copies available

in yellow Global Analysis

1979

ON THE PREVALENCE OF APERIODICITY IN SIMPLE SYSTEMS

by

Edward N. Lorenz

0. Introduction

As the lone meteorologist at a seminar of mathematicians, I feel that a few words regarding my presence may be in order. Let me begin with some remarks about the mathematics of meteorology.

One of the most familiar problems of interest to meteorologists is weather forecasting. Mathematically this is an initial-value problem. The atmosphere and its surroundings are governed by a set of physical laws which in principle can be expressed as a system of integro-differential equations. At the turn of the century, the forecast problem was identified by Bjerknes [1] as the problem of solving these equations, using initial conditions obtained from observations of current weather. Detailed numerical procedures for solving these equations were formulated during World War I by Richardson [8], but the practical solution of even rather crude approximations had to await the advent of computers.

Another problem of interest is climate. This is a problem in dynamical systems. The climate is often identified with the set of all long-term statistical properties of the atmospheric equations. It is commonly assumed that one can devise a finite system of ordinary differential equations whose typical solutions nearly duplicate those of the more realistic system. In the phase space of such a system, weather forecasting, particularly at short range, is a local problem, while climate is global.

The atmosphere is a forced dissipative system; the forcing is thermal, while the dissipation is thermal and mechanical. Any system of equations whose general solution can hope to approximate the atmosphere must likewise contain forcing and dissipation. The various orbits in phase space are therefore not confined to separate energy surfaces, as they would be in a conservative system.

During my early exposure to theoretical meteorology, I had wondered whether there might instead be a single surface which a few special orbits would occupy, and which the remaining orbits would approach. I had even hoped to discover some smooth function which would vanish on this surface, and would assume positive values on one side and negative values on the other. Needless to say I was unsuccessful, and, in the light of more recent results, the search for such a function seems rather naive. I presently turned to other matters.

My return to dynamical systems was prompted by an interest in weather forecasting rather than climate. By the middle 1950's "numerical weather prediction", i.e., forecasting by numerically integrating such approximations to the atmospheric equations as could feasibly be handled, was very much in vogue, despite the rather mediocre results which it was then yielding. A smaller but determined group favored statistical prediction, and especially prediction by linear regression, using large numbers of predictors. Apparently because of a misinterpretation of a paper by Wiener [12], the idea became established that the statistical method could duplicate the performance of the dynamical method, despite the essential nonlinearity of the dynamic equations. I was skeptical, and decided to test the idea by applying the statistical method to a set of artificial data, generated by solving a system of equations numerically. Here the dynamical method would consist of solving the equations all over again, and would obviously give perfect results. I doubted very much that the statistical method would do likewise.

The first task was to find a suitable system of equations to solve. In principle any nonlinear system might do, but a system with some resemblance to the atmospheric equations offered the possibility of some useful by-products. The system would have to be simple enough to be able to simulate a rather long stretch of weather with a reasonable amount of computation. Moreover, the general solution would have to be aperiodic, since the statistical prediction of a periodic series would be a trivial matter, once the periodicity had been detected. It was not obvious that these conditions could be met.

At about this time our group was fortunate enough to secure its own digital computer, which still sits across the hall from my office. The computer was slow by today's standards, but we were competing with no one for its use. Moreover, its very slowness

...ing printed as it was produced, and we could stop the computation and introduce new numbers whenever the present output appeared uninteresting.

We first chose a system which had been used for numerical weather prediction. The system represented the three-dimensional structure of the atmosphere by two horizontal surfaces, and we proceeded to expand the horizontal field of each atmospheric variable in a series of orthogonal functions. We then reduced the system to manageable size by discarding all terms of the series except those representing the largest few horizontal scales, and programmed the resulting system for the little computer.

For a while our search produced nothing but steady or periodic solutions, but at last we found a system of twelve equations whose solutions were unmistakably aperiodic. It was now a simple matter to put the statistical forecasting method to test, and we found, incidentally, that it failed to reproduce the numerically generated weather data.

During our computations we decided to examine one of the solutions in greater detail, and we chose some intermediate conditions which had been typed out by the computer and typed them in as new initial conditions. Upon returning to the computer an hour later, after it had simulated about two months of "weather", we found that it completely disagreed with the earlier solution. At first we suspected machine trouble, which was not unusual, but we soon realized that the two solutions did not originate from identical conditions. The computations had been carried internally to about six decimal places, but the typed output contained only three, so that the new initial conditions consisted of old conditions plus small perturbations. These perturbations were amplifying quasi-exponentially, doubling in about four simulated days, so that after two months the solutions were going their separate ways.

It soon became evident that the instability of the system was the cause of its lack of periodicity. The variables all had limited ranges, so that near repetitions of some previous conditions were inevitable. Had the system been stable, the difference between the original occurrence and its near repetition would not have subsequently amplified, and essentially periodic behavior would have resulted.

I immediately concluded that, if the real atmospheric equations behaved like the model, long-range forecasting of specific weather conditions would be impossible.

The observed aperiodicity of the atmosphere, once the normal diurnal and annual variations are removed, suggests that the atmosphere is indeed an unstable system. The inevitable small errors in observing the current weather should therefore amplify and eventually dominate.

Still, I felt that we could better appreciate the problems involved by studying a simpler example. The ideal system would contain only three variables, whence we could even construct models of orbits in phase space, or of the surface, if any, which these orbits would approach. However, our attempts to strip down the twelve-variable system while retaining the aperiodicity proved fruitless.

The break came when I was visiting Dr. Barry Saltzman, now at Yale University. In the course of our talks he showed me some work on thermal convection, in which he used a system of seven ordinary differential equations [5]. Most of his numerical solutions soon acquired periodic behavior, but one solution refused to settle down. Moreover, in this solution four of the variables appeared to approach zero.

Presumably the equations governing the remaining three variables, with the terms containing the four variables eliminated, would also possess aperiodic solutions. Upon my return I put the three equations on our computer, and confirmed the aperiodicity which Saltzman had noted. We were finally in business.

1. A Physical System with a Strange Attractor.

In a changed notation, the three equations with aperiodic solutions are

$$dX/dt = -\sigma X + \sigma Y \quad , \qquad (1.1)$$

$$dY/dt = -XZ + rX - Y \quad , \qquad (1.2)$$

$$dZ/dt = XY - bZ \quad . \qquad (1.3)$$

Although originally derived from a model of fluid convection, (1.1)-(1.3) are more easily formulated as the governing equations for a laboratory water wheel, constructed by Professor Willem Malkus of M.I.T. to demonstrate that such equations are physically realizable. The wheel is free to rotate about a horizontal or tilted axis. Its circumference is divided into leaky compartments. Water may be introduced from above, whereupon the wheel can become top-heavy and begin to rotate. Different compartments will then move into position to receive the water. Depending upon the values of the constants of the apparatus, the wheel may be observed to remain at rest, rotate continually in one direction or the other, or reverse its direction at regular or irregular intervals.

