Available online at www.sciencedirect.com
ScienceDirect
Procedia Computer Science 00 (2011) 1–10
Procedia Computer Science

Block-Entropy Analysis of Climate Data

J. Walter Larson[a,b,c], Peter R. Briggs[d], Michael Tobis[e]

[a]*Mathematics and Computer Science Division, Argonne National Laboratory*
[b]*Computation Institute, University of Chicago/Argonne National Laboratory*
[c]*Research School of Computer Science, The Australian National University*
[d]*CSIRO Marine and Atmospheric Research*
[e]*Institute for Geophysics, Jackson School of Geosciences, University of Texas at Austin*

Abstract

We explore the use of block entropy as a dynamics classifier for meteorological timeseries data. The block entropy estimates define the entropy growth curve $H(L)$ with respect to word length L. For a finitary process, the entropy growth curve tends to an asymtotic linear regime $H(L) = E + h_\mu L$, with entropy rate h_μ and excess entropy E. These quantities apportion the system's information content into "memory" (E) and randomness (h_μ). We discuss the challenges inherent in analyzing weather data using symbolic techniques. We apply the block entropy-based techniques to Australian daily precipitation data from the Patched Point Dataset station record collection and version 3 of the Australian Water Availability Project analysis dataset. Preliminary results demonstrate $h_\mu$and E are viable climatological classifiers for precipitation, with station records from similar climatic regimes possessing similar values of $h_\mu$and E. The entropy rates of convergence analysis rules out finite order Markov processes for orders falling within the range of block sizes considered. Differences between entropy parameters associated with station records and analyses at station locations may provide clues regarding analysis error sources.

Keywords:
Symbolic Dynamics; Timeseries Analysis; Climate Predictibility

1. Introduction

The climate system is complex, and its underlying dynamics are at best a manifestation of finite-dimensional chaos. Much effort has been devoted to climate modeling, numerical weather prediction, data assimilation, and statistical analysis, but the prediction horizon for skillful weather forecasting is around two weeks. Markov models of climate processes assume an underlying fusion of determinism and randomness in the system's dynamics. A method for potentially disentangling the "random" and "ordered" elements in meteorological timeseries is thus attractive. Model-data evaluation provides another motivation for techniques capable of separating sources of discrepancies between model output and observations.

One possible approach is information theory, which provides powerful tools for quantifying the information content of systems (the shannon entropy) and methodologies for quantifying agreement between corresponding data (comparing timeseries or spatial patterns element by element using mutual information) or probability mass or density functions (the Kullback-Leibler divergence). Information theory has been applied to the climate system in many ways, with studies of similarity [1, 2], predictability [3], complexity [4], and variability [5]. These tools are at their strongest when applied to *discrete* data; that is, data involving a finite number of state values.

Another powerful technique is symbolic dynamics [6, 7, 8], by which a continuous timeseries $x(t)$ is reduced to a *sequence* of *symbols* and the progression of these symbols is analyzed. This reduction technique is called *coarse-graining* of *symbolization* [9]. The symbols correspond to discretization of the range x into a finite, integral number of N *states* $\mathcal{A} = \{S_1, \ldots, S_N\}$, where $\mathcal{A}$ is the sequence's *alphabet*; the size of this alphabet is $|\mathcal{A}| = N$. The time domain t is discretized as well, with $t \to \{t_1, \ldots, t_N\}$, and $t_{i+1} - t_i = \Delta t \quad \forall i$. Time evolution for the sequence becomes a *shift operation* $t_i \to t_{i+1}$. Thus, transformation of $x(t)$ into a symbol sequence yields $\{X_1, X_2, \ldots, X_j \ldots\}$ with $X_i \in \mathcal{A} \quad \forall i$.

Symbol-based analysis of continuous-valued data may seem drastic, but has many strengths. It is less sensitive to measurement noise than are continuum-based techniques (e.g., spectral analysis). Moreover, reduction from floating-point data to a limited symbol alphabet can increase the computational speed while reducing the memory footprint. Symbol techniques thus are well suited to emerging novel computing architectures such as field programmable gate arrays, graphical processing units, and the next generation of leadership-class supercomputers with low memory per node. Symbolic dynamics has been applied widely—and with great success—to communications, bioinformatics, dynamical systems theory, and many other fields [9].

The main challenge in applying symbolic dynamics to a continuous timeseries is the choice of a symbolization scheme. Common approaches include *threshold crossings* (e.g., crossing the mean or median of a timeseries) and *behavioral classes* (e.g., DNA bases). The alphabet size is indicative of how much information is retained after symbolization. A symbolization's resulting dynamics may be highly sensitive to threshold choice. We will discuss this issue further in the context of weather data in section 3.

We are exploring the use of symbolic-information-theoretic techniques to form an integrated diagnostic for a meteorological timeseries' dynamics. In particular, we are interested in how "randomness" and "structure" are apportioned in in the climate system. The present work uses block-entropy analysis combined with knowledge of entropy rates of convergence to apportion randomness versus order in a meteorological timeseries and quantify how randomness (embodied in the *entropy rate* h_μ) is exchanged for order (quantified by the *excess entropy* E) as longer blocks are used in the analysis. Block entropy has been used previously to study climate. Nicolis et al. [10] analyzed Swiss daily weather data that were coarse-grained into three classes (*convective*, *advective*, and *mixed*) and estimated block entropies and entropy rates, and ruled out a first order Markov process as the system's underlying dynamics. We have applied block-entropy analysis to Australian rainfall timeseries and have found that the excess entropy and entropy rate for a station timeseries are a classifier for the climate at that location. We also employ these tools to rule out Rth order Markov processes, where R is the length of the longest blocks used in our analyses.

