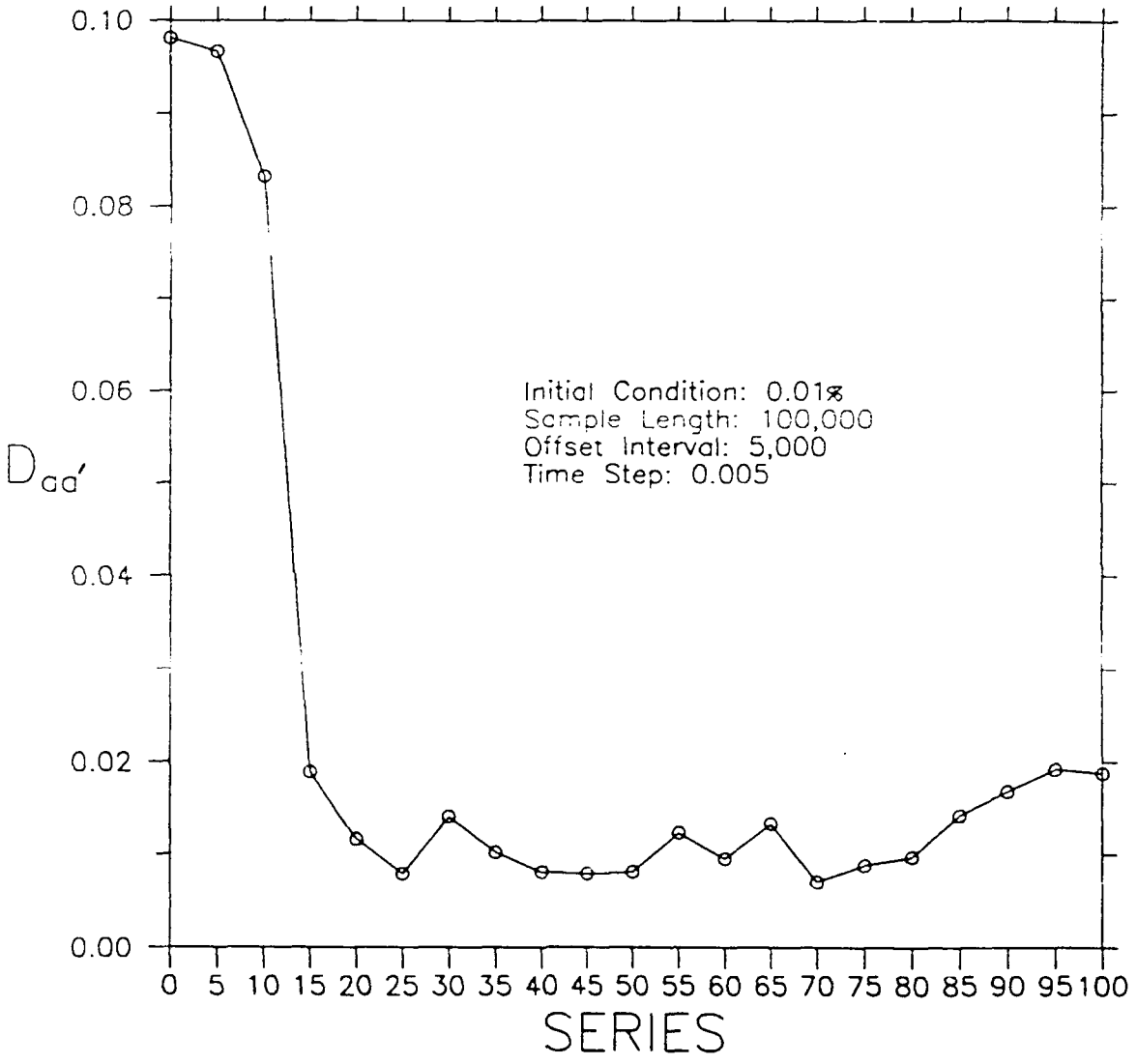


Figure 3.9: The offset mean absolute difference value $D_{aa'}$ as a function of the offset histogram series (in thousands) for the 0.01% initial condition. Note that at 15, the curve has first reached a tolerance value $D_{aa'}$ of approximately 0.02. This suggests that to avoid the transient portions of the solution in this case, we must remove the first 15,000 points of the series. Thus, values of $D_{aa'}$ above 0.02 are related to the transient, while those below are functions of characteristic intrinsic variabilities.

of the transient; these differences continue to decrease as we remove an increasing number of initial points from the data.

What is at first puzzling, however, is that we seem to obtain a significantly different duration for the transient portion using this method than we obtained using the spike method. For this particular initial condition, we discovered from the spike technique that the transient is gone once 15,000 points have been removed. Figure 3.9 shows a continued sharp decrease in mean-square differences up to this interval; however, this decrease continues through the first 25,000 points to a relatively stable difference *floor* value ($D_{aa'} \sim 0.01$), largely contradicting the transient duration results from the other method. Upon closer inspection, however, we observe that once approximately 15,000 points have been removed, the difference value $D_{aa'} = 0.0189$ (Table 3.1) has *first* reached an error threshold or *tolerance*. We note that the $D_{aa'}$ maximum threshold value is 0.0192 at x=95 in Table 3.1. Within this tolerance level ($D_{aa'} < 0.02$), we argue that the small differences between the histograms are most likely functions of their intrinsic structural variabilities, although their *average* difference values appear closer to 0.01. For values above this tolerance level ($D_{aa'} \geq 0.02$), we conclude that the histogram differences we observe are most likely caused by the transient. Based on this argument, then, the duration of the transient portion of this

Table 3.1: The offset mean absolute difference values $D_{aa'}$ for three sample lengths within the 0.01% initial condition as a function of the offset histogram intervals (H_a and $H_{a'}$). The asterisks note the maximum threshold values for each sample length; values in bold type flag the point interval at which the threshold values are first reached from the initial portion of the series. This point interval (15,000) represents the number of points that must be omitted from the initial part of the series in order to avoid the unwanted transient solutions.

Offset Histogram Intervals		Sample Length		
		100,000	200,000	300,000
H_a	$H_{a'}$		$D_{aa'}$	
0	5,000	0.0981	0.0498	0.0331
5,000	10,000	0.0967	0.0492	0.0327
10,000	15,000	0.0832	0.0418	0.0275
15,000	**20,000**	**0.0189**	**0.0132**	**0.0062**
20,000	25,000	0.0116	0.0091	0.0058
25,000	30,000	0.0079	0.0081	0.0066
30,000	35,000	0.0141	0.0064	0.0047
35,000	40,000	0.0102	0.0076	0.0048
40,000	45,000	0.0081	0.0068	0.0049
45,000	50,000	0.0079	0.0038	0.0063
50,000	55,000	0.0081	0.0130	0.0045
55,000	60,000	0.0123	0.0121	0.0027
60,000	65,000	0.0095	0.0078	0.0046
65,000	70,000	0.0133	0.0105	0.0035
70,000	75,000	0.0070	0.0056	0.0065
75,000	80,000	0.0088	0.0135*	0.0067*
80,000	85,000	0.0097	0.0089	0.0062
85,000	90,000	0.0142	0.0092	0.0055
90,000	95,000	0.0168	0.0111	0.0060
95,000	100,000	0.0192*	0.0068	0.0041
100,000	105,000	0.0188	0.0090	0.0048

series is again approximately 15,000 points, in virtual agreement with the results that we obtained using the spike method.

Varying the size of the sample length provides us with even more convincing evidence. Figure 3.10 shows a superposition of the plots for three sample lengths: 100,000, 200,000, and 300,000 points; their $D_{aa'}$ values are given in Table 3.1. All three show strikingly similar behavior, although they differ in the relative values of the mean absolute difference $D_{aa'}$ and in their associated tolerance levels. This *decrease* in difference value with an *increase* in the number of points used is an interesting finding, but not extremely meaningful, as we should expect this behavior when we add more points to the normalized histogram. What is most intriguing is that, despite the size of the sample length, the tolerance level is reached once approximately 15,000 points have been removed from the data.

We next determine whether this method gives consistent results for the time series produced using other initial conditions. Figures 3.11 and 3.12 show the superposition of the results obtained using three different sampling intervals for both the 0.10 and 1.0 percent initial condition cases; their $D_{aa'}$ values are given in Tables 3.2 and 3.3. Using the same argument as above, we observe that their associated tolerance levels are reached by approximately 10,000 points in the 0.10 percent case and by

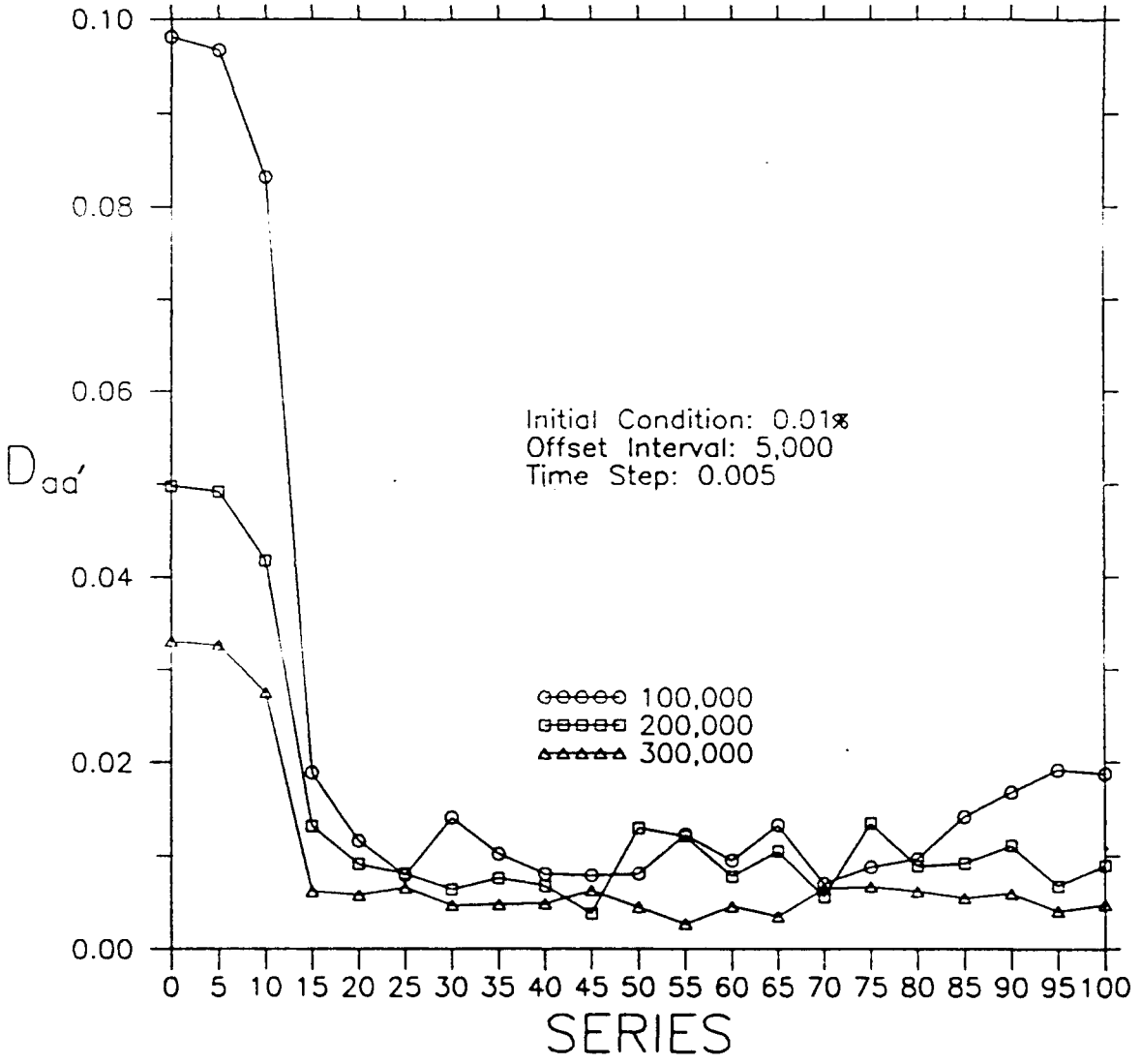


Figure 3.10: The offset mean absolute difference value $D_{aa'}$ as a function of the offset histogram series (in thousands) for the 0.01% initial condition and three separate sample lengths. Despite the dependence of $D_{aa'}$ values upon the sample length, the interval at which the separate tolerance values is first reached remains the same in each case; again this 15,000 points represents the number that must be removed from the initial part of the series in order to avoid the unwanted transient portions of the solution.

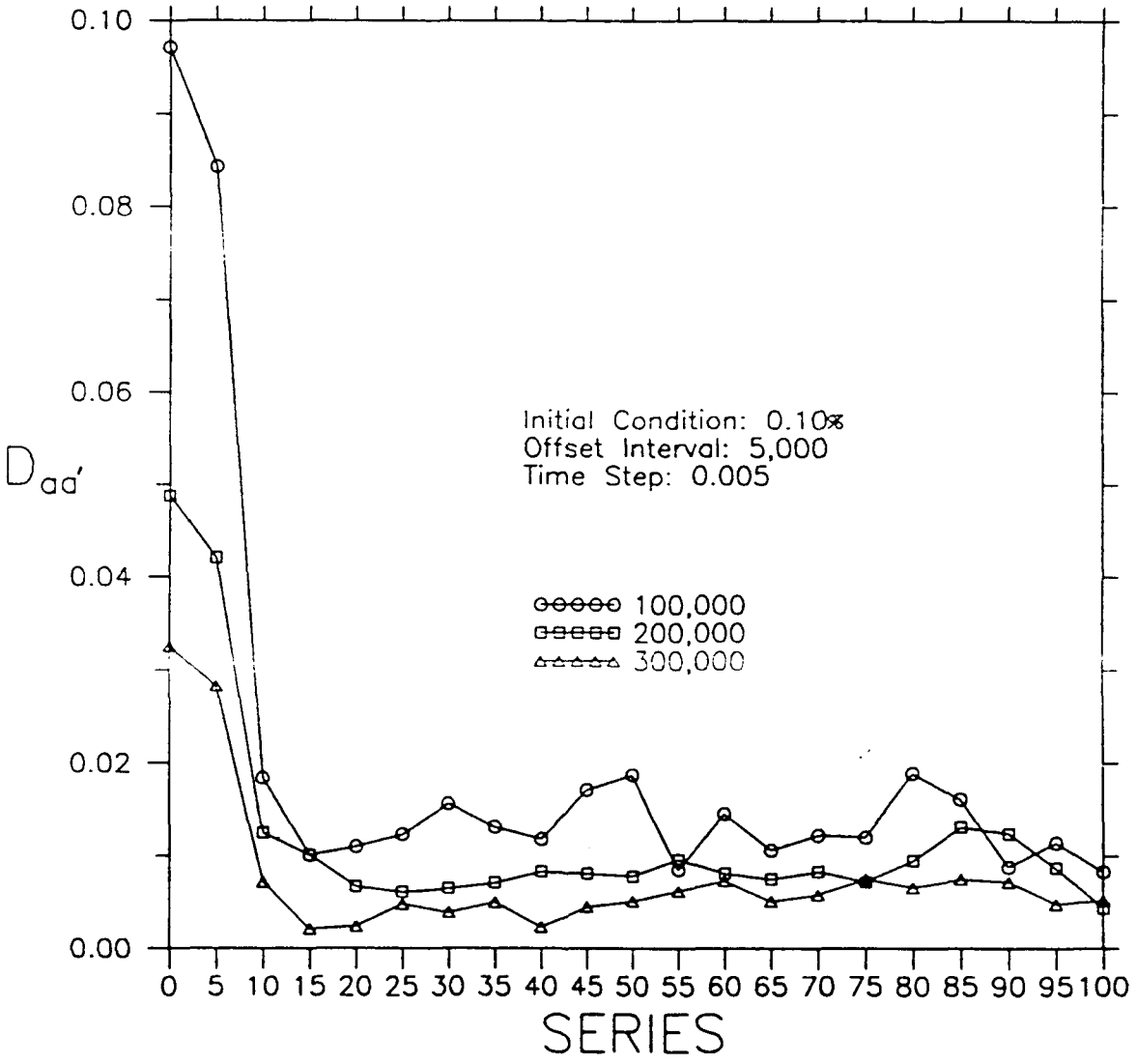


Figure 3.11: The offset mean absolute difference value $D_{aa'}$ as a function of the offset histogram series (in thousands) for the 0.10% initial condition and three separate sample lengths. Consistent with that of the 0.01% case, the interval at which the tolerance is first reached is the same for all three sample lengths. Only now the point interval, which represents the transient portions of the solution, is approximately of length 10,000.

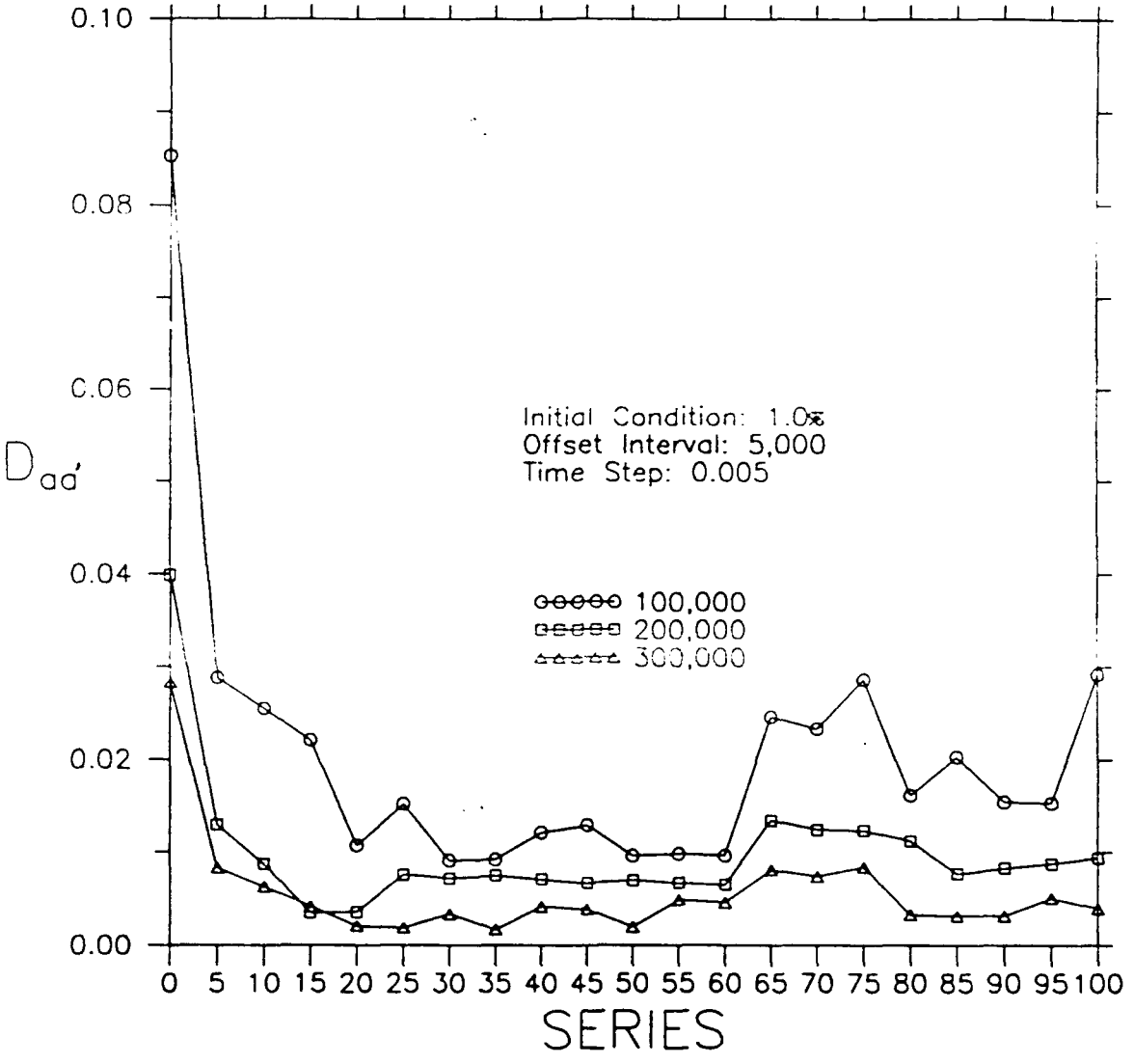


Figure 3.12: The offset mean absolute difference value $D_{aa'}$ as a function of the offset histogram series (in thousands) for the 1.0% initial condition and three separate sample lengths. Consistent with the other two initial condition cases, the point interval at which the tolerance is first reached is the same for all three sample lengths; in this case, the interval that represents the transient portions of the solution is approximately 5,000 points in length.

Table 3.2: The offset mean absolute difference values $D_{aa'}$ for three sample lengths within the 0.10% initial condition as a function of the offset histogram intervals (H_a and $H_{a'}$). The asterisks note the maximum threshold values for each sample length; values in bold type flag the point interval at which the threshold values are first reached from the initial portion of the series. This point interval (10,000) represents the number of points that must be omitted from the initial part of the series in order to avoid the unwanted transient solutions.

Offset Histogram Intervals		Sample Length		
		100,000	200,000	300,000
H_a	$H_{a'}$		$D_{aa'}$	
0	5,000	0.0971	0.0488	0.0325
5,000	10,000	0.0843	0.0421	0.0283
10,000	**15,000**	**0.0184**	**0.0125**	**0.0072**
15,000	20,000	0.0101	0.0101	0.0021
20,000	25,000	0.0110	0.0067	0.0024
25,000	30,000	0.0123	0.0061	0.0048
30,000	35,000	0.0156	0.0065	0.0039
35,000	40,000	0.0131	0.0071	0.0050
40,000	45,000	0.0118	0.0083	0.0023
45,000	50,000	0.0171	0.0081	0.0045
50,000	55,000	0.0187	0.0078	0.0051
55,000	60,000	0.0085	0.0096	0.0062
60,000	65,000	0.0145	0.0081	0.0073
65,000	70,000	0.0106	0.0075	0.0051
70,000	75,000	0.0122	0.0083	0.0058
75,000	80,000	0.0120	0.0072	0.0075*
80,000	85,000	0.0189*	0.0095	0.0066
85,000	90,000	0.0161	0.0131*	0.0075*
90,000	95,000	0.0088	0.0124	0.0072
95,000	100,000	0.0114	0.0087	0.0048
100,000	105,000	0.0083	0.0044	0.0052

Table 3.3: The offset mean absolute difference values $D_{aa'}$ for three sample lengths within the 1.0% initial condition as a function of the offset histogram intervals (H_a and $H_{a'}$). The asterisks note the maximum threshold values for each sample length; values in bold type flag the point interval at which the threshold values are first reached from the initial portion of the series. This point interval (5,000) represents the number of points that must be omitted from the initial part of the series in order to avoid the unwanted transient solutions.

Offset Histogram Intervals		Sample Length		
		100,000	200,000	300,000
H_a	$H_{a'}$		$D_{aa'}$	
0	5,000	0.0853	0.0399	0.0283
5,000	**10,000**	**0.0288**	**0.0129**	**0.0083**
10,000	15,000	0.0255	0.0087	0.0063
15,000	20,000	0.0221	0.0035	0.0042
20,000	25,000	0.0107	0.0036	0.0021
25,000	30,000	0.0152	0.0076	0.0019
30,000	35,000	0.0091	0.0072	0.0034
35,000	40,000	0.0092	0.0075	0.0017
40,000	45,000	0.0121	0.0071	0.0042
45,000	50,000	0.0129	0.0067	0.0039
50,000	55,000	0.0096	0.0070	0.0020
55,000	60,000	0.0098	0.0067	0.0049
60,000	65,000	0.0096	0.0065	0.0046
65,000	70,000	0.0246	0.0134*	0.0081
70,000	75,000	0.0233	0.0124	0.0074
75,000	80,000	0.0286	0.0123	0.0084*
80,000	85,000	0.0161	0.0112	0.0033
85,000	90,000	0.0203	0.0077	0.0032
90,000	95,000	0.0154	0.0083	0.0032
95,000	100,000	0.0153	0.0088	0.0051
100,000	105,000	0.0292*	0.0094	0.0040

approximately 5,000 points in the 1.0 percent case. These results also agree with those obtained from the spike method, and so provide us with a more robust and reliable way to detect the presence of unwanted, nonchaotic portions of a time series. Most important from these results is that both methods presented here demonstrate that, in order to avoid unwanted transient contamination, eliminating the first 20,000 points from the data series as we did in Chapter 2 is in fact adequate for all three sets of initial conditions.

The two methods described in this chapter for detecting and identifying the duration of nonchaotic portions of a solution work well for this model because these transient portions exist totally in the initial part of the time series. We note that for other series, which may be generated from another model or obtained from observations, the transient behavior may not be as well-behaved. We therefore caution that, although these methods show merit in the Lorenz model, limitations may become apparent when we apply these methods to a more complicated time series. We address this concern in some detail in the conclusions. For now, we are quite confident in the ability to identify transient-free, model-generated data.

Our next task is to identify the minimum number of points that are required for us to have an adequate sample of the chaotic attractor. To do this, we use the simple

Histogram Measure to look for convergence of the histogram shape as points are added to the series. If convergence does occur, then the resulting control histograms become key elements in the development of sampling strategies. In pursuit of a control histogram, we also use information that we obtain from the more conventional Correlation Dimension Measure, and we address in detail the numerous sampling issues involved in creating an adequate data base.

CHAPTER 4

SAMPLING ISSUES RELATED TO FINDING CONVERGENT MEASURES

Now that we have ensured that we have data representing a chaotic attractor, we seek the minimum number of data points needed to quantify essential information about this attractor. To do this, we use the Histogram Measure to define a convergent or control histogram. However, finding this histogram requires a more extensive and time-consuming effort than might be expected, because obtaining convergence depends greatly upon the way in which we sample the attractor. Therefore, in finding criteria for convergence, we focus on a detailed discussion of the relevant sampling issues, relying heavily on results that we obtain from correlation dimension calculations. Furthermore, we comment throughout on the importance of quantifying these sampling issues and on their relations to the predictability characteristics of the Lorenz model.

4.1. The Histogram Measure

A common property of most fractal dimension measures that are used to quantify chaotic attractor structure is the independence of the dimension value as the initial conditions are varied (i.e. *insensitivity to initial conditions*). In this section, we first determine whether the Histogram Measure displays this same independence. Once we accomplish this, we then seek convergence within each of the initial conditions.

To find convergence *among* initial conditions, we obtain histograms from time series of fixed length for each of the three initial conditions, and then we compare these histograms with each other. Throughout this chapter, the values of B, P, and R remain the same as those used in Chapter 2, and the time step is 0.005; we also use the 1-3-1 weighted filter to smooth the data. We recall that in Chapter 2 we employed this comparative procedure using the Histogram Measure for the last 80,000 points of a 100,000-point data set. The distinct nonconvergent behavior that we obtained led us to conclude that either we detected a significant initial condition dependence or we simply did not use enough points to sample the attractor. To determine which possibility is correct, we work here with vastly longer time series, superimposing the histograms

using the three initial conditions for series lengthening by blocks of 100,000 points. Based on our results in Chapter 3, we discard the first 20,000 points in order to avoid virtually all of the unwanted transient portions of the solutions. We do this for all of the results that we discuss in this chapter.

For each successive interval increase, the structural differences between the three histograms slowly decrease. However, as seen in Figure 4.1, we observe that even after using the last 980,000 points of a 1,000,000-point series, *detectable structural differences remain.* This is a fascinating, yet disturbing result that contradicts the notion of measure independence with initial condition. In addition, this result raises the issue of just what we mean by convergence. Some might argue that this degree of convergence is good enough for their purposes. However, seeking to *unambiguously* find convergence, we argue that we do not yet have convergent histograms. With a time step of 0.005, this 980,000 points corresponds to a dimensionless time t* of 4900. The dimensionless time t* is related to the real time t by

$$t = \frac{t^* z^2}{\pi^2 \kappa (1 + a^2)} . \qquad (4.1)$$

For atmospheric application, z is defined to be the height of the boundary layer (1

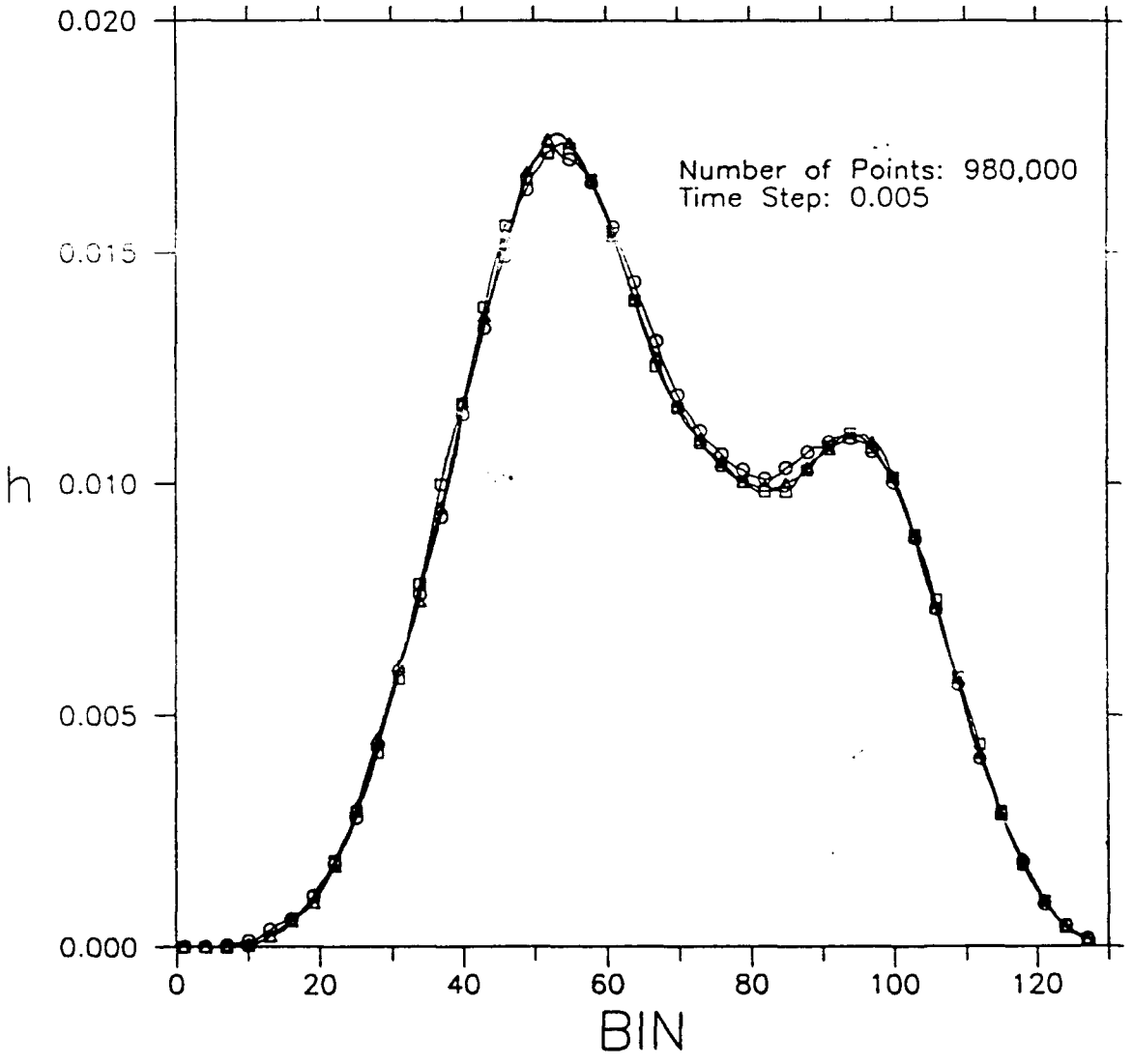


Figure 4.1: The three initial condition histograms superimposed on each other representing the last 980,000 points of a 1,000,000-point data set. Although the curves are relatively similar, they do not exhibit unambiguous convergence.

kilometer), κ is the thermal dissipation rate (30 $m^2 s^{-1}$), and a is the domain aspect ratio for which $a^2 = 0.5$ (Shirer 1987). This expression yields a value for t that is approximately 30 hours--much longer than the convective time scale of the atmosphere. This result is troublesome, because convergence requires much longer to achieve than the time (~ 4 hours) for which the governing Boussinesq system equations are themselves valid. Moreover, this result is inconsistent with that of the more standard fractal measures and also brings into question the validity of the use of Monte Carlo techniques to determine the predictability of an operational model. We discuss this technique in greater detail later in this chapter. For now, we raise a red flag. Assuming that the Histogram Measure is accurately quantifying the attractor, we consider two distinct possibilites: 1) The Histogram Measure is so sensitive that the differences it detects represent some large, intrinsic variability within the attractor that other measures can not detect or 2) We still have not sampled the attractor adequately to find the reasonable convergence that does actually exist.

