ANTHROPOLOGY

Disentangling the evolutionary drivers of social complexity: A comprehensive test of hypotheses

Peter Turchin[1,2,3], Harvey Whitehouse[3]*, Sergey Gavrilets[4], Daniel Hoyer[5,6,7], Pieter François[3], James S. Bennett[8], Kevin C. Feeney[9], Peter Peregrine[10], Gary Feinman[11]†‡, Andrey Korotayev[12], Nikolay Kradin[13], Jill Levine[14], Jenny Reddish[1], Enrico Cioni[5], Romain Wacziarg[15], Gavin Mendel-Gleason[9], Majid Benam[1]

During the Holocene, the scale and complexity of human societies increased markedly. Generations of scholars have proposed different theories explaining this expansion, which range from broadly functionalist explanations, focusing on the provision of public goods, to conflict theories, emphasizing the role of class struggle or warfare. To quantitatively test these theories, we develop a general dynamical model based on the theoretical framework of cultural macroevolution. Using this model and Seshat: Global History Databank, we test 17 potential predictor variables proxying mechanisms suggested by major theories of sociopolitical complexity (and >100,000 combinations of these predictors). The best-supported model indicates a strong causal role played by a combination of increasing agricultural productivity and invention/adoption of military technologies (most notably, iron weapons and cavalry in the first millennium BCE).

INTRODUCTION

During the Holocene—roughly, the past 10,000 years—the scale and complexity of human societies have been utterly transformed. This transformation was a multidimensional process (*1*–*4*). The social scale at which humans interact and cooperate increased by six orders of magnitude, from societies of hundreds (or a few thousand) to hundreds of millions and even billions (*4*). A particular form of political organization, the state, arose in the mid-Holocene, eventually becoming the dominant form of social organization across the world. Other dimensions of change include not only increasingly productive economies and widespread adoption of writing and literacy but also deeper inequalities and entrenched class hierarchies (*5*, *6*).

Generations of historians, anthropologists, and philosophers have offered a diversity of theories to account for these marked changes in the social scale and complexity of human political formations. The view that agriculture was a necessary condition for the evolution of complex societies, which was crystallized in the work of early anthropologists Childe (*7*), White (*8*), and Service (*9*), is implicitly, and often explicitly, endorsed today by most scholars of the past. According to some, food production was not only a necessary but also a sufficient cause (*10*). Widespread adoption of agriculture was generally tied to increasing sedentarization, storable food surpluses, and human population explosion. Surpluses from agriculture supported division of labor into increasingly specialized units (*11*). This, in turn, allowed for the emergence of full-time craftspeople and inventors, which drove the cumulative growth of technology. Productive economies also undergirded the appearance of rulers and elites, full-time bureaucrats, military officers, and soldiers, resulting in political centralization, social stratification, and increasingly violent conflict.

The argument that agriculture contributed to the rise of sociopolitical complexity has a long history and remains generally accepted by archaeologists despite some widely discussed objections, citing evidence from complex hunter-gatherer societies (*12*, *13*). While agriculture may not be necessary or sufficient for foraging bands to increase in size and social complexity, the upper threshold for this growth appears to be set much lower than for agricultural societies. However, social scientists are more divided on the nature of factors over and above the effects of agriculture in driving the rise and spread of states (*6*, *8*, *9*, *14*). Two contrasting approaches to this question had crystallized by the 1960s: functionalist and conflict theories (*15*). Functionalist theories argued that the state evolved and expanded in response to the organizational challenges of increasingly large and complex societies, such as the need to expand long-distance trade (*9*), to manage productivity risk (*5*), or to develop and to maintain irrigation infrastructure (*16*). Conflict theories have tended to emphasize either internal conflict resulting from inequality and class struggle (*14*, *17*) or external warfare (*18*, *19*). Most current theories combine functionalist and conflict perspectives (*20*). Previous empirical tests of these theories have been hampered by two major problems: the paucity of large-scale time-resolved—"diachronic"—data that capture changes in various characteristics of social complexity and the lack of a general conceptual framework with appropriate mathematical models and statistical tests to analyze these data.

The first problem is now being overcome with the development of Seshat: Global History Databank (*21*), which traces, over the past 10,000 years, the developmental sequences of societies in a stratified sample of 35 natural geographic areas (NGAs). These NGAs were selected to maximize diversity, covering 10 world regions spanning

[1]Complexity Science Hub Vienna, Vienna, Austria. [2]Department of Ecology and Evolutionary Biology, University of Connecticut, Storrs, CT, USA. [3]Centre for the Study of Social Cohesion, School of Anthropology and Museum Ethnography, University of Oxford, Oxford, UK. [4]Department of Ecology and Evolutionary Biology, Department of Mathematics, Center for the Dynamics of Social Complexity, University of Tennessee, Knoxville, TN, USA. [5]Seshat: Global History Databank, Evolution Institute, San Antonio, FL, USA. [6]George Brown College, Toronto, Canada. [7]Evolution Institute, San Antonio, FL, USA. [8]University of Washington, Seattle, WA, USA. [9]TerminusDB, Dublin, Ireland. [10]Lawrence University, Appleton, WI, USA. [11]Field Museum of Natural History, Chicago, IL, USA. [12]National Research University Higher School of Economics, Moscow, Russia. [13]Institute of History, Archaeology and Ethnology, Far East Branch of the Russian Academy of Sciences, Vladivostok, Russia. [14]Evolution Institute, Tampa, FL, USA [15]University of California Los Angeles, Anderson School of Management, Los Angeles, CA, USA.
*Corresponding author. Email: harvey.whitehouse@anthro.ox.ac.uk
†Present address: Department of Anthropology, University of Illinois-Chicago, Chicago, IL, USA.
‡Present address: Department of Anthropology, Northwestern University, Evanson, IL, USA.

all world continents (fig. S1A). In each of these 10 world regions, sociopolitical complexity emerged relatively early in at least one NGA, relatively late in at least one other, and somewhere in the middle for at least one further NGA. The aim of this sampling strategy was to capture as much variation as possible in the evolution of sociopolitical complexity over the course of world history, thus helping us to resolve the endogeneity and small sample size problems that have typically hampered previous efforts to analyze the evolution of sociopolitical complexity. The most recent data release (21), which we use here, includes information about >100 variables for 373 societies with temporal coverage between 9600 BCE and 1900 CE (fig. S1B).

The second problem is resolved with the rise of cultural evolution theory (22), which provides a robust framework to conceptualize and investigate long-term sociocultural change in human societies. A key element of this theory for studies of social complexity is cultural macroevolution (23, 24). Building on an analogous distinction between microevolution (i.e., genetic and phenotypic changes within populations) and macroevolution (i.e., changes at or above the species level) in evolutionary biology, cultural microevolution is defined as the change in the frequency of cultural variants within a population (25), and cultural macroevolution is defined as large-scale changes in cultural traits of whole groups (23). This is an important distinction as we recognize that the specific combinations and processual relationships between the different dimensions of social complexity combined and recombined in different ways across space and over time.

The process of cultural macroevolution of polities (i.e., politically independent societies, such as chiefdoms and states) can be formalized mathematically as a nonlinear dynamical system modeled, for example, by difference equations

$$
\begin{aligned}
X_{t+1} &= f(X_t, Y_t, \ldots Z_t, U_t, \ldots V_t) \\
Y_{t+1} &= g(X_t, Y_t, \ldots Z_t, U_t, \ldots V_t) \\
&\quad \cdots \\
Z_{t+1} &= h(X_t, Y_t, \ldots Z_t, U_t, \ldots V_t)
\end{aligned}
\tag{1}
$$

where $X_t, Y_t, \ldots Z_t$ are state variables reflecting cultural characteristics of a polity at time t that are treated as endogenous variables representing dynamical feedbacks; $U_t, \ldots V_t$ are exogenous factors that are not involved in feedback loops; and $f, g, \ldots h$ are nonlinear functions specifying how variables interact with each other. The time step is one century, here chosen to operationalize the temporal resolution of the Seshat Databank. Continuous time analogs of this model include the Ornstein-Uhlenbeck process and its generalizations, which investigators have used to model biological macroevolutionary processes (26). These nonlinear dynamical models allow us to capture the "descent with modification" nature of the evolutionary process, because current values of state variables are conditioned not only on potential causal factors ("modification") but also on their own past values ("descent"). Exogenous variables can represent a variety of processes: white noise, random walks (successive values of V_t are autocorrelated), or a single discontinuous change of the environment. Terms representing spatial diffusion and descent from a common ancestor can also be added and analyzed (see Materials and Methods for details). Depending on the specifics of interaction functions ($f, g, \ldots h$), state variables can undergo an unbiased random walk (neutral evolution), a biased random walk (directional selection), or fluctuate around an equilibrium determined by other state variables (stabilizing selection).

Given diachronic data, this general model (Eq. 1) can be used to investigate a variety of causal scenarios that may give rise to correlation between variables, such as X and Y. At least four scenarios can be distinguished (arrows indicate the direction of causation):

1) $X_t \rightarrow Y_{t+1}$ (with no feedback from Y to X)
2) $Y_t \rightarrow X_{t+1}$ (with no feedback from X to Y)
3) $X_t \rightarrow Y_{t+1}$ and $Y_t \rightarrow X_{t+1}$ (mutual causation)
4) $X_{t+1} \leftarrow Z_t \rightarrow Y_{t+1}$ [no direct causal effect between X and Y but correlation between them because they share common driving factor(s), Z]

Our conceptual framework permits the same variable to be both a response and a predictor, enabling us to capture scenarios of mutual causation. This formalism, also known as Wiener-Granger causality (27, 28), allows us to restate the variety of theories of social evolution, proposed by past and contemporary social scientists, into a common framework so that they can be tested against each other using time-resolved data.

We acknowledge that "causality" is a notoriously difficult statistical, and philosophical, problem (29). Here, we use a particular notion, that of evolutionary causality. Our definition of causality is made mathematically explicit by the general macroevolutionary model (Eq. 1), which underlies the statistical approach of dynamic regression (DR) and leverages the availability of time-resolved data. We stress that this approach to causation is quite different in its goals and statistical methods from that of directed acyclic graphs (DAGs) (30), as we explain more fully in Materials and Methods.

In summary, the core of this article is a "large-n" statistical analysis of a diverse world sample of historical societies, going back in time from c. 1900 to as far as the Seshat data allow. In Discussion, we supplement this comprehensive (both in its geographic reach and the variety of hypotheses tested) analysis with a few individual case studies, highlighting the most important general insights that have emerged from it.

Data variables for testing theories
Main response variables
To describe social complexity, we use three measures: social scale [Scale; the first principal component (PC1) of log-transformed polity population, polity territory, and the population of the largest settlement; for a detailed description of how all variables were defined and measured, see Materials and Methods], vertical or hierarchical complexity (Hier; an average number of levels in administrative, military, and settlement hierarchies), and specialization of governance (Gov; combining 11 measures of government sophistication). These three measures serve as the main response (dependent) variables in the statistical analysis.

Predictor variables
We use five sets of measures related to agriculture, functionalist theories, internal conflict, external conflict, and religion. Our quantitative proxy of the effects of agriculture is its productivity (tons per hectare), Agri. Because there can be a substantial time lag between the transition to agriculture and the rise of large-scale complex societies (11), a second proxy is the antiquity of agriculture (years since adoption of agriculture), AgriLag.

We use a variety of proxies for processes proposed by functionalist (integrative) theories. The provision of public goods and infrastructure (Infra) aggregates 12 Seshat variables, including the presence or absence of water supply systems, food storage sites, roads, bridges, canals, ports, markets, and postal service. The hydraulic society theory

of Wittfogel (16) focuses on one particular state function and is proxied by the presence of Irrigation. Other integrative theories propose economic development as a major driver for the evolution and growth of states. We use three diverse proxies for various aspects of economic development. One is the population of a polity's largest settlement (Cap), because economic historians often use urbanization as a proxy for economic growth (31). The second is a binary variable for the presence of spaces facilitating trade and exchange, Market. The third focuses on the sophistication of the means of exchange, which we proxy using an aggregate of Seshat variables capturing the types of Money used in a polity. Last, some integrative theories postulate that sophisticated institutions of governance evolved in response to the need to manage information flows within a polity (5). We proxy this hypothesis with another aggregate measure of 13 variables capturing information complexity (Info).

A related set of hypotheses focuses on increasing social scale as the driving force behind increasing political complexity and state formation. Social scale itself may be a result of the transition to agriculture, which resulted in an order of magnitude (or greater) increase in sustainable population density. Thus, the "scalar stress hypothesis" (32) posits that governance institutions and efficiently transmitted identity markers evolved to coordinate the work and solve inevitable conflicts among groups of people, which were too large to be integrated by face-to-face interactions (33, 34). We proxy this hypothesis by polity population (Pop). Another direct measure of scale, polity territory (Terr), has also been proposed as a possible driver (35). Separately, an influential current in the evolutionary theorizing of religion proposes that belief in all-knowing, morally concerned, punitive deities—"Big Gods"—increased the ability of groups to sustain large-scale social organizations, as well as successfully scale up and expand by facilitating group cohesion and cooperation within a shared ideological framework (36, 37). We proxy this hypothesis with the synthetic variable moralizing supernatural punishment (MSP), which aggregates several Seshat variables coding for religious characteristics (38).