2. Symbolic Dynamics, Information Theory, and Complexity

In this section we provide a brief theoretical background to the methods developed and applied in this paper.

2.1. Informaton Theory

Information theory provides mathematical tools for structural description and intercomparison of probability mass functions (PMFs) and probability density functions (PDFs) for discrete and continuous variables, respectively. Information theory was orginally developed to model communications channels [11], but the formalism has roots in Boltzmann's statistical mechanical derivation of thermodynamic entropy. Below, we define information-theoretic quantities for discrete variables; a continuum analogue for (1) exists but must be used with care. Complete introductory discussions can be found in widely-used textbooks [12, 13].

Consider a discrete variable X capable of taking on the $N > 0$ values $\{x_1, \ldots, x_N\}$. The PMF $\vec{\pi} \equiv \{\pi_1, \ldots \pi_N\}$ contains the probabilities of X taking on respectively the values $\{x_1, \ldots, x_N\}$. A PMF satisfies the following conditions: *suitability* as a probability, specifically, $0 \leq \pi_i \leq 1 \quad \forall i \in \{1, \ldots, N\}$, and *normalization*, $\sum_{i=1}^{N} \pi_i = 1$. The *Shannon entropy* (SE), or simply *entropy* is

$$H(X) = -\sum_{i=1}^{N} \pi_i \log \pi_i. \tag{1}$$

The base of the logarithm in (1) determines the units of $H(X)$, with bases 2 and e yielding *bits* and *nats*, respectively. Unless otherwise specified, all logarithms used henceforth are base 2, and we measure all information-theoretic quantities in bits.

Given a symbol sequence $\mathcal{X} = X_0, X_1, \ldots, X_{\mathscr{L}}$, we can analyze it by computing *word statistics*. Here, a "word" is a string of length L symbols, with each symbol drawn from some alphabet $\mathcal{A}$. If $|\mathcal{A}| = N$ symbols, then the number of possible words of length L is $\mathcal{N}_W(L) = N^L$. The distribution of words of length $L \ll \mathscr{L}$ within $\mathcal{X}$ can be measured as follows. A sliding window of length L shifts through $\mathcal{X}$ one symbol at a time; each shift reveals a new word of length L, which is identified with index $w \in \{1, \ldots, N^L\}$, and its counter n_w is incremented. A total of $\mathscr{L} - L + 1$ words is sampled, resulting in a set of word counts $\{n_1, \ldots, n_{N^L}\}$. A PMF $\vec{\pi} \equiv (\pi_1, \ldots \pi_{N^L})$ is constructed from the word counts, with $\pi_w = n_w/(\mathscr{L} - L + 1)$; note $\vec{\pi}$ meets the boundedness and normalization conditions that define a PMF. The *block entropy* $H(L)$ for $\mathcal{X}$ can be computed by using (1). Computing $H(L)$ for a range of L-values yields the *entropy growth curve* [14].

2.2. Entropy Rates of Convergence

Crutchfield and Feldman [14] proposed a framework for assessing randomness versus order based on *entropy rates of convergence* of the entropy growth curve. In their schema, a process is *finitary* (*infinitary*) if a finite (infinite) amount of information is required to predict it; periodic, Markov, and hidden Markov models are examples of finitary processes. For a finitary process, the entropy growth curve $H(L)$ eventually enters a linear asymptotic regime, with $H(L) = E + h_\mu L$, where $h_\mu$is the *entropy rate* and E is the *excess entropy*. E measures the effective memory of the system. The entropy rate h_μ quantifies a system's irreducible randomness. The entropy rate $h_\mu \equiv lim_{L\to\infty} H(L)/L$, a limit that is guaranteed to exist for a stationary source [12]. Key fundamental quantities are $H(L)$ and its *discrete derivatives* and *discrete integrals*. For a function $F(L)$, its discrete derivative is defined by $\Delta F(L) \equiv F(L) - F(L-1), \quad L > 1$. Higher-order discrete derivatives can be defined by $\Delta^{(n)}F(L) = \Delta \circ \Delta^{(n-1)}F(L)$ (e.g., $\Delta^2 F(L) = F(L) - 2F(L-1) + F(L-2)$). Discrete definite integration of a function $F(L)$ between the limits $L = A$ and $L = B$ is accomplished by the sum $\sum_{L=A}^{B} F(L)$. Definitions and further details regarding the properties of discrete derivatives and integrals can be found in Section IIIA of [14]. Analysis of the entropy growth curve using discrete derivatives and integrals of $H(L)$ recovers previously known complexity, predictability, and memory metrics, and yields a new quantity called the *transient information*.

