Supplementary Information for: The Past as a Stochastic Process

David H. Wolpert,[1,2,3] Michael H. Price,[4] Stefani A. Crabtree,[1,5] Timothy A. Kohler,[1,6,7] Jürgen Jost,[1,8] James Evans,[1,9] Peter F. Stadler,[1,10,11] Hajime Shimao,[1] and Manfred D. Laubichler[1,3]

[1]*Santa Fe Institute*[*]
[2]*Complexity Science Hub Vienna*
[3]*Arizona State University*
[4]*Channel Cognition, LLC*
[5]*Utah State University*
[6]*Washington State University*
[7]*Crow Canyon Archaeological Center*
[8]*The Max Planck Institute for Mathematics in the Sciences*
[9]*University of Chicago*
[10]*Leipzig University*
[11]*The Max Planck Institute for Mathematics in the Science*

I. TYPES OF DYNAMICS

A. Which stochastic process model to use

Although we believe there are many advantages to using stochastic processes, there are also potential limitations and important considerations in choosing the type of stochastic process model to use. In this appendix, we discuss some of these issues, and provide further details on examples discussed in the main text. In many cases, the examples offer a concrete illustration of the limitations and pitfalls that we must otherwise discuss fairly abstractly (and, of course, of the benefits). One major set of issues has to do with the sparsity of archaeological data, which means that a formal stochastic model may not capture every salient aspect of the socio-political-environmental dynamics, and implies that, a *priori*, we should not necessarily propose detailed models with too many parameters and explicit features. This is one benefit using first order Markov models rather than more complex models with more delays. This problem can be somewhat mitigated by using prior information in a Bayesian modeling framework.

We should regardless be aware of the shortcomings of first order Markov models. We treat potentially deterministic fluctuations whose underlying causes we do not grasp as stochastic. Some models may be "blind" to details of human agency. Since the data and models operate at a rather coarse level, as will be discussed below, it is possible to violate the Markovian assumption. Similarly, even though the some underlying variables, such as humans, polities, dollars, etc., are discrete, we often work with continuous state spaces to simplify the mathematics. Fortunately, this is usually a good and legitimate approximation. In spite of these and other shortcomings, these first order Markov models offer crucial advantages, as they allow us to capture dominant features and qualitative aspects in a robust manner. They can be flexibly adapted to add complexity when new data come in. They are simple enough for understanding the dynamics they generate. They may allow us to identify driving forces and critical turning points in historical processes.

There are many variants of the basic first-order Markov process given in Eq. 1. For example, if the state space is discrete rather than real-valued, then the integral in Eq. 1 gets replaced by a sum. If in addition one models a historical process as evolving in discrete time, e.g., years, then the derivative on the LHS of Eq. 1 gets replaced by a discrete-difference. In fact, that is also the setting of the formal example discussed below, as well as the *discrete-time* Markov chains discussed in Section 2B and Section 2D of the main text.

It is important to note though that there are subtle assumptions that arise if we use a discrete-time Markov chain. It turns out that a sizeable portion of all discrete-time Markov chains are theoretically impossible, if one presumes that the true underlying Markov process is actually continuous in time. As a striking example, suppose we have a system with only two states, $\{-1, 1\}$, e.g., due to coarse-graining. The simple discrete-time Markov chain that flips the two possible states, sending $-1 \leftrightarrow +1$, cannot even be approximated as arising from an underlying continuous-time Markov process over those two states [1, 2]. In fact, the set of all discrete-time Markov chains that cannot even be approximated with a continuous-time Markov process has nonzero measure (according to any of the usual measures over the space of stochastic matrices defining discrete-time Markov chains). The same is true if we restrict attention to discrete-time Markov chains that (unlike the bit flip) are highly non-deterministic. Since the physical world is in fact continuous in time, this means that if one wishes to fit a discrete-time Markov model to time series generated by some evolving physical system — including sociopolitical systems — one should exercise great care, to avoid accidentally selecting a Markov chain that is physically impossible.

In addition, even in the context of continuous-time models, the assumption of a first-order Markov process is

* david.h.wolpert@gmail.com; http://santafe.edu/~dhw

a very strong one. Formally, it means that knowledge of the current state of $\mathbf{x}$ suffices to compute the probabilities of its future states. This amounts to assuming that, due to the nature of the variables in $\mathbf{x}$, knowledge of past values of $\mathbf{x}$ will not lead to better predictions of future values, beyond knowing the current value. This assumption is satisfied when all relevant details of the state are known. However, if we coarse-grain the state, that is, lump similar values together into some kind of coarse or macro state, then the dynamics of those states need no longer be Markovian. (See [3] for a systematic analysis of coarse graining.) Likewise, if there are hidden variables that cannot be directly observed, but influence the dynamics of the observable state $\mathbf{x}$, the latter's dynamics need not be Markovian. Generally, when our information about the current state is incomplete, be it due to coarse graining, hidden variables, or some other factor, we may have to draw upon the memory of the states $\mathbf{x}$ to make better predictions.