Internal conflict theories emphasize inequality and conflict between social classes as a major driver for the growth of states (14, 17). One proxy for this hypothesis is social stratification (Class). Here, we use the data on emergent stratification among archaeologically known societies collected by Peregrine (39). Another line of scholarship focuses on the length of chains of command, arguing that the more levels of control and command in a hierarchy, the more power accrues to the individuals occupying the top levels, who will favor centralization and state-level institutions that would protect their advantageous position. We measure this "iron law of oligarchy" (40) with hierarchical complexity (Hier). Another theory emphasizing internal conflict postulates that grain is more storable than root crops, making it easier to appropriate by emerging elites who use this control to accumulate wealth and power, institutionalizing these privileges within state structures (12). The binary Grain variable is coded as 0 if the main carbohydrate sources is a root crop (yam, sweet potato, and taro) and 1 if it is a cereal (wheat, rice, maize, millet, and rye). Last, whereas the Big Gods hypothesis emphasizes an integrative function of religion, it can also reinforce extreme social stratification and inequality. Thus, the "social control hypothesis" proposes that human sacrifice (HS) bolsters the power of elites by legitimating their authority (41)—due to the ritual and ideological significance of the sacrifice as a means of communicating with (or appeasing) supernatural beings—and by motivating compliance via

divinely endorsed intimidation (42). We use the binary Seshat variable HS as a proxy for this hypothesis.

External conflict theories propose that competition between societies, usually taking the form of warfare, imposes a selection regime that weeds out relatively dysfunctional, poorly organized, and internally uncooperative polities, favoring those with larger populations and effective, centralized, and internally specialized institutions (6, 43–45). The main proxy for the conflict hypothesis is the Seshat measure of the realized sophistication and variety of military technologies used by polities, MilTech (46). A large variety of sophisticated means of attack and defense serves as a quantitative proxy for the intensity of warfare in the environment of the polity, because people tend to invest in expensive armor and defenses when their societies are threatened by their neighbors. Warfare intensity is often measured in archaeological datasets using evidence of violent death, such as cranial trauma, but in this case, our concern is not so much to measure rates of death due to intergroup conflict but levels of cooperative investment in strengthening the group's military preparedness and effectiveness in the face of existential threats. Furthermore, we explore the relationship between MilTech and a quantitative measure of warfare intensity, finding it to be approximately linear and characterized by a high correlation coefficient (more than 0.9), suggesting that MilTech captures increasing intensity of interstate conflict and threat (details provided in the "MilTech and war severity" section in the Supplementary Materials)

In addition to the MilTech variables, we have also included technologies contributing to mounted warfare. This is because previous analyses suggest that the invention of effective horse riding in the Pontic-Caspian steppes, combined with iron metallurgy allowing for more effective weapons and armor, elevated the intensity of warfare as it spread from the steppes south to the belt of farming societies, triggering the formation of particularly large states (43, 47). Iron implements boosted agricultural production as well, linking the effects of growing productivity to interstate conflict. We use the data in (48, 49) to construct a synthetic variable (IronCav) that captures the spread of these two key technologies. Although the IronCav variable used in our analyses played a notably important role in Eurasian history, it should be emphasized that the MilTech variable is a fully global one, and our analysis captures these variables across all world regions and time periods in the Seshat database from the Neolithic to the Industrial Revolution (roughly 10,000 years of global history in total).

Model selection and analysis

Model selection (choosing which terms to include in the regression model) was accomplished by exhaustive search: regressing the response variable on all possible linear combinations of predictor variables. This means that we tested >100,000 special hypotheses (this number is further increased because, in addition to the 17 possible predictors, we also investigated the effects of various autoregressive and nonlinear terms; see Materials and Methods). The degree of fit was quantified by the Akaike information criterion (AIC), which penalizes models with too many fitted coefficients. Possible nonlinear effects were checked by adding quadratic terms to the regression model. Standard diagnostic tests were performed for the best-fitting models (50). To check for cross-equation error correlations, we fitted a "seemingly unrelated regression" (51). Missing data values, estimate uncertainties, and expert disagreements in the predictors were dealt with by multiple imputation (52, 53). Because diagnostic tests

indicated that the distribution of residuals is not Gaussian, we used nonparametric bootstrap to estimate the *P* values associated with various regression terms (see Materials and Methods).

RESULTS

Our analysis identified two classes of predictors that, in combination, have a consistent effect on the three complexity variables: external conflict and agriculture (Table 1). The strongest support is for the IronCav variable, which codes for the joint spread of cavalry warfare and iron metallurgy. MilTech, the proxy for warfare intensity, has an additional effect on all three response variables. The productivity of agriculture (Agri) and the antiquity of agriculture (AgriLag) are also selected as predictors for all responses. The rest of the hypotheses are not supported by this analysis. As Table 1 shows, some predictors are sometimes selected in the best models by AIC, but these effects are statistically weak and inconsistent (several are negative),

as expected when multiple model specifications are fitted. This main result is robust to alternative model specifications, as extensively detailed in Supplementary Results.

How strong is the statistical support for models including both agricultural productivity and conflict, compared with theoretical alternatives? An instructive comparison is with the agriculture-plus-functionalism models, because the Seshat project has invested a similar level of effort to conceptualize and code these variables. We compared these two classes of models using the difference between the AIC values for alternative models, delAIC. In general, when AIC for the best-supported model is lower than the AIC of an alternative model by 10 units (delAICs > 10), the alternative model has essentially no statistical support (*54*). We found that the best agriculture-plus-functionalism model formulations were characterized by delAICs varying between 23.19 and 62.65, depending on the response variable (see "Comparison between functionalist and external conflict proxies" in the Supplementary Materials). Such a big

Table 1. Results of fitting DR models for the three response variables and 17 predictor variables. The three columns on the right indicate which predictors were selected in the best model (lowest AIC) and their estimated effect. Empty cells indicate that the predictor was not selected (see details in Supplementary Results). Symbols explanation:

(−)	Negative effect, not significant at the *P* < 0.05 level
(+)	Positive effect, not significant at the *P* < 0.05 level
+	*P* < 0.05
++	*P* < 0.01
+++	*P* < 0.001
++++	*P* < 0.00001
+++++	*P* < 0.000001
NA	Predictor the same as response (Hier) or used in calculating response (Scale); omitted from regressions

	Hypothesis	Variable	Hypothesis class	Scale	Hier	Gov
1	Productivity of agriculture	Agri	Agriculture	+	++	++
2	Antiquity of agriculture	AgriLag	Agriculture	+++	+++	(+)
3	Provision of public goods	Infra	Functional			
4	Hydraulic society	Irrigation	Functional			
5	Urbanization	Cap	Functional	NA		
6	Trade	Market	Functional			
7	Economic exchange	Money	Functional			
8	Information system	Info	Functional		(+)	(+)
9	Scalar stress	Pop	Social scale	NA		
10	Territorial expansion	Terr	Social scale	NA		
11	Social stratification	Class	Conflict—internal	(+)		
12	Iron law of oligarchy	Hier	Conflict—internal		NA	
13	Cereal crops	Grain	Conflict—internal			(−)
14	Big Gods	MSP	Religion, functional	(−)		
15	Social control	HS	Religion, conflict		(+)	
16	Warfare intensity	MilTech	Conflict—external	(+)	++	++
17	Military revolution	IronCav	Conflict—external	+++++	+++++	+++++

difference in AIC is a very strong statistical evidence against functionalist theories.

Crucially, the DR approach, based on the general model of cultural macroevolution (Eq. 1), allows us to distinguish correlation from causation by separating the influence of potential causal factors on the response variables rather than relying on "static" correlations, where the direction of causality remains ambiguous. For example,

when we plot the three response measures against the two major quantitative predictors (warfare and agriculture) at the same time, we observe variable degrees of synchronous correlation (Fig. 1). The correlation with Agri, in particular, is not strong (R^2 ranging from 0.24 to 0.32). However, when Agri is included in the DR model together with warfare intensity measures, we see a strong predictive effect of this variable on all response measures at the next time step.
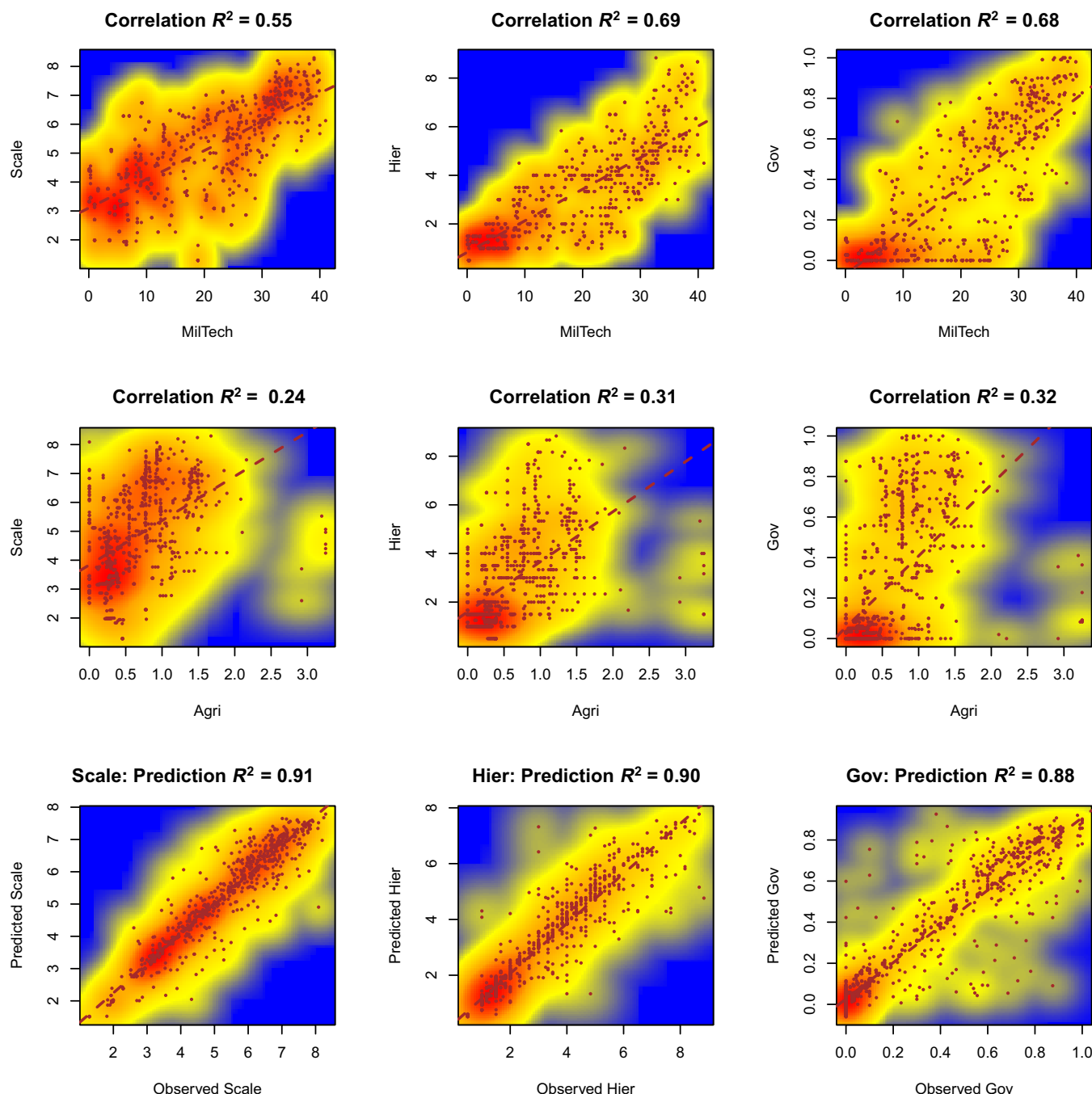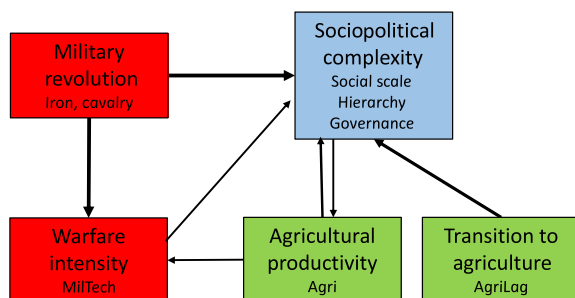

Fig. 1. Distinguishing between correlation and causation. (**Top**) Pairwise (synchronous) correlations between the three response variables (Scale, Hier, and Gov) and MilTech. (**Middle**) Pairwise (synchronous) correlations between the three response variables (Scale, Hier, and Gov) and Agri. (**Bottom**) Out-of-sample prediction accuracy of response variables estimated by k-fold cross-validation using DRs with time-lagged measures of agriculture and warfare (see Materials and Methods). Background colors indicate the density of data points, with red as the highest density. Dashed lines are linear regressions.

Coefficients of determination in the best-fitting models for all three responses are high (R^2 ranging from 0.89 to 0.92; see Supplementary Text), which includes the effects of nonlinear autocorrelation terms on the response variables (even excluding these autocorrelation terms, we find that the proportion of the explained variance remains high, with R^2 ranging from 0.58 to 0.71). Furthermore, k-fold cross-validation, which estimates the capacity of the macroevolutionary model for out-of-sample prediction (see Materials and Methods), yields similarly high prediction R^2 (Fig. 1), further supporting the interpretation that interpolity conflict and agricultural productivity are major causal drivers of social complexity across our diverse, global sample.