The first two discrete derivatives of $H(L)$ are the entropy gain $h_\mu(L)$ and predictabillity gain $\Delta^2 H(L)$, respectively:

$$h_\mu(L) = \Delta H(L); \tag{2}$$

and

$$\Delta^2 H(L) = h_\mu(L) - h_\mu(L-1), \quad L > 0. \tag{3}$$

By definition, $\Delta H(0) = \log_2 |\mathcal{A}|$, because we have no measurements to tell us anything to dispute the notion that the sequence has a uniform distribution of symbols from its alphabet [14]. $\Delta H(1) = H(1)$ because $H(0) = 0$ by definition; in short, no measurements mean no information. The entropy gain measures how much more information appears embodied in the system when viewed using words of length L rather than $L-1$. The entropy gain is a finite-L estimate $h_\mu(L)$ of h_μ. Another finite-L estimate $h_\mu{}'$ can be computed by ignoring the excess entropy; $h_\mu{}' = H(L)/L, L \geq 1$. The two estimates have differing convergence rates to h_μ; $h_\mu{}'(L) \geq h_\mu(L) \geq h_\mu$ [14]. For a finitary process, $h_\mu(L)$ decays toward its limit $h_\mu$faster than $1/L$; for some constant A, $h_\mu(L) - h_\mu < A/L$ [14]. The predictability gain measures how much more quickly the entropy rate is approaching its asymptotic value h_μ when words of length L rather than $L-1$ are considered; for a finitary process, $\Delta^2 H(L) \leq 0$, and equality is satisfied only when the asymptotic linear entropy growth regime is reached. For a finitary process, all higher-order discrete derivatives satisfy $\lim_{L\to\infty} \Delta^{(n)} H(L) = 0; \forall n \geq 2$.

Three integral quantities—the excess entropy E, the total predictability G, and the transient information T—parameterize the system's predictability and how the entropy growth curve approaches the linear regime (Table 1). The total predictability G quantifies the amount of nonrandom behavior in the system. The synchronization information measures the degree of difficulty encountered to *synchronize*—that is, to get into the linear asymptotic regime of the entropy growth curve—to a system.

A finite-L estimator of E is

$$E(L) = H(L) - L\Delta H(L). \tag{4}$$

The total predictability G satisfies the condition

$$\log_2 |\mathcal{A}| = |G| + h_\mu. \tag{5}$$

Table 1: Integrals of Entropy Rates of Convergence

Quantity	Definition
Generic form	$\mathcal{I}_n = \sum_{L=L_n}^{\infty} \left[\Delta^{(n)}(L) - \lim_{L\to\infty} \Delta^{(n)} H(L)\right]$
Excess Entropy	$E = \sum_{L=1}^{\infty} \left[\Delta H(L) - h_\mu\right]$
Total Predictability	$G = \sum_{L=1}^{\infty} \Delta^2 H(L)$
Transient Information	$T = \sum_{L=0}^{\infty} \left[E + h_\mu L - H(L)\right]$

Table 2: Sample E and h_μ Values for Binary Processes

Process Type	E	h_μ
Coin Flip:		
$p_{\mathrm{HEADS}} = 0.5$	0.0	1.0
$p_{\mathrm{HEADS}} = 0.7$	0.1187	0.8813
Period-P	$\log_2 P$	0
Order-R Markov	$H(R) - Rh_\mu$	constant > 0
Finitary: (exp. decay)	$\frac{H(1)-h_\mu}{1-2^{-\gamma}}$	$h_\mu(L) - h_\mu = A2^{-\gamma L}$
Infinitary: (log. growth)	$c_1 + c_2 \log L$	constant > 0

In the present work, we focus exclusively on h_μ, E, and $\Delta^2 H(L)$ to study the symbolic dynamics of rainfall data; other quantities were presented for the sake of completeness. In principle, we could compute finite-L estimates of T and G, but this is deferred for future work. The values (E, h_μ) are a strong classifier for processes (Table 2) [14]. In particular, we note that purely random independent identically distributed (IID) processes may be distinguished from each other, and that finite-order Markov processes can be identified.

3. Practical Considerations

We now address practical considerations in applying symbol-based analysis to meteorological timeseries data.

3.1. Symbolization

Creating symbolic sequences from continuous data is easily accomplished by using a *partition*. On the other hand, finding a partition that captures completely the continuous system's dynamics—a *generating partition*—is extremely difficult [15, 16]. Threshold crossing-based partitions are particularly prone to creating symbol sequences whose subsequent analysis can yield erroneous results [17]. Weather observations, however, offer alternatives to threshold-crossing based partitioning through *observation conventions*. Numerous examples exist. For example the binary wet/dry partition has been used to study precipitation occurrence complexity [4], and to model it as a Markov process [18]. Cloud amount is frequently reported or visualized in *octas* (9 symbols; 0=no cloud, 8=totally overcast), *tenths* (11 symbols; 0=no cloud, 10=totally overcast), and *percent* (101 symbols; 0=no cloud, 100=totally overcast). Atmospheric turbulence is classified by using six *Pasquill stability classes*, which are integrated measures derived from surface wind speed, incoming solar radiation, and nighttime cloud cover. Seasonal forecasting systems typically employ either *ternary* or *quintenary* partitions based on terciles or quintiles, respectively. Given the perils of a misplaced threshold partition, we have chosen to concentrate on partitions derived from observing and forecasting conventions; although these partitions may not be 'generating' with respect to the underlying system's smooth dynamics—assuming they exist—they do result in meteorologically relevant symbolic systems whose dynamics are of interest.

3.2. Finite Sequence Length

At first glance the meteorological observational record appears long, but in terms of timeseries analysis techniques most datasets are undesirably short; the longest weather station record is the Central England Temperature (CET) Record, which has monthly observations from the year 1659 to the present and daily observations from 1772 to the present [19]. Even for a binary alphabet the number of possible L-words scales as 2^L, meaning that the simplest possible word analysis of the CET will suffer undersampling for $L > 16$, and redundancy within the CET will cast suspicion on values of $L = 16$ and less; how much less depends on the entropy rate $h_\mu$of the sample. The longest hourly precipitation observation records span ≈100 years, but this number of observations (≈876 600) still places word analyses of $L \geq 20$ under suspicion. In all cases, one can envision a predictability "horizon" time $\tau_h = \log_{|\mathcal{A}|} \mathscr{L} \Delta t$,

where $\mathscr{L}$ is the number of observations and Δt is the spacing between them. Computation of block entropies $H(L)$ for increasing L will eventually encounter problems associated with the sequence length $\mathscr{L}$; eventually, $H(L)$ will saturate, with $h_\mu(L)$ approaching zero, and E will approach the constant value $\log \mathscr{L}$ [20]. Thus, finite sequence length can defeat entropy rate-based analyses because there may not be enough data to provide word statistics of sufficient quality at word lengths associated with the asymptotic linear regime.

