Designing Chaotic Models

EDWARD N. LORENZ

Massachusetts Institute of Technology, Cambridge, Massachusetts

(Manuscript received 14 July 2004, in final form 1 October 2004)

ABSTRACT

After enumerating the properties of a simple model that has been used to simulate the behavior of a scalar atmospheric quantity at one level and one latitude, this paper describes the process of designing one modification to produce smoother variations from one longitude to the next and another to produce small-scale activity superposed on smooth large-scale waves. Use of the new models is illustrated by applying them to the problem of the growth of errors in weather prediction and, not surprisingly, they indicate that only limited improvement in prediction can be attained by improving the analysis but not the operational model, or vice versa. Additional applications and modifications are suggested.

1. Introduction

The time-dependent behavior of a real or hypothetical physical system can in principle be duplicated by solving an appropriate set of mathematical equations. Sometimes we can find a general solution explicitly, but, with the advent of high-speed computers, we have gained a practical means for determining particular solutions of many otherwise intractable sets. In particular, we can deal with chaotic systems, that is, those where slightly different states generally evolve in due time into states that are perhaps unrecognizably different. Our preliminary task of establishing the equations, given the physical system, may nevertheless continue to entail some difficulties. Approximations are commonly needed, and the equations then become a model of the system.

When the system is sufficiently simple, the process of formulating the equations may be fairly straightforward. If, for example, the system is an ordinary pendulum, whose motion, incidentally, is not chaotic, two independent investigators may well arrive at the same set of equations, particularly if they both disregard friction and external driving and treat the entire mass as being concentrated at a point. One could then say that the investigators have simply discovered the equations that were waiting to be found.

For more involved systems a unique model may be a rarity; witness, for example, the multiplicity of operational and other models purporting to approximate the global circulation of the atmosphere. Each of these has retained certain physical features deemed by the investigator or group of investigators to be rather important, while omitting some thought to be less essential, and each employs devices, for example, representation by gridpoint values or spherical-harmonic coefficients, that appeal most to the investigator. Here one can say that the models have been designed rather than discovered; indeed, a model can bear the mark of its designer just as surely as a song can bear the mark of its composer.

When we construct operational global circulation models, we typically attempt to duplicate the atmosphere and its surroundings as closely as is feasible, given the present knowledge and the available computing facilities. On other occasions, perhaps in the interest of speeding the computation or simplifying the interpretation of the output, we often deliberately use something less than the best attainable approximation. Besides discarding some features outright, we often effectively remove others by simply reducing the spatial resolution. Sometimes, however, we replace the terms that describe a particular process by terms that cannot justifiably be said to do so; they simply possess some key properties in common with the terms that they replace. It is models of this sort that will occupy our attention now.

The suitability of a particular model depends critically upon the purpose for which it is to be used. Perhaps we may wish to gain some idea as to why a certain phenomenon occurs. In this case we must not use a model that will not produce the phenomenon, nor may we use one that forces the phenomenon to occur. For example, if we are interested in why the transport of angular momentum across middle latitudes in the at-

Corresponding author address: Edward N. Lorenz, Dept. of Earth, Atmospheric, and Planetary Sciences, Massachusetts Institute of Technology, Room 54-1622, Cambridge, MA 02139.
E-mail: jmsloman@mit.edu

mosphere tends to be poleward, we must not use a model where the orientation of sinusoidal trough and ridge lines is specified to be from south to north, producing no transport, or southwest to northeast, demanding a northward transport (cf. Starr 1948); evidently we must allow the orientation to vary. Naturally, with a simplified model we do not eliminate the possibility that a phenomenon will appear for the wrong reason.

Frequently, when one designs and describes a model, subsequent investigators will appropriate it, as often as not for applications that had not been envisioned by the original investigator. Established models may prove attractive because their key properties are known, or simply because they are ready to use, but it also seems likely that merely having appeared in print endows a model with a certain amount of legitimacy.

My starting point will be a very simple model that has been previously used to advantage (Lorenz 1996; Lorenz and Emanuel 1998, hereafter L96 and LE98, respectively), and that subsequent investigators (e.g., Hansen and Smith 2000) have found suitable for their needs. I shall begin by documenting its typical behavior.

As with any model, there are certain potential studies for which this one is not altogether suitable. One such study is the problem of the appearance and growth of errors in routine weather forecasting when both analysis error and model error are present. Here, of course, we need one model—the "perfect" model—just to produce the simulated true states, and a "model error" will mean the failure of a simulated operational forecasting model to duplicate the perfect one. I shall note certain shortcomings of the original model for pursuing this problem and describe my attempts to redesign it to eliminate them.

I shall include a brief application of the new models to the analysis error–model error problem, but the principal interest will be in the designing process itself. I hope that the new models will also satisfy some of the needs of future investigators.

2. Model I

The equations of the first model, which will be called Model I, are

$$dX_n/dt = -X_{n-2}X_{n-1} + X_{n-1}X_{n+1} - X_n + F. \quad (1)$$

The system is supposed to represent a one-dimensional atmosphere; F is a positive constant, t is time, and the N variables $X_1, \ldots, X_N$ are to be identified with the values of some unspecified scalar atmospheric quantity at N equally spaced points about a latitude circle, which will be called grid points, even though the "grid" is one-dimensional. Thus, when $n < 1$ or $n > N$, X_n will mean X_{n+N} or X_{n-N}. The symbol X, without a subscript, will frequently be used to denote the set of variables X_n.

The quadratic, linear, and constant terms are supposed to represent advection, thermal or mechanical damping, and external forcing. Note that no coefficients appear in the quadratic and linear terms; they were removed by scaling to minimize the amount of computation. The time unit is therefore the damping time, assumed to be five days. As we shall see, a typical cross-longitude profile of X will consist of an irregular chain of waves, if N and F are sufficiently large.

Equation (1) cannot be derived from any dynamic equation of which I am aware; it is the result of an attempt to formulate the simplest possible set of N dissipative chaotically behaving differential equations that is unchanged by cyclically permuting the variables. It nevertheless shares certain properties with the atmosphere, or at least with some much larger atmospheric models. There follow its relevant properties, many of which were documented in LE98.

Let r and s^2 denote averages of X_n and X_n^2 over n and regard $s^2/2$ as the total energy. Then

$$d(s^2/2)/dt = -s^2 + Fr; \quad (2)$$

the quadratic terms in Eq. (1) have canceled and, like advection in many atmospheric models, they do not add or remove energy. Since $s \geq r$, values of X that make $s > F$ will make the right side of Eq. (2) negative, that is, they will make s^2 decrease; hence s and r can never exceed F after transients have died out.

If R and S^2 denote long-term averages of r and s^2; that is, averages of X_n and X_n^2 over n and over a long enough time to make average time derivatives negligible, the long-term variance σ^2 of X is $S^2 - R^2$. From Eq. (2), $S^2 = FR$, whence $\sigma^2 = R(F - R)$. Since variances are nonnegative, R must lie between 0 and F and, when $R = F/2$, σ acquires its greatest possible value $F/2$. The only obvious steady solution, $X_n = F$, makes $R = F$, $S = F$, and $\sigma = 0$.

Since the forcing F is constant, it cannot directly inject irregularities into the profile of X. Instead, its effect is to increase X_n uniformly, and hence to increase the mean r. However, if waves are nearly absent, so that the values of X_n are all close to r, the small departures x_n of X_n from r will approximately satisfy the linear perturbation equation

$$dx_n/dt = r(x_{n+1} - x_{n-2}) - x_n. \quad (3)$$

If $x_n = \exp(ht)\cos(kn - mt)$, so that the perturbation is sinusoidal, with a wavelength of $2\pi/k$ gridpoint spacings,

$$h = r(\cos k - \cos 2k) - 1. \quad (4)$$

The factor $\cos k - \cos 2k$ assumes its maximum value 9/8 when $\cos k = 1/4$, or $2\pi/k = 4.77$, so that waves of a suitable length will grow if $r > 8/9$. The actual wavelength must be a divisor of N; this will happen for a length fairly close to 4.77 points if N is fairly large, and finite-amplitude waves will develop when r slightly ex-

