Universal abundance fluctuations across microbial communities, tropical forests, and urban populations

Ashish B. George[1*] and James O'Dwyer[1*]
[1] *Department of Plant Biology and Carl R. Woese Institute for Genomic Biology, University of Illinois at Urbana-Champaign, Urbana, IL 61801, USA.*
*Correspondence to: ashish.b.george@gmail.com, jodwyer@illinois.edu
(Dated: September 14, 2022)

The growth of complex populations, such as microbial communities, forests, and cities, occurs over vastly different spatial and temporal scales. Although research in different fields has developed detailed, system-specific models to understand each individual system, a unified analysis of different complex populations is lacking; such an analysis could deepen our understanding of each system and facilitate cross-pollination of tools and insights across fields. Here, we analyze time-series data of the human gut microbiome, tropical forest, and urban employment using a shared framework. We demonstrate that a single, three-parameter model of stochastic population dynamics can reproduce the empirical distributions of population abundances and fluctuations in all three data sets. The three parameters characterizing a species measure its mean abundance, deterministic stability, and stochasticity. Our analysis reveals that, despite the vast differences in scale, all three systems occupy a similar region of parameter space when time is measured in generations. In other words, although the fluctuations observed in these systems may appear different, this difference is primarily due to the different observation periods for each system. Further, we show that the distribution of temporal abundance fluctuations is described by just two parameters and derive a two-parameter functional form for abundance fluctuations to improve risk estimation and forecasting.

INTRODUCTION

The dynamics of complex populations is studied in fields ranging from microbiology to economics. These studies have culminated in theoretical and computational models, with various levels of system-specific detail, that have made progress towards explaining system behavior and developing conceptual understanding. This includes the consumer-resource and Lotka-Volterra models of microbial communities; economic, econometric, and physics-based models of urban dynamics; and ecological models of forests based in niche and neutral theory[1–6]. Quantitative analysis, inspired by these models, has also unearthed statistical patterns in the data, hinting at simple emergent behavior of each population[7–10].

What is lacking, however, is an investigation of emergent behavior using a common framework across different populations and an analysis of the temporal fluctuations in these populations. Classic models in statistical physics, such as diffusion and Ornstein-Uhlenbeck process, have successfully described the behavior of diverse physical systems[11–13]. Their success demonstrates that some emergent properties are determined by only a few key underlying details of the dynamics. Efforts at applying this philosophy in other fields have sometimes succeeded[4,14–16], but are often hindered by the lack of data and the incorporation of many system-specific details that make models difficult to analyze. In this paper, we undertake a statistical physics-inspired investigation of three different complex populations: microbial communities, tropical forests, and human cities, through a common framework. Our analysis will aim to uncover the key underlying similarities and differences between the populations; this will not only deepen our understanding of each system, but also facilitate the cross-pollination of tools and techniques between research fields.

All three complex populations we consider fluctuate in time. Large fluctuations in these populations are associated with catastrophic events such as disease, economic or ecological collapse[17–20]; hence understanding these fluctuations is crucial for risk assessment, quantitative biological methods, and forecasting[17,21–23]. Yet, many models of these complex populations study the equilibrium and steady-state properties such as the average population abundance[1,2,20,24–27]. This restriction is in part because 1) analyzing dynamical properties of complex models is harder than analyzing their steady-state behavior and 2) temporal data required to fit and validate complex models has been lacking. We will tackle these two challenges by 1) adopting a minimal model which makes analysis of fluctuations feasible and 2) utilizing data from three separate systems to fit and validate the model. Generating a reliable null model for population fluctuations that enables improved risk estimation is therefore a major goal of the study.

We analyze time-series data of abundances in 3 different systems: the human gut microbiome, tropical forests, and human cities. These systems span spatial scales from a single body to an entire country, temporal scales from days to years, and are studied in separate research fields. Nevertheless, we harness the similarities in structure of the data to analyze the same emergent features of all three systems. We compare the observed features to predictions from a model of stochastic population dynamics with just three population-specific parameters. This simplicity allows us to make analytical predictions for emergent features that we can fit and validate using the limited data. Remarkably, the model is able to capture most of the observed variation, despite its simplicity.

