Complexity and Predictability of Hourly Precipitation

J. B. ELSNER

Department of Meteorology, Florida State University, Tallahassee, Florida

A. A. TSONIS

Department of Geosciences, University of Wisconsin—Milwaukee, Milwaukee, Wisconsin

(Manuscript received 3 December 1991, in final form 20 April 1992)

ABSTRACT

Recent studies have shown how concepts from information theory can be applied to climate models to better understand the problem of climate prediction. This paper describes how information theory, specifically the concept of entropy, can be used in the analysis of short-term precipitation records. The ideas are illustrated through analysis and comparisons of two long, hourly precipitation records. From the results it is concluded that the records are not periodic and are definitely more complex than records of random origin. This complexity, however, arises from underlying deterministic rules indicating the potential for predictability.

1. Introduction

In climate and weather studies, much effort is spent on distinguishing signal (deterministic) from "noise" (stochastic). The premise is that these two components are related in a trivial (read: linear) way and thus can be separated using linear methods. The role of nonlinearities in generating irregularity and variability is, however, now widely recognized (Shukla 1985). Recently it has been accepted that in order to describe complex dynamical behavior such as weather and climate—where the system intermittently becomes more sophisticated instead of more random—it is important to have a measure of the degree or level of complexity (see, e.g., Crutchfield and Packard 1983; Wolfram 1984; Grassberger 1986; Loyd and Pagels 1988; Li 1991).

Deterministic chaos provides a useful model for variability of weather and climate (Lorenz 1984) since it not only gives rise to aperiodicity in space and time but also displays sensitivity to initial conditions (Nicolis 1990). Motivation, therefore, comes from the instinct (derived from chaos theory) that linear methods, or methods that have their root in linear statistics, may be insufficient in many situations for explaining weather and climate variability. We thus seek more general methods for describing such variability.

The objective of the present paper is to examine the possibility of using the concept of entropy for the problem of assessing complexity and predictability of precipitation records. Goals are similar to those of Leung and North (1990), who introduced information theory for the study of climate prediction. The major contribution of the present work is the application of these important concepts to actual data records. The paper presents a new data analysis never performed before on precipitation records.

In particular, we are interested in a multiple application of a definition of entropy that will be useful for defining complexity. Simple deterministic dynamical systems can exhibit very complex behavior often resembling that of random processes. In an effort to distinguish one from the other, new approaches that estimate the complexity associated with periodic, random, and chaotic sequences have lately been developed. After providing a data description in section 2, a hierarchical approach for the estimation of complexity is presented and employed on long and continuous short-term precipitation records in section 3. It is concluded that the records are not periodic and definitely more complex than records of random origin. In section 4 interpretation of the results is presented centering on the idea of underlying deterministic rules and predictability. A summary is provided in section 5.

2. Data

Since the definitions described in the next section are based on a limit as the length of record goes to infinity, we searched for long, continuous data records.

Corresponding author address: Dr. James B. Elsner, Department of Meteorology, B-161, Florida State University, Tallahassee, FL 32306.

For this pilot study we chose two stations, Milwaukee, Wisconsin (MKE; ~43°N, 88°W)—influenced primarily by midlatitude weather systems—and West Palm Beach, Florida (PBI; ~27°N, 80°W)—influenced by both midlatitude and tropical weather systems. Both stations have 40 years of nearly continuous hourly observations of accumulated precipitation. The number of missing hours for both stations was less than 0.1% of the total. Missing reports were treated as no precipitation except in situations where it was obvious that a missing report occurred during a wet period, in which case a representative amount was assumed based on earlier and later reports. The period of data coverage for the MKE record is August 1948-November 1987 (344 786 hours) and for the PBI record is January 1949-June 1988 (346 224 hours). Precipitation is accumulated for each hour, after which a depth in hundredths of inches is measured. Precipitation falling as sleet, snow, or hail is melted before a depth is recorded. A small portion of both hourly records used in this study is shown in Fig. 1.