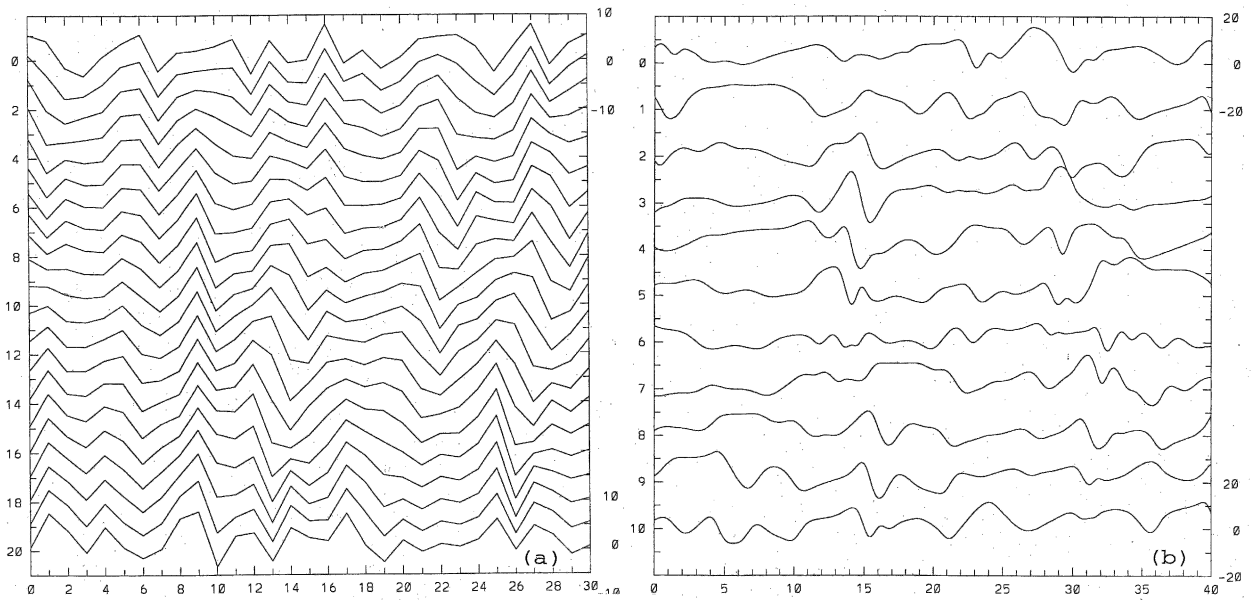

FIG. 1. (a) Profiles of X produced by Eq. (1) with $N = 30$ and $F = 10$, at 6-h intervals for 5 days. Scale at bottom is gridpoint number. Numbers at upper and lower right are scales for top and bottom profiles. Numbers at left indicate chronological order of profiles. Each number is placed on the zero line for its profile; note that most profiles begin with $X_0 > 0$. (b) Time series of $X_0, \ldots, X_{10}$ produced by Eq. (1) with $N = 30$ and $F = 10$ extending 40 days. Scale at bottom is time in days. Numbers at upper and lower right are scales for X_0 and X_{10}. Numbers at left are gridpoint numbers.

ceeds 8/9. This in turn will happen eventually if F exceeds 8/9.

Although we are interested mostly in positive values of F, it is worth noting that $\cos k - \cos 2k$ assumes its greatest negative value -2 when $k = \pi$, or $2\pi/k = 2$. Waves about 2 grid intervals long will therefore develop if $r < -1/2$; this must happen eventually if $F < -1/2$.

The remaining properties of Model I are most readily found by numerical integration, where the standard fourth-order Runge–Kutta scheme with a time step of 1/40 unit or 3 h will invariably be used. This requires choosing values for N and F. The case $N = 3$ is of little interest since the quadratic terms cancel. In a work performed before Eq. (1) was formulated for general values of N (Lorenz 1984), with a system equivalent to Eq. (1) with N fixed at 4, I found that chaos would appear with $F > 11.84$ or $F < -60.7$ and studied the case $F = -100$ in detail (y_n and c in that work equal $-X_{-n}$ and $-F$ in the present one). Here, only positive values of F and larger values of N are considered. Values of F slightly exceeding 8/9 produce periodic behavior but, as long as $N \geq 12$, chaos is found when $F > 5$.

In the principal example, $N = 30$ and $F = 10$, giving 12° longitudinal resolution (we used $N = 40$ and $F = 8$ in LE98). To eliminate transient effects each variable was first chosen randomly from a uniform distribution between 0 and 1, and then integrated for 2 yr. Here a "year" consists of twelve 30-day months, or 72 time units, or 2880 time steps.

Figure 1a shows a sequence of cross-longitude profiles of X, which are complete states, for five days at 6-h intervals. The straight-line segments connecting successive computed values have been introduced solely to show which point lies on which profile; they do not imply that X varies linearly between grid points or is even defined there. The initial profile (at top) contains eight distinct crests and troughs—reasonably close to the six or seven that might have been expected from the theoretical 4.77-point wavelength. The major crests and troughs generally maintain their identities through the five days, although the trough initially at grid point 19, for example, disappears by day 4. Minor crests and troughs frequently appear and disappear. The irregularity of the profiles, and their irregular modulation as time advances, suggests chaos.

Figure 1b is like Fig. 1a, but with the axes of n and t interchanged. The curves are time series of the consecutive variables, $X_0, \ldots, X_{10}$, extending for 40 days; the left eighth of Fig. 1b and the left third of Fig. 1a contain the same information. The curves are smooth but irregular, with maxima and minima occurring every few days and major extremes farther apart. Consecutive curves do not appear highly correlated, and computations indicate that the long-term correlation coefficients at spatial lags of 1 through 5 grid points are 0.05, −0.33, −0.11, 0.03, and 0.05.

To quantify the chaos inherent in Eq. (1) we go to Fig. 2a, which shows error-growth curves over a 40-day period, for $N = 30$, for selected values of F. In each case a 2-yr integration from random conditions has produced the initial "true" state. For the initial "assumed"

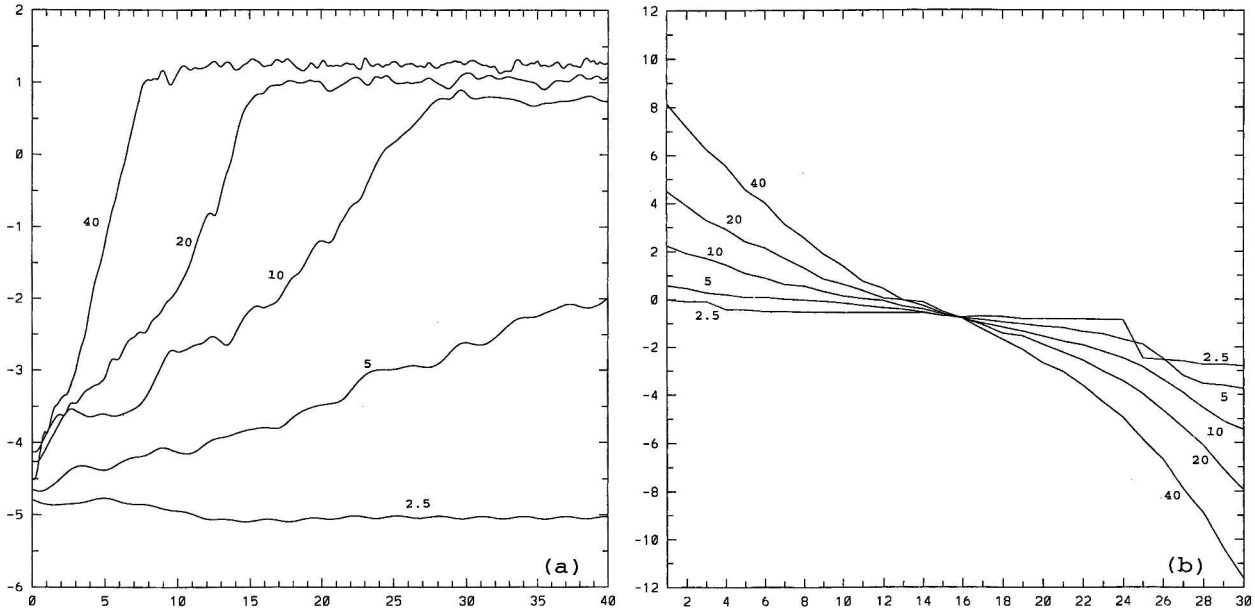

FIG. 2. (a) Growth curves for initially small errors in X extending 40 days, each produced by a single pair of solutions of Eq. (1), with $N = 30$ and with values of F indicated by numbers beside curves. Scale at bottom is time in days. Scale at left is base-10 logarithm of root-mean-square error. (b) Lyapunov exponents (natural logarithms of Lyapunov numbers) of Eq. (1) with $N = 30$ and with values of F indicated by numbers beside curves. For each F, values of successive exponents are shown connected by line segments. Scale at bottom is serial number of exponent. Scale at left is value of exponent.

state a randomly chosen value from a uniform distribution extending from -0.0001 to 0.0001 has been added to each variable. The quantity shown is the root-mean-square difference between the true and assumed values of the 30 variables.

For $F = 2.5$ the solution is not chaotic and the curve initially descends, leveling off when presumably the only remaining error is a phase error. At $F = 5$ the threshold for chaos is just exceeded, and by day 40 the curve is still far short of its limiting value, or saturation, but each remaining curve amplifies quasi-exponentially at its own rate and reaches saturation before 40 days, thereafter proceeding horizontally. As expected, larger values of F produce faster error growth. For $F = 10$ the slope before saturation implies a doubling time of about 1.5 days—a reasonable simulation of the behavior of synoptic-scale errors in the atmosphere (Simmons et al. 1995). By averaging many cases, smoother curves could have been produced.

Figure 2b shows the complete set of Lyapunov exponents for the same values of F. For each F the exponents have been evaluated from one integration that uses the same true initial state as in Fig. 2a and extends for two more years, and 30 additional integrations in each of which a single variable is perturbed initially.

