

Can We Define Climate Using Information Theory?

J. Walter Larson

Mathematics and Computer Science Division, Argonne National Laboratory, 9700 S. Cass Avenue, Argonne, IL 60439, USA

Computation Institute, University of Chicago, 5735 South Ellis Avenue, Chicago, IL 60637, USA

School of Computer Science, The Australian National University, Canberra, ACT 0200, Australia

E-mail: larson@mcs.anl.gov

Abstract. The standard definition of climate is, by convention, based on a thirty-year sample. But why? One way to define the sampling period for constructing climatologies is to ask: What is a sufficient sample to construct probability density functions (PDF) for key meteorological variables? One method for judging the sufficiency of a sample to construct a PDF is to use information theory. I propose a framework for evaluating climatic sampling periods based on level of detail and associated uncertainties in the estimated PDF, the Shannon entropy growth curve and its discrete derivative, and Kullback-Leibler divergence-based statistics for quantifying the information gain as the sampling period is expanded by a specified amount. I apply this approach to daily data from the Central England Temperature (CET) record spanning the period 1772–2006. PDF estimation is performed by using an optimal binning technique derived from Bayesian principles to determine a uniform binning strategy that maximizes the posterior probability given the data sample; this technique identifies the known heavy truncation of the CET data and yields insight into the PDF structure with estimated uncertainties for a sampling period spanning 1–235 years. Ensemble-generated statistics from windowed resampling and Monte Carlo calculations of neighboring estimated PDFs are computed, resulting in confidence intervals for all the structural quantities in the framework. I use these statistics to compare the relative confidence associated with a number of popular sampling periods.

1. Introduction

It is standard practice to compute climatic quantities by using a thirty-year sample of meteorological data—a convention defined by the World Meteorological Organization. The choice of thirty years as the sampling period was decided in 1935 at the International Meteorological Organization’s conference in Warsaw [1, 2]. Sampling periods considered ranged from eleven years (the sunspot cycle) to fifty years (to better capture interdecadal variability), but no record exists about the advocates for a particular period or the supporting arguments they offered [2]. The UK Met Office (UKMO) claims on its web site that climate is defined using a thirty-year sample to eliminate year-to-year variations in the moments [3].

One way to decide how long a record is required to describe the climate is to construct probability density functions (PDFs) for multiple meteorological fields over an increasingly long period of time and compare them with PDFs for longer periods or the complete data record. The question then becomes: How long a sample period is required to derive PDFs

that are variable by variable sufficiently complete to describe the whole record? Information theory provides quantitative measures for the content of a PDF, and assessing the skill with which a candidate PDF $Q(x)$ models the “ideal” PDF $P(x)$. The *Shannon entropy* (SE) $H(x)$ measures the information content of $P(x)$, and the Kullback-Leibler divergence (KLD) quantifies the disparity between $Q(x)$ and $P(x)$.

Using the SE, the KLD, and a PDF estimation technique derived from Bayesian principles, I propose a method for assessing climate definitions by answering the following questions:

- Q1** How much detail is present in, and how confident can we be of an estimate of, the PDF for a variable for a given sample of W years?
- Q2** What is the information content H of a given sample of W years?
- Q3** How well does a sample of W years reflect the whole available data record of $Y > W$ years?
- Q4** What is the incremental change in H with respect to W ?
- Q5** What is the informativeness of expanding a sample of W years to $W + 1$ years?

I address **Q1** using the uncertainty measures associated with the PDF estimation technique described in Section 4, **Q2** and **Q4** using the SE and its discrete derivative $\Delta H(W)$, and **Q3** and **Q5** using the KLD. I apply this technique to the longest existing daily meteorological observation dataset, the Central England Temperature record (CET).

2. Information Theory

Information theory [4] is a means for quantifying overall structure in probability mass functions (PMFs) and PDFs for discrete and continuous variables, respectively. Its antecedents lie in Boltzmann’s statistical mechanical definition of thermodynamic entropy. Reference [4] gives a detailed discussion of information theory. Previous applications of information theory to climate include predictability ([5] and references therein), model evaluation [6, 7], and variability [7].