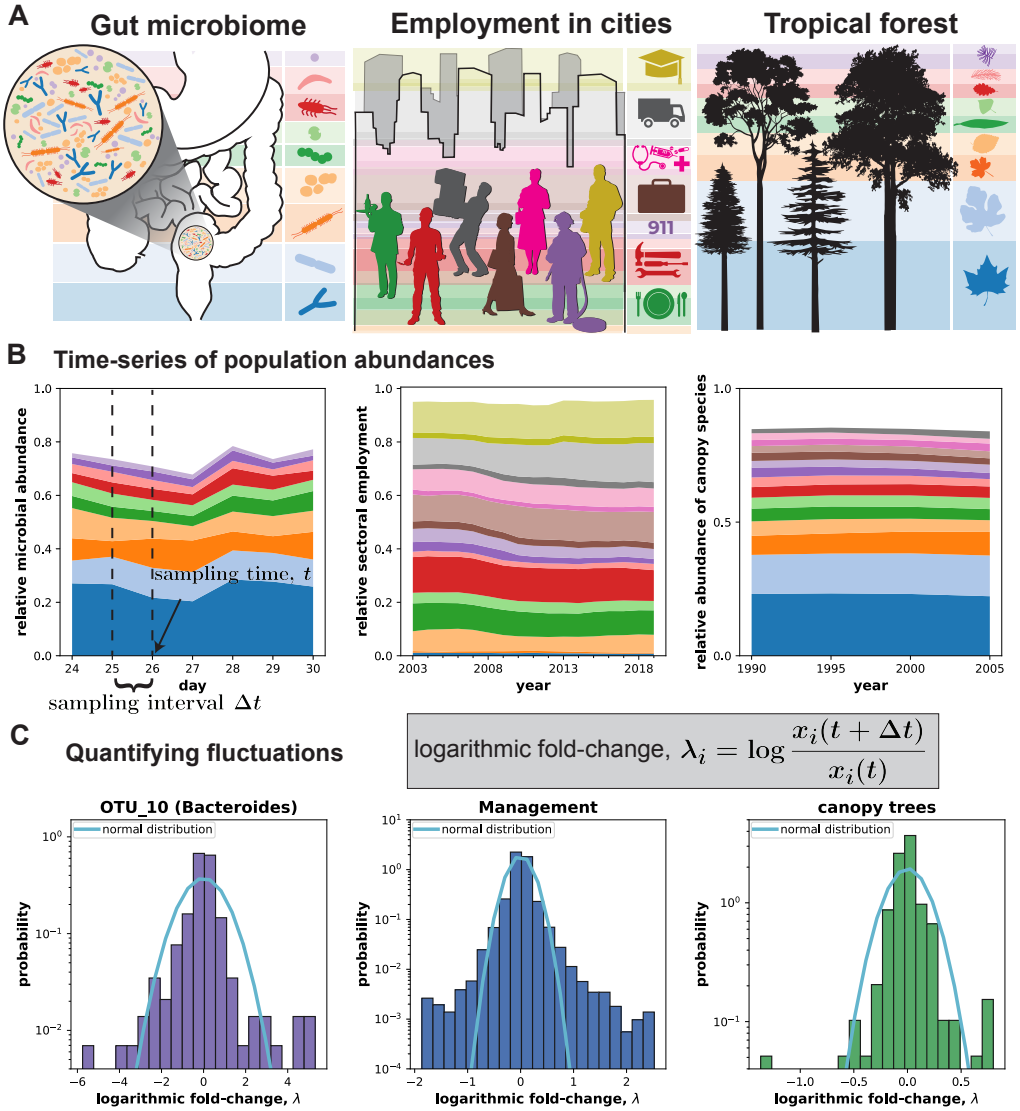

FIG. 1. **Time series data from three complex populations spanning spatio-temporal scales. A)** We analyze time-series population data from three complex populations: the human gut microbiome, employment in U.S. cities, and trees in a tropical forest. Each complex population is composed of sub-populations, i.e., microbial species in the microbiome, employment within different economic sectors in cities, and tree species in four trait-based clusters in a tropical forest. **B)** Abundances of the sub-populations (species/sectors), sampled at periodic intervals, continue to fluctuate in time. Data shows microbiome abundance over a week, employment over 17 years for San Diego metropolitan area, and abundance of canopy tree species. **C)** The abundance fluctuations are measured using the logarithmic fold-change, $\lambda_i = \log \frac{x_i(t+\Delta t)}{x_i(t)}$, where x_i is the abundance of microbial species, an economic sector across cities, or tree species in a height cluster. The distribution of logarithmic fold change (LFD) for one microbial species, the management sector across all cities, and all canopy tree species is shown. The fit by the normal distribution to the LFD is shown by the cyan line. The fit is unable to capture the large fluctuations in the tails, central peak of the distribution, or both, illustrating that fluctuations in these complex populations are not normally distributed. Only the 10 (15) most abundant microbial (tree) species are shown in panel B for clarity.

MODEL AND DATA

To develop a unified understanding of complex population dynamics across spatio-temporal scales, we analyze time-series population data from three disparate systems: microbial communities in the human gut, trees in a tropical forest, and employment in U.S. cities (see Fig.1A). Traditionally studied by different fields, we interrogate the data using a shared framework by taking advantage of the similarities in the structure of the data. Each data set contains the relative abundance of population categories within each community at a sequence of time points (see Fig.1B).

More specifically, the gut microbiome dataset records the relative abundances of microbial species in the human

gut sampled at daily intervals for almost a year[28]. The data from the Barro-Colorado Island (BCI) forest records the number of tree species within a 50 hectare plot on the island, sampled at 5-year intervals for two decades. We group trees based on clustering along trait axes into four clusters following recent work[29]. The four clusters reflect the classical grouping of trees based on height-niches into shrubs, understory treelets, midstory trees, and canopy trees. The employment dataset records the number of people employed in different economic sectors, classified according to North American Industry Classification System (NAICS), in 383 US cities (Metropolitan Statistical Areas). The data is sampled monthly for 17 years. We analyze the relative employment in a sector (employment in a sector in a city divided by total employment in the city) to remove the effect of varying city sizes. Similarly, we analyze the relative abundance of a tree species within a height cluster. Since we analyze relative abundance data in all three three complex populations, for brevity, 'abundance' will mean relative abundance going forward, unless otherwise specified.

In contrast to the static predictions from most deterministic models, the data in Fig.1B shows that population abundances continue to fluctuate in time. Furthermore, the strength of these fluctuations differs between the three systems with the largest fluctuations observed in the microbiome. To quantify the strength of fluctuations, we measured the logarithmic fold-change in abundance over a time interval Δt, defined as

$$\lambda_i = \log \frac{x_i(t+\Delta t)}{x_i(t)}, \tag{1}$$

where $x_i(t)$ denotes the (relative) abundance of a species/sector i at time t. The empirical distribution of logarithmic fold-change or the Logarithmic Fold-change Distribution (LFD) is shown for a single species/sector from each system in Fig.1C. The comparison with the fit of the logarithmic fold-change by the normal distribution (equivalent to fitting fold-change by the lognormal distribution) illustrates that fluctuations cannot be understood as an outcome of just multiplicative noise (see Methods). Further, the fit by the normal distribution indicates that fluctuations much larger than expected from a normal distribution may occur in some of these complex populations. Understanding the distribution of abundance fluctuations is required to quantify the likelihood these large fluctuations which have a major impact on risk-estimates and time-series forecasting.

We now develop a simple model that is capable of predicting statistical features of all three data-sets, including the LFD. To achieve our goal of describing all three data sets, we keep the model both as generally-applicable and simply-formulated as possible. The model assumes that the abundance of a species/sector i, x_i fluctuates around an equilibrium value, x_i^*, determined by the metabolic, ecological, or economic niche the species/sector occupies. We do not explore the system-specific mechanisms (resource competition, metabolic/economic interactions, etc.) that determine the particular equilibrium value[1,2,26,30–32]. Fluctuations then occur due to the stochastic processes governing population growth and decline. Deviations from the equilibrium value result in a linear restoring force, described by $(x_i^* - x_i)/\tau_i$, where τ_i is the timescale of return to equilibrium. Based on these assumptions, we call the model the Stochastic Linear-Response Model (SLRM) of population dynamics.