The most striking feature, aside perhaps from the smooth progressions from the largest to the smallest exponent, is the common intersection of the curves, at a point separating the largest 15 exponents from the smallest 15. For $F = 2.5$, where the solution is periodic with a chain of six waves, the leading exponent is 0 and the last six exponents fall well below the others. For $F > 2.5$ the number of positive exponents increases from 6, when $F = 5$, to 12, when $F = 40$. The model might therefore appeal to someone requiring a system whose fractional dimension is high, at least when evaluated by the formula of Kaplan and Yorke (1979). The leading exponent 2.2 when $F = 10$ corresponds to a 1.6-day doubling time, in good agreement with the smaller sample of Fig. 2a.

There remains the question of the values of R and S^2, for which analytic reasoning yielded only upper and lower bounds. Extended computations through the chaotic range up to $F = 100$ reveal an almost perfect linear relationship between $\log R$ or $\log S$ and $\log F$, and, very closely, $R = a^2F^{1/3}$ and $S^2 = a^2F^{4/3}$, with a^2 close to 1.2. I know of no analytic explanation for this finding.

If in Eq. (1) $F = f^{-3}$, and then $t = f^2\tau$ and $X_n = f^{-2}Y_n$, we obtain

$$dY_n/d\tau = -Y_{n-2}Y_{n-1} + Y_{n-1}Y_{n+1} - f^2Y_n + f. \quad (5)$$

With this rescaling, arbitrarily large values of F can be treated without computational difficulty by making f arbitrarily small, whereupon the mean of Y_n should approach 0 while the mean of Y_n^2 should remain nearly independent of f. In the limit we encounter a nondissipative system.

3. Introducing spatial continuity: Model II

With $N = 30$ and $F = 10$, Eq. (1) simulates typical atmospheric wavelengths and error-growth rates rea-

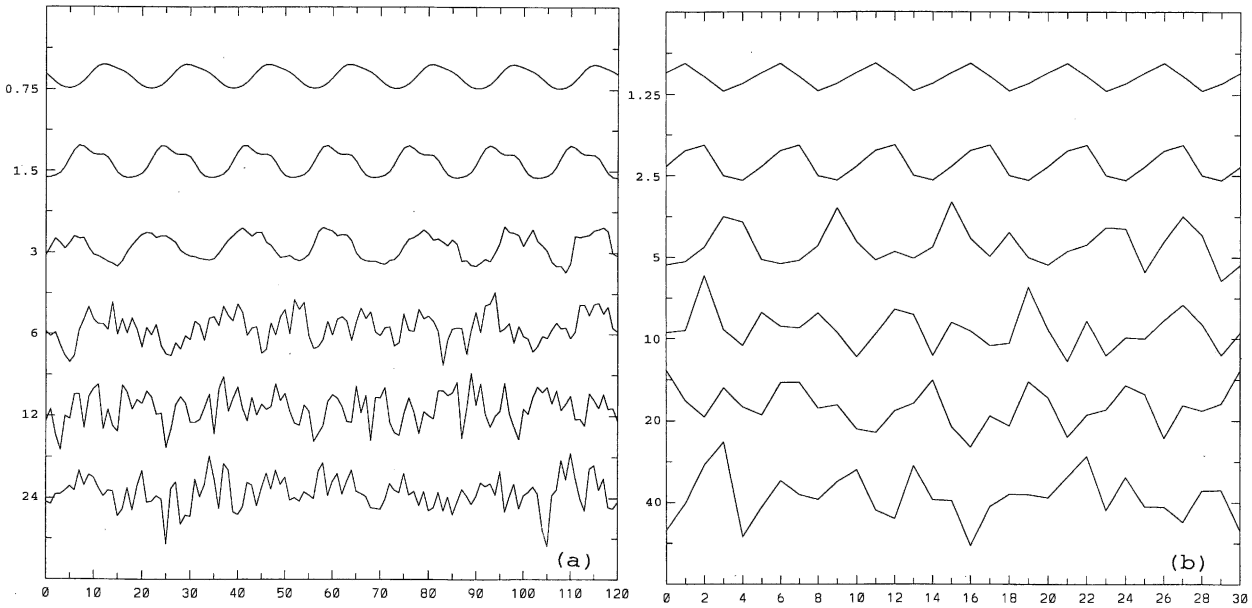

FIG. 3. (a) Profiles of X produced by Eq. (6) with $N = 120$, $j = 8$, $i = 7$, and values of F indicated by numbers at left. Scale at bottom is gridpoint number. (b) Same as (a) but with $N = 30$, $j = 2$, and $i = 1$.

sonably well, but there are applications where the 12° longitudinal resolution would be quite inadequate. In the analysis error–model error problem, for example, one would need to begin with observations that, in the real world, would ordinarily fall between grid points. If one planned to use Eq. (1) with $N = 30$ as the operational model, one would first have to use Eq. (1), perhaps with a new value of F, or else some other equation approximating Eq. (1), as the perfect model, to produce true gridpoint values. One would then encounter the problem of interpolating between these values to produce observations—a task not faced in the real world.

It is always possible to introduce some interpolation scheme as part of the perfect model, and in certain applications this might be the most satisfactory solution. Wholly apart from any particular problem, however, it may be difficult to identify a truly good scheme for interpolating between quantities like X_n and X_{n+1}, which are almost as poorly correlated as random numbers. What would be preferable is a system that will produce rather smooth variations from one longitude to the next, so that almost any reasonable interpolation scheme, and perhaps even linear interpolation, will give nearly the same result. Equivalently, one would like the system to make values at consecutive grid points rather highly correlated. This implies that the typical wavelength must be much greater than four grid intervals. If chains of six or more waves are to be retained, this demands a much larger value of N.

Retaining Eq. (1) and increasing N will not solve our problem. The waves will still be about four grid intervals long; there will simply be more waves. It follows that, if Eq. (1) with a large value of N has been chosen as the perfect model, Eq. (1) with a smaller value of N cannot constitute a realistic operational model. For some applications, then, a new equation is needed.

False starts are common in model designing, and this account would be incomplete if I did not demonstrate how a plausible assumption can lead to one. One way to construct a new system would be to introduce two numbers, i and j, and replace Eq. (1) by

$$dX_n/dt = -X_{n-i}X_{n-j} + X_{n-i+j}X_{n+j} - X_n + F. \quad (6)$$

Note that the quadratic terms still do not add or remove energy and that Eq. (6) becomes Eq. (1) if $i = 2$ and $j = 1$. If instead $i = 8$ and $j = 7$, for example, the coefficient of r in Eq. (4) is replaced by $\cos k - \cos 8k$, which assumes its maximum value 1.925 when the wavelength $2\pi/k$ equals 16.24 grid lengths. The steady solution is therefore unstable when F exceeds 0.519, and, if one chooses $N = 120$, a chain of seven or eight waves should develop.

The problem may seem to be solved, but Fig. 3a shows what actually happens. The curves are profiles produced by Eq. (6) with $N = 120$, $i = 8$, and $j = 7$ for the indicated values of F. Each curve has been normalized by division by $F^{2/3}$. For $F = 0.75$, a bit above the value where waves first appear, one sees the expected chain of seven nearly sinusoidal waves, while the waves are more distorted when $F = 1.5$ and especially when $F = 3$. At $F = 6$ chaos has set in, and the seven waves are scarcely detectable, while shorter waves are very active, and in the final two profiles the short waves completely dominate. Interpolation is therefore no more feasible, when chaos is well developed, than with Eq. (1), and the new system fails to serve its intended purpose.

For comparison, Fig. 3b shows similar profiles produced by Eq. (1) with $N = 30$. Beginning with $F = 1.25$, again shortly beyond where waves can develop, we see six waves; a different initial state, incidentally, would have produced seven. Again the waves become more irregular as F increases and chaos ensues, but this time there is only a modest increase in the prevailing wavenumber. Linearization about the mean value seems to predict the general appearance of the chaotic regime fairly well when $i = 2$ and $j = 1$, but it proves worthless when $i = 8$ and $j = 7$.

The seeming success of the linearization procedure in the one case may be fortuitous. In either figure, the waves seem to acquire a length of somewhat over three grid intervals when F becomes fairly large. This is close enough to the 4.77 intervals predicted for Fig. 3b for the procedure to seem moderately good, but it completely fails to resemble the 16.24 intervals predicted for Fig. 3a.

Equation (6) has been tested with other values of i and j, with no success. Apparently the simplicity of Eqs. (1) or (6) has to be abandoned, and I shall do this by replacing the pair of products in Eq. (6) by the sum of many such pairs of products. A possible advantage to this scheme is that many of the same products can be made to appear in the derivatives of X_n and X_{n+1}. Conceivably the derivatives would then be correlated, and X_n and X_{n+1} might also be correlated.