The equations are written for a wheel of radius a with a horizontal axis, and with its mass confined to the rim. Its angular velocity $\Omega(t)$ may be altered by the action of gravity g on the nonuniformly distributed mass and by frictional damping proportional to Ω. The mass $\rho(t,\theta)$ per unit arc of circumference may be altered by a mass source increasing linearly with height, a mass sink proportional to ρ, and, at a fixed location in space, by rotation of the wheel. Here t is time and θ is arc of circumference, measured counterclockwise. The wheel then obeys the equations

$$d(a^2\bar{\rho}\,\Omega)/dt = -\,g\,a\,\overline{\rho\cos\theta} - k\,a^2\bar{\rho}\,\Omega \quad , \qquad (1.4)$$

$$\partial\rho/\partial t + \Omega\,\partial\rho/\partial\theta = A + 2B\sin\theta - h\rho \quad , \qquad (1.5)$$

representing the balances of angular momentum and mass, where $\overline{}$ denotes an average with respect to θ, and A, B, k, and h are positive constants. From (1.5) it follows that $\bar{\rho}$ approaches A/h exponentially; assuming that $\bar{\rho}$ has reached A/h, (1.4) and (1.5) yield the three ordinary differential equations

$$d\Omega/dt = -k\,\Omega - (gh/aA)\,\overline{\rho\cos\theta} \quad , \qquad (1.6)$$

$$d\,\overline{\rho\cos\theta}\,/\,dt = -\,\Omega\,\overline{\rho\sin\theta} - h\,\overline{\rho\cos\theta} \quad , \qquad (1.7)$$

$$d\,\overline{\rho\sin\theta}\,/\,dt = \Omega\,\overline{\rho\cos\theta} - h\,\overline{\rho\sin\theta} + B \quad . \qquad (1.8)$$

With a suitable linear change of variables, (1.6)-(1.8) reduce to (1.1)-(1.3), with $b=1$.

In the convective model the motion takes place between a warmer lower surface and a cooler upper surface, and is assumed to occur in the form of long rolls with fixed parallel horizontal axes and quasi-elliptical cross sections. The water wheel is therefore like a "slice" of a convective roll. The variables X, Y, Z measure the rate of convective overturning and the horizontal and vertical temperature variations. The damping results from internal viscosity and conductivity, and σ denotes the Prandtl number, while r is proportional to the Rayleigh number. Because the horizontal and vertical temperature structures differ, Y and Z need not damp at the same rate, whence b need not equal unity. The equations may afford a fair representation of real convection when r is near unity, but they become unrealistic when r is large, since real convective rolls would then break up into smaller eddies.

Although we have discussed (1.1)-(1.3) in detail elsewhere [4], we shall repeat some of the results needed for the later discussion. First, it follows that

$$\tfrac{1}{2}\, d\,[X^2 + Y^2 + (Z-\sigma-r)^2] / dt = -[\sigma X^2 + Y^2 + b(Z-\tfrac{1}{2}\sigma - \tfrac{1}{2}r)^2] + b(\tfrac{1}{2}\sigma + \tfrac{1}{2}r)^2 \quad . \tag{1.9}$$

The ellipsoid E in (X,Y,Z) - phase-space defined by equating the right side of (1.9) to zero passes through the center of the sphere S_o whose equation is $X^2 + Y^2 + (Z-\sigma-r)^2 = c^2$, and hence lies wholly in the region R_o enclosed by S_o, provided that c exceeds the maximum diameter of E. It follows from (1.9) that every point exterior to E, and hence every point exterior to S_o, has a component of motion toward the center of S_o, so that every orbit ultimately becomes trapped in R_o.

Next, if S is a surface enclosing a region R of volume V,

$$dV/dt = -(\sigma + b + 1)V \quad . \tag{1.10}$$

Hence, following the passage of time intervals Δt, $2\Delta t$, ..., S is carried into surfaces S_1, S_2, ... enclosing regions R_1, R_2, ... of volumes V_1, V_2, ..., where $V_n \to 0$ exponentially. If $S = S_o$, $R_o \supset R_1 \supset R_2 \supset \ldots$, whereupon every orbit is ultimately trapped in a set $R_\infty = R_o \cap R_1 \cap R_2 \cap \ldots$ of zero volume. This set could be a point, a curve, a surface, or a complex of points, curves, or surfaces.

The attractor set is R_∞ , or a portion of R_∞.

Eqs. (1.1)-(1.3) possess the obvious steady solution $X = Y = Z = 0$; this becomes unstable when $r > 1$. In this event there are two additional steady solutions $X = Y = \pm (br-b)^{\frac{1}{2}}$, $Z = r-1$; these become unstable when r passes its critical value

$$r_c = \sigma(\sigma + b + 3) (\sigma - b - 1)^{-1} \quad . \tag{1.11}$$

This can occur only if $\sigma > b + 1$. We shall consider only solutions where $r \geqq r_c$; these are most readily found by numerical integration.

In the first example we shall use Saltzman's values $b = 8/3$ and $\sigma = 10$, whence $r_c = 470/19 = 24.74$; as in [4] we shall use the slightly supercritical value $r = 28$. Here we note another lucky break; Saltzman used $\sigma = 10$ as a crude approximation to the Prandtl number (about 6) for water. Had he chosen to study air, he would probably have let $\sigma = 1$, and the aperiodicity would not have been discovered.

For advancing in time we use the alternating 4-cycle scheme [6], equivalent to a fourth-order Runge-Kutta scheme, with a time increment $\delta t = 0.005$. Our initial point $X = Y = 6.0$, $Z = 13.5$ lies on the parabola passing through the fixed points.

Fig. 1.1 shows the variations of X, Y, and Z from $t = 9$ to $t = 18$; the behavior seems to be typical. Evidently Z is always positive, and possesses a succession of unambiguously defined maxima and minima, spaced at fairly regular but not exactly equal intervals. In absolute value X and Y behave somewhat like Z, but they change sign rather irregularly.

Fig. 1.2 shows the projection of the orbit on the Y-Z plane, from $t = 9$ to $t = 14$. The three unstable fixed points are at 0, C, and C'. The curve spirals outward rather regularly from C' or C until it reaches a critical distance, whereupon it crosses the Z-axis and merges with the spiral about C or C'.