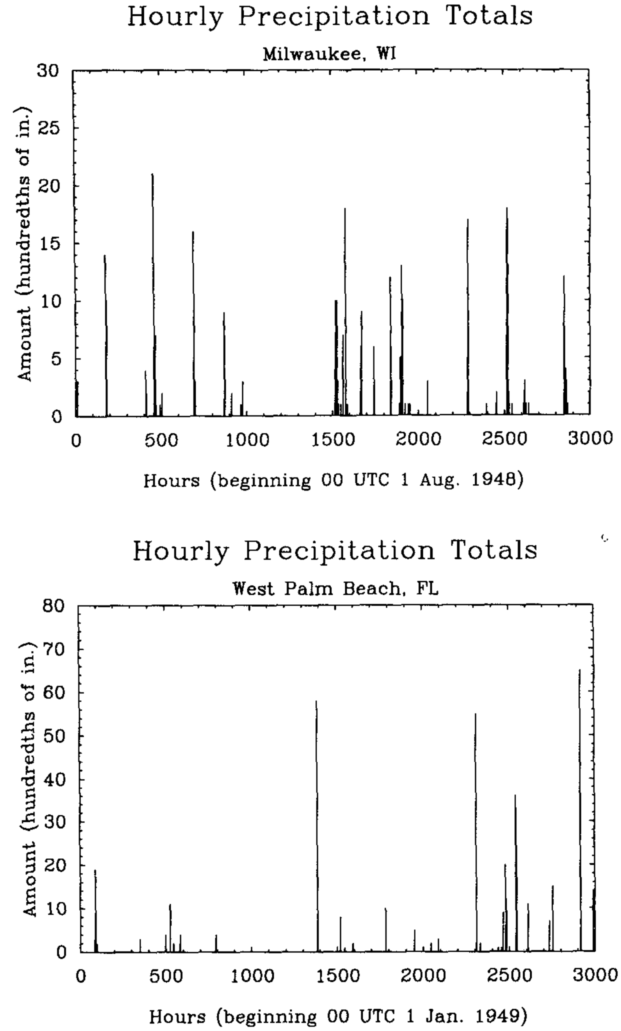

FIG. 1. (a) A portion of the hourly precipitation record from Milwaukee, Wisconsin (MKE). The time axis is given as successive hours beginning 1 August 1948 and the vertical axis is the hourly accumulated precipitation totals in hundredths of inches. (b) Same as (a) except for West Palm Beach, Florida (PBI), beginning 1 January 1949.

3. A measure of complexity

What is meant by simplicity? What is meant by complexity? Is it possible to objectively measure complexity? Recent efforts to understand deterministic chaos have certainly muddled the issue. Simple systems with only a few degrees of freedom can show very complex behavior. When faced with behavior that is labeled as "complex," it would be instructive to have some objective measure of just how complex it is. Classical ideas of complexity are based on the concept of entropy or information (Kolmogorov 1965; Chaitin 1966). With such definitions, pure randomness will maximize complexity. As Grassberger (1986) points out, however, since the generation of randomness is characterized by a lack of rules, a useful definition of complexity must allow one to recognize randomness as a rather simple pattern. Indeed, Grassberger argues, with the help of three two-dimensional patterns, that the pattern one would call the most complex is neither the one with the highest entropy nor the one with the lowest; however, this notion must be somehow coupled to the intuition inherent in the definition of entropy that randomness is more complex than periodicity.

In an attempt to reconcile these intuitive notions, more complicated measures of complexity have recently emerged. In particular, D'Allessandro and Politi (1990, hereafter referred to as DP) have developed a unique approach in which complexity is measured by using a hierarchy of numbers as opposed to a single measure. They demonstrate the usefulness of such a definition for describing iterative maps, while cautioning that the problem of characterizing complexity is by no means solved. Nevertheless, it is noted that their method may provide a different way of viewing physical systems. In this section we describe the procedure of DP and apply it to estimating the complexity of short-term precipitation at two stations—one in the midlatitudes and the other in the subtropics.

a. Method

The problem of defining complexity can be reduced to characterizing the data as a binary sequence of zeros and ones. Consider the following two binary sequences, for example:

0101010101010101

1001101001011001.