There are undoubtedly many effective ways to choose the products. For the method chosen, which is influenced by the contemplated application, one first introduces the symbol Σ' to denote a modified summation, like the ordinary summation that is denoted by Σ except that the first and last terms are to be divided by 2. One chooses a number K, much smaller than N and let $J = K/2$ if K is even and $J = (K - 1)/2$ if K is odd. Then, for any two sets of variables X and Y, one defines

$$[X, Y]_{K,n} = \sum_{j=-J}^{J}{}' \sum_{i=-J}^{J}{}' (-X_{n-2K-i}Y_{n-K-j} + X_{n-K+j-i}Y_{n+K+j})/K^2 \quad (7)$$

if K is even, with Σ' replaced by Σ if K is odd. The equation for Model II, where the only set of variables is X, will be

$$dX_n/dt = [X, X]_{K,n} - X_n + F. \quad (8)$$

Note that setting $K = 1$ makes $J = 0$; hence $[X, X]_{1,n}$ represents the single pair of products appearing in Eq. (1). Model II then reduces to Model I.

In numerical computations one lets

$$W_n = \sum_{i=-J}^{J}{}' X_{n-i}/K, \quad (9)$$

whereupon

$$[X, X]_{K,n} = -W_{n-2K}W_{n-K} + \sum_{j=-J}^{J}{}' W_{n-K+j}X_{n+K+j}/K. \quad (10)$$

Again Σ should replace Σ' if K is odd. Since W_{n+1} may be evaluated rather quickly by subtracting and adding terms from and to W_n, while the summation in Eq. (10) may be similarly evaluated from the previous summation, integration of Eq. (8) may be accomplished with only about four times as much computing, per variable, as is required for Eq. (1), but, in view of the need to use more variables, the entire integration process can be considerably slower.

Figure 4a shows typical profiles produced by Eq. (8) when $N = 240$ and $F = 10$ for selected values of K. When $K = 2$, there are nearly as many waves as if Eq. (1) had been retained. Increasing K to 4 decreases the number, but there are still too many compared with Fig. 1a. When $K = 8$, one evidently succeeds in producing an acceptable number of major waves, although weaker smaller-amplitude waves are superposed. In drawing the curve I have, as usual, connected the successive values of X_n with straight-line segments, but these are hard to detect. Any other reasonable interpolation procedure would have produced an indistinguishable curve. Increasing K to 16, 32, or 64 lengthens the waves still more, and, evidently, one can produce any wavenumber desired by choosing K judiciously.

Since the ratio N/K is 30 in the third profile, whose dominant wavenumber agrees most closely with Fig. 1a, where $N = 30$, there is a suggestion that the appearance of a profile may depend largely upon N/K. Figure 4b is constructed with $F = 10$ and $N/K = 30$ in each profile, and with N successively doubling from 30 in the leading profile to 960 in the final one. The conjecture seems to be well supported; the profiles in Fig. 4b show little resemblance to any profile in Fig. 4a except the third one.

With $N = 960$ and again with $K = 32 = N/30$ and $F = 10$, Fig. 5a has been constructed in the manner of Fig. 1a; it shows profiles produced by Eq. (8) at 6-h intervals for five days. Again, at least for the five days, the major crests and troughs retain their identities, while minor ones come and go. One can conclude that Model II is ready for some applications for which Model I would have been inadequate.

For Model I the doubling time for small errors, as seen in Fig. 2a, depends strongly upon F, but is nearly independent of N if N is not too small. For Model II, with $K > 1$, it also proves to depend strongly upon F while being nearly independent of N and K if N/K is not too small, but, for a given value of F, it is much smaller when $K > 1$ than when $K = 1$. Thus, for the values used in Fig. 5a, the doubling time is about four days—considerably longer than expected in the atmosphere. It can be restored to a more nearly atmospheric value by increasing F.

Figure 5b is constructed like Fig. 5a, again with $N =$

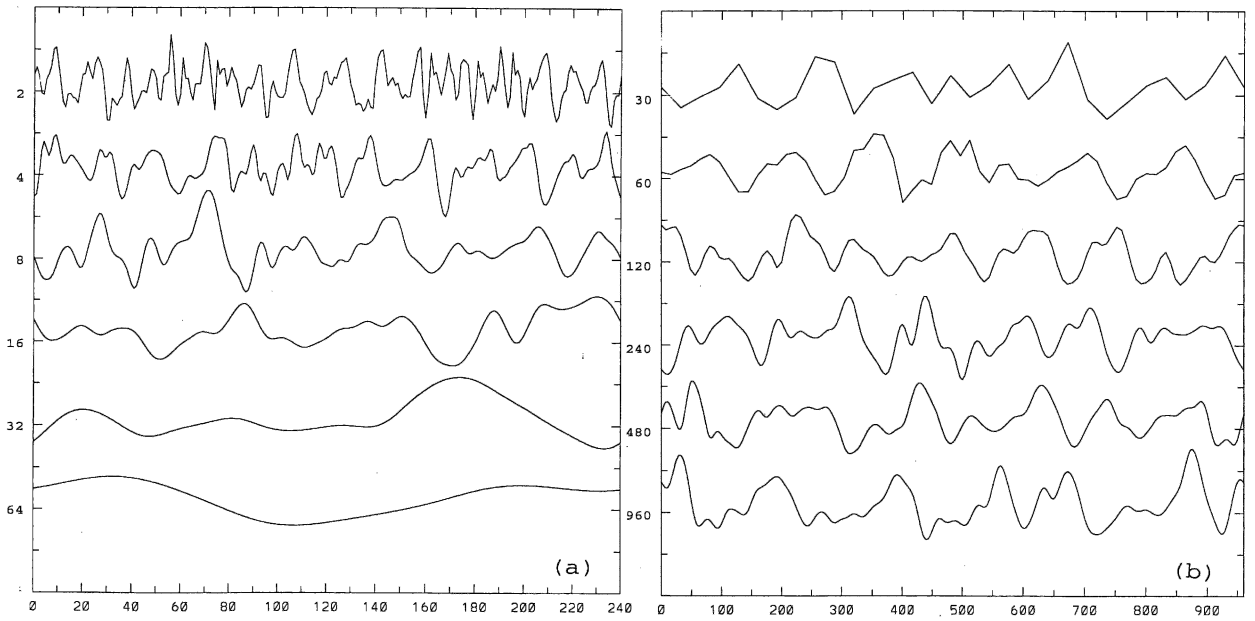

FIG. 4. (a) Profiles of X produced by Eq. (8) with $N = 240$, $F = 10$, and values of K indicated by numbers at left. Scale at bottom is gridpoint number. (b) Profiles of X produced by Eq. (8) with values of N indicated by numbers at left and with $K = N/30$ and $F = 10$. Scale at bottom is gridpoint number for bottom curve.

960 and $K = 32$, but with $F = 15$. There is still a suggestion of six or seven longer waves, but the shorter waves are more in evidence. Note that, with N so large, even these shorter waves are 30 or more grid intervals long—the point-to-point variations are very smooth. The doubling time has been reduced to about two days.

Apparently, in trying to make the curves produced by Model II look like reasonable spatial interpolations of the kind of curve produced by Model I, one must choose between too long a doubling time (smaller F) or unanticipated shorter waves (larger F). The value $F = 15$ is a compromise.

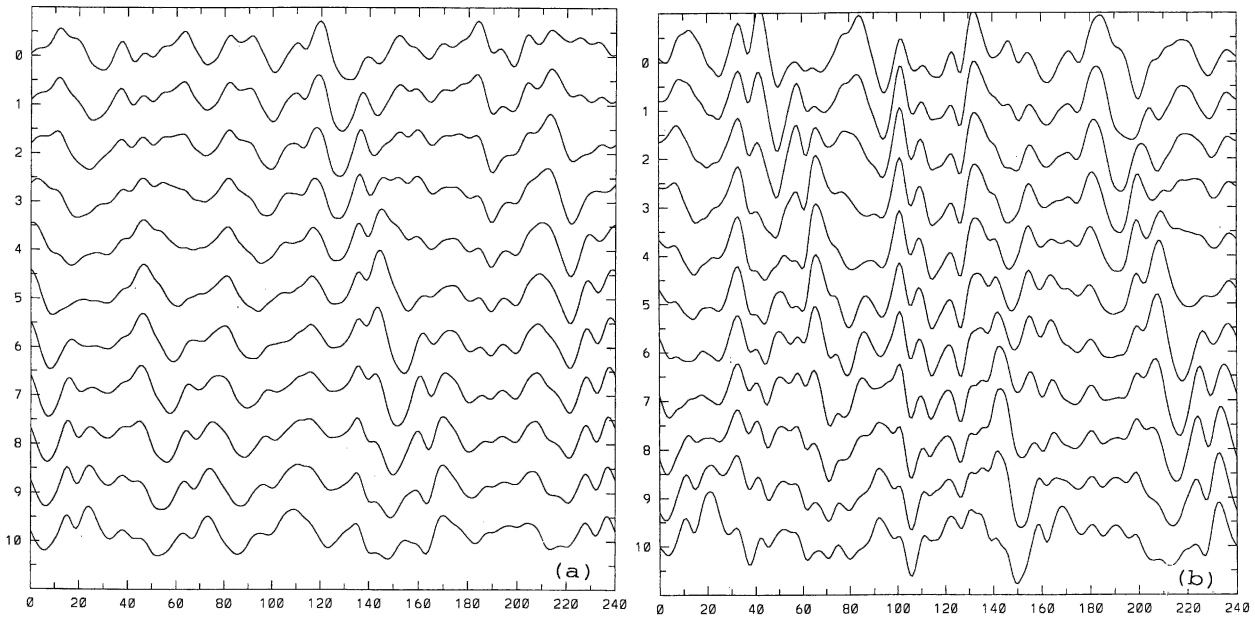

FIG. 5. (a) Profiles of X produced by Eq. (8) with $N = 240$, $K = 8$, and $F = 10$, at 12-h intervals for 5 days. Scale at bottom is gridpoint number. Numbers at left indicate chronological order of profiles. (b) Same as (a) but with $F = 15$.