Our analysis indicates an unexpectedly simple web of causation between a few key variables (Fig. 2), considering that we tested >100,000 of combinations of 17 possible predictors. Most of the causal influences are unidirectional. For example, a separate analysis of the evolutionary causes responsible for the increase in our measure of warfare intensity (46) finds that none of the dimensions of social complexity are included in the best models with MilTech as the response variable. That analysis finds that MilTech is not affected by any polity characteristics such as territory or population size, governance or administrative complexity, monetary sophistication, and others; instead, its evolution is governed by major technological revolutions (particularly mounted warfare and iron metallurgy), overall world population, centrality of location with respect to the major communication routes within Afro-Eurasia, and, weakly, by agricultural productivity (46). MilTech, thus, acts as an exogenous variable with respect to social complexity. This reflects what we know about historical improvement and spread of military technologies, especially in the premodern era (individual weapons and armor could be produced as easily by stateless societies).

The case for unidirectional (rather than mutual) causality is even clearer with our second warfare proxy, IronCav, particularly its cavalry component. Mounted warfare was invented only once—by stateless people inhabiting the Pontic-Caspian steppes—and spread to the far ends of Afro-Eurasia and subsequently to all major world regions. Furthermore, agrarian empires, such as China, had to go to great lengths to secure plentiful supplies of horses needed for their cavalries. Cavalry, thus, is an exogenous measure of warfare intensity—a variable that excludes the possibility of reverse causality (from social complexity to intensity of warfare).

The case of agricultural productivity is different, because we find that not only Agri has a consistent positive effect on all three response measures but also Gov has a positive effect on Agri, although not a particularly strong one (see analysis in Supplementary Text). This is the only instance of possible mutual causation that our analysis detects.

Last, antiquity of agriculture is an exogenous variable because the adoption of agriculture typically precedes the appearance of large-scale societies by many centuries and, sometimes, millennia. The most common time lag from agriculture to large states is two millennia (Fig. 3). This worldwide analysis, based on regions defined by the ArchaeoGLOBE Project (55), is consistent with our finding that, based on 35 Seshat regions, AgriLag is a strong predictor of Scale, in particular. However, Fig. 3 also confirms that, while agriculture is a necessary condition for the rise of large-scale societies, it is not a sufficient one, because 23% of world regions, where agriculture was common before 500 BCE, failed to develop macrostates before 1500 (before the European expansion).

Macroevolution of social scale

Historians and archaeologists (56) have noted that sociocultural evolution during the Holocene was not a gradualistic process but rather involved phases of rapid change interspersed by long-term relatively stable periods, a pattern resembling "punctuated equilibrium" in biological macroevolution (57). The observed evolutionary dynamics of the maximum mean Scale (averaging the three largest Seshat polities extant at each century interval to reduce idiosyncratic fluctuations due to peculiarities of the largest one) illustrates this observation (Fig. 4A). A sustained increase during the third millennium BCE is followed by stagnation during the second millennium BCE. Another phase of rapid change during the first millennium BCE is
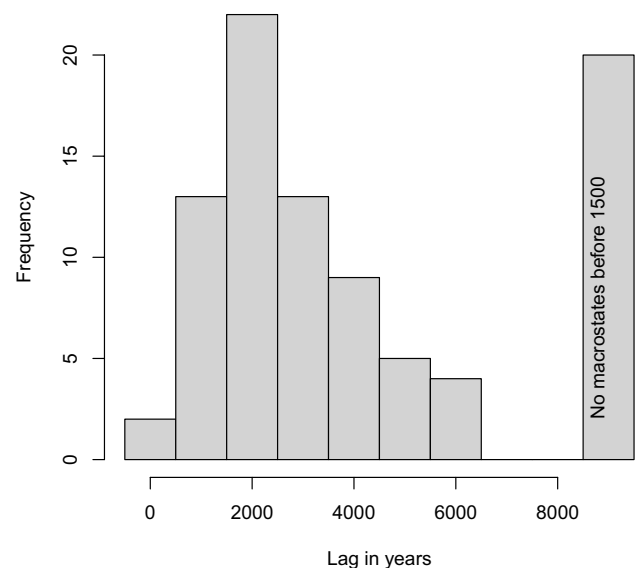

Fig. 2. Proposed web of causation affecting the evolution of sociopolitical complexity, indicated by our analysis. The thickness of arrows indicates the strength and consistency of the effect. The reciprocal causality arrow from the sociopolitical complexity to agricultural productivity is mediated by the Gov → Agri effect. Note that each arrow includes an explicit time dimension, that is, it has the form $X_t \rightarrow Y_{t+1}$.

Fig. 3. Time lags between the adoption of agriculture and the appearance of macrostates. Macrostates are states controlling territories of at least 100,000 km². The sample is based on 88 ArchaeoGLOBE regions, in which agriculture became common by 500 BCE. Only macrostates forming before 1500 (and, thus, before European colonization) are included in this analysis. Data sources are as follows: adoption of agriculture (55) and macrostates (63).
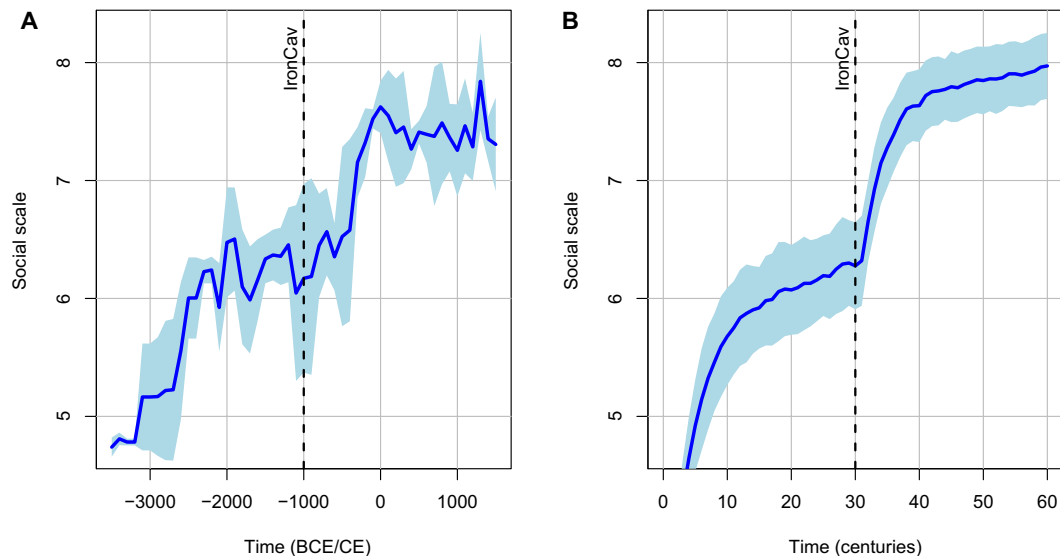
Fig. 4. A comparison between the observed and predicted evolution of largest polities. Scale integrates log-transformed polity population, territory, and the largest settlement; thus, a unit of change corresponds to 10-fold increase in untransformed quantities. (**A**) Observed macroevolution of the maximum Scale (averaging the three largest polities between 3500 BCE and 1500 CE). (**B**) Evolutionary dynamics of the maximum Scale predicted by Eq. 2, assuming that IronCav changes at $t = 30$. In both (A) and (B), solid curve denotes the mean and shading denotes the means ± SD.

followed by fluctuations around an apparent equilibrium during the next 1500 years.

Our statistical results suggest that the abrupt shift from stagnation in the second millennium BCE to rapid change in the following millennium is due to the introduction of iron and cavalry in Eurasia around 1000 BCE, followed by their rapid spread within the Imperial Belt of Afro-Eurasia. We can test this explanation quantitatively with the macroevolutionary model (Eq. 1). Using the terms that were selected in the best-supported model for Scale leads to the following equation for its dynamics

$$
\begin{aligned}
\text{Scale}_{t+1} = & -0.24 + 1.2\,\text{Scale}_t - 0.04\,\text{Scale}_t^2 + 0.10\,\text{Agri}_t + \\
& 0.00002\,\text{AgriLag}_t + 0.16\,\text{IronCav}_t + \varepsilon_t
\end{aligned}
\tag{2}
$$

where ε_t represents a stochastic error term. We focus on the effect of IronCav, the factor that has the largest effect on Scale (both the largest standardized regression coefficient and highest t value). During the first period of the simulation, we run the model for IronCav = 0 and then abruptly switch to IronCav = 2, representing the joint arrival of iron and cavalry in the region. We keep Agri at a constant value but allow AgriLag to increase (assuming that agriculture is adopted at time 0).

The empirically based model (Eq. 2) predicts that trajectories will tend to increase to an equilibrium, set by the values of predictors, and then fluctuate around it (see fig. S5). Thus, for periods when other drivers do not change, the model predicts a stabilizing behavior set by these predictors due to the quadratic form of Scale_t. For a direct comparison between the observed Scale dynamics and that predicted by Eq. 2, we sample its trajectories in the same way as with the data (see the Supplementary Materials for details). We observe that an abrupt change in IronCav results in a substantially higher equilibrium level, which induces a period of directional change followed by fluctuations around the new level (Fig. 4B). Although this pattern resembles a punctuated equilibrium, it is described more

aptly as stabilizing selection around equilibria set by the predictor variables. Also note that the equilibrium level continuously grows (but at a lower rate) as a result of increasing AgriLag.

DISCUSSION

As stated in Introduction, the core of this article is a large-n analysis using a global databank. Our results indicate that the general model of cultural macroevolution (Eq. 1) provides a productive analytic framework for statistical analysis of time-resolved historical data, such as have been gathered in Seshat. Although we tested 17 proxies suggested by five major classes of theories of social evolution (as well as >100,000 combinations), this analysis identified an unexpectedly simple web of causation (Fig. 2), in which the chief drivers of increasing social complexity and scale are agriculture and warfare. Variation between world regions in the timing of changes in predictor variables thus offers a series of natural experiments that not only make the dataset informative but can also serve as regional case studies to check the general results of the global analysis. Below, we sketch how these focused studies could be formulated.

The evolution of the largest territorial polities during the past 5000 years was characterized by a series of upsweeps, followed by periods of relative stability, or even decline (also see figs. S4 and S5) (*58*). For example, the spread of bronze metallurgy within Afro-Eurasia from c. 3000 BCE, associated with a sudden proliferation of hand-held weapons such as (bronze) swords (fig. S5A), resulted in the first appearance of macrostates (polities controlling a territory larger than 100,000 km²), Akkad in Mesopotamia and Old Kingdom in Egypt. The next military revolution was associated with the spread of chariot warfare, which required horses to pull them and powerful composite bows used by archers to shoot from these mobile platforms. The spread of chariot warfare also triggered the need to manufacture personal armor, which resulted in the proliferation of shields and helmets (fig. S5). Late Bronze Age polities, such as the Hittites in

Anatolia and the Shang in China were an order of magnitude larger than preceding ones, with one, New Kingdom Egypt, breaking through the megaempire threshold (1 million km^2). These empires, however, collapsed during the Crisis of the Late Bronze Age, resulting in a reduction of the maximal polity area (fig. S4).

The next military revolution resulted from the joint spread of horse riding and iron metallurgy. It also triggered another increase in the sophistication of armor, as indicated by the rapid spread of breastplates and limb protection pieces. The Cavalry Revolution led to the rise of very large empires, whose size topped 3 million km^2. In each of the major Eurasian subregions, these megaempires arose three or four centuries following the appearance of cavalry (table S6; the lags between iron and megaempire are much more variable, suggesting that horse is a better predictor of megaempire than iron). Such a temporal lag is consistent with the speed of change predicted by the empirically based macroevolutionary model (Fig. 4). Note that innovations in military technology resulted in more rapid evolutionary change (shorter lags), compared to the adoption of agriculture (Fig. 3).

Following the IronCav revolution, the maximum imperial territory fluctuated around the same level, 3 million km^2, for nearly two millennia. This dynamic equilibrium was broken by the Gunpowder Revolution, which resulted in yet another increase in maximum territory (43). The time lag between the appearance of effective gunpowder weapons and the rise of European colonial empires was also 300 to 400 years (58).

We focus here on Afro-Eurasia, because that is where the largest territorial states were located until more recently in world history. In contrast, the North American continent did not develop an indigenous megaempire comparable to Rome before 1500, although Central Mexico acquired agriculture at approximately the same time as Southern Europe (c. 6000 BCE). However, on a smaller scale, there are remarkable parallels. The rise of the Aztec Empire resulted in a substantial upsweep of social scale and complexity in the Basin of Mexico (fig. S2D). For example, the population of Tenochtitlan was between 150,000 and 300,000 (for comparison, the population of Naples, the largest city in the Habsburg Empire, which conquered Mexico, was approximately the same, 224,000). The Aztec upsweep was preceded by a number of military innovations, including bows and arrows, which arrived in the region from the North c. 1100 and stone-bladed broadswords (59). Together with already available thrusting spears and sophisticated armor, these military technologies were highly effective. Nevertheless, the Mexica controlled a tiny territory, by the standards of Eurasian empires (less than 30,000 km^2). The main impediment was a severe limitation on their logistics arising from the need to move troops and supplies by foot (59, 60). It is probably not a coincidence that the only megaempire in the Americas, the Incas, was in the area where domesticated transport animals (llama) were available. More generally, domestication of llama in the Andes, the use of atlatls in Mexico and elsewhere in the Americas, and the spread of the "Asian War Complex," which included the backed and recurved bow, armor, wrist guards, and other features, through North America starting 700 CE (61) are examples of other military "mini revolutions" that took place outside Eurasia and are captured in the MilTech variable.