Ideally, we could do this by fitting a higher-order Markov model to the data. Such an approach is closely related to delay-embedding techniques [4]. (Note that delay embeddings capture chaotic dynamics, which is not possible with first-order Markov processes.) However, in practice, the amount of data one needs to fit an order-n model grows exponentially with the the size of the space and the value of n — using a larger value of n will result in a poor statistical fit. Especially in the context of fitting historical data-sets, where data is quite sparse, this can mean that for purely statistical reasons we have to either choose $n = 1$, or adopt a careful Bayesian analysis if we wish to fit the data with an $n > 1$ model. (However, see [5] for a recent example of trying to fit higher-order models with non-Bayesian methods even when data are sparse, in the specific context of historical data.)

In practice, it is probably most common to use cross-validation to determine n, as well the other hyperparameters in one's model, even if one adopts a Bayesian approach. It is worth noting that there are alternative approaches though, which don't involve cross-validation. For example, in a hierarchical Bayesian approach, one would average over the hyperparameters according to a hyperprior. As another example, one could set hyperparameters using the semi-Bayesian approach of ML-II. (As a technical comment, the use of Bayesian "Occam factors" should not be used to choose n, since they build in a bias to low-dimension models; see [6].)

B. Noise versus chaos versus bifurcations

Often stochastic processes can be viewed as a deterministic evolution of the variable $\mathbf{x}$ with noise superimposed. In particular, for a broad class of functions $\mathbf{W}_{\boldsymbol{\theta}}(\mathbf{x}|\mathbf{x}')$, Eq. 1 in the main text, which involves the dynamics of a time-dependent probability distribution $p_t(\mathbf{x})$, can be reformulated as a noisy equation for the time-dependent state of the system, $\mathbf{x}(\mathbf{t})$:

$$\frac{d}{dt}\mathbf{x}(\mathbf{t}) = A(\mathbf{x}(\mathbf{t}), \mathbf{t}) + B(\mathbf{x}(\mathbf{t}), \mathbf{t})\xi(t) \tag{1}$$

where the $\xi(t)$ is the Wiener noise process, and the functions $A(.)$ and $B(.)$ are determined by $\mathbf{W}_{\boldsymbol{\theta}}(\mathbf{x}|\mathbf{x}')$. (This is the "Langevin equation", discussed in the text.) $A(\mathbf{x}(\mathbf{t}))$ can be viewed as the deterministic dynamics of the variable $\mathbf{x}(\mathbf{t})$, with $B(\mathbf{x}(\mathbf{t}))$ determining the amount of noise superimposed on that dynamics.

The conceptual distinction between deterministic and stochastic dynamics can get blurred in practice because the deterministic dynamics may be chaotic and therefore appear random. Chaos essentially means that very small fluctuations can get amplified so that from very similar initial conditions very different endpoints can be reached after a long enough time, even if the dynamics are deterministic. Fortunately, people have developed sophisticated methods to distinguish chaotic and stochastic components in time series [7]. The formal concept of a stochastic process can accommodate both deterministic and stochastic features. In such a process, the probabilities for a state variable $\mathbf{x}$ change in time according to some rule that is described by some parameter θ. The state $\mathbf{x}$ is observed, while the parameter θ defines the model and can only be statistically inferred, but not directly observed. θ itself may also change, in which case the process is called non-stationary, but usually on a slower time scale than $\mathbf{x}$. In nonlinear dynamics, the qualitative properties of the dynamics may change at particular values of θ. One speaks of bifurcations. That is, a very small variation of the parameter can send the dynamics into completely different regimes. Thus, while in chaotic dynamics, the future of a trajectory may depend very sensitively on the initial conditions, at bifurcations, the dynamics depends very sensitively on a parameter value. It is clearly important to identify such bifurcation points in historical dynamics.

II. PREVIOUS EXAMPLES CONSIDERING HISTORY AS A DYNAMIC PROCESS

For some time a few archaeologists and historians have been graphing behaviors of societies through time in small state spaces, usually considering two variables at a time. Such phase plots implement part of the program we discuss here, since they make it possible to describe trajectories through time in these small state spaces, though they do not generally attempt the fundamental step of formulating the stochastic process model that underlies the behaviors such plots reveal. They are descriptive, graphic devices that do serve to identify semi-cyclic tendencies and possible discontinuities through time. Examples include [8], who examine the frequency of interpersonal violence against population size through time, and [9], in which the number of communities and the population size of a study area are plotted against each

