Atmospheric Analogs and Recurrence Time Statistics: Toward a Dynamical Formulation

C. NICOLIS

Institut Royal Météorologique de Belgique, Brussels, Belgium

(Manuscript received 26 June 1996, in final form 19 June 1997)

ABSTRACT

A dynamical approach to atmospheric analogs extending the statistical formulation by Toth and Van den Dool is developed. Explicit analytical formulas for the probability and the mean recurrence time of analogs displaying the system's intrinsic time scales are provided for both discrete and continuous time dynamical systems and are evaluated numerically on a representative model. The analysis reveals strong dependence of recurrence times of analogs on the local properties of the attractor and a pronounced variability around their mean. Finally, the formulation is extended to stochastically forced systems such as a red noise atmosphere.

1. Introduction

It is now recognized that initially close states of the atmosphere diverge subsequently in an exponential-like fashion. Nevertheless, if the difference between them is small enough the phase space trajectories emanating from these initial configurations of the system will, for some time and for all practical purposes, be indistinguishable from each other. The importance of such natural close states or ''analogs'' for the prediction of short-term weather fluctuations has been recognized for a long time (see, e.g., Barry and Perry 1973). Subsequently, this idea was used in a variety of problems such as the estimation of atmospheric predictability (Lorenz 1969), seasonal forecasting (Barnett and Preisendorfer 1978), probabilistic temperature forecasts (Kruizinga and Murphy 1983), and cluster analysis (Wallace et al. 1991).

Ordinarily, the identification and classification of atmospheric analogs relies on the historical record of observations. As the full set of variables characterizing the state of the atmosphere is very large, the selection of analogs is carried out on a reduced phase space spanned by a limited number of variables or empirical orthogonal functions (see, e.g., Ruosteenoja 1988; Toth 1991a). This subspace is endowed with a suitable metric, and states within a prescribed distance are qualified as analogs. Unfortunately, as most of the libraries available are rather short (10–100 years) it turns out to be very hard to find analogs within a reasonably small distance, even when attention is limited to a single type of observable such as the 500-mb geopotential height (Van den Dool 1994).

In view of the above limitations, theoretical ideas allowing one to arrive at a priori estimates of the probability of analogs of a given quality and of the waiting times necessary for their realization become highly desirable. An important development in this direction has been the observation that in an averaged sense, the distribution of circulation patterns in phase space is practically indistinguishable from a multinormal distribution (Toth 1991b). Van den Dool (1994) utilized this information to estimate the analog waiting time, M. The basic quantity involved in his estimate is the probability $P(\varepsilon)$ that two arbitrary chosen states are within a distance ε (the confidence interval chosen to define an analog) at a given point in physical space. To evaluate $P(\varepsilon)$ one integrates the probability density function $\rho(\delta y)$ of differences of the relevant variable values over an interval of order ε. Given a library of M years and denoting by m the number of independent cases in the time windows considered, the chance, α, of finding a match in the library resembling a selected base case is, then,

$$\alpha = 1 - (1 - P^N(\varepsilon))^{mM}, \tag{1}$$

where N is the number of independent degrees of freedom. Inverting this equation one finds, for N sufficiently large (Van den Dool 1994),

$$M \geq \frac{|\ln(1 - \alpha)|}{mP^N(\varepsilon)}, \tag{2}$$

leading to astronomical waiting time estimates for any reasonably small value of ε.

Implicit in the above reasoning is the idea that the waiting time M is determined by the univariate probability distribution $\rho(\delta y)$—a static quantity assumed, in

Corresponding author address: Dr. Catherine Nicolis, Institut Royal Météorologique de Belgique, Avenue Circulaire 3, B-1180 Bruxelles, Belgium.
E-mail: cnicolis@oma.be

addition, to be normal. Now, in our view this description needs to be enlarged for two reasons. First, the waiting time M, should bear in one way or the other the signature of the underlying dynamics, and yet, in Eq. (2) no intrinsic timescale is apparent. Second, accepting for a moment that 1 yr is a ''natural'' time step, one is entitled to inquire whether the *full*, multivariate probability $\rho(\mathbf{y})$ should not also enter in the estimate of M. This would dispense us from making the assumption of normal distribution at the very detailed level of the individual variables since, as pointed out by Toth (1991b), normality holds only in the averaged sense of distances from the climatological mean. The principal goal of the present work is to enlarge the description of analogs along the above indicated lines, in particular, by formulating the analog problem as a problem of recurrence time statistics of a dynamical system.

In section 2 the formulation of recurrence for discrete time dynamical systems is laid down. An explicit general formula is derived and compared with the formulation of Eqs. (1)–(2). In section 3 the formulation is extended to continuous time systems. A formula originally due to Smoluchowski is proposed for the mean recurrence times of such a system, which incorporates the system's intrinsic timescales through the presence of *two-time* probability distributions. The formulation is applied, in section 4, to Lorenz's three-variable thermal convection model (Lorenz 1963). The dependence of recurrence times on the local properties of the attractor, and their variance around the mean are computed by direct numerical simulation and compared with Eqs. (1)–(2). In section 5 the formulation is extended to stochastically forced systems, such as a red noise atmosphere. The main conclusions are drawn in section 6.

Reassessing the theoretical basis of analogs will also allow one to cope with the fact that the phase space attractors associated to atmospheric dynamics are most probably fractal objects, implying that the invariant probability $\rho(\mathbf{y})$ cannot really be a smooth function of the multinormal type. Evidence will also be given of a wide dispersion of waiting times around their mean, an aspect that, given the small number of samples available, is usually not touched upon in analog studies.

2. Recurrence time statistics in discrete time dynamical systems: Formulation

Suppose that a dynamical system like the atmosphere is probed every τ time units. Let $\mathbf{x} = (x_1, \cdots, x_n)$ be the set of the state variables, taken to lie within a cell C of phase space Γ whose linear dimension ε is the confidence interval chosen to define analogs. If the evolution laws are deterministic, the state at time τ will be uniquely defined in terms of $\mathbf{x}$, by the action of a certain operator $\mathbf{F}^\tau$

$$\mathbf{x}_\tau = \mathbf{F}^\tau(\mathbf{x}). \tag{3}$$

Upon repeated action of $\mathbf{F}^\tau$ the system will leave, as a rule, the cell C. However, if the evolution operator defines an *ergodic transformation*, the system will be bound to return to cell C, given enough time (Kac 1959). One may derive an explicit formula for the mean recurrence time $\langle\theta_\tau\rangle$ of cell C as follows. Let $C_1, C_2, \cdots, C_k$ be the sets of points such that

$$C_1 : \mathbf{x} \in C, \qquad \mathbf{x}_\tau \in C \qquad \text{(cell } C_1\text{)}$$

$$C_2 : \mathbf{x} \in C, \qquad \mathbf{x}_\tau \notin C, \qquad \mathbf{x}_{2\tau} \in C \qquad \text{(cell } C_2\text{)}$$

$$C_k : \mathbf{x} \in C, \qquad \mathbf{x}_\tau \notin C, \cdots \mathbf{x}_{(k-1)\tau} \notin C, \qquad \mathbf{x}_{k\tau} \in C. \qquad \text{(cell } C_k\text{)}$$