For a discrete variable $X \in \{x_1, \dots, x_N\}$, the (PMF) $\vec{\pi}$ is defined such that π_i is the probability that $X = x_i$; the probabilities π_i satisfy the conditions $0 \leq \pi_i \leq 1, \forall i \in \{1, \dots, N\}$, and $\sum_{i=1}^N \pi_i = 1$. The SE $H(X)$ is

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

The base of the logarithm in (1) determine the units of $H(X)$; for bases 2 and e , $H(X)$ is measured in *bits* and *nats*, respectively. The normalization properties of the probabilities π_i imply $H(X) \geq 0$. The SE quantifies the “surprise” in X . The maximum value of $H(X)$ occurs for a uniform distribution $\pi_1 = \dots = \pi_N = 1/N$; for a system with N states, $0 \leq H(X) \leq \log N$.

Consider two distinct PMFs for X , $\vec{\zeta}$ and $\vec{\pi}$. The additional information required to describe $\vec{\pi}$ given $\vec{\zeta}$ is the KLD

$$D_{KL}(\vec{\pi} \parallel \vec{\zeta}) = \sum_{i=1}^N \pi_i \log \left[\frac{\pi_i}{\zeta_i} \right]. \quad (2)$$

Information theory is also applicable to a continuous variable x ; one replaces the PMFs $\vec{\pi}$ and $\vec{\zeta}$ with PDFs $p(x)$ and $q(x)$, respectively; and the summations over discrete states in (1) and (2) become integrals with respect to the continuous variable x [4]. The main challenge is that an integral formulation of the SE can be negative or infinite; the reason is that the probability density function $p(x)$ may locally exceed unity. Furthermore, numerical computation of the SE and KLD is necessarily sensitive to the discretization $dx \rightarrow \Delta x$; this is more the case with the SE than the KLD. Thus, one must take care in discretizing or binning continuous data to form a PMF or PDF, respectively. This issue is discussed in detail in Section 4.

In this study I propose a framework for evaluating climate sampling periods based on the SE and the KLD. Consider a given climate variable (e.g., a station record) for which a long-term sample of Y years is available. Suppose we draw from this large sample a contiguous (in time) window spanning a period of $W < Y$ years. Computing the SE for this sample, $H(W)$, where the argument to H denotes the sample size, provides the answer to **Q2**. One can answer **Q3** by computing the KLD $D_{KL}(P \parallel Q)$ for the PDF's $Q(W)$ and $P(Y)$ estimated from the windows spanning W and Y years, respectively. The marginal increase in SE from **Q4** is determined by computing the *discrete derivative* [8] $\Delta H(W) \equiv H(W) - H(W - 1), W \geq 1$. Note that $\Delta H(1) = H(1)$, since we go from knowing nothing about the system to knowing what a one year sample will tell us. The answer to **Q5** is the KLD $D_{KL}(P(W + \Delta W) \parallel Q(W))$.

3. The Central England Temperature Record

The (CET) the longest observational record for surface air temperature. The original CET compiled by Manley are monthly averages beginning in 1659. The daily CET span the period 1772–present. In this study, I used data for the period 1772-2006, obtained from the UKMO Hadley Centre [9] operated by the British Atmospheric Data Centre. Parker et al. [10] describe the data selection and processing methods used to create the CET daily record. Archived temperatures in the record are rounded to the nearest 0.1°C.

4. Numerical Method for PDF and Entropy Estimation

The central problem for computing the SE (1) for a continuously valued quantity is choosing a robust discretization scheme that allows PDF and subsequent entropy estimation. The main perils one faces in PDF estimation are choosing too few (*oversmoothing*) or too many (*overfitting*) bins. Because the SE is sensitive to the number of bins used to compute it from a sample of a continuous variable, care is required in this choice. Many methods exist for binning data to estimate PDFs; Scott [11] provides an excellent overview of the problem.