The introduction of the horse to North America by the Spaniards in the 16th century provides us with yet another natural experiment (43). The spread of horses and horse riding from Mexico into the Great Plains resulted in a sociocultural evolution there with notable parallels to steppe confederations of the Old World. The most powerful nomadic confederation was the Comanche "Empire" (62). During the 18th century, the Comanches became a hegemonic power, controlling the entire southern Great Plains. Their raids reached deep into Mexico and Texas. However, the arrival of cavalry in the Americas occurred late, and the effects of the Cavalry Revolution were soon overtaken by the gunpowder revolution. During the 19th century, the Comanches (and other Native American polities in the Plains) were overrun by the steamroller of the United States, resulting in the rise of the most powerful modern megaempire.

Another notable example of the influence of military technology on the evolution of large-scale societies is Hawaii. Before the arrival of the Cook Expedition in 1778, Hawaiian Islands were ruled by four or more chiefdoms, focused on the main islands of Kaua'i, O'ahu, Maui, and Hawai'i. The Big Island (Hawai'i) alternated several times between being united under one ruler and split into two or more smaller polities. Very soon after Western arms became available, the ruler of one polity, Kamehameha I, used them to unify the entire archipelago within his kingdom (15).

Here, we lack space to continue this survey, instead referring readers to another publication (43). However, it is clear that detailed case studies focusing on specific historical societies undergoing evolutionary transitions are a key complement to large-n analyses.

An agenda for future research

Here, we have proposed a general approach for studying historical processes that combines the use of nonlinear dynamical systems, large-scale historical datasets, and a systematic statistical testing of alternative causal hypotheses. Our approach has allowed us to compare quantitatively all major theoretical approaches to the evolution of human social complexity within a single framework and to lay the ground for more nuanced and precise theories to be rigorously tested in the future.

Our analysis confirms that increasing agricultural productivity is necessary but not sufficient to explain the growth in social complexity. Furthermore, analysis indicates that this increase was not driven by factors associated with either functionalist or internal conflict theories. Instead, external (interpolity) conflict and key technical innovations associated with increasing warfare intensity appear to be the primary drivers of state growth, along with the growing population and resource base provided by increasing agricultural productivity. Our analyses help clarify why a mechanistic model that privileges warfare and military revolutions (63) and agriculture (64) has offered compelling, if provisional, interpretations for what drove the rise, spread, and equilibrium levels of social complexity in Afro-Eurasia in the ancient and medieval periods, as well as worldwide during the early modern period. Although factors such as infrastructure provision, market and monetary exchange, and ideological developments do not appear to play a significant causal role in propelling subsequent advances in social scale, hierarchical complexity, or governance sophistication, they likely are integral elements that support and maintain the results of that growth, which would account for the relationship observed between these factors in previous scholarship.

We expect that future analyses, using additional time-resolved data and advanced analytic methods, will help clarify whether and to what extent these different factors are critical for reinforcing or stabilizing states at different levels of complexity and modes of organization. We reiterate that the present study is limited to a sample

of 35 world regions during the Holocene. Given the current interest in Mesolithic fishing societies that have achieved large-scale, hierarchical social formations in areas not currently covered by the Seshat database, it would be desirable to add key regions in which these societies flourished (such as the West and Northwest Coasts of North America, Peru, Chulmun Korea, and the Baltic). More generally, we call for a much more thorough sampling of Africa, North America, and South America (the Seshat project is already expanding our coverage of Subsaharan Africa). It is possible that future efforts to expand the geographical scope and temporal depth of the Seshat database, as well as filling gaps in the regional histories already covered, will alter the results reported here. The Seshat database is continuously evolving as it grows and as additional evidence comes to light. Nevertheless, this study showcases an ambitious new approach to quantifying global history over thousands of years, allowing us to test theories of cultural macroevolution more comprehensively than ever before, thereby unleashing the explanatory power of history as part of a broader scientific framework.

MATERIALS AND METHODS

For a brief introduction to the Seshat Project and the detailed explanation of how historical information is coded into the Seshat Databank, see Supplementary Methods.

Defining the response variables

Social complexity is a characteristic that has proven difficult to conceptualize and quantify (*1*). It is clear that social complexity has many dimensions or manifestations (*65*). While several researchers proposed synthetic, integrative measures that capture multiple dimensions of social complexity (*2*, *39*, *66*), a more common approach has been to use a single proxy measure, such as the population size of the largest settlement (*2*), the number of decision-making levels (*67*), the number of levels of settlement hierarchy (*68*), or the extent of controlled territory (*63*). Others have criticized these approaches on the grounds that these proposed measures focus too much on size and hierarchy (*69*).

Similar definitional problems bedevil the study of the "state." Some anthropologists focus on social scale and simply define the state as a regionally organized society with a population of hundreds of thousands or more (*5*, *70*). Another approach privileges political centralization. The state, then, is a polity with three levels of administration above the local community, whereas simple and complex chiefdoms are characterized by one and two levels above the local community [this is the approach taken by cross-cultural ethnographic databases (*71*, *72*)]. Such an approach, however, fails to distinguish between early states and super-complex chiefdoms (such as nomadic imperial confederations in the Great Eurasian Steppe) that may have three or more levels of organization.

The third approach is to define the state as a politically centralized territorial polity with internally specialized administrative organization (*73*). The emphasis on internal specialization of administration arose as a result of the desire by archaeologists and political anthropologists to distinguish between chiefdoms and states: "A chiefdom can be recognized as a cultural development whose central decision-making activity is differentiated from, although it ultimately regulates, decision-making regarding local production and local social processes; but it is not itself internally differentiated. It is thus externally but not internally specialized" (*73*).

These approaches differ from the traditional definition of the state in historical sociology, going back to Weber (*74*), according to which it is a polity that maintains a monopoly on the legitimate use of violence. However, there are two problems with defining the state on the basis of legitimate use of violence. First, premodern states often did not concern themselves with such monopoly or, if they did, were quite inefficient at maintaining it (*75*). Second, for preliterate societies without records, it is usually impossible to determine whether the polity or ruler claimed a monopoly on legitimate violence. In other words, using the Weber definition, we would be limited to studying only very modern states.

Here, we adopt an inclusive approach that allows us to determine how different dimensions of social complexity have evolved. Accordingly, we aggregate 17 Seshat variables into three integrated measures: social scale, hierarchical complexity, and internal specialization of governance (see below for details).

Another definitional note is that we use polity, defined as an independent political unit, as a general term for a variety of political organizations, ranging from autonomous villages (local communities) through simple and complex chiefdoms to states and empires. Thus, we do not impose a hard distinction on which polities are states and which are not. Instead, the response variables define a three-dimensional phase space in which polities reside and evolve. Whether there are concentrations or other kinds of structure in this space becomes an empirical question. We now discuss how the three response variables are defined in Seshat. Short names of variables are given in parentheses following the long names.

Social scale (Scale)

This synthetic variable combines the effects of
1) Polity population
2) Polity territory
3) The population of the largest settlement

We log-transform (base 10) these three constituent variables and submit them to the principal components analysis. Scale is the PC1. For ease of interpretation, we scale PC1 to the same range as polity population. Thus, Scale = 3, for example, corresponds to polities with a population of around 1000, and Scale = 6 corresponds to polity population of 1 million.

Hierarchical complexity (Hier)

We measure hierarchical complexity by averaging the number of levels in
1) Military hierarchy
2) Administrative hierarchy
3) Settlement hierarchy

(the last one is a particularly useful measure for archaeologically known societies). This measure includes the lowest level (e.g., private soldier or lowest clerk); thus, the minimum value of 1 corresponds to nonhierarchical societies. The Seshat Databank also has an additional variable, the number of levels in religious hierarchy, but analysis indicates that although military, administrative, and settlement levels are tightly correlated, the relationship between the combined measure and religious levels is much more variable, suggesting that religious levels are indicator of a different polity characteristic.

Specialization of governance (Gov)

Specialization of governance is an example of a complex Seshat variable, which is broken down into simpler components for the data collection stage. All component variables not only are binary and code each characteristic as present or absent but also allow us to reflect uncertainty about these estimates and disagreement between sources.

After data are gathered for each component, in preparation for analysis, we assemble binary variables into a quantitative measure that is more suitable for statistical analysis (4).

The Seshat Databank captures different dimensions of internal specialization of governance with 11 variables. The first four variables code for the presence/absence of professional military officers, soldiers, religious specialists, and administrative specialists (bureaucrats). For example, we code "full-time bureaucrats" as absent if administrative duties are performed by generalists, such as chiefs and subchiefs. We also code it absent if state officials perform multiple functions, e.g., combining administrative tasks with military or priestly duties.

The next two variables code for bureaucracy characteristics: presence/absence of an examination system and of merit promotion. These two variables can be coded present only if full-time bureaucrats are present. In addition, "present" codes require evidence of formal and institutionalized examination and merit systems.

The next variable, specialized government buildings, is particularly useful for societies known only from their archaeological record. These buildings are where administrative officials are located and must be distinct from the ruler's palace. They may be used for document storage, registration offices, a treasury, and so on. Defense structures (walls and towers) or state-owned/operated workshops are excluded.

The final four variables code for the characteristics of the legal system: formal legal code, professional judges, professional advocates, and specialized buildings used for legal purposes (courts).

The 11 variables on which Gov is based are the following:
1) Professional officers
2) Professional soldiers
3) Professional priests
4) Full-time bureaucrats
5) Specialized buildings used for government
6) Examination system
7) Merit promotion
8) Formal legal code
9) Full-time judges
10) Professional lawyers
11) Courts (specialized buildings used for administering justice)
Gov is constructed by summing together the 11 codes and scaling it from 0 to 1.

Outlining hypotheses and defining predictor variables
In this section, we discuss all hypotheses that we test in the statistical analysis. For all hypotheses, we also identify predictor variables proxying the hypothesized evolutionary mechanism. As before, the variable short name is provided in the parenthesis.
Agriculture hypotheses (1 and 2)
Plant and animal domestication was "the most momentous change in Holocene human history … because it provides most of our food today, it was prerequisite to the rise of civilization, and it transformed global demography" (10). The view that agriculture was the necessary condition for the evolution of complex societies crystallized in the work of early anthropologists Childe (7), White (8), and Service (9). It is implicitly, and often explicitly, held by most scholars of the past today. In a more extreme view, food production was not only a necessary but also a sufficient cause [e.g., (10)]. According to this view, widespread practice of agriculture resulted in sedentarization, storable food surpluses, and human population explosion. Surpluses resulting from agriculture could be used to feed full-time craftspeople

and inventors, which drove cumulative growth of technology. Surpluses also made possible the appearance of rulers and elites, full-time bureaucrats, military officers, and soldiers, thus resulting in social stratification, political centralization, and interstate warfare. In this view, therefore, transition to agriculture inevitably leads to the rise of large-scale complex societies. However, the time lag between these two processes can be quite substantial.

A time lag between the switch to productive economies and the appearance of centralized polities (chiefdoms) or states is also postulated by theories in which agriculture is a necessary but not sufficient condition. Although the rate of cultural evolution is generally faster than biological evolution (76), the development of social norms and institutions for collective action is not straightforward and may require long periods of cultural experimentation (77). Furthermore, norms and institutions may need to build on preceding innovations and, thus, accumulate over generations (78). Differences in the time that has been available to societies to develop the institutions, which underpin stable large-scale organization, may therefore play an important role in explaining the distribution of these societies (79).

Hypothesis 1: Productivity of agriculture (Agri). Although at the most basic level agriculture can be conceptualized as a binary variable— absent before the Neolithic Revolution and present after it—this approach works well only in those cases when a population with fully developed agricultural technologies colonizes new lands, where agriculture was previously absent. In most other cases, agricultural technologies developed gradually, passing through multiple phases. Thus, cultivation typically precedes domestication (that is, genetic and morphological changes of the cultivated crops or herded animals), but there is no fixed time period that needs to elapse between these events. Furthermore, what constitutes "domestication" varies from species to species. For vegetatively propagated crops, such as root crops (potato, sweet potato, taro, and yams), banana, and sugarcane, the domestication syndrome is only beginning to be defined (80).

The Seshat project has developed a sophisticated approach to estimating how the productivity of agriculture has evolved in each of the Seshat NGAs on which the Seshat sample of past polities is based (81). The approach that we used to obtain these estimates quantitatively combined the influences of production technologies (and how they change with time), climate change, and effects of artificial selection into a relative yield coefficient, indicating how agricultural productivity changed over time in each NGA between the Neolithic and the 20th century. We then use estimates of historical yield in each NGA to translate the relative yield coefficient into an estimated yield (tons per hectare per year) trajectory. We tested the proposed methodology with independent data and concluded that while much more work is needed to refine this approach, it provides reasonable approximation of agricultural productivities in world history (81).

Hypothesis 2: Antiquity of agriculture (AgriLag). As discussed above, several theories postulate a time lag between the switch to productive economies and the appearance of chiefdoms and states. We use the data from the ArchaeoGLOBE Project (55), which coded all world regions for several variables. We focus on the date when extensive agriculture becomes common in a region encompassing a Seshat NGA. AgriLag is then calculated as the time difference between this region-specific date and the date associated with the response variable.