By definition, the mean recurrence time in cell C is

$$\langle\theta_\tau\rangle = \frac{\tau}{\mu(C)} \sum_{k=1}^{\infty} k\mu(C_k), \tag{4}$$

where the *measure* $\mu(C)$ of cell C is the integral over C of the invariant probability $\rho(\mathbf{x})$:

$$\mu(C) = \int_C d\mathbf{x}\, \rho(\mathbf{x}). \tag{5}$$

It is convenient to introduce the *characteristic function* $\chi(\mathbf{x})$ of cell C

$$\chi(\mathbf{x}) = \begin{cases} 1, & \text{if } x \in C \\ 0, & \text{otherwise.} \end{cases} \tag{6}$$

It is then a matter of algebra to show that (Kac 1959)

$$\sum_{k=1}^{\infty} k\mu(C_k) = 1 - w, \tag{7a}$$

where

$$w = \lim_{n\to\infty} \int_\Gamma d\mathbf{x}\, \rho(\mathbf{x})(1 - \chi(\mathbf{x}))(1 - \chi(\mathbf{x}_\tau)) \cdots \times (1 - \chi(\mathbf{x}_{n\tau})). \tag{7b}$$

Clearly, w represents the probability that neither $\mathbf{x}$ nor any of its future images are in cell C. In an ergodic process this probability is zero. Substituting (7a) into (4) one obtains then

$$\langle\theta_\tau\rangle = \frac{\tau}{\mu(C)}. \tag{8}$$

Identifying τ with the sampling time this relation should be equivalent with Eq. (2). In this respect, the following comments are in order:

1) Equations. (2) and (8) are in qualitative agreement in the sense that they both express waiting or recurrence times in terms of quantities related to the invariant probability distribution.
2) Since the probability $P(\varepsilon)$ in Eq. (2) is, by construction, of order ε, Eq. (2) predicts a scaling of M with respect to the confidence interval in the form

$$M \approx \frac{1}{\varepsilon^N}, \tag{9a}$$

where N is the number of independent degrees of freedom. On the other hand, since the support of $\rho(\mathbf{x})$ in Eq. (5) is the system's attractor, an order of magnitude estimate of $\mu(C)$ in Eq. (8) is $\mu(C) \approx \varepsilon^d$, hence

$$\langle \theta_\tau \rangle \approx \frac{\tau}{\varepsilon^d}, \tag{9b}$$

where d is the attractor dimension. This statement clarifies Van den Dool's (1994) idea that analogs are to be sought in a subspace spanned by a limited number of ''relevant'' variables or by some ''dominant'' empirical orthogonal functions, by identifying this subspace as the part of phase space spanned by the system's attractor.

3) According to Eq. (8), the waiting (or recurrence time) depends on the *multivariate* probability $\rho(\mathbf{x})$ on the attractor [see Eq. (5)]. In contrast, Eq. (2) is based entirely on the univariate probability density, $\rho(\delta x)$. This difference relates to the assumptions of factorization and normality made at the outset in the statistical approach. The advantage of Eq. (8) is to be independent of such properties, which as pointed out previously (Toth 1991a) can be valid only in an average sense anyway. An analytical argument substantiating this idea further is developed in the appendix. It must be borne in mind, however, that in practice real world datasets may be too small and noisy to allow for a more accurate estimate of $\mu(C)$ beyond the crude one, $\mu(C) \approx \varepsilon^d$.

3. Recurrence time statistics in continuous time dynamical systems

We turn now to the more realistic picture of the atmosphere as a continuous time dynamical system. Indeed, while the typical sampling time of the data used in a library is of the order of the day, the intrinsic timescale relevant for the dynamics is substantially larger. In reality one is therefore monitoring, through the record, a continuous time dynamical system even though one chooses for practical purposes to look at the system through a particular window. Let the corresponding evolution laws be

$$\frac{d\mathbf{x}}{dt} = \mathbf{f}(\mathbf{x}), \tag{10}$$

where, depending on the case, $\mathbf{f} = (f_1, \cdots, f_n)$ may represent the right-hand sides of the primitive equations, of the thermodynamic equation etc., evaluated at suitable grid points. We are interested, again, in the event that a phase space trajectory emanating from a cell C is reentering this cell for the first time. Now, in the limit $\tau \to 0$ (continuous sampling) definition (4) and Eq. (8) applied straightforwardly give the trivial (and wrong) result $\langle \theta_\tau \rangle \to 0$. On the other hand it is clear that in this limit cell C_1 introduced in section 2 must be omitted from the counting, otherwise one will have no way to express appropriately that the system first leaves cell C_1 before returning to it. A formulation accounting for this point has been worked out by Smoluckowski. We do not dwell here on the details of the method, referring the interested reader to Kac's (1959) classical monograph. We merely write the result,

$$\langle \theta \rangle = \tau \frac{1 - P(C)}{P(C) - P(C, 0; C, \tau)}, \tag{11}$$

where τ is again the sampling time (now with the possibility $\tau \to 0$), $P(C)$ is the probability to be in cell C, and $P(C, 0; C, \tau)$ is the joint probability to be in C both initially and at time τ.

To extend the results of section 2 to continuous time systems one needs now to evaluate the two quantities $P(C)$ and $P(C, 0; C, \tau)$ on the system's attractor. As in section 2 it is again convenient to introduce the characteristic function $\chi(\mathbf{x})$ of cell C through Eq. (6). The invariant probability $P(C)$ becomes, then,

$$P(C) = \int_\Gamma d\mathbf{x}\, \rho(\mathbf{x}) \chi(\mathbf{x}), \tag{12a}$$

where Γ is the total phase space available to the system. Likewise, the two-time distribution $P(C, 0; C, \tau)$ becomes

$$P(C, 0; C, \tau) = \int_\Gamma d\mathbf{x}\, \rho(\mathbf{x}) \chi(\mathbf{x}) \chi(\mathbf{x}_\tau), \tag{12b}$$

where $\mathbf{x}_\tau(\mathbf{x})$ denotes the phase space point on which $\mathbf{x}$ is projected after a time interval τ, according to the evolution equations (10).

Let us choose, for concreteness, a cell C in the form of a hypercube of relative side ε (i.e., divided by the size of the attractor) including a point $\mathbf{x}_0 = (x_{10}, \cdots, x_{n0})$ as its downleftmost summit. The presence of characteristic function $\chi(\mathbf{x})$ restricts, then, the integration in (12a) to the limits $\{x_{0i}, x_{0i} + \varepsilon\}$,

$$P(C) = \int_{x_{01}}^{x_{01}+\varepsilon} dx_1 \cdots \int_{x_{0n}}^{x_{0n}+\varepsilon} dx_n\, \rho(\mathbf{x}). \tag{13a}$$

Since the support of $\rho(\mathbf{x})$ is the system's attractor, a good estimate of $P(C)$ is, again [see discussion preceding (9b)]

$$P(C) \approx \varepsilon^d. \tag{13b}$$

Consider now Eq. (12b). The presence of both $\chi(\mathbf{x})$ and $\chi(\mathbf{x}_\tau)$ implies that the integration must be compatible with the fact that the intervals $\{x_{\tau i}(\mathbf{x}_0), x_{\tau i}(\mathbf{x}_0 + \varepsilon)\}$ must overlap with $\{x_{0i}, \mathbf{x}_{0i} + \varepsilon\}$. In plain terms this means that integration in (12b) bears on the *intersection* $C \cap C_{-\tau}$ of cell C and the set of those points that will be projected into C after a time interval τ. This set is