4. Introducing multiple scales: Model III

In the forthcoming application, I shall assume that the observations are free of instrumental error, but may nevertheless be unrepresentative. In the real world this can happen when, for example, spatially small-scale activity is superposed on larger-scale synoptic features. The local values that the instruments record are sums of large-scale and small-scale contributions, and the large-scale contribution alone, which one may wish to use in an analysis, is not always readily extracted.

Model III will be designed to include both large-scale and small-scale activity with waves much shorter than the shorter ones in Fig. 5b. Before introducing it, I shall describe another system that includes both large and small scales; it is a modification of one presented in L96, which, like Model I, has been used to advantage by other investigators (e.g., Boffetta et al. 1998). One introduces sets of variables X and Y, each defined at the same set of N longitudes, and attempt to formulate the equations so that X will exhibit mainly large-scale activity, while Y will contain the small scales.

Temporarily let X be governed by Eq. (8) with N large and N/K fairly large. Next note that solutions of the equation

$$dY_n/dt = b^2[Y, Y]_{1,n} - bY_n + F \tag{11}$$

are identical to those of Eq. (1) except that the variables fluctuate b times as rapidly and their amplitude is reduced by the factor b. Waves a few grid intervals long must still predominate. With a fairly large value of b, this is precisely the way that one would like the small-scale variables to behave.

Temporarily let Y be governed by Eq. (11). Then, following L96, couple the variables by introducing linear terms that together do not alter the large-scale plus small-scale energy, and omit the external forcing term in Eq. (11), letting the small-scale activity be driven entirely by the coupling. One obtains the system

$$dX_n/dt = [X, X]_{K,n} - X_n - cY_n + F, \tag{12a}$$

$$dY_n/dt = b^2[Y, Y]_{1,n} - bY_n + cX_n, \tag{12b}$$

where, like b, the coupling coefficient c is an additional parameter of the model.

Note that the coupling term cX_n in Eq. (12b)—the driving for Y—is a large-scale term; hence, like the forcing in Eq. (1), it cannot directly produce small-scale waves. Instead, wherever it is positive at a considerable number of consecutive grid points, its effect is to increase Y_n at these points and, therefore, to increase the mean of Y_n over the same points. If small-scale activity is nearly absent, the departures of Y_n from their local mean will obey an equation much like Eq. (3), with a derived equation much like Eq. (4), and in due time waves with lengths of several grid intervals should develop. Similarly, where cX_n is negative at many consecutive grid points, waves about two grid intervals long should eventually develop. The larger the value of N, the greater the disparity between the lengths of the long and short waves, that is, the stronger the spectral gap.

Equations (12) nevertheless possess an unrealistic property that disqualifies them for the desired Model III. The large-scale and small-scale features are represented by separate sets of variables X and Y instead of appearing as superposed features of a single set, say Z. What one would like is first a procedure for expressing given variables Z_n as sums of two quantities, say X_n and Y_n, whose respective profiles will consist mainly of long waves and short waves. One then needs an equation governing Z_n that will make the quantities X_n and Y_n derived from Z_n behave somewhat as X_n and Y_n do in Eqs. (12).

The former task is the easier. One could subject the profile of Z to a complete Fourier analysis at each time step and let X be the sum of the longest-wave components, while Y is the sum of the remaining ones. I balk at such a procedure, which would be computationally costly and would remove from the model much of what simplicity still remains.

Instead, one can introduce a number I and a pair of constants α and β and let

$$X_n = \sum_{i=-I}^{I}{}'(\alpha - \beta|i|)\, Z_{n+i}, \tag{13a}$$

$$Y_n = Z_n - X_n. \tag{13b}$$

One wishes to choose I, α, and β so that X will be effectively a smoothed version of Z with the short waves filtered out, whereupon these waves will appear in Y.

With a sufficiently large spectral gap, whose occurrence will depend upon the equation still to be chosen, one can do this by choosing α and β so that X_n will equal Z_n whenever Z varies quadratically over the longitudes $n - I$ through $n + I$; X_n will do so if $\Sigma'(\alpha - \beta|i|) = 1$ and $\Sigma' i^2(\alpha - \beta|i|) = 0$, whereupon

$$\alpha = (3I^2 + 3)/(2I^3 + 4I), \tag{14a}$$

$$\beta = (2I^2 + 1)/(I^4 + 2I^2). \tag{14b}$$

Waves whose lengths are exact divisors of I will be completely eliminated from X and will therefore show up in Y, while waves of comparable lengths will be largely eliminated. If I is too large, there may not be many intervals of length $2I$ where Z_n varies nearly quadratically after smoothing. A value of I between 10 and 20 may thus be optimal. The separation of scales will not be complete, but it need not be; it is also not complete in Eqs. (12).

A procedure for constructing a governing equation for Z that naturally suggests itself consists of formally adding two equations like Eqs. (12a) and (12b). However, adding Eqs. (12) as they stand will not suffice. The direct effect of the coupling term cX_n in Eq. (12b), if the equations are not added, is to inject long waves into the profile of Y; these will then enable short waves to de-

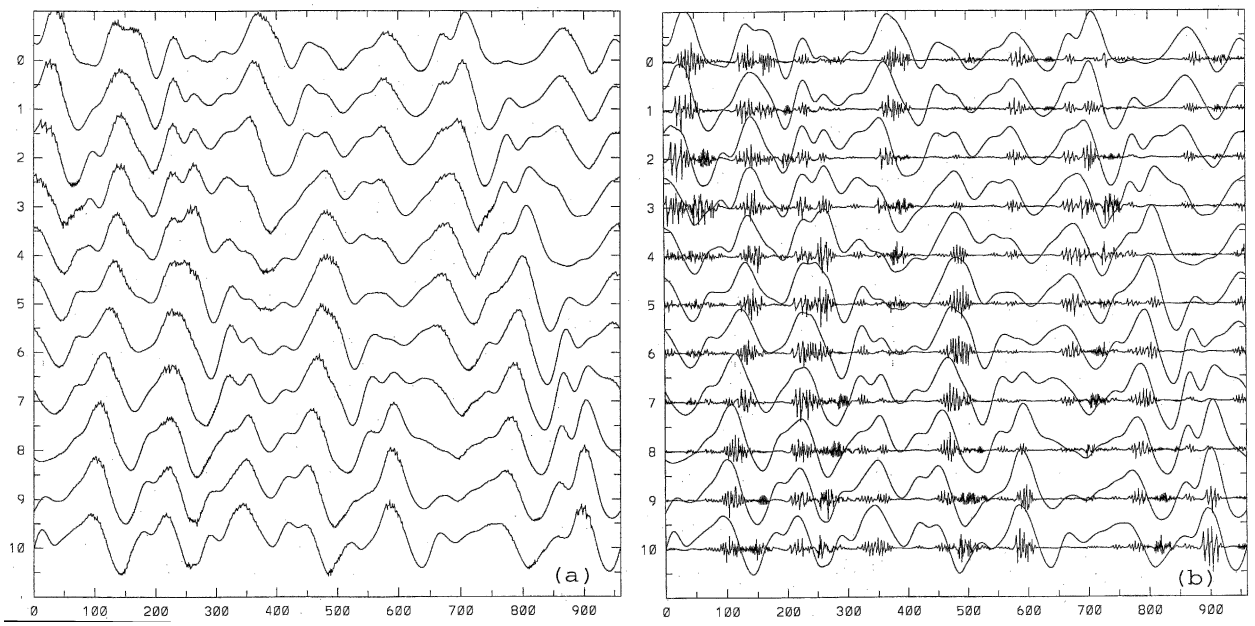

FIG. 6. (a) Profiles of Z produced by Eq. (15) with $N = 960$, $K = 32$, $J = 12$, $F = 15$, $b = 10$, and $c = 2.5$, at 12-h intervals for 5 days. Scale at bottom is gridpoint number. Numbers at left indicate chronological order of profiles. (b) Same as (a) but with superposed profiles of X (the smoother curves) and Y and with Y profiles vertically stretched four times.

velop. If instead the equations are added, these long waves will be injected into the profile of $X + Y$, if indeed, the coupling term is not completely canceled by being added to the dissipation term $-X_n$ in Eq. (12a). Presumably they will appear in X rather than Y when Z is next analyzed, and there will be nothing to enable the short waves in the profile of Y to grow.