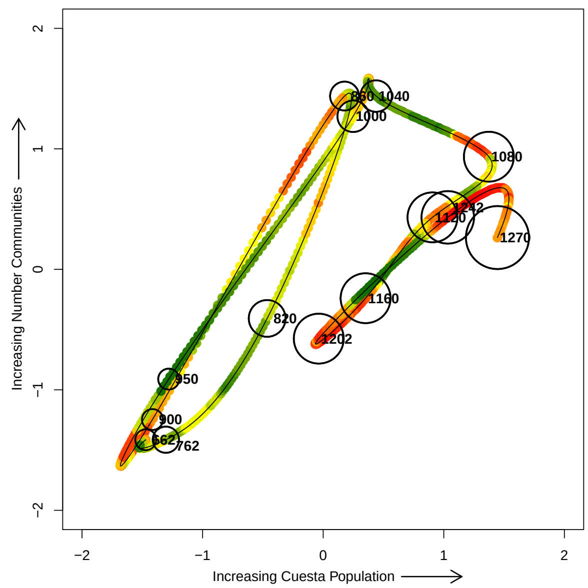


FIG. 1: Relationship through time between number of communities and population on the Mesa Verde cuesta, Colorado shown in z-score space across the periods used by the Village Ecodynamics Project. This area was first densely colonized by farming populations ca. A.D. 600 (at bottom left of figure, unlabeled), and was completely depopulated by ca. AD. 1285. Each point is plotted at the midpoint (AD) of each of the 14 periods recognized; the size of each bubble is proportional to average community population at that time. A 20-year smoothed maize productivity niche is shown on a red (low productivity) to green (high productivity) spectrum. Source: [9, Figure 6]

.

other, using line colors and symbol sizes to put these into the contexts of estimated maize production levels and average community sizes (Fig. 1). In the case considered in that figure, a population bearing a new sociopolitical system intruded on this area between the AD 1040 and 1080 points. From the perspective of this study area, that immigration can be considered as an exogenous perturbation that changed the subsequent evolution of the social and settlement system by (among other things) allowing larger communities to be supported (hence, changing θ).

Peter Turchin and colleagues have investigated the dynamics of agrarian states with special attention to their demography, social structure, surplus extraction, and sociopolitical instability [10]. These variables, and their proxies, are expected to exhibit a patterned relationship with each other through time based on theory developed in [11], [12]. These examples go further than the archaeological cases mentioned above by positing formal models, though there is no attempt to quantify the fit between any of these models and the empirical examples, or to derive models directly from the data.

III. USING PC1/PC2 HINGES AS REFERENCE POINTS FOR THE APPEARANCE OF MORALIZING GODS

So long as one can evaluate the value of Seshat's PC1 for the polities recorded in that data set, one can quantify how "complex" those polities were when they underwent events recorded in that data set. In earlier work with Seshat this was done simply by identifying PC1 with complexity. In particular, [13] used a time series fit to Seshat data to determine when "jumps" occurred in the complexity of individual polities. The times of those jumps were then compared to the time of appearance of Moralizing Gods (MGs) in the polity. [14] The conclusion was that jumps in complexity are preconditions for the appearance of MGs.

Note that this approach assumes that all intervals in PC1 space correspond to the same amount of "social complexity", since it identifies large changes in PC1 during small times as "jumps in complexity". In addition, in order to assign a time of the complexity jump to any particular polity based on its time series, which is then used to determine whether the jump was before the MG onset for that polity, any time series was discarded from the data set unless it had a continuous sequence of PC1 values stretching to before the MG onset. This introduces statistical artifacts.[15]

The discovery of hinge points provides an alternative way to investigate the relationship between jumps in social complexity and the appearance of MGs, without these two shortcomings.

As shown in Figure 2, MGs arise in polities both before and after the polity undergoes a "jump in complexity". At least based on the complexity thresholds found by analyzing Seshat [16], there are instances in which MGs arise before the threshold at PC1 $= -2.5$, and instances in which MGs arise after the threshold at PC1 $= -0.5$.

From a social science perspective, this suggests that a discontinuity in social complexity neither causes the onset of MGs nor requires them, contradicting several other analyses in the literature (including one that was based on the same Seshat dataset). In terms of building an underlying stochastic process model, these results are broadly consistent with the hypothesis that the onset of MGs is a jump of the first kind, where a large change of $\mathbf{x}(t)$ occurs with small but non-negligible probability under the transition matrix $W_\theta(\mathbf{x}|\mathbf{x}')$.[17]

Note though that precisely because such a Poisson process is independent of the current value of $\mathbf{x}(t)$, as well as previous values, it is very challenging to distinguish the hypothesis that the onset of MGs is a Poisson process from the hypothesis that it is be an exogenous perturbation. This is a significant difference from the hinge points themselves. Because the thresholds captured in those hinge points *do* depend on the current value of $\mathbf{x}(t)$, (by definition), and since the precise value of t differs widely from one polity to another, those hinge points seem much more certain to be a jump, albeit of the second kind.