Functional hypotheses (3 to 8)
The next set of hypotheses focuses on integrative or managerial theories of the state.

Hypothesis 3: Provision of public goods (Infra). Functionalist theories explain the rise of the state as a (at least partial) solution to the various challenges and problems facing societies. In particular, it is well known that provision of public goods is a very problematic issue for societies due to the free-rider problem (*82*). According to these managerial theories, the state is needed to solve the coordination and cooperation challenges needed to build and maintain costly infrastructure, such as roads, bridges, and postal stations; to buffer the population against famine by building food storage facilities; and to provide other useful goods such as public markets and drinking fountains. Accordingly, we use Infra, which aggregates 12 binary Seshat variables: irrigation systems, drinking water supply systems, food storage sites, markets, roads, bridges, canals, ports, markets, postal courier service, postal stations, and a general postal service.

Hypothesis 4: Hydraulic society (Irrigation). The variable Infra that we use to proxy a variety of public goods that need to be supplied by the polity may inappropriately combine different kinds of function, suggesting that we should also investigate possible effects of components separately. Of particular interest is the variable Irrigation that serves as a proxy for the theory of oriental despotism by Wittfogel (*16*). Other binary variables constituting the overall measure of Infra will also be investigated. Because Irrigation is a component of Hier, we do not use both predictors in the same regression (see Supplementary Results).

Hypothesis 5: Urbanization (Cap). The administrative demands of increasingly sophisticated and diverse economic activities may have played a significant role in propelling the evolution of the state (*5*). Economic historians have proposed a number of proxies that can be used to trace economic development in the long run. One popular indicator is the degree of urbanization. The logic is that cities, especially large cities, had to rely on imported food to exist. This means that the level of agricultural productivity had to be high enough to support city dwellers who did not produce their own food. Transportation needed to be efficient enough to get the food and other raw materials produced in the countryside to the cities. In other words, only well-developed (and reasonably well-managed) economies make large cities possible. In addition, a substantial proportion of urban population is usually involved in various crafts and trades, thus providing another index of economic sophistication. We proxy this explanation with log-transformed population of the largest settlement, Cap.

Hypothesis 6: Trade (Market). Marketplaces increase the efficiency of trade but often require an overarching authority for preventing crime (theft) and resolving disputes (*83*). Solving the coordination and cooperation challenges needed for efficient management of trade may be another important function of the state. For this reason, we add another proxy, a binary Seshat variable that codes for the presence or absence of markets in the focal polity (Market). Note that the variable Market (similarly to Irrigation) is also used in a more synthetic measure Infra.

Hypothesis 7: Economic exchange (Money). A related hypothesis focuses on the degree of economic exchange itself and the means used to facilitate or regulate exchange (*84*). Scholars argue that increasing the scope and ease of exchange has two main benefits: More robust exchange supports larger populations and more extensive territories by facilitating the distribution of food and other goods, thus supporting rising social scale and administration as explained above, and it is easier for polities to control and derive revenue from the movement of goods (especially when exchange is monetized) than from agrarian production alone, which generally leads to larger, more spe-

cialized fiscal administration and allows for a greater scope of state activity. We know that many different methods were used to support economic exchange over time and in different parts of the globe: cowries, ingots of metal, coins, paper bills, credit cards, and so on. Some means of exchange are more efficient than others, and as a result, when they appear, they tend to replace the less-efficient ones. For example, the most important unit of wealth in early Rome was cattle (pecunia means cattle). Later on, Romans used bronze ingots and then coins. Thus, data on how the means of exchange changed over time provide another reasonable proxy for the sophistication of economy. As the proxy for this hypothesis, we use Money, which combines information from six binary Seshat variables. The Money scale reflects the "most sophisticated" monetary instrument present in the coded society (0, none; 1, articles; 2, tokens; 3, precious metals; 4, foreign coins; 5, indigenous coins; 6, paper currency).

Hypothesis 8: Information system (Info). Some managerial theories postulate that sophisticated institutions of governance evolved under the pressure of the need to manage information flows within a polity (*5*). In addition, governance institutions, such as bureaucracy, may require the prior appearance of writing and may further require the presence of an intellectually sophisticated, literate segment of the population, from whom bureaucrats can be recruited. We proxy this hypothesis with a measure of information complexity (Info) that combines data from 13 binary Seshat variables. The first four provide the basis for measuring the sophistication of the writing system (mnemonic devices, non-written records, script, and written records), while the additional nine variables code for the presence or absence of various kinds of texts (lists, calendar, sacred texts, religious literature, practical literature, history, philosophy, scientific literature, and fiction). For details, see (*50*).

Social scale hypotheses (9 and 10)

The next set of hypotheses focuses on the effect of various aspects of the social scale.

Hypothesis 9: Scalar stress (Pop). A number of theories propose some aspect of social scale as the driving force behind state formation. Perhaps the most obvious aspect of scale is the total population of a polity. Although humans have evolved a remarkable capacity to cooperate in large groups, compared to other mammals, once the size of the group exceeds a few hundreds, the ability to coordinate its activities by means of face-to-face interactions sharply diminishes (*33*, *34*). Political centralization is one possible, culturally evolved mechanism that could allow human societies to break through the limit imposed by face-to-face sociality (*47*). However, centralization alone is not enough. Once the earliest centralized societies—chiefdoms—exceeded a certain population threshold, they required internal specialization to continue functioning in a reasonably efficient manner. Societies that failed to develop complex division of labor could not pass this population threshold and either remained small or succumbed to better organized ones. The term "scalar stress" was proposed by Johnson (*32*). The proxy variable to test this hypothesis is the log-transformed (base 10) polity population (Pop).

Hypothesis 10: Territorial expansion (Terr). Spencer (*35*) proposed that a major factor behind the rise of primary states in the half-dozen areas where they first developed was not the total population, per se, but rather a significant aggressive expansion of the territory that needed to be governed, which required the delegation of specialized parcels of military and civil authority to distant conquered/subjugated regions, allowing for effective management from the political center of the nascent state. This proposal gains some empirical support

from the archaeological records of these primary states (*35*). However, a focus on just primary/pristine states provides us with a very small sample for testing the broader applications of this theory. In the present analysis, we use a large sample of historic polities of varying complexity not to test the hypothesized relationship between territorial expansion and primary state formation but rather to test whether this pattern is/can be generalized to more developed/complex or "descendant" states. We proxy this hypothesis with log-transformed polity territory (Terr).

Internal conflict hypotheses (11 to 13)

Hypothesis 11: Social stratification (Class). Inequality plays an important explanatory role in theories that emphasize internal conflict between social classes (*14*). Drennan (*17*) proposes the following sequence of events that transforms autonomous villages into chiefdoms:

> *This … sequence begins with the emergence of patterns of economic inequality in a small autonomous village. Such patterns of differing wealth would tend to concentrate population in that village as those of greater wealth take advantage of the opportunities their wealth provides to make others dependent upon them. Such concentration of dependents would be encouraged by the wealthy since it provides enhanced opportunities for still further acquisition of wealth. This process would eventually involve the incorporation of existing small neighboring villages into the system or the founding of additional small villages by people from the emergent center so as to increase the resource base for wealth accumulation.*

We proxy this hypothesis with social stratification (Class), which was defined by Murdock and Provost (*66*) and coded by Peregrine (*39*) for archaeological societies. This variable takes three values: egalitarian (no classes), two classes, and three or more classes.

Hypothesis 12: Iron law of oligarchy (Hier). Another approach to testing theories that propose internal conflict as a major driver for the evolution of the state is a measure that focuses on the length of chains of command. The idea behind this is that the more levels of control and command there are in a hierarchy, the more power accrues to the individuals occupying the top levels. Inevitably, those at the top of these hierarchies will be tempted to use it to their personal advantage. This dynamic is sometimes referred to as "the iron law of oligarchy" (*40*). To preserve and protect their high status, power, and wealth, the individuals at the top of the hierarchy (the elites) should favor the rise of the hierarchically organized state, including its coercive apparatus (e.g., professional military and police, courts and judges, etc.) and ideological machinery (e.g., professional priests). We will use the measure of hierarchical complexity (Hier), discussed above. Note that Hier is also one of the response variables (see above). Given the time-resolved nature of Seshat data and the statistical analysis that capitalizes on this feature (dynamical regression; see Statistical analysis: Dynamic regression in Materials and Methods), it is appropriate to include a variable in the analysis as both the response and the predictor. We use time lags to break endogeneity, so that variable X_t enters the model as the response and $X_{t-\tau}$ as the predictor (see Statistical analysis: Conceptual overview in Materials and Methods). This feature of our approach enables us to resolve the cases of mutual causality (when factor X influences the evolution of Y and Y influences change in X).

Hypothesis 13: Cereal crops (Grain). Several theories postulate that grain is more storable than root crops, making it easier to appropriate by emerging elites (*12*, *85*). The binary Grain variable is coded as 0 if the main carbohydrate sources is a root crop (yam, sweet potato, and taro) and 1 if it is a cereal (wheat, rice, maize, millet, and rye).

Religion hypotheses (14 and 15)

Hypothesis 14: Big Gods (MSP). Religious constructs relating to supernatural agency, the afterlife, and ritual efficacy have been documented across the ethnographic record and likely have deep roots in our species' evolutionary history (*86*, *87*). By contrast, moralizing religions, in which moral behavior is at the center of religious life, appear to be a much more recent cultural innovation (*37*, *88*, *89*). The question why moralizing religions have grown over time, becoming the predominant form of religious practice around the world today, has a long history (*90*). An influential current in the evolutionary theorizing of religion proposes that belief in all-knowing, morally concerned, punitive deities—Big Gods—facilitated increases in social complexity (*36*, *37*). One formulation of the Big Gods theory (*37*) begins with the premise that religious beliefs and behaviors originated as an evolutionary by-product of ordinary cognitive tendencies, such as mind-body dualism (*91*) or teleological reasoning (*92*). These intuitive biases were exploited by culturally evolved beliefs in supernatural surveillance and punishment because these beliefs increased the ability of groups to sustain complex social organizations and successfully scale up and expand. Competition among cultural groups gradually aggregated these elements into cultural packages, in the form of organized religions. Thus, Big Gods coevolved with larger and more complex societies (*37*). A variant of the Big Gods theory proposes that "broad supernatural punishment" (including nonagentic forces such as karma) contributed to the rise of sociopolitical complexity (*89*, *93*). We proxy this hypothesis with the synthetic variable MSP, which aggregates seven Seshat binary variables coding for religious characteristics (*38*).

Hypothesis 15: Social control (HS). The social control hypothesis proposes that ritual HS bolsters the power of elites by legitimating their authority (*41*), due to the ritual and ideological significance of HS as a means of communicating with (or appeasing) supernatural beings, and motivating compliance via intimidation (*42*). This hypothesis therefore predicts a positive relationship between HS and social complexity. Support for this hypothesis comes from an analysis of data on 93 traditional Austronesian cultures (*41*). We use the binary Seshat variable HS as a proxy for this hypothesis.

External conflict hypotheses (16 and 17)

Hypothesis 16: Warfare intensity (MilTech). Several conflict theories propose interpolity competition (which includes, but is not limited to, warfare) as the main evolutionary driver for the rise of the state. Probably the most notable example of this mechanism in action is the rise of the modern European state. Military historians and historical sociologists have argued that the military revolution in early modern Europe transformed the scale of war and led to an increase in the authority of the state (*20*). As Tilly (*84*) famously stated, "War made the state and the state made war." Cultural multilevel selection (CMLS) theory has generalized this explanation beyond its focus on the post-1500 period. According to this theory, competition between societies, usually taking the form of warfare or at least the threat of annihilation from war, imposes a selection regime that weeds out dysfunctional, poorly organized, and internally uncooperative polities, favoring those with larger populations and effective centralized and internally specialized institutions—states (*6*, *43–45*).

These theoretical considerations propose the following causal sequence: Development of new military technologies makes armed conflict deadlier and, thus, pose a greater existential risk to the societies involved, which makes the selection pressures imposed by potential combatants more intense and that, in turn, leads to the rise

of increasingly better-organized centralized societies. The CMLS theory is global in scope and potentially applies to all world regions and time periods in the Holocene (roughly the past 10,000 years). In addition to the early-modern military revolution resulting from the development of gunpowder weapons, other similar revolutions include those resulting from the spread of cavalry and iron weapons/armor, as well as chariots and bronze weapons/armor before that.

The main proxy for the CMLS hypothesis is the Seshat measure of the realized sophistication and variety of military technologies used by polities, MilTech. This measure aggregates 46 Seshat binary variables coding for the presence or absence of various types weapons, armor, projectiles, and defensive structures; the use of metals for making weapons and armor; and of transport animals used for military logistics (46). The adjective "realized" refers to our approach in constructing these 46 variables that assigns 1 when there is evidence that a particular weapon, projectile, etc. was used by the coded society and 0 when such evidence is absent. The reason for this "strong evidence" rule is that our focus is not on whether a technology was known but whether it was used. A large variety of sophisticated means of attack and defense, thus, serves as a quantitative proxy for the intensity of warfare in the environment of the polity. People tend to invest in expensive defenses when their societies are threatened by their neighbors. However, we should also note that there is a feedback loop from advances in military technologies and the intensity of warfare. In particular, our analysis of the evolution of MilTech in the Seshat sample indicated that these new technologies as horse riding and iron metallurgy result in strong advances of other aspects of military technologies (46). Thus, we add a supplementary hypothesis, Cavalry/Iron Revolution, as a check for the CMLS hypothesis (see below).