Assuming that population growth and decline occur in proportion to the abundance, we can write down the stochastic differential equation of the SLRM governing population abundances:

$$\frac{dx_i}{dt} = \frac{x_i^*}{\tau_i} - \frac{x_i}{\tau_i} + \sqrt{2\sigma_i x_i}\,\eta(t), \tag{2}$$

where σ_i captures the strength of population fluctuations, and $\eta(t)$ is delta-correlated gaussian noise. The model resembles the classic Ornstein-Uhlenbeck process, used to describe many stochastic quantities in physics and finance[11,13,33], with one crucial difference: the scaling of the stochastic fluctuations with the square root of x_i. This scaling reflects our assumption on the stochastic processes underlying population growth and decline; specifically, the square root scaling arises when growth and decline is a consequence of many independent, random events and the number of such events is proportional to x_i, and is commonly used in many stochastic population models[34]. Interpreting x_i^*/τ_i as an effective immigration or supply term, we can recast the model as a birth-death-immigration model[35]. A similar model has also been analyzed in the context of bond interest rates and singular diffusion processes[36,37].

In ecological terms, the SLRM resembles a scenario where species occupy well-separated niches, with equilibrium abundance x_i^* for species i. This idea of niche-separation determining model parameters is reflected in our assumptions. Specifically, in the microbiome, each species is described by its own SLRM parameters; in urban employment, each sector is described by its own SLRM parameters (independent of city); in forests, each trait-based cluster is described by its own SLRM parameters (independent of species index within the cluster). This assumption is reflected in our choice of showing the empirical LFD for a single sector across all cities, a single microbial species in the gut, and all tree species within a trait cluster in Fig.1C. While inter-species interactions may occur in each system, incorporating interactions drastically increases the model complexity hindering parameter-inference[23,38]. Further, species interactions may not be necessary to describe population fluctuations in these systems, as we will show.

The minimal nature of the model allows us to make analytical predictions for two key quantities that characterize its long-term behavior. First, at long times ($t > \tau_i > 0$) the distribution of population abundances will converge to a steady-state distribution. The steady-state distribution arises from the balance of stochastic fluctuations that kick the population from its determin-

istic equilibrium value and the restoring force towards the equilibrium. The analytical form of the distribution is (see SI)

$$P_{\mathrm{ss},i}(x) = \frac{(\sigma_i\tau_i)^{-x_i^*/\sigma_i\tau_i}}{\Gamma(x_i^*/\sigma_i\tau_i)} x^{\frac{x_i^*}{\sigma_i\tau_i}-1} e^{-\frac{x}{\sigma_i\tau_i}}, \qquad (3)$$

which is a Gamma distribution defined by two parameters: $\sigma_i\tau_i$ and the Equilibrium to Noise Ratio (ENR), $x_i^*/\sigma_i\tau_i$, and $\Gamma()$ is the gamma function.$\sigma_i\tau_i$ quantifies the effective strength of the noise i.e., the fluctuations over the equilibration time period. The ENR, $x_i^*/\sigma_i\tau_i$, quantifies the relative strength of deterministic restoring force to stochastic fluctuations.

Second, we can analytically derive how abundances near steady-state fluctuate in time. Specifically, we quantify the temporal fluctuations by measuring the logarithmic fold-change of population abundances between two time points Δt time units apart, denoted as $\lambda_i = \log \frac{x_i(t+\Delta t)}{x_i(t)}$. The distribution of logarithmic fold-changes λ_i or Logarithmic Fold-change Distribution (LFD), near the steady-state is given by (see SI)

$$P_{\mathrm{fluc},i}(\lambda,\Delta t) = \frac{\Gamma\left(\frac{x_i^*}{\sigma_i\tau_i}+\frac{1}{2}\right)}{2\sqrt{\pi}\Gamma\left(\frac{x_i^*}{\sigma_i\tau_i}\right)} \left(e^{\frac{\Delta t}{\tau_i}}-1\right)^{\frac{x_i^*}{\sigma_i\tau_i}} e^{\frac{\Delta t}{2\tau_i}} \cosh\frac{\lambda}{2} \left(\frac{1}{e^{\frac{\Delta t}{\tau_i}}(\cosh\frac{\lambda}{2})^2-1}\right)^{\frac{x_i^*}{\sigma_i\tau_i}+\frac{1}{2}}. \qquad (4)$$

Despite its imposing appearance, the equation depends only on two parameters: the ENR, $x_i^*/\sigma_i\tau_i$, and the equilibration timescale, τ_i, which scales the time interval Δt. This distribution quantifies the likelihood of large abundance fluctuations in the model. In particular, the tails of the distribution at large λ decay exponentially at a rate proportional to the ENR, i.e, $\sim e^{-\frac{x_i^*}{\sigma_i\tau_i}|\lambda|}$, which ensures the moments of the distribution are well-defined. Note that this exponential decay for the logarithmic fold-change at large λ corresponds to a power-law decay for the fold-change, $\rho = \frac{x_i(t+\Delta t)}{x_i(t)}$, at large ρ with exponent $-\frac{x_i^*}{\sigma_i\tau_i}-1$.

RESULTS

Model captures empirical abundance fluctuations across diverse populations

In line with our goal of characterizing abundance fluctuations in complex populations, we plot the empirical Logarithmic Fold-change Distribution (LFD) in the three systems in Fig. 2. Fig. 2A,B,C plots the LFD of urban employment, aggregated across all US cities, for three different sectors. Fig. 2D,E,F plots the LFD of 3 different species in the human gut microbiome. Fig. 2G,H plots the LFD of all the tree species belonging to two height cluster in the forest. The fold-change in abundance was calculated for time intervals of 1 year, 1 day, and 5 years for the city, microbiome, and forest data.

All LFDs were roughly centered around $\lambda_i = 0$ (average λ_i was 0.001, 0.004, and 0.05 for city, microbiome, and forest), but exhibited a large variation relative to the mean (coefficient of variation $\gg 1$; see SI Fig.S1). This relative dominance of fluctuations over systematic trends not only makes it easier to study fluctuations easier, but also underscores the need to understand fluctuations better. The Stochastic Linear Response Model (SLRM), which neglects any systematic trends in the mean, is a good candidate model of population fluctuations.