For this study, I have chosen Knuth's [12] optimal binning scheme, which is derived from a Bayesian approach. The Bayesian priors are the sample data \vec{d} , and the assumption of a piecewise-constant PDF with uniform bins constitute I . The number of bins M maximizes a marginal posterior probability function $p(M|\vec{d}, I)$:

$$p(M|\vec{d}, I) \propto \left(\frac{M}{V}\right)^N \frac{\Gamma(\frac{M}{2})}{\Gamma(\frac{1}{2})^M} \frac{\prod_{k=1}^M \Gamma(n_k + \frac{1}{2})}{\Gamma(N + \frac{M}{2})}, \quad (3)$$

where $\Gamma(\cdot)$ is the Gamma function, N is the sample size, V is the sample range, and n_k is the number of counts in each bin. The value of M that maximises $p(M|\vec{d}, I)$ yields the most probable piecewise constant, uniform-bin-width PDF based on the sample \vec{d} . It is easier computationally to maximize the logarithm of RHS of (3). The mean bin probability π_i and its variance σ_i^2 are

$$\pi_i = \left(n_i + \frac{1}{2}\right) / \left(N + \frac{M}{2}\right) \quad (4)$$

$$\sigma_i^2 = \left(n_i + \frac{1}{2}\right) \left(N - n_i + \frac{M-1}{2}\right) \left(N + \frac{M}{2} + 1\right)^{-1} \left(N + \frac{M}{2}\right)^{-2}. \quad (5)$$

Note that even when data is absent from a bin, $\pi_i \neq 0$, and the π_i are normalized by construction. The average signal-to-noise ratio SNR for the estimated PMF is $\sum_{i=1}^M \pi_i / \sigma_i$. The optimal value of M from (3), the resulting π_i in (4), and the associated SNR answer **Q1**.

The aforementioned rounding of the CET data presents a problem for the optimal binning scheme described; the truncation of the data constitutes a fine-scale feature in addition to the

broad-scale shape of the PDF. This fine-scale feature causes the log posterior function to have numerous local maxima and to exhibit growth at large numbers of bins (cf. Figure 5D in [12]). In fact, the optimal binning scheme is an excellent detector of severely truncated data [13]. The truncation effects can be removed through data smoothing by adding a random uniformly distributed deviate that brackets the rounded value by half the truncation value; this will not replace lost information [13] but will allow us to estimate the larger-scale structure of the PDF. In this study I have smoothed the CET data by adding to each value T_i a uniform random deviate $\delta_i \in [-0.05, 0.05)$; this smoothing allows the optimal binning scheme to better identify the large-scale structure of the sample’s PDF (cf. Figure 1B in [12]) and, when truncated to the nearest 0.1°C , yields the original CET timeseries.¹

The SE and KLD values in this study are calculated by first estimating the uniformly binned PDF through maximization of the logarithm of (3), yielding the number of uniform bins M and the bin probabilities (4); these estimates are used to compute each PDF’s SNR . Uncertainty estimates for the SE and KLD are computed through windowed resampling of the data and Monte Carlo ensemble statistics using a large number (10,000) of “neighboring” PDFs determined from (4) and (5). Random number generation to determine the neighboring PDFs is accomplished by using standard techniques. Data sampling is implemented by using a sliding window technique in which a window of W years is moved one year at a time across the total dataset of Y years. This process results in $Y - W + 1$ samples for a window spanning W years; for the CET’s $Y = 235$ years, $W = 1$ and $W = 10$ years result in 235 and 226 distinct samples (or 2.35×10^6 and 2.26×10^6 PDFs), respectively. Ensemble statistics are computed for the SE and KLD quantities, including the mean, variance, minima/maxima, and key percentiles.

5. Results

Figure 1 shows the results of the optimal binning scheme applied to the CET 1772–2006. The number of bins M grows steadily with increasing window size W but shows some variation for fixed W (Figure 1(a)); the central curve is the mean $\langle M(W) \rangle$, the box indicates one standard deviation on either side of $\langle M(W) \rangle$, and the whiskers indicate maximum and minimum values. The range of values $M(W)$ narrows with larger samples as the range between extreme temperature values become more consistent across each windowed sample; for $W \geq 176$ years, all samples contain the minimum and maximum values in the full CET, resulting in the same dynamic range and yielding $M = 41$ bins. The SNR grows steadily with W (Figure 1(b)), with a narrowing of the range of values of SNR . The growth in M and SNR w.r.t. W indicates increasing detail in the PDF, combined with increasing confidence in the PDF’s individual bin probabilities (Figures 1(c)–1(f)).