Ideally, we would measure warfare intensity directly using the information about harmful consequences of warfare for individuals, groups, and polities. The Seshat project developed 17 such "severity of warfare" indicators and invested large effort into coding them for the Seshat sample. However, we found that finding information about these aspects of the past proved to be exceedingly difficult and, in many instances, impossible. As a result, these 17 variables have a high proportion of missing values. As a result, they are not suitable for use in general analyses of the evolution of social scale and complexity; however, we can use them as a check of how well our main variable, MilTech, captures the intensity of conflict in a set of polities for which we were able to code warfare severity. This analysis (see the "MilTech and War Severity" section in the Supplementary Materials) indicates that MilTech is a good proxy for war severity, because the relationship between the two variables is approximately linear and the degree of correlation is high (over 0.9 for well-coded polities).

Details of the Seshat variables that feed into MilTech and how they are aggregated are in (46). The principal components analysis shows that the six aggregated measures are closely correlated with each other, and the PC1 captures 75% of variance. Thus, inevitable errors of coding affecting any of the specific variables tend to be compensated by the information contained in other variables. Thus, basing this proxy on 46 variables coming in six different classes builds in redundancy and, thus, increases the robustness of this measure.

Analysis in (46) also shows that the evolution of MilTech is affected by regional and global factors, rather than by the characteristics of the polity. This will be important for building the causal web of interactions.

Hypothesis 17: Military revolution (IronCav). According to the Cavalry Revolution theory, the invention of effective horse riding in the Pontic-Caspian steppes, combined with powerful recurved bows and iron-tipped arrows, had several consequences. First, it elevated the intensity of warfare as it spread from the steppes south to the belt of farming societies (43). Second, it triggered a process of military innovation, because the threat of nomadic warriors armed with this advanced (for the period) military technology spurred the development of countermeasures designed to mitigate the cavalry advantage, while also producing an incentive to adapt cavalry in areas further and further away from the location of their initial invention along the Steppe. The history of the military use of the horse went through several stages: the use of the chariot, the development of riding, the formation of light auxiliary cavalry, the development of nomadic riding, the appearance of the hard saddle, armored cataphracts, stirrups, and, lastly, heavy cavalry—the main branch of troops across Afro-Eurasian societies between c. 550 and 1400 CE (94). As a result, effective horse riding had far-reaching consequences for the evolution of military technologies, and specifically armor, projectiles such as crossbows, and fortifications.

Invention of iron metallurgy had a similarly widespread effect. Multiple authors (95, 96) have suggested that the availability of iron had a huge impact on the evolution of military technologies, because this strong and malleable material served as an input for a host of important technologies, military and otherwise, throughout the period under investigation here. Iron metallurgy and horse riding together worked synergistically, as iron arrowheads had greater penetrating power than stone and were cheaper to produce than bronze. Furthermore, iron played an important role in later cavalry-related developments such as the evolution of the saber and heavily armored cataphracts and knights (94).

We use the data from (48) for the Cavalry variable and data from (49) for the Iron variable. For the maps of spread of these two technologies, see (46). A potentially confounding factor is that these two variables, Cavalry and Iron, are highly correlated, and it may be difficult to estimate their effects separately (this is known as the problem of collinearity). To address this potential issue, we created a synthetic variable, IronCav, which combines the two effects (by adding Cavalry and Iron together). IronCav, thus, takes the maximum value for societies with both mounted warfare and iron weapons, intermediate value for societies having one characteristic and not the other, and minimum value for societies with neither characteristic. We explored with DRs whether IronCav turns out to be a better predictor than either of its constituent variables, reported below.

Statistical analysis
Conceptual overview
As noted in Introduction, our statistical approach to causation is based on DR (described in detail in the "Dynamic regression" section) rather than DAGs. Because the DAG approach recently gained much popularity, here, we explain how the goals and the logic of the DR approach are different. Note that the DR framework is based on the ideas of Wiener (27), which were later developed by Granger (28). This approach has also been used in the statistical analysis of animal population dynamics (97).

The most important difference between DR and DAG is that the latter does not explicitly include the time dimension. Thus, instead of causal links in DR, such as $X_t \rightarrow Y_{t+1}$, in DAG, causal connections are denoted without time subscripts, as $X \rightarrow Y$. Because of this difference, DAGs have to be acyclic. In other words, scenarios of mutual causation cannot be investigated. Furthermore,

the main goal in DAG is estimation of the causal effect. This approach is appropriate if we need to know, for example, by how many years a particular drug would increase life expectancy when we only have observational data. To answer this question, an analyst must assume a particular DAG—its form is underdetermined by (time-unresolved) data.

The goal of the DR approach is different, because we aim to use data to adjudicate between different theories of social evolution, each proposing a different causal graph (an example is in Fig. 2). This is generally impossible to do with static (time-unresolved) data, which is why the goal of the Seshat project from its inception was to collect time series data. Unlike with DAGs, the main goal of the DR approach is model selection, choosing which predictor terms should be included on the right hand sides in Eq. 1. We are also interested in estimation, because we want to compare the numerical strengths of different factors, but this goal is secondary to model selection, as we first need to determine which causal graph should be used for coefficient estimation. Thus, the DAG approach, excellent as it is, differs in goals and technics from the DR approach. The DR approach was designed to resolve questions of causation in evolutionary processes (descent with modification) that unfold slowly in time. It allows us to deal with these complications as mutual causation loops and temporal autocorrelations arising from the inertial nature of evolution, as well as (with fairly straightforward extensions, see the "Dynamic regression" section below) with spatial diffusion and phylogenetic effects. A model's ability to predict data is interesting not in itself but as a tool for adjudicating between different theories. A more precise term for this approach is retrodiction, because even when we use out-of-sample prediction, it is about the past, not the future (see the "k-fold cross-validation" section).

Time-resolved data are what enables tests of evolutionary theories against each other, but it does not solve all possible problems in the analysis of evolutionary causation. For this reason, the specific results reported in this article are tentative and contingent on additional data and improved analytic approaches. One recurrent problem with historical data is the gaps in the knowledge of historians and archaeologists, resulting in missing data. We dealt with this problem, as well as uncertainty in estimates and expert disagreements, by multiple imputation (see the "Multiple imputation and non-parametric bootstrap" section). While this is a valid statistical technique, additional research by expert scholars aiming to fill the gaps in the database would be a much more satisfying long-term solution. One of the goals of the Seshat project has been highlighting gaps in our knowledge to motivate such research.

Another fundamental difficulty is the "hidden variable" problem or omitted variable bias (98). This happens when our analysis implicates X as a causal factor for Y, while, in reality, the true cause is a variable not included in the analysis, Z, with which X is closely correlated. The only solution for this problem is gathering data on as many potential predictors as possible. This is why the Seshat project defined, and gathered data on, proxies for all major classes of theories that have been proposed by social scientists so far. However, we acknowledge that the set of proxies used in this article is just the beginning. This is where we see the most fruitful area for future research: defining additional proxies for theoretically postulated mechanisms, gathering data on these variables, and rerunning analyses to find whether new variables turn out to be better predictors of social complexity and scale than the ones we have tested so far.

A related potential problem is that of an "uninformative dataset," which happens when there is not enough variation in either potential predictors or response variables (or both). This is why we need data that sample as many different evolutionary trajectories as possible. To illustrate this point, consider the effect of cavalry on the evolution of social complexity. If our dataset only contained Eurasian societies after horse riding spread everywhere, then we would not have enough variation to statistically detect the effect of this driver. The inclusion in the analysis of Eurasian societies before 1000 BCE and, most crucially, New World societies where cavalry arrived very late is key. In essence, the spread of horse-based warfare, which happened at very different times in different parts of the world, is a "natural experiment" that allows us to estimate its effect on the evolution of social complexity. However, because cavalry and iron metallurgy are so closely correlated in our dataset, we were unable to disentangle the effects of these two technologies. We need additional information to do so [see (58)]. A general conclusion from this discussion is that more work is needed not only to eliminate gaps and to gather new data on additional variables (as we called for in the previous paragraphs) but also to sample more evolutionary trajectories in as diverse settings as possible.

In summary, the DR analysis attempts to distinguish correlation from causation by estimating what influence potential causal factors at a previous time has on the response variable at a later time. While an improvement over static correlations, where causal direction remains ambiguous, this method is, nevertheless, insufficient for making absolute claims of causality. Further scrutiny will be required to provide additional support for the provisional causal interpretations suggested in our article.

Dynamic regression

The conceptual modeling framework of cultural macroevolution (see Eq. 1) suggests analysis of data with DR models of the following form [see also (50)]

$$Y_{i,t} = a + \sum_{\tau} b_{\tau} Y_{i,t-\tau} + c \sum_{i \neq j} \exp\left[-\frac{\delta_{i,j}}{d}\right] Y_{j,t-1} +$$
$$h \sum_{i \neq j} w_{i,j} Y_{j,t-1} + \sum_{k} g_k X_{k,t-1} + \epsilon_{i,t}$$

Here, $Y_{i,t}$ is the response variable (Scale, Hier, or Gov) for location (NGA) i at time t. We construct a spatiotemporal series for response and predictor variables by following Seshat polities (or quasi-polities, such as archaeologically attested cultures) that occupied a specific NGA at each century mark during the sampled period. Thus, the time step in the analysis is 100 years.

On the right-hand side, a is the regression constant (intercept). The next term captures the influences of past history ("autoregressive terms"), with $\tau = 1, 2, \ldots$ indexing time-lagged values of Y (as time is measured in centuries, $Y_{i,t-1}$ refers to the value of the response 100 years before t).

The third term represents potential effects resulting from geographic diffusion. We used a negative exponential form to relate the distance between location i and location j, $\delta_{i,j}$, to the influence of j on i. Unlike a linear kernel, the negative exponential does not become negative at very large $\delta_{i,j}$, instead approaching 0 smoothly. The third term, thus, is a weighted average of the response variable values in the vicinity of location i at the previous time step, with weights falling off to 0 as distance from i increases. Parameter d measures how steeply the influence falls with distance, and parameter c is a regression coefficient measuring the importance of geographic diffusion.

For an overview of potential effects resulting from geographic diffusion, see (*98*).

The fourth term detects autocorrelations due to any shared cultural history at location i with other regions j using the phylogeny variable. Here, w represents the weight applied to the phylogenetic (linguistic) distance between locations (set to 1 if locations i and j share the same language, 0.5 if they are in the same linguistic genus, and 0.25 if they are in the same linguistic family). Linguistic genera and families were taken from *The World Atlas of Language Structures* and Glottolog (*99*).

The next term on the right-hand side represents the effects of the main predictor variables X_k, with g_k as regression coefficients. These variables (described in the "Outlining hypotheses and defining predictor variables" section) are of primary interest because they enable us to test various theories about the evolution of social scale and complexity. Last, $\varepsilon_{i,t}$ is the error term. We also include quadratic versions of these terms at a time lag (the "Dynamic regressions" section in the Supplementary Materials) to explore nonlinear responses to response and predictor factors.

Model selection

Model selection (choosing which terms to include in the regression model) was accomplished by exhaustive search: regressing the response variable on all possible linear combinations of predictor variables. This means that we tested >100,000 special hypotheses (this number is further increased because, in addition to the 17 possible predictors, we also investigate the effects of various autoregressive and nonlinear terms). The degree of fit was quantified by the AIC, which penalizes models with too many fitted coefficients. Possible nonlinear effects were checked by adding quadratic terms to the regression model. Standard diagnostic tests were performed for the best-fitting models (*50*). To check for cross-equation error correlations, we fitted a seemingly unrelated regression (*51*).

Multiple imputation and nonparametric bootstrap

Missing values, estimated uncertainty, and expert disagreement in the predictors (independent variables) were dealt with by multiple imputation (*52*). The response (dependent) variable, however, is not imputed, because such a procedure can result in biased estimates.

Imputation involves replacing missing entries with plausible values, and this allows us to retain all cases for the analysis. We use the approach of multiple imputation, in which analysis is done on many datasets, each created with different imputed values that are sampled in probabilistic manner. This approach results in valid statistical inferences that properly reflect the uncertainty due to missing values (*53*). Our procedure followed the approach introduced in (*4*):

1) Expert disagreement. In cases where experts disagree, each alternative coding has the same probability of being selected. Thus, if there are two conflicting codings presented by different experts and we create 20 imputed sets, then each alternative will be used roughly 10 times.

2) Uncertainty. Values that are coded with a confidence interval are sampled from a Gaussian distribution whose mean and variance are estimated, assuming that the interval covers 90% of the probability. For example, if a value of 1000 to 2000 was entered for the polity population variable, then we would draw values from a normal distribution centered on 1500 with an SD of 304. Thus, in 10% of cases, the value entered into the imputed set will be outside the data interval coded in Seshat. For categorical or binary variables, we sample coded values in proportion to the number of categories that are presented as plausible. For example, if Hier was coded as [2;3], that is, our degree of knowledge does not allow us to tell whether its value was 2 or 3 at a particular time, then the imputed data will contain "2" for roughly half the sets and "3" for the rest.