The first sequence can be described simply as repetitions of the "word" 01, whereas for the second string no simple description can be made. To overcome the difficulty of defining complexity as previously outlined, the complexity of a sequence can be measured by a hierarchy of numbers, the first of which is equivalent to the entropy, where entropy can be defined as the logarithm of the number of words of length n that are found in the sequence (admissible words) as n becomes large (Maddox 1990). At the second level there is an analogous definition but this time for forbidden words of increasing word length.

Specifically, we let $N_p(n)$ be the number of possible words of length n. For $n = 1$, the number of possible words is 2, namely, 0 and 1, and for $n = 2$, $N_p(n) = 4$ (i.e., 00, 01, 10, and 11). One can think of $N_p(n)$ as the size of the dictionary for a given word length, which for binary sequences equals 2^n. Then, following closely the notation of DP and given a long sequence, we let $N_a(n)$ be the number of admissible words of length n contained in the sequence. Words of length n are formed by successively shifting a window of size n through the sequence. For example, the sequence 0010101 contains three different words of length two, namely, 00, 01, and 10, out of the set of size four of possible words. The word 11 is missing (or forbidden), since it is not contained in the sequence. In general, we assume that we have an infinite sequence and if a word is not included then it is termed forbidden. With these definitions it is possible to define first-order complexity as

$$C^1 = \lim_{n\to\infty} \log[N_a(n)]/n, \qquad (1)$$

which is equivalent to the topological entropy. It is a useful notion of complexity, for it allows one to easily distinguish between periodic sequences (C^1 small) and random sequences (C^1 large), but it has limitation since, as DP note, it fails to recognize random sequences as fairly simple objects.

This problem is handled at the next level, where the possibility of rules governing the sequence is examined. One way of defining a rule for a given sequence is that it should not contain irreducible forbidden words, where irreducible means that the word does not contain any shorter forbidden word. In the prior example, the word 11 is a forbidden word of length two. Since 11 consists of a single admissible word twice (i.e., two 1's), besides being a forbidden word it is also an irreducible forbidden word of length two. At word length three we discover more forbidden words, some of which include the shorter forbidden words and thus are reducible (e.g., 011 is reducible while 000 is irreducible).

Similar to what was done for admissible words, the second-order complexity is defined as

$$C^2 = \lim_{n\to\infty} \log[N_{if}(n)]/n, \qquad (2)$$

where $N_{if}(n)$ is the number of irreducible forbidden words of length n. For an infinite and completely random sequence there should be no forbidden words and thus C^2 will be zero, whereas for a chaotic dynamical system C^2 should in general be nonzero.

A general description of the procedure is as follows. First, at level one it is possible to establish whether the sequence is more complex than a periodic sequence by comparing the number of admissible words as the length of the word increases. Similarly then, at the second level it is possible to establish whether the sequence is more complex than a random sequence by comparing the number of irreducible forbidden words as a function of word length. In this two-step procedure it is possible to determine an objective measure of complexity on a scale in which periodicity is more simple than randomness and randomness is more simple than deterministic chaos. In fact, the second-order complexity C^2 is related to the chaotic properties of the underlying attractor via the relation

$$C^2 = \lambda_+ D/(1 + D), \qquad (3)$$

where D is the dimension of the attractor and λ_+ is the positive Lyapunov exponent (DP).

Along these lines, we mention the recent interest in the study of Markov chains for understanding the behavior of chaotic dynamical systems (e.g., Nicolis 1990; Destexhe 1990). Assuming an adequate partition of the data, the order of the corresponding Markov process will give useful information about the time correlations of the underlying dynamical system. We have not explored this idea and only note that in the way in which complexity was defined here, only the structure of the sequence is considered without taking into account the probabilities associated with the different words (orbits) as is the case with Markov chains.

b. Results

We apply the preceding definitions to two long hourly precipitation records (MKE and PBI). First we reduce the data to binary sequences and make some general comments concerning the records as viewed in this manner. The raw data are reduced to a binary sequence by assigning zeros and ones to hours with no measurable precipitation and to hours with measurable precipitation, respectively.