The red lines in Fig. 2 show the fit of the observed LFDs with the SLRM prediction, P_{fluc} (Eq. (4)). A separate fit is performed for each species/sector, allowing the inference of corresponding population parameters. Despite having only two free parameters, the SLRM is able to fit the observed fluctuations across these diverse systems. We compared the SLRM fit to fits by two other candidate distributions using the Akaike Information Criterion[39], as an additional test of the fit. The two candidate distributions (normal distribution and Laplace distribution) were chosen based on their success at fitting empirical fluctuations and alternative models (see Methods and Ref.[7,34,40]). The SLRM prediction outperformed the other candidate distributions in the majority of cases (see SI Fig.S2).

We then further quantified the fit, independent of comparisons with other models, by testing whether the observed data was likely to have been generated by the model. Specifically, for each fit and corresponding parameter estimates, we compared the likelihood of the observed data to the likelihood of 100 random samples of the same size as the data generated from the fitted distribution. If the likelihood of the observed data was smaller than the 95% of the samples, we concluded that the data was unlikely to have been generated by the fitted distribution and rejected the fit. The percentile score in Fig.1I quantifies the percentage of random samples with a larger likelihood than the observed data. All fits had a percentile score <95%, meaning that SLRM passes our goodness of fit test in all cases. Thus P_{fluc} serves as a reasonable null expectation for the distribution of population fluctuations.

Comparing the three systems in Fig. 2, we notice that the scale of fold-change on the x-axis appears larger for the microbiome (Fig. 2D-F) than the two macroscopic systems (Fig. 2A-C,G-H). These differences in the data are reflected in the the parameters values of the fit, and we explore this parameter variation in the next section. Importantly, despite these differences in scale and shape of the data, P_{fluc} provides a good fit to data from all three systems.

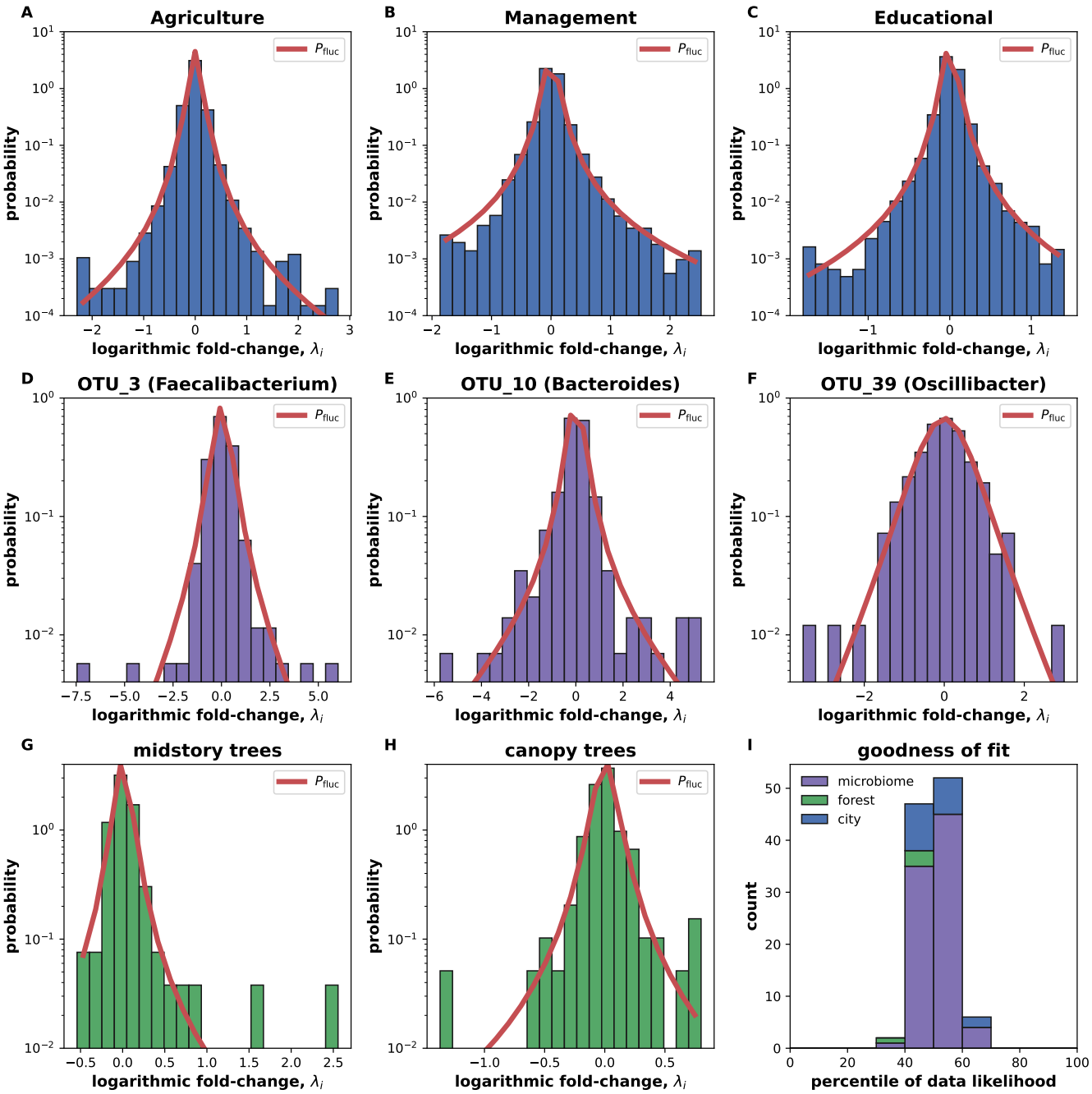

FIG. 2. **Universal distribution of fluctuations across complex populations.** The histogram of the empirical Logarithmic Fold-change Distribution (LFD), i.e., the distribution of $\lambda_i = \log \frac{x_i(t+\Delta t)}{x_i(t)}$, in each system is plotted and fit with the model prediction P_{fluc} from Eq. (4) (red line). The empirical distribution and fits are shown for: **(A-C)** employment in 3 different sectors aggregated across U.S. cities, **(D-F)** abundances of 3 microbial species in the human gut, and **(G,H)** abundances of tree species within two height clusters in the BCI forest. **I)** The percentile score, which quantifies the goodness of fit, shows that the likelihood of observing the data is similar to the likelihood of a random sample from the fitted distribution (of the same size as the data). Specifically, the percentile score quantifies the percentage of random samples with a higher likelihood than the observed data. A percentile score $> 95\%$ indicates a poor fit; all fits had a percentile score $< 95\%$, as shown in the stacked histogram. The fold-change in abundance was calculated for time intervals of 1 year, 1 day, and 5 years for the city, microbiome, and forest data.