It is desirable to estimate at what word length L^* one can no longer trust completely $H(L)$ and associated statistics. There are $\mathscr{L}/L$ independent (i.e., nonoverlapping) words of length L in a sequence of length $\mathscr{L}$, from which one could reasonably require $\mathscr{L} \geq L|\mathcal{A}|^L$. For an IID process, this constraint is reasonable. For a process with temporal data dependencies, the entropy rate—the average amount of information per symbol—will affect L^*; Lesne et al. [20] propose a condition for L-words to have good statistics that is derived from this assumption

$$\mathscr{L} h_\mu \geq L|\mathcal{A}|^L \ln |\mathcal{A}|, \tag{6}$$

with equality satisfied by the cutoff value L^*. In the analysis presented here, we will place the most trust in quantities associated with word lengths $L < L^*$.

3.3. Implementation

We have constructed a prototype word analysis system capable of handling (modulo computational resources) alphabets of arbitrary size. This system provides facilities for symbolization of continuously valued timeseries; object representations for symbol sequences and word logs (a bundle of PMFs or other word statistics for a range of word lengths for a sequence); and methods for computing word count quantiles, $C^{(1)}$and $C^{(2)}$complexities, and entropy rate of convergence statistics. It can also use estimates of h_μ to determine L^*. The present Fortran implementation has been tested extensively on a number of cases, including key examples from [14].

4. Case Study: Australian Daily Rainfall

We apply the techniques discussed in Sections 2 and 3 to Australian daily rainfall amounts from the Patched Point Dataset (PPD [21]) and Phase 3 data from the Australian Water Availability Project (AWAP [22]).

4.1. Datasets Used in this Study

The PPD data comprises all variables available from Australian Bureau of Meteorology stations; the term "point" identifies the location with the station record, and "patched" indicates that missing observations are either filled in with temporally interpolated station data or redistribution of weekend totals across the two or three days as appropriate. By definition the PPD contains only "perfect" records—as a result of patching—for the time the station was in operation. The PPD data cover the period 1889–present. The AWAP data were created by using a topographically aware technique that combines empirical interpolation with function fitting to map in situ and satellite observations of rainfall, surface air temperature, and vapor pressure [22]. The AWAP dataset is available online, with fields available for the period 1900–present on a $0.05° \times 0.05°$latitude-longitude grid ($\approx$5 km spacing). These fields were used to compute timeseries of rainfall at locations corresponding to the PPD stations. In broad terms, one would expect the PPD data to reflect more accurately the time progression between days with and without rainfall.

Table 3 contains a summary of the 14 PPD station locations used in this study, along with results of the analysis discussed in Section 4.2. These stations were chosen because they had long observational records with relatively few missing observations; the intent is to minimize the influence of patching on our results, and thus they are the closest to uninterrupted daily rainfall station records available for Australia. Note that there are no "wet tropical" stations in the record as there are insufficient data for long uninterrupted records.

PPD and AWAP timeseries data from three representative locations—Bourke, Brisbane, and Hobart—are shown in Figures 1(a)–1(c) and 1(d)–1(f), respectively.

In our analysis, PPD and AWAP data were coarse-grained into a binary alphabet by assigning a '1' ('0') to observed nonzero (zero) rainfall. This partition is unlikely to be a generating partition, but it has precedent [4] and is certainly a meteorologically relevant observation convention because the binary dynamics of the succession of wet and dry days of wide interest to meteorologists and water resource researchers.

Table 3: Description of and Results for PPD Stations

				PPD Results			AWAP Results		
Station	ID No	Latitude	Longitude	L^*	h_μ	E	L^*	h_μ	E
Bourke	48013	-30.09°	145.94°	12	0.5082	0.2000	12	0.6166	0.3150
Broken Hill	47037	-31.89°	141.99°	12	0.3991	0.3312	12	0.5721	0.3014
Hobart	94029	-42.89°	147.33°	13	0.9022	0.4473	12	0.8545	0.2850
Melbourne	86071	-37.87°	144.97°	13	0.8781	0.4543	12	0.8456	0.3288
Perth	9034	-31.96°	115.87°	12	0.7247	0.4558	12	0.7274	0.5127
Sydney (Obs. Hill)	66062	-33.86°	151.21°	13	0.8489	0.4653	12	0.8234	0.3228
Charters Towers	34002	-20.08°	146.26°	12	0.5318	0.3402	12	0.5950	0.3943
Adelaide	23000	-34.93°	138.59°	13	0.7709	0.5312	12	0.7725	0.4149
Brisbane	40214	-27.48°	153.03°	13	0.7807	0.4846	12	0.7831	0.3916
Toowoomba	41103	-27.58°	151.93°	12	0.7641	0.2770	12	0.8106	0.3622
Dubbo	65012	-32.24°	148.61°	12	0.6748	0.2397	12	0.7686	0.3011
Coonabarabran	64008	-31.27°	149.27°	12	0.6832	0.2339	12	0.7961	0.3196
Newcastle	61055	-32.92°	151.80°	13	0.8423	0.4499	12	0.8626	0.2955
Wilcannia	46043	-31.56°	143.37°	12	0.4695	0.1742	12	0.5666	0.2666