The coupling process needs to be reformulated. One way to do this would be to redefine Y to include a small fraction, say c', of the long waves that are presently allocated to X; short waves in Y could then amplify. Equivalently, one can simply replace Y_n by $Y_n + c'X_n$ in the quadratic terms in Eq. (12b). The same purpose proves to be accomplished by replacing $[Y, Y]_{1,n}$ by $[Y, Y + c'X]_{1,n}$ rather than $[Y + c'X, Y + c'X]_{1,n}$. Upon adding the new equations, and letting $c = b^2c'$ be the new coupling coefficient, one obtains, for Model III,

$$dZ_n/dt = [X, X]_{K,n} + b^2[Y, Y]_{1,n} + c[Y, X]_{1,n} - X_n - bY_n + F. \quad (15)$$

Note that $\alpha = 1$ and $\beta = 1$ when $I = 1$, whereupon $X_n = Z_n$ and $Y_n = 0$. Model III then reduces to Model II.

Figure 6a has been constructed from Model III with the format of Figs. 1a and 5 with $N = 960$, $K = 32$, $I = 12$, $F = 15$, $b = 10$, and $c = 2.5$, and it shows profiles of Z at 12-h intervals for five days. The major crests and troughs maintain their identities for several days, and sometimes for all five. The intermediate-scale waves that were so prominent in Fig. 5b are considerably weaker, but very short waves appear at many of the crests and troughs of the long waves. Figure 6b shows superposed profiles of the quantities X and Y into which Z has been analyzed; the Y profile has been vertically stretched four times. The X profiles appear smooth, while the regions where Y is active tend to follow the crests and troughs of X.

In Fig. 7a a portion of Fig. 6a is enlarged near the lower left corner; the top, middle, and bottom curves are 45° segments (grid points 80–200) of the bottom three profiles of Z in Fig. 6a, which span 24 h, while in between profiles of Z at 4-h intervals have been inserted. The contrasting scales are quite evident. Figure 7b shows the corresponding profiles of X and Y, this time with no vertical stretching of Y. The smoothness of the X profiles attests to the effectiveness of the filtering procedure. The Y waves are patently shorter in the trough of the X waves than on the crest. Individual crests and troughs in the short waves often do not last through the 24 h, in contrast to the regions of short-wave activity, which can persist as long as the long-wave crests and troughs.

Figure 8a shows superposed separate error-growth curves for X and Y; each has the format of the curves in Fig. 2a. The first five days have been stretched horizontally in Fig. 8b. The initial small error was placed entirely in Y; it undergoes immediate rapid growth, doubling in less than 3 h until it approaches saturation before two days. Meanwhile, an error in X appears and increases equally rapidly, presumably because of its coupling with the error in Y but, after the Y error reaches saturation, the X error proceeds to amplify at

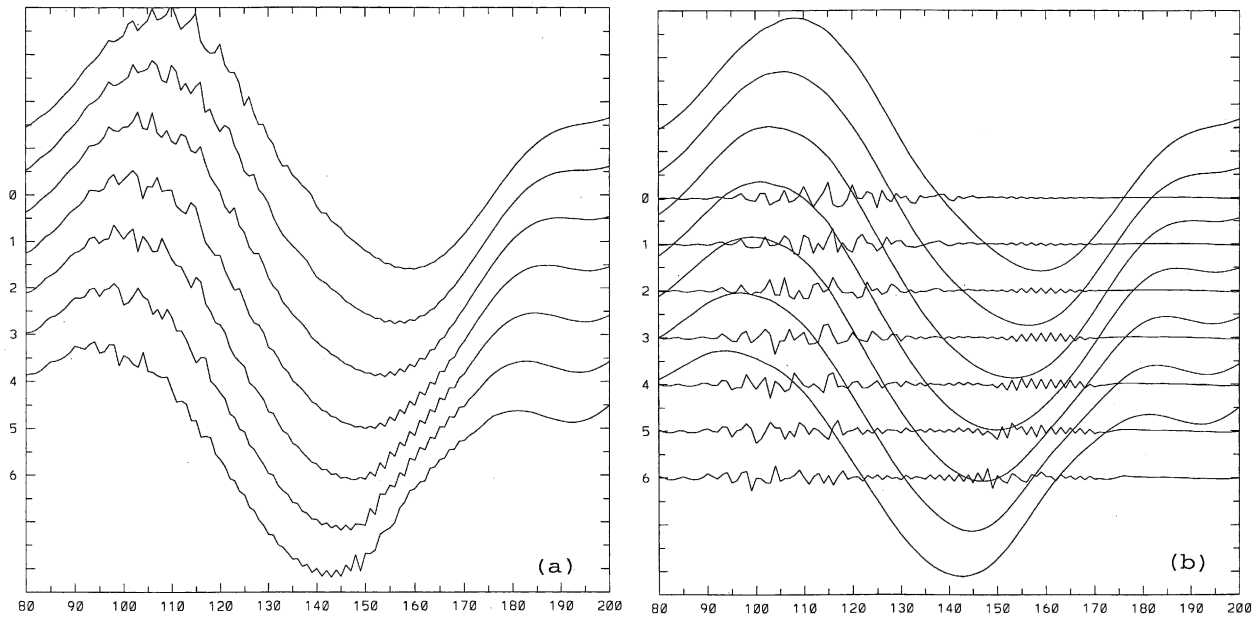

FIG. 7. (a) As in Fig. 6a but for 45° sectors of profiles of Z, at grid points indicated by scale at bottom, at 4-h intervals for final day of Fig. 6. (b) As in (a) but for superposed 45° sectors of profiles of X (the smoother curves) and Y and without vertical stretching of Y.

its own rate, first doubling in a day or so and then growing more slowly, reaching saturation in about 25 days. I conclude that Model III possesses a number of potentially desirable properties that are not found in Models I and II.

5. An application

I shall now illustrate the effectiveness of Models II and III by applying them to a problem that was in mind when designing them—the behavior of errors in opera-

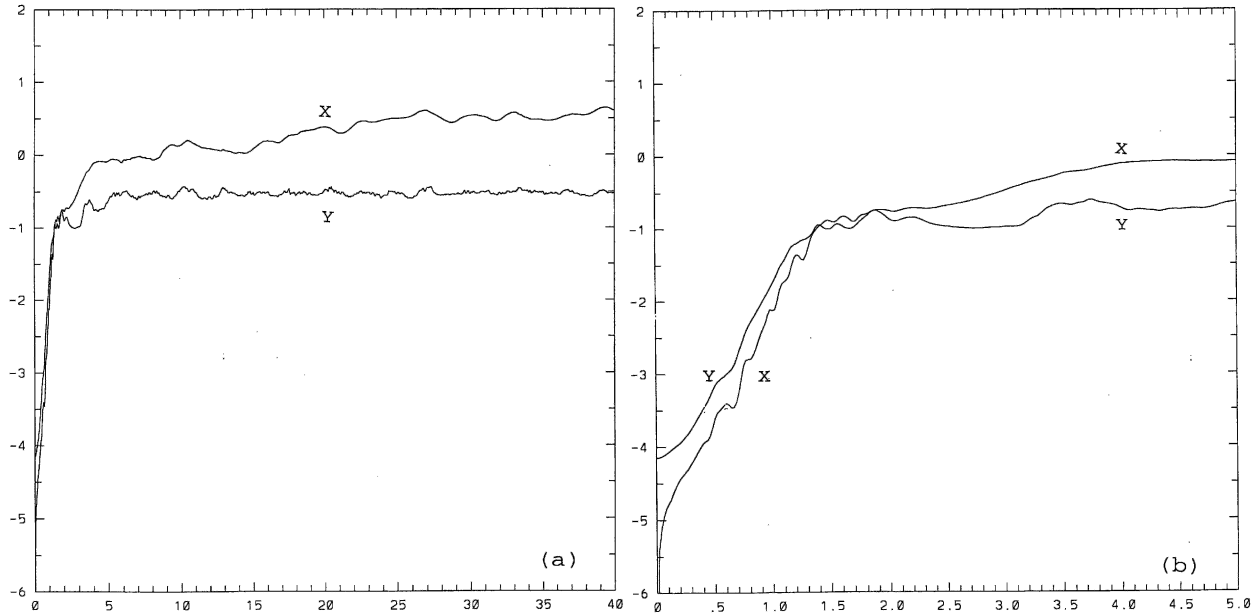

FIG. 8. (a) Growth curves for errors in X and Y extending 40 days, produced by a single pair of solutions of Eq. (15) with $N = 960$, $K = 32$, $J = 12$, $F = 15$, $b = 10$, and $c = 2.5$. The initially small errors were confined to Y. Scale at bottom is time in days. Scale at left is base-10 logarithm of rms error. (b) Horizontal magnification of first 5 days of (a).

tional forecasting when both analysis error and model error are present. I shall examine how these errors at various forecast ranges decrease as the analyses or operational models are replaced by successively better ones.