Universal fluctuations appear different in micro- and macroscopic systems due to observation timescale

Fitting P_{fluc} to the empirical LFD provides us with maximum likelihood estimates for the equilibration timescale τ_i and the ENR $x_i^*/\sigma_i\tau_i$. We compare the dimensionless versions of these inferred parameters across the three systems. Specifically, we compare the ratio of the equilibration timescale to time interval of observation τ/t_{obs} and the ENR across the three systems.

Fig 3A plots the ENR and τ/t_{obs} across the three systems. While the ENR spans a similar range across all three data sets, the micro- and macroscopic systems differ in the inferred values of τ/t_{obs}. For microbial populations, $\tau/t_{obs} < 1$, which means that the duration over which we observe the system, about a year, is longer than its equilibration timescale which is around 10 days. In contrast, cities and forests have $\tau/t_{obs} > 1$, implying that the duration of observation, two decades, is much shorter than the equilibration timescale. This difference in τ/t_{obs} is responsible for the difference in shapes of the fitted P_{fluc} in Fig. 2. Thus, the duration of observation is an important difference between micro- and macroscopic systems.

An alternative way to compare these populations is to use the ratio of the equilibration timescale to the generation time, τ/t_{gen} (see Fig 3B). The generation times in the 3 systems were estimated as 4.2 hours, 10 years, and 55 years for microbes, employment, and forests respectively (see Methods and Refs.[41–44]). Remarkably, when viewed in terms of generation time, all 3 systems occupy a similar region of parameter space. Hence, fluctuations in the three systems are described by the same distribution over a similar parameter range when time is measured in generations, highlighting the similarities in the emergent behavior across the systems.

Note that a small subset of microbes in Fig 3B appear as outliers, with $\tau/t_{gen} < 1$. For these microbes, the empirical LFD is better-fit by a normal distribution than the model prediction P_{fluc} (see SI Fig.S3). This may be due to processes not incorporated in the model, such as environmental stochasticity (see Methods), which has been used to explain certain properties of complex populations [8,34,45,46]. But, we did not require it to fit the majority of our data.

For the remaining data, the equilibration time scale is on the order of 10 generations for all three systems. Therefore, when viewed in terms of generation time rather than observation time, emergent fluctuations in the three systems are highly similar.

The steady-state abundance distribution

From Fig. 3, we see that the observation duration for the microbiome is longer than its equilibration time scale ($\tau/t_{obs} < 1$). Hence the temporal trajectory of abundances of each microbial species should trace out the cor-

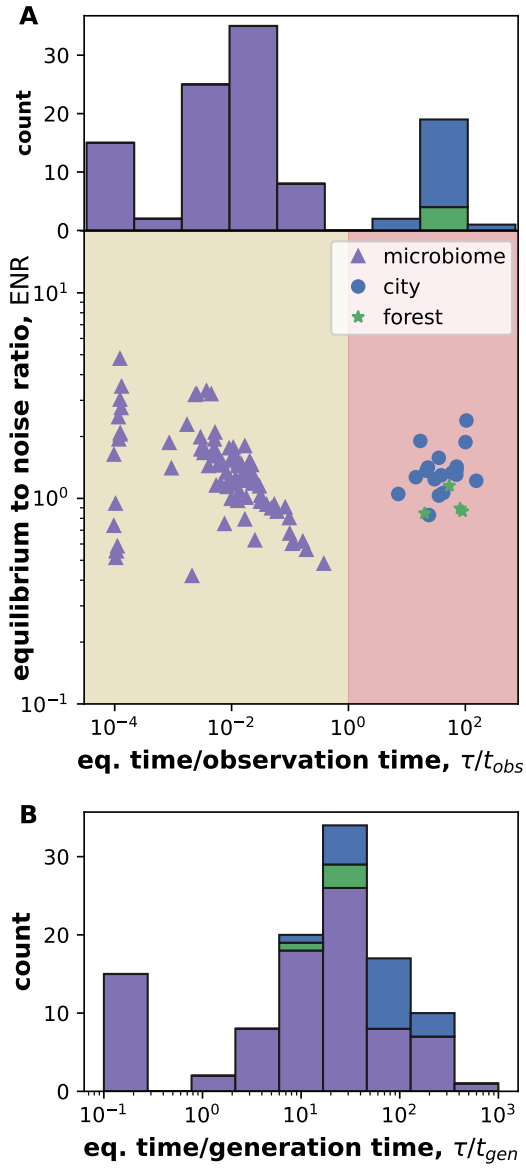

FIG. 3. **Timescale of observation differentiates microscopic and macroscopic systems; generation time unifies them. A)** Plot of the dimensionless parameters ENR ($x_i^*/\sigma_i\tau_i$) and τ/t_{obs} inferred from fitting the empirical LFD. τ/t_{obs} is the ratio of the equilibration timescale to time duration of observation. All three systems span similar values in ENR, but τ/t_{obs} differentiates the microscopic and macroscopic systems. Microbiome data has $\tau/t_{obs} < 1$, i.e., we observe the microbiome for longer than its equilibration times. The macroscopic systems have $\tau/t_{obs} > 1$, i.e., we observe cities and forests for shorter than their equilibration times. **B)** Measuring time in generations reveals the similarities between systems. The stacked histogram of inferred τ/t_{gen}, the ratio of the equilibration timescale to the generation time, shows that all 3 systems inhabit a similar parameter range, with an equilibration timescale on the order of 10 generations. A small group of microbes appear as outliers with $\tau/t_{gen} < 1$; these are the microbes for which the LFD was well-fit by a normal distribution (see SI Fig.S3). The estimated generation time, based on Refs.[41–44], was 4.2 hours, 10 years, and 55 years for microbiome, cities, and forests (see Methods).