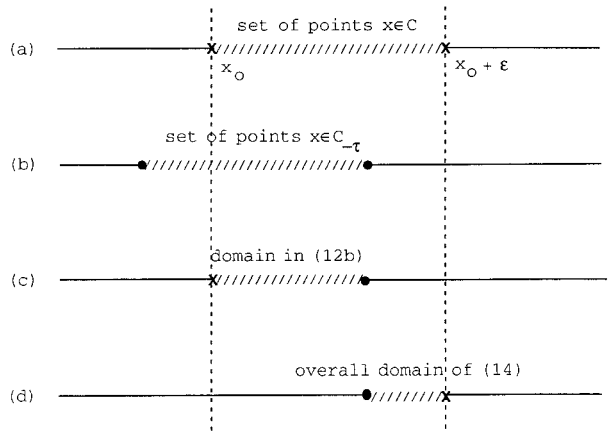

FIG. 1. Schematic representation of the steps involved in the evaluation of Eq. (11) in a one-dimensional phase space. (a) The analog considered lies in cell C between x_0 and $x_0 + \varepsilon$. (b) Set of points in cell $C_{-\tau}$ evolving to C after a time interval τ. (c) The intersection $C \cap C_{-\tau}$ provides the domain of integration of Eq. (12b). (d) The difference between C and the intersection $C \cap C_{-\tau}$ represents the domain of integration of the denominator of Eq. (11).

referred to as the preimage $C_{-\tau}$ of cell C. The situation is illustrated in Fig. 1 for a one-dimensional space. To evaluate the mean recurrence time from Eq. (11), one actually needs the difference between (12a) and (12b). This amounts to integrating the invariant density $\rho(\mathbf{x})$ over a support of the type indicated in Fig. 1d, which is clearly $O(\tau)$ and can therefore allow, in principle, for cancellation with the τ factor in the numerator of Eq. (11). To estimate the ε dependence of this overall expression we argue as follows: Consider the limit $\tau \to 0$ (continuous sampling). The denominator, Q, of Eq. (11) is then of the form

$$Q = \int_{x_{01}}^{x_{01}+\varepsilon} dx_1 \cdots \int_{x_{0n}}^{x_{0n}+\varepsilon} dx_n \rho(\mathbf{x}) - \int_{\max(x_{01}, x_{01}+a_1\tau)}^{\min(x_{01}+\varepsilon, x_{01}+\varepsilon+a_1'\tau)} dx_1 \cdots \int_{\max(x_{0n}, x_{0n}+a_n\tau)}^{\min(x_{0n}+\varepsilon, x_{0n}+\varepsilon+a_n'\tau)} dx_n \rho(\mathbf{x}), \tag{14}$$

where $\{a_i\}$, $\{a_i'\}$ are related to $\{x_{0i}\}$ and ε in a complicated way that needs not be specified at this stage. We now expand (14) in powers of τ, a process that is legitimate as long as ρ is finite and integrable. The zero-order term will cancel, while the first-order one will be a sum of n contributions in which one of the n integrations over the x_i will disappear, whereas the others will bear on the *original* limits $\{x_{0j}, x_{0j+\varepsilon}\}$. Clearly, each of the factors in Eq. (14) is of order ε^d. The suppression of one coordinate in the integration over phase space in Eq. (14) would affect the exponent d in this scaling by a term δ between 0 and 1, depending on whether the direction in question is practically parallel or, on the contrary, transversal to the motion on the attractor. We express this complex dependence by the formal relation

$$Q = \sum_{i=1}^{n} \tau v_i \varepsilon^{d-\delta_i},$$

where v_i is the velocity along the direction i. Clearly, in the limit of small ε this expression will be dominated by the term corresponding to $\delta_i = 1$. As this term refers to a direction along the trajectory one will have then

$$Q \approx \tau \overline{v} \varepsilon^{d-1} \qquad (\varepsilon \to 0), \tag{15}$$

where $\overline{v}$ is the mean phase space velocity. Combining this with Eqs. (11) and (13b), we obtain

$$\langle \theta \rangle \approx \frac{1 - \varepsilon^d}{\overline{v} \varepsilon^{d-1}}. \tag{16}$$

This result clarifies further Van den Dool's estimate [Eq. (2)]. Regarding dependence on ε, it leads to a reduction by a factor ε^{-1} in the denominator and to the presence of the extra factor $1 - \varepsilon^d$ in the numerator while showing, once again, that what really matters is the attractor dimension, which is the only intrinsic way to identify the number of relevant variables involved in the dynamics. Naturally Eq. (16) is only a rough estimate. In reality, as the developments of this section amply illustrate, recurrence time statistics strongly depends on the local properties of the attractor, and hence on the nature (e.g., scale) of the regime of interest. This point will be developed further in the next section, where the ideas of this section will be illustrated on a simple model.

4. A case study: Lorenz's three-variable thermal convection model

Lorenz's three-variable model reads (Lorenz 1963; Sparrow 1982)

$$\frac{dx}{dt} = \sigma(-x + y)$$
$$\frac{dy}{dt} = rx - y - xz$$
$$\frac{dz}{dt} = xy - bz, \tag{17}$$

where the variable x measures the rate of convective (vertical) turnover, y the horizontal temperature variation, and z the vertical temperature variation. The parameters σ and r are proportional, respectively, to the Prandtl number (depending entirely on the intrinsic properties of the fluid) and to the Rayleigh number (incorporating the effect of the thermal constraint). Finally, the parameter b accounts for the geometry of the convective pattern.

As is well known, for $\sigma = 10$, $r = 28$, and $b = 8/3$, Eqs. (17) give rise to the classical Lorenz attractor, the first clear-cut demonstration of deterministic chaos in dissipative autonomous dynamical systems.

The distribution of Euclidean distances around the

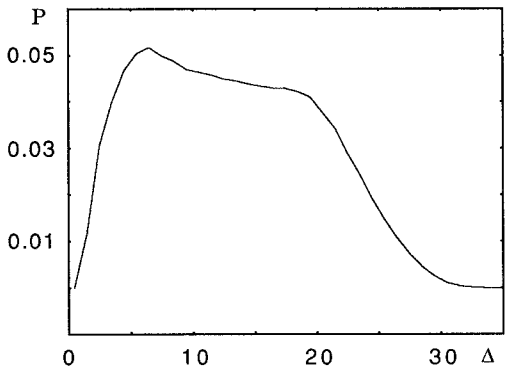

FIG. 2. Probability density of the Euclidean distances, Δ, around the climatological mean of the Lorenz model, Eqs. (17), with $b = 8/3$, $\sigma = 10$, and $r = 28$ as obtained from 200 000 data points sampled every 0.01 time units.

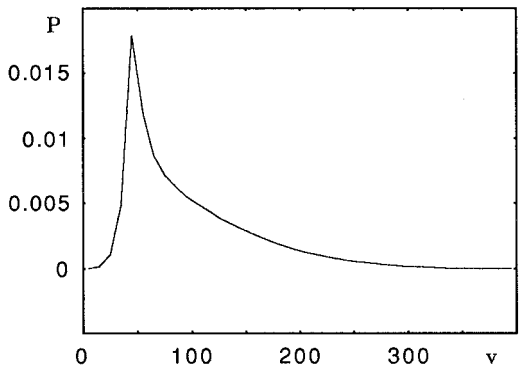

FIG. 3. Probability density of $Pv(t)$ as obtained from model (17) after an integration of 1000 time units. Parameter values as in Fig. 2.

''climatic'' (long time) average of the Lorenz attractor induced by the probability $\rho(x, y, z)$ is depicted in Fig. 2. We observe that this distribution is not—nor is it supposed to be—a normal distribution as in Toth (1991b) since, for one thing, the number of degrees of freedom involved is small and the variables are highly correlated. This property is expected to extend to all systems possessing a low-dimensional attractor and is further discussed in the appendix.