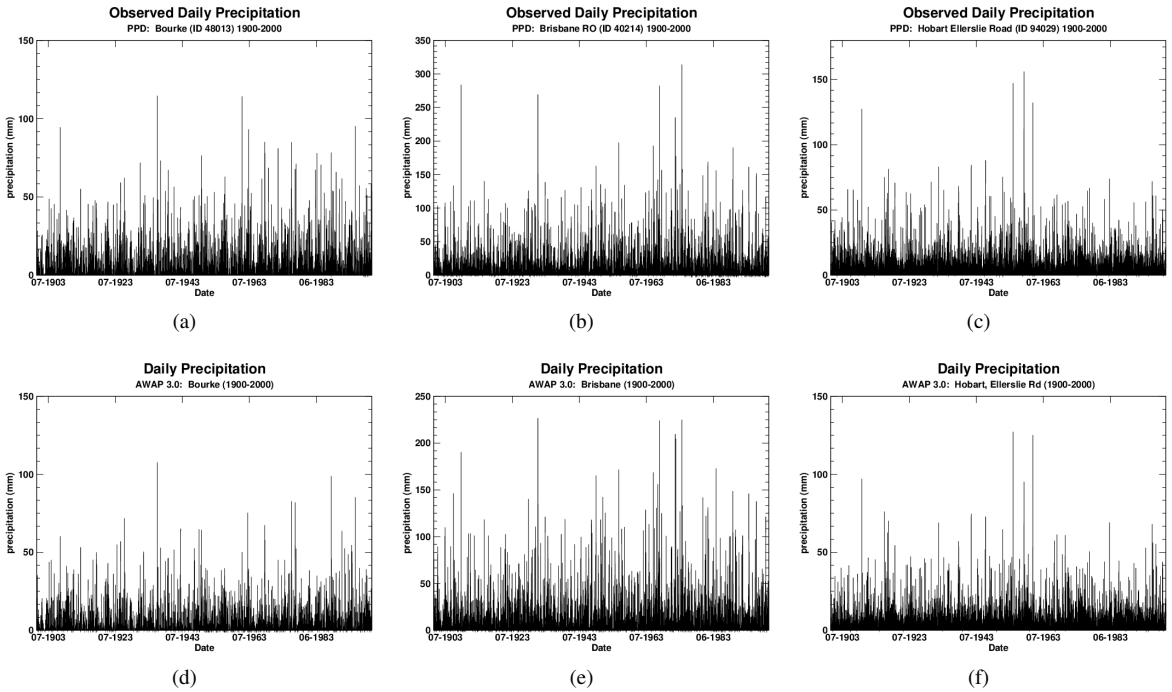


Figure 1: Sample daily precipitation timeseries 1900-2000: PPD station data from (a) Bourke, (b) Brisbane, and (c) Hobart; AWAP 3.0 data from (d) Bourke, (e) Brisbane, and (f) Hobart.

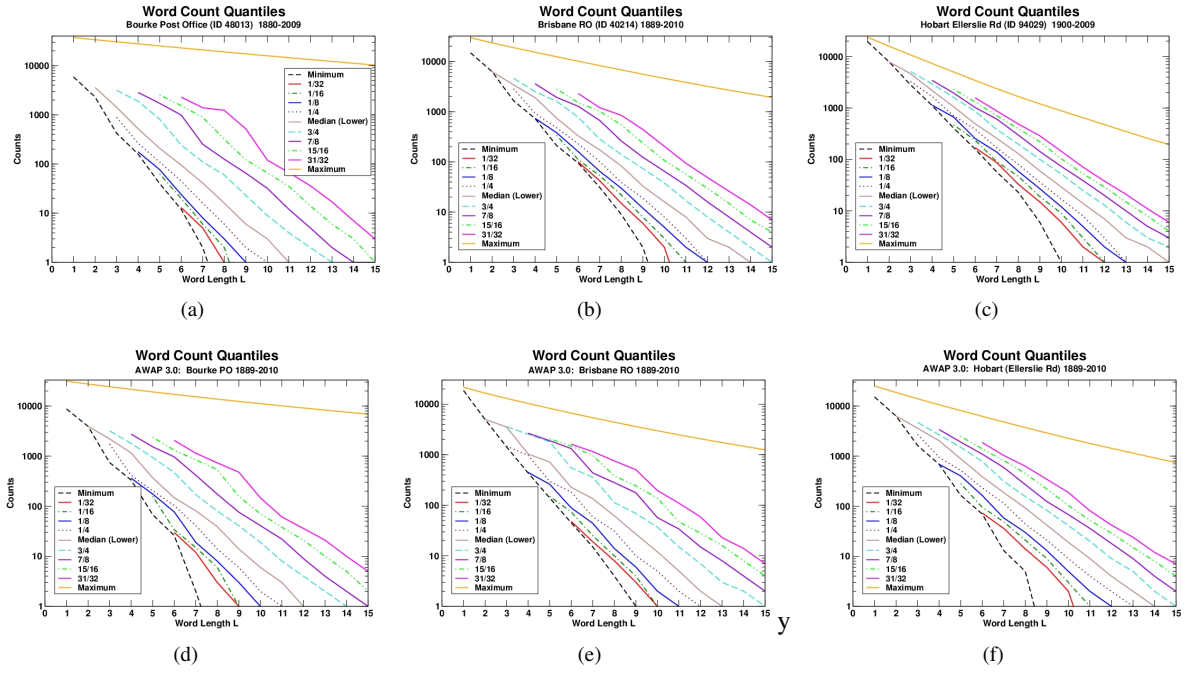


Figure 2: Quantiles for word counts for L-words: PPD station data from (a) Bourke, (b) Brisbane, and (c) Hobart; AWAP 3.0 data from (d) Bourke, (e) Brisbane, and (f) Hobart.