We first note that for the MKE record approximately 93.3% of the hours are without precipitation, leaving a total of 23 136 wet hours, or better than 2.5 years of precipitation, and for the PBI record approximately 95.0% of the hours are without precipitation, leaving a total of 17 150 wet hours, or approximately 2 years of rainfall.

Having coarse grained the precipitation records into binary sequences, we first count the number of admissible words in the sequence as the word size increases. For the MKE record the first nonadmissible (forbidden) words occur at word length nine and there are eight of them, namely 010101001, 010101011, 010110001, 011010101, 101010101, 101011011, 110010101, and 110101010. For the PBI record the first forbidden words occur at word length eight and there are three of them, namely 10100101, 10101010, and 11010101. None of these forbidden words contains greater than three consecutive hours of either precipitation or no precipitation, indicating that the nature of precipitation on this time scale over both Milwaukee and West Palm Beach tends to be continuous rather than temporally sporadic.

Since the number of admissible words increases with increasing word length we can immediately recognize the hourly precipitation records as more complex than a periodic sequence. For comparison, we created a periodic sequence in which the ratio of the number of wet hours to the number of dry hours is roughly equivalent to that of the precipitation records. The periodic sequence consists of 14 consecutive hours of no precipitation, followed by one hour of precipitation, followed by 14 more hours of no precipitation, and so on. A comparison is made in Fig. 2 where the solid line and dashed line represent the scaling rates of the number of admissible words for the precipitation records and the dashed-dotted line represents the scaling rate for the synthetic periodic sequence. For both random or deterministic chaotic sequences the number of admissible words will continue to increase, whereas for any periodic or quasi-periodic sequence the number of admissible words is limited.

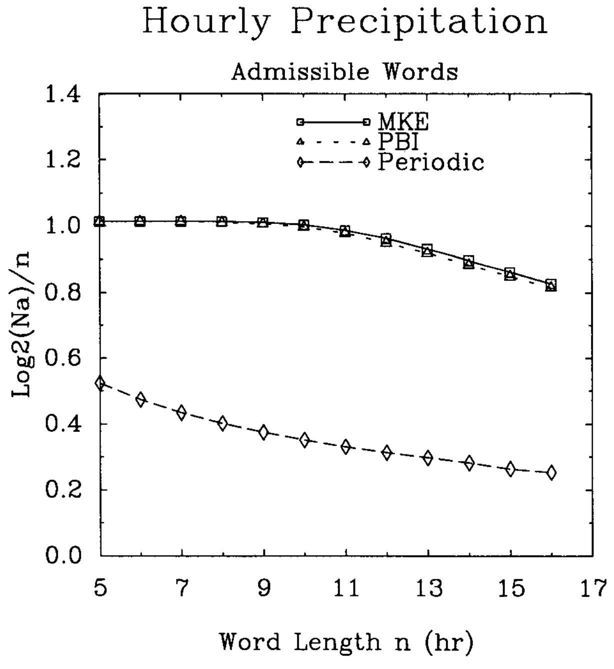

FIG. 2. Growth rate of the number of admissible words as a function of word length for the precipitation records (MKE: solid line; PBI: dashed line) and for a periodic sequence (dash-dotted line). Note the stable growth rate for the precipitation records contrasts with a decay for the synthetic periodic record.