An estimate for the mean recurrence time can be obtained by following the lines leading to Eqs. (15)–(16). In the present case the dimensionality of the attractor is $d \sim 2.06$. To estimate the mean phase space velocity $\overline{v}$, we first observe, using Eqs. (17), that the instantaneous velocity field is

$$\mathbf{v}(t) = \{\sigma(-x + y),\ rx - y - xz,\ xy - bz\}$$

with

$$v(t) = |\mathbf{v}(t)| = [\sigma^2(-x + y)^2 + (rx - y - xz)^2 + (xy - bz)^2]^{1/2}.$$

Figure 3 depicts the probability density of $v(t)$. We observe a large dispersion around the mean, and a marked asymmetry toward high values of v. The mean and most probable velocities turn out to be $\overline{v} \approx 95.7t^{-1}$ and $55t^{-1}$, respectively (t^{-1} denotes here inverse units of time), whereas the standard deviation is $57.3t^{-1}$. For comparison, the value of velocity at the ''climatological'' mean of the model (0, 0, 23.5) is 62.5 t^{-1}.

Since ε in Eq. (16) is between 0 and 1, the above figures must be normalized by the size of the system's attractor before a realistic estimate of $\langle\theta\rangle$ for the model at hand can be performed. A detailed view of the attractor shows a span of about 39 along x, of 54 along y, and of 47 along z. This leads us to an estimate of $\overline{v} \approx 2.2\ t^{-1}$ in Eq. (16) yielding

$$\langle\theta\rangle \approx \frac{1 - \varepsilon^{2.06}}{2.2\varepsilon^{1.06}}. \tag{18}$$

We shall now resort to a numerical simulation of the recurrence process and the concomitant search of analogs. We first address recurrence in a three-dimensional box surrounding a point chosen to correspond to the ''climatological'' mean of the model ~(0, 0, 23.5). The sides of the box have been normalized by the size of the attractor in such a way that they correspond to the same ε along the directions x, y, and z. The system is started initially in the box just defined and Eqs. (17) are integrated over a long time period. The passage times t_i of the phase space trajectory from the box are monitored and the corresponding recurrence times $\Delta t_i = t_{i+1} - t_i$ recorded. Clearly, the states at t_i and t_{i+1}, $i = 0, 1, \cdots$ can be regarded as analogs for the dynamics on the Lorenz attractor. The experiment is conducted until a large number of recurrences is achieved, and repeated for box sizes ranging from 1/20 of the attractor size ($\varepsilon = 0.05$) to one-half the size of the attractor ($\varepsilon = 0.5$) with a step $\Delta\varepsilon = 0.025$. Notice that the recurrence times evaluated in this way are necessary multiples of the time step used in the numerical integration method, the smallest possible value being twice this step.

Figure 4 depicts the recurrence time distribution for $\varepsilon = 0.05$ (a), $\varepsilon = 0.15$ (b), and $\varepsilon = 0.5$ (c) as obtained from 20 000 realizations. The distributions are rather delocalized entailing the existence of a nonnegligible variance around the mean, which will be discussed further. The empty circles in Fig. 5a depict the dependence of the mean recurrence time $\langle\theta\rangle$ on the box size ε as obtained from 100 000 realizations, whereas the full circles reproduce the rough estimate obtained from the theoretical expression (18). The overall correspondence between these two curves is qualitatively acceptable. As ε increases, that is, as the analogs become coarser, the distance between the curves tends to zero, although at first sight the agreement would be expected to be better for small ε. This trend may be attributed to the fact that in deducing Eq. (18) an averaging approximation has been made and hence any trace of the local character of the attractor has been wiped out. Quantitatively, it is largely due to the numerator of $\langle\theta\rangle$, which goes to zero as ε tends to unity. The crosses in Fig. 5a reproduce

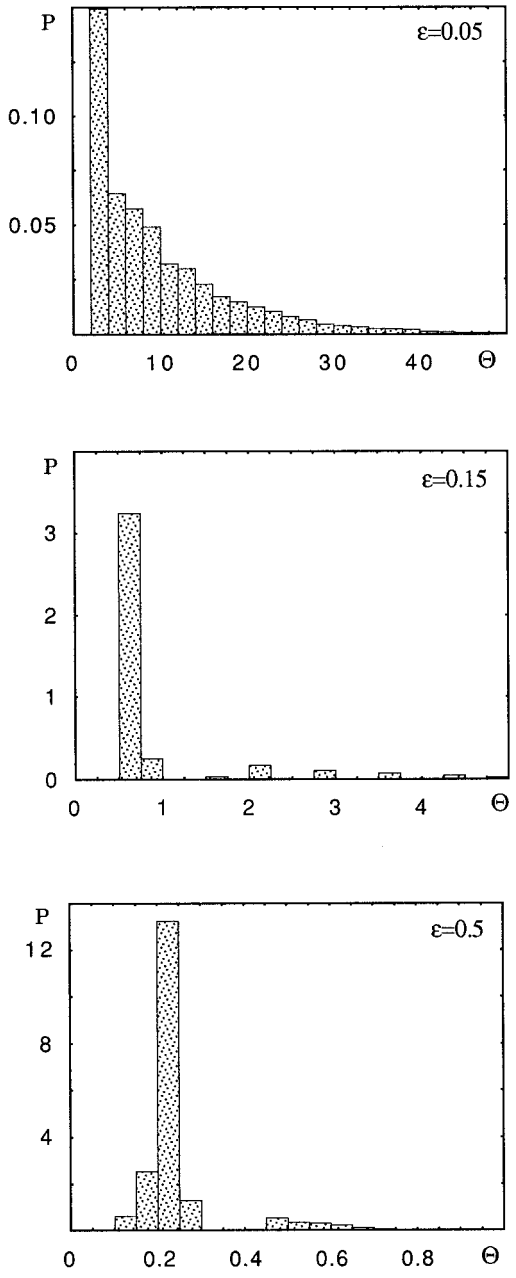

FIG. 4. Probability density of recurrence times of analogs surrounding the climatological mean of model (17) as obtained from 20 000 recurrences. The distances ε in the three phase space directions have been normalized by the attractor's size (a), $\varepsilon = 0.05$; (b), $\varepsilon = 0.15$; and (c), $\varepsilon = 0.5$. Parameter values as in Fig. 2.

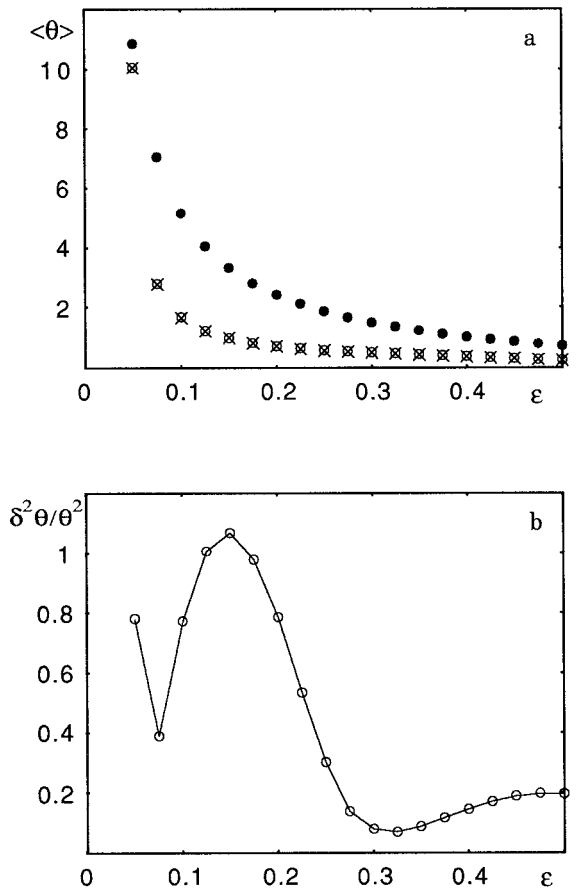

FIG. 5. (a) Dependence of the mean recurrence time on the coarseness, ε, of the analog considered for model (17) as obtained from direct numerical simulation of the trajectory (open circles), using Smoluchowski's formula (crosses) and the rough estimate, Eq. (18), (full circles). (b) Dependence of the relative variance on ε. Parameter values as in Fig. 2. Number of realizations considered in each ε: 100 000.