For the period 1889–1999, there are $\mathscr{L} = 40541$ daily observations. For a binary alphabet, words with $L \geq 16$ should not be considered because there are not sufficient data to ensure that every L-word might potentially be observed. Figure 2 displays quantiles of word counts of varying length for PPD and AWAP data. Each curve corresponds a fraction f of the 2^L possible L-words; for example, the $(\frac{31}{32})$ curve corresponds to the $\{8, 16, 32\}$th highest counts encountered words for $L = \{8, 9, 10\}$. Word counts decline with increasing L, with half of the words seen only once or not at all for in the PPD (AWAP) L-values of $\{11, 14, 15\}$ ($\{12, 13, 14\}$) for Bourke, Brisbane, and Hobart, respectively. The word quantile curves present a dilemma: Is a given word unobserved because the system's governing dynamics exclude it, or is its absence a consequence of sample size? We present word statistics up to $L = 15$, but will treat statistics for words with $L \geq L^*$ as defined by (6) with skepticism.

4.2. Symbolic Dynamics of Australian Daily Precipitation

Table 3 includes values of L^* computed by using its definition (6) and substituting the instantateous entropy rate $h_\mu(L)$ for h_μ; the reader should keep these "cutoff" values in mind when viewing Figures 3 and 4, placing confidence only in statistics with $L \leq L^* - 1$. We have included the higher L values in these plots to illustrate the key pitfall associated with block entropy analysis: misplaced faith in word statistics lacking underlying sample support. In the PPD data, all of the coastal stations except Perth have $L^* = 13$, while Perth and the inland stations have $L^* = 12$; this is due to the differences in entropy rates between these groups of stations, with the inland stations having less inherently random (and thus more temporally dependent) occurrence values. The AWAP data all have $L^* = 12$; for the coastal stations, this disparity with the PPD may largely be due to the shorter—by eleven years—sample size.

Figure 3 shows the entropy growth curves $H(L)$, entropy rates $h_\mu(L)$, and finite-L excess entropies for PPD and AWAP timeseries data. PPD station data show stratification associated with inland/coastal location and distance to the coast, with clustering of outback stations showing the lowest block entropies, inland stations occuping a broad midrange, and coastal stations having the highest values (Figure 3(a)); this stratification is duplicated in the instantaneous entropy rates $h_\mu(L)$ (Figure 3(b)). There is slight overlap between the coastal and inland station clusters, with Perth having slightly lower block entropy and entropy rate than does Toowoomba. Excess entropy estimates for the PPD data (Figure 3(c)) show clustering at low values for the outback stations, but the inland and coastal clusters are disrupted. Hobart shows the lowest non-outback values of $E(L)$ for $L \leq 11$, but the values rise rapidly between $L = 11$