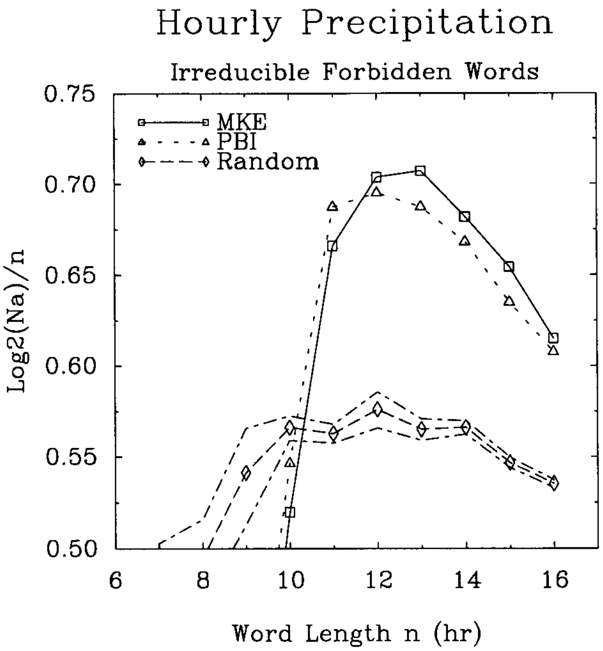

FIG. 3. Growth rate of the number of irreducible forbidden words as a function of word length for the precipitation records (MKE: solid line; PBI: dashed line) and for a random sequence (dash-dotted line). The random sequence curve is generated by pseudorandomly permuting the MKE precipitation record five times and averaging the number of irreducible forbidden words at each word length. A confidence line (dotted) is estimated by ± one standard deviation from the mean of the five random permutations. If we replace the limit with the supremum then from the graphs we can conclude that the precipitation records have larger second-order complexity measures than a random sequence.

The next step is to monitor the growth rate of the number of irreducible forbidden words. For the MKE record, since at word length nine we find the first forbidden words, all of them must be irreducible. For comparison we pseudorandomly permute the MKE hourly observations, creating a sample ($N = 5$) of random records having the same length and distributions as the precipitation record (results are nearly identical if we randomly permute the PBI record). The number of irreducible forbidden words for the MKE precipitation record is represented by a solid line in Fig. 3, the number of irreducible forbidden words for the PBI record is represented by a dashed line, and the average number of irreducible forbidden words for the random sequences is indicated by a dash-dotted line. The dotted lines indicate the error bars (one standard deviation). The precipitation and random records show a large increase in the number of irreducible forbidden words for relatively short word lengths reaching a maximum near word length 12. Since we are working with finite length sequences, we can replace the limit in Eq. (2) with the supremum. By this definition it is clear that the precipitation records are more complex than similar random sequences. It thus appears that, at least for these particular examples, the definitions of DP are adequate for describing the outcome of a complicated deterministic process as more complex than one that is inherently random.

We have also tested DP's notion of complexity against random strings with serial correlations identical to the actual records. To do this we divided the precipitation records into subsequences of length two and randomly shuffled the subsequences, while preserving the order within each subsequence. This procedure destroys the long-range correlations while maintaining the shortest-range correlations. We have also divided the precipitation records into subsequences of length five and randomly shuffled. We then repeated the analysis of irreducible forbidden words as described above for these two random autocorrelated sequences. Results show that the suprema of the autocorrelated noises are below the suprema of the precipitation records by at least a factor of 2. Further, a sensitivity analysis, where we increased the number of "missing" reports by a factor of 2 and then 4 and repeated the analysis, indicated only a few percent difference in the value of the suprema in Fig. 3.

4. Discussion

The interpretation of these results centers on the possibility of underlying hidden rules operating on increasing time scales. If there is a rule for words of length four—for instance, the word 0101 is not allowed—then there will be words of length five such as 10101 that will not be found in the record. This word is reducible; that is, it contains a forbidden word of length four. In addition to rules for words of length four there may be rules for words of length five (longer time scale) but these will generate irreducible forbidden words. Because the random sequence has finite length, there will be some forbidden words, but no rules; therefore, the number of irreducible forbidden words will be less than a sequence that is dictated to some extent by rules, as must be the case for the precipitation record. This explanation points to a limiting factor of this methodology: the length of the available record. If the record is short, then the number of irreducible forbidden words as a result of rules will be swamped by those as a result of the relatively small sequence length.