$\langle\theta\rangle$ as obtained using Smolukowski's formula [Eq. (11)]. In this case the stationary probability distribution $P(C)$ of the particular cell C considered, as well as its joint probability $P(C, 0; C, \tau)$, have been estimated from a long time integration of the model involving 100 000 passages through cell C. By identifying τ in the formula to the integration step the corresponding mean recurrence time has been evaluated. The agreement with the values obtained by counting the time spent by the trajectory outside the cell C (empty circles) is complete.

Figure 5b shows the properties of the variance of θ normalized by the square of the mean recurrence time $\langle\theta\rangle$. We observe that the dependence on ε here is far more complex than the one depicted in Fig. 5a. Specifically, contrary to $\langle\theta\rangle$, the normalized variance experiences (at least for the step $\Delta\varepsilon$ considered) a maximum value at $\varepsilon = 0.15$ and two minima at $\varepsilon = 0.1$ and $\varepsilon = 0.325$, whereas for intermediate values of ε the dependence is close to linear. Moreover, the variability seems to be much more important for good analogs (small ε) than for mediocre ones (large ε). This last point suggests that, contrary to what one could anticipate, the identification of good analogs in the Lorenz attractor does not necessarily entail better predictability, at least as far as waiting times of the analog to recur are concerned.

We next introduce a coverage of the attractor in the xy plane by 20^2 boxes spanning the entire region of

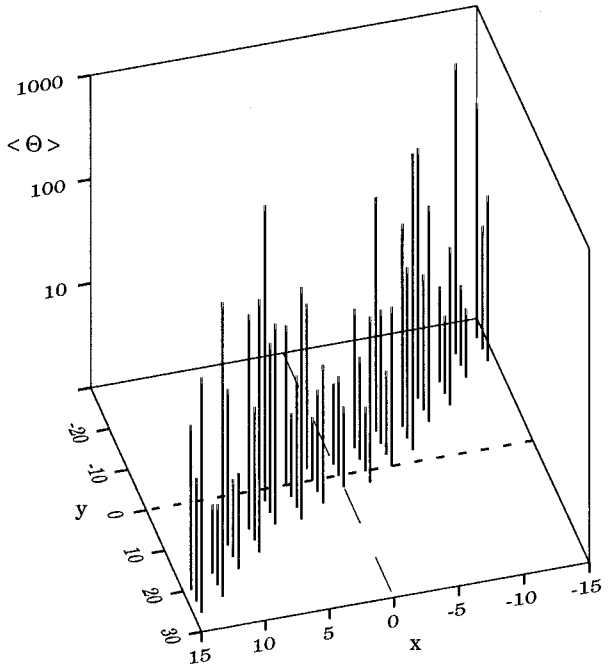

FIG. 6. Distribution of the mean recurrence times of analogs in the *xy* plane for $z = 23.5 \pm 0.05\ Z$ (Z being the size of the attractor in the z direction) for model (17) as obtained from a partitioning of the attractor's *xy* plane into 20^2 cells of equal relative size and 10^6 realizations.

these phase space variables $\sim(-20 \leq x \leq 20,\ -27 \leq y \leq 27)$. For consistency with the previous experiments care is taken that the climatological mean is in the center of one of the grids of the partition. The system is integrated and the local recurrence times for a given z range are registered as in the previous experiment. Figure 6 depicts the distribution of $\langle\theta\rangle$ in the *xy* plane for $z = 23.5 \pm 0.05\ Z$, Z being the size of the attractor in the z direction as obtained from 10^6 realizations. We observe that among the 400 partitions available in the *xy* plane only about a quarter of them are visited by the trajectory, a characteristic reminiscent of the deterministic nature of the underlying system. Moreover, the values of the mean recurrence times span more than two orders of magnitude, giving a mean in the entire *xy* plane of about 100 time units.

Finally, a coverage of the entire attractor by 19^3 grids is performed and the mean recurrence times of these three-dimensional phase space subvolumes are estimated. Figures 7a and 7b depict the probability density of the mean recurrence times $\langle\theta\rangle$ and of the relative variances, respectively, over the entire attractor as obtained from 10^7 realizations. We retained only values that have been obtained through an averaging over a number of occurrences such that the statistics remain robust as the integration time considered increases. A reasonable lower limit was found to be about 500 occurrences, giving a total number of subvolumes visited by the trajectory of about 600. The mean recurrence time over the entire attractor is found to be $\sim$60 time units. It is worth noticing that in Fig. 7a the probability (open circles) varies with $\langle\theta\rangle$, definitely slower than exponentially (full line). It is much closer to a power law (dotted line), though not perfectly fitted for short $\langle\theta\rangle$. Given that recurrence is a discrete time dynamics, one can argue that a purely exponential dependence on $\langle\theta\rangle$ could be reminiscent of a Markovian process (Nicolis et al. 1997). On the other hand, power laws are characteristic signatures of intermittency. The behavior found here is probably indicative of a situation closer to an infinite memory intermittent dynamics than a memoryless Markovian process.

The distribution of the relative variance, $\delta^2\theta/\langle\theta\rangle^2$, over the attractor (Fig. 7b) reveals that most of the probability density is concentrated around a relatively sharp peak around $\sim$2, although large excursions up to three times this value are possible. This implies serious limitations in the practical usefulness of the mean recurrence times in prediction.

5. Stochastic dynamical systems: The case of a red noise atmosphere

The earth's energy and momentum budget or the mass budget of minor constituents are continuously perturbed by local imbalances of the various fluxes or by changes in the boundary conditions. In many instances such forcings can be accounted for, at least qualitatively, by adding a random noise term in the balance equations. This formulation has been used for modeling, among others, surface temperature anomalies (Frankignoul and Hasselmann 1977; Kim and North 1991), Quaternary glaciations (Nicolis 1982), or El Niño events (Vallis 1988). In this section we address the recurrence and analog problem for such systems and comment on the insight afforded by this analysis in the statistical prediction of the corresponding events. We consider the simplest possible case of a single observable, or more precisely of its deviation $x = X - \overline{X}_\infty$, from the ergodic mean $\overline{X}_\infty$,

$$\frac{dx}{dt} = -\lambda x + F(t), \tag{19a}$$

where λ^{-1} is the relaxation time associated to the damping of x and $F(t)$ a Gaussian white noise process:

$$\langle F(t)\rangle = 0$$
$$\langle F(t)F(t')\rangle = q^2\delta(t - t'); \tag{19b}$$

here the brackets denote statistical average over the various realizations of F and q^2 is the variance of the noise.