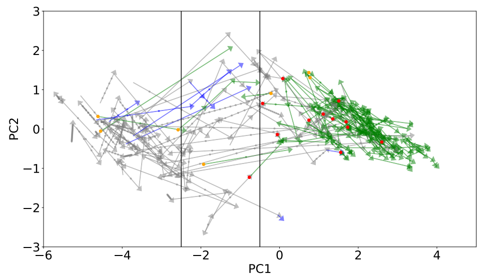


FIG. 2: Movement plot for the moralizing gods material. The x-axis is the PC1 value and the y-axis is the PC2 value. This is the original Seshat data. The first vertical line marks the scale threshold and the second line marks the information threshold [16]. The arrow colors indicate the status of the of the MoralisingGods variable at the beginning of the movement, where grey indicates missing/unknown, blue indicates absent, and green indicates present. The orange points mark transitions from unknown to present and the red points mark transitions from absent to present. The root cause of the retraction of the original moralising gods paper [18] is that transitions from unknown to present were treated as absent.

IV. A 1-DIMENSIONAL RANDOM WALK AS A SIMPLE STOCHASTIC PROCESS

We illustrate stochastic processes using perhaps the simplest such process there is. A random variable x that at every discrete time step assumes some integer value, and from one time step to the next can change its value by ± 1, The probability of going up is p, and that of going down hence $1 - p$. Such jumps at different times are independent of each other. Only x, but not p can be directly observed. p can only be estimated from the observed data of x. Suppose we observe a finite time series of points generated by sampling such a process, but cannot directly observe p. If p changes at some time t – an exogenous perturbation that we do not observe – then the dynamics after t will be different. But based only on the time series, no matter how different it may look before and after t, we cannot definitely conclude that p changed, either at t or some other time. We can only suspect such a change when the time series starts to look qualitatively different. The best we can do is statistically estimate that such a change occurred. Fortunately, there do exist powerful techniques for such estimates.

Next, suppose that p never changes, but we coarse-grain x into bins of width 5. Then if we know that the system is currently in the bin $\{0, 1, 2, 3, 4\}$, in order to predict the probability that at the next time step the system will be in bin $\{5, 6, 7, 8, 9\}$ we need to estimate the relative probabilities of which precise point in $\{0, 1, 2, 3, 4\}$ the system is currently in. In general, we will assign a non-zero probability to the event that the system is currently at the precise point 4, and therefore assign a nonzero probability that at the next time step the system is in $\{5, 6, 7, 8, 9\}$. On the other hand, if we also know that at the previous time step the system was in the bin $\{-5, -4, -3, -2, -1\}$, then in fact we know the system is currently at the point 1, with probability 1, and so cannot be in the bin $\{5, 6, 7, 8, 9\}$ in the next time step. So knowing something about the past state of the system, in addition to knowing its current state, changes the relative probabilities of its future states. This illustrates that coarse-graining the observed time series will in general change a Markovian dynamics into a non-Markovian one.

V. SESHAT: GLOBAL HISTORY DATABANK

Begun in 2009, the Seshat project has pursued the ambitious goal of developing a dataset to test theories about sociocultural evolution by cataloging the development of human civilization from the dawn of the Neolithic to the Industrial Revolution [19]. Two foundational elements of Seshat are the Natural Geographic Area (NGA) and polity. An NGA is a roughly 100 km by 100 km geographic area delimited by natural geographical features such as river basins, coastal plains, valleys, islands, and so forth [20]. A polity is an independent political unit that controls territory, and can range in size from small groups organized in local, independent local communities to territorially expansive, multi-ethnic empires. At any given time, exactly one polity controls an NGA, though that controlling polity may have its base or capital outside the NGA. For example, the Konya Plain NGA is located in the Central Anatolia Region of contemporary Turkey and has an area of 28,900 square km. It was controlled by the Hittite Empire (a polity) in 1000 BC and by the Eastern Roman Empire (also a polity) in AD 500. Seshat data used in recent analyses (i.e., the CCs in the World Sample 30; see below) are coded at 100-year intervals. Thus, a useful way to think about the Seshat data is as a data matrix in which each row consists of an NGA-polity-time triplet.

The Seshat database contains an ever-growing set of variables, now well over 1500, and cases, coded through time for each polity in consultation with archaeological and historical experts. For example, for the Hittite Empire in 1000 BC, the following variables related to money have these codes:

Neo-Hittite Empire in 1000 BC (Population 1.3 - 2.0 million)

Articles	Present
Tokens	Absent
Precious Metals	Present
Foreign coins	??
Indigenous coins	??
Paper currency	Absent
...	...

For comparison, when the Konya Plain was controlled by the Eastern Roman Empire in AD 500, it had a population of about 15 million people.