Following a brief initial interval the orbit should be virtually confined to the attractor set. Fig. 1.3 shows the topography of the attractor, as seen from

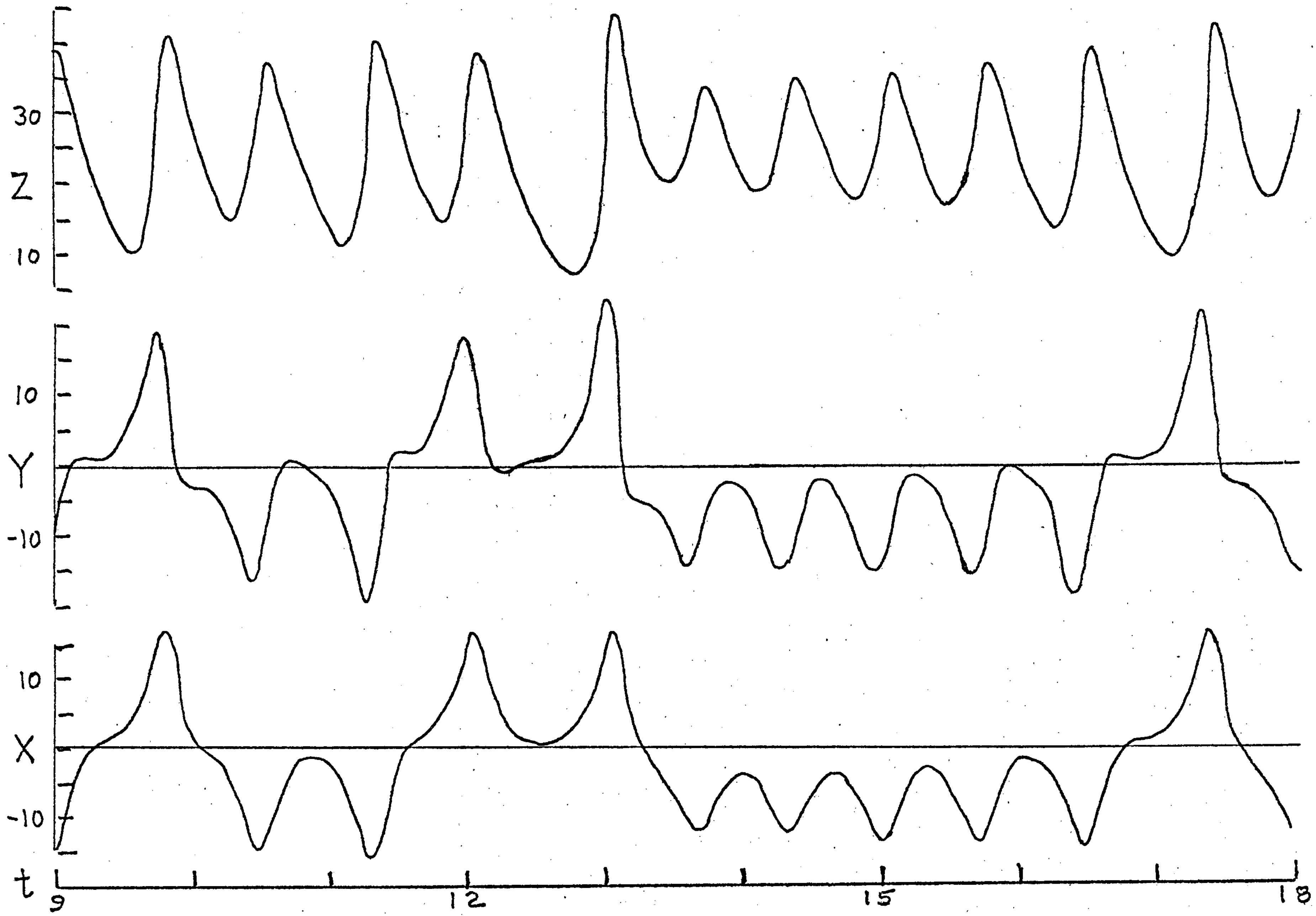

Fig. 1.1. Variation of X, Y, Z for particular solution of Eqs. (1.1)-(1.3).

the positive X-axis; the curves are contours of X. Where there are two values of X, the higher one occurs if the orbit is just completing a circuit about C. As one follows an orbit, the two sheets of the attractor appear to merge; however, this would require pairs of orbits to merge, which is impossible. Hence what appears to be a single sheet must be composed of two sheets, extremely close together, so that what looks like two merging sheets must contain four sheets. Continuing with this reasoning, we find that these four sheets must be eight sheets, then sixteen, etc, and we conclude that there is actually an infinite complex of sheets. The closure of these sheets forms the attractor set; a curve normal to the sheets would intersect it in a Cantor set. Attractors of this sort have become known as strange attractors [9].

The regularity of the spirals about C and C' in Fig. 1.2 implies that the value Z_n of Z at its n^{th} maximum determines with fair precision the value Z_{n+1} at the following maximum, as well as indicating whether Y will change sign before the next maximum occurs. Fig. 1.4 is constructed as a scatter diagram of successive maxima of Z , but in fact reveals no scatter. It appears to define a difference equation

$$Z_{n+1} = F(Z_n) \quad , \tag{1.12}$$

whose analytic form cannot however easily be determined. We shall base our subsequent conclusions rather heavily on the appearance of Fig. 1.4, and on the assumption that it is for practical purposes a curve.

Maxima of Z are intersections of the orbit with the conic $bZ = XY$. The curve in Fig. 1.4 is therefore a form of Poincaré map; we shall call it a Poincaré curve. The conic intersects a surface of constant $Z > 0$ in a hyperbola. Since the attractor has zero volume, it intersects the hyperbola in a set of measure zero, which must be a Cantor set. The orbits emanating from this Cantor set reintersect the conic in a set whose Z-coordinates form another Cantor set. It follows that a vertical line in Fig. 1.3 intersects the Poincaré