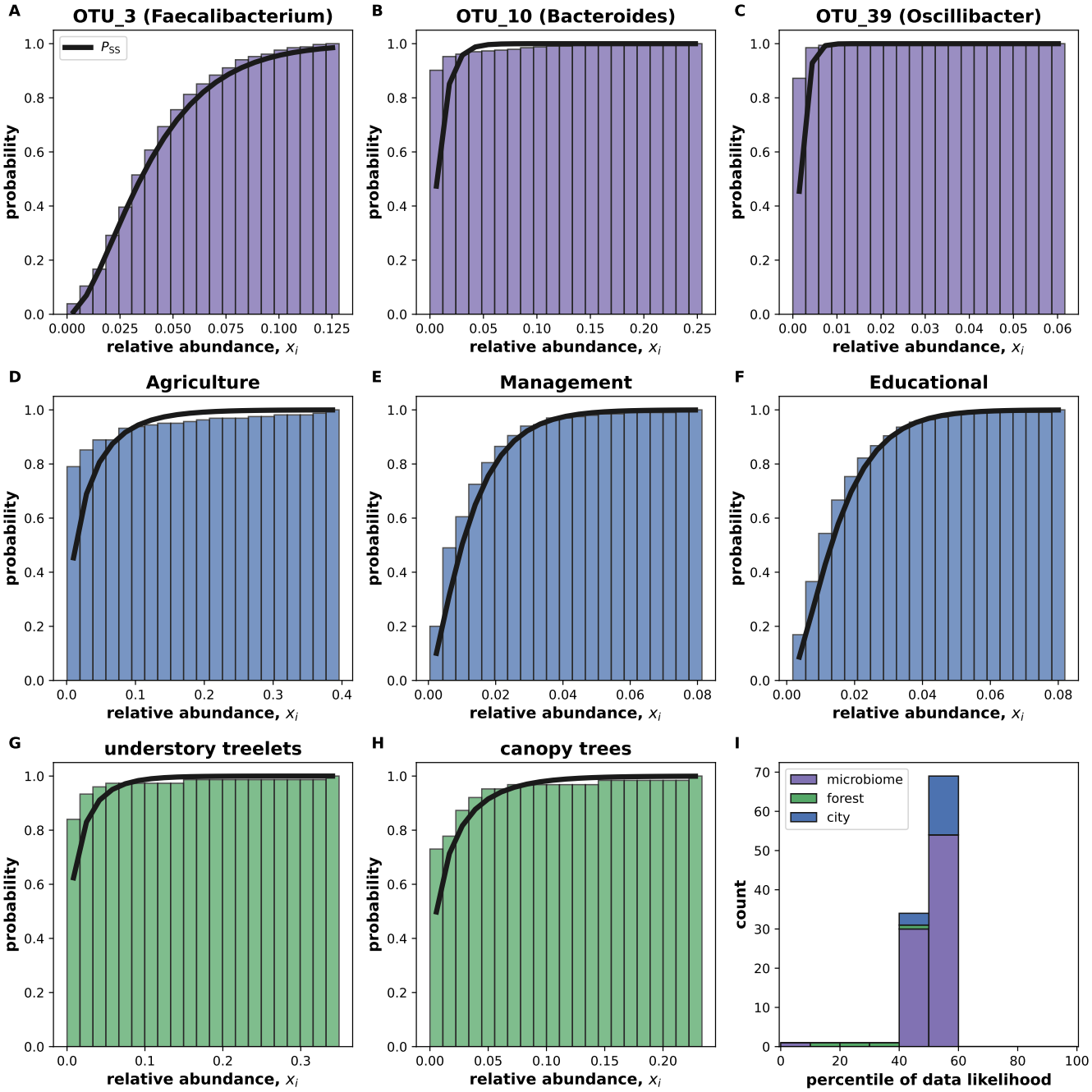

FIG. 4. **The distribution of abundances is well-fit by the SLRM steady-state distribution. A-C)** The distribution of microbial abundances over time for three different species is fit by the two-parameter gamma distribution predicted by the SLRM (black line). We fit the temporal distribution of microbial abundances because we observe the microbiome for far longer than its equilibration time scale. For cities and forests, we observe the system for intervals shorter than the equilibration time. **D-H)** Hence, we fit the cross-sectional distribution of abundances, i.e, employment in a sector across all cities and abundance of all tree species in a height cluster. Data and fit for three sectors and two height clusters are shown. **I)** The percentile score, which quantifies the goodness of fit, shows that the likelihood of observing the data from the fitted distribution is similar to the likelihood of a random sample of the same size as the data. All fits had a percentile score $< 95\%$, as seen from the stacked histogram.

responding steady-state distribution, which is a gamma distribution described by two parameters (Eq. (3)). We plot the distribution of relative of microbial species and fit it with the two-parameter gamma distribution (black line), via maximum likelihood in Fig.4A,B,C.

For city and forest data, the observation timescale is longer than the equilibration timescale ($\tau/t_{obs} > 1$), and so the temporal trajectory of abundances will not converge to the steady-state distribution. Instead, we plot the cross-sectional distribution of abundances, i.e., the relative abundance of a sector in a city aggregated across all cities and the abundances of tree species within a height cluster, at a random time point (Fig. 4D-H). The fit with the gamma distribution is the black line.

To test whether the observed data could have been generated by the model, we repeat the procedure used in Fig. 2. Specifically, we compared the likelihood of the observed data to the likelihood of randomly generated samples from the fitted distribution. All data sets had a likelihood comparable to a random sample, as evidenced by the percentile scores (from 1000 random samples of the same size) shown in Fig.4I. Together, these observations indicate that the abundances in these systems are well described the gamma distribution. This conclusion is further supported by previous research that used a gamma distribution to successfully fit cross-sectional microbial abundances across microbiomes[8].

The steady-state distribution (SSD) and logarithmic fold-change distribution (LFD) predicted by the model (Eqs. (3), (4)) share a parameter, the ENR ($x_i^*/\sigma_i\tau_i$). We compare the maximum likelihood estimates of the ENR obtained by the separate fits of the empirical abundance distribution and LFD in Fig. 5. The discrepancy in inferred ENR values could arise from limitations of the data or the model.