4.2. The Correlation Dimension

To further analyze the above dilemma, we employ the correlation dimension ν. The advantage of working with this standard measure is that its behavior is well-known--the dimension value of the Lorenz attractor is approximately 2.06

(Grassberger and Procaccia 1983b). After finding the conditions necessary to produce convergence to this optimum value of ν, we apply that knowledge toward finding convergence in the Histogram Measure.

In Section 2.3, we described in detail the method used to calculate the correlation dimension. Here, we add a new feature by calculating ν using ten successively lengthening series for the attractor. In doing so, we can observe the convergent properties of this measure. Using data from the interval 990,001 to 1,000,000 time steps, we superimpose ν values for integer distances of radius ε for the ten interval lengths ranging from 1,000 to 10,000 points. Figures 4.2-4.4 show results using an embedding dimension d_E of seven for each of the three initial conditions. We recall that the time step t_S is 0.005 and that we use an optimal attractor reconstruction time lag Δt of 20. All three figures show similar behavior, with pronounced peaks and valleys. Figures 4.2 and 4.4, corresponding to the 1.0 and 0.01 percent perturbations, show the most distinct similarities; each has two distinct maxima and two minima, and they occur in the same range of ε values. However, determining an appropriate scaling region seems unclear from these figures. Both maxima have peak values corresponding to ν between 2.1 and 2.4, and the minima are much lower. These results appear suspect, not only because of the particular values that we obtain, but because using increasingly

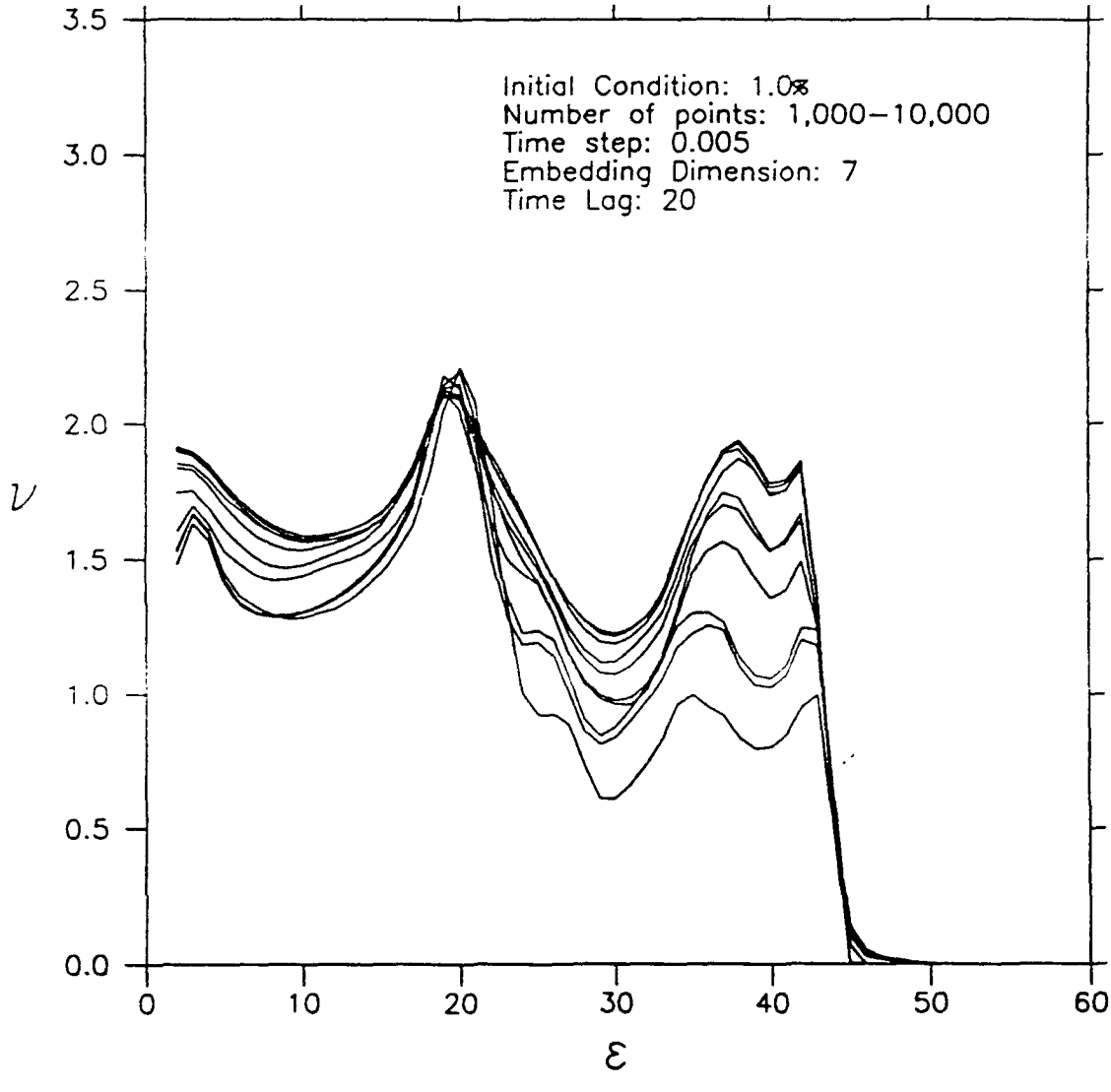


Figure 4.2: The correlation dimension value as a function of bin distance using the slope formula for a 1.0% initial condition and integer bin distances. Note the distinct dependence of the correlation dimension value on the interval length. Also note that the values are generally below two, inconsistent with the accepted value of 2.06.

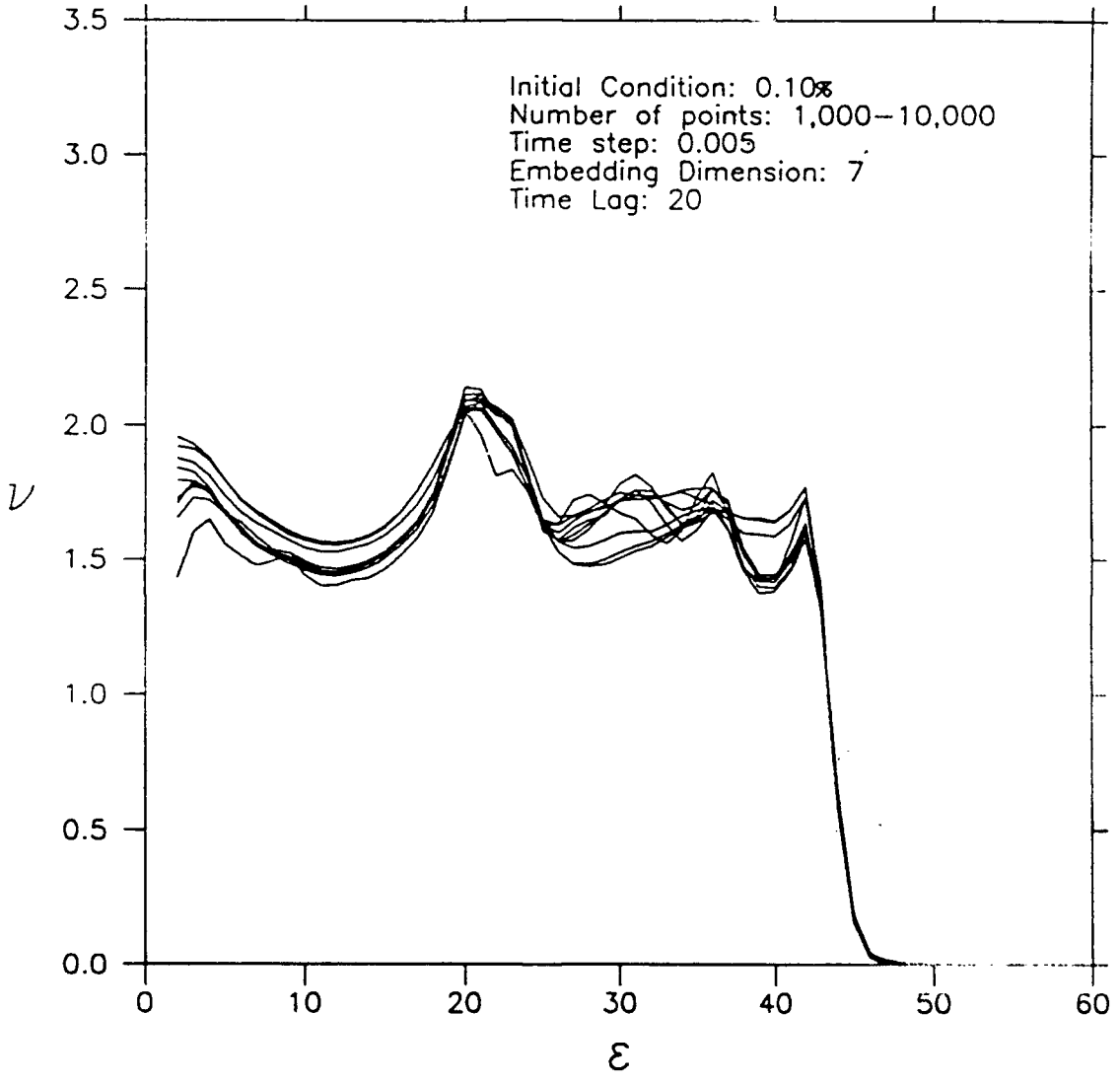


Figure 4.3: The correlation dimension value as a function of bin distance using the slope formula for a 0.10% initial condition and integer bin distances. Again, there is a dependence of the correlation dimension value on the interval length and its values are consistently below two.

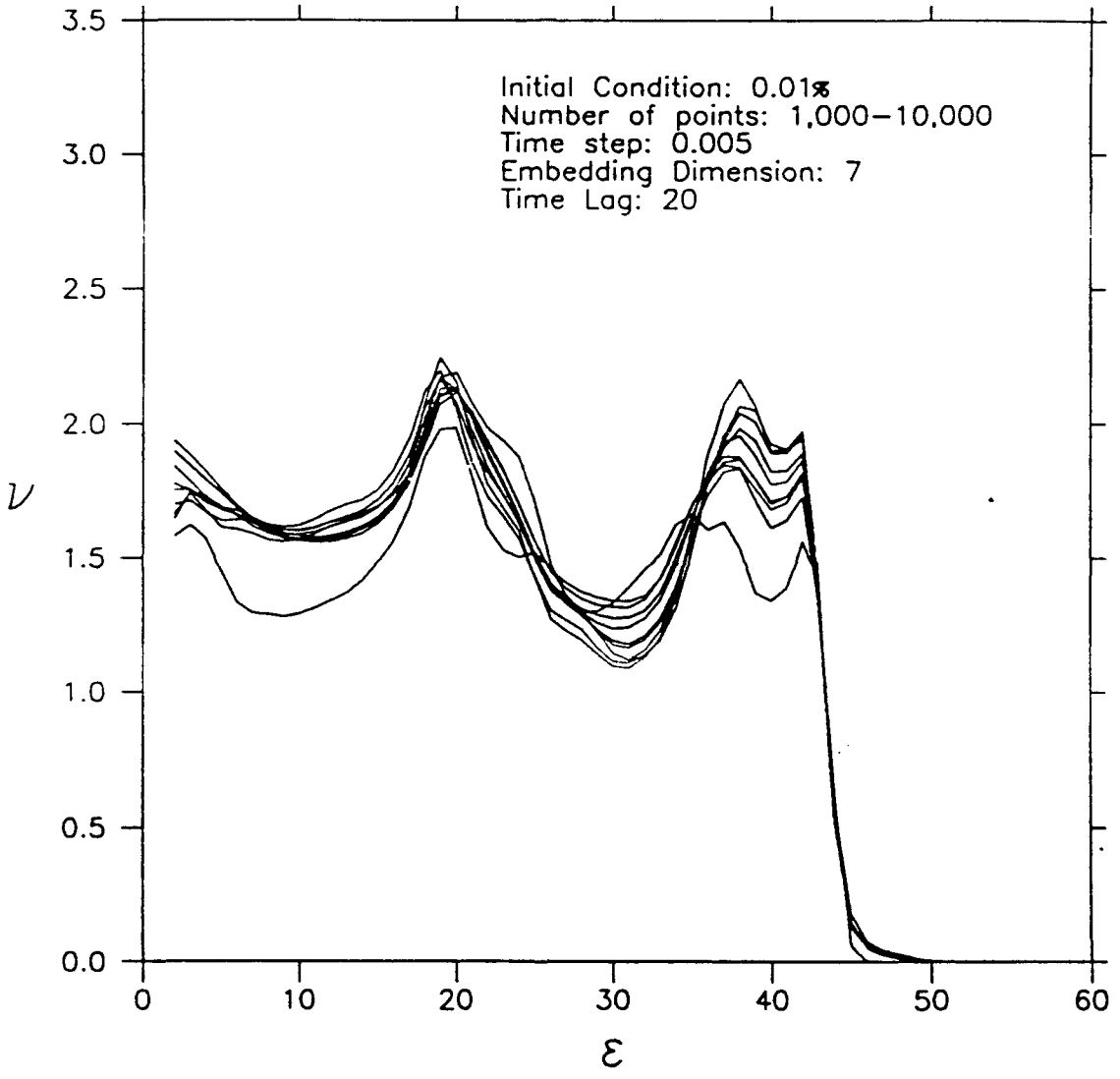


Figure 4.4: The correlation dimension value as a function of bin distance using the slope formula for a 0.01% initial condition and integer bin distances. As with the other two initial conditions, there is a distinct dependence of the correlation dimension value on the interval length; in addition, the values remain inconsistent with the accepted correlation dimension value.

long data intervals does not clearly lead to any convergence. This result is disturbing, as calculating ν for 10,000 points is expensive.

This lack of convergence may be the result of using too small an embedding dimension. A Whitney Embedding Theorem would imply that $d_E = 7$ should be sufficient to capture this attractor in artificial phase space and has been shown to have merit in studies using observed atmospheric data (e.g., Fraedrich 1986).

To test the possibility that the value of d_E should be larger, we perform the same calculations for an embedding dimension of nine; the results are shown for each initial condition in Figures 4.5-4.7. We observe that changing the embedding dimension value has a profound effect on the correlation dimension plots. The maximum that is located in the largest ε distances (40-45) has a much larger value than that seen in Figures 4.2-4.4. This value, approaching or even exceeding 3.0, obviously is unrepresentative of the actual dimension, which must be less than three. However, the other maximum now has a much broader peak spread over a greater range of ε distances; most importantly its maximum value has generally decreased to between 2.0 and 2.2, values that are more consistent with the accepted correlation dimension value. Perhaps this maximum in bins 20 to 26 does occur in a true scaling region and shows that the true value for ν is $2.06 \pm \sigma$, where σ is the variance.

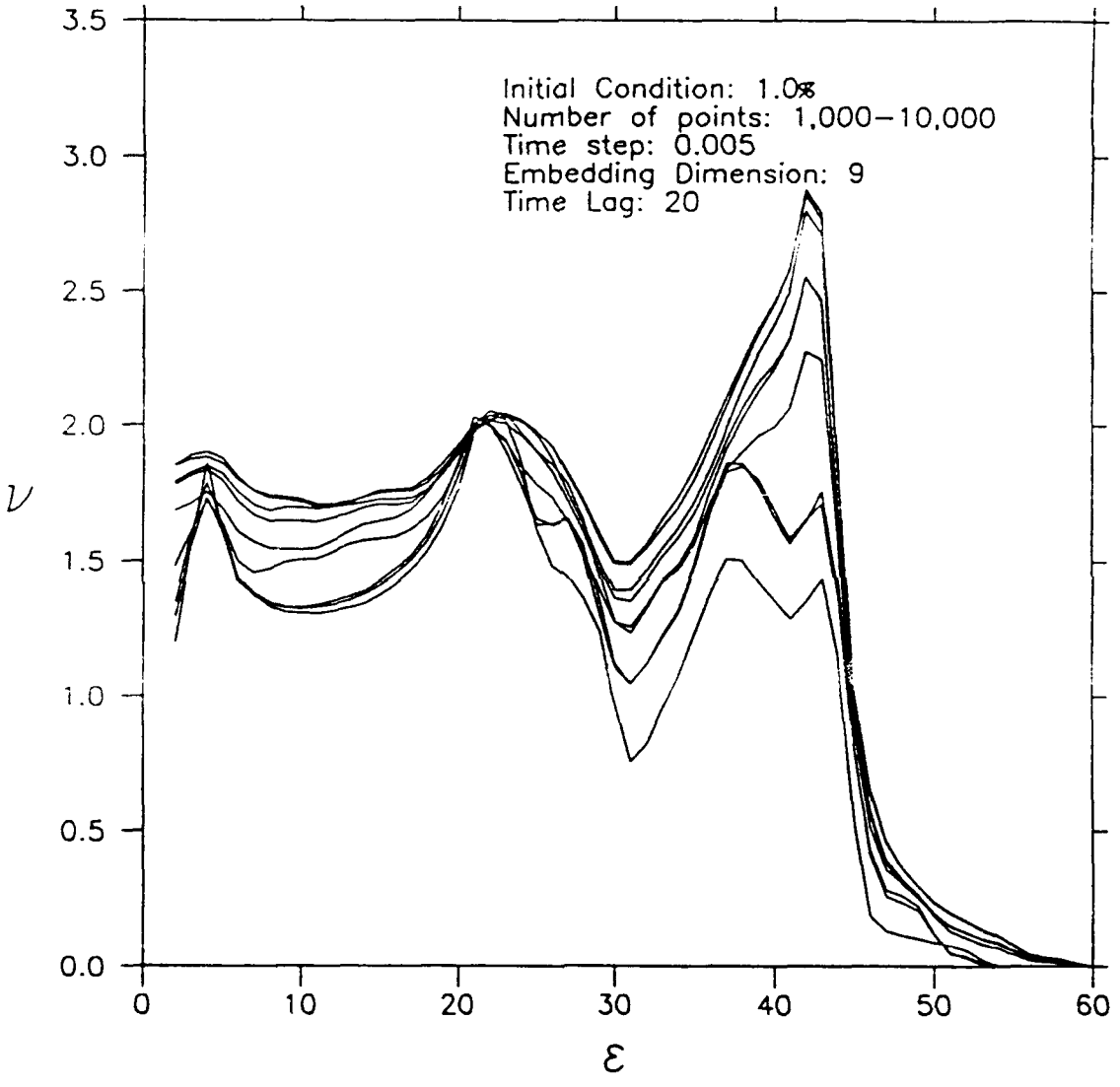


Figure 4.5: The correlation dimension value as a function of bin distance using the slope formula for a 1.0% initial condition and integer bin distances. Increasing the embedding dimension value from seven to nine changes the behavior markedly. It is heartening to note a reasonably good convergence to a dimension value near 2.06 in the bin distance range 20–26.

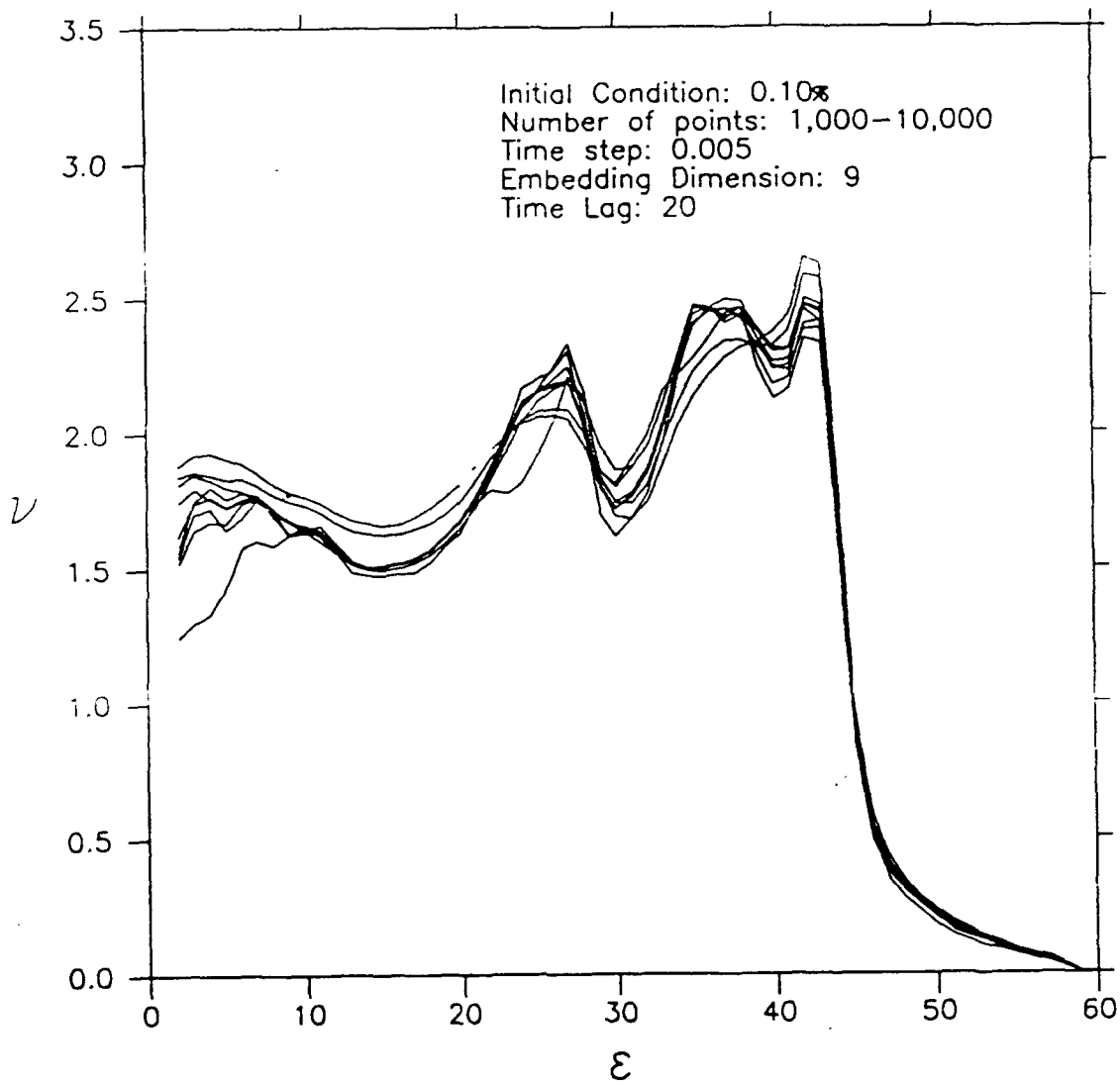


Figure 4.6: The correlation dimension value as a function of bin distance for a 0.10% initial condition and integer bin distances. For this case, it is harder to argue for a reasonable convergence to a correlation dimension value near 2.06.

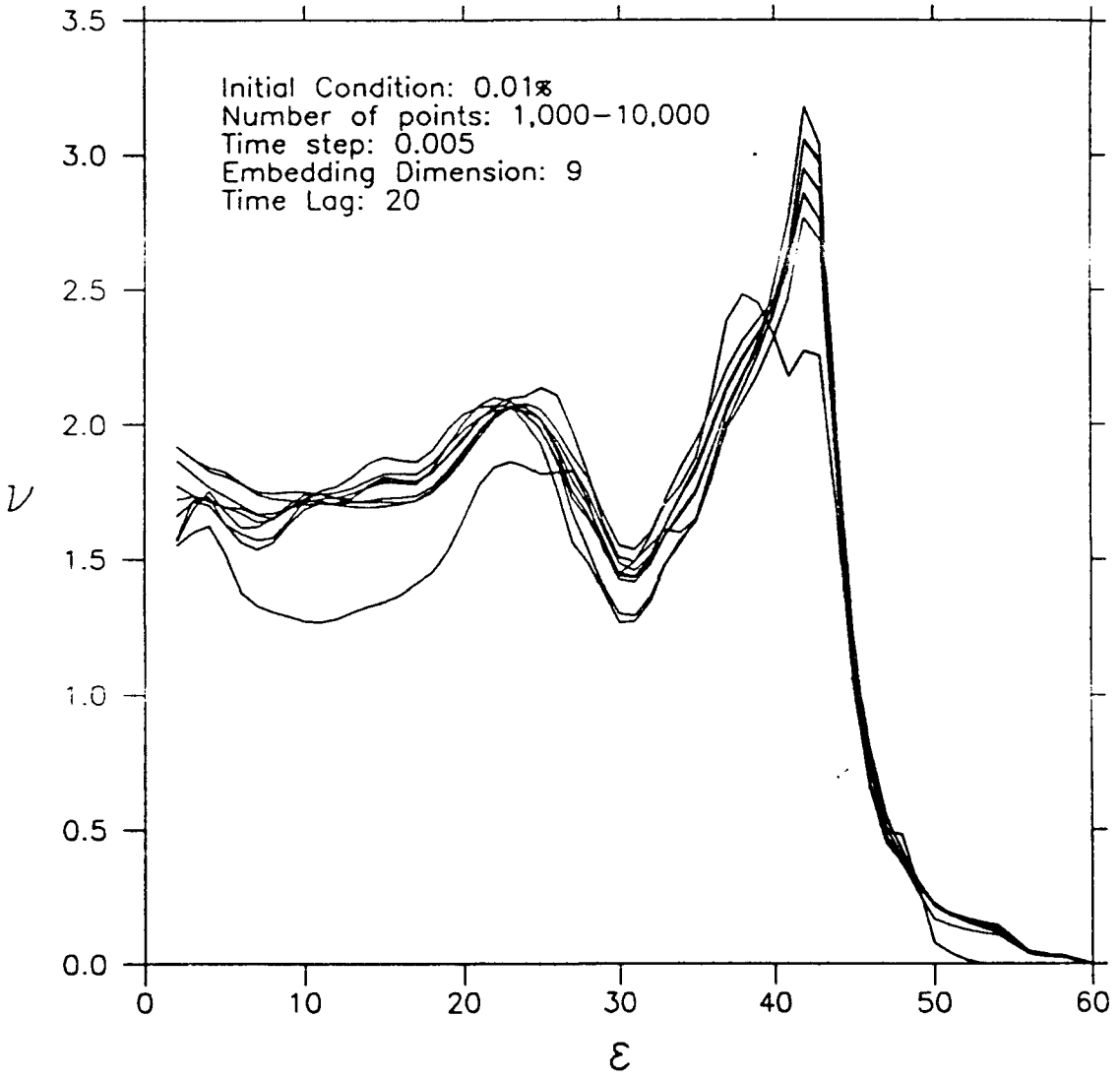


Figure 4.7: The correlation dimension value as a function of bin distance for a 0.01% initial condition and integer bin distances. Similar to the 1.0% case, increasing the embedding dimension produces a reasonably good convergence to a correlation dimension value near 2.06 in a similar bin distance range (22–28).

The above conclusion, however, is a troubling one for many reasons. First, the correlation dimension changes dramatically with changes in embedding dimension. If something like the Whitney Embedding Theorem is indeed valid, then both the values of d_E that we used should have captured similar ν behavior. Having to test values of d_E greater than nine to determine when the value of ν stabilizes is extremely time-consuming and highly impractical. Besides, for a fixed interval, increasing the embedding dimension decreases the sample size and thus most likely decreases the validity of the results. Second, there continues to be a lack of convergence among not only the ten separate data intervals, but also among the three sets of initial conditions shown in Figures 4.5-4.7. Finally, the apparent scaling region that we have found is at a much larger, macroscale distance ε than that reported by others (Nese 1985) and is also inconsistent with the definition of ν, which is formally valid only in the limit as ε approaches zero. In our plots, the values of ν are well below its accepted value of 2.06, ranging from 1.5 to 1.9 at the smaller ε distances.

The suspicion above most likely to be correct is that using integer values of ε actually represents too coarse a resolution to approximate the true fractal structure of the chaotic attractor. To find a more representative depiction of this structure, we use a smaller bin distance ε for estimating the cumulative correlation $C(\varepsilon)$ from which we

obtain ν. By decreasing the size of the ε bins, we now look for dimension estimates in the microscale range. Using the same data set as above, we now increment the ε value in hundredths. Figure 4.8 shows a 1.0 percent initial condition correlation dimension plot for ε values ranging between 0.01 and five. Comparison of Figure 4.8 with Figure 4.5 reveals a marked microstructure not seen in the larger scale. Unfortunately, that structure continues to show little if any convergence to a value of ν near the accepted value of 2.06. The other two initial conditions display similar behavior, and so these results are discouraging.

We consider yet another factor. As with the early histogram plots, these correlation dimension plots may be characterized by some noisy contamination. This is a plausible argument, as these small ε distances most likely produce a data set containing some intrinsic roundoff error. An even more satisfying argument is that, within the smallest bins, the percentage of total information contained within them is very small. We need only look at Figure 2.14, the plot of ln C(ε) versus ln ε, to verify this notion. In effect, there is most likely a distinct undersampling problem at the smallest ε distances.