The existence of deterministic rules relates to predictability. The fact that the analysis indicates or not the existence of rules allows a distinction to be made between "simple" sequences (random) and "complex" sequences (precipitation). The finding that MKE's precipitation, however, appears to be more complex (higher second-order complexity) than the precipitation of PBI indicates more rules governing the fall of precipitation over MKE at this short time scale. More rules suggest a greater deterministic component and perhaps better predictability.

5. Summary

Time series that are characterized by broadband spectra are notoriously difficult to analyze. Traditional methods such as Fourier analysis or other linear transforms and filters usually fail to offer many new insights into the underlying structure of such time series (Savit and Green 1991). Motivated by chaos theory, in this paper we have explored a method from information theory to measure the degree to which short-term precipitation records are predictable—in other words, the extent to which the records are driven by reproducible deterministic dynamics given previous values of the record.

In particular, we have applied the hierarchical procedure of D'Alessandro and Politi (1990) to estimate the complexity of precipitation at two stations. Data were hourly precipitation amounts at Milwaukee, Wisconsin (MKE), and West Palm Beach, Florida (PBI), over the past 40 years. We conclude that, as a result of underlying rules inherent in precipitation formation mechanisms, the observed records, in terms of definitions employed here, are more complex than a record of random origin with the MKE precipitation displaying a bit more complexity than the PBI precipitation. The consequence of this result is that at the next step it may be possible to extract the rules inherent in the dynamical system and thus ultimately improve prediction. For instance, in the aforementioned example, if one observes the sequence 1010, then the probability that the next value is 1 is quite small. Here lies the potential of such techniques for improving predictability.

Finally, we mention that by applying the preceding method to a number of different stations over, say, the United States, we will be able to construct a contour map of precipitation complexity. This will allow one to determine regions of relatively high predictability from those of relatively low predictability. Our current research is in this area.

Acknowledgments. Thanks are extended to Mr. J. Roth for retrieving the MKE data, and to Dr. T. Gleason and the Florida Climate Center for retrieving the PBI data used in this study.

REFERENCES

Chaitin, G., 1965: On the length of programs for computing finite binary sequences. *J. Assoc. Comput. Mach.*, **13**, 547–569.

Crutchfield, J. P., and N. H. Packard, 1983: Symbolic dynamics of noisy chaos. *Physica D*, **7**, 201–223.

D'Alessandro, G., and A. Politi, 1990: Hierarchical approach to complexity with applications to dynamical systems. *Phys. Rev. Lett.*, **64**, 1609–1612.

Destexhe, A., 1990: Symbolic dynamics from biological time series. *Phys. Lett. A*, **143**, 373–378.

Grassberger, P., 1986: Toward a quantitative theory of self-generated complexity. *Int. J. Theor. Phys.*, **25**, 907–938.

Kolmogorov, A. N., 1965: Three approaches to the definition of the concept "Quantity of information." *Prob. Inf. Transm.*, **1**, 1–7.

Leung, L.-Y., and G. R. North, 1990: Information theory and climate prediction. *J. Climate*, **3**, 5–14.

Lloyd, S., and H. Pagels, 1988: Complexity as thermodynamic depth. *Ann. Phys.*, **188**, 186–213.

Li, W., 1991: On the relationship between complexity and entropy for Markov chains and regular languages. *Comput. Sys.*, **5**, 381–399.

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

Maddox, J., 1990: Complicated measures of complexity. *Nature*, **344**, 705.

Nicolis, C., 1990: Chaotic dynamics, Markov processes and climate predictability. *Tellus*, **42A**, 373–378.

Savit, R., and M. Green, 1991: Time series and dependent variables. *Physica D*, **50**, 95–116.

Shukla, J., 1985: Predictability. *Advances in Geophysics, Issues in Atmos. and Ocean Modeling. Part B: Weather Dynamics*, S. Manabe, Ed. Academic Press, 87–123.

Wolfram, S., 1984: Computational theory of cellular automata. *Commun. Math. Phys.*, **96**, 15–57.