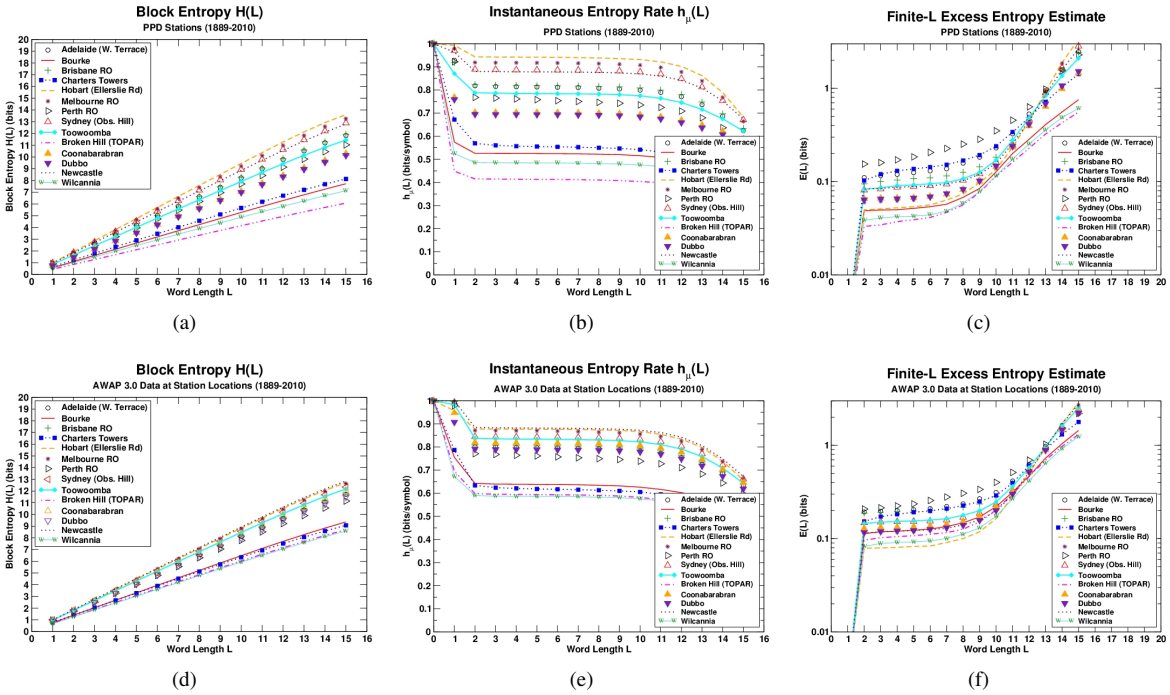


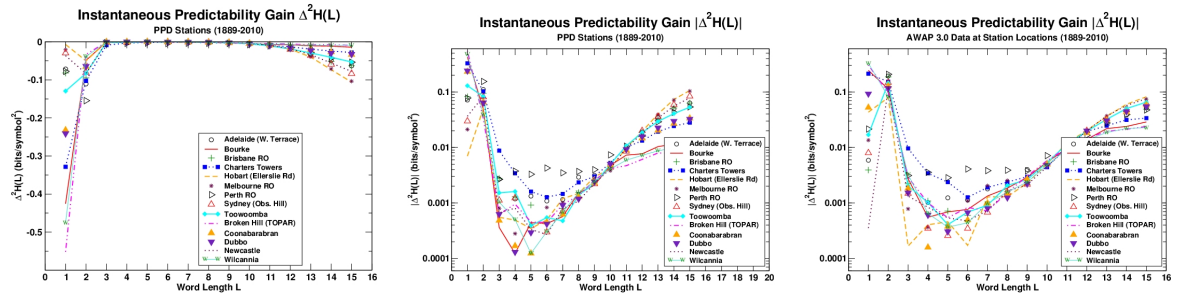
Figure 3: Entropy growth curves for L-words: (a) block entropy $H(L)$, (b) entropy rate $h_\mu(L)$, and (c) excess entropy estimate $E(L)$ for PPD data; and (d) block entropy $H(L)$, (e) entropy rate $h_\mu(L)$, and (f) excess entropy estimate $E(L)$ for AWAP 3.0 data.

and $L = 12$; the reliability of higher values is questionable because of finite sequence length effects. For much of the range of L, Perth has the highest excess entropy, and thus precipitation occurrence at this location has the most "memory." Block entropies computed from AWAP data (Figure 3(d)) show a "low" cluster comprising the outback stations and Charters Towers and a "high" cluster of the remaining stations. Entropy growth curves for outback (non-outback) locations are generally shifted upward (downward) and their ranges narrowed with respect to their PPD counterparts. $H(L)$ curves for Perth and Charters Towers show the best agreement between the PPD and AWAP data. The shifts in the AWAP block entropy curves is caused largely by a narrowing in the range of $h_\mu(L)$ (Figure 3(e)); instantaneous entropy rates for Charters Towers and Perth track their PPD counterparts well. AWAP excess entropies (Figure 3(f)) are shifted downward and broadened with respect to the PPD $E(L)$ curve.

The $h_\mu(L)$ curves are nearly straight lines (Figures 3(b) and 3(e)), which imply low values of the second discrete derivative $\Delta^2 H(L)$ (e.g., Figure 4(a) for the PPD). Semilog plots of the predictability gain $|\Delta^2 H(L)|$ for the PPD and AWAP data are shown in Figures 4(b) and 4(c), respectively. For both datasets, the predictability gain shows a wide range of values for $L < 4$, which narrows dramatically for $4 < L \leq 8$, and is narrowest at $L = 10$. For $L > 10$, stations are clustered as with $H(L)$. The PPD outback stations have the lowest values, the inland stations cluster in the midrange, and the coastal stations cluster at the high end, with the inland and coastal clusters overlapping. The AWAP data show a similar clustering, though the range of values is much narrower, with the outback cluster shifted upward. In both datasets, Perth shows the highest predictability gain for much of the L-range; and the "bump" at $L = 6$ implies analysis using blocks of this length, as opposed to $L = 5$, is significantly more informative than going from $L = 6$ to $L = 7$.

Table 3 contains lower- and upper-bound estimates of E h_μ, respectively. Broadly put, both randomness and memory are substantially overestimated in the AWAP data for the outback stations. For the inland non-outback stations the AWAP data produce overestimates in $h_\mu$and E with respect to the PPD. For coastal locations, generally AWAP data is associated with mild underestimates in $h_\mu$and substantial underestimates in E; the underestimates in memory may be a consequence of interpolation error.

In sum, the entropy rates of convergence indicate a long, fairly uniform random regime low L. The PPD station data fall into clusters determined by these statistics' values, the AWAP data less so. Shortly before the word length



(a) (b) (c)

Figure 4: Predictability gain versus word size L: (a) $\Delta^2 H(L)$ for PPD stations; and semilog plots of $|\Delta^2 H(L)|$ for (b) PPD stations, and (c) AWAP 3.0 data.

cutoff values L^* are reached, there is some weakening of the system's inherent randomness, which is exchanged for a corresponding growth in memory. The slow—but continuing—changes in $h_\mu$with respect to L rule out Rth order Markov processes, where $R = L$. It is hard to disentangle early finite-sample-size effects from a genuine aquisition of skill without a more substantial uncertainty analysis of $H(L)$.

5. Conclusions and Future Work

We have developed a symbolic dynamics analysis system for studying word complexity and entropy rates of convergence for coarse-grained timeseries data and have applied it to observed and analyzed, long-term Australian daily rainfall records. We have found that the randomness observed that is used to justify Markov-chain modeling of precipitation occurrence may not be in the assumed asymptotic linear block entropy regime one would want present to justify this approach. Our results are, however preliminary; the progression is away from what appears to be a linear "random" block entropy regime commences immediately before finite-sample-size effects emerge. More work is required to attach rigorous uncertainty estimates to our estimates of the entropy rate h_μ and the excess entropy E. In our work here, we refrained from trying to hunt for finite-order Markov models by fitting the data to the formula for E presented in Table 2 because we want greater confidence in the entropy growth curve itself first.