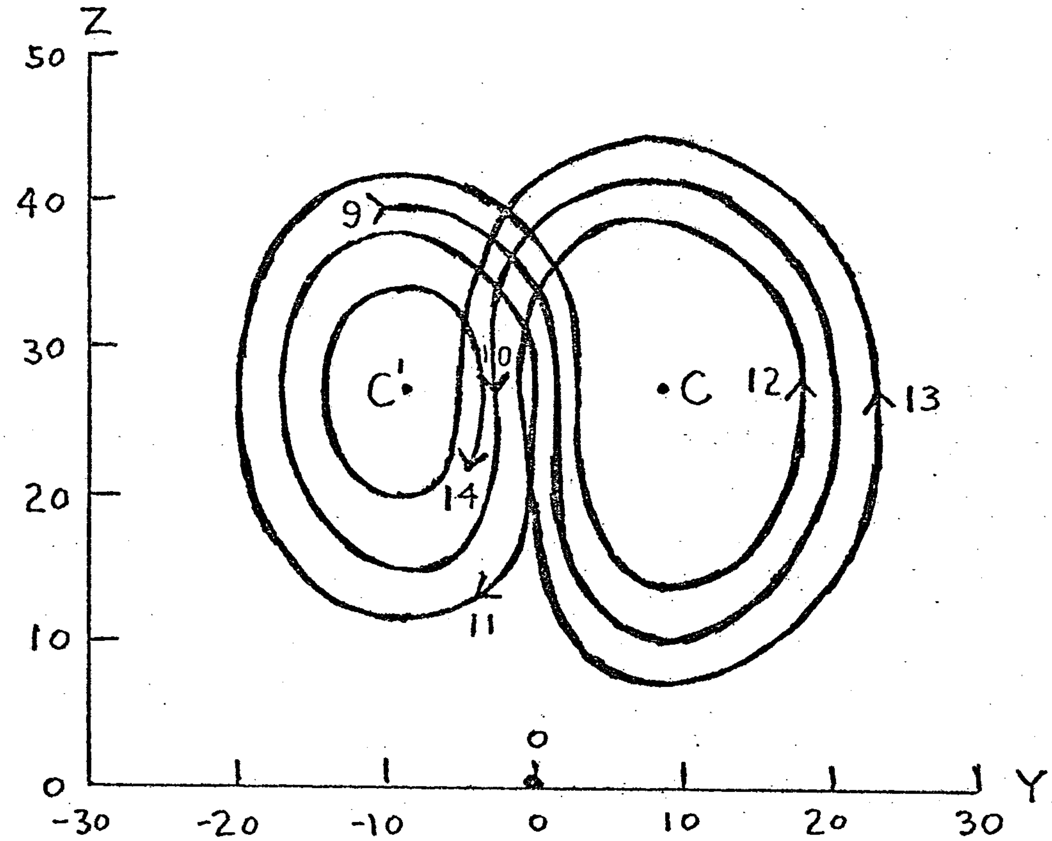

Fig. 1.2. Projection of segment of solution of Eqs. (1.1)-(1.3) on Y-Z plane. Numbers 9-14 indicate values of t. Unstable fixed points are at 0, C, and C'.

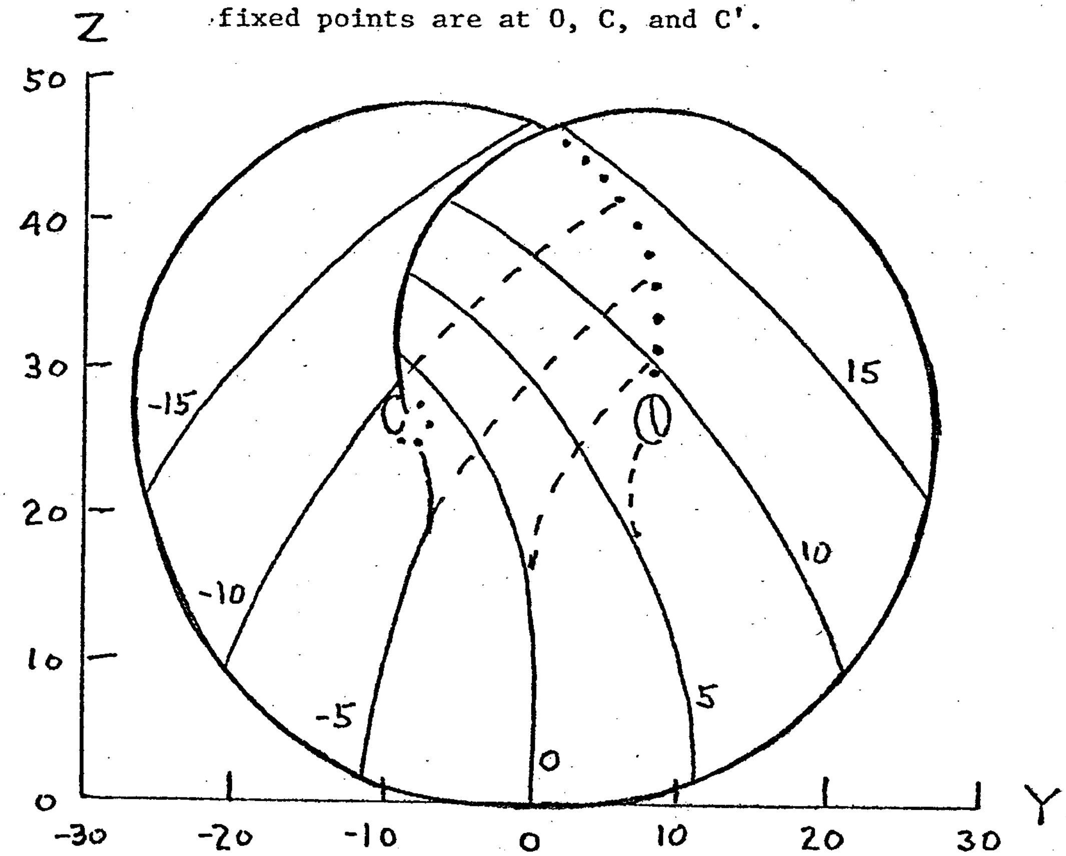

Fig. 1.3. Topography of the attractor for Eqs. (1.1)-(1.3). Solid lines are contours of X; dashed lines are contours of lower value of X where two values occur. Heavy curve is natural boundary of attractor.