The application here consists of two experiments. In the first one uses Model II, with rather large values of N and K, say N_0 and K_0, as the perfect model, first to produce a set of true initial states, on which the simulated analyses will be based, and then, for each initial state, to produce true future states, against which the simulated forecasts will be verified. One chooses $N_0 = 960$ so that successive grid points are 3/8 of a degree of longitude apart, while $K_0 = 32$. The second experiment is identical with the first except that the perfect model is Model III.

In either experiment one introduces a succession of operational models. Each one is Model II, with small values of N and K for the leading model and successively large values for the subsequent ones; N is always a divisor of N_0, while $K/N = K_0/N_0$ and the N longitudes constitute an equally spaced subset of the N_0 longitudes used in the perfect model.

If in Model II, with N even and $J = K/2$ even, one approximates X_n by $(X_{n-1} + X_{n+1})/2$ whenever n is odd, and if $x_{n/2} = X_n$ when n is even, I find from Eqs. (7) and (8), or (9) and (10), that the model is simply approximated by Model II, with $N/2$ and $K/2$ replacing N and K and x replacing X. If N is rather large and X_n varies rather smoothly with n, the approximation should be rather good. It is therefore convenient to form each new model by doubling the old values of N and K. For the leading model let $N = 30$ and $K = 1$; then the model thus reduces to Model I, and is presumably a rather poor approximation to the perfect model. The model is denoted by m30. Subsequent models, denoted by m60, . . . , m480, should afford successively better approximations; m960 will denote the perfect model. In every model I let $F = 15$.

Likewise, one introduces a succession of analyses based on a succession of sets of observations. For the leading set one chooses, as observations, the exact values of X_n at a small number M of longitudes, chosen randomly from the N_0 longitudes of the perfect model. For convenience I have let $M = 30$, the value of N in the leading operational model, although it would not be necessary for M to be a divisor of N_0. The analysis, which consists of values of X_n at all N_0 grid points, is produced by first choosing the observed values of X_n at the M observation points. For each remaining grid point, say point j, one finds the third-degree polynomial in n whose values at the two observation points to the west and the two to the east of point j are the observed values of X_n at these points. One then chooses the value of the polynomial at point j as the analyzed value and denotes the analysis by a30. I have made no attempt to use a refined scheme like four-dimensional variational assimilation, which would undoubtedly produce superior analyses and lead to smaller prediction errors, but would also make the model error contribute to the analysis error.

For the subsequent sets of observations, one successively adds more randomly chosen longitudes to the ones already chosen, doubling the total number M at each step, and proceed as before, obtaining analyses a60, . . . , a480; a960 denotes the perfect analysis or true initial state.

To begin either experiment, one chooses states at random and integrates forward with the perfect model for 120 days to suppress transient effects, using 3-h time steps in the first experiment and half-hour steps in the second. With the resulting state as the first initial true state, one integrates forward with each combination of analysis and operational model for seven days, thus obtaining data for the first member of the set. For each additional member, up to a total of 50, one obtains an initial true state by taking the final true state from the previous member and integrating forward for 28 days.

Separately for each combination of analysis, model, and time step, the root-mean-square difference is determined between the predicted and the true state, the averaging being performed over all grid points used in the particular operational model, and over all members of the set. These differences–the "errors"–constitute our output.

Table 1 presents a sampling of the output for the first experiment. It shows the errors produced by different combinations of model and analysis, including the perfect ones. The values are arranged into four blocks, corresponding to the four selected forecast ranges 0, 1, 3, and 7 days. In each block, separate columns refer to separate models, and separate rows to separate analyses. The saturation value for the errors is about 8.0 units.

The 0-day forecast in the leading block is simply the analysis, so that the numbers compare analyses with the perfect one, and, since the forecast models have yet to be applied, the separate columns should be the same except for the differences in sampling. We see that a30 and even a60 are rather poor approximations to a960, while a480 is nearly perfect.

For each remaining block, the combination a30–m30 produces essentially useless forecasts. If one follows down the main diagonal from a30–m30 to a480–m480, improvement is found at every step; a240–m240 produces good forecasts at 5 days while a480–m480 is excellent even at 7 days. However, if one begins with any number on the main diagonal and proceeds to the right, that is, if the model is improved but the analysis is retained, the improvement in the forecast is hardly detectable. If, instead, one proceeds downward from the diagonal by a single row, improving the analysis but not the model, the forecast is generally improved unless near-saturation has already been reached, but moving another row downward yields no further substantial gain. One can conclude that there is a limit to how

TABLE 1. Rms errors in prediction at indicated ranges (0, 1, 3, or 7 days) produced by indicated models (m30–m960) with indicated analyses (a30–a960) when the perfect model (m960) is Model II with $N = 960$, $K = 32$, and $F = 15$.

	m30	m60	m120	m240	m480	m960
			0 days			
a30	5.53	5.52	5.52	5.52	5.52	5.52
a60	3.12	3.11	3.11	3.11	3.11	3.11
a120	1.15	1.17	1.17	1.17	1.17	1.17
a240	0.21	0.21	0.23	0.23	0.23	0.23
a480	0.02	0.01	0.02	0.02	0.02	0.02
a960	0.00	0.00	0.00	0.00	0.00	0.00
			1 day			
a30	6.75	6.46	6.46	6.46	6.46	6.46
a60	4.88	4.02	4.00	4.01	4.01	4.01
a120	3.78	1.70	1.52	1.51	1.51	1.51
a240	3.73	0.88	0.32	0.25	0.25	0.25
a480	3.75	0.86	0.21	0.05	0.02	0.02
a960	3.75	0.86	0.21	0.05	0.01	0.00
			3 days			
a30	7.61	7.11	7.09	7.12	7.13	7.13
a60	7.29	5.69	5.51	5.51	5.51	5.51
a120	7.22	3.77	2.99	2.90	2.89	2.89
a240	7.20	2.54	0.78	0.51	0.49	0.49
a480	7.23	2.52	0.64	0.16	0.05	0.03
a960	7.23	2.52	0.63	0.15	0.03	0.00
			7 days			
a30	8.08	7.71	7.66	7.77	7.77	7.78
a60	8.02	7.27	7.32	7.38	7.38	7.39
a120	8.11	6.28	5.40	5.28	5.30	5.31
a240	8.25	5.74	2.94	1.97	2.02	2.05
a480	8.42	5.71	2.60	0.85	0.22	0.16
a960	8.38	5.71	2.60	0.84	0.18	0.00

TABLE 2. Rms errors in prediction at indicated ranges (0, 1, 3, or 7 days) produced by indicated models (m30–m960) with indicated analyses (a30–a960), when the perfect model (m960) is Model III with $N = 960$, $K = 32$, $I = 12$, $F = 15$, $b = 10$, and $c = 2.5$.

	m30	m60	m120	m240	m480	m960
			0 days			
a30	4.05	4.07	4.08	4.08	4.08	4.08
a60	1.84	1.83	1.83	1.83	1.83	1.83
a120	0.62	0.60	0.60	0.60	0.60	0.60
a240	0.31	0.32	0.30	0.30	0.30	0.30
a480	0.20	0.20	0.20	0.20	0.20	0.20
a960	0.00	0.00	0.00	0.00	0.00	0.00
			1 day			
a30	4.53	4.38	4.38	4.38	4.38	4.35
a60	2.77	2.36	2.33	2.32	2.32	2.28
a120	2.09	1.23	1.11	1.10	1.10	1.04
a240	2.01	1.07	0.92	0.90	0.90	0.78
a480	2.03	1.05	0.89	0.88	0.87	0.61
a960	1.99	1.04	0.89	0.87	0.87	0.00
			3 days			
a 30	5.49	5.18	4.91	4.85	43.84	4.37
a 60	4.94	4.17	3.73	3.65	3.63	3.08
a120	5.00	3.35	2.70	2.58	2.55	1.92
a240	5.01	3.34	2.64	2.52	2.50	1.58
a480	5.03	3.33	2.64	2.52	2.50	1.24
a960	5.04	3.30	2.64	2.52	2.49	0.00
			7 days			
a30	7.46	6.63	6.27	6.19	6.17	5.01
a60	7.20	6.24	5.56	5.45	5.43	3.93
a120	7.17	6.25	5.54	5.34	5.30	3.13
a240	7.26	6.24	5.56	5.36	5.31	2.64
a480	7.17	6.25	5.64	5.43	5.39	2.32
a960	7.06	6.21	5.62	5.44	5.37	0.00

much improvement in forecasting can be realized by improving only the model or only the analysis. This seems to be generally recognized by operational forecasters, and I cannot claim that this illustration has led to any new discoveries.

Table 2, summarizing the second experiment, has the same format as Table 1. The perfect model is Model III, with $N = 960$, $K = 32$, $I = 12$, $F = 15$, $b = 10$, and $c = 2.5$.