To reduce this potential source of error and to provide some consistency with our smoothed Histogram Measure, we now use a bin smoother to determine a value of

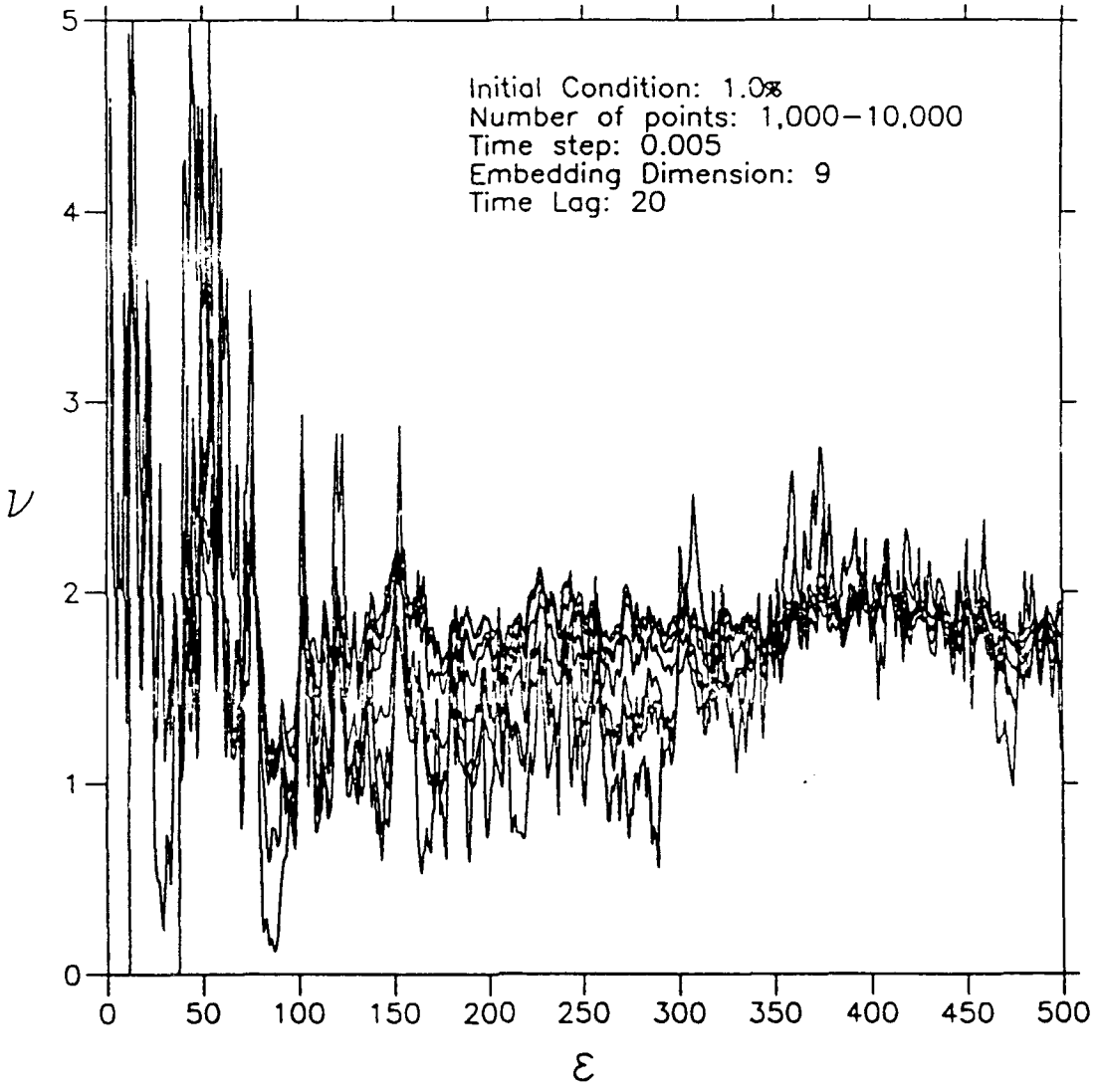


Figure 4.8: The correlation dimension value as a function of bin distance for a 1.0% initial condition and bin distances in hundredths ranging from 0.01 to five. It is discouraging to note the lack of convergence in the smallest of bin distances to a correlation dimension value above two.

ν. This smoother averages the cumulative correlation integral $C(\varepsilon)$ values over the number of bins we specify before we employ the slope finder to calculate the dimension. Although we tried many smoothers, all produce relatively disappointing results. Figure 4.9 shows the 1.0 percent plot for a five-bin grouping on the data and given by

$$\overline{C(\varepsilon)} = \left[\frac{C(\varepsilon\text{-}2) + C(\varepsilon\text{-}1) + C(\varepsilon) + C(\varepsilon\text{+}1) + C(\varepsilon\text{+}2)}{5} \right] . \qquad (4.2)$$

We have used this *nonweighted* smoothing function for simplicity. Unfortunately, we still have found no pronounced convergence of ν near to 2.06 and no consistency among initial conditions.

Efforts to find convergence to an acceptable correlation dimension value up to this point have been fruitless. Before we give up hope, however, there is one other sampling issue we have yet to discuss: the time step t_S that we use to produce the model-generated data sets. Fortunately, it is this particular issue that yields the resolution of the problem.

The original time step value of 0.005 was used in order to sample the attractor very accurately. We now choose a significantly larger time step of 0.05 to produce a

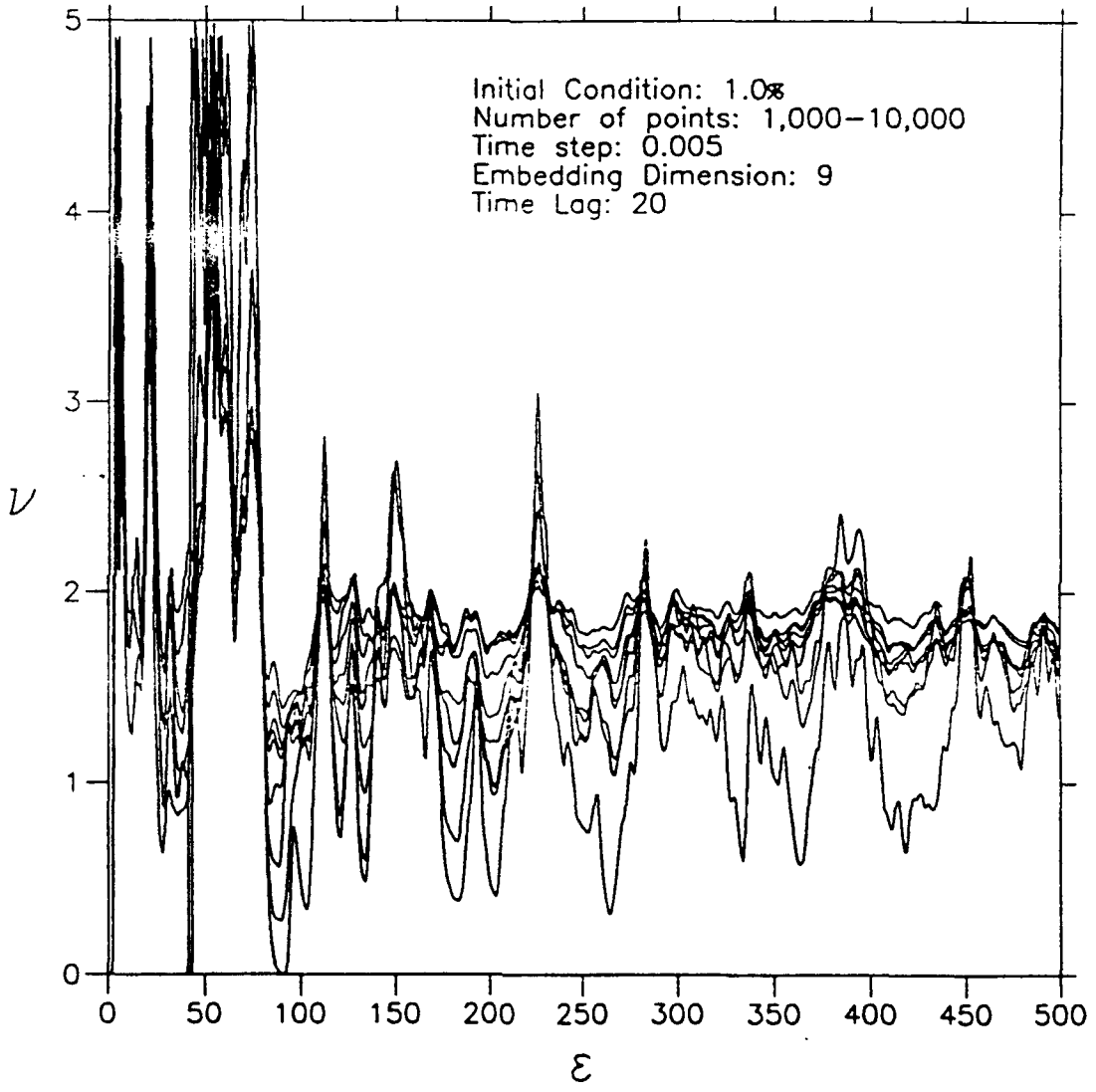


Figure 4.9: The correlation dimension value as a function of bin distance for a 1.0% initial condition with bin distances in hundredths ranging from 0.01 to five, after a nonweighted five–bin smoother has been applied to the data. Despite the smoothing of the data, there still is no significant convergence to a correlation dimension value above two.

new data set, and then we redo the correlation dimension calculations using the data from 990,001 to 1,000,000 time steps. Before we can accomplish this, however, we must ensure that the time lag Δt is appropriate for the new time step value. After testing numerous values, we decide that the optimum reconstruction occurs at a Δt value of 2, as seen in Figure 4.10. We first test the new time series by specifying integer ε values and d_E of nine. These correlation dimension plots are shown for ten data intervals and for all three initial conditions in Figures 4.11-4.13. Although the general shapes here are similar to those in Figures 4.5-4.7, encouraging differences are apparent. First, we are pleased to note a significant increase in the value of ν at the smallest ε distances (0.01-5), increasing from between 1.5 and 1.8 to between 2.0 and 2.1. Second, we observe a strikingly greater degree of convergence displayed by superimposing the results from the ten separate data intervals that we have chosen. Unlike the results that we obtained with the smaller time step, we now observe that a 1,000-point data set exhibits behavior that is more consistent with that given by a 10,000-point set. However, we still note that some structural differences remain among the three initial conditions. Even so, we have renewed hope that we can obtain some convergent behavior in ν to a value near 2.06 at small ε distances.

Utilizing the same microscale ε distances and five-bin smoother (4.2), we

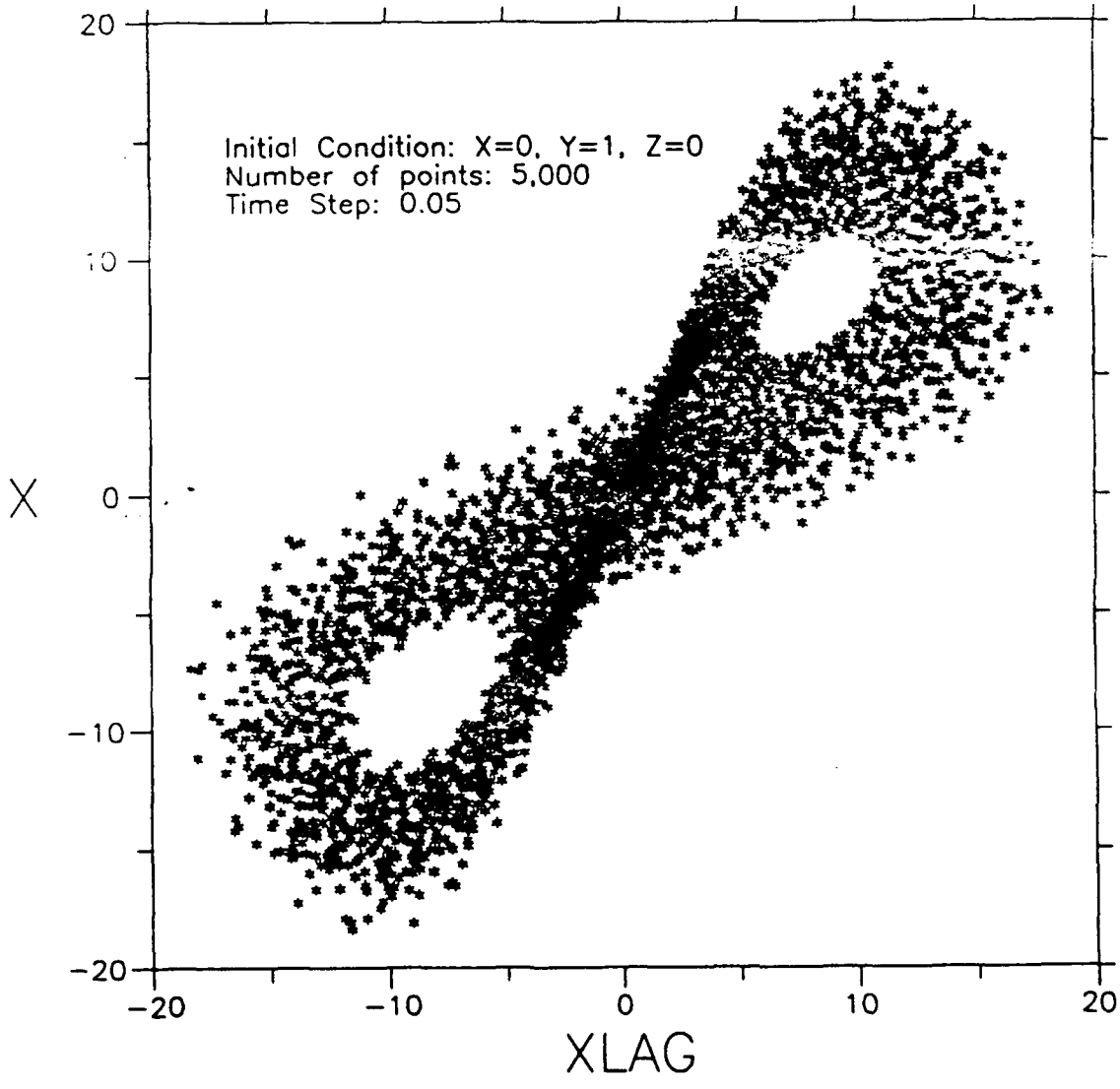


Figure 4.10: The reconstructed Lorenz attractor X against the lagged series XLAG for a time lag value of 2. For this larger time step value, this lag value represents the optimum one to use when calculating the correlation dimension.

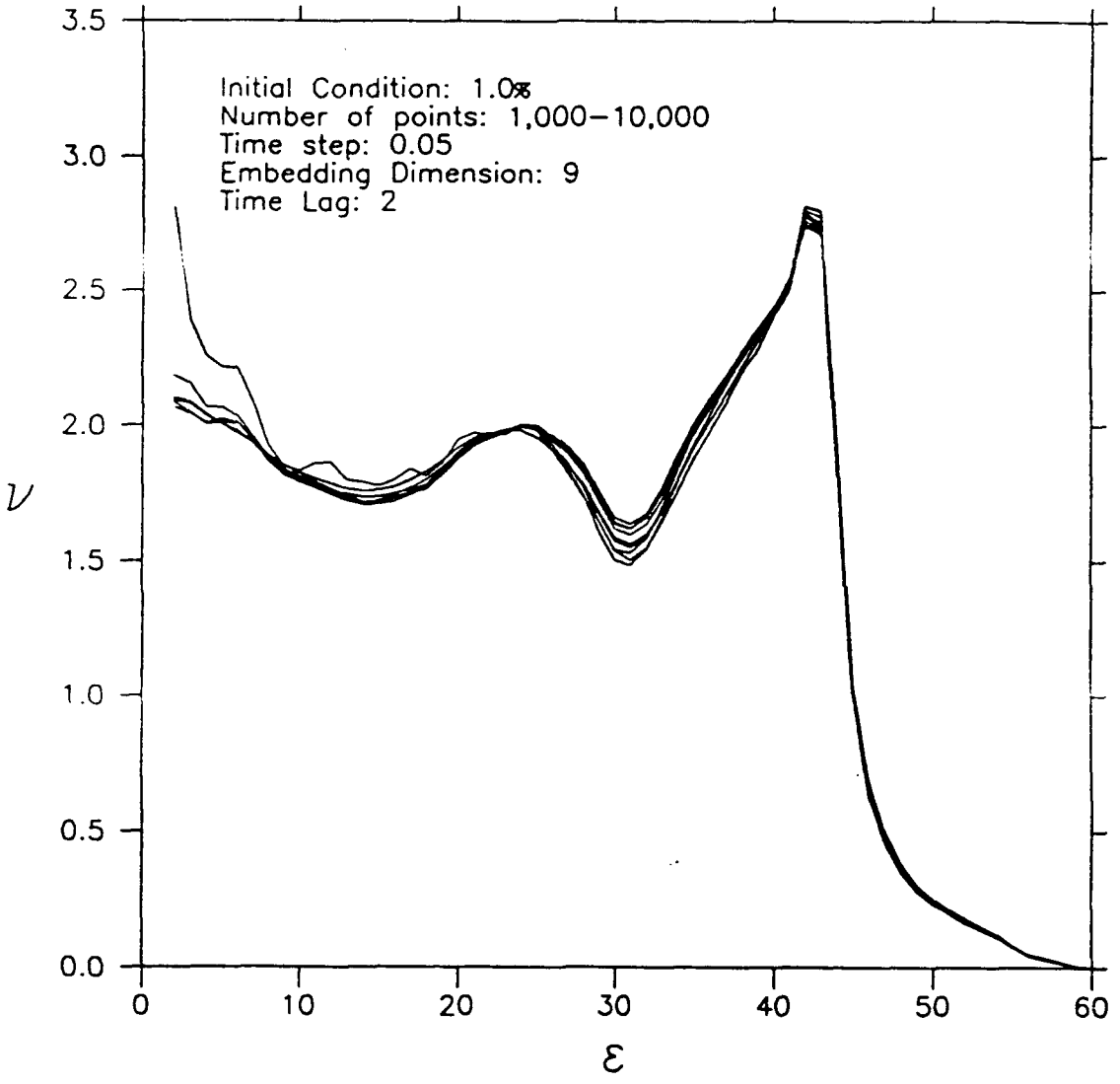


Figure 4.11: The correlation dimension value as a function of bin distance for a 1.0% initial condition and integer bin distances. Increasing the time step value changes the behavior dramatically. Unlike the great dependence of correlation dimension value upon interval length that was seen earlier with the smaller time step, all of the curves converge quite well. In addition, there is now convergence to a value above two for the smallest bin distances.

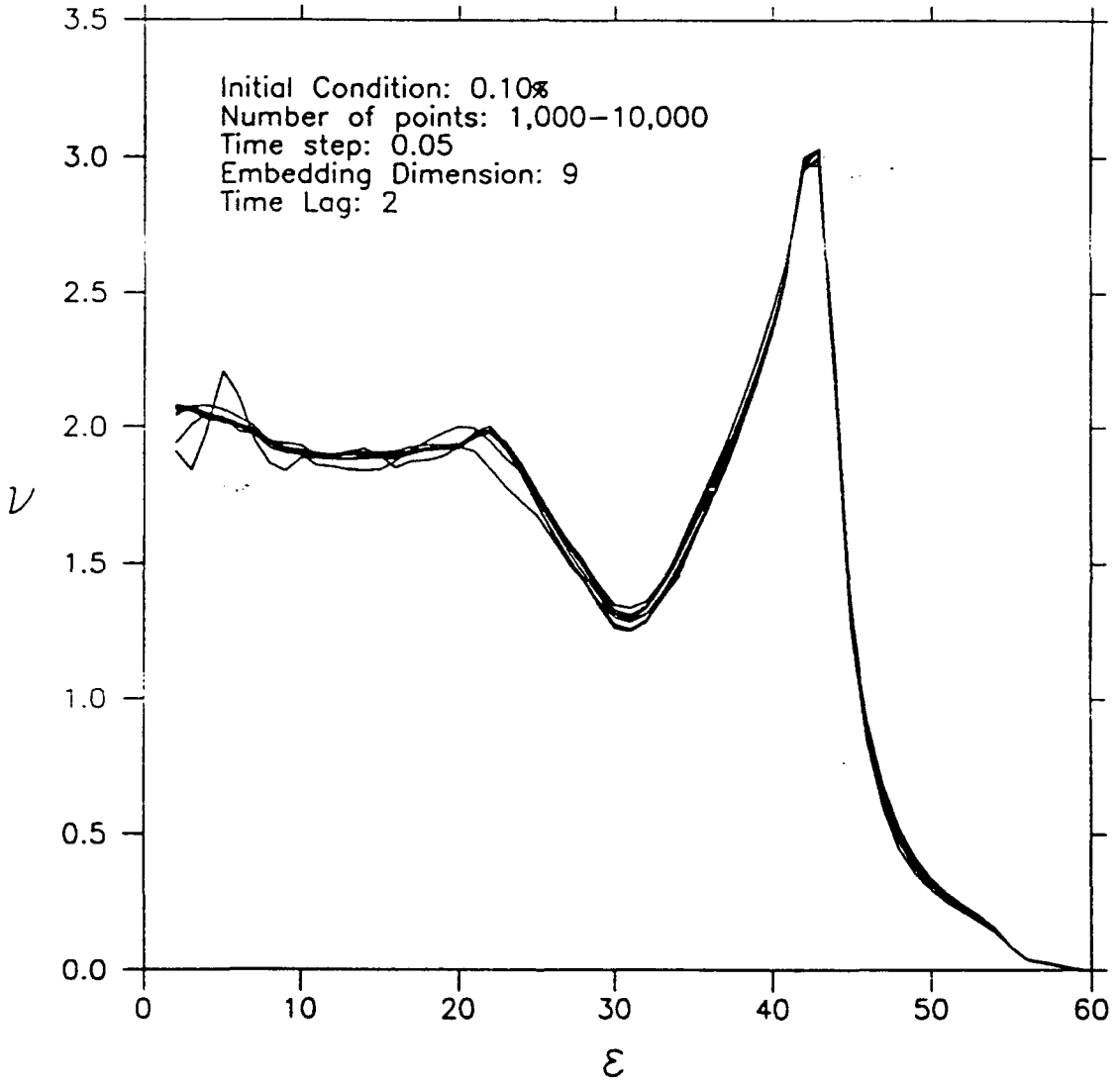


Figure 4.12: The correlation dimension value as a function of bin distance for a 0.10% initial condition and integer bin distances. Again, with the larger time step, there is strong convergence between all the data intervals and the correlation dimension values for the smallest distance bins are reassuringly above two.

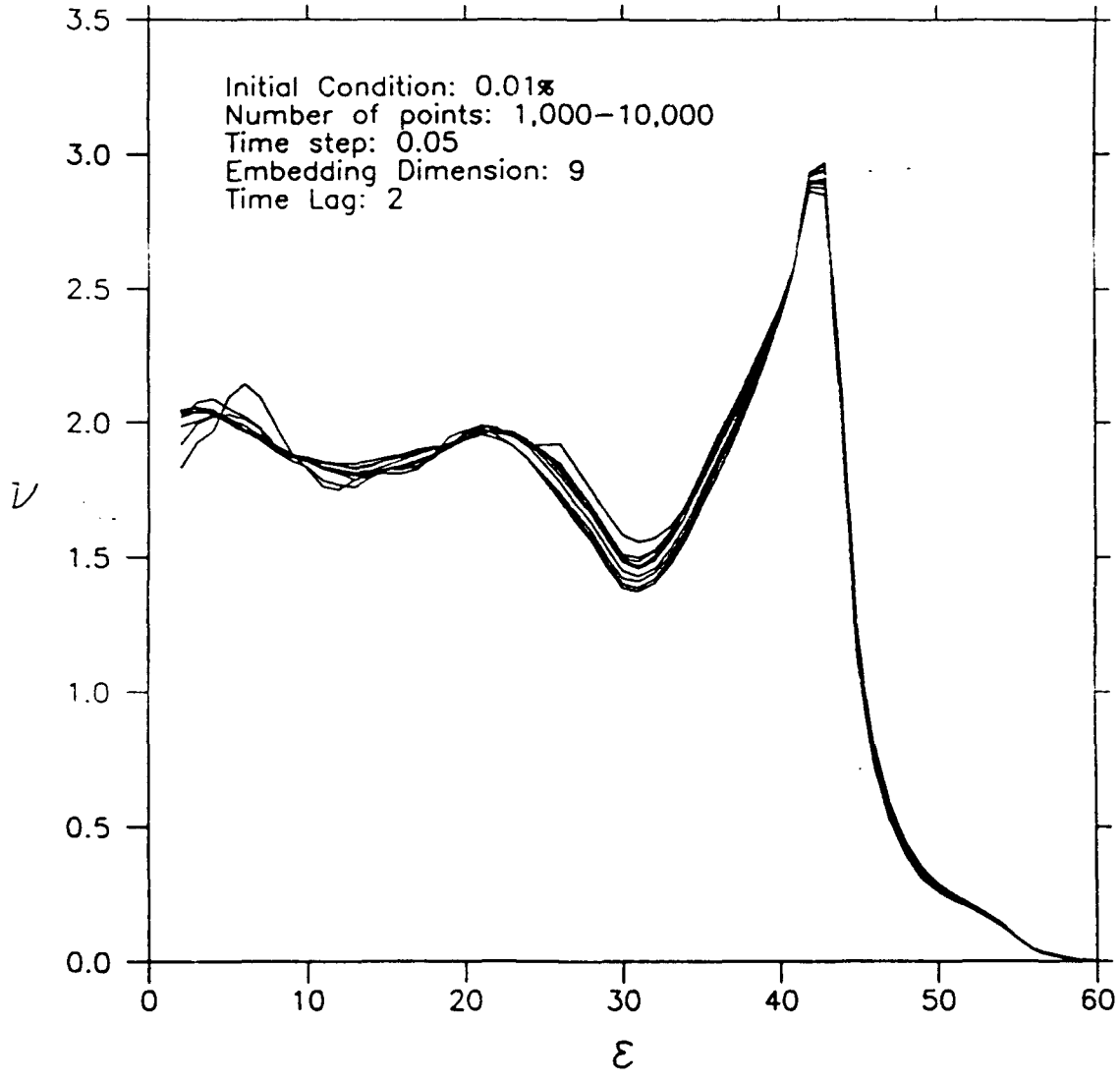


Figure 4.13: The correlation dimension value as a function of bin distance for a 0.01% initial condition and integer bin distances. Consistent with the other two initial conditions using the larger time step, the data intervals exhibit strong convergence and the correlation dimension values for the smallest distance bins remain above two.

calculate ν again for an ε distance range from 0.01 to five for each of the three initial conditions. These very encouraging results are shown in Figures 4.14-4.16. We argue in all three cases that after some initial noisy behavior for $\varepsilon \leq 1$, we achieve some convergent behavior near $\nu = 2.06$ once we have used approximately 3,000 points. For $1 \leq \varepsilon \leq 5$, the variance σ of ν about 2.06 is remarkably small, leading us to conclude that this may indeed be the elusive scaling region that we have been seeking. Further tests reveal that beyond $\varepsilon = 5$, the value of ν begins decreasing slowly below a value of two. Even more remarkably and reassuringly, we note that this convergence shows little, if any, difference among the three initial conditions.

4.3. The Histogram Measure Revisited

Concluding that we have now found time series that produce a reasonable convergence of ν to its widely accepted value, we now use the same series to look for histogram convergence to within a very small tolerance among the three initial conditions. Two possibilities exist:

1) We *will* find reasonable convergence in the three histograms,

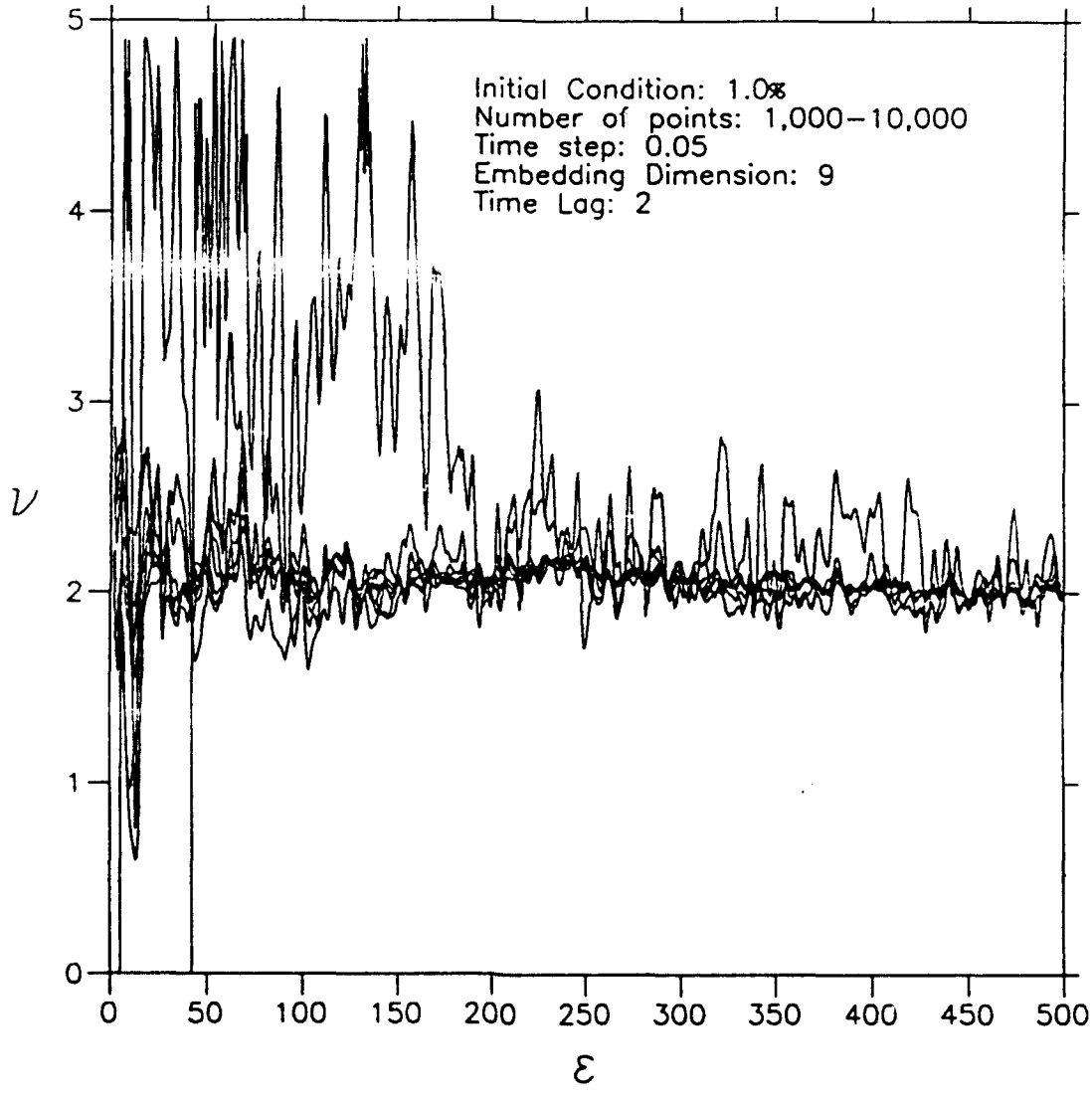


Figure 4.14: The correlation dimension value as a function of bin distance for a 1.0% initial condition with bin distances in hundredths ranging from 0.01 to five, after a nonweighted five–bin smoother has been applied to the data. Using the larger time step to produce the data sets yields a correlation dimension value that converges remarkably to within a small variance about its accepted value of 2.06 for values greater than approximately 100.