Seshat is constantly evolving as new data are added and old data re-assessed. This includes the addition of new variables and new NGAs. Some relatively recent articles used a fixed version with 30 NGAs called the World Sample 30. NGAs for this sample were chosen to maximize geographic extent and diversity in social organization. In particular, 3 NGAs were chosen from each of 10 world regions (Africa, Europe, Central Eurasia, Southwest Asia, South Asia, Southeast Asia, East Asia, North America, South America, and Oceania-Australia) and the 3 NGAs in each region were selected such that sociopolitical complexity arrived relatively early, intermediate, and late. The more recent Equinox version of the dataset has five additional NGAs. Both the original and Equinox versions provided a summary version of the dataset in which variables were aggregated into distinct "Complexity Characteristics" (CCs). For the original dataset, 51 variables were collapsed into 9 CCs:

1. Population: Population of the entire polity
2. Territorial Area: Territorial extent (area) of the polity
3. Capital Population: Population size of the largest urban center (usually the capital)
4. Hierarchical Levels: Number of types of settlements (e.g., hamlets to cities) and levels of administrative hierarchy
5. Government: Aspects of government and bureaucracy, such as the presence of a legal code and merit promotion
6. Infrastructure: Presence of bridges, roads, irrigation, etc.
7. Information Technology/Writing: Presence and type of writing and recording systems
8. Texts: Presence of specialized literature, including scientific texts, histories, calendars, fiction, etc.
9. Money: The monetary system—presence of local/foreign currencies, paper currency, tokens of exchange, etc. (see Neo-Hittite example above)

The Equinox dataset collapses Information Technology/Writing and Texts into a single CC. Each of the CCs is normalized to lie between 0 and 1, and for each polity-time pair in the dataset there exists an observation for each CC. One challenge of working with the Seshat data is that, especially for earlier polities, there is insufficient evidence to code many variables—and often disagreements among experts. Rather than assign a value for each CC and create one canonical imputation, the Seshat team assigned a distribution of possible values for the original dataset. These distributions are then sampled 20 times, to produce 20 replicates of the dataset [20, p. 7]. The imputation is performed on each replicate, producing 20 different sets of CC values. These replicates are used to produce confidence intervals on the proportion of variance explained by each PC, the component loadings, and the values of the PCs for each polity. The Equinox dataset, however, only contains one imputation per observation.

VI. A HIDDEN MARKOV MODEL OF AGE-SPECIFIC DEMOGRAPHY

Consider an age structured population of reproductive females (including juveniles) where the age class j contains individual between $\Delta a \cdot j$ and $\Delta a \cdot (j+1)$ years of age, with Δa being the age-spacing; it is straightforward to generalize this to males and post-reproductive females, as well as to account for other types of population structure, such as spatial location and socioeconomic status[21, 22]. The column vector $\mathbf{z}_t$ gives the number, or proportion, of females in each age class at time step t. F_j is the age-specific fertility of females in age class j, accounting only for female offspring. The age-specific survival of females in age class j is P_j. Given these definitions, the population projection matrix for reproductive females is

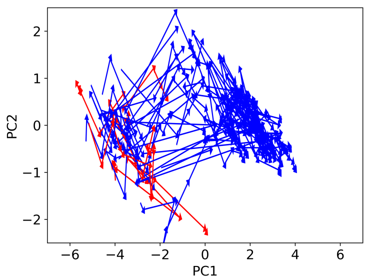
$$\mathbf{A} = \begin{pmatrix} F_0 & F_1 & F_2 & F_3 & \cdots & F_J \\ P_0 & 0 & 0 & 0 & \cdots & 0 \\ 0 & P_1 & 0 & 0 & \cdots & 0 \\ 0 & 0 & P_2 & 0 & \cdots & 0 \\ 0 & 0 & 0 & \ddots & \cdots & 0 \\ 0 & 0 & 0 & \ddots & P_{J-1} & 0 \end{pmatrix}, \tag{2}$$

where J is the last reproductive age class. The population project equation is

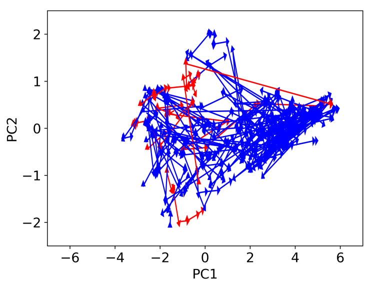
$$\mathbf{z}_{t+1} = \mathbf{A}\,\mathbf{z}_t. \tag{3}$$

These preceding equations succinctly summarize three effects: Females get older, that is, advance from one age class to the next, give birth to young females, and may die.