To check whether the model is able to simultaneously fit both the LFD and abundance distribution, we performed a modified version of our goodness of fit test we used previously. First, using the ENR inferred from fitting the abundance, we compare the likelihood of fitting the LFD by P_{fluc} to the likelihood of 100 random samples from the fitted distribution of the same size. Then we fit the other way around, i.e., using the ENR inferred from fitting the LFD, we compare the likelihood fitting the abundance by P_{ss} to the likelihood of 1000 random samples from the fitted distribution of the same size. If the likelihood of the observed data was within 95% of the likelihood of the samples for either of these comparisons, then we conclude that the inferred parameters were consistent, i.e., the ENR obtained from fitting one distribution is able to provide a reasonable fit of the other distribution. The data points deemed consistent is depicted by filled markers in Fig. 5. If the data likelihood was smaller than 95% in both cases, we term the inferred parameters as inconsistent. The data points deemed inconsistent are shown as unfilled markers in Fig. 5.

The ENR estimates for the majority of the data were consistent (Fig. 5). A minority of the microbes (28/85)

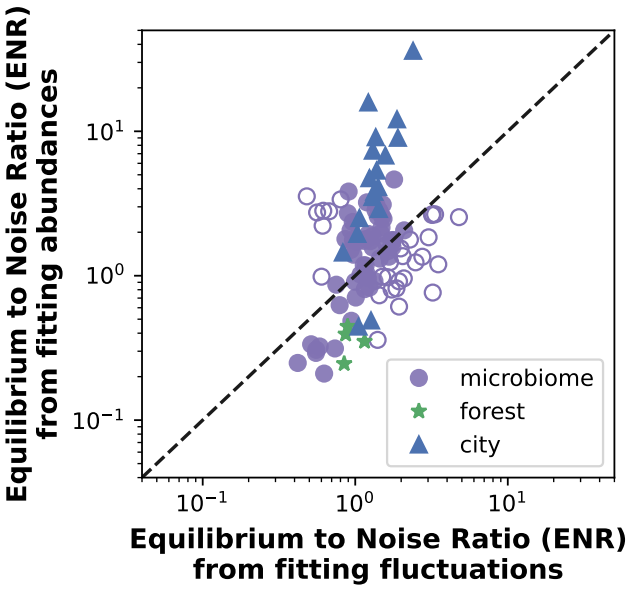

FIG. 5. **The consistency of inferred parameters from independent fits of the distributions of fluctuation and abundance.** Independent estimates of the Equilibrium to Noise Ratio (ENR= $x_i^*/\sigma_i\tau_i$) can be inferred from fitting the empirical Logarithmic Fold-change Distribution (LFD) or ftting the abundance distribution. If a fit of the LFD using the ENR value inferred from fitting the abundance passes the goodness of fit test, or vice-versa, we deem the inferred values of ENR as consistent. Filled points denote consistent estimates and unfilled points denote inconsistent data points. All city sectors and forest species were consistent, and 57 of 85 microbial species were consistent.

were rejected as being inconsistent. This could partly be due to temporal correlations in the data used to fit the abundance distribution, which we neglected. These correlations exist over timescales of τ_i and only vanish when $\Delta t \gg \tau_i$. For cities and forests, we used cross-sectional data, which does not suffer from this drawback. All employment data, including the apparently large outliers in Fig. 5, were termed consistent. The consistency of these large outliers was because although the abundance distribution of some sectors were well fit by large ENR values, substantially smaller ENR values also provided a reasonable fit and could not be rejected using the few hundred observations. Overall, the majority of the observed species/sectors in the three systems followed the expected relationship between the distribution of abundances and the distribution of fluctuations predicted by the model.

Variation of model parameters explains Taylor's law

In addition to providing a simple 2-parameter null models for fluctuations and abundances in complex populations, the SLRM can also help understand other empirical patterns in the data. Prior research in micro-

A

parameter	Coefficient of Variation
$x_i^*/\sigma_i\tau_i$	0.42
$\sigma_i\tau_i$	1.7
x_i^*	1.6
τ_i	2.1
σ_i	3.3
x_i^*/τ_i	2.6

B

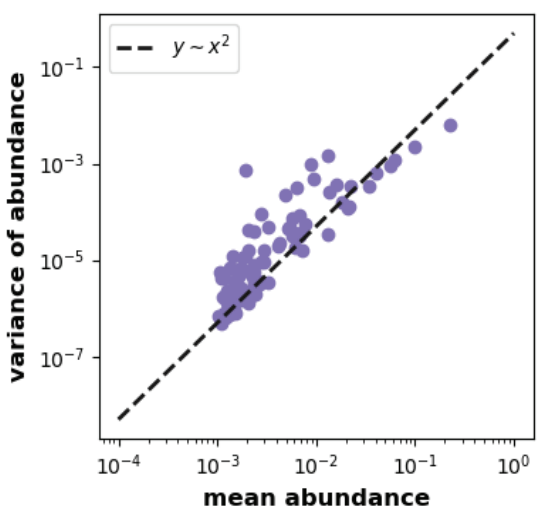

FIG. 6. **Power-law scaling of temporal mean and variance of abundance (Taylor's law) in microbiome due to approximate conservation of ENR. A)**The inferred ENR ($x_i^*/\sigma_i\tau_i$) in the microbiome data is relatively constant compared to other parameters, as evidenced by the tabulated coefficient of variation. **B)** The properties of the gamma distribution (which fits the abundance distribution) define the mean abundance of species i, as x_i^*, and the ratio of variance to mean squared as the ENR $x_i^*/\sigma_i\tau_i$. Thus, due to the approximate conservation of the ENR, we get the power-law scaling (with exponent 2) of variance with mean referred to as Taylor's law. The geometric mean of the ENR inferred from fitting the LFD and abundance distribution was used when estimating the variation of different parameter combinations.

biome data has revealed an approximate power-law scaling (with exponent 2) of the variance vs. mean of the abundance, called Taylor's law[7,8] (Fig. 6), which may arise due to various mechanisms[47–49]. Examination of inferred model parameters provides an alternative explanation for this empirical observation.