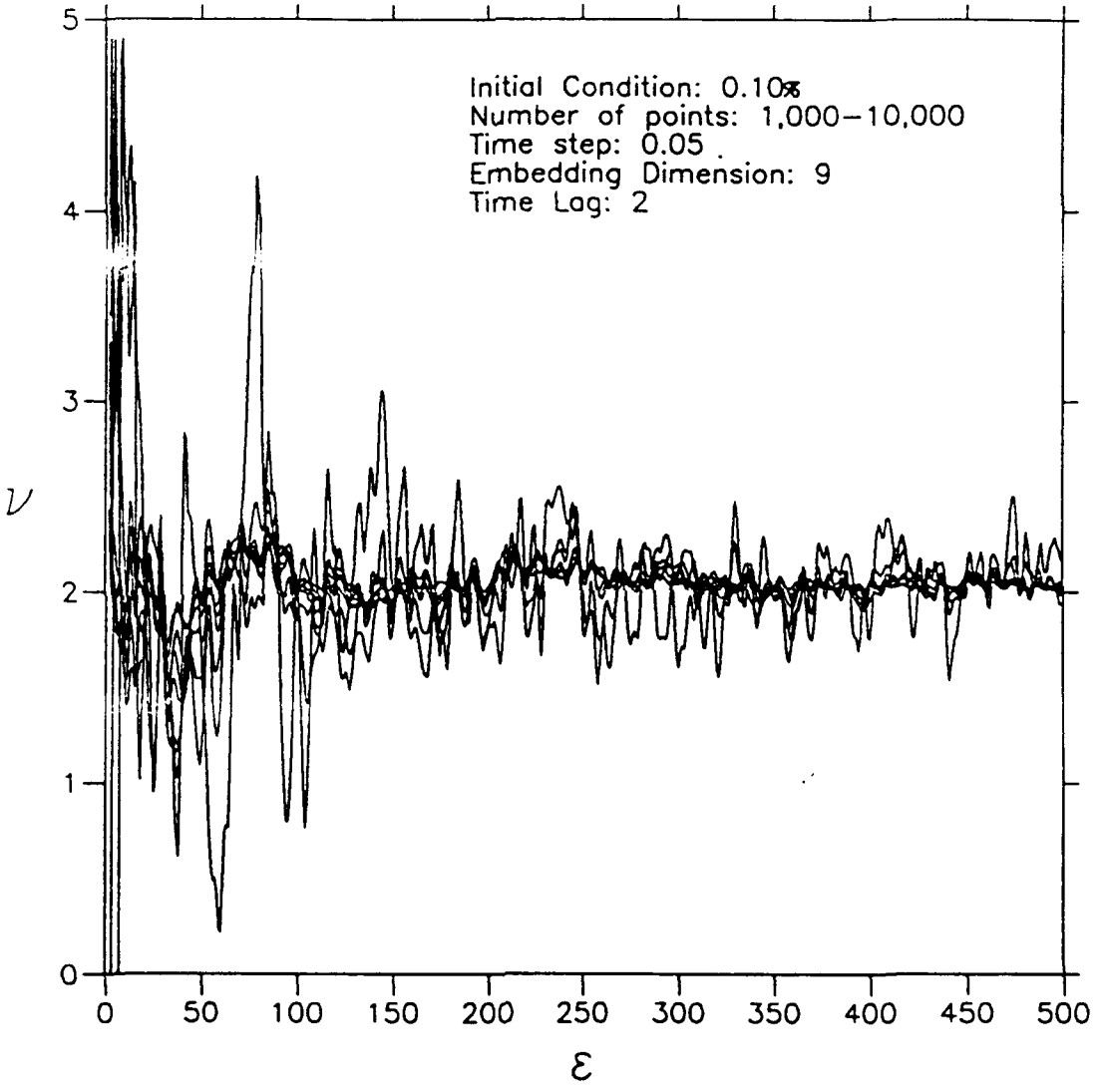


Figure 4.15: The correlation dimension value as a function of bin distance for a 0.10% initial condition with bin distances in hundredths ranging from 0.01 to five, after a nonweighted five–bin smoother has been applied to the data. Again, using the larger time step data sets yields correlation dimension values that converge to 2.06, to within a remarkably small variance.

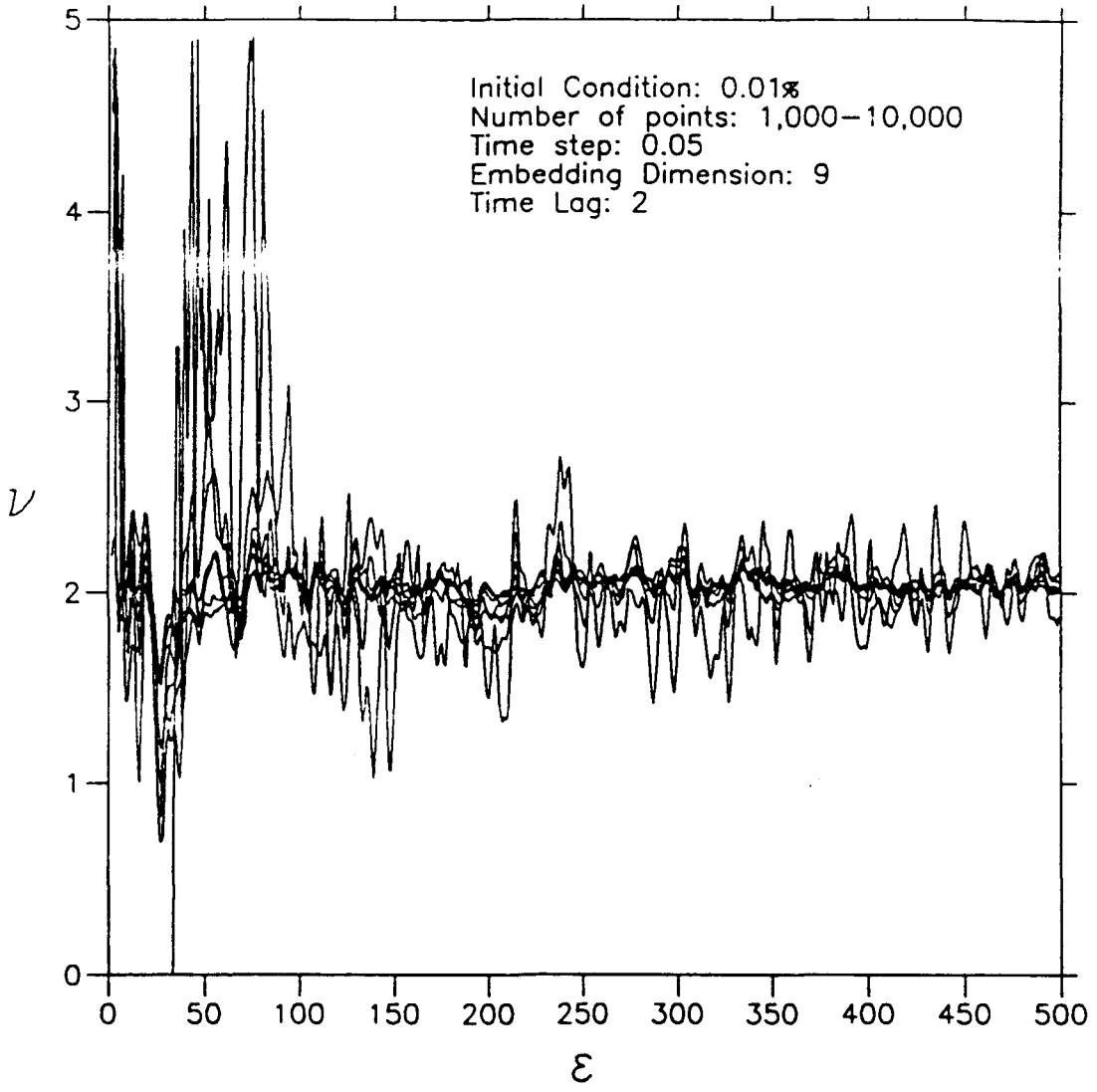


Figure 4.16: The correlation dimension value as a function of bin distance for a 0.01% initial condition with bin distances in hundredths ranging from 0.01 to five, after a nonweighted five–bin smoother has been applied to the data. Consistent with the data sets produced with the larger time step for the other two initial conditions, there is eventual convergence to a correlation dimension value that fluctuates closely about 2.06.

indicating that the histogram structure is relatively independent of initial condition. This finding is essential to the eventual development of sampling strategies and suggests that predictability estimates involving comparison of the divergence rates between numerous solutions having different initial conditions, as in the Monte Carlo technique, do have merit.

2) We *will not* find reasonable convergence in the histograms, because there are large, detectable, intrinsic differences in these numerically generated data sets. This finding would imply that developing representative sampling strategies would be very difficult and would lead us to conclude that predictability estimates would vary greatly with initial condition.

To look for convergence in the Histogram Measure, we calculate histograms using the 0.05 data set for the three initial conditions and compare their structures with each other. For now, we do not change the bin width value b_w, as was useful earlier when we calculated v. For the remainder of the chapter, we use $b_w = 0.4$. Figure 4.17 shows the histograms for the three initial conditions superimposed for the last 980,000 points of a 1,000,000-point data set. Unlike Figure 4.1 that shows a distinct lack of

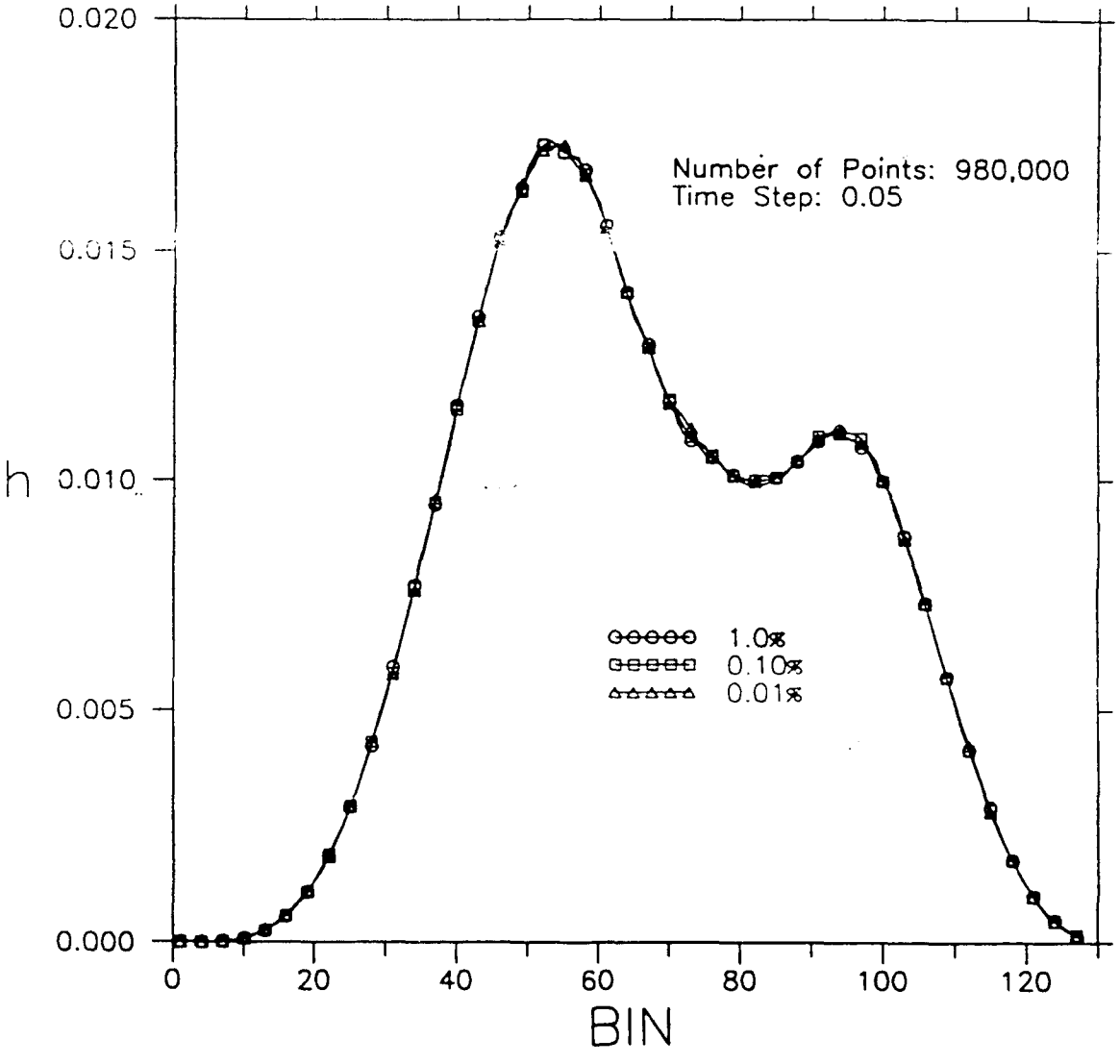


Figure 4.17: The three initial condition histograms superimposed on each other representing the last 980,000 points of a 1,000,000-point data set. Note that the degree of convergence exhibited by the three histograms is remarkably good when using the larger time step value.

convergence between the three histograms, this figure shows that the degree of convergence between the three initial conditions is extremely good. Indeed, we argue that the histograms *do* exhibit reasonable convergence.

Thus, as in the calculation of ν, we also conclude that the key issue in our efforts to detect convergence in the histograms is the time step size that we used to sample the attractor. Therefore, using 980,000 points at a time step of 0.005 most likely corresponds to a series length of only 98,000 points sampled at a time step of 0.05. Thus our early failures indicate a *distinct undersampling problem.* Our newly found results lead us to conclude that the robustness of some aspects of the Lorenz attractor is extremely sensitive to the manner in which we sample it; hence, finding an optimum sampling is crucial to its quantitative study. More specifically, we conclude that sampling more leaves of the attractor less accurately by using a coarser temporal resolution provides a more accurate representation of its chaotic characteristics than does sampling fewer leaves more accurately using a finer temporal resolution. This theory indicates that a certain minimum number of circuits around the attractor is necessary to define fully its quantitative information, a point that we address further later in the chapter.

Now that we have obtained what we believe to be a reasonable convergence of

the three histograms, we next wish to determine some minimum number of points that yields this convergence. To accomplish this objective, we first superimpose histograms that were obtained using the same series length for the three initial conditions. Figures 4.18-4.22 show these superimposed histogram plots for data sets lengthening by successive 100,000-point intervals ranging from 80,000 to 480,000 points. We observe that although the 480,000-point series still exhibits a pronounced convergence, there is a noticeable increase in the differences among the histograms with decreasing numbers of points; indeed the 80,000-point results exhibit very little convergence (Figure 4.18). Unfortunately, this method is quite subjective, and so it is difficult to determine which data interval is the minimum one necessary for producing convergence. As a result, we employ a more objective method to help us.

4.3.1. The Average Mean Absolute Difference Method

Instead of comparing superimposed histogram structures, we now define a measure of the average mean absolute difference among the three histograms. We recall that we used a similar method in Section 3.2 to find the transient portion of a series.

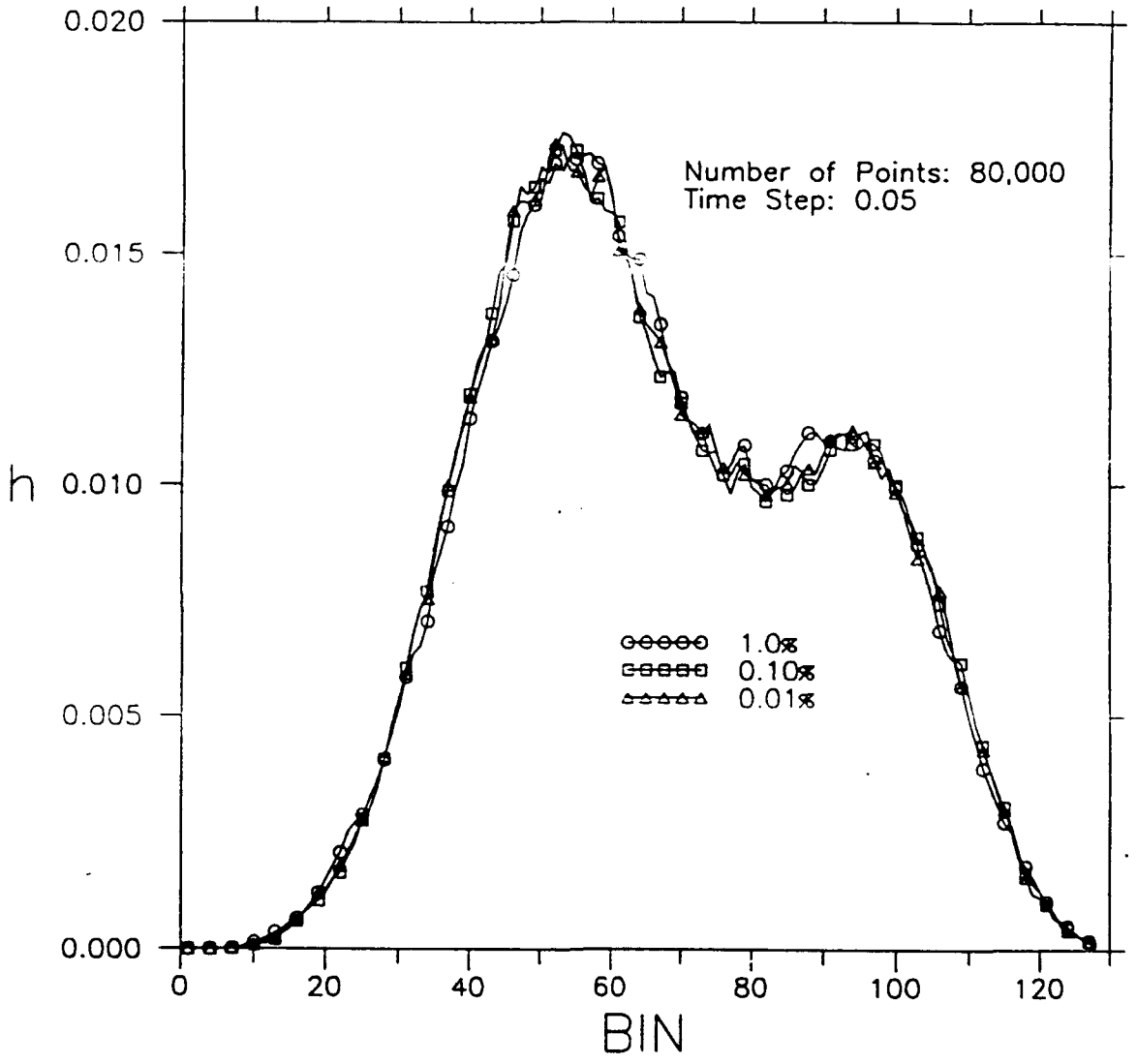


Figure 4.18: The three initial condition histograms superimposed on each other representing the last 80,000 points of a 100,000–point data set.

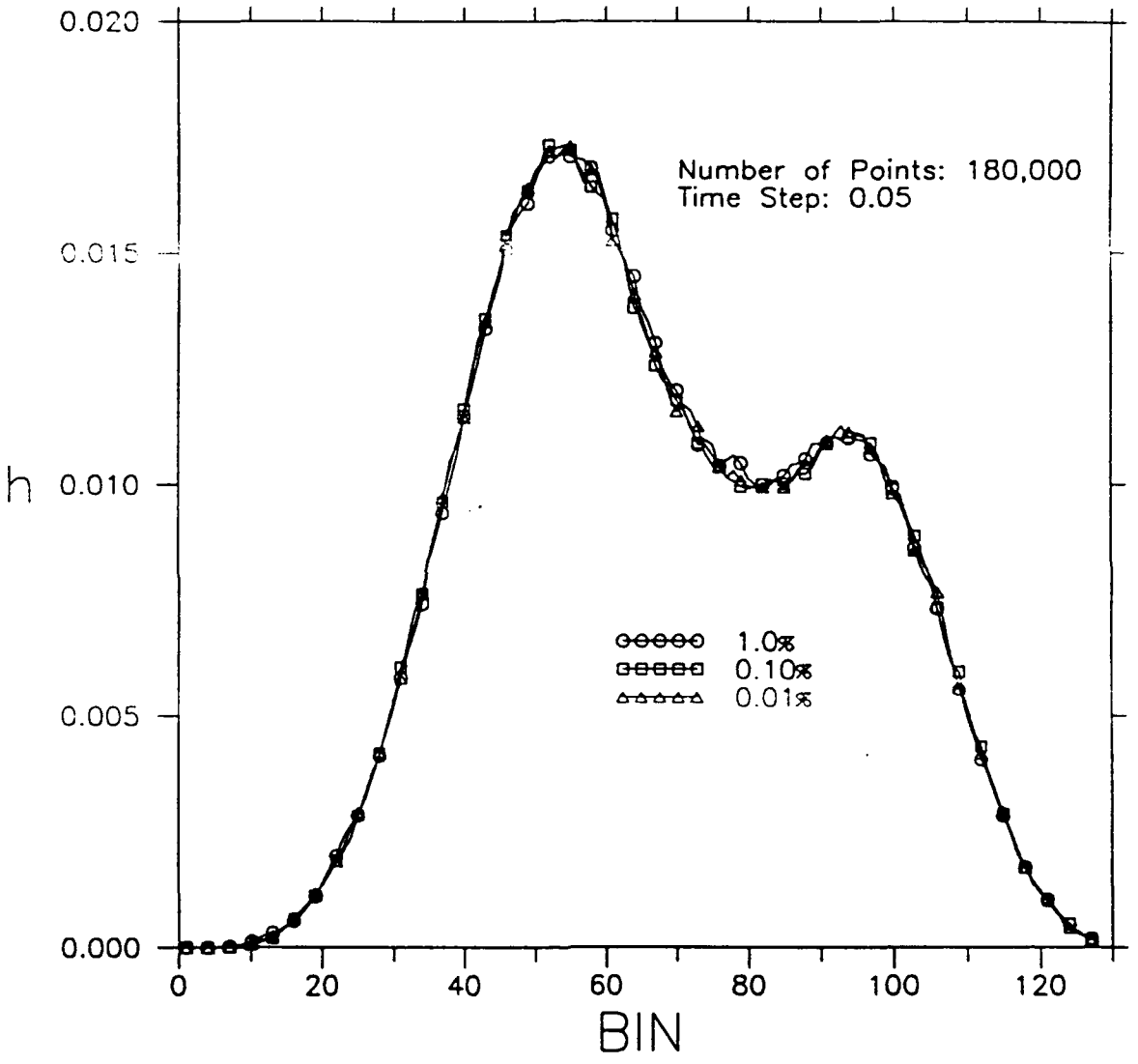


Figure 4.19: The three initial condition histograms superimposed on each other representing the last 180,000 points of a 200,000-point data set.

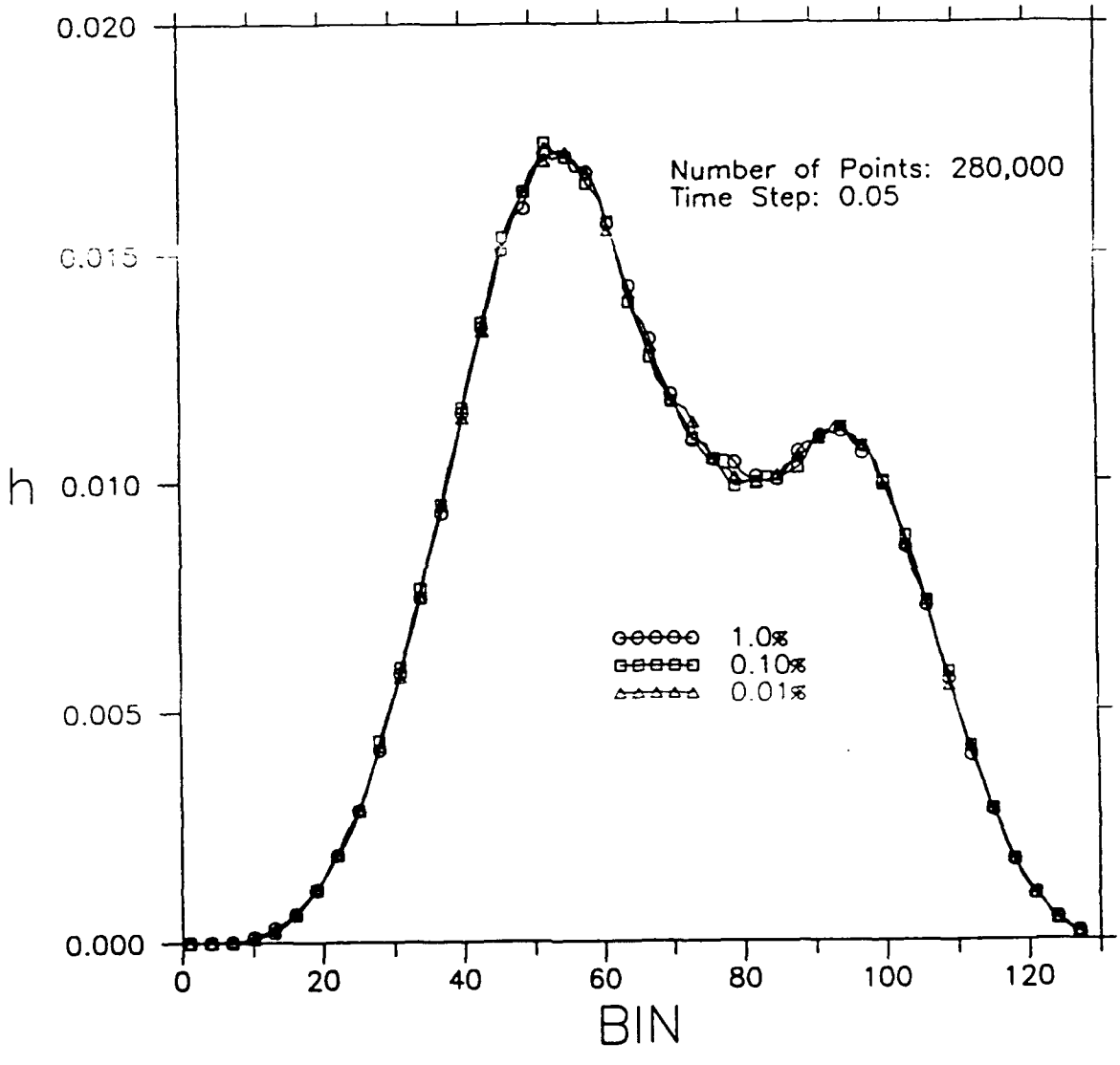


Figure 4.20: The three initial condition histograms superimposed on each other representing the last 280,000 points of a 300,000-point data set.

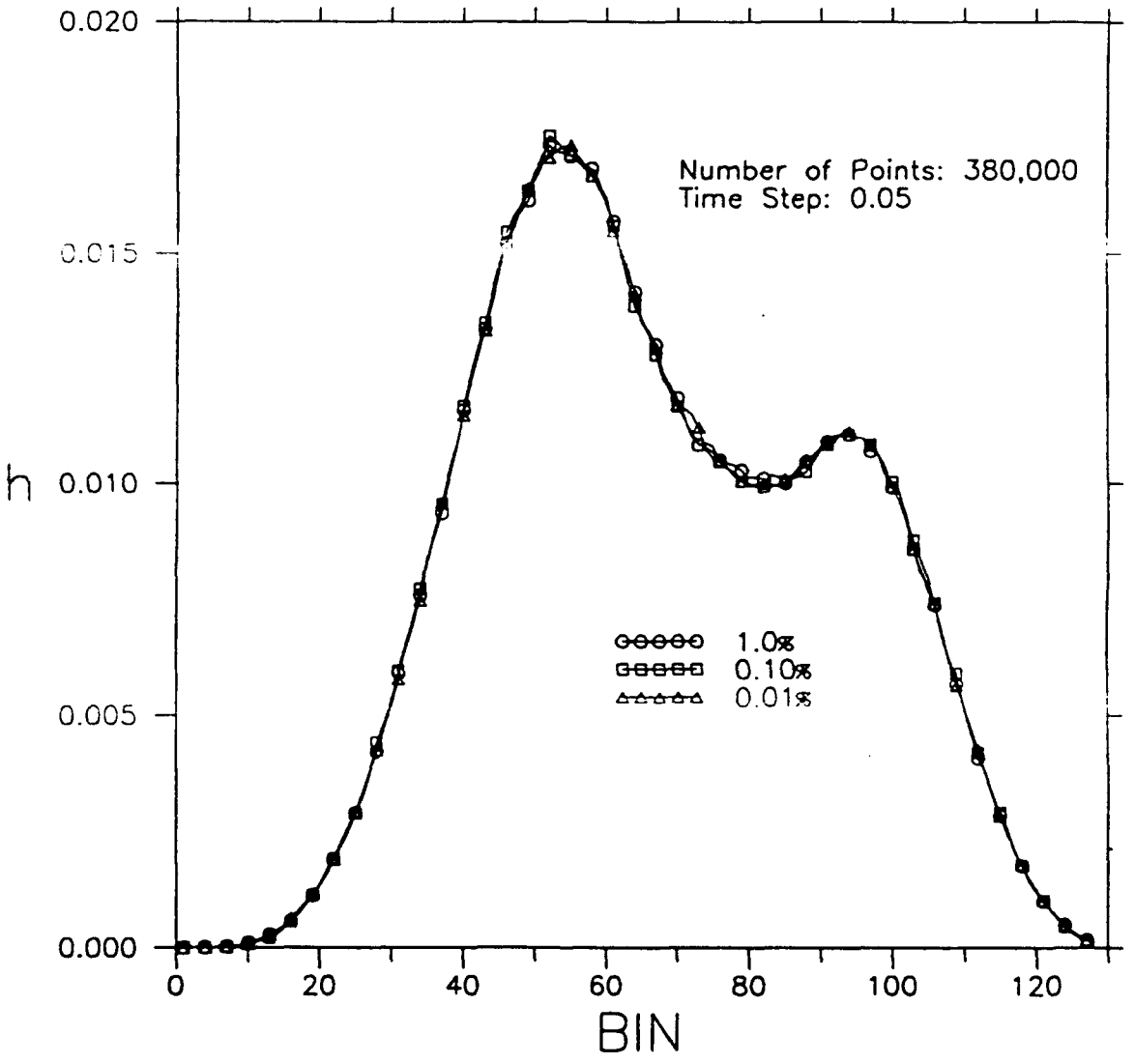


Figure 4.21: The three initial condition histograms superimposed on each other representing the last 380,000 points of a 400,000-point data set.

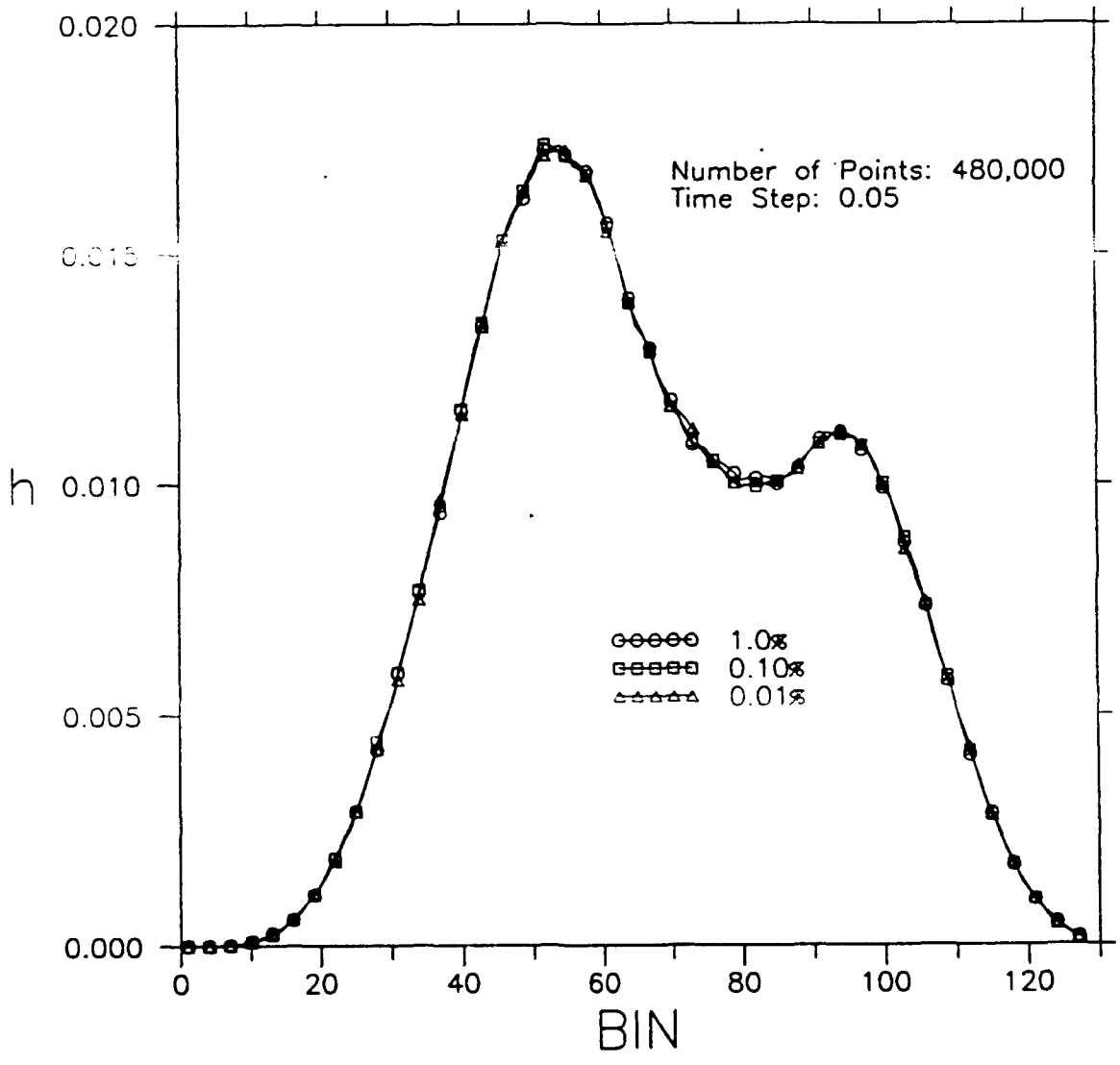


Figure 4.22: The three initial condition histograms superimposed on each other representing the last 480,000 points of a 500,000-point data set.