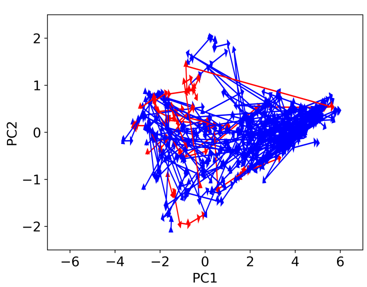
Stable demography: If demographic rates are stable [23] the population vector $\mathbf{z}_t$ converges to a constant growth rate and stable age distribution. The growth "rate" (more precisely, the per period growth factor) equals λ, the dominant left eigenvalue of $\mathbf{A}$. The stable age distribution ($\mathbf{u}$) equals the corresponding dominant



(a) Original data (and thus original NGAs)



(b) Equinox data with only original NGAs



(c) Equinox data with original NGAs and new Equinox NGAs

FIG. 3: Movement plots in PC1-PC2 space. Each arrow is the movement in the space for one NGA from one timestamp to the next timestamp 100 years later. Blue arrows are for Old World NGAs and red arrows are for New World NGAs. The top plot is identical to that in [16]. The middle plot uses the new Seshat Equinox dataset, but the new NGAs added to Equinox have been removed. The bottom plot uses the new dataset and utilizes all Equinox NGAs. The key difference between the original datasest and the Equinox dataset is that the latter dataset uses one fewer CCs, collapsing the two original information processes CCs into a single CC. These figures can be reproduced using the code in the publication github repository.

right eigenvector and the age specific reproductive value ($\mathbf{v}$) equals the corresponding left eigenvector [22, 24].

Time dependence: Rather than assuming a constant population projection matrix, we now add a subscript to allow time-dependence, $\mathbf{A}_t$. The time-dependent population projection equation is

$$\mathbf{z}_{t+1} = \mathbf{A}_t\, \mathbf{z}_t. \tag{4}$$

One can easily expand the preceding model to a full demographic model that includes males and post-reproductive females. Parameters such as the sex ratio at birth and age-specific mortality for the additional segments of the population can be inferred from known demographic statistics and/or observed data. The full demographic specification for time period t is $\mathcal{D}_t = \{\mathbf{F}_t, \mathbf{P}_t\}$, where $\mathbf{F}_t$ is the time-dependent vector of age-specific fertilities and $\mathbf{P}_t$ is the time-dependent vector of age-specific survivals, and the demographic model for the period 0 to t is $\mathcal{G}_t = \{\mathcal{D}_0, \mathcal{D}_1, \cdots, \mathcal{D}_{t-1}\}$. An implicit assumption in this definition is that the population vector at time step 0 is set assuming stable demography defined by the demographic specification $\mathcal{D}_0$. This assumption could be relaxed by letting $\mathbf{z}_0$ be a free parameter in the demographic model $\mathcal{G}_t$.

A. From demographic model to likelihood and a hidden Markov model

To link the preceding model to the stochastic process framework suggested in the main text, we now interpret the population projection matrices $\mathbf{A}_t$ as being determined by latent variables in a hidden Markov model. Further, we assume that there exists an external variable, let's say for the sake of concreteness a categorical climate variable (likely slowly changing) that can be directly or indirectly observed, indexed by $k = 1 \cdots K$, where for each climate state, k, different transition probabilities apply:

$$\mathbf{p}_t = \left[\sum_{k=1}^{K} c_k\, \mathbf{W}^{(k)}\right] \mathbf{p}_{t+1}, \tag{5}$$

where c_k is an indicator variable for the categorical climate state and $\mathbf{W}^{(k)}$ is the transition matrix that applies for climate state k. There are many alternatives to this approach, including modeling the projection matrices $\mathbf{A}_t$ as depending on a continuous, exogenous climate parameter vector $\boldsymbol{\theta}$. However, the model we describe is sufficiently simple to be motivating yet sufficiently complex to be realistic. We assume that there exists a set of N reference dynamics indexed by $n = 1 \cdots N$, where each reference dynamics, $\mathcal{D}^{(n)}$, is for a distinct annual demographic state – for example, one n could correspond to famine, another warfare, etc. Next define the vector

$$\mathbf{q} = \begin{bmatrix} m_1 & m_2 & \cdots & m_T \end{bmatrix}^T \quad (6)$$

to be the population dynamics state, n, for the time periods 1 through T. Given the preceding formulation, it is straightforward to calculate the probability of any given $\mathbf{q}$ given the set of transition probabilities (to be inferred) and the set of reference dynamics (pre-specified); we denote this probability by $p(q|\{\mathbf{W}^{(k)}\}, \{\mathbf{A}^{(n)}\})$, where we indicate the set of transition matrices with $\{\mathbf{W}^{(k)}\}$ and the set of reference population project matrices with $\{\mathbf{A}^{(n)}\}$. One can then sum over the set of valid vectors $\mathbf{q}$ and use Equations 2 and 3 to calculate the probability of each $\mathbf{z}_t[j]$. Finally, this can be linked to available archaeological data to calculate a likelihood function, which can be used for either maximum likelihood estimation or Bayesian inference. For example, given a set of radiocarbon determinations (e.g., as in [25]) one can assume that the probability a given sample is from a given year is proportional to the total population size in that year, the sum of the elements of $\mathbf{z}$. Similarly, if skeletal age-at-death is known, and a rough date estimate is available from associated artifacts, one can calculate the relative probability of being in the pertinent age class (and sex class if the model is suitably generalized) from the population vectors by summing across years as determined by the associated artifacts. Naturally, one can use multiple types of data as part of a single likelihood calculation to improve inference of the underlying model parameters; a major benefit of the approach we have described is that it is straightforward to accommodate additional types of data in the likelihood calculation in order to further improve inference (e.g., isotopic data to inform on migration, health data such as from linear enamel hypoplasias (LEHs) to improve inference on mortality, and both ancient and modern genetic data).