3) Missing data. For missing data, we impute values as follows. Suppose that for some polity, we have a missing value for variable A and coded values for variables B to H. We select a subset of cases from the full dataset in which all values of A to H variables have values and build a regression model for A. Not all predictors B to H may be relevant to predicting A, and, thus, the first step is selecting which of the predictors should enter the model. Once the optimal model is identified, we estimate its parameters. Then, we go back to the polity (where variable A is missing) and use the known values of predictor variables for this polity to calculate the expected value of A using estimated regression coefficients. However, we do not simply substitute the missing value with the expected one (because, as explained above, this is known to result in biased estimates). Instead, we sample from the regression residuals and add it to the expected value. We applied the same approach to each missing value in the dataset, yielding an imputed dataset without gaps.

The overall imputation procedure was repeated 20 times, yielding 20 imputed sets that were used in regression analysis as tests for the hypotheses.

Another source of potential bias is the violation of the assumptions of the statistical model needed to calculate confidence intervals and associated P values. Regression diagnostics indicate that the distribution of residuals violates the normality assumption (see Supplementary Results). Furthermore, for any data coming from the same geographic locality (NGA), it is possible that values are not truly independent because of memory effects. While we estimate and, when appropriate, model short-term memory effects by fitting autoregressive terms, there is also a possibility, which cannot be discounted, that there is a longer-term memory in the system.

We used nonparametric bootstrap to deal with this problem. However, instead of sampling (with replacement) each data point (polity-century), we sample with replacement the whole block of data associated with each NGA. Thus, the bootstrap procedure we used mimics the process by which we constructed the Seshat sample (see the "A brief introduction to the Seshat Global History Databank" section in the Supplementary Materials).

We combined multiple imputation with bootstrap. First, we created 20 imputed datasets, as described above. Second, we resampled, with replacement, NGAs in each imputed dataset 500 times, for a total of $20 \times 500 = 10,000$ bootstrapped datasets. We then calculated the statistics of interest (regression coefficients associated with various predictors) and constructed the frequency distribution of the 10,000 bootstrapped values. The P value is approximated by the proportion of statistic values greater than 0 (if the hypothesis we test is that, then the effect of the predictor is positive) or less than 0 (otherwise). The 95% confidence interval is then approximated by eliminating the smallest 250 and largest 250 values.

Calculating P values and confidence intervals assuming normality is expected to yield more liberal estimates, while resampling whole blocks of data for each NGA is a more conservative approach. Using these two approaches permits us to bracket the true values. The analysis sequence, thus, follows a two-phase approach. In the first phase (model selection), we check which predictors need to be included in the regression model, which autoregressive terms need to be explicitly modeled, the linearity of the relationships between the response and predictors, a test for possible omitted variables, and NGA fixed effects. As a result, we run many regressions, identify the "best

model" (with the smallest AIC), and sort the rest by increasing delAIC (difference from the best model). Once this model selection and testing phase is accomplished, the second phase (confidence tests) uses nonparametric bootstrap to approximate the *P* values and confidence intervals.

All analyses were performed using R version 4.0.2. R scripts and data files are published as a supplement to the article.

k-fold cross-validation

It is well known that the regression coefficient of determination, R^2, is an upwardly biased measure of the capacity of the fitted regression model to predict out-of-sample data (data that were not used in estimating the model). Adding more independent variables to the regression always results in a higher R^2 even when the variable has no effect. The improvement is achieved by a more complex model fitting the noise, rather than capturing the signal in the data. The standard approach for obtaining an unbiased measure of the capacity of the model to predict novel data is cross-validation, in which the model is fitted on one part of the dataset and its predictive ability is tested on another part, which was not used in estimating model coefficients. Such a straightforward approach, however, is very wasteful of data points, which are always in limited supply. This limitation can be overcome by a statistical technique known as *k*-fold cross-validation (*100*).

As noted above, simple cross-validation estimates the true predictability, characterizing a statistical model by splitting data into two sets. The parameters of the statistical model are estimated on the fitting set. Next, this fitted model is used to predict the data in the testing set. Because the prediction is evaluated on the "out of sample" data (data that were not used for fitting the model), cross-validation results give us a much better idea of the signal/noise ratio in the data compared to the coefficient of determination, R^2.

The accuracy of prediction is often quantified with the coefficient of prediction (*97*)

$$\rho^2 = 1 - \frac{\sum_{i=1}^n (Y_i^* - Y_i)^2}{\sum_{i=1}^n (\overline{Y} - Y_i)^2}$$

where Y_i is the observations from the testing set (the omitted values), Y_i^* is the predicted value, $\overline{Y}$ is the mean of Y_i, and n is the number of values to be predicted. The coefficient of prediction ρ^2 equals 1 if all data are perfectly predicted and 0 if the regression model predicts and the data average (in other words, if the model is simply $Y_i^* = \overline{Y}$). Unlike the regression R^2, which can vary between 0 and 1, prediction ρ^2 can be negative, when the regression model predicts data worse than the data mean. Prediction ρ^2 becomes negative when the sum of squares of deviations between predicted and observed is greater than the sum of squares of deviations from the mean.

In *k*-fold cross-validation rather than having a single fitting set and one testing set, we divide the data into *k* sets. In the analysis of Seshat data, we divide our dataset into 10 sets, for each of the 10 world regions. Next, we set aside one region, for example, Africa, and use the other nine regions to fit a regression model for the variable of interest. Here, we focus on the three measures of social complexity. After fitting a regression model for Scale, for example, using the data from nine regions (omitting Africa), we predict the values of *Y* (Scale) for Africa using the known values for other variables in African polities and the fitted regression coefficients. Next, we omit another region, for example, Europe, and repeat the exercise. At the end, we have predicted all data points by the out-of-sample method while fitting the model on $^9/_{10}$ of data at any given step.

SUPPLEMENTARY MATERIALS

REFERENCES AND NOTES

1. R. H. McGuire, Breaking down cultural complexity: Inequality and heterogeneity. *Adv. Archeol. Method Theory* **6**, 91–142 (1983).
2. G. Chick, Cultural complexity: The concept and its measurement. *Cross-Cult. Res.* **31**, 275–307 (1997).
3. G. M. Feinman, in *Cooperation and Collective Action: Archaeological Perspectives*, D. M. Carballo, Ed. (University of Colorado Press, 2013), pp. 299–307.
4. P. Turchin, T. E. Currie, H. Whitehouse, P. François, K. Feeney, D. Mullins, D. Hoyer, C. Collins, S. Grohmann, P. Savage, G. Mendel-Gleason, E. Turner, A. Dupeyron, E. Cioni, J. Reddish, J. Levine, G. Jordan, E. Brandl, A. Williams, R. Cesaretti, M. Krueger, A. Ceccarelli, J. Figliulo-Rosswurm, P. J. Tuan, P. Peregrine, A. Marciniak, J. Preiser-Kapeller, N. Kradin, A. Korotayev, A. Palmisano, D. Baker, J. Bidmead, P. Bol, D. Christian, C. Cook, A. Covey, G. Feinman, Á. D. Júlíusson, A. Kristinsson, J. Miksic, R. Mostern, C. Petrie, P. Rudiak-Gould, B. ter Haar, V. Wallace, V. Mair, L. Xie, J. Baines, E. Bridges, J. Manning, B. Lockhart, A. Bogaard, C. Spencer, Quantitative historical analysis uncovers a single dimension of complexity that structures global variation in human social organization. *Proc. Natl. Acad. Sci. U.S.A.* **115**, e144–e151 (2018).
5. A. W. Johnson, T. Earle, *The Evolution of Human Societies: From Foraging Group to Agrarian State* (Stanford Univ. Press, ed. 2, 2000).
6. K. Flannery, J. Marcus, *The Creation of Inequality: How Our Prehistoric Ancestors Set the Stage for Monarchy, Slavery, and Empire* (Harvard Univ. Press, 2012).
7. V. G. Childe, *Man Makes Himself* (Watts & Company, 1936).
8. L. A. White, *The Evolution of Culture* (McGraw-Hill, 1959).
9. E. R. Service, *Origins of the State and Civilization: The Process of Cultural Evolution* (Norton, 1975).
10. J. Diamond, Evolution, consequences and future of plant and animal domestication. *Nature* **418**, 700–707 (2002).
11. O. Borcan, O. Olsson, L. Putterman, Transition to agriculture and first state presence: A global analysis. *Explor. Econ. Hist.* **82**, 101404 (2021).
12. J. C. Scott, *Against the Grain: A Deep History of the Earliest States* (Yale Univ. Press, 2017).
13. D. Graeber, D. Wengrow, *The Dawn of Everything: A New History of Humanity*. (Farrar, Straus and Giroux, 2021).
14. M. H. Fried, *The Evolution of Political Society: An Essay in Political Anthropology* (Random House, 1967).
15. P. V. Kirch, *How Chiefs Became Kings: Divine Kingship and the Rise of Archaic States in Ancient Hawai'i* (University of California Press, 2010).
16. K. A. Wittfogel, *Oriental Despotism: A Comparative Study of Total Power* (Oxford Univ. Press, 1957).
17. R. D. Drennan, in *Chiefdoms in the Americas*, R. D. Drennan, C. A. Uribe, Eds. (University Press of America, 1987).
18. F. Oppenheimer, *The State; Its History and Development Viewed Sociologically* (Free Life Editions, 1975).
19. R. L. Carneiro, A theory of the origin of the state. *Science* **169**, 733–738 (1970).
20. M. Mann, *The Sources of Social Power. I. A History of Power From the Beginning to A.D. 1760* (Cambridge Univ. Press, 1986).
21. P. Turchin, D. Hoyer, J. Bennett, K. Basava, E. Cioni, K. Feeney, P. Francois, S. Holder, J. Levine, S. Nugent, J. Reddish, C. Thorpe, S. Wiltshire, H. Whitehouse, The Equinox2020 seshat data release. *Cliodynamics* **11**, 41–50 (2020).
22. P. J. Richerson, M. H. Christiansen, *Cultural Evolution: Society, Technology, Language, and Religion (Strüngmann Forum Reports)* (MIT Press, 2013).
23. N. Eldredge, *Macroevolution in Human Prehistory*, A. Prentiss, I. Kuijt, J. C. Chatters, Eds. (Springer, 2009).
24. A. Mesoudi, *Cultural Evolution: How Darwinian Theory Can Explain Human Culture and Synthesize the Social Sciences* (Chicago Univ. Press, 2011).
25. L. L. Cavalli-Sforza, M. W. Feldman, *Cultural Transmission and Evolution: A Quantitative Approach* (Princeton Univ. Press, 1981).
26. J. M. Beaulieu, D.-C. Jhwueng, C. Boettiger, B. C. O'Meara, Modeling stabilizing selection: Expanding the Ornstein–Uhlenbeck model of adaptive evolution. *Evolution* **66**, 2369–2383 (2012).
27. N. Wiener, The theory of prediction, in *ModernMathematics for Engineers*, E. Beckenbath, Ed. (McGraw-Hill, 1956).
28. C. W. J. Granger, Investigating causal relations by econometric models and cross-spectral methods. *Econometrica* **37**, 424–438 (1969).
29. M. Bunge, *Causality and Modern Science: Third Revised Edition* (Dover, 1979).
30. J. Pearl, *Causality: Models, Reasoning and Inference* (Cambridge Univ. Press, ed. 2, 2009).
31. N. Nunn, Historical Developement, in *Handbook of Economic Growth* (Elsevier, 2014), vol. 2A, pp. 347–402.

32. G. A. Johnson, Organizational structure and scalar stress, in *Theory and Ecplanation in Archaeology*, C. Renfrew, M. Rowlands, B. A. Segraves-Whallon, Eds. (Academic Press, 1982), pp. 389–421.

33. S. Gavrilets, A. Vose, The dynamics of Machiavellian intelligence. *Proc. Natl. Acad. Sci.* **103**, 16823–16828 (2006).

34. R. I. M. Dunbar, S. Shultz, Evolution in the social brain. *Science* **317**, 1344–1347 (2007).

35. C. S. Spencer, Territorial expansion and primary state formation. *Proc. Natl. Acad. Sci. U.S.A.* **107**, 7119–7126 (2010).

36. G. E. Swanson, *The Birth of the Gods: The Origin of Primitive Beliefs* (University of Michigan Press, 1960).

37. A. Norenzayan, A. F. Shariff, W. M. Gervais, A. K. Willard, R. A. McNamara, E. Slingerland, J. Henrich, The cultural evolution of prosocial religions. *Behav. Brain Sci.* **39**, e1 (2016).

38. P. Turchin, H. Whitehouse, J. Larson, E. Cioni, J. Reddish, D. Hoyer, P. E. Savage, R. Alan Covey, J. Baines, M. Altaweel, E. Anderson, P. K. Bol, E. Brandl, D. Carballo, G. Feinman, A. Korotayev, N. Kradin, J. Levine, S. Nugent, A. Squitieri, V. Wallace, P. François, Explaining the rise of moralizing religions: A test of competing hypotheses using the Seshat Databank. SocArXiv Preprint (2021); https://doi.org/10.31235/osf.io/2v59j.

39. P. N. Peregrine, Atlas of cultural evolution. *World Cultures* **14**, 2–88 (2003).

40. R. Michels, *Political Parties: A Sociological Study of The Oligarchical Tendencies of Modern Democracy* (Hearst's International Library, 1915).