Here, we define a similar difference expression D_{ik} to that used in (3.1) and is given by

$$D_{ik} = \frac{1}{128} \sum_{j=1}^{128} \left[\left(H_i(z_j) - H_k(z_j) \right)^2 \right]^{1/2}, \qquad (4.3)$$

where j represents the bin number and H_i and H_k are the histograms for any of the three initial conditions with i or k = 1,2,3. We use (4.3) to calculate the three possible differences (D_{12}, D_{13}, D_{23}) and then sum them to produce the average mean absolute difference value D_{avg} given by

$$D_{avg} = \frac{D_{12} + D_{13} + D_{23}}{3}. \qquad (4.4)$$

In Figure 4.23, this average difference value is plotted as a function of series length L_{ser}. The ordinate represents D_{avg} and the abscissa denotes the series length in increments of 50,000 points up to length 800,000. We recall that with the first 20,000 points discarded, the value of 30 actually represents the last 30,000 points of a 50,000-point series and so on. We observe the existence of an exponential type of decrease in the average difference values D_{avg} among the three initial conditions as the interval length increases. Convergence to a relatively constant D_{avg} value is apparent in

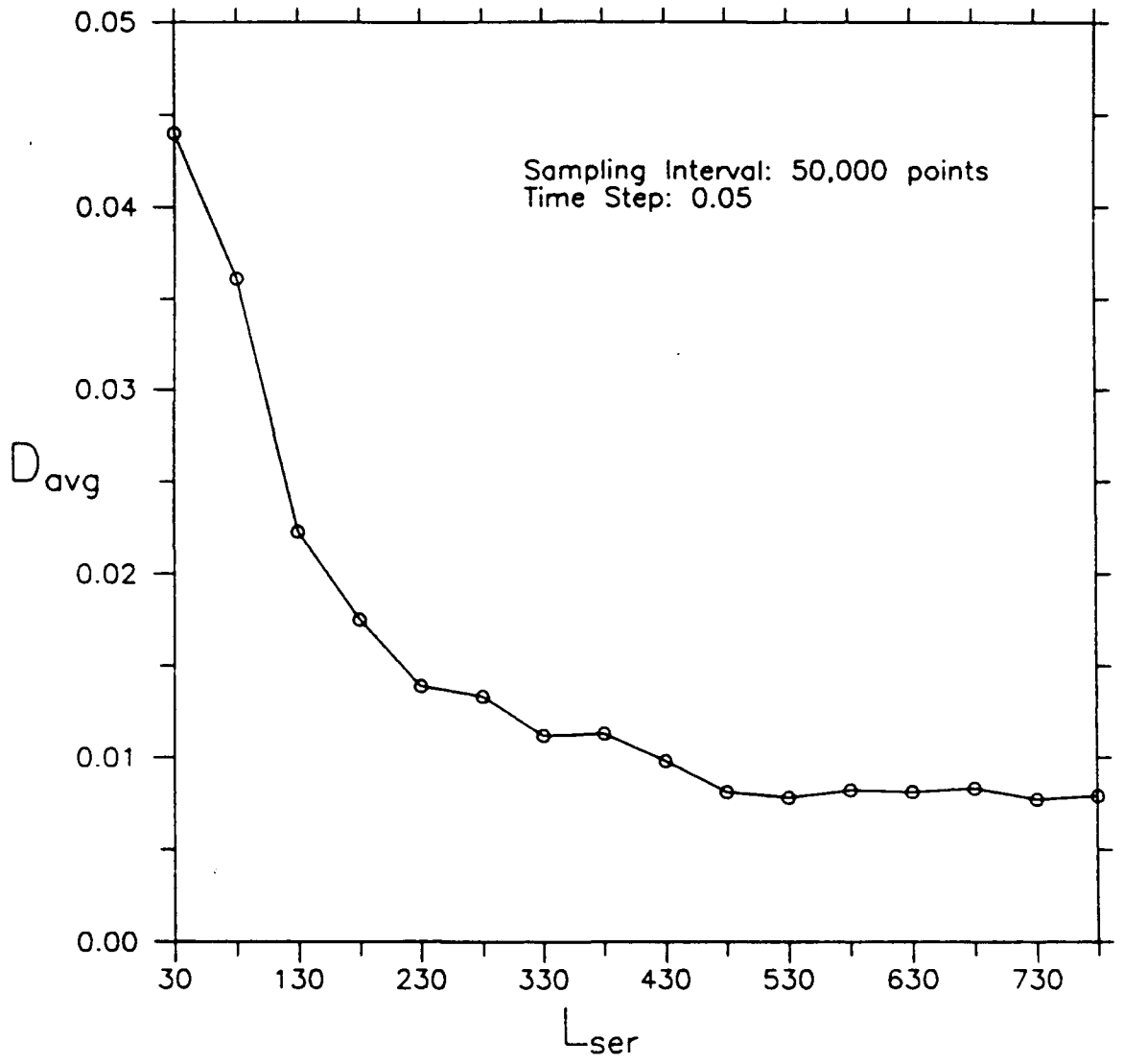


Figure 4.23: The average mean absolute difference values D_{avg} between the three initial condition histograms as a function of series length L_{ser} (in thousands). Note that the difference values decrease to a floor with an average value of approximately 0.008 by 480,000 points.

Figure 4.23; this value ($D_{avg} \sim 0.008$) represents the *average minimum intrinsic difference* among the histograms and may be related to using finite temporal approximations to describe solutions of an ordinary differential system. Having successfully quantified a difference value representing a reasonable histogram convergence, we seek the series that first displays this minimum difference or *floor* value; after using this number of points, we feel confident that we have found the greatest achievable convergence among the three initial conditions. Using this argument and Table 4.1, we conclude that 480,000 points represents the minimum series length needed to specify a control histogram because a longer series does not yield a better one.

There is another way to view these results. Although we are quite certain of convergence by 480,000 points, if similar results hold for larger numerical models that are much more expensive to run, then we would need to be able to make estimates with far fewer points. We may interpret the results in Figure 4.23 as giving a relation between the numerical accuracy of the Histogram Measure and the number of points we may consider. Wishing to use as few points as possible, and still have reasonable confidence in our results, we notice that the bulk of the difference between the three initial conditions has been eliminated by 230,000 points and yields an average mean

Table 4.1: The average mean absolute difference values D_{avg} as a function of the series length L_{ser}. Note a continuous decrease in values up to a series length of 480,000 points, after which the values fluctuate about approximately 0.008. This value represents the minimum average intrinsic difference between the histograms and suggests that beyond 480,000 points, no additional information can be gained.

L_{ser}	D_{avg}
30,000	0.0440
80,000	0.0361
130,000	0.0223
180,000	0.0177
230,000	0.0139
280,000	0.0133
330,000	0.0113
380,000	0.0112
430,000	0.0098
480,000	0.0081
530,000	0.0078
580,000	0.0082
630,000	0.0081
680,000	0.0083
730,000	0.0077
780,000	0.0079

absolute difference value of approximately 0.014. Thus, by using half the number of points, we have only increased our average difference value by approximately 0.006, which is rather small. Because there exists a tradeoff between obtaining the maximum possible accuracy in defining convergence and the expense we incur by doing so, we must determine the optimal choice for our particular application.

Another issue that we consider is the rate of decrease in difference value as a function of series length. Quantifying this decrease certainly has ramifications for estimating the predictability of an attractor. We discuss this issue in detail later in this chapter.

4.3.2. The Asymptotic Mean Absolute Difference Method

Having quantified convergence among the three initial conditions, we feel confident that there must also exist a convergent or control histogram *within* each of the three cases. We recall from Figures 4.14-4.16 that the correlation dimension calculations for the 0.05 time step data set showed a pronounced convergence to its accepted value of 2.06 within each of the initial conditions by approximately 3,000

points. Understanding the importance of using the larger time step to produce data sets yielding convergence in the value of ν, we now hope to quantify successfully histogram convergence using the data set produced with the larger time step.

As we did previously, we first seek a qualitative judgement about the convergent properties of the histograms. We do so by superimposing five histograms obtained for successively longer series for a single initial condition. These series range from the first 80,000 to 480,000 points--again we recall that the initial 20,000 points for all series have been removed. Figures 4.24-4.26 show the results that we obtain for each of the three initial conditions; all indicate a distinct convergence in structure. This finding certainly supports the existence of convergence by 480,000 points, as found in Figure 4.23.

To quantify this convergence, we use a difference method similar to the ones already used in (3.1) and (4.3); only now we calculate the mean absolute difference between the normalized histograms H_a and H_b for *series of increasing lengths* within one initial condition, where the asymptotic mean absolute difference D_{ab} is given by

$$D_{ab} = \frac{1}{128} \sum_{j=1}^{128} \left[\left(H_b(z_j) - H_a(z_j) \right)^2 \right]^{1/2}, \qquad (4.5)$$

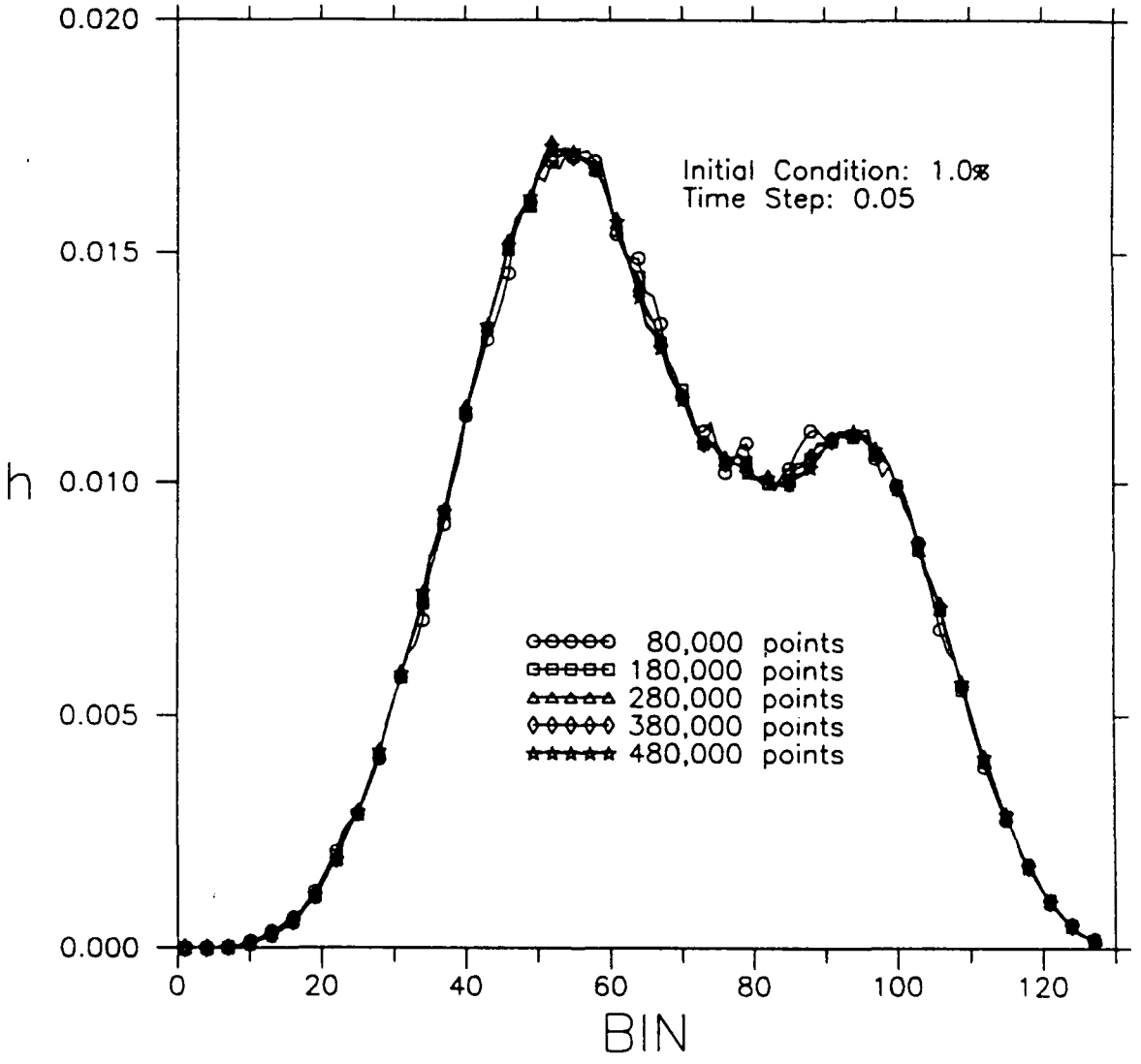


Figure 4.24: The histograms superimposed on each other representing the five data set lengths within the 1.0% initial condition.

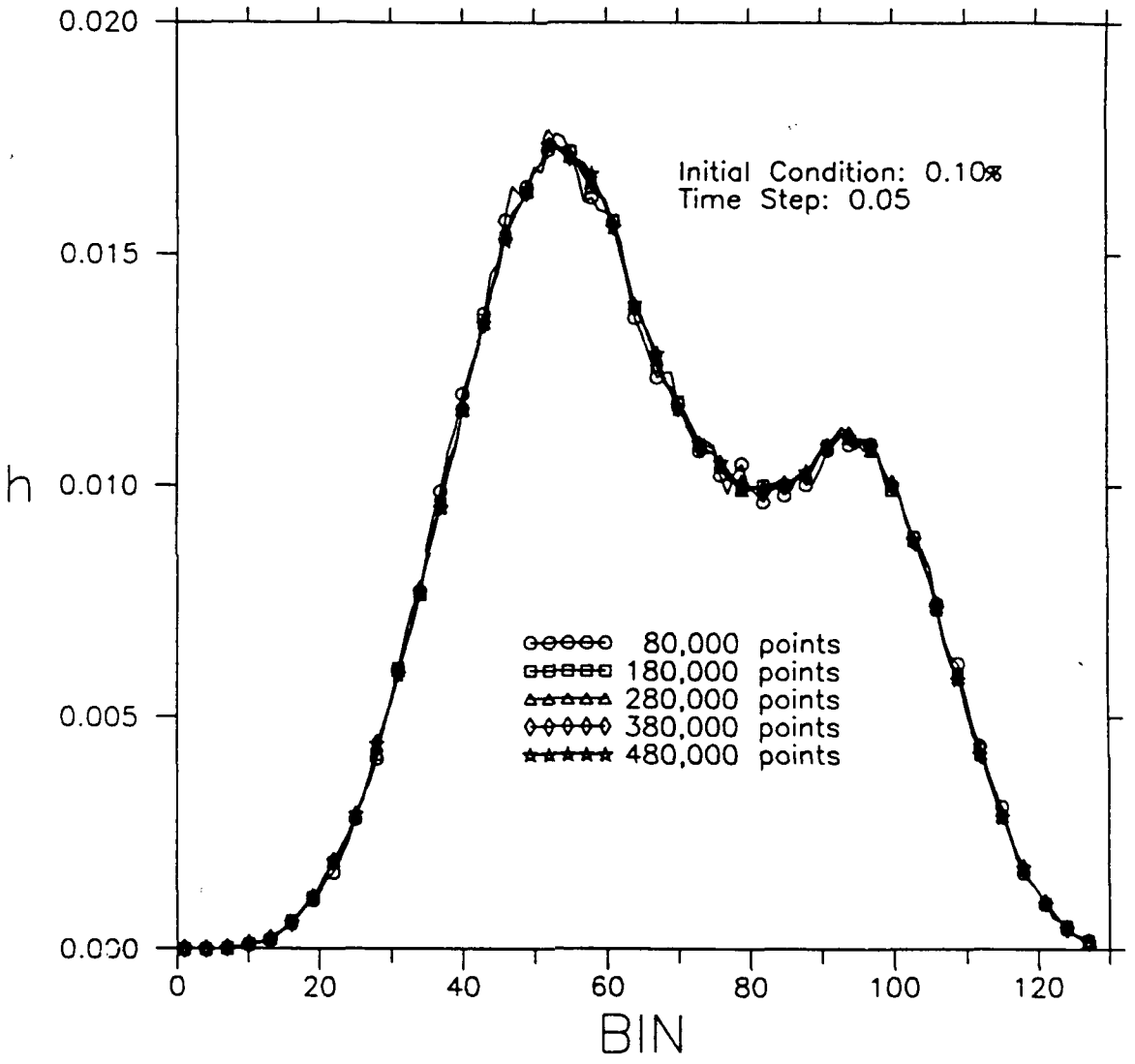


Figure 4.25: The histograms superimposed on each other representing the five data set lengths within the 0.10% initial condition.

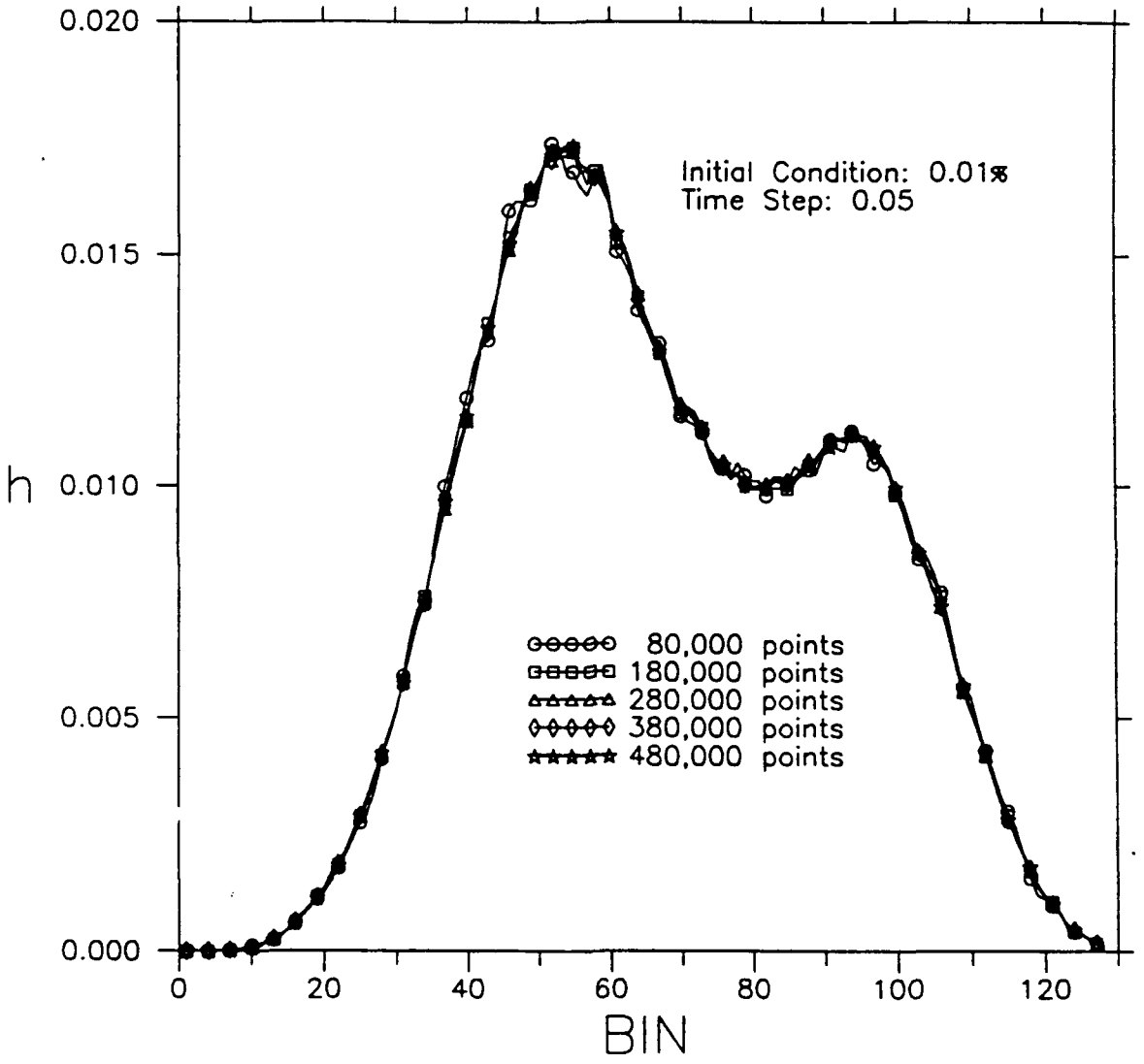


Figure 4.26: The histograms superimposed on each other representing the five data set lengths within the 0.01% initial condition.

Figure 4.27 shows these results for each of the three initial conditions. As in Figure 4.23, the ordinate is simply the magnitude D_{ab} of the difference values. The abscissa now has a different interpretation than that of D_{avg}; each of the values of D_{ab} represents a comparison between two sample lengths. For example, at the value of 30 is given the comparison of histograms between the last 30,000 points of a 50,000-point series and the last 80,000 points of a 100,000-point series; at the value of 80 is given the comparison between the 80,000-point and 130,000-point histograms and so on.

What initially strikes us is the consistent behavior displayed by all three initial conditions. As before, we observe an early, exponential-type decrease in difference values with increasing series lengths. Employing the same argument that we used earlier to define histogram convergence, we note that the differences in all three cases approach a distinct floor. However, given the behavior of all three cases, we are less certain of a definitive floor value and therefore the length of the data series at which the histogram structures converge than we were with D_{avg}. In Table 4.2, the difference values are given for all three cases. To determine the minimum length of the data set for which we find convergence, we use the *larger* of the two series interval values that are compared with one another. Using this convention, we find that convergence occurs by 430,000 points at values generally near $D_{ab} = 0.0026$. However, in the next interval

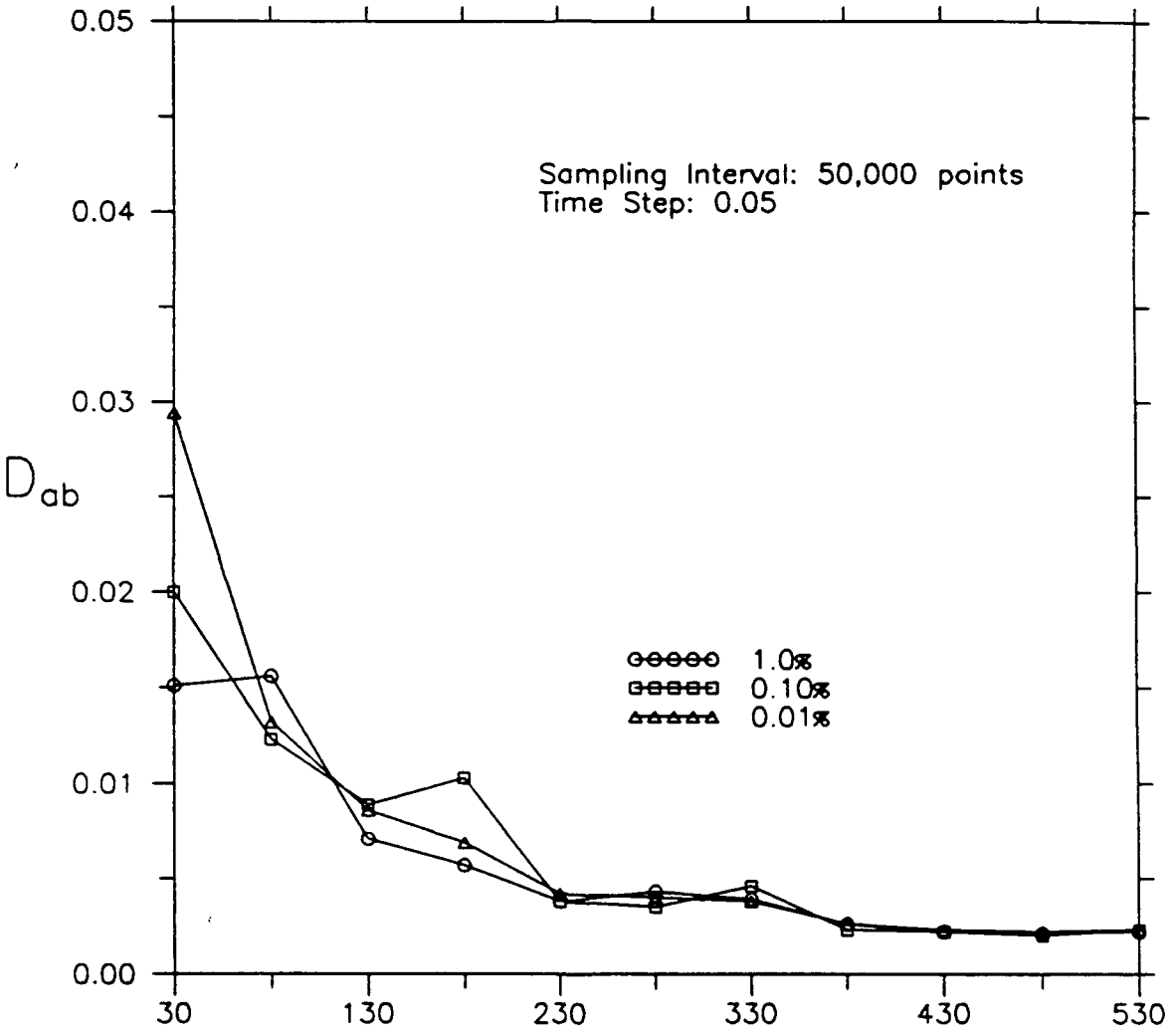


Figure 4.27: The asymptotic mean absolute difference values D_{ab} for each of the initial condition cases as a function of increasing series length comparison (in thousands) for a sampling interval of 50,000 points. Note that all sets of difference values decrease to a similar floor value.

Table 4.2: The asymptotic mean absolute difference values D_{ab} for all three initial conditions and a sampling interval of 50,000 points as a function of increasing histogram series length comparison (H_a and H_b). The average D_{ab} value ($\overline{D_{ab}}$) has been computed for each initial condition over a range of series intervals in which the difference values fluctuate about a minimum range; these values represent the average minimum intrinsic difference between histograms of increasing series length for this particular sample size. The series length comparison interval displayed in bold type represents that after which no additional information can be gained (480,000 points).

Histogram Series Length		**D_{ab}**		
H_a	**H_b**	**1.0%**	**0.10%**	**0.01%**
30,000	80,000	0.0151	0.0200	0.0294
80,000	130,000	0.0156	0.0123	0.0132
130,000	180,000	0.0071	0.0089	0.0086
180,000	230,000	0.0057	0.0103	0.0069
230,000	280,000	0.0038	0.0038	0.0042
280,000	330,000	0.0043	0.0035	0.0040
330,000	380,000	0.0039	0.0046	0.0038
380,000	430,000	0.0026	0.0023	0.0026
430,000	**480,000**	**0.0022**	**0.0022**	**0.0023**
480,000	530,000	0.0021	0.0023	0.0022
530,000	580,000	0.0023	0.0023	0.0023
305,000	330,000	0.0026	0.0026	0.0023
$\overline{D_{ab}}$ in range (430,000-580,000):		0.0022	0.0022	0.0023

(430,000-480,000 points), in two of the three cases, there is a continued drop to difference values closer to 0.0022. Beyond this point interval, the values generally fluctuate in the range between 0.0021 and 0.0023. As a result, although we are confident that a floor does exist, there is some subjectivity in determining its value.

We believe that a less ambiguous way to determine the value of the floor is to *average* the difference values of D_{ab} once we determine the range of points within which these D_{ab} values exist. This average value ($\overline{D_{ab}}$) assures us not only of better finding a minimum data set that defines convergence in each case, but also corresponds more closely with the definition of convergence given by the other difference method D_{avg}. Averaging these values for each of the three cases, we note more consistent minimum difference values between them: $\overline{D_{ab}} = 0.0022$ for the 1.0 and 0.1 percent cases and $\overline{D_{ab}} = 0.0023$ for the 0.01 case. Taking the larger of the two values as our difference floor (0.0023), we conclude that in all three cases the histogram structures converge by 480,000 points, a value consistent with the one that we found with D_{avg}. Using more points than this will yield little, if any, additional improvement in the information content of the histogram. As before, we see a similar tradeoff between the degree of accuracy we wish to achieve and the expense in generating a long series.

More specifically, although 480,000 points represents the minimum data set required to achieve maximum convergence, we observe that using 280,000 points is quite adequate for capturing the bulk of the convergence within all of the initial conditions.

Although these are good results, we have not fully addressed this convergence issue. Unlike the convergence that we found *among* the three initial conditions at a fixed data interval, we can calculate the convergence *within* each initial condition by incrementing the sample by various amounts. Up to now, the results that we have obtained are based solely upon the comparison of histograms with a sampling interval of 50,000 points.