Figure 2 shows percentiles of H for the CET. The median value of $H(W)$ climbs rapidly for $1 \leq W \leq 10$ years, continues to climb steadily for $10 < W \leq 60$ years, climbs less rapidly for $60 < W \leq 100$ years, and converges to a near-constant value $H \approx 4.55$ bits beyond the century timescale (Figure 2(a)). The range of values of $\Delta H(W)$ narrows rapidly for $1 \leq W \leq 10$ years, and then drops off steadily with increasing sample size W (Figure 2(b)). The interdecile ranges of ΔH at $W = 30$ and $W = 50$ years correspond roughly to the interquartile ranges at $W = 10$ and $W = 30$ years, respectively.

Figure 3 shows percentiles for the total and incremental information gains $D_{KL}(P(Y) \parallel Q(W))$, and $D_{KL}(P(W) \parallel Q(W - 1))$, respectively. The representativeness of a subsample of W years increases rapidly for $1 \leq W \leq 10$ years, with the median of $D_{KL}(P(Y) \parallel Q(W))$ dropping by a factor of 10 over this range, and continuing to decrease steadily out to $W = 100$ years (Figure 3(a)). The incremental information gain $D_{KL}(P(W) \parallel Q(W - 1))$ also drops in a

¹ This smoothing has been tried for multiple random smoothings of the CET, and the resulting quantities computed from these datasets are consistent with the results presented here.

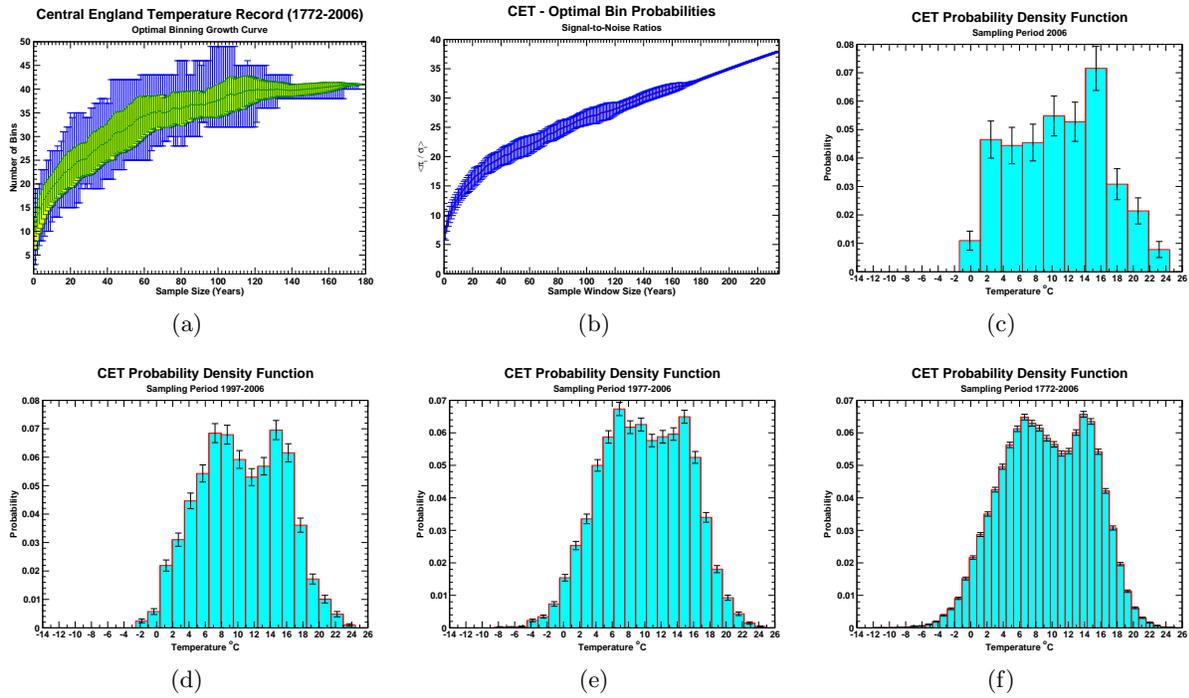


Figure 1. Optimal binning of the CET data: a) growth curve for number of bins, b) SNR , and estimated PDFs for windowed samples covering periods c) the year 2006, d) 1997-2006, e) 1977-2006, and f) 1772-2006.