41. J. Watts, O. Sheehan, Q. D. Atkinson, J. Bulbulia, R. D. Gray, Ritual human sacrifice promoted and sustained the evolution of stratified societies. *Nature* **532**, 228–231 (2016).

42. B. G. Trigger, *Understanding Early Civilizations* (Cambridge Univ. Press, 2014).

43. P. Turchin, A theory for formation of large empires. *J. Global Hist.* **4**, 191–217 (2009).

44. E. M. Redmond, C. S. Spencer, Chiefdoms at the threshold: The competitive origins of the primary state. *J. Anthropol. Archaeol.* **31**, 22–37 (2012).

45. G. M. Feinman, D. M. Carballo, Collaborative and competitive strategies in the variability and resiliency of large-scale societies in Mesoamerica. *Econ. Anthropol.* **5**, 7–19 (2018).

46. P. Turchin, D. Hoyer, A. Korotayev, N. Kradin, S. Nefedov, G. Feinman, J. Levine, J. Reddish, E. Cioni, C. Thorpe, J. S. Bennett, P. Francois, H. Whitehouse, Rise of the war machines: Charting the evolution of military technologies from the Neolithic to the Industrial Revolution. *PLOS ONE* **16**, e0258161 (2021).

47. P. Turchin, *Ultrasociety: How 10, 000 Years of War Made Humans the Greatest Cooperators on Earth* (Beresta Books, 2016).

48. P. Turchin, T. E. Currie, E. A. L. Turner, Mapping the spread of mounted warfare. *Cliodynamics* **7**, 217–227 (2016).

49. E. A. L. Turner, Discovering the anvil age: A map of the spread of Iron metallurgy across Afro-Eurasia. *Cliodynamics* **11**, 21–40 (2020).

50. P. Turchin, Fitting dynamic regression models to seshat data. *Cliodynamics* **9**, 25–58 (2018).

51. W. H. Greene, *Econometric Analysis* (Pearson Education, ed. 5, 2005).

52. D. B. Rubin, *Multiple Imputation for Nonresponse in Surveys* (Wiley, 1987).

53. Y. C. Yuan, *Multiple Imputation for Missing Data: Concepts and New Development (Version 9.0)* (SAS Institute, 2011).

54. K. P. Burnham, D. R. Anderson, *Model Selection and Inference: A Practical Information-Theoretic Approach* (Springer, 1998).

55. L. Stephens, D. Fuller, N. Boivin, T. Rick, N. Gauthier, A. Kay, B. Marwick, C. G. Armstrong, C. M. Barton, T. Denham, K. Douglass, J. Driver, L. Janz, P. Roberts, J. D. Rogers, H. Thakar, M. Altaweel, A. L. Johnson, M. M. Sampietro Vattuone, M. Aldenderfer, S. Archila, G. Artioli, M. T. Bale, T. Beach, F. Borrell, T. Braje, P. I. Buckland, N. G. Jiménez Cano, J. M. Capriles, A. Diez Castillo, Ç. Çilingiroğlu, M. Negus Cleary, J. Conolly, P. R. Coutros, R. A. Covey, M. Cremaschi, A. Crowther, L. der, S. di Lernia, J. F. Doershuk, W. E. Doolittle, K. J. Edwards, J. M. Erlandson, D. Evans, A. Fairbairn, P. Faulkner, G. Feinman, R. Fernandes, S. M. Fitzpatrick, R. Fyfe, E. Garcea, S. Goldstein, R. C. Goodman, J. Dalpoim Guedes, J. Herrmann, P. Hiscock, P. Hommel, K. A. Horsburgh, C. Hritz, J. W. Ives, A. Junno, J. G. Kahn, B. Kaufman, C. Kearns, T. R. Kidder, F. Lanoë, D. Lawrence, G. A. Lee, M. J. Levin, H. B. Lindskoug, J. A. López-Sáez, S. Macrae, R. Marchant, J. M. Marston, S. McClure, M. D. McCoy, A. V. Miller, M. Morrison, G. Motuzaite Matuzeviciute, J. Müller, A. Nayak, S. Noerwidi, T. M. Peres, C. E. Peterson, L. Proctor, A. R. Randall, S. Renette, G. Robbins Schug, K. Ryzewski, R. Saini, V. Scheinsohn, P. Schmidt, P. Sebillaud, O. Seitsonen, I. A. Simpson, A. Sołtysiak, R. J. Speakman, R. N. Spengler, M. L. Steffen, M. J. Storozum, K. M. Strickland, J. Thompson, T. L. Thurston, S. Ulm, M. C. Ustunkaya, M. H. Welker, C. West, P. R. Williams, D. K. Wright, N. Wright, M. Zahir, A. Zerboni, E. Beaudoin, S. Munevar Garcia, J. Powell, A. Thornton, J. O. Kaplan, M. J. Gaillard, K. Klein Goldewijk, E. Ellis, Archaeological assessment reveals Earth's early transformation through land use. *Science* **365**, 897–902 (2019).

56. C. S. Spencer, in *Handbook of Evolutionary Research in Archaeology*, A. M. Prentiss, Ed. (Springer, 2019), pp. 183–213.

57. S. J. Gould, N. Eldredge, Punctuated equilibria: The tempo and mode of evolution reconsidered. *Paleobiology* **3**, 115–151 (1977).

58. P. Turchin, S. Gavrilets, Tempo and mode in cultural macroevolution. *Evol. Psychol.* **19**, 147470492110666 (2021).

59. R. Hassig, *Mexico and the Spanish Conquest* (Longman, 1994).

60. A. A. Alves, *Brutality and Benevolence: Human Ethology, Culture, and the Birth of Mexico* (Greenwood Press, 1996).

61. H. Maschner, O. K. Mason, The bow and arrow in northern north america. *Evol. Anthropol.* **22**, 133–138 (2013).

62. P. Hamalainen, *The Comanche Empire* (Yale Univ. Press, 2008).

63. P. Turchin, T. E. Currie, E. A. L. Turner, S. Gavrilets, War, space, and the evolution of old world complex societies. *Proc. Natl. Acad. Sci.* **110**, 16384–16389 (2013).

64. J. S. Bennett, Retrodicting the rise, spread, and fall of large-scale states in the old world. *PLOS ONE* **17**, e0261816 (2022).

65. G. M. Feinman, The emergence of social complexity: Why more than population size matters, in *Cooperation and Collective Action: Archaeological Perspectives*, D. M. Carballo, Ed. (University Press of Colorado, 2013), pp. 35–56.

66. G. P. Murdock, C. Provost, Measurement of cultural complexity. *Ethnology* **12**, 379–392 (1973).

67. T. E. Currie, R. Mace, Political complexity predicts the spread of ethnolinguistic groups. *Proc. Natl. Acad. Sci.* **106**, 7339–7344 (2009).

68. C. S. Spencer, E. M. Redmond, Multilevel selection and political evolution in the Valley of Oaxaca, 500–100 B.C. *J. Anthropol. Archaeol.* **20**, 195–229 (2001).

69. R. Blanton, L. Fargher, *Collective Action in the Formation of Pre-Modern States* (Springer, 2008).

70. G. M. Feinman, J. Marcus, *Archaic States* (School of American Research Press, 1998).

71. G. P. Murdock, *Ethnographic Atlas* (University of Pittsburgh Press, 1967).

72. G. P. Murdock, D. R. White, Standard cross-cultural sample. *Ethnology* **8**, 329–369 (1969).

73. H. T. Wright, Recent research on the origin of the state. *Ann. Rev. Anthropol.* **6**, 379–397 (1977).

74. M. Weber, *Wirtschaft und Gesellschaft* (Verlag von J. C. B. Mohr, 1922).

75. R. L. Carneiro, The Chiefdom: Precursor of the state, in *The Transition to Statehood in the New World*, G. D. Jones, R. D. Kautz, Eds. (Cambridge Univ. Press, 1981), pp. 37–39.

76. P. J. Richerson, R. Boyd, *Not by Genes Alone: How Culture Transformed Human Evolution* (University of Chicago Press, 2005).

77. H. T. Wright, Early state dynamics as political experiment. *J. Anthropol. Res.* **62**, 305–319 (2006).

78. T. E. Currie, P. Turchin, J. Bednar, P. J. Richerson, G. Schwesinger, S. Steinmo, R. Wacziarg, J. J. Wallis, Evolution of institutions and organizations, in *Complexity and Evolution: Toward a New Synthesis for Economics*, D. S. Wilson, A. Kirman, Eds. (MIT Press, 2016), pp. 199–234.

79. T. E. Currie, P. Turchin, S. Gavrilets, History of agriculture and intensity of warfare shaped the evolution of large-scale human societies in Afro-Eurasia. SocArXiv Preprint (2019); https://osf.io/preprints/socarxiv/9kmrw/.

80. T. Denham, H. Barton, C. Castillo, A. Crowther, E. Dotte-Sarout, S. A. Florin, J. Pritchard, A. Barron, Y. Zhang, D. Q. Fuller, The domestication syndrome in vegetatively propagated field crops. *Ann. Bot.* **125**, 581–597 (2020).

81. P. Turchin, T. Currie, C. Collins, J. Levine, O. Oyebamiji, N. R. Edwards, P. B. Holden, D. Hoyer, K. Feeney, P. François, H. Whitehouse, An integrative approach to estimating productivity in past societies using Seshat: Global History Databank. *Holocene* **31**, 1055–1065 (2021).

82. M. Olson, *The Logic of Collective Action: Public Goods and the Theory of Groups* (Harvard Univ. Press, 1965).

83. R. E. Blanton, Cooperation and the moral economy of the marketplace, in *Merchants, Markets, and Exchange in the Pre-Columbian World*, K. G. Hirth, J. Pillsbury, Eds. (Dumbarton Oaks, 2013), pp. 23–48.

84. C. Tilly, *Coercion, capital, and European states, AD 990–1990* (Blackwell, 1990).

85. J. Mayshar, O. Moav, Z. Neeman, L. Pascali, Cereals, appropriability, and hierarchy. Discussion Paper No. 10742, Centre for Economic Policy Research. (2015).

86. P. Boyer, *Religion Explained: The Evolutionary Origins of Religious Thought* (Basic Books, 2001).

87. R. W. Hood, P. C. Hill, B. Spilka, *The Psychology of Religion: An Empirical Approach* (The Guilford Press, ed. 4, 2009).

88. R. N. Bellah, *Religion in Human Evolution: From the Paleolithic to the Axial Age* (Harvard Univ. Press, 2011).

89. J. Watts, S. J. Greenhill, Q. D. Atkinson, T. E. Currie, J. Bulbulia, R. D. Gray, Broad supernatural punishment but not moralizing high gods precede the evolution of political complexity in Austronesia. *Proc. R. Soc. B* **282**, 20142556 (2015).

90. D. S. Wilson, *Darwin's Cathedral: Evolution, Religion, and The Nature of Society* (University of Chicago Press, 2002).

91. J. M. Bering, The folk psychology of souls. *Behav. Brain Sci.* **29**, 453–462 (2006).

92. D. Kelemen, Are children "Intuitive Theists"?: Reasoning about purpose and design in nature. *Psychol. Sci.* **15**, 295–301 (2004).

93. B. Raffield, N. Price, M. Collard, Religious belief and cooperation: A view from Viking-Age Scandinavia. *Religion Brain Behav.* **9**, 2–22 (2019).

94. S. A. Nefedov, *War and Society (In Russian: Voyna i obschestvo. Faktornyi analiz istoricheskogo protsessa)* (Territoriya buduschego, 2009).

95. R. Drews, *The End of the Bronze Age: Changes in Warfare and the Catastrophe ca. 1200 BC* (Princeton Univ. Press, 1993).

96. J. Kim, Elite strategies and the spread of technological innovation: The spread of iron in the bronze age societies of Denmark and Southern Korea. *J. Anthropol. Archaeol.* **20**, 442–478 (2001).

97. P. Turchin, *Complex Population Dynamics: A Theoretical/Empirical Synthesis* (Princeton Univ. Press, 2003).

98. E. A. Eff, P. W. Routon, Farming and fighting: An empirical analysis of the ecological-evolutionary theory of the incidence of warfare. *Struct. Dyn.* **5**, 1–33 (2012).

99. M. S. Dryer, M. Haspelmath, *The World Atlas of Language Structures Online* (Max Planck Institute for Evolutionary Anthropology, 2013).

100. R. Kohavi, A study of cross-validation and bootstrap for accuracy estimation and model selection. *Proc. Int. Joint Conf. Artificial Intell.* **2**, 1137–1143 (1995).

101. P. Turchin, R. Brennan, T. Currie, K. Feeney, P. Francois, D. Hoyer, J. Manning, A. Marciniak, D. Mullins, A. Palmisano, P. Peregrine, E. A. L. Turner, H. Whitehouse, Seshat: The global history databank. *Cliodynamics* **6**, 77–107 (2015).

102. P. François, J. G. Manning, H. Whitehouse, R. Brennan, T. Currie, K. Feeney, P. Turchin, A macroscope for global history: Seshat Global History Databank, a methodological overview. *Digital Hum. Quart.* **10**, 4 (2016).

103. S. Hall, C. Moskovitz, M. Pemberton, *Understanding Text Recycling: A Guide for Researchers* (Text Recycling Research Project, 2021).

104. F. E. Reed, *CENTENNIA for Windows* (Clockwork Software Inc., 1996).