To investigate this sensitivity, we employ the same tests as those used above for sampling intervals of 25,000 and 100,000 points. Figures 4.28 and 4.29 show these results. In comparing these figures with Figure 4.27, we observe distinct similarities and differences when the sampling interval is changed. Although all three cases display the same exponentially decreasing behavior, the most reassuring behavior among the three is their convergence to a reasonably similar average minimum difference value $\overline{D_{ab}}$. The difference values are given for the 25,000-point and 100,000-point intervals in Tables 4.3 and 4.4, respectively. Employing the same method as was used above, we

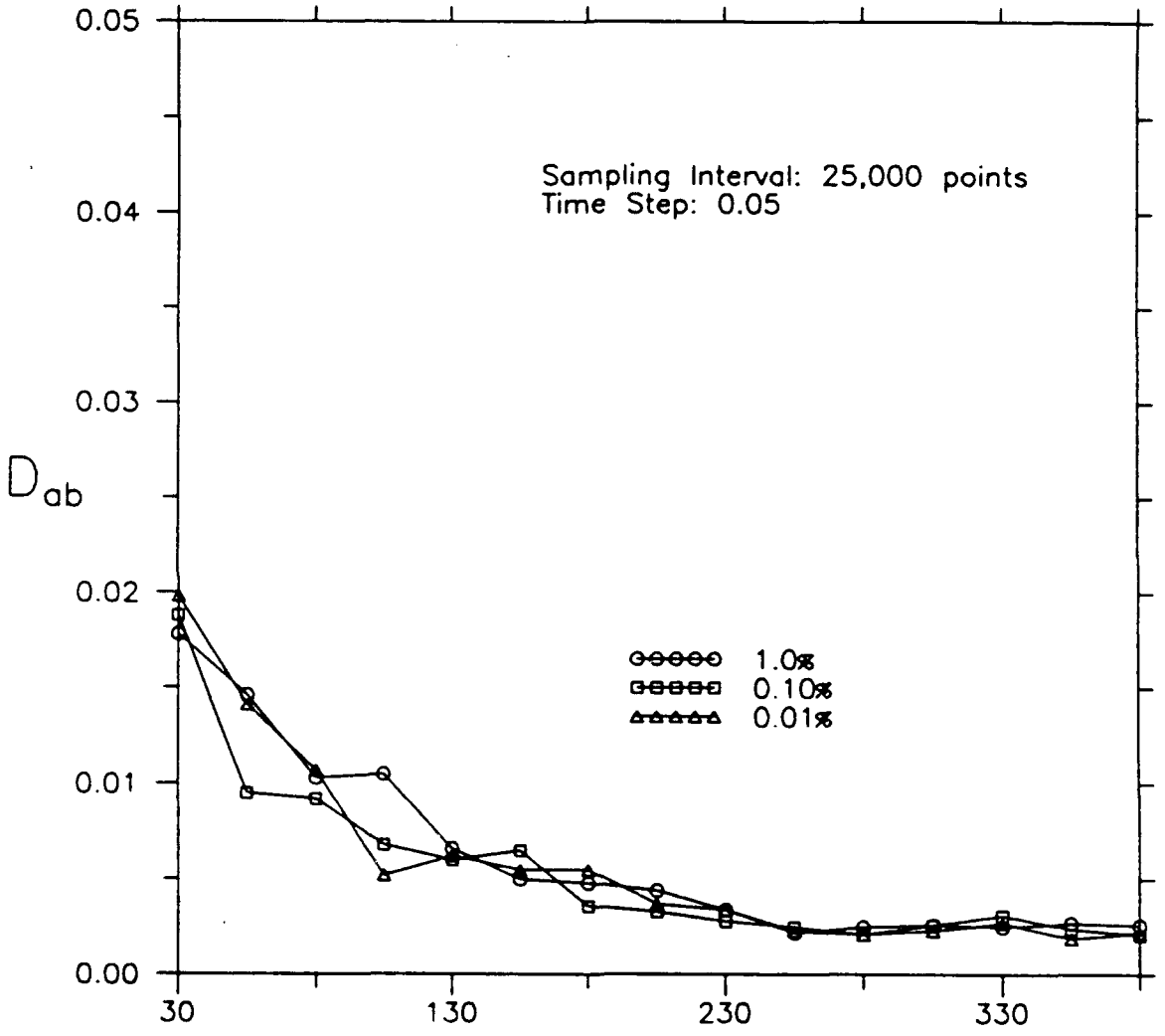


Figure 4.28: The asymptotic mean absolute difference values D_{ab} for each of the initial condition cases as a function of increasing series length comparison (in thousands) for a sampling interval of 25,000 points. Despite the change in sampling size, all three sets of difference values again decrease to similar floor values.

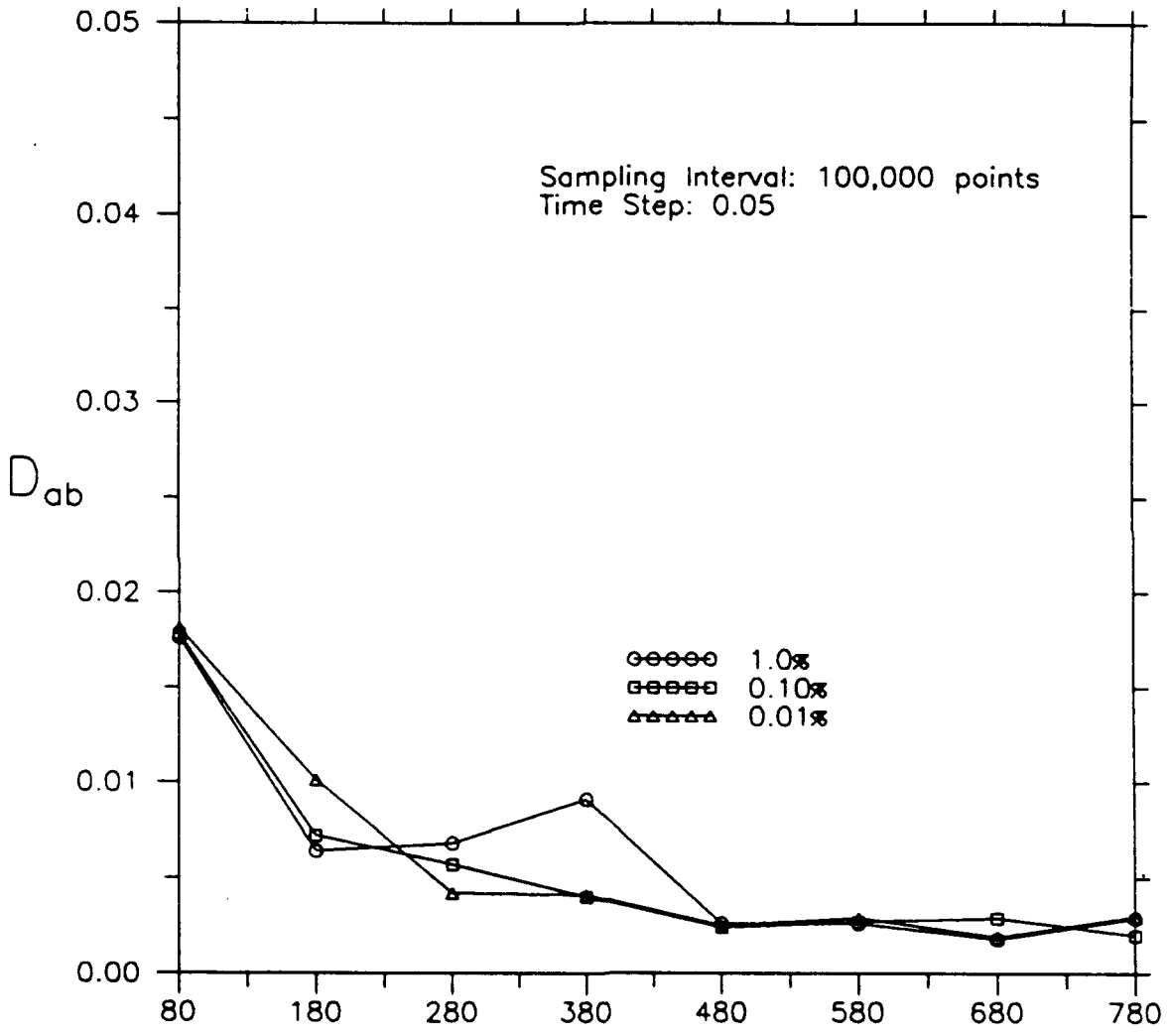


Figure 4.29: The asymptotic mean absolute difference values D_{ab} for each of the initial condition cases as a function of increasing series length comparison (in thousands) for a sampling interval of 100,000 points. As with the other two sample sizes, all three sets of difference values again decrease to similar floor values.

Table 4.3: The asymptotic mean absolute difference values D_{ab} for all three initial conditions and a sampling interval of 25,000 points as a function of increasing histogram series length comparison (H_a and H_b). The average D_{ab} value ($\overline{D_{ab}}$) has been computed for each initial condition over a range of series intervals in which the difference values fluctuate about a minimum range; these values represent the average minimum intrinsic difference between histograms of increasing series length for this particular sample size. The series length comparison interval displayed in bold type represents that after which no additional information can be gained (280,000 points).

Histogram Series Length		**D_{ab}**		
H_a	**H_b**	**1.0%**	**0.10%**	**0.01%**
30,000	55,000	0.0178	0.0188	0.0198
55,000	80,000	0.0146	0.0095	0.0141
80,000	105,000	0.0103	0.0092	0.0107
105,000	130,000	0.0105	0.0068	0.0052
130,000	155,000	0.0066	0.0060	0.0062
155,000	180,000	0.0050	0.0065	0.0055
180,000	205,000	0.0048	0.0036	0.0055
205,000	230,000	0.0044	0.0033	0.0037
230,000	255,000	0.0034	0.0028	0.0034
255,000	**280,000**	**0.0022**	**0.0025**	**0.0023**
280,000	305,000	0.0025	0.0021	0.0021
305,000	330,000	0.0026	0.0026	0.0023
330,000	355,000	0.0025	0.0031	0.0027
355,000	380,000	0.0027	0.0024	0.0019
380,000	405,000	0.0026	0.0021	0.0022
$\overline{D_{ab}}$ in range (255,000-405,000):		0.0025	0.0025	0.0023

Table 4.4: The asymptotic mean absolute difference values D_{ab} for all three initial conditions and a sampling interval of 100,000 points as a function of increasing histogram series length comparison (H_a and H_b). The average D_{ab} value ($\overline{D_{ab}}$) has been computed for each initial condition over a range of series intervals in which the difference values fluctuate about a minimum range; these values represent the average minimum intrinsic difference between histograms of increasing series length for this particular sample size. The series length comparison interval displayed in bold type represents that after which no additional information can be gained (580,000 points).

Histogram Series Length		**D_{ab}**		
H_a	**H_b**	**1.0%**	**0.10%**	**0.01%**
80,000	180,000	0.0176	0.0178	0.0181
180,000	280,000	0.0064	0.0072	0.0101
280,000	380,000	0.0068	0.0057	0.0042
380,000	480,000	0.0091	0.0040	0.0041
480,000	**580,000**	**0.0025**	**0.0024**	**0.0025**
580,000	680,000	0.0026	0.0027	0.0029
680,000	780,000	0.0018	0.0029	0.0019
780,000	880,000	0.0029	0.0020	0.0030
880,000	980,000	0.0025	0.0025	0.0027
$\overline{D_{ab}}$ in range (480,000-980,000):		0.0025	0.0025	0.0026

observe that the values of $\overline{D_{ab}}$ range between 0.0023 and 0.0025 for the 25,000-point samples (Table 4.3) and between 0.0024 and 0.0026 for the 100,000-point samples (Table 4.4). This similarity is a fascinating discovery and leads us to conclude, given these limited results, that the minimum difference values $\overline{D_{ab}}$ between histogram structures appear to be relatively *independent* of both the initial condition value and the sampling interval that we use.

The existence of a nonzero minimum value in the histogram difference suggests that there is some small variability intrinsic to the analysis. We wonder, however, if this value is indeed intrinsic to the Lorenz model solutions or if it is related directly to the way in which we have defined the measures that we use. The value of this minimum difference is most likely a function of many factors. First, it may be a function of the time step chosen; we have already verified qualitatively that the minimum series length needed for convergence is highly dependent upon this value. Second, it is most likely a function of the bin width b_w because changing the bin width changes the resolution at which the attractor leaves are being separated; we expect that the difference values would increase as b_w decreases. Third, the difference value may be a result of the way in which we defined our expressions for D_{avg} and D_{ab}; their average minimum values

differ by approximately a factor of three. Finally, the minimum difference values that we have obtained could be the result of the fact that the ideal histogram is simply an approximation of the derivative of a nondifferentiable function. Thus, the limit of the histograms as the bin width $b_w \to 0$ and as the series length $L_{ser} \to \infty$ is strictly a generalized function. Employing a true and continuous function such as the area under the histogram curve may yield eventually a zero difference floor. We elaborate on these possibilities further in the conclusion.

The most remarkable difference that we observe when we compare the three sample size cases in Figures 4.27-4.29 is the data interval at which we feel confident in defining histogram convergence. We recall in the 50,000-point case that we achieved a reasonable convergence once the series had approximately 480,000 points. To determine the data set representing the minimum for histogram convergence, we employ the same method as that above. Using the largest of the three $\overline{D_{ab}}$ values ($\overline{D_{ab}}$ = 0.0025) in Table 4.3, we observe that adding data in 25,000-point increments, as shown in Figure 4.28, yields convergence at a value representing the last 280,000 points of a 300,000-point series for all three initial conditions. Using the same procedure for the 100,000-point increment cases seen in Figure 4.29, we observe that convergence occurs when we have obtained the last 580,000 points of a 600,000-point series (Table

4.4); this value is somewhat larger than that found when we compared the three initial conditions in Figure 4.23. These differences seem counterintuitive, we might expect, using round-off error arguments, that if differences do occur, then the 100,000-point case would show a faster convergence to some minimum difference value than would the 25,000-point case. What does seem clear, however, is that there is a *dependence* of he minimum data set necessary to produce histogram convergence on the sampling interval that we use, at least for intervals less than or equal to 100,000 points. These results indicate the variabilities that are produced using a "snapshot" or "stroboscopic" view to capture the attractor data. The causes of these variabilities are unclear. We theorize that there may exist attractor fluctuations or smearing of its approximation owing to roundoff error. Based on these results, we conclude that the data that we have produced with this model are extremely sensitive to the sample length that we use to quantify histogram convergence. Figure 4.30 highlights the distinct similarities and differences by showing a superposition of the results from the three sampling intervals in Figures 4.27-4.29 for the 0.01 percent initial condtion.

These results are exciting, yet are conflicting and confusing in numerous ways that we have already briefly mentioned. What does seem certain and is most important to us from all the results that we have obtained thus far in this chapter is that, despite

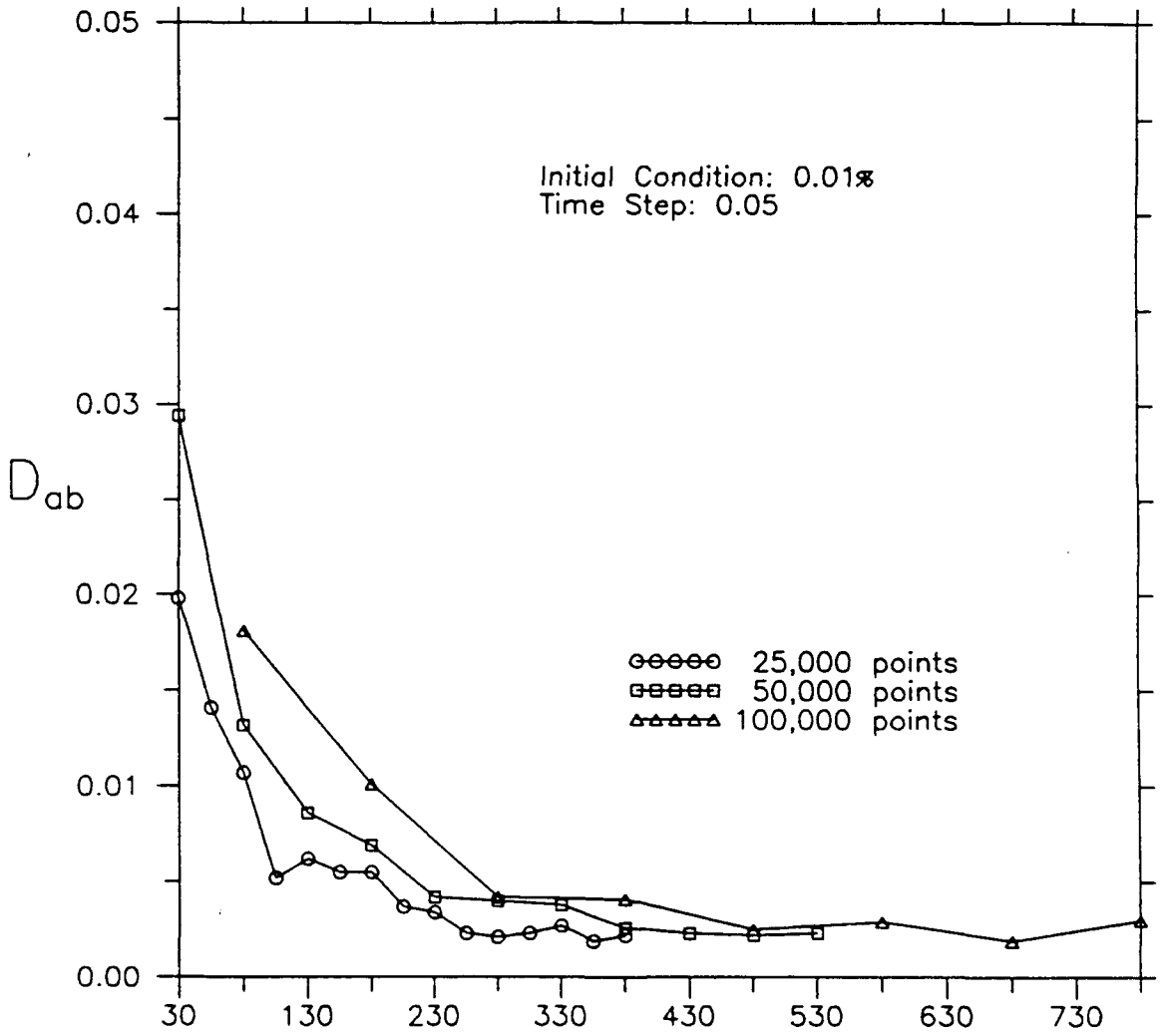


Figure 4.30: The asymptotic mean absolute difference values D_{ab} for the 0.01% initial condition case as a function of increasing series length comparison (in thousands) for all three sampling intervals. Note their decrease to a similar floor value despite the change in sample size.

the dependence of the rate of convergence on the series increment used, we now have a quantitative estimate for the minimum series length that we need to define histogram convergence. For this Lorenz model, we are confident that we have a control histogram if we use at least 580,000 points to define it, as long as we choose a time step value of 0.05. This number of points assures us of a reasonable histogram convergence whether we use a single initial condition or an average of three initial conditions. Applied to atmospheric models, these results tell us that both single and multiple initial condition integration methods that are presently used to estimate predictability have merit.

Although we have achieved the primary goal of finding the minimum series length necessary for histogram convergence, there are other issues related to this convergence that require further quantitative study. These issues not only justify the extreme sensitivities related to the way that we sample the model data, but also help us quantify predictability estimates. These results further describe the attractor in the low-order Lorenz model and provide useful information that has potential applications to larger, more complicated time series.

4.3.3. Finding Convergence as a Function of Time

Throughout this chapter, we have quantified convergent behavior in the histograms by calculating their difference values using data sets generated with a time step of 0.05. Now we focus on the behavior of the histogram convergence as a function of the time step value and the total time of record. By *quantifying* these behaviors, we seek to determine whether using a larger time step is indeed most beneficial for finding convergence as well as to express convergence as a function of time, not series length.

4.3.3.1. Differences Based on Time Step Value

We begin this section by comparing the average mean absolute differences D_{avg} (4.4) for the two time step values: 0.05 and 0.005. In Figure 4.31, we present those results. Whereas the larger time step curve shows a smooth, exponential-type decrease to a minimum value by 480,000 points, the smaller time step curve shows a relatively rugged, generally decreasing behavior with no discernible minimum, even by 980,000 points. This quantitative result agrees with the nonconvergence in histograms that we

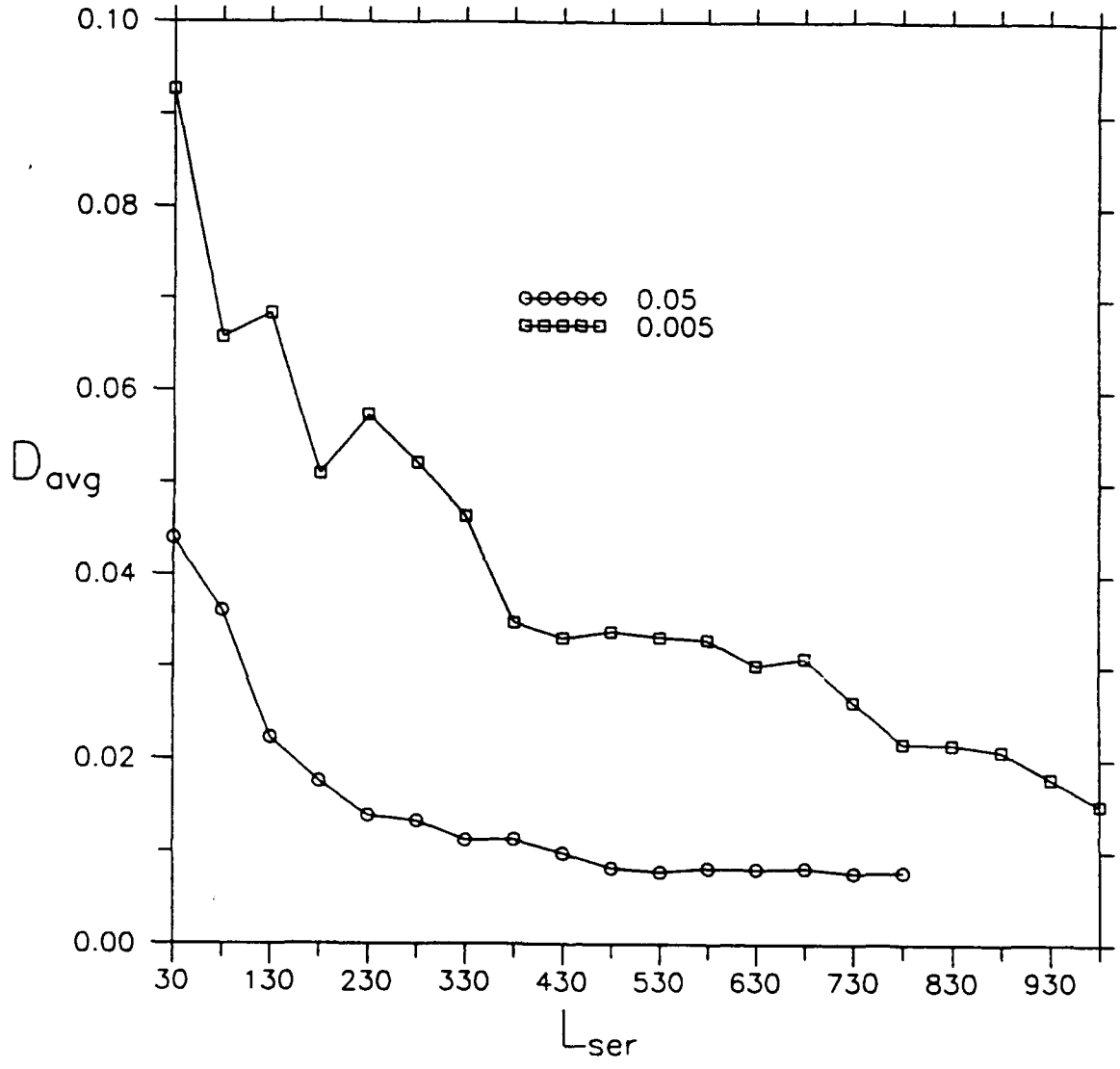


Figure 4.31: The average mean absolute difference values D_{avg} as a function of series length L_{ser} (in thousands) for the two time step values. Note that convergence to a definitive floor is not apparent in the smaller time step (0.005) curve.

saw earlier in Figure 4.1. Both sets of difference values are provided in Table 4.5. Because the smaller time step values continue to decrease, we believe that convergence as we have defined it does exist, but with a much greater number of points. In the next subsection, we seek this number. For now, suffice it to say that if reasonable convergence does exist with a data series produced with a time step of 0.005, then we believe that the required series length is quite large. If this length proves consistent with previous hypotheses, then we expect that this value would be approximately 4,800,000 points.

There is another observation that warrants further explanation. Comparing the two sets of data in Table 4.5 more closely, we observe significantly large differences in D_{avg} values for each series interval. Upon comparing these values over all the intervals, we conclude that when using the larger time step, we obtain the same degree of accuracy with only 15 to 25 percent of the points than we would need if using the smaller time step. This observation provides further evidence of a distinct undersampling problem when using too fine a temporal resolution and reinforces the conclusion that we made earlier concerning the benefits of sampling the model data at a

Table 4.5: The average mean absolute difference values D_{avg} as a function of the series length L_{ser} for both values of time step (0.05 and 0.005). Note that, unlike the distinct convergence to a difference floor value that is exhibited when using the larger time step, there is no corresponding convergence of the data to a minimum value when using the smaller time step (out to 980,000 points).

L_{ser}	D_{avg}	
	0.05	**0.005**
30,000	0.0440	0.0927
80,000	0.0361	0.0657
130,000	0.0223	0.0683
180,000	0.0177	0.0509
230,000	0.0139	0.0574
280,000	0.0132	0.0522
330,000	0.0113	0.0464
380,000	0.0112	0.0348
430,000	0.0098	0.0331
480,000	0.0081	0.0338
530,000	0.0078	0.0332
580,000	0.0082	0.0329
630,000	0.0081	0.0300
680,000	0.0083	0.0309
730,000	0.0077	0.0262
780,000	0.0079	0.0217
830,000		0.0216
880,000		0.0209
930,000		0.0180
980,000		0.0182

coarser resolution.

We now make the same comparative calculations using D_{ab} (4.5) for each of the three initial conditions. In this section, since all three initial condtions show relatively similar behavior, for brevity, we show only the results for the 0.01 percent initial condition case at a sampling interval of 50,000 points. In the next subsection, we comment further on the other two cases and the importance of including those results. The difference values for this case are plotted in Figure 4.32 and given in Table 4.6. Again, significant differences between the two time step curves exist and seem somewhat consistent with previous results; the smaller time step yields larger difference values and displays a much rougher behavior than that seen with the larger time step. However, the separation between the two curves is much less dramatic than that seen in the previous case (Figure 4.31). What first strikes us is the apparent convergence displayed by the smaller time step data that was not seen before. Using the same method that we employed in the previous section to find $\overline{D_{ab}}$, we obtain a value of $\overline{D_{ab}}$= 0.0027 for the smaller time step (t_S = 0.005). As a result, we are confident of histogram convergence by 680,000 points at a $\overline{D_{ab}}$ value that closely corresponds to that of the larger time step ($\overline{D_{ab}}$ = 0.0023). As above, we still find histogram convergence

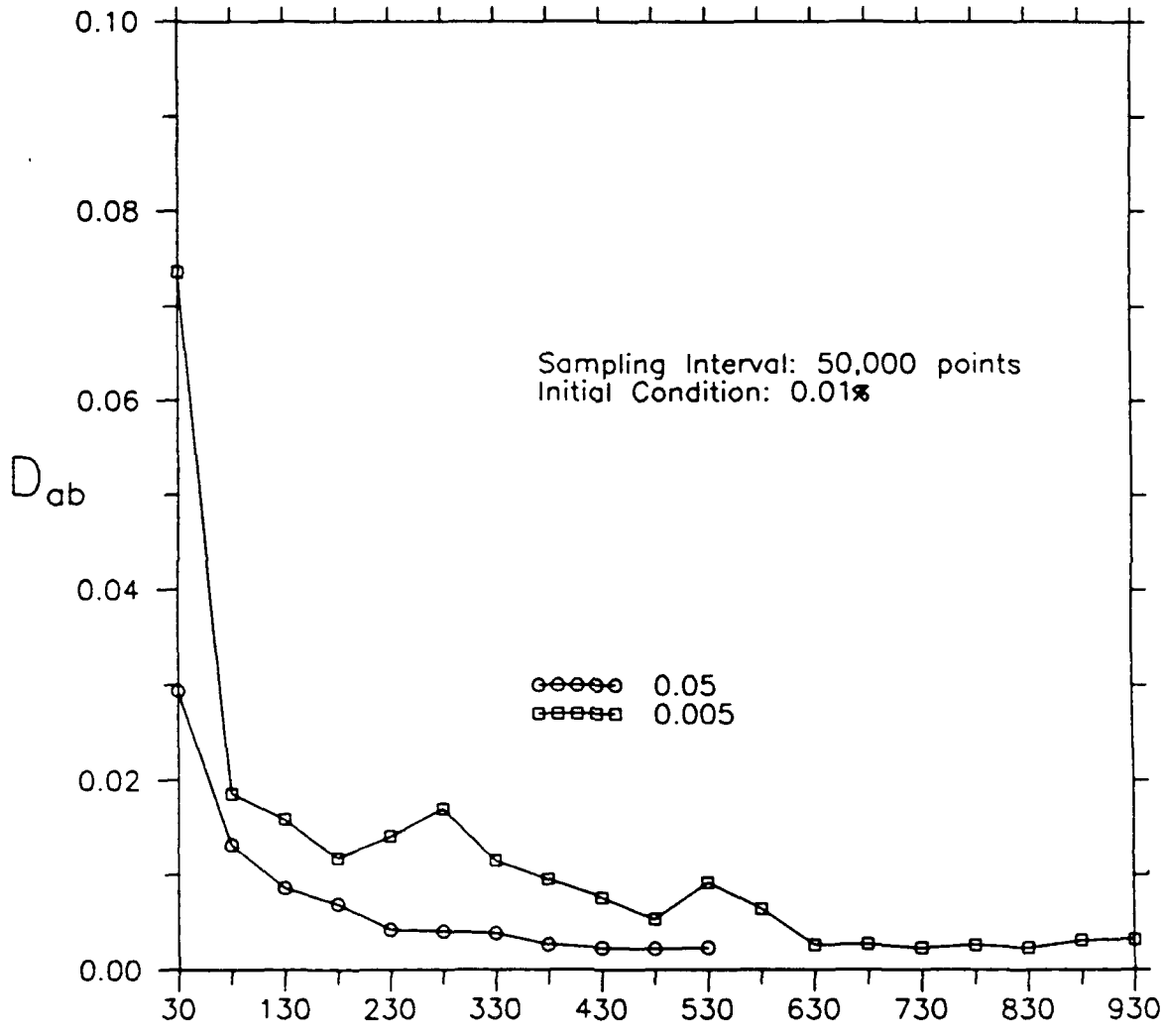


Figure 4.32: The asymptotic mean absolute difference values D_{ab} as a function of increasing series length comparison (in thousands) for the 0.01% initial condition at a sampling interval of 50,000 points. Unlike that seen for D_{avg}, the D_{ab} values generated with the smaller time step (0.005) do exhibit convergence to a floor.

Table 4.6: The asymptotic mean absolute difference values D_{ab} as a function of increasing histogram series length comparison (H_a and H_b) for both values of time step (0.05 and 0.005) and the 0.01% initial condition. Note that, unlike the nonconvergent behavior in D_{avg} values exhibited when using the smaller time step, their D_{ab} values do satisfy the criteria for convergence, although at a larger series interval than that exhibited when using the larger time step value (680,000 vs. 480,000).

Histogram Series Length		**D_{ab}**	
H_a	**H_b**	**0.05**	**0.005**
30,000	80,000	0.0294	0.0736
80,000	130,000	0.0132	0.0186
130,000	180,000	0.0086	0.0159
180,000	230,000	0.0069	0.0118
230,000	280,000	0.0042	0.0141
280,000	330,000	0.0040	0.0170
330,000	380,000	0.0038	0.0115
380,000	430,000	0.0026	0.0095
430,000	480,000	0.0023	0.0076
480,000	530,000	0.0022	0.0053
530,000	580,000	0.0023	0.0092
580,000	630,000		0.0064
630,000	680,000		0.0026
680,000	730,000		0.0027
730,000	780,000		0.0023
780,000	830,000		0.0026
830,000	880,000		0.0023
880,000	930,000		0.0031
930,000	980,000		0.0032
$\overline{D_{ab}}$ in range (430,000-580,000):		0.0023	
$\overline{D_{ab}}$ in range (630,000-980,000):			0.0027

with fewer points (480,000) when using the larger time step than when using the smaller time step (680,000). However, it is troubling that the advantage in using the larger time step in this case appears greatly reduced. This limited result even suggests, at least for the smaller time step, that using a single initial condition run to identify convergence might be superior to using comparisons of multiple initial conditions--results inconsistent with those obtained when using the larger time step. Given these inconsistencies, it seems that it is more important to quantify histogram convergence as a function of *total time of record.*

4.3.3.2. Differences Based on Total Time

To provide an improved basis for comparison of the convergent histogram behaviors for the two time step values, we compute the total elapsed dimensionless time t_{tot}. To calculate this value, we simply use the time step value t_s and the length of series L_{ser} via

$$t_{tot} = t_s * L_{ser} \quad . \tag{4.6}$$

Figure 4.33 displays the average difference values D_{avg} between the three initial conditions as functions of total time t_{tot} for both time steps.