The fact that the technique clusters station records by their climatic properties is compelling. We believe block entropy analysis may provide a useful organizing priciple for clustering of climate data and are eager to apply this technique to a wider variety of station records, reanalysis datasets, and climate model integration outputs. We look forward to seeing plots of large numbers of records in an E–$h_\mu$plane to see if our suspicions are well-founded.

A comprehensive uncertainty quantification framework will comprise symbol-based tests for stationarity (for example, use of the Kullback-Leibler divergence to quantify nonstationarity as information gain when the PMF for a sample from one time period is used to model the PMF of a different time window); Monte Carlo tests for robustness to "bit-flipping," which will quantify likely effects of misidentified precipitation occurrences known to exist in Australian station records [23]; and integrated correlation time-based estimates for likely errors in emperical PMFs and their resulting block entropies [20]. Once this framework is in place, we can return to the question of dynamics classification and deploy this system on much larger datasets.

Acknowledgments

This work was supported by the U.S. Department of Energy, under Contract DE-AC02-06CH11357.

References

[1] N. A. Bagrov, Statistical entropy as an indicator of similarity or difference of meteorological fields, Meteorologiya i Gidrologiya (1).
[2] J. Shukla, T. DelSole, M. Fennessy, J. Kinter, D. Paolino, Climate model fidelity and projections of climate change, Geophysical Research Letters 33 (2006) L07702.

[3] T. DelSole, M. Tippett, Predictability: Recent insights from information theory, Reviews of Geophysics 45 (2007) RG4002.
[4] J. B. Elsner, A. A. Tsonis, Complexity and predictability of hourly precipitation, J. Atmos. Sci. 50 (3) (1993) 400–405.
[5] J. W. Larson, Can we define climate using information theory?, IOP Conference Series: Earth and Environmental Science 11 (1) (2010) 012028.
[6] M. Morse, G. A. Hedlund, Symbolic dynamics, American Journal of Mathematics 60 (4) (1938) 815–866.
[7] B.-L. Hao, Elementary Symbolic Dynamics and Chaos in Dissipative Systems, World Scientific, Singapore, 1989.
[8] B. Marcus, S. Williams, Symbolic dynamics, Scholarpedia 3 (11) (2008) 2923.
[9] C. S. Daw, C. E. A. Finney, E. R. Tracy, A review of symbolic analysis of experimental data, Review of Scientific Instruments 74 (2) (2003) 915–930.
[10] C. Nicolis, W. Ebeling, C. Baraldi, Markov processes, dynamic entropies and the statistical prediction of mesoscale weather regimes, Tellus A 49 (1) (1997) 108–118. doi:10.1034/j.1600-0870.1997.00008.x.
[11] C. E. Shannon, A mathematical theory of communication, Bell System Technical Journal 27 (1948) 379–423.
[12] T. M. Cover, J. A. Thomas, Elements of Information Theory, 2nd Edition, Wiley-Interscience, New York, 2006.
[13] F. M. Reza, An Introduction to Information Theory, Dover, New York, 1994.
[14] J. P. Crutchfield, D. P. Feldman, Regularities unseen, randomness observed: Levels of entropy convergence, Chaos: An Interdisciplinary Journal of Nonlinear Science 13 (1) (2003) 25–54. doi:10.1063/1.1530990.
[15] M. B. Kennel, M. Buhl, Estimating good discrete partitions from observed data: Symbolic false nearest neighbors, Phys. Rev. Lett. 91 (8) (2003) 084102. doi:10.1103/PhysRevLett.91.084102.
[16] Y. Hirata, K. Judd, D. Kilminster, Estimating a generating partition from observed time series: Symbolic shadowing, Phys. Rev. E 70 (1) (2004) 016215. doi:10.1103/PhysRevE.70.016215.
[17] E. M. Bollt, T. Stanford, Y.-C. Lai, K. Zyczkowski, What dynamics do we get with a misplaced partition? on the validity of threshold crossings analysis of chaotic time-series, Physica D 154 (2001) 259–286.
[18] M. F. Hutchinson, A point rainfall model based on a three-state continuous Markov occurrence process, Journal of Hydrology 114 (1–2) (1990) 125–148. doi:DOI: 10.1016/0022-1694(90)90078-C.
[19] D. E. Parker, T. P. Legg, C. K. Folland, A new daily central England temperature series 1772–1991, International Journal of Climatology 12 (1992) 317–342.
[20] A. Lesne, J.-L. Blanc, L. Pezard, Entropy estimation of very short symbolic sequences, Physical Review E 79 (4) (2009) 046208.
[21] S. J. Jeffrey, J. O. Carter, K. B. Moodie, A. R. Breswick, Using spatial interpolation to construct a comprehensive archive of australian climate data, Environmental Modelling and Software 16 (4) (2001) 309–330.
[22] D. A. Jones, W. Wang, R. Fawcett, High-quality spatial climate data-sets for australia, Australian Meteorological and Oceanographic Journal 58 (4) (2009) 233–248.
[23] N. R. Viney, B. C. Bates, It never rains on Sunday: The prevalence and relavence of untagged multi-day rainfall accumulations in the Australian high quality data set, International Journal of Climatology 24 (2004) 1172–1192.

Government License