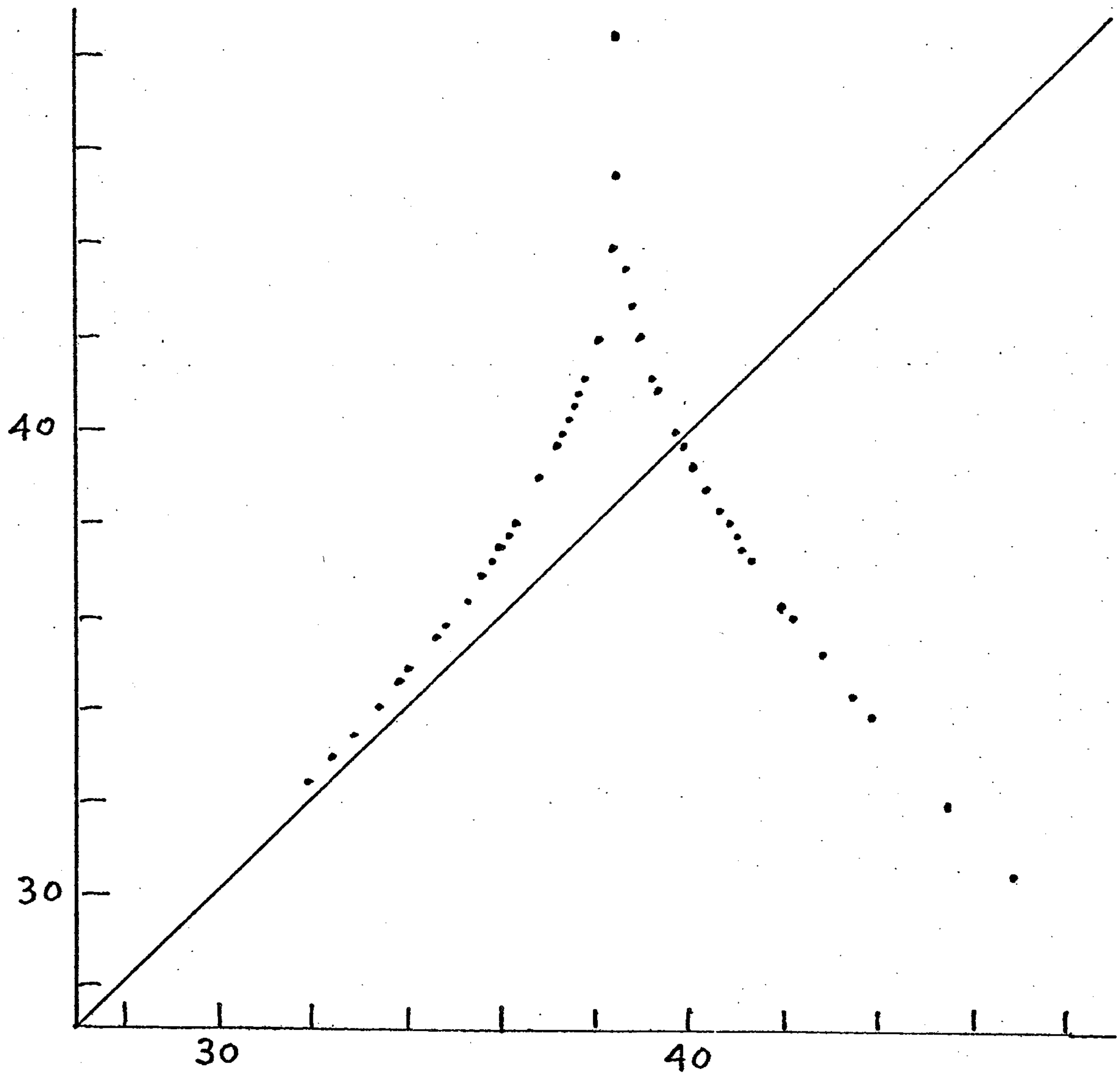

Fig. 1.4. "Scatter diagram" for successive maxima of Z for Eqs. (1.1)-(1.3), revealing lack of scatter.

curve in a Cantor set, so that the curve is really composed of a Cantor set of curves. However, the extreme horizontal distance between two curves on the same side of $Z_n = 38.5$ is about 10^{-4} times the distance between two curves on opposite sides, so that the Poincaré curve is closely approximated by a pair of merging curves, i.e., a single curve with a cusp.

A sequence $Z_1, Z_2, \ldots$ of successive maxima may be exactly periodic, i.e., $Z_m = Z_n$ with $m \neq n$. It may be asymptotically periodic, i.e., asymptotic to a periodic sequence $Z_1', Z_2', \ldots$, in which case the latter sequence must be stable. Finally it may be aperiodic. The corresponding orbits will likewise be periodic (exactly or asymptotically) or aperiodic.

Assuming reasonable smoothness of F, there will be a finite number of exactly periodic sequences of a given period, and hence a countable number altogether. Thus almost all sequences are either asymptotically periodic or aperiodic. If no periodic sequences are stable, almost all sequences are aperiodic.

If a maximum Z_n is altered slightly, Z_{n+1} will be altered by the same amount, multiplied by the slope $\lambda_n = d\,Z_{n+1}/d\,Z_n$ of the Poincaré curve. An exactly periodic sequence of period N is therefore unstable or stable according to whether or not the product Λ_N of the slopes at the N points exceeds unity in absolute value. Since it appears from Fig. 1.4 that the slope exceeds unity everywhere, there are no stable periodic sequences, and the general solution of (1.1)-(1.3) is aperiodic.

2. Aperiodicity in a Quadratic Difference Equation

In the previous section we used a difference equation, defined graphically rather than analytically, to draw conclusions regarding a system of differential equations. We now turn to an analytically defined difference equation.

The single first order equation

$$x_{n+1} = a(x_n - x_n^2) \quad , \tag{2.1}$$

to which any quadratic difference equation may be reduced by a linear change of variables, has been extensively studied as the "simplest" nonlinear equation; a comprehensive discussion is given by Guckenheimer [2]. The solutions of (2.1), like those of (1.12), may be exactly or asymptotically periodic, or aperiodic.

We shall replace (2.1) by

$$X_{n+1} = \tfrac{1}{2} X_n^2 - A \quad , \tag{2.2}$$

where $X_n = a(1-2x_n)$ and $A = \frac{1}{2} a^2 - a$. The variable X_n is then the slope of the graph of X_{n+1} against X_n. It is evident that if $-\frac{1}{2} \leqq A \leqq 4$ and $|X_0| \leqq a$, $|X_n| \leqq a$ for all n; if also $A > 0$, $|X_n| \leqq A$ for large enough n. If

$$\Lambda_N = \prod_{n=1}^{N} X_n \quad , \tag{2.3}$$

a periodic solution with $X_N = X_0$ is unstable or stable according to whether or not $|\Lambda_N| < 1$.

Our principal concern is with the probability P that if A and X_0 are chosen randomly from (0, 4) and (0, A), the sequence $X_0, X_1, X_2, \ldots$ will be aperiodic; specifically we are interested in whether $P = 0$ or $P > 0$. In an earlier study [5] we conjectured that $P > 0$. We are as yet unable to prove the conjecture, so we shall simply present supporting evidence, which will at times lack the rigor needed for a proof.

Our interest in this question stems from the existence of relations between difference and differential equations, as illustrated by Eqs. (1.1)-(1.3) and (1.12). We believe that the answer for a large class of difference equations is the same as the answer for (2.2), and that many systems of differential equations, including some representing physical systems, give rise to difference equations of this sort. In a sense, then, we are asking whether aperiodicity is an exceptional or a normal phenomenon.

The point where $X_n = 0$ is called a _singularity_. For any A, we shall call the solution with $X_0 = 0$ the _singular_ solution. A useful theorem [3, 5, 11] tells us that if a stable periodic solution exists, the singular solution approaches it asymptotically. A corollary is that there is at most one stable periodic solution.

We shall call a value of A <u>periodic</u> if a stable periodic solution exists, and <u>aperiodic</u> otherwise. If the (Lebesque) measure of the set of aperiodic values of A in $(0, 4)$ exceeds zero, $P > 0$.

It is easily shown that a stable solution of period 1 (steady) exists if $-\frac{1}{2} < A < \frac{3}{2}$; this bifurcates to a period 2 which is stable if $\frac{3}{2} < A < \frac{5}{2}$, and thence to period 4, 8, ... , the sequence of intervals terminating at $A = 2.802$. Within $(2.802, 4)$ there are <u>some</u> aperiodic values of A.

Numerical solutions of (2.2) suggest that if A' and A'' are distinct aperiodic values of A , the corresponding singular solutions eventually acquire opposite signs. It follows that for some intermediate value A_c , the singular solution is exactly periodic, and stable, since $\Lambda_N = 0$. Such a value of A will be called <u>central</u>.

By continuity there is an interval enclosing A_c where $|\Lambda_N| < 1$, whence a continuum of periodic values of A separates A' from A'' . The set of aperiodic values is therefore nowhere dense.