Comparing these two curves now in terms of total time, we notice a dramatically different relationship between them than we obtained when viewing them as a function of series length. Both sets of difference values are given in Table 4.7. Consistent with Figure 4.33 and (4.6), the larger time step curve exhibits convergence at t_{tot} = 24,000. However, unlike the nonconvergence that we found with the smaller time step value in Figure 4.31, we now find a reasonable convergence to that value by t_{tot} = 19,900. This t_{tot} value corresponds to a series length L_{ser} of 3,800,000 points and represents a value approximately 20 percent smaller than the value that we previously thought necessary to produce histogram convergence. An even more intriguing behavior is that, although using the smaller time step yields a much higher difference value initially, we see that its decrease is much more rapid than that for the larger time step--so much so that for values of $t_{tot} \geq$ 1,900, the difference values for the smaller time step are less that those for the larger time step, indicating that in fact the 0.005 data set is more accurately representing the attractor than is the 0.05 data set.

Using the same procedures with the 0.01 percent initial condition curves seen in Figure 4.32, we produce plots in Figure 4.34 of the difference values D_{ab} as functions

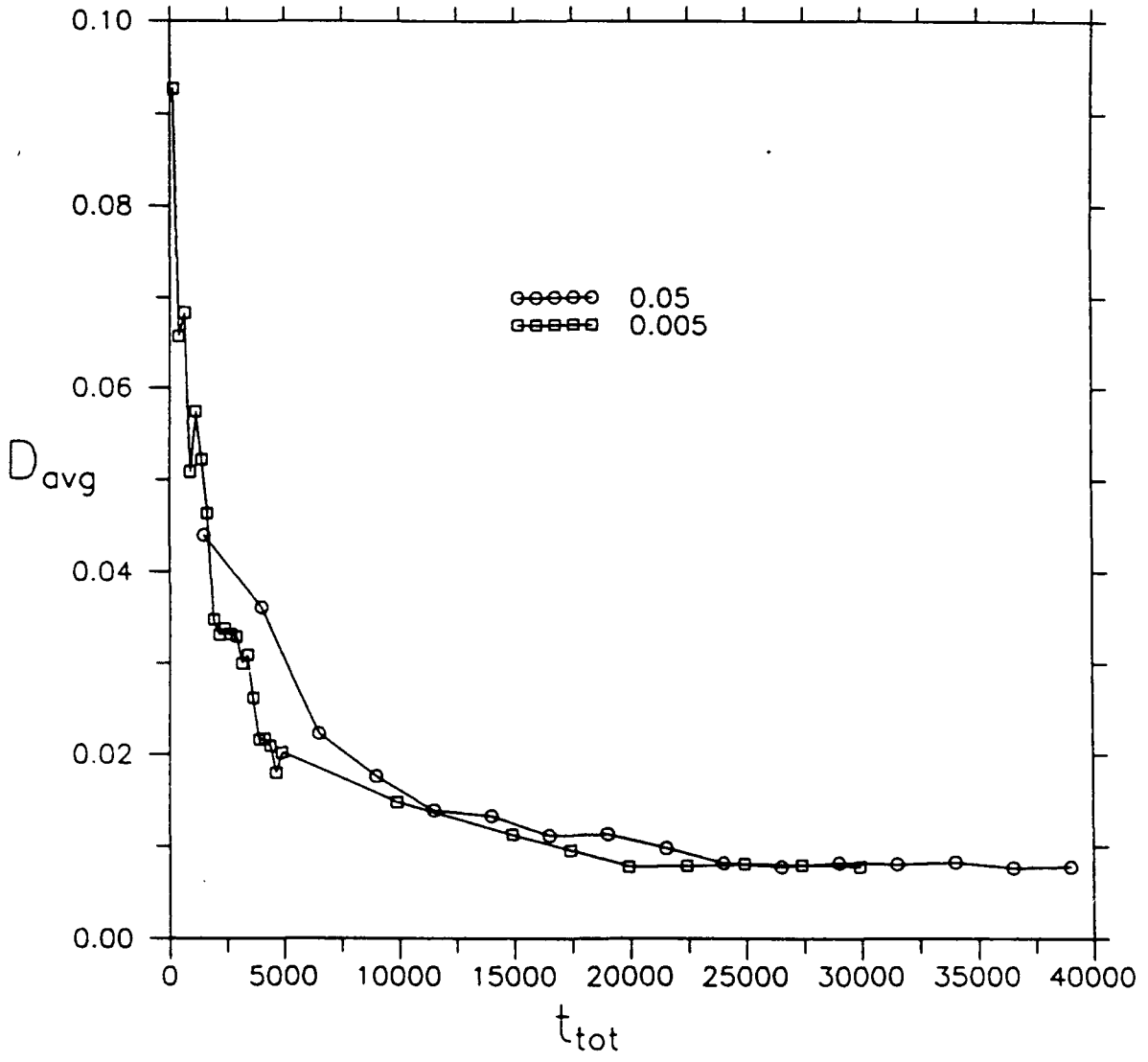


Figure 4.33: The average mean absolute difference values D_{avg} as a function of total time t_{tot} for the two time step values. In terms of total time, the smaller time step (0.005) curve exhibits convergence to the floor sooner (19,900) than does that of the larger time step (24,000).

Table 4.7: The average mean absolute difference values D_{avg} as a function of the total time of record t_{tot} for both values of time step (0.05 and 0.005). When viewed in terms of total time instead of series length, the difference values generated when using the smaller time step display convergence to a minimum value faster than does using the larger time step (19,900 vs. 24,000).

0.05		0.005	
t_{tot}	D_{avg}	t_{tot}	D_{avg}
1,500	0.0440	150	0.0927
4,000	0.0361	400	0.0657
6,500	0.0223	650	0.0683
9,000	0.0177	900	0.0509
11,500	0.0139	1,150	0.0574
14,000	0.0132	1,400	0.0522
16,500	0.0113	1,650	0.0464
19,000	0.0112	1,900	0.0348
21,500	0.0098	2,150	0.0331
24,000	0.0081	2,400	0.0338
26,500	0.0078	2,650	0.0332
29,000	0.0082	2,900	0.0329
31,500	0.0081	3,150	0.0300
34,000	0.0083	3,400	0.0309
36,500	0.0077	3,650	0.0262
39,000	0.0079	3,900	0.0217
		4,150	0.0216
		4,400	0.0209
		4,650	0.0180
		4,900	0.0182
		9,900	0.0148
		14,900	0.0113
		17,400	0.0096
		19,900	0.0079
		22,400	0.0079
		24,900	0.0081
		27,400	0.0081

of increasing total time comparison between histograms H_a and H_b; these values are given in Table 4.8. Here, we observe an even greater disparity in convergent behavior between the two time step curves than we saw in Figure 4.33. First, the smaller time step curve shows an even greater rate of decrease in difference value; this value becomes less than that of the larger time step by roughly a time value of 500. Second, the smaller time step curve exhibits a pronounced convergence to an average minimum difference value $\overline{D_{ab}} = 0.0026$ at a time of 3,400. Because a time of 24,000 in the 0.05 curve is required to obtain its value of $\overline{D_{ab}}$, we conclude that histogram convergence occurs at a rate nearly *seven times faster* when using the smaller time step than that obtained when using the larger time step.

This large disparity between the two time step values in the total times necessary to obtain convergence profoundly contradicts the results that we obtained when viewing convergence in terms of series length. However, this result is based on the behavior of only one of the three initial conditions. To ensure that we are not basing conclusions on a potentially anamolous case, we produce Figure 4.35 that shows a plot of D_{ab} as a function of increasing total time comparison for all three initial conditions; the values are given in Table 4.9. Using the same method that we described earlier to

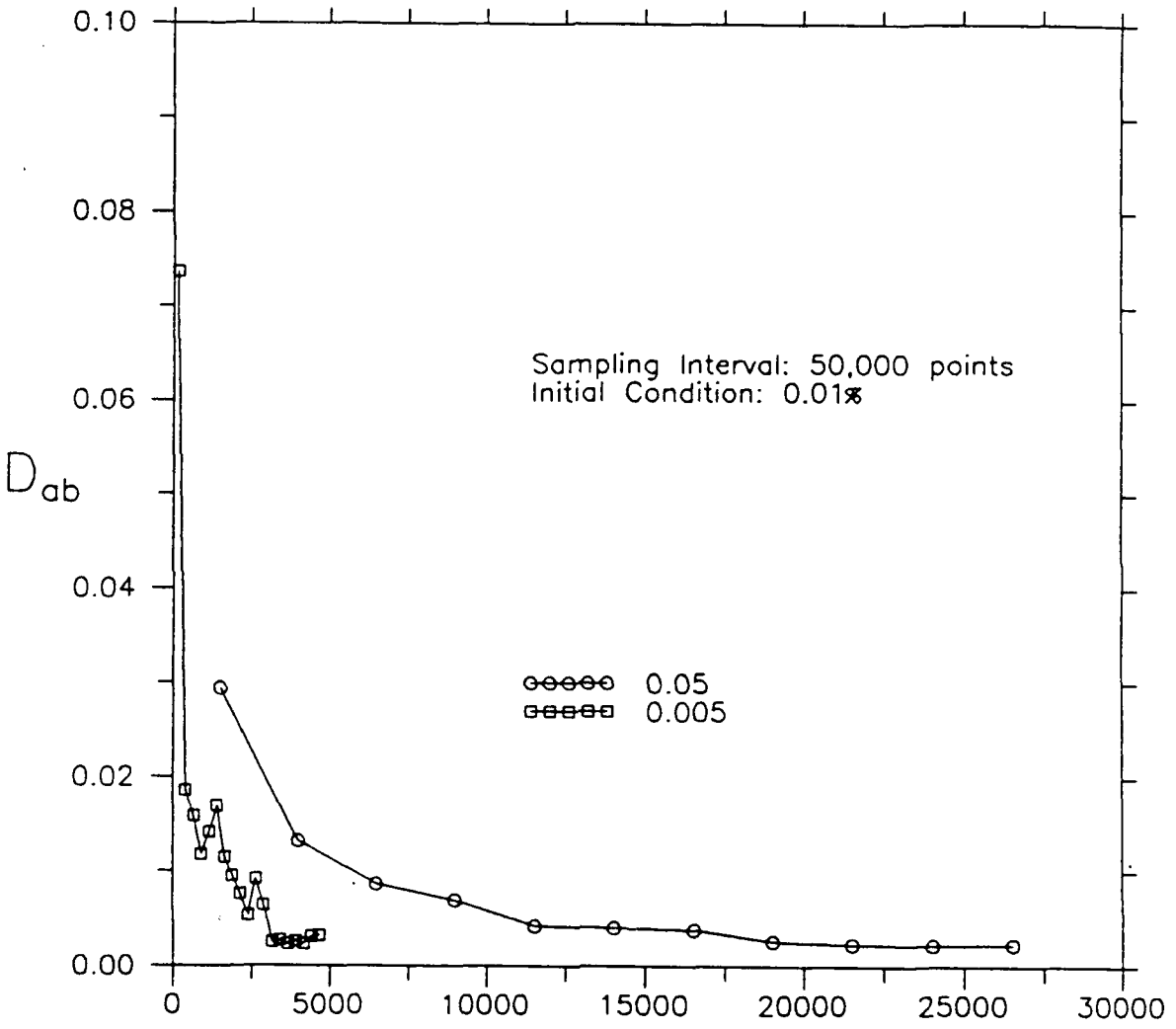


Figure 4.34: The asymptotic mean absolute difference values D_{ab} as a function of increasing time comparison for the two time step values for the 0.01% initial condition. Note that the smaller time step curve exhibits convergence to the floor approximately seven times faster (3,400) than does that of the larger time step curve (24,000).

Table 4.8: The asymptotic mean absolute difference values D_{ab} as a function of increasing histogram time comparison for both values of time step (0.05 and 0.005) and the 0.01% initial condition. For this case, the difference values generated with the smaller time step exhibit convergence to a minimum nearly seven times faster than that produced with the larger time step (3,400 vs. 24,000).

0.05			**0.005**		
Histogram Total Time		**D_{ab}**	**Histogram Total Time**		**D_{ab}**
H_a	**H_b**		**H_a**	**H_b**	
1,500	4,000	0.0294	150	400	0.0736
4,000	6,500	0.0132	400	650	0.0186
6,500	9,000	0.0086	650	900	0.0159
9,000	11,500	0.0069	900	1,150	0.0118
11,500	14,000	0.0042	1,150	1,400	0.0141
14,000	16,500	0.0040	1,400	1,650	0.0170
16,500	19,000	0.0038	1,650	1,900	0.0115
19,000	21,500	0.0026	1,900	2,150	0.0095
21,500	24,000	0.0023	2,150	2,400	0.0076
24,000	26,500	0.0022	2,400	2,650	0.0053
26,500	29,000	0.0023	2,650	2,900	0.0092
			2,900	3,150	0.0064
			3,150	3,400	0.0026
			3,400	3,650	0.0023
			3,650	3,900	0.0026
			3,900	4,150	0.0023
			4,150	4,400	0.0021
			4,400	4,650	0.0031
			4,650	4,900	0.0032
$\overline{D_{ab}}$ in range (19,000-26,500):		0.0023			
$\overline{D_{ab}}$ in range (3,150-4,900):					0.0027

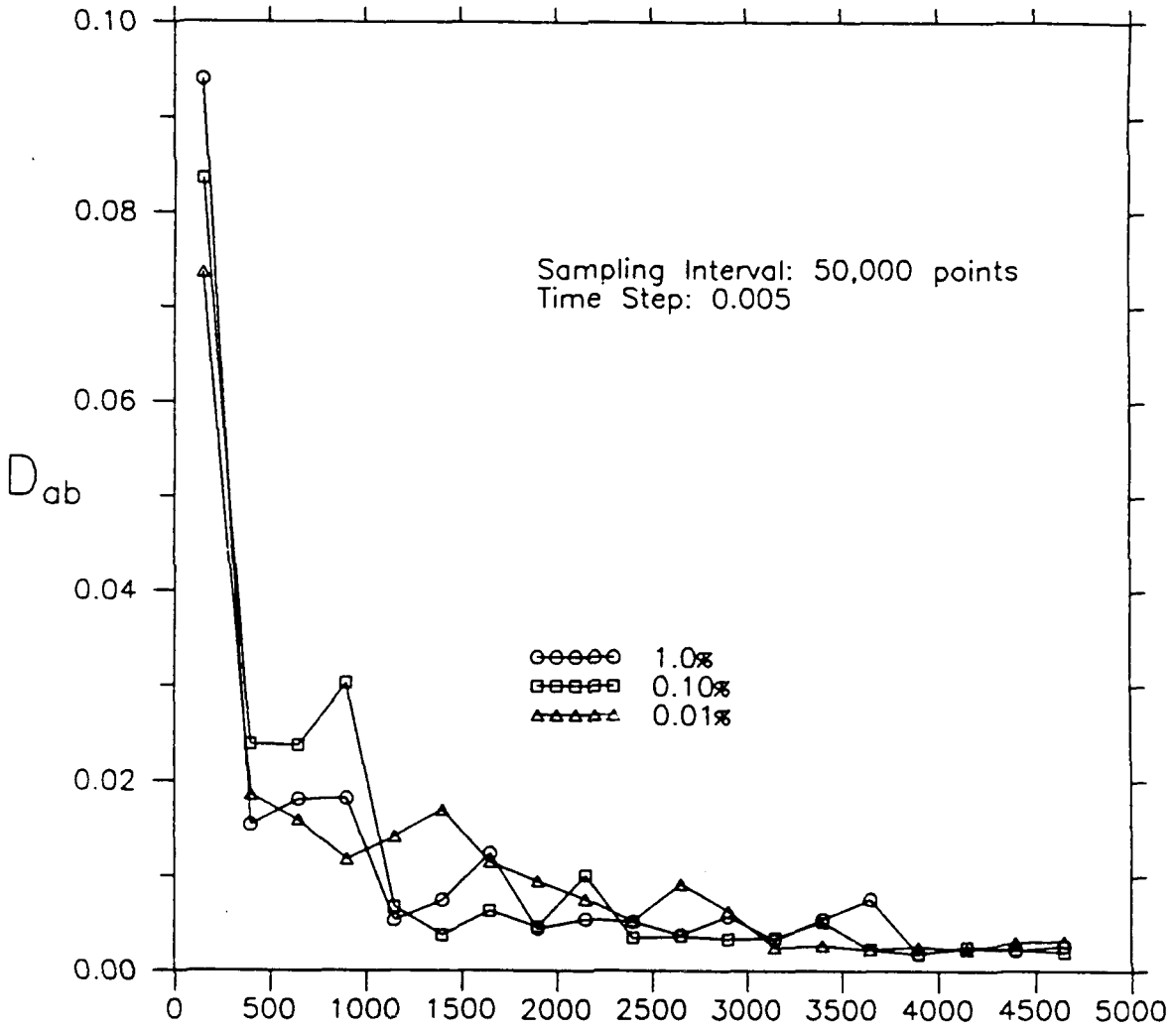


Figure 4.35: The asymptotic mean absolute difference values D_{ab} as a function of increasing time comparison for the smaller time step curves of all three initial conditions. All three curves exhibit convergence to the floor at nearly the same elapsed time.

Table 4.9: The asymptotic mean absolute difference values D_{ab} as a function of increasing histogram time comparison for all three initial conditions and generated with the smaller time step value (0.005). Within all three sets of initial conditions, convergence to their minimum difference values range from a total time of 3,400 for the 0.01% case to a time of 4,150 for the 1.0% and 0.10% cases.

Histogram Total Time		**D_{ab}**		
H_a	**H_b**	**1.0%**	**0.10%**	**0.01%**
150	400	0.0942	0.0836	0.0736
400	650	0.0154	0.0239	0.0186
650	900	0.0181	0.0238	0.0159
900	1,150	0.0182	0.0303	0.0118
1,150	1,400	0.0054	0.0068	0.0141
1,400	1,650	0.0075	0.0038	0.0170
1,650	1,900	0.0124	0.0064	0.0115
1,900	2,150	0.0045	0.0047	0.0095
2,150	2,400	0.0055	0.0101	0.0076
2,400	2,650	0.0052	0.0035	0.0053
2,650	2,900	0.0038	0.0037	0.0092
2.900	3,150	0.0058	0.0034	0.0064
3,150	3,400	0.0033	0.0035	0.0026
3,400	3,650	0.0055	0.0052	0.0027
3,650	3,900	0.0076	0.0026	0.0023
3,900	4,150	0.0019	0.0018	0.0026
4,150	4,400	0.0026	0.0026	0.0023
4,400	4,650	0.0023	0.0023	0.0031
4,650	4,900	0.0027	0.0022	0.0032
$\overline{D_{ab}}$ in range (3,900-4,900):		0.0024	0.0023	
$\overline{D_{ab}}$ in range (3,150-4,900):				0.0027

find the value of $\overline{D_{ab}}$, we find $\overline{D_{ab}}$ = 0.0024 and 0.0022 for the 1.0 and 0.10 percent cases, respectively; the time values for both cases equal 4,150, a value quite close to that for the 0.01 percent case. Assuming that this value (4,150) is a conservative estimate for the dimensioness total time needed to find convergence in all three initial conditions, we still find that convergence occurs *five times faster* using the smaller time step value than that for the larger one. Thus, we now feel quite confident that a large disparity in the total times necessary for obtaining convergent series does exist and is a function of the time step that we use.

What is clear, based on these results, is that the expectation of the benefits of choosing a larger time step over that of a smaller value to produce optimum convergence in the histograms as theorized in Section 4.2 was not necessarily correct. It seems now that choosing an optimum time step for sampling the attractor depends upon the limitations under which we are placed. More specifically, when choosing a time step value, there is a crucial tradeoff between the degree of accuracy that we require and the number of points that we desire to use. If we want to find histogram convergence to a high degree of accuracy, despite the large number of data points, then we would choose a smaller time step; if we want to limit the length of the series, at the expense of accuracy, then we would choose a larger time step. Using either of these two

time steps eventually yields histogram convergence to within a small variability that is intrinsic to the model.

From this argument, it seems that the optimum way to represent the attractor is to use the smaller time step value and then to sample the model-generated series using a certain frequency, say every tenth point. In doing so, we obtain the best of both characteristics, achieving a high degree of accuracy even though using fewer points. This conjecture, at least for the Lorenz model, certainly lends credibility to the use of sampling strategies in order to find optimum subsets of the adequate data samples. We expand on these ideas in the conclusion.

4.3.4. Relation to Minimum Circuits Around the Lorenz Attractor

Early in this chapter, we argued that sampling at a coarser temporal resolution produces a better quantitative representation of the attractor data than does using a finer resolution. This argument led us to conclude that it is more important to sample the entire attractor more frequently than to sample separate leaves more accurately. Since then, we have come to understand that choosing an optimum time step value is a

function of the limitations with which we are most comfortable. In this section, we seek to find the minimum number of circuits around the attractor that are necessary to define the histogram convergence that we have quantified as functions of both values of time step.

We begin with the results that we obtained when using the larger time step (t_s = 0.05). To calculate the minimum number of circuits required about the attractor, we need to know two separate quantities. One is the total dimensionless time t_{tot} that is necessary to produce convergence of the histograms to their minimum difference value. Taking the most conservative estimate of the values that we obtained using the larger time step, we choose a value of 24,000.

The other quantity, the average time t_{attr} required to traverse once around the attractor, involves some calculation. We find this by defining a reference point on the attractor and then by determining the time t(δ) at which the trajectory returns to within some minimum proximity δ of that point. For simplicity, we define this reference point to be on the Z-axis (Z = 25). Using a reference point anywhere else on the attractor would yield difficulties because there are two lobes about which the trajectories travel. Besides, the dynamics of the Lorenz system is strongly tied to the trajectory behavior near the Z-axis that is on the stable manifold (Nese 1987). Upon finding t(δ), we can

calculate this average time via

$$t_{attr} = t(\delta) * t_s \ . \qquad (4.7)$$

Analysis of the time series reveals that $t(\delta)$ is approximately 32; thus, the average time t_{attr} to traverse once around the attractor is approximately 1.6.

Having these two quantities, we can now calculate the minimum number C_{min} of circuits that are required to define convergence for the Lorenz system, given by

$$C_{min} = \frac{t_{tot}}{t_{attr}} \ . \qquad (4.8)$$

Substitution of the above values for t_{tot} and t_{attr} yields a value for C_{min} of 15,000. For a time step value of 0.05, this number represents the minimum number of times that we must sample completely around the attractor in order to be confident of convergence in the histograms.

To find the value of C_{min} for the smaller time step case, we use the same method. Although t_s changes, the total time t_{attr} needed to traverse once around the attractor does not; that value of 1.6 is a relatively constant one. Recalling the large

disparity in convergence times between D_{avg} and D_{ab} in the smaller time step case, we should expect a large disparity in the values of C_{min} as well. Using the time estimate of 4,150 that we obtained for $\overline{D_{ab}}$ (Table 4.9), we use (4.8) to find $C_{min} \sim 2{,}600$. We obtain a more conservative estimate of C_{min} when we use the value of t_{tot} obtained from D_{avg}. With $t_{tot} = 19{,}900$ (Table 4.7), we use (4.8) to find $C_{min} \sim 12{,}400$. As with the results above using the larger time step value, C_{min} represents the minimum number of circuits required to capture histogram convergence.

Besides giving a nice quantitative relation between the Histogram Measure and the structure of the Lorenz attractor itself, we believe that the value of C_{min} has many applications. The most interesting one is a possible link between histogram convergence and the *spatial resolution* given by the time step of the attractor. Although we do not address it further, the value of C_{min} is obviously dependent upon other factors such as changes in forcing value, integration method, and so on.

4.3.5. Rate of Decay of New Information Gain

Earlier in this chapter, we commented briefly on the exponential form of the decrease in the mean absolute difference values to a minimum or floor as the series

lengthens. We believe that by successfully quantifying this predictive behavior for the Lorenz model, we might use these results as stepping stones eventually to quantify the predictability characteristics of larger, more complicated time series, whether model-generated or from observations. In this section, we obtain this quantification.

Because of their monotonic appearance, we look strictly at the behavior exhibited by the larger time step $t_s = 0.05$ curves. To accurately calculate this behavior, we must first subtract the value of the floor at every data point; in effect, we must create a new set of data having a zero floor. Once we have accomplished this, we must determine what type of mathematical law this curve appears to follow. Given the nature of the model with which we are working, we expect that all of the curves will satisfy either of two types of laws: power or exponential.

To determine which law better describes the behavior, we use the new data set that we obtain by subtracting the average difference floor and then determine best fit lines for the result. If the data behave in some power law relation, then the best fit would be given by a log-log display of the data; if we have an exponential relation, then the best fit would be given by a log-linear display. Figure 4.36 shows a log-log view of the new data set that was produced from Figure 4.23, in which we compared the average difference values D_{avg} for all three initial conditions in 50,000-point blocks.

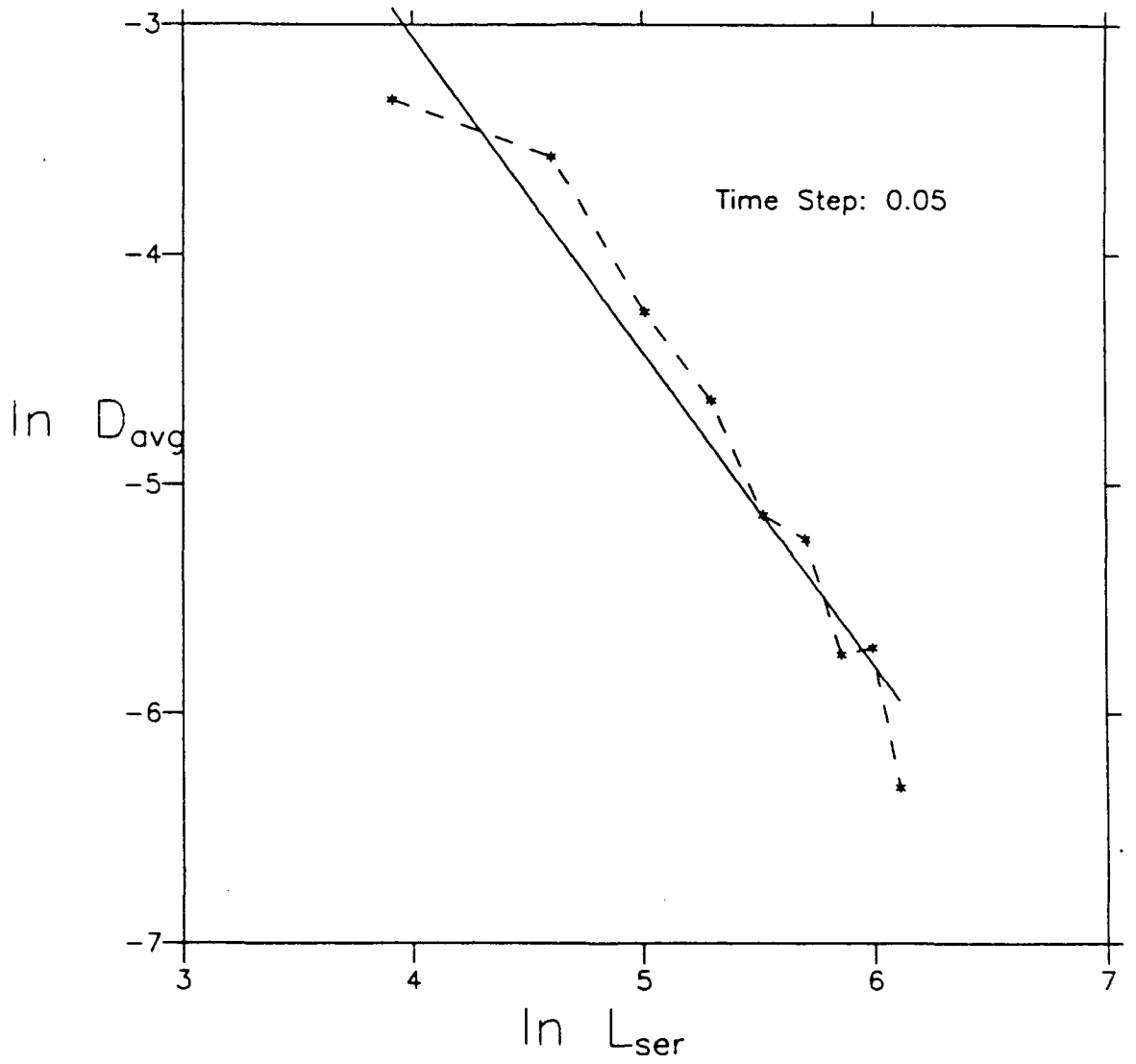


Figure 4.36: The log–log representation of D_{avg} as a function of series length L_{ser}. The best fit line (solid) is not a reasonably good one and thus the curve does not exhibit a power law relation.

Figure 4.37 shows the same data set, only now using a log-linear representation of it. Although this latter fit is not a perfect one, it certainly captures the typical behavior of the difference values much better than does the log-log display. For histogram convergence between the three initial conditions, then, the behavior of the difference values D_{avg} with total series length x is given by the exponential relation

$$D_{avg}(x) = 0.041\, e^{-0.0073x} + 0.008\,, \qquad (4.9)$$

in which x is expressed in thousands.

Armed with this knowledge, we wish to use the same approach to find the relation for each of the initial conditions; we hope to find a similar relation to that in (4.9). Since the behaviors have been shown to be quite similar between each of the three initial conditions, for brevity we address only one. Although the result is not shown, consistent with the D_{avg} data, the power law again simply does not fit the data very well. We recall that the series length at which we were assured of convergence was a function of how we sampled the data. Using the same procedures as above to obtain the new data sets, we produce Figures 4.38-4.40 that exhibit the best exponential fit for the 0.01 percent initial condition data sampled at every 25,000, 50,000, and

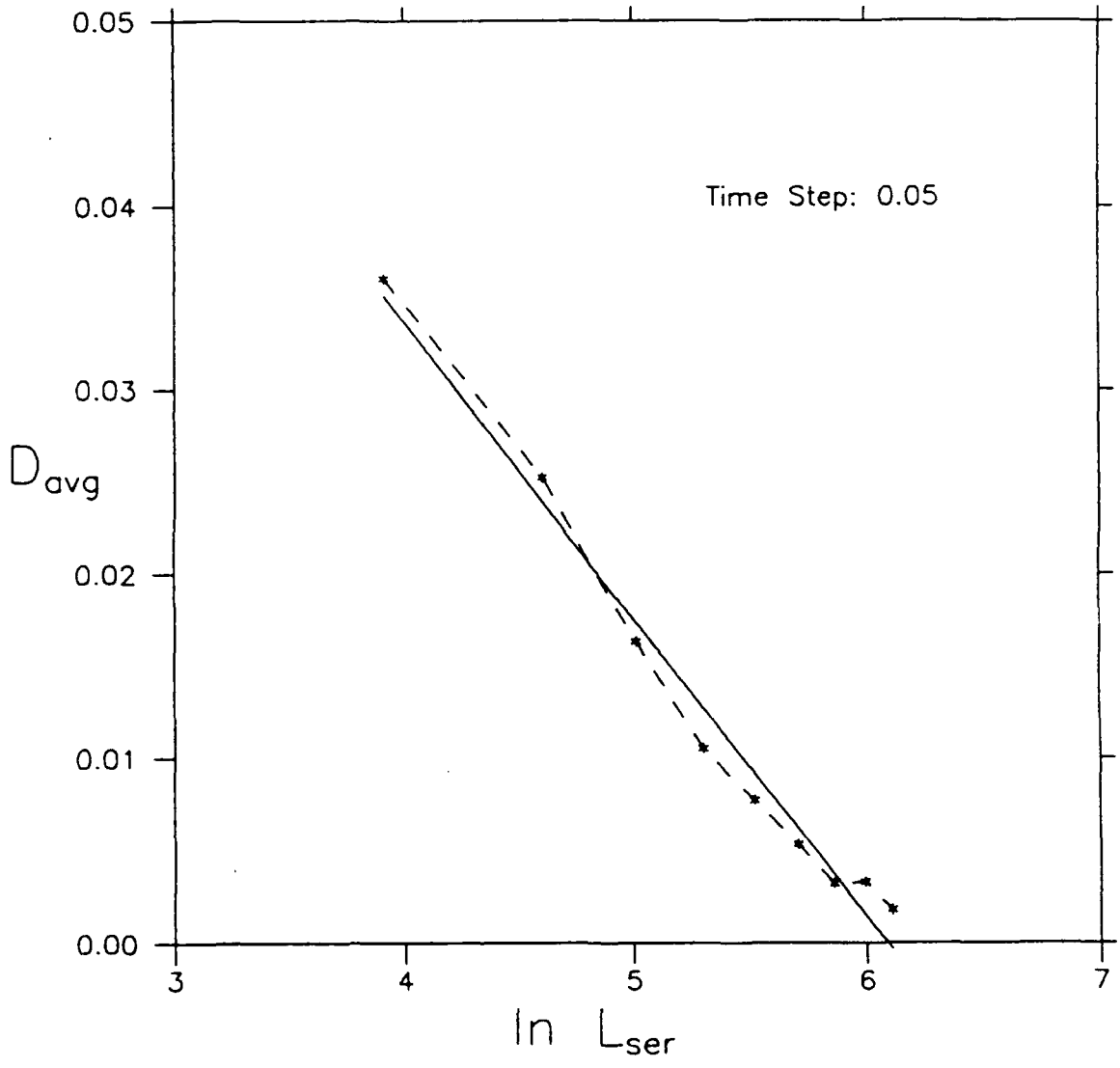


Figure 4.37: The log–linear representation of D_{avg} as a function of series length L_{ser}. The best fit line (solid) in this case is much better than that for the log–log display. Thus, the curve better follows an exponential relation.