B. Source code and the Python bighist package

We put a good deal of effort into software engineering for this project, efforts which we hope will benefit the broader community. Most importantly, we created a Python package called bighist that adopts a data abstraction framework that is viable for many, perhaps even most, analyses in big history, cultural evolution, etc. This abstraction is implemented in the class bighist.StratifiedTimeSeries. The strata are distinct subseries, or "samples" of the time trajectories. For Seshat, these are the NGAs. For each time in a sub-series, there can be multiple observations of the data vector, which provides support for multiple imputations. The bighist package provides utility functions to load Seshat data, do dimensionality reduction using PCA, and make movement plots. We used it to create Figures 1 and 2 in the main text and Figure 3 in this supplement. Our vision of the package is that it can provide a unifying framework for analyses where the underlying dataset matches the data abstraction assumptions, including the types of stochastic process analyses we discuss in our article. In addition, we hope to add additional datasets to bighist, further simplifying the process of doing the types of analyses we discuss. To ensure the code is high quality and works as intended, bighist utilizes unit tests and adheres as closely as possible to Python Enhancement Proposal (PEP) recommendations for syntax and formatting. The source code for bighist is available here

The README file for the preceding bighist provides provides installation instructions. Our source code for this article, available at the following link, provides further instructions and example code here

For the moralizing gods material we utilized the R code here. Note well that all of the code in the following repository is deprecated (replaced by the preceding github repository) aside from the moralizing gods code. A warning of this is provided in the README of the repository.

For Figure 3 of the main text, we utilized the npde (Nonparametric Differential Equations with Gaussian Processes) code located here:

https://github.com/jinhongkuan/npde

Finally, the following repository has a power calculation that implements some of the ideas we discuss for using stochastic latent variable models to fit demographic processes. Note well that this is a separate piece of work – i.e., we do not use any results from that repository in this article. We link to it because it demonstrates that the overall approach is viable and the code may be of interest to some readers.

https://github.com/MichaelHoltonPrice/kelmelis_maya_nsf

[1] Jeremy A. Owen, Artemy Kolchinsky, and David H. Wolpert, "Number of hidden states needed to physically implement a given conditional distribution," New Journal of Physics **21**, 019501 (2018).

[2] David H. Wolpert, Artemy Kolchinsky, and Jeremy A. Owen, "A space-time tradeoff for implementing a function with master equation dynamics," Nature Communications **10**, 1727 (2019).

[3] Oliver Pfante, Nils Bertschinger, Eckehard Olbrich, Nihat Ay, and Jürgen Jost, "Comparison between different methods of level identification," Advances in complex systems **17**, 1450007 (2014).

[4] Elizabeth Bradley and Holger Kantz, "Nonlinear time-series analysis revisited," Chaos: An Interdisciplinary Journal of Nonlinear Science **25**, 097610 (2015), https://doi.org/10.1063/1.4917289.

[5] Nishant Malik, "Uncovering transitions in paleoclimate time series and the climate driven demise of an ancient

civilization<? a3b2 show [editpick]?>," Chaos: An Interdisciplinary Journal of Nonlinear Science **30**, 083108 (2020).

[6] David H Wolpert *et al.*, "On the bayesian occam factors argument for occam's razor," (1995).

[7] Holger Kantz and Thomas Schreiber, *Nonlinear Time Series Analysis* (Cambridge Univ. Press, 1997).

[8] Timothy A. Kohler, Sarah Cole, and Stanca Ciupe, "Population and Warfare: A Test of the Turchin Model in Pueblo Societies," in *Pattern and Process in Cultural Evolution*, edited by Stephen Shennan (University of California Press, Berkeley, 2009) pp. 277–295.

[9] Kelsey M. Reese, Donna M. Glowacki, and Timothy A. Kohler, "Dynamic Communities on the Mesa Verde Cuesta," American Antiquity **84**, 728–747 (2019).

[10] Peter Turchin and Sergey A. Nefedov, *Secular Cycles* (Princeton University Press, Princeton, NJ, 2009).

[11] Jack A. Goldstone, *Revolution and Rebellion in the Early Modern World* (University of California Press, Berkeley and Los Angeles, 1991).