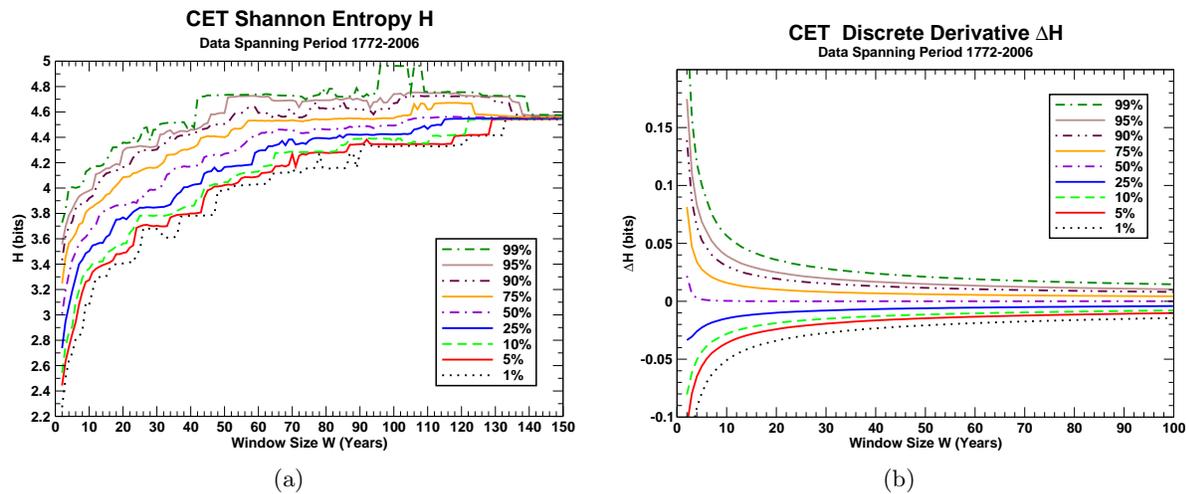


Figure 2. Shannon Entropy estimation for CET: a) SE as a function of sampling window size, and b) Marginal Entropy increase ΔH .

similar fashion as adding one year to a sample becomes less significant (Figure 3(b)). The 90th percentile at $W = 50$ and $W = 30$ years correspond roughly to the 30th and 10th percentiles at $W = 30$ and $W = 10$ years, respectively.