For example, $X_3=0$ when $A = A_c = 3.510$ and $X_0=0$, and the interval where period 3 is stable extends from $A = A_a = 3.5$ (exactly) to $A = 3.538$. For slightly higher values of A , period 6, then period 12, etc. are stable, and for still higher values up to $A = A_e = 3.581$, the singular solution is <u>semiperiodic</u> of period 3, i.e., there are three nonoverlapping intervals such that $X_0, X_3, X_6, \ldots$ occupy one, $X_1, X_4, \ldots$ occupy another, and $X_2, X_5, \ldots$ occupy the other. Within the semiperiodic range there are some periodic values of A , the periods being multiples of 3. For the aperiodic but semiperiodic values, a variance spectrum would contain lines superposed on a continuun. We shall call the interval from A_a to A_e a <u>semiperiodic band</u>.

A similar semiperiodic band encloses each other central value of A . The band for period 1 is the entire interval $(-\frac{1}{2}, 4)$, since a completely aperiodic considered semiperiodic of period 1. Moreover, every semiperiodic band (except for period 2) is virtually a small copy of $(-\frac{1}{2}, 4)$, containing within it the same structure. Thus there are bands within bands within bands, etc. A band which lies within no other band except $(-\frac{1}{2}, 4)$ will be called <u>prime</u>; other bands will be called <u>composite</u>. The period of a composite band is obviously a composite number; the converse does not hold.

Because of the similarity of the bands, the measure of the aperiodic values of A is positive if and only if the measure of the values of A exceeding 3/2 and not contained in prime bands is positive. We might then attempt to answer our question by summing the lengths of the prime bands. Table 1 presents these for periods ≤ 7; the band for period 2 ends at 3.0874; and it is evident that the remaining bands do not fill much of the space in (3.0874, 4). However, for any large period there exist a few prime bands, located very close to prime bands of much lower period, which, although narrow, are exceptionally wide for their period. For example, of the 26,817,356,775 bands of period 41, whose average width is certainly $< 3.4 \times 10^{-11}$, one, with $A_c = 3.49788$, has a width of 1.73×10^{-7}. We have not been able to show that these exceptional bands, taken together, do not fill the space which the "normal" bands leave nearly empty.

Our conjecture that $P > 0$ was originally prompted by the observation that when a value of A in (3, 4) was chosen at random, the resulting singular solution was usually aperiodic. We must therefore note that with the usual computer precision most solutions become incorrect before 100 iterations. The inevitable round-off errors introduced in the early iterations amplify by a factor whose average may approach 2.0 per iteration, until the noise drowns the signal. Indeed, May [7] regards the computer solutions as simulations, and suggests that there may be periodicities considerably higher than 100 which the simulations fail to reveal.

To test this possibility we have repeated some of the computations with a special multiple-precision program, using as many as 500 decimal places, and carrying upper and lower bounds to the true value of X. These bounds remain close together for 1000 and sometimes 3000 iterations. For no tested values of

Table 1. Limiting values A_a, A_e, central values A_c, and widths $A_e - A_a$ of prime semiperiodic bands of period $\leqq 7$, for Eq. (2.2).

N	A_a	A_c	A_e	A_e-A_a
2	1.50000	2.00000	3.08738	1.58738
7	3.14943	3.14978	3.15255	0.00312
5	3.24879	3.25083	3.26672	0.01793
7	3.34791	3.34813	3.34991	0.00200
3	3.50000	3.50976	3.58066	0.08066
7	3.66458	3.66463	3.66502	0.00044
5	3.72117	3.72156	3.72466	0.00349
7	3.76958	3.76961	3.76978	0.00020
6	3.81450	3.81456	3.81503	0.00053
7	3.85428	3.85430	3.85441	0.00013
4	3.88110	3.88160	3.88552	0.00442
7	3.90740	3.90741	3.90747	0.00007
6	3.93353	3.93355	3.93369	0.00016
7	3.95436	3.95436	3.95438	0.00002
5	3.97082	3.97085	3.97108	0.00026
7	3.98363	3.98363	3.98364	0.00001
6	3.99275	3.99275	3.99277	0.00002
7	3.99819	3.99819	3.99819	0.00000

A where we had not found a periodicity less than about 30 did we discover any higher periodicities. If the interval (3, 4) is filled by semiperiodic bands, the periods must be high indeed.

What we did generally observe in these solutions was that the product Λ_N continued to increase quasi-exponentially with N. The periodic bands seem to consist of those rare values of A where, after many iterations, we suddenly encounter a value of X so close to zero that it cancels the remaining factors in Λ_N. Encountering a value which partially cancels the product, and then another value which completes the cancelation, is also possible but seems less likely.

Accordingly, for our final bit of evidence supporting our conjecture we have constructed a statistical model of the difference equation (2.2). We take $A > 0$ and choose $X_1, X_2, \ldots$ randomly and independently from the interval $(-A, A)$. Letting Λ_N again be given by (2.3), we seek the probability $P(A)$ that $|\Lambda_N| > 1$ for all N. This model cannot prove or disprove our conjecture, since successive values of X generated by (2.2) are not independent, and the distribution of these values of X in $(-A, A)$ is not uniform. The model can be regarded as highly indicative.

We find that $P(A) = 0$ if $A \leq e$, but $P(A) = 1 - A'/A > 0$ if $A > e$, where $A' < e$ is a number such that $(\log A')/A' = (\log A)/A$. For example, if $A = 4$, $b = 2$ and $P = 1/2$; if $A = 3.375$, $A' = 2.25$ and $P = 1/3$. To establish this result we let P_N be the probability that $|\Lambda_N| < 1$ for $n \leq N$, and note that $1 - P_1 = 1/A$, while by direct integration $P_N - P_{N+1}$ equals $1/A$ times a function of $(\log A)/A$. Hence $1 - P(A)$ and $1 - P(A')$ differ only by the factor A'/A. Since the (geometric) mean of X is <1 when $A < e$, the result follows.

The implication is that for (2.2), in the vicinity of $A = 3.375$ about one third of the values of A should be aperiodic, while near $A = 4$ about one half should be aperiodic. The more general implication is that aperiodicity is a normal phenomenon. It is remarkable that aperiodic values for (2.2) first appear at $A = 2.802$, which is so close to e. Actually the numerical solutions suggest that nearly all values of A near 4 are aperiodic; the discrepancy may occur because the model has assumed a uniform distribution of X, while in reality the larger values tend to occur more frequently.

3. Some Attractors are Stranger than Others

It is apparent that there is a wide variety of systems of equations with aperiodic general solutions. There should therefore be a wide variety of strange attractors.

Let us consider (1.1)-(1.3) for other values of the parameters. For the values of b and σ previously used, but for $r = r_c$, the Poincaré curve would look about like Fig. 1.4, but with unit slope at the fixed point. Since it would still be concave upward everywhere, no stable periodic solutions would be introduced.