A more fundamental view of Eqs. (19) is to interpret x as a collective, slowly varying quantity forced by a large number of faster varying ones, undergoing chaotic dynamics in a high-dimensional phase space. A similar representation has been proposed some time ago for the Lorenz model, the difference with Eq. (19) being that,

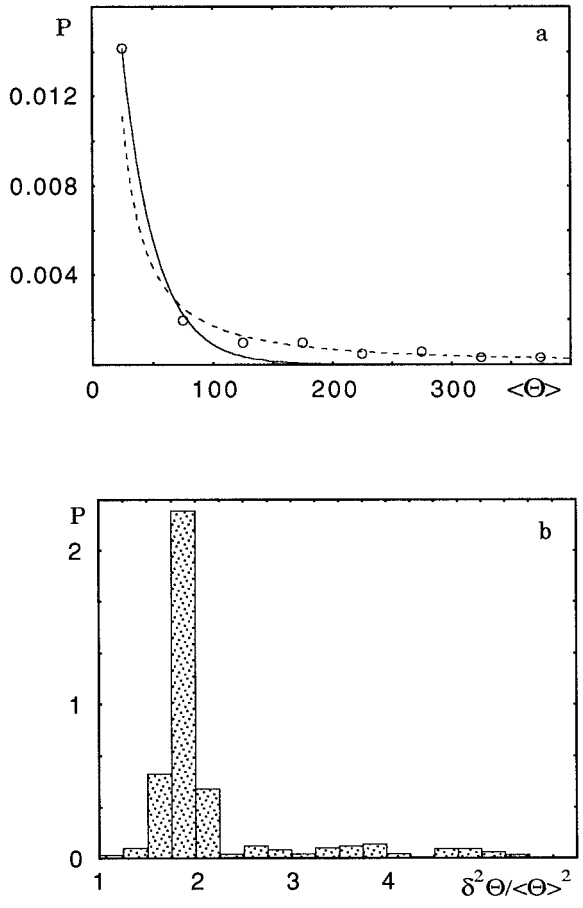

FIG. 7. Probability density of (a) $\langle\theta\rangle$ (open dots), best fit with an exponential function (full line) and a power law (dotted line) and (b) of $\delta^2\theta/\langle\theta\rangle^2$, as obtained from a partitioning of the entire attractor of model (17) into 19^3 cells of equal relative size and 10^7 realizations.

owing to the small number of variables involved, the noise term F itself is not white but exhibits, rather, memory effects (Nicolis and Nicolis 1986).

Equations (19) define the Ornstein–Uhlenbeck process—the most typical example of red noise. It is well known that the invariant probability density, instantaneous variance, and two-time conditional probability density for this process read (Gardiner 1983)

$$\rho(x) = \left(\frac{\lambda}{2\pi q^2}\right)^{1/2} e^{-(\lambda/q^2)x^2} \tag{20a}$$

$$\sigma^2(t) = \frac{q^2}{2\lambda}(1 - e^{-2\lambda t}) \tag{20b}$$

$$\rho(x, t\,|\,x_0, 0) = \frac{1}{(2\pi\sigma^2(t))^{1/2}} e^{[-(x-x_0e^{-\lambda t})^2]/[2\sigma^2(t)]}. \tag{20c}$$

We now evaluate the quantities appearing in the Smoluchowski formula [Eq. (11)]. The cell C reduces here to a one-dimensional interval, which we express in the general form $(a - \varepsilon/2, a + \varepsilon/2)$, taking for simplicity ε small. This entails that the numerator in Eq. (11) can be replaced by unity. Using the explicit forms (20a), (20c) we may write the two terms in the denominator of Eq. (11) as

$$P(C) = \int_{a-(\varepsilon/2)}^{a+(\varepsilon/2)} dx\,\rho(x) \tag{21a}$$

$$\begin{aligned} P(C, 0; C, \tau) &= \int_{a-(\varepsilon/2)}^{a+(\varepsilon/2)} dx_0\,\rho(x_0) \int_{a-(\varepsilon/2)}^{a+(\varepsilon/2)} dx\,\rho(x, \tau\,|\,x_0, 0) \\ &= \frac{1}{2}\int_{a-(\varepsilon/2)}^{a+(\varepsilon/2)} dx_0\,\rho(x_0)\left\{\mathrm{erf}\left[\frac{a + \frac{\varepsilon}{2} - x_0e^{-\lambda\tau}}{(2\sigma^2(\tau))^{1/2}}\right]\right. \\ &\quad \left. + \mathrm{erf}\left[\frac{x_0e^{-\lambda\tau} - a + \frac{\varepsilon}{2}}{(2\sigma^2(\tau))^{1/2}}\right]\right\}, \end{aligned} \tag{21b}$$

where erf$w$ represents the error function

$$\mathrm{erf}w = \frac{2}{\pi^{1/2}}\int_0^w dy\,e^{-y^2}. \tag{22}$$

For any given (small) ε, we are interested in the behavior of (21b) for small τ. As seen from Eq. (20b), in this limit $\sigma^2(t) \sim q^2\tau$ and the argument of erf in (21b) tends to infinity except for values of x_0 belonging to a thin layer of width $2^{1/2}\sigma(\tau)$ near $x_0 = a + \varepsilon/2$, for which it tends to zero. Accordingly, the function itself switches rapidly from 1 to 0. For the two pieces of (21b) the transition takes place, respectively, at $x_0 = a \pm \varepsilon/2 \mp 2^{1/2}\sigma$. Equation (21b) may therefore be replaced, effectively, by

$$P(C, 0; C, \tau) = \frac{1}{2}\left[\int_{a-(\varepsilon/2)}^{a+(\varepsilon/2)-2^{1/2}\sigma} dx_0\,\rho(x_0) + \int_{a-(\varepsilon/2)+2^{1/2}\sigma}^{a+(\varepsilon/2)} dx_0\,\rho(x_0)\right]. \tag{23}$$

Expanding this expression for small τ (and hence small σ) yields

$$P(C, 0; C, \tau) = P(C) - 2^{-1/2}q\tau^{1/2}\left[\rho\left(a + \frac{\varepsilon}{2}\right) + \rho\left(a - \frac{\varepsilon}{2}\right)\right]. \tag{24}$$

When inserted in the Smoluchowski formula, Eq. (11), this expression still leaves us with a τ dependence,

$$\langle\theta\rangle \sim \frac{\tau^{1/2}}{2^{-1/2}q\left[\rho\left(a + \frac{\varepsilon}{2}\right) + \rho\left(a - \frac{\varepsilon}{2}\right)\right]}. \tag{25}$$

One is thus led to the following conclusions:

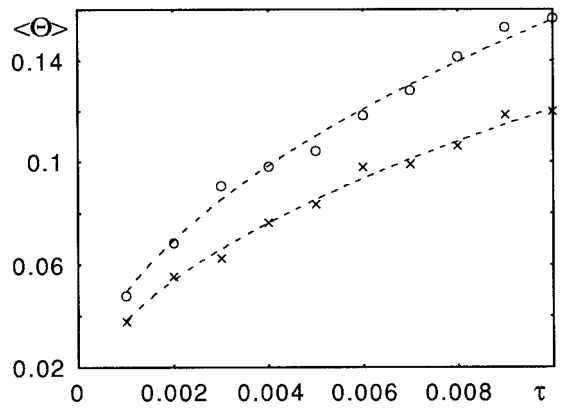

FIG. 8. Dependence of $\langle\theta\rangle$ on the integration step τ of Eqs. (19) defining an Ornstein–Uhlenbeck process for an analog x_0, $-0.01 \leq x_0 \leq 0.01$ (empty dots) and $-0.02 \leq x_0 \leq 0.02$ (crosses). Dotted line represents the power law predicted by Eq. (25). Parameter values $\lambda = 1$, $q^2 = 10^{-3}$. Number of realizations considered in each τ: 10 000.