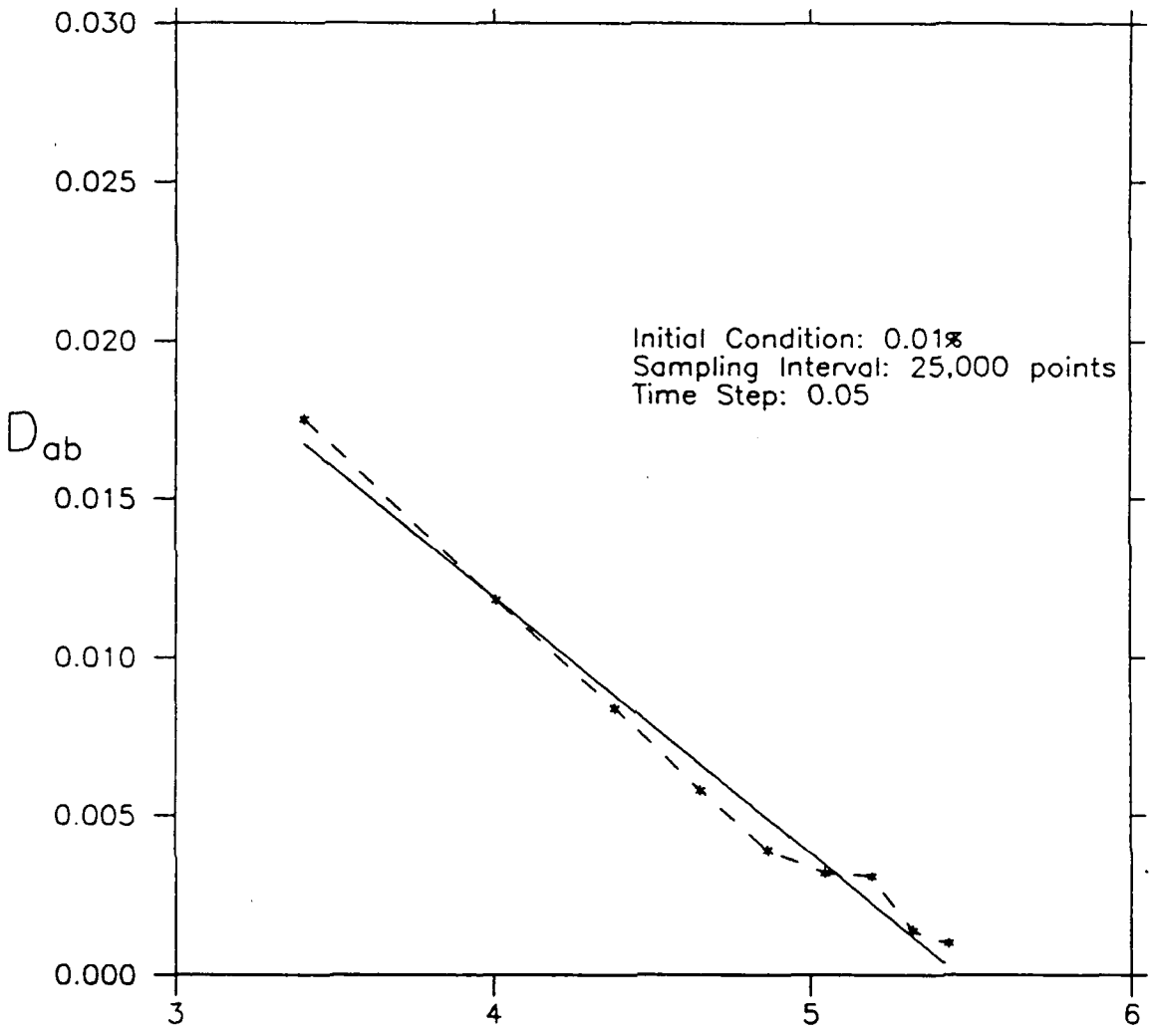


Figure 4.38: The log–linear representation of D_{ab} as a function of increasing series length comparison for the 0.01% initial condition and 25,000–point sampling interval. Consistent with the D_{avg} results, this fit is a reasonably good one.

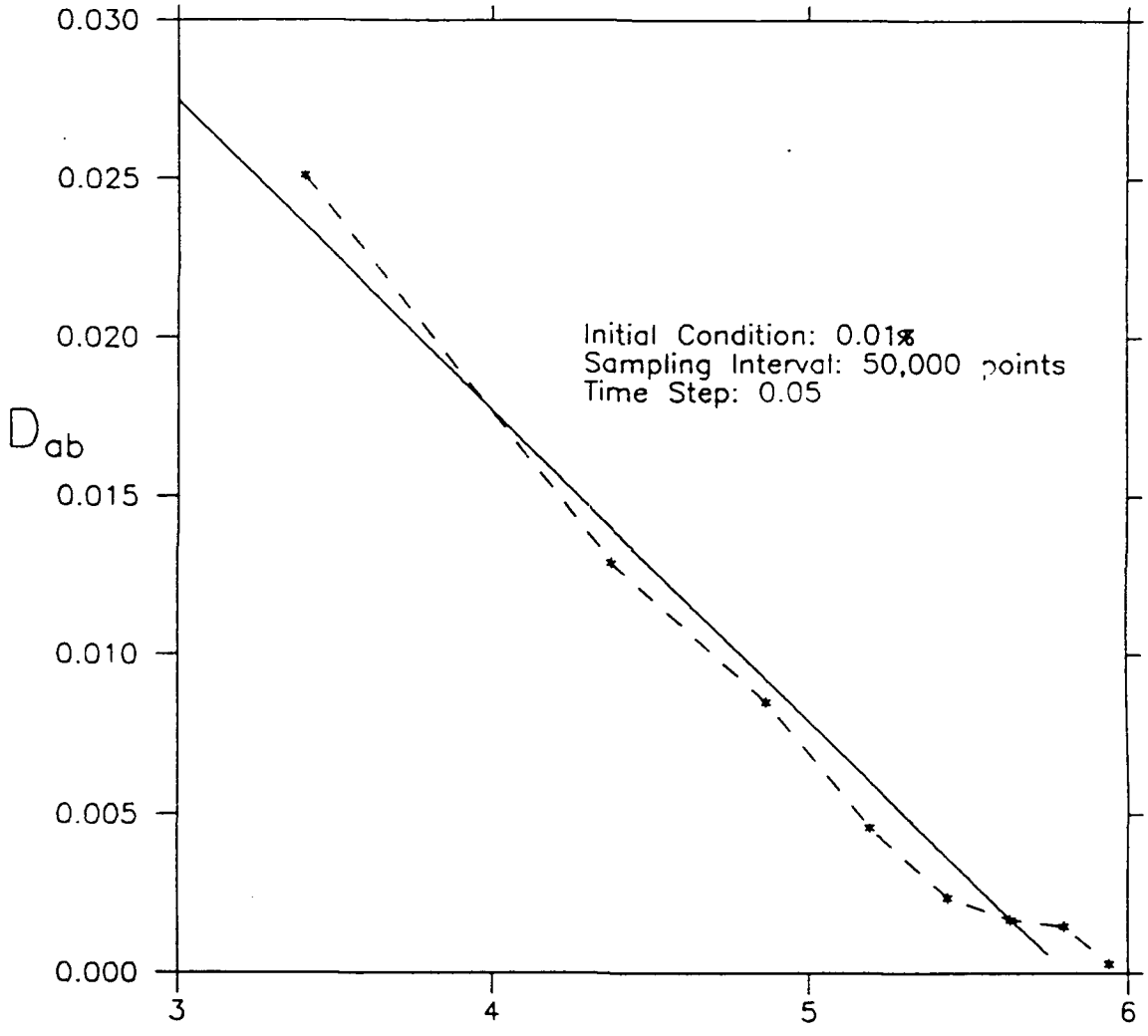


Figure 4.39: The log–linear representation of D_{ab} as a function of increasing series length comparison for the 0.01% initial condition and 50,000–point sampling interval. Note again a relatively good fit to the data.

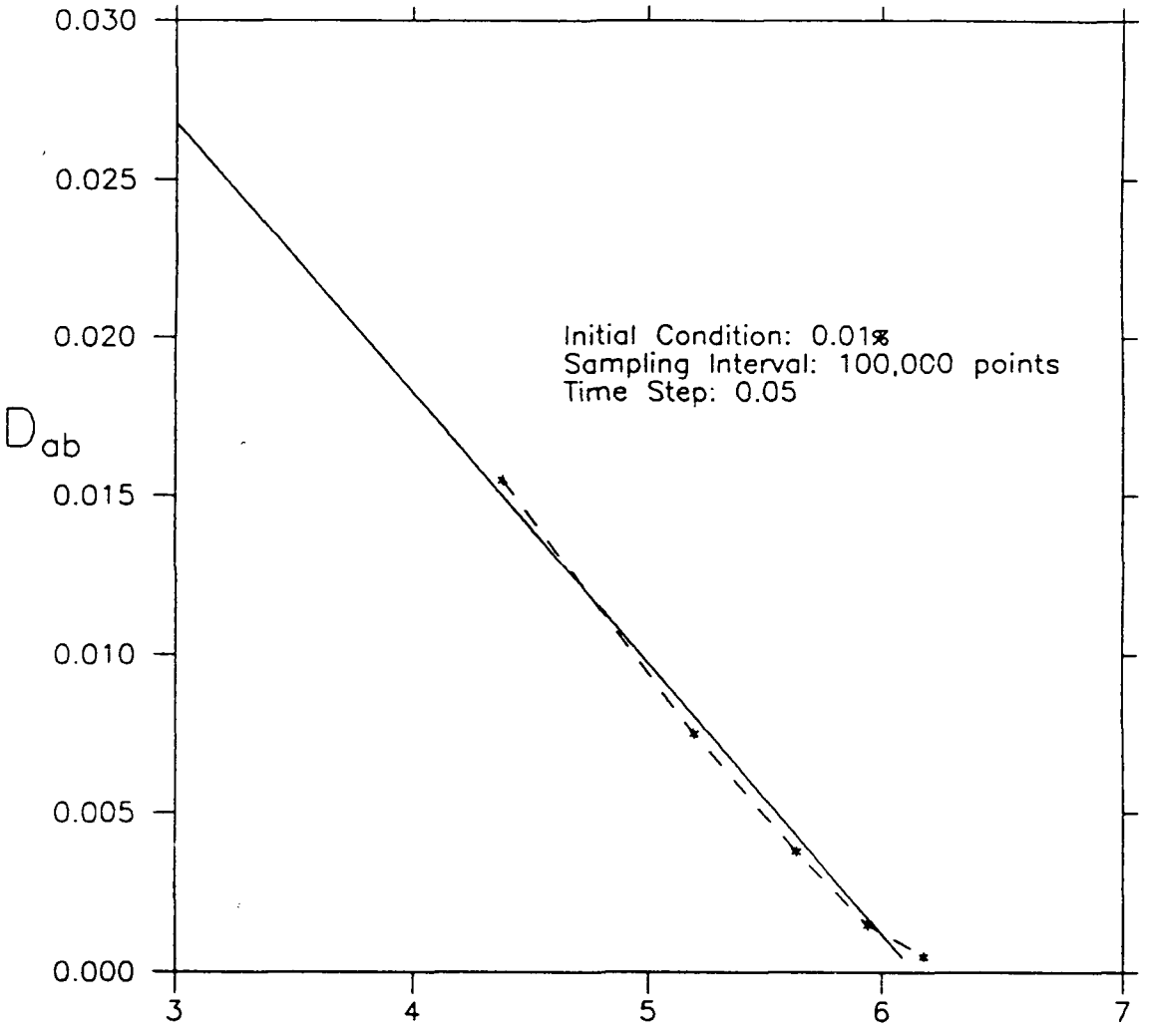


Figure 4.40: The log–linear representation of D_{ab} as a function of increasing series length comparison for the 0.01% initial condition and 100,000–point sampling interval. This fit is an extremely good one.

100,000 points respectively. Again, we conclude that the exponential fits of the data are relatively good ones. However, the values of the exponential decay rate are functions of sample size. The expressions for each are given in Table 4.10a. These results are somewhat disappointing, as we hoped to see a universal relation independent of both the sample size and the initial condition.

What may prove more successful is to reexpress these relations in terms of time instead of series length. We can convert the expressions for D_{avg} and D_{ab} into ones involving time by recognizing that the series length x is simply the ratio of the dimensionless time t and the time step t_S. For each expression, then, we obtain the time t such that

$$t = x * t_S , \tag{4.10}$$

When we apply this relation to the earlier expression for $D_{avg}(x)$ in (4.9), we obtain:

$$D_{avg}(t) = 0.041\, e^{-0.176t} + 0.008 , \tag{4.11}$$

Applying the same relation (4.10) to the nine D_{ab} expressions representing all combinations of the three initial conditions and the three sampling intervals, we

Table 4.10: The exponential expressions quantifying the rate of decay of new information gain

(a) within the 0.01% initial condition as a function of sampling interval. Note that in terms of series length x, the decay rate is largely dependent on the size of the sample.

Sampling Interval	Expression
25,000	$D_{ab}(x) = 0.024\, e^{-0.013x} + 0.0023$
50,000	$D_{ab}(x) = 0.026\, e^{-0.010x} + 0.0023$
100,000	$D_{ab}(x) = 0.028\, e^{-0.008x} + 0.0026$

(b) for all three initial conditions and all three sampling intervals, In terms of time t, the decay rate is relatively independent of both the sampling interval and the initial condition that is used.

Sampling Interval	Initial Condition	Expression
25,000	0.01%	$D_{ab}(t) = 0.024\, e^{-0.168t} + 0.0023$
50,000	0.01%	$D_{ab}(t) = 0.026\, e^{-0.181t} + 0.0023$
100,000	0.01%	$D_{ab}(t) = 0.028\, e^{-0.189t} + 0.0026$
25,000	0.10%	$D_{ab}(t) = 0.023\, e^{-0.169t} + 0.0025$
50,000	0.10%	$D_{ab}(t) = 0.020\, e^{-0.158t} + 0.0022$
100,000	0.10%	$D_{ab}(t) = 0.023\, e^{-0.168t} + 0.0025$
25,000	1.0%	$D_{ab}(t) = 0.022\, e^{-0.177t} + 0.0025$
50,000	1.0%	$D_{ab}(t) = 0.023\, e^{-0.182t} + 0.0022$
100,000	1.0%	$D_{ab}(t) = 0.021\, e^{-0.166t} + 0.0025$

produce the expressions that are shown in Table 4.10b. We observe that all nine expressions show relatively similar exponential decreases as functions of time. These nine separate expressions averaged together satisfy the relation

$$D_{ab}(t) = (0.024 \pm 0.004)\, e^{-(0.175 \pm 0.017)t} + (0.0024 \pm 0.0002)\,, \qquad (4.12)$$

Thus to within only a 10% variability, we have found an expression quantifying the rate of decay of new information gain that is *independent of both initial condition and sample size.*

The above method seems viable for quantifying the decay rate of information gain for D_{ab}. However, we believe that a better, more reliable estimate is obtained by averaging *all nine data sets together* and then finding *one best fit* to that data set. When we do this, we obtain the relation

$$D_{ab}(t) = 0.024 \exp^{-0.172t} + 0.0023\,, \qquad (4.13)$$

Comparing the expression for D_{avg} (4.11) with the one for D_{ab} (4.13) yields a fascinating result. Although the values of their floors and intercepts differ, their values of exponential decay rate of new information gain are the same to *within only a 2%*

difference. This is exciting, as we have found a general expression that is not only independent of the initial condition that we use and the way that we sample the model-generated data, but also on whether we use a *single or multiple initial condition* integration method.

One parameter that we have not taken into consideration is the time step value. We recall that in the beginning of this section we justified the decision to use the larger time step value because of its monotonic appearance. However, despite the greater irregularities seen in the smaller time step (0.005) curves, we speculate that there may well exist in them similar relations to those found in the larger time step curves, although their predictability estimates may be more suspect, because we are fitting exponential curves to data sets that are less well-behaved.

Having quantified the above behavior for the low-order Lorenz model, we believe that similar predictability estimates are possible when applying these procedures to longer time series. Quantifying this rate of decay in information gain to within a small tolerance is an exciting prospect; these expressions may even represent less expensive analogies to the Lyapunov exponents (Osledec 1968) or local divergence rates (Nese 1989) that are widely used to quantify the stability properties of dynamical systems.

Based on numerous successes in quantifying adequate data samples through histogram convergence, we feel quite confident in utilizing the simple, relatively inexpensive Histogram Measure as a reference for the eventual development of optimum sampling strategies. Although not attempting to do this here, we have successfully addressed many of the sampling issues that must be understood before determining such optimum strategies for a particular model. By doing so, we have laid the groundwork for related studies using longer, more complicated time series that are either model-generated or observed.

CHAPTER 5

SUMMARY OF RESULTS AND RELATED CONCLUSIONS

In this study, because of the current arguments over what constitutes adequate samples of chaotic time series upon which to quantify chaos measures, we have determined objectively the amount of data necessary to achieve such samples. Before accomplishing that, however, we distinguished the chaotic solutions from the transient, or nonchaotic, solutions that are inherent in chaotic time series. This is a crucial consideration, as transients contaminate the chaotic characteristics of any time series. Because current conventional measures that are used to quantify chaotic time series are quite expensive to calculate, we have developed a new, relatively inexpensive measure, the Histogram Measure, that allows us to successfully work with large data sets. This measure quantifies the structure of an attractor by giving the distribution of trajectory distances from the phase space origin. We have also demonstrated the links between this measure to that of the more conventional Correlation Dimension Measure

(Grassberger and Procaccia 1983b) that is used to quantify chaotic time series, and we have successfully quantified a unique predictability estimate, that of loss of information gain as functions of series length and elapsed time.

We have demonstrated remarkable success in using the simple, inexpensive Histogram Measure to distinguish between the chaotic and nonchaotic portions of a time series. We find that the duration of the transients can be determined by either noting the interval at which the histogram structures appear to converge or more objectively, by noting when the bin values of the histograms differ by less than a specified tolerance after sufficient portions of data have been eliminated from the initial part of the series. We also find that this transient duration is highly dependent upon the value of the initial condition, but largely independent of sample size. For the three initial conditions values with which we work, we are confident that transient solutions have been eliminated once we have removed at least 15,000 points from the initial portion of the series. Despite this success, we caution that the determination of transients and their duration with these methods is series dependent, since the transients in other chaotic time series may be more strongly masked.

Once we have eliminated definitively the nonchaotic poritons of these series, we have successfully found adequate data sets of the Lorenz (1963) Rayleigh-Bénard

convection model of chaos. Quantifying these required data sets, however, is not straightforward, as the model data are extremely sensitive to the manner in which they are sampled, as they are functions of sample size, initial condition, and time step. The most notable of these sensitivities is in the time step used to generate the data set. When viewing histogram convergence as a function of series length, we find that sampling the data with the larger time step value (0.05) produces convergence with fewer points than does sampling at the smaller time step (0.005). In effect, sampling every point in the 0.05 case is most likely analogous to sampling at every tenth point in the 0.005 case, thereby giving validity to the notion of optimal sampling strategies. In terms of total time, however, we have shown that the data sets produced with the smaller time step converge five to seven times faster than do the larger time step data sets, indicating that sampling with a smaller time step yields a better, more accurate representation of the attractor. Thus, there are benefits and tradeoffs related to the way that we sample; the choice of time step with which to sample the attractor can be tailored to the specific needs and limitations of the user.

The existence of relatively stable minimum differences in the histograms, both among and within initial conditions, suggests that there is a variability intrinsic to the Lorenz model upon which we can not improve significantly beyond a certain time or

number of data points. This first suggests that the best that we can do is to quantify adequate data sets to within a reasonable tolerance; once at that tolerance value, using larger data sets does not provide better estimates. This also suggests that a definitive cost/benefit analysis is possible--in effect, the degree of error that we are willing to tolerate when making predictability estimates from chaotic time series is balanced against the expense of generating enormous data sets. For example, we recall in Section 4.3.1. that the mean average absolute difference value D_{avg} when displayed as a function of series length (Figure 4.23) exhibits a sharply decreasing, exponential behavior to a stable floor at a value of approximately 0.008 by 480,000 points. However, by using less than one half the number of points (230,000), we have only increased the error to approximately 0.014. This suggests an important tradeoff between obtaining the maximum possible accuracy in defining convergence and the relative cost that we incur by doing so.

Another intriguing result of this study demonstrates the similarities in sampling between the Histogram Measure and the more conventional Correlation Dimension Measure ν that is used commonly to quantify estimates of chaotic time series. After conducting numerous tests in order to find reasonable convergence in the correlation dimension, we determine that choosing a larger time step value is the key issue.

Applying this knowledge to finding convergence with the Histogram Measure, we observe the same dramatic improvements in the degree of convergence exhibited by the data series. Even more remarkably, we note that just as the optimum correlation dimension plots (Figures 4.14-4.16) exhibit a small variability about the value 2.06 of ν that is usually stated for the Lorenz model, the Histogram Measure flags unique intrinsic variabilities also with values that are dependent on the manner in which the data are sampled. These are exciting results, as we have developed a measure possibly comparable to the standard Correlation Dimension Measure that does not exhibit similar severe, cost-limiting constraints on the amount of data that can be sampled. However, the full extent of the relation between the two measures remains inconclusive. Preliminary tests show that reducing the bin width within the Histogram Measure to the optimum one (0.01) used in the Correlation Dimension Measure yields a significantly noisier histogram structure with a relatively large variability about the mean. In addition, attempts to quantify the fractal dimension of the histogram structure itself have not yielded any significant success. Further work is necessary in these areas to determine the full extent of the benefits that the Histogram Measure provides over that of the more standard fractal dimension measures.

Quantifying the rate of decay of information gain of the data have also yielded

fascinating results. We have found that, given a particular time step with which to sample the data, there is a time-dependent, exponential decrease of information to be gained that is relatively independent of not only the initial condition and sample size, but also of the type of series comparison used (single or multiple initial condition). This result certainly verifies that the use of Monte Carlo techniques, which involve comparison of the divergence rates between solutions having different, randomly chosen initial conditions, have merit for estimating the predictability characteristics of chaotic time series. More importantly, this result suggests that we can optimize the choice of time step with which to sample the data; the larger the exponential decrease to a reasonable tolerance, the faster the series will converge to an adequate sample. This predictability estimate may possibly provide an inexpensive analog to the widely used Lyapunov exponents (Osledec 1968) or local divergence rates (Nese 1989).

Having utilized the simple, inexpensive Histogram Measure to successfully obtain adequate data sets that are sampled at every point, we theorize that optimal subsets of these data may be possible that optimize chaos estimates with far fewer points, at least to within suitably small tolerances. This reasoning stems from our results in Chapter 4 from which we conclude that the data sets that exhibit greater degrees of convergence and that are generated by using a larger time step (0.05) may

indeed be analogous to data sets produced using every tenth point having a time step of 0.005. Work is currently in progress, using the Histogram Measure and other statistical methods, to begin quantifying these subsets of the data and their resulting reliabilities in quantifying chaos estimates. If these sampling strategies prove successful for yielding adequate data sets generated with a low-order model, then we have reason to believe that similar applications can be made to chaotic time series that are longer, more complicated, and more operational in nature.

REFERENCES

Abarbanel, H., R. Brown and J.B. Kadtke, 1989: Prediction in chaotic nonlinear systems: Time series analysis for nonperiodic evolution. Institute for Nonlinear Science, 1-78.

Abraham, N.B., A.M. Albano, B. Das, G. De Guzman, S. Yong, R.S. Goggia, G.P. Puccione and J.R. Tredicce, 1986: Calculating the dimension of attractors from small data sets. Phys. Lett. A, 114, 217-221.

Ben-Mizrachi, A., I. Procaccia and P. Grassberger, 1984: Characterization of experimental (noisy) strange attractors. Phys. Rev., 29, 975-977.

Brandstäter, A., J. Swift, H.L. Swinney, A. Wolf, J.D. Farmer, E. Jen and P.J. Crutchfield, 1983: Low dimensional chaos in a hydrodynamic system. Phys. Rev. Lett., 51, 1442-1445.

Dutton, J.A. and R. Wells, 1984: Topological issues in hydrodynamic predictability. In Predictability of Fluid Motions, pp. 11-44, G. Holloway and B.J. West (Eds.). AIP Conference Proceedings, No. 106, American Institute of Physics, New York.

Ellner, S., 1988: Estimating attractor dimensions from limited data: A new method, with error estimates. Phys. Lett. A, 133, 128-133.

Feigenbaum, M.J., 1978: Quantitative universality for a class of nonlinear transformations. J. Stat. Phys., 19, 25-52.

Fraedrich, K., 1986: Estimating the dimensions of weather and climate attractors. J. Atmos. Sci., 43, 419-432.

Fraedrich, K., 1987: Estimating weather and climate predictability on attractors. J. Atmos Sci., 44, 722-728.

Fraedrich, K., 1988: El Niño southern oscillation predictability. Mon. Wea. Rev., 116, 1001-1012.

Gleick, J., 1987: Chaos: Making a New Science. Viking Penguin, Inc., New York, 352 pp.

Gollub, J.P. and H.L. Swinney, 1975: Onset of turbulence in a rotating fluid. Phys. Rev. Lett., 35, 927-934.

Grassberger, P., 1986: Do climatic attractors exist? Nature, 323, 609-612.

Grassberger, P. and I. Procaccia, 1983a: Characterization of strange attractors. Phys. Rev. Lett., 50, 346-349.

Grassberger, P. and I. Procaccia, 1983b: Measuring the strangeness of strange attractors. Physica, 9D, 189-208.

Guckenheimer, J. and G. Buzyna, 1983: Dimension measurements for geostrophic turbulence. Phys. Rev. Lett., 51, 1438-1441.

Henderson, H.W. and R. Wells, 1988: Obtaining attractor dimensions from meteorological time series. Adv. Geophys., 30, 205-237.

Hénon, M., 1976: A two-dimensional mapping with a strange attractor. Communications in Mathematical Physics, 50, 69-77.

Hense, A., 1987: On the possible existence of a strange attractor for the southern oscillation. Contr. Atm. Phys., 60, 34-47.

Higgins, R.W., 1987: From the equations of motion to spectral models. Chapter 3 in Nonlinear Hydrodynamic Modeling: A Mathematical Introduction, H.N. Shirer, (Ed.), Lecture Notes in Physics, 271, Springer-Verlag, Heidelberg, 47-69.

Krishna Mohan, T.R., J. Subba Rao and R. Ramaswamy, 1989: Dimension analysis of climatic data. J. Climate, 2, 1047-1057.

Lorenz, E.N., 1963: Deterministic nonperiodic flow. J. Atmos. Sci., 20, 130-141.

Lorenz, E.N., 1982: Atmospheric predictability experiments with a large numerical model. Tellus, 34, 505-513.

Lorenz, E.N., 1984: Irregularity: A fundamental property of the atmosphere. Tellus, 36A, 98-110.

Mandlebrot, B.B., 1977: The Fractal Geometry of Nature. Freeman, New York.

Nese, J.M., 1985: Phase space structure and dimension of attractors of finite spectral models. M.S. Thesis, Department of Meteorology, The Pennsylvania State University, University Park, PA, 180 pp.

Nese, J.M., 1987: Diagnosing the structure of attractors. Chapter 16 in Nonlinear Hydrodynamic Modeling: A Mathematical Introduction, H.N. Shirer, (Ed.), Lecture Notes in Physics, 271, Springer-Verlag, Heidelberg, 412-443.

Nese, J.M., 1989: Quantifying local predictability in phase space. Physica D, 35, 237-250.

Nese, J.M., J.A. Dutton and R. Wells, 1987: Calculated attractor dimensions for low-order spectral models. J. Atmos. Sci., 44, 1956-1970.

Nicolis, C. and G. Nicolis, 1986: Is there a climatic attractor? Nature, 311, 529-532.

Osledec, V.I., 1968: A multiplicative ergodic theorem. Lyapunov characteristic numbers for dynamical systems. Trans. Moscow Math. Soc., 19, 197-231.

Packard, N.H., J.P. Crutchfield, J.D. Farmer and R.S. Shaw, 1980: Geometry from a time series. Phys. Rev. Lett., 45, 712-716.

Rössler, O.E., 1981: Chaos and chemistry. Springer Series Synergetics, 12, 79-87.

Ruelle, D., 1979: Ergodic theory of differential dynamical systems. Institut des Hautes-Études Scientifiques Publications Mathématiques, 50, 27-58.

Shirer, H.N., (Ed.), 1987a: Nonlinear Hydrodynamic Modeling: A Mathematical Introduction, Lecture Notes in Physics, 271, Springer-Verlag, Heidelberg, 546 pp.

Shirer, H.N., 1987b: A simple nonlinear model of convection. Chapter 2 in Nonlinear Hydrodynamic Modeling: A Mathematical Introduction, H.N. Shirer (Ed.), Lecture Notes in Physics, 271, Springer-Verlag, Heidelberg, 22-46.

Sparrow, C., 1982: The Lorenz Equations: Bifurcations, Chaos, and Strange Attractors. Applied Mathematical Sciences, 41, Springer-Verlag, Heidelberg, 269 pp.

Takens, F., 1981: Lecture Notes in Mathematics, D.A. Rand and L.S. Young (Eds.), Springer, New York, 366 pp.

Thomson, D.W. and H.W. Henderson, 1989: Attactor dimensions and statistical properties of surface and profiler-measured tropospheric winds. Preprints, Ninth Symposium on Turbulence and Diffusion, Roskilde, Denmark, Amer. Meteor. Soc., Boston, MA, 220-223.