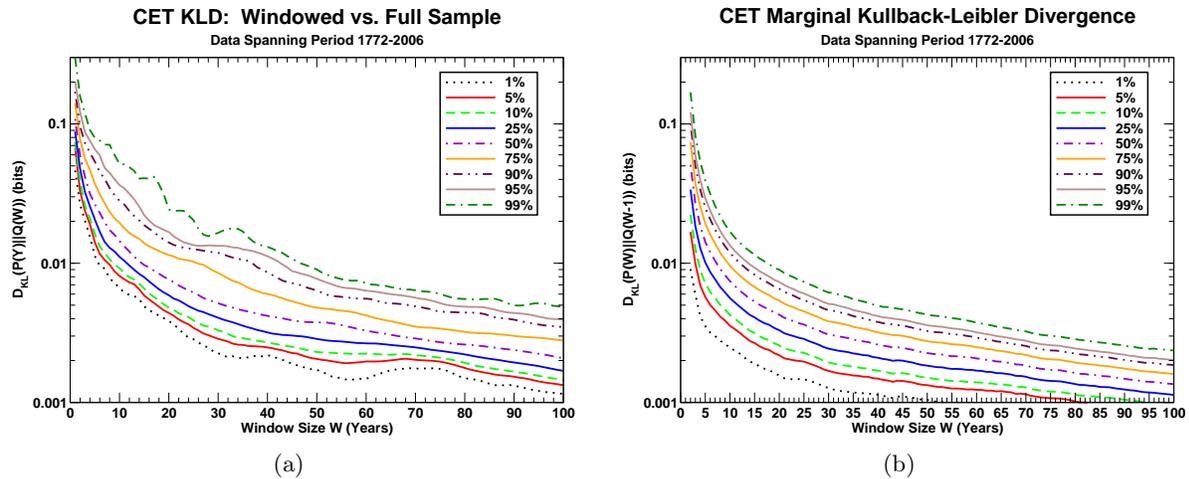


Figure 3. Kullback-Leibler Divergences for the CET: a) KLD for sample of W years versus the whole dataset; b) Marginal KLD $D_{KL}(P(Y) \parallel Q(W))$.

6. Conclusions

A information-theoretic framework for evaluating the definition of climate based on sampling periods has been presented, and applied to the CET. The framework allows direct comparison between sampling periods, allowing one to decide the relative risk or gain in selecting shorter or longer sampling periods, respectively. The results presented here are preliminary. The overall utility of this framework must be evaluated through application to a wider variety of observational station records and to reanalyses, and this will be an area of future investigation.

ACKNOWLEDGMENTS

This work was supported by the Office of Advanced Scientific Computing Research, Office of Science, U.S. Department of Energy (DOE), under Contract DE-AC02-06CH11357.

References

- [1] International Meteorological Organization 1937 *Proceedings of the Meetings in Danzig and Warsaw 29–31 August and 12 September 1935* (Leyden: Secretariat of the IMO)
- [2] Hulme M, Dessai S, Lorenzoni I and Nelson D R 2009 *Geoforum* **40** 197–206
- [3] UK Meteorological Office 2010 Climate Averages <http://www.metoffice.gov.uk/climate/uk/averages/> [Accessed 27 February 2010]
- [4] Cover T M and Thomas J A 2006 *Elements of Information Theory* 2nd ed (New York: Wiley-Interscience)
- [5] DelSole T and Tippett M 2006 *Reviews of Geophysics* **45** RG4002
- [6] Shukla J, DelSole T, Fennessy M, Kinter J and Paolino D 2006 *Geophysical Research Letters* **33** L07702
- [7] Larson J W 2009 *Proceedings of the 18th World IMACS Congress and MODSIM09 International Congress on Modelling and Simulation* ed Anderssen R S, Braddock R D and Newham L T H (Modelling and Simulation Society of Australia and New Zealand and International Association for Mathematics and Computers in Simulation) pp 2639–2646
- [8] Crutchfield J P and Feldman D P 2003 *Chaos: An Interdisciplinary Journal of Nonlinear Science* **13** 25–54
- [9] British Atmospheric Data Centre, Hadley Centre, UK Meteorological Office 2009 Historical Central England Temperature (CET) Data <http://badc.nerc.ac.uk/data/cet/>
- [10] Parker D E, Legg T P and Folland C K 1992 *International Journal of Climatology* **12** 317–342
- [11] Scott D W 1992 *Multivariate Density Estimation: theory, practice, and visualization* (New York: Wiley)
- [12] Knuth K H 2006 Optimal data-based binning for histograms <http://arxiv.org/abs/physics/0605197>
- [13] Knuth K H, Castle J P and Wheeler K R 2006 *Proceedings of the 17th meeting of the International Association for Statistical Computing—European Regional Section: Computational Statistics (COMPSTAT 2006)* ed Rizzi A and Maurizio V (Springer)

Disclaimer: The submitted manuscript has been created by UChicago Argonne, LLC, Operator of Argonne National Laboratory (“Argonne”). Argonne, a U.S. Department of Energy Office of Science laboratory, is operated under Contract No. DE-AC02-06CH11357. The U.S. Government retains for itself, and others acting on its behalf, a paid-up nonexclusive, irrevocable worldwide license in said article to reproduce, prepare derivative works, distribute copies to the public, and perform publicly and display publicly, by or on behalf of the Government.