For any b, r_c is large if σ is near $b+1$ or if σ is large, and there is a value

$$\sigma_m = b+1 + [2(b+1)(b+2)]^{\frac{1}{2}} \tag{3.1}$$

for which r_c assumes a minimum value r_{cm}. To keep our study manageable, we shall vary b, letting $\sigma = \sigma_m$ and $r = r_{cm}$ in all cases. It is of interest that when $b = 8/3$, the values $\sigma = 10$ and $r = 28$ used previously are not far from $\sigma_m = 9.52$ and $r_{cm} = 24.72$.

We shall again base our conclusions mainly on the Poincaré maps. For purposes of comparison we shall divide X, Y, Z by their values at C, so that the fixed points become (0,0,0), (1,1,1), and (-1,-1,1).

In our numerical integrations we have chosen initial conditions on the parabola $Y = X$, $Z = X^2$ passing through the fixed points. We assume that after passing one or two maxima of Z the orbit is close to the attractor, and we study the remainder of the solution. By suitably adjusting the initial point along the parabola, we can force the orbit to visit the rarely visited portions of the attractor.

Fig. 3.1 shows the Poincaré curves for $b = 2$, 1, and 1/2. The curve for $b = 2$ is much like Fig. 1.4, and there is no possibility of a stable periodic solution. For $b = 1$ and 1/2, singularities have appeared, at $Z = 1.75$ and $Z = 1.60$, and the possibility of periodic solutions with $|\Lambda_N| < 1$ arises. Actually these do not occur, because solutions with points close to the singularity also contain points close to the cusp.

However, for $b = 1/2$ the solution with a maximum of Z exactly at 1.60 has the eighth subsequent maximum close to 1.60. If the curve were a single curve rather than a Cantor set of curves, we could be sure that by changing b slightly we would obtain a solution of period 8 with $\Lambda_8 = 0$. Lacking this

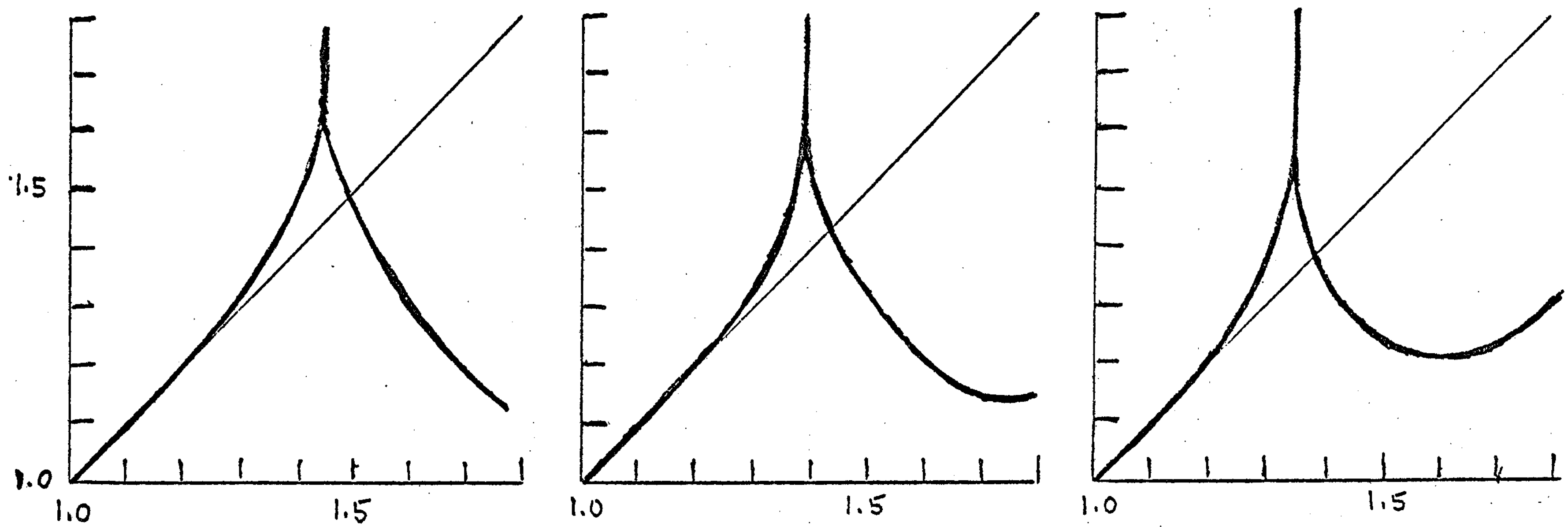

Fig. 3.1. Poincaré curves for Eqs. (1.1)-(1.3) for b = 2 (left), b = 1 (center), and b = 1/2 (right), with $\sigma = \sigma_m$ and $r = r_{cm}$.

assurance, we find for b = 1/2, by adjusting the initial conditions along the parabola, a solution where the third maximum of Z exactly hits the singularity. We then adjust b and repeat the process, until we have found the desired solutic of period 8. Finally we extend the solution through several more periods to be sure that it is indeed stable.

With some search we located a stable period 8 at b = 0.498007. Varying b by intervals of 0.25×10^{-6}, we found no period 8 at b = 0.49800625, and a stable period 8 only from 0.49800650 to 0.49800750. The solution becomes a stable period 16 from 0.49800775 to 0.49800825, and is semiperiodic at 0.49800850 and 0.49800875, with apparently a stable period 24 at the latter value. At 0.49800900 the stable periodicity is gone. We have, in fact, found an extremely narrow semiperiodic band of b which evidently possesses the same structure as the semiperiodic bands of A noted in the previous section.

Lowering b to 1/4 and then 1/8, we encounter in Fig. 3.2 some possibly unexpected additional cusps. Points near or to the right of these cusps can be reached only after a nearly direct hit on the first cusp, and represent extremely rare events.