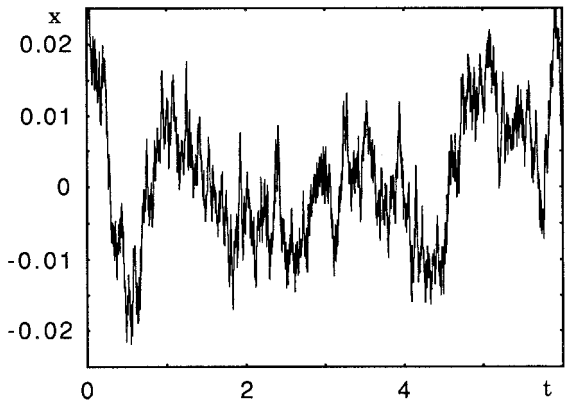

FIG. 9. A realization of the Ornstein–Uhlenbeck process, Eqs. (19), using a time step $\tau = 0.001$ time units, illustrating the tendency of the trajectory to recur infinitely often to any given initial state. Parameter values as in Fig. 7.

1) $\langle\theta\rangle$ is not determined entirely by the system's intrinsic parameters but keeps an explicit τ dependence. In particular, if τ is chosen to be the integration step used in solving Eq. (19), $\langle\theta\rangle$ is resolution dependent. We have evaluated $\langle\theta\rangle$ explicitly from the stochastic trajectory generated by Eq. (19). Specifically, we have chosen an analog of size ε in the space of the variable x surrounding the origin, which corresponds to the stable steady state of Eq. (19) in the absence of the random force, $F(t)$. Adopting $\lambda = 1$ and $q^2 = 10^{-3}$, a large number of realizations, starting at $x = 0$, have been generated and the recurrence of the trajectory for each realization to the analog considered has been registered.

Figure 8 depicts the mean recurrence times obtained for $\varepsilon = 0.02$ (crosses) and $\varepsilon = 0.04$ (circles) using 10 000 realizations as a function of the integration time step τ. The results confirm entirely the theoretical scaling law (dashed line) predicted by Eq. (25). A close look at a single realization, depicted in Fig. 9, allows us to reach a qualitative explanation: We see that the process recurs many times around a given local ''macrostate'' associated to the smoothing of the fluctuations, before being suddenly driven to a new cluster of states. Actually one can show that for a diffusion process (of which the Ornstein–Uhlenbeck process is a particular case) the system recurs infinitely often to any given initial state, entailing that $\langle\theta\rangle$ should tend to zero in the limit $\tau \to 0$, as stipulated by Eq. (25) (Feder 1988). As a by-product, events associated with finite excursions of the relevant variables—including extreme events for which recurrence times are finite and often very long—cannot be modeled by a linear regression equation driven by white noise.

2) $\langle\theta\rangle$ remains finite in the limit $\varepsilon \to 0$. This reflects the fact, well known in the literature on stochastic process (Feller 1967), that diffusion process in one and two dimensions recurs with probability one in any given state. This would be impossible in a deterministic aperiodic process [cf. Eq. (17)].

3) Recurrence is conditioned by the noise strength q rather than by the correlation time λ^{-1}. At first sight, this conclusion seems surprising since one would tend to predict the waiting time of analogs on the basis of the correlation time of the underlying processes. In actual fact, the result is more in line with the analysis of the preceding sections, in the sense that information on the complexity of the underlying dynamics (accounted by the fractal attractor in section 3 and 4) is now compressed into the single variable q.

6. Concluding remarks

Despite their importance in the understanding of atmospheric variability, analogs remain poorly understood, owing primarily to the limited amount of information available in the relevant data libraries. Statistical ideas, based in one way or another to central-limit-type theorems, have so far dominated theoretical approaches to the problem.

In this paper we attempted a dynamical approach to analogs, with special emphasis on the probability of their occurrence and the associated waiting times. The main quantities appearing in the theory are the invariant probability for the system to be in a part of phase space whose linear dimension is given by the confidence interval used to define the analogs, ε, and the probability to remain in this part during the sampling time τ. This allowed us to arrive at scaling laws expressing the mean analog waiting times in terms of ε, the attractor dimension d, the number of variables involved, and, when the continuous time limit $\tau \to 0$ was taken, the system's intrinsic timescales. The theory was fully corroborated by numerical simulations on Lorenz's three-variable thermal convection model. The simulations brought out two further important features: a strong dependence of waiting times on the local properties

of the attractor and their high variability around the mean in a given phase space cell of fixed size.

The above results provide some interesting new insights on analogs in the real atmosphere. First, the scaling law derived extends the result previously obtained by Van den Dool (1994) while providing a firm connection with the underlying dynamics. Second, our findings draw attention on the limitations of the usefulness of the mean recurrence times in prediction since the dispersion around this mean is typically comparable to the mean.

The extension of the original purely deterministic formulation to a red noise atmosphere led to a rather unexpected resolution dependence of mean waiting times. On the practical side, this entails that simple regression models driven by noise are not suitable for describing the recurrence dynamics of finite size disturbances, of which extreme events provide a particularly important class.

Future investigations in this area should aim at the analysis of more detailed models of atmospheric dynamics in the perspective of our formulation. The main problem here is how to cope with the large number of variables involved. We foresee two ways out: either the dynamics happens to reduce to a low-dimensional attractor or, otherwise, a cluster analysis is first performed and serves to define the ''states'' on which the analog problem is to be formulated.

Stochastic models of the diffusion type are bound to present the behavior found in section 5. An alternative can be envisaged in connection with the problem of transitions between simultaneously stable states. It would consist in smoothing out small-scale variability and view the transitions as a jump process similar in some respect to a random telegraph signal, but in which the transition probabilities should bear the signature of the small-scale variability. Such ''adiabatic approximations'' have already been suggested in the context of stochastic resonance models of Quaternary glaciations (Nicolis 1982).

Acknowledgments. I am indebted to referee B for his thorough analysis of the manuscript and his insightful comments. This work is supported, in part, by the Belgian Federal Office of Scientific, Technical and Cultural Affairs under the *Pôles d'Attraction Interuniversitaires* program.

APPENDIX

Extended Central Limit Theorem

In this appendix it is shown how a Gaussian distribution governing distances in phase space can emerge out of N mutually independent variables x_i having a common distribution $f(x)$ that need not be Gaussian. We follow a method similar to that used in the proof of the classical central limit theorem (Feller 1971).