[12] Peter Turchin, *Historical Dynamics: Why States Rise and Fall* (Princeton University Press, Princeton, NJ, 2003).

[13] Harvey Whitehouse, Pieter Francois, Patrick E Savage, Thomas E Currie, Kevin C Feeney, Enrico Cioni, Rosalind Purcell, Robert M Ross, Jennifer Larson, John Baines, *et al.*, "Complex societies precede moralizing gods throughout world history," Nature **568**, 226–229 (2019).

[14] Presence of moralizing gods has now been added as variable in Seshat, to allow this analysis by [13], but was not a component of the PC values discussed in this section.

[15] To see this, as a null hypothesis, suppose that all jumps in social complexity occur during the interval from -1.0 to 1.0 in PC1. Suppose as well that MG onset PC1 values occur independently, by sampling a Gaussian centered on PC1 $= 0$. Finally, suppose that due to archaeological artifacts, half of all time series have a continuous sequence of PC1 values stretching back only to PC1 $= 1.0$, and the other half stretch back to before PC1 $= -1.0$. Then it will appear that there is probability 2/3 of MG onset occurring after the jump in social complexity, even though the real probability is 1/2.

[16] Jaeweon Shin, Michael Holton Price, David H. Wolpert, Hajime Shimao, Brendan Tracey, and Timothy A. Kohler, "Scale and information-processing thresholds in Holocene social evolution," Nat. Comm. **11**, 2394 (2020).

[17] Intuitively, this hypothesis is similar to supposing that the onset of MGs is a Poisson process, with a rate that is non-negligible across a broad span of PC1 values. We emphasize though that we have not yet done a proper statistical test of this hypothesis.

[18] Harvey Whitehouse, Pieter François, Patrick E Savage, Thomas E Currie, Kevin C Feeney, Enrico Cioni, Rosalind Purcell, Robert M Ross, Jennifer Larson, John Baines, *et al.*, "Retraction note: Complex societies precede moralizing gods throughout world history," Nature **595**, 320–320 (2021).

[19] Peter Turchin, Rob Brennan, Thomas E. Currie, Kevin C. Feeney, Pieter Francois, Daniel Hoyer, J. G. Manning, Arkadiusz Marciniak, Daniel Mullins, Alessio Palmisano, Peter Peregrine, Edward A. L. Turner, and Harvey Whitehouse, "Seshat: The global history databank," Cliodynamics **6**, 1–107 (2015).

[20] Peter Turchin, Thomas E. Currie, Harvey Whitehouse, Pieter François, Kevin Feeney, Daniel Mullins, Daniel Hoyer, Christina Collins, Stephanie Grohmann, Patrick Savage, Gavin Mendel-Gleason, Edward Turner, Agathe Dupeyron, Enrico Cioni, Jenny Reddish, Jill Levine, Greine Jordan, Eva Brandl, Alice Williams, Rudolf Cesaretti, Marta Krueger, Alessandro Ceccarelli, Joe Figliulo-Rosswurm, Po-Ju Tuan, Peter Peregrine, Arkadiusz Marciniak, Johannes Preiser-Kapeller, Nikolay Kradin, Andrey Korotayev, Alessio Palmisano, David Baker, Julye Bidmead, Peter Bol, David Christian, Connie Cook, Alan Covey, Gary Feinman, Árni Daníel Júlíusson, Axel Kristinsson, John Miksic, Ruth Mostern, Cameron Petrie, Peter Rudiak-Gould, Barend ter Haar, Vesna Wallace, Victor Mair, Liye Xie, John Baines, Elizabeth Bridges, Joseph Manning, Bruce Lockhart, Amy Bogaard, and Charles Spencer, "Quantitative historical analysis uncovers a single dimension of complexity that structures global variation in human social organization," Proceedings of the National Academy of Sciences **115**, E144–E151 (2018), http://www.pnas.org/content/115/2/E144.full.pdf.

[21] Andrei Rogers, "The Multiregional Matrix Growth Operator and the Stable Interregional Age Structure," Demography **3**, 537–544 (1966).

[22] Hal Caswell, *Matrix Population Models: Construction, Analysis and Interpretation*, 2nd ed. (Sinauer, Sunderland, MA, 2001).

[23] Certain technical constraints on $\mathbf{A}$ must also be satisfied. In particular, $\mathbf{A}$ must be irreducible and primitive, which is almost always true.

[24] Ronald A. Fisher, *The genetical theory of natural selection* (Dover, New York, 1958).

[25] Michael Holton Price, José M. Capriles, Julie A. Hoggarth, R. Kyle Bocinsky, Claire E. Ebert, and James Holland Jones, "End-to-end Bayesian analysis for summarizing sets of radiocarbon dates," Journal of Archaeological Science **135**, 105473 (2021).