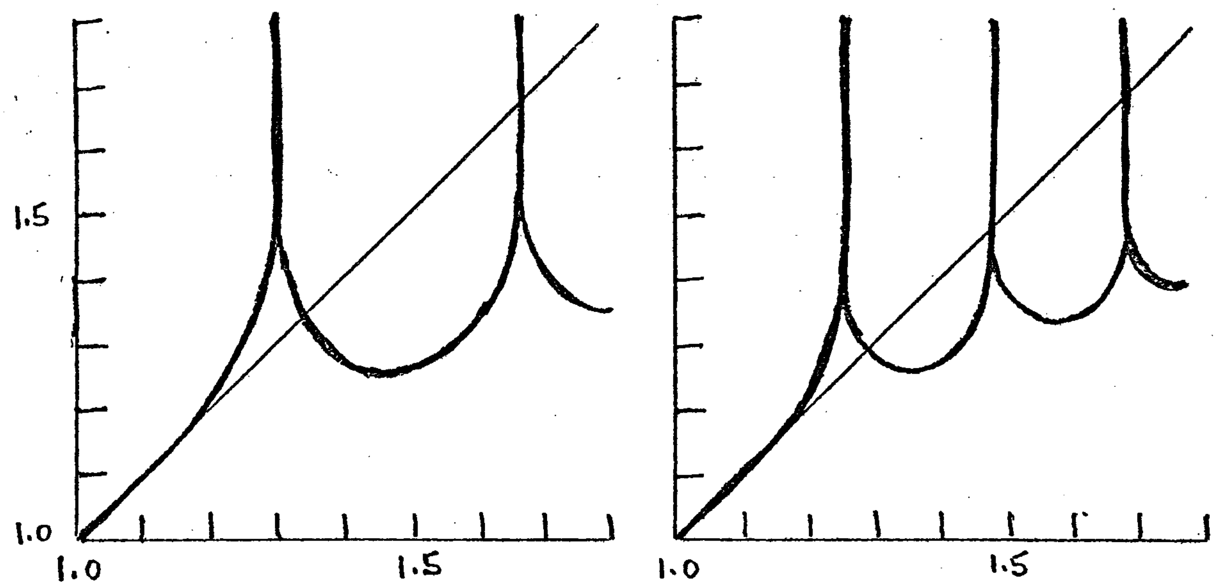

Fig. 3.2. Poincaré curves for Eqs. (1.1)-(1.3) for b = 1/4 (left) and b = 1/8 (right), with $\sigma = \sigma_m$ and $r = r_{cm}$.

For $b = 1/4$, points near the singularity at $Z = 0.145$ are preceded by points near the first cusp, but not so near as when $b = 1/2$. The semiperiodic bands near $b = 1/4$ should therefore be somewhat wider. We find, in fact, a stable period 3 from 0.2596 to 0.2609, becoming period 6 from 0.2609 to 0.2612. Another stable period 3 occurs from 0.2685 to 0.2697, becoming period 6 from 0.2697 to 0.2704. One periodic solution has one point slightly to the left of the cusp; the other has one slightly to the right.

Near $b = 1/8$ stable periodic solutions are abundant. In fact, at $b = 0.115$ the fixed point acquires a slope of -1, and for $b < 0.115$ period 1 is stable. Period 2 is stable from 0.120 to 0.135, period 4 is stable at 0.140, and at 0.145 the solution is semiperiodic with period 2. Periodicity disappears by 0.150. At 0.165 and 0.170 a second period 2 appears, becoming period 4 at 0.175 and 0.180. The solution is semiperiodic at 0.185, and aperiodic at 0.190.

We find, then, that the range of b from 0.1 to 1.0 is teeming with semiperiodic bands. It seems probable that, as with the difference equation (2.2), any two aperiodic values of b are separated by a semiperiodic band, in most cases very narrow. Below 0.25 most values of b are periodic, and above 0.5 most are aperiodic. In any event, aperiodicity is not an exceptional phenomenon, even below $b = 1.1$. Above $b = 1.1$ the singularity disappears, and all values of b are aperiodic.

Since period 2 is stable at $b = 1/8$, one may wonder how Fig 3.2 can show a curve instead of just two points. When the general solution is aperiodic, say at $b = 0.15$, an arbitrary orbit rapidly approaches the attractor set. At $b = 1/8$ such an orbit approaches a set which is essentially an analytic extension of the attractor set from higher values of b. Only somewhat later does it become trapped by the stable periodic orbit, which is the true attractor. Fig. 3.2 describes its behavior in the meantime.

A point on the Poincaré map corresponds to a segment of an orbit between two maxima of Z. For $b = 2$, as with Figs. 1.3 and 1.4, points to the right or left of the cusp correspond to segments which do or do not cross from one wing of the attractor (the region around C or C') to the other. The cusp corresponds to an orbit which makes a direct hit on the origin and terminates. Orbits emanating from the origin form a natural boundary for the attractor, and leave holes surrounding C and C'.

For b = 1 or b = 1/2, where a singularity occurs, the orbits emanating from the origin still form natural boundaries, but the holes at C and C' , as viewed from the X-axis, are bounded by another orbit which corresponds to the singularity. Lines parallel to the X-axis are tangent to the surface at the edge of a hole, and away from the edge there are two values of X for a given Y and Z , on orbits corresponding to points to the left and right of the singularity. As one continues around C or C' , these two orbits appear to merge, i.e., the curved surface becomes folded, while they also appear to join up with an orbit from the other wing.

At b = 1/4, points to the right of the second cusp correspond to orbits, including those emanating from the origin, which cross the X-Y and Y-Z planes and then cross back again in descending from a maximum of Z ; seen from the Z-axis they spiral downward. The second cusp itself corresponds to another direct hit on the origin; hence the two cusps must have equal heights. A topographic map of the attractor would, in some locations, have to show four values of X , which would be on orbits proceeding from a large maximum and a small maximum of Z in one wing of the attractor, and two distinct intermediate maxima in the other wing. At b = 0.4, where the second cusp first appears, the trajectory emanating from the origin would run back into it; above and below b = 0.4, the attractors are thus topographically distinct.

Strange attractors appear to be characteristic of forced dissipative systems with aperiodic general solutions, such as systems describing turbulent flow. Presumably the attractor can become stranger as the number of variables increases. What we have shown is that we need not go beyond three equations, nor even change the form of the equations, to find more complicated attractors than the one which we originally presented.

Acknowledgement. I wish to thank Dr. James Curry for many stimulating conversations, and for critically reading the manuscript. The work was supported by the Climate Dynamics Program of the National Science Foundation under Grant NSF-g 77 10093 ATM.

REFERENCES

1. V. Bjerknes, Das Problem der Wettervorhersage, betrachtet vom Standpunkt der Mechanik und der Physik, Meteorol. Zeit. (1904), 107.

2. J. Guckenheimer, On bifurcation of maps of the internal, Invent. Math., 39 (1977), 165-178.

3. G. Julia, Mémoire sur l'itération des fonctions rationelle, J. Math. Pures. Appl., 4 (1918), 47-245.

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

5. E.N. Lorenz, The problem of deducing the climate from the governing equations, Tellus, 16 (1964), 1-11.

6. E.N. Lorenz, An N-cycle time-differencing scheme for stepwise numerical integration. Mon. Weather Rev., 99, 644-648.

7. R.M. May, Simple mathematical models with very complicated dynamics, Nature, 261 (1976), 459-467.

8. L.F. Richardson, Weather prediction by numerical process. Cambridge Univ. Press (1922).

9. D. Ruelle and F. Takens, On the nature of turbulence, Commun. Math. Physics, 20 (1971), 167-192.

10. B. Saltzman, Finite amplitude free convection as an initial value problem -I, J. Atmos. Sci., 19 (1962), 329-341.

11. D. Singer, Stable orbits and bifurcation of maps of the interval, SIAM J. Appl. Math., 35 (1978), 260-267.

12. N. Wiener, Nonlinear prediction and dynamics. Proc. 3rd Berkeley Sympos. Math. Statistics and Probability, Vol. 3 (1956), 247-252.

MASSACHUSETTS INSTITUTE OF TECHNOLOGY