We define distances u in the N-dimensional space of $\{x_i\}$ through the Euclidean norm

$$u = \sqrt{x_1^2 + \cdots + x_N^2}. \tag{A1}$$

By definition, the probability density ρ of the variable $w = u^2$ is (assuming for simplicity that $\{x_i\}$ are distances from their mean values, so that their range of variation is $-\infty$ to ∞)

$$\rho(w) = \int_{-\infty}^{\infty} dx_1 \cdots \int_{-\infty}^{\infty} dx_N \times \delta(w - (x_1^2 + \cdots + x_N^2)) f(x_1) \cdots f(x_N). \tag{A2}$$

Using the familiar representation of the delta function one may further write Eq. (A2) as

$$\rho(w) = \frac{1}{2\pi} \int_{-\infty}^{\infty} dk \exp(ikw) \times \left(\int_{-\infty}^{\infty} dx \exp(-ikx^2) f(x) \right)^N. \tag{A3}$$

In the following we assume that $f(x)$ decays to 0 sufficiently rapidly as $|x| \to \infty$ to have finite moments up to fourth order. Expanding $\exp(-ikx^2)$ in powers of $(-ikx^2)$ and retaining the first few terms, one then obtains

$$\begin{aligned} &\int_{-\infty}^{\infty} dx \exp(-ikx^2) f(x) \\ &\quad \approx \int_{-\infty}^{\infty} dk \left(1 - ikx^2 - \frac{k^2x^4}{2}\right) f(x) \\ &\quad = 1 - ikm_2 - \frac{k^2}{2} m_4, \end{aligned} \tag{A4}$$

where m_2 and m_4 being the second and fourth moments.

The expression involved in the integral over k in Eq. (A3) becomes

$$\begin{aligned} &\left(\int_{-\infty}^{\infty} dx \exp(-ikx^2) f(x) \right)^N \\ &\quad = \left(1 - ikm_2 - \frac{k^2}{2} m_4\right)^N \\ &\quad = \exp N \ln\left(1 - ikm_2 - \frac{k^2}{2} m_4\right). \end{aligned} \tag{A5}$$

Expanding the logarithm and retaining dominant terms yields

$$\begin{aligned} &\ln\left(1 - ikm_2 - \frac{k^2}{2} m_4\right) \\ &\quad \approx -ikm_2 - \frac{1}{2} k^2 (m_4 - m_2^2). \end{aligned} \tag{A6}$$

Substituting (A5)–(A6) into (A3) one obtains

$$\rho(w) = \frac{1}{2\pi}\int_{-\infty}^{\infty} dk\ \exp[ik(w - Nm_2)] \times \exp\left[-\frac{N}{2}k^2(m_4 - m_2^2)\right]. \quad \text{(A7)}$$

Knowing that the Fourier transform of a Gaussian is itself Gaussian, we conclude that $\rho(w)$ is a Gaussian distribution with mean M and variance V given by

$$M = Nm_2$$
$$V = N(m_4 - m_2^2). \quad \text{(A8)}$$

Alternatively, the normalized variable

$$N^{1/2}\left(\frac{1}{N}\sum_{i=1}^{N} x_i^2 - m_2\right)$$

is a normal distribution with zero mean and variance equal to $m_4 - m_2^2$.

It is important to realize that the truncation of (A6) to terms of order k^2 is legitimate only in the limit of large N, in the sense that higher-order terms would give in this limit additional contributions to (A7) that would be negligibly small.

The distribution $\tilde{\rho}(u)$ of the norm u, Eq. (A1), follows straightforwardly from (A7)–(A8) and the change of variable $w = u^2$

$$\tilde{\rho}(u) = \rho(w(u))\left|\frac{dw(u)}{du}\right|$$

or, in a more explicit form,

$$\tilde{\rho}(u) = A^{-1}u\ \exp\left[-\frac{(u^2 - Nm_2)^2}{2N(m_4 - m_2^2)}\right], \quad \text{(A9)}$$

where A is a normalization factor. As it stands, this distribution is not Gaussian for the variable u. However, setting

$$\tilde{\rho}(u) = A^{-1}\exp[V(u)], \quad \text{(A10)}$$

where

$$V(u) = -\frac{(u^2 - Nm_2)^2}{2N(m_4 - m_2^2)} + \ln u,$$

one easily checks that $V(u)$—and hence $\tilde{\rho}(u)$ itself exhibits for $N \to \infty$ a sharp maximum at $u^* = (Nm_2)^{1/2}$. Expanding around this maximum reduces then $\tilde{\rho}(u)$ asymptotically to a Gaussian distribution.

The extent to which the large N limit is realized in a problem of atmospheric interest depends, of course, on the number of relevant independent degrees of freedom or, alternatively, on the attractor dimension. As we saw in section 4, the statistics of the Lorenz model is definitely non-Gaussian. This conclusion obviously extends to all phenomena amenable to a low-dimensional attractor. On the other hand, the full-scale analysis of a variety of atmospheric phenomena is likely to introduce a high-dimensional dynamics. In such cases Gaussian statistics can legitimately be applied for global quantities as u, although there is no reason for the individual variables $\{x_i\}$ themselves to be Gaussian.

REFERENCES

Barnett, T. P., and R. W. Preisendorfer, 1978: Multifield analog prediction of short-term climate fluctuations using a climatic state vector. *J. Atmos. Sci.,* **35,** 1771–1787.

Barry, R. G., and A. H. Perry, 1973: *Synoptic Climatology Methods and Applications.* Methuen, 555 pp.

Feder, J., 1988: *Fractals.* Plenum, 283 pp.

Feller, W., 1967: *An Introduction to Probability Theory and Its Applications.* Vol. I. Wiley, 509 pp.

——, 1971: *An Introduction to Probability Theory and Its Applications.* Vol. II. Wiley, 669 pp.

Frankignoul, C., and K. Hasselmann, 1977: Stochastic climate models. Part II: Application to sea-surface temperature anomalies and thermocline variability. *Tellus,* **29,** 289–305.

Gardiner, C., 1983: *Handbook of Stochastic Methods.* Springer, 442 pp.

Kac, M., 1959: *Probability and Related Topics in Physical Sciences.* Interscience, 266 pp.

Kim, K.-Y., and G. R. North, 1991: Surface temperature fluctuations in a stochastic climate model. *J. Geophys. Res.,* **96,** 18 573–18 580.

Kruizinga, S., and A. H. Murphy, 1983: Use of an analogue procedure to formulate objectively probabilistic temperature forecasts in the Netherlands. *Mon. Wea. Rev.,* **111,** 2244–2254.

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

——, 1969: Atmospheric predictability as revealed by naturally occurring analogs. *J. Atmos. Sci.,* **26,** 636–646.

Nicolis, C., 1982: Stochastic aspects of climatic transitions—Response to a periodic forcing. *Tellus,* **34,** 1–9.

——, and G. Nicolis, 1986: Effective noise of the Lorenz attractor. *Phys. Rev.,* **34A,** 2384–2390.

——, W. Ebeling, and C. Baraldi, 1997: Markov processes, dynamic entropies, and the statistical prediction of mesoscale weather regimes. *Tellus,* **49A,** 108–118.

Ruosteenoja, K., 1988: Factors affecting the occurrence and lifetime of 500 mb height analogs: A study based on a large amount of data. *Mon. Wea. Rev.,* **116,** 368–376.

Sparrow, C., 1982: *The Lorenz Equations.* Springer, 269 pp.

Toth, Z., 1991a: Estimation of atmospheric predictability by circulation analogs. *Mon. Wea. Rev.,* **119,** 65–72.

——, 1991b: Circulation patterns in phase space: A multinormal distribution? *Mon. Wea. Rev.,* **119,** 1501–1511.

Vallis, G. K., 1988: Conceptual models of El Niño and the Southern Oscillation. *J. Geophys. Res.,* **93,** 13 979–13 991.

Van den Dool, H. M., 1994: Searching for analogs, how long must we wait? *Tellus,* **46A,** 314–324.

Wallace, J. M., X. Cheng, and D. Sun, 1991: Does low frequency atmospheric variability exhibit regime-like behavior? *Tellus,* **43A,** 16–26.