Looking at the first block one sees that a480 is again a very good analysis, although its departure from the true state is an order of magnitude greater than in Table 1. Perhaps surprisingly, a120 and even a60 and a30 appear better in the second experiment than in the first. The apparent explanation is contained in Figs. 5b and 6a, which were actually produced by the perfect models for the two experiments. Figure 6a, relevant to Table 2, shows noticeable small-scale activity, which is missing in the poorer analyses, but Fig. 5b reveals even stronger intermediate-scale activity, which the poorer analyses presumably do not capture.

Proceeding to the remaining blocks, one again sees that the forecasts are not improved when we move to the right from the main diagonal, or downward by more than one row. One also sees that at the longer ranges, with any operational model, it makes little differences whether the analysis is a120, a240, or a480, while, with any analysis, it makes little difference whether the forecast is with m120, m240, or m480. At 7 days all the forecasts are poor except those made with the perfect model. The perfect model and m480 differ here mainly in that only the former includes the dynamics of the small scales. In the first experiment, both the perfect model and m480 include the dynamics of the intermediate scales, and they perform equally well regardless of the analysis.

Perhaps the errors shown in Table 2 could be slightly reduced by basing the analyses on a smoothed true state, or by parameterizing the small-scale effects in the models. The latter might be accomplished by slightly reducing the forcing or increasing the dissipation.

6. Concluding remarks

I have begun with a simple set of equations, previously used by myself and others to investigate various atmospheric problems. I have noted that these equations, as they stand, are inappropriate for certain stud-

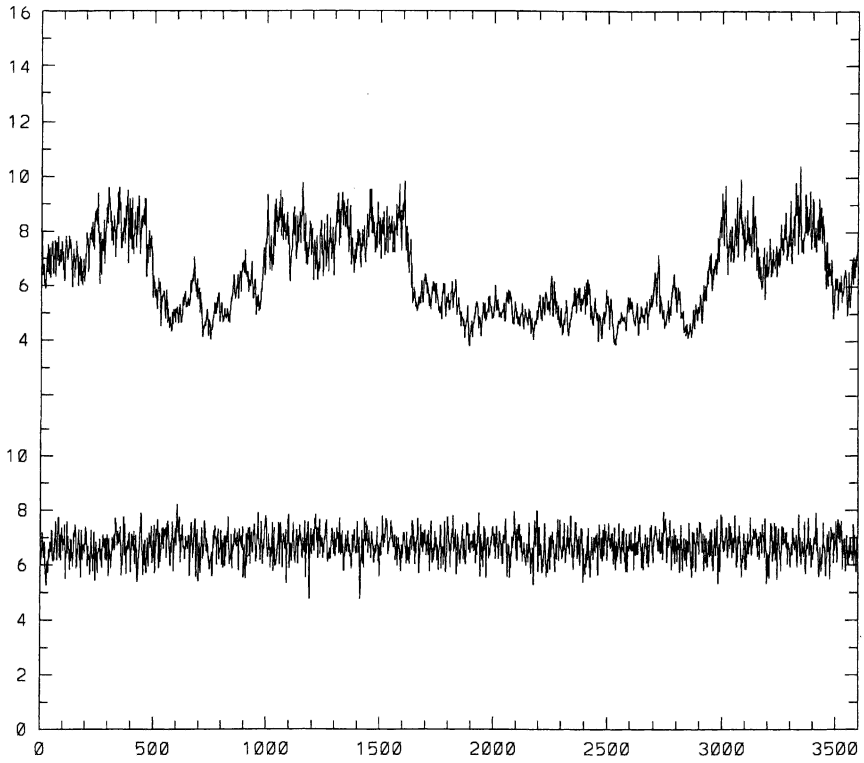

FIG. 9. Time series of s produced by Eqs. (15) and (1) with F variable (upper curve)and with $N = 30$, $H = 15$, $g = 0.1$, and $h = 5$ and by Eq. (1) with F constant (lower curve) with $N = 30$ and $F = 15$. Scale at bottom is time in days. Numbers (0–10) at lower left are scale for lower curve. Numbers (4–16) at upper left are scale for upper curve.

ies, and have introduced modifications that, with some sacrifice of simplicity, render them more suitable. I have applied the modified equations to a specific problem, but, since my main interest has been in the modifying process per se, I have not pushed the application to the point of producing new significant results.

There are numerous other problems that one could pursue with Model II. One of these is ensemble forecasting, where many different schemes for selecting the ensemble members could be explored. Another is data assimilation, where one could examine many promising variants of the procedure.

Either of these problems would ordinarily involve vast amounts of computation, and the main advantage of using a simplified model would be economy. In other problems it might be beneficial to use different combinations of the constants when we do not know which combinations are really appropriate. In the real atmosphere, for example, would smaller-scale systems behave chaotically or regularly if they could be decoupled from the synoptic scales? Does their actual behavior, since they are coupled, depend upon the answer to this question? One could approach these questions by comparing Model III with the values of the constants used in producing Figs. 6, 7, and 8, with an altered form where the advective term $+b^2[Y, Y]_{1,n}$ in Eq. (15) is retained but the damping term $-bY_n$ is replaced by $-b'Y_n$, with b' chosen large enough to remove the chaos from the decoupled Eq. (11).

I have not developed anything resembling a general theory of model design. What successes I have enjoyed have resulted from trial and error, but not, however, from random trial and error. Each satisfactory attempt has been guided by the detailed analysis of previous failures. I make no claim to have discovered the ideal equations.

There are further problems for which different modifications of Model I are in order. For example, the models appear not to produce significant variations with periods of several months or longer, at least when chaos is fully developed, and hence they are unsuitable for problems where progressive changes from one regime of behavior to another play a role. With little additional effort one could produce long-term variability by letting F in Eq. (1) vary with time instead of being a constant and adding a single equation governing F.

Since, very nearly, $S^2 = a^2F^{4/3}$ in Model I, s, if not constant, must undergo fluctuations about $aF^{2/3}$, and, if $G = (s/a)^{3/2}$, we can expect G to undergo fluctuations about F. In a modification of Model I, F can be forced by a constant, say H, but also by G, as long as this positive feedback does not force F all the way to ∞ or

close enough to 0 to eliminate the chaos. One way to do this (there are many) is to let

$$dF/dt = -F + (1 + g)(G - H) - g(G - H)^3/h^2 + H. \quad (16)$$

With $g > 0$, but not too large or too small, and h not too large, values of F will tend to avoid H and cluster around $H - h$ or $H + h$. In Fig. 9, the upper curve is a 10-yr time series of s produced by Eq. (16) [and Eq. (1) with F variable], with $N = 30$, $H = 15$, $g = 0.1$, and $h = 5$. The regimes, sometimes lasting 2 years or longer, with s near $a(\mathrm{H} - h)^{2/3} = 5.08$ or $a(H + h)^{2/3} = 8.07$ are much in evidence. By contrast, the lower curve is produced with Model I, and while the short-term fluctuations of s, this time about $aF^{2/3} = 6.66$, are as strong as in the upper curve, the regimes are absent.

There are presumably numerous other ways to alter Model I, II, or III advantageously. I do not suggest adding a second horizontal dimension, which does not offer obvious advantages over using some form of the vorticity equation. Possibly there is some use for a model with one horizontal and one vertical dimension.

I leave it to the reader or potential user to identify new uses for the models already introduced, or new profitable ways to modify them. Some of these, including the addition of stochastic noise, have already been suggested by an anonymous reviewer.

Acknowledgments. In choosing final forms for the equations, I have benefited greatly from frequent discussions with my colleague James A. Hansen. The work has been supported by the Large-Scale Dynamic Meteorology Program, Lower Atmosphere Research Section, Division of Atmospheric Sciences, National Science Foundation under Grant ATM-0216866.

REFERENCES

Boffetta, G., P. Giuliani, G. Paladin, and A. Vulpiani, 1998: An extension of the Lyapunov analysis for the predictability problem. *J. Atmos. Sci.,* **55,** 3409–3416.

Hansen, J. A., and L. A. Smith, 2000: The role of operational constraints in selecting supplementary observations. *J. Atmos. Sci.,* **57,** 2859–2871.

Kaplan, J. L., and J. A. Yorke, 1979: Chaotic behavior of multidimensional difference equations. *Lecture Notes in Mathematics,* H.-O. Peitgen and H.-O. Waters, Eds., Springer-Verlag, 204–227.

Lorenz, E. N., 1984: The local structure of a chaotic attractor in four dimensions. *Physica D,* **13,** 90–104.

——, 1996: Predictability—A problem partly solved. *Proc. Seminar on Predictability,* Vol. 1, Reading, Berkshire, United Kingdom, ECMWF, 1–18.

——, and K. A. Emanuel, 1998: Optimal sites for supplementary weather observations: Simulation with a small model. *J. Atmos. Sci.,* **55,** 399–414.

Simmons, A. J., R. Moreau, and T. Petrolagis, 1995: Error growth and estimates of predictability from the ECMWF forecasting system. *Quart. J. Roy. Meteor. Soc.,* **121,** 1739–1771.

Starr, V. P., 1948: An essay on the general circulation of the earth's atmosphere. *J. Meteor.,* **5,** 39–43.