To investigate why Taylor's law arises, we compare the variation of the inferred model parameters across microbial species, using the coefficient of variation (Fig. 6A). The ENR ($x_i^*/\sigma_i\tau_i$) has a substantially lower coefficient of variation than the other parameters, and so can be considered to be approximately constant. By well-known properties of the Gamma distribution, the mean abundance of a species i is x_i^* and the ratio of variance to mean squared is the ENR, $x_i^*/\sigma_i\tau_i$. In microbiome data, the ENR is constant, and so the ratio of variance to mean squared remains fixed while the mean abundance varies, which leads to the observation of an approximate power-law scaling of the variance with mean. Hence, the approximate constancy of the ENR provides an alternative explanation for the observed power-law known as Taylor's law in microbiome data. The approximate constancy of ENR implies that the distribution of fluctuations (LFD) of different microbial species will be similar when time intervals are measured in terms of τ_i.

DISCUSSION

Understanding the behavior of diverse complex populations is challenging, but critical for progress in many fields. The unified approach to studying these populations adopted here complements detailed investigations of each specific population. By showing that the emergent distributions of population abundances and fluctuations in all three populations can be reproduced by a single model, our analysis highlighted emergent properties that are independent of system-specific details. Our analysis also revealed that when using generations to measure time, all three systems occupy a similar range of parameters. The difference in observation duration, measured in generations, is responsible for the apparent differences in fluctuations between systems. It suggests that fluctuations in cities and forests over decades and centuries would resemble the microbiome fluctuations more closely.

Importantly, our analysis provides a simple two-parameter functional form for the distribution of fold-changes in the populations and relates this to the abundance distribution. The predicted distribution of fluctuations P_{fluc} (Eq.(4)) is able to fit fluctuations in all three systems despite having only two-parameters. Further, P_{fluc} arises from a well-defined model and suggests a plausible mechanism; it is not simply chosen from the vast library of statistical distributions historically examined. This distribution can serve as an important null model for fluctuations in complex populations, where understanding the likelihood of large fluctuations is crucial. For employment fluctuations, large fluctuations impact urban planning and economic stability; for forests, large fluctuations impact ecological management strategies; for microbiome, large fluctuations can cause dysbiosis, which affects the health of the host[17,18,20]. A two-parameter

null model for fluctuations in complex populations is useful in practical, data-limited settings; it can estimate the risk of large fluctuations more accurately and improve quantitative methods (see SI Fig. S4).

Intuitively, the SLRM can be understood as the linearization of a more complex nonlinear model around its equilibrium when species interactions are neglected (see SI). This simplification drastically reduces the number of parameters, making parameter inference from available data feasible. Deviations from the model predictions, could indicate the presence of species interactions, which are often modeled by Lotka-Volterra, consumer-resource and other models of higher order species interactions[1,2,50], or other mechanisms. Such models could potentially be parameterized by using specialized methods with additional data[23,38].

Due to its simplicity, the Stochastic Linear Response Model (SLRM) can serve as a useful base model for further research. It can be easily augmented with additional mechanisms, such as environmental fluctuations or species interactions, which could be tested with additional data (see SI). Other modifications could help understand evolving populations. For the timescales examined, we assumed that the equilibrium value x^* remains constant. Over longer timescales, however, the equilibrium value could change due to biological evolution, climate change, or socio-technological revolution. Investigating the model when x^* changes in time could help understand emergent dynamics in complex populations over evolutionary timescales and presents an interesting direction for future research.

To butcher two well-worn phrases, all models are wrong, some are useful, and some are unreasonably effective. We believe the SLRM falls into the latter two categories, and that its surprising effectiveness across such diverse datasets points to something universal about the way complex populations fluctuate. The SLRM also provides valuable two parameter null models for the distributions of abundances and fluctuations in complex populations, which are of particularly utility in data-limited scenarios for forecasting and risk-analysis. The unified analysis of the three systems highlights the similarities and differences between the systems and paves the way for cross-pollination of tools and techniques between difference fields.

METHODS

Data processing

1. City data

City data was obtained from Quarterly Census of Employment and Wages from the U.S. Bureau of Labor Statistics (https://www.bls.gov/cew/). The data provides the employment classified into industrial sectors by the North American Industry Classification System (NAICS) at county level in the U.S. (see SI Table S1). We aggregated data at the county level to 383 Metropolitan Statistical Areas (MSA), which we call cities. MSAs are independent statistical units defined by the Census Bureau that encompass a central city and the geographical areas connected to the city. We obtained the list of counties in each MSA in 2017 from U.S. Census Bureau, County Business Patterns program, and used this to calculate employment at MSA level. This approach allowed us to maintain a consistent definition of MSAs across the entire time-series. The employment data is recorded at monthly intervals from 2003 to 2019. Since many industries, such as agriculture and accommodation, display seasonal trends in employment withing a year, we used a Δt of 1 year for calculating the empirical LFD. We plot and fit data with only non-zero abundance.

For privacy reasons, sectoral employment data at some points are suppressed, and so we remove these points from our analysis. Since this suppression increases at finer levels of NAICS classification, we analyzed sectors classified at the two digit level. This also made our analysis robust to changes in the NAICS classification scheme that impacted the temporal continuity of the data at finer resolution. We used 18 of the 21 NAICS categories at the two-digit level (see SI Table S1). We removed three categories: '81' (other) ,'92' (public administration) and '99' (unclassified). Public administration and government employment was removed because much of this sector is not reported due to governmental regulations. Relative sectoral employment was calculated for each city by dividing sectoral employment in the city with the total employment reported in the city. Note that due to data suppression and removal of three naics categories, the sum of the relative employment in a city over all sectors we analyze need not equal one. We assume that each sectors is described by a set of SLRM parameters that is independent of the city; this provides us with enough data to fit the SLRM predictions.

2. Microbiome data

Microbiome data from Ref.[28] was obtained and processed as in Ref.[7]. We consider only the gut microbiome data for individual M3 since it was substantially longer than other time-series. The data was collected in time intervals of one day with some gaps in sampling. There were a total of 336 time-points recorded.

To process the data, first, read counts at each time point were normalized to obtain relative abundance. Only prevalent or relatively abundant species were used for our analysis. Specifically, we considered species only if they were present in more than half of the time points and their average abundance was greater than 10^{-3}. 85 species that met this criterion. To calculate the empirical LFD, we used abundances that were collected $\Delta t = 1$ days apart. We plot and fit data with non-zero relative abundance. Each species has its own set of SLRM pa-

rameters.

3. Forest data

The Barro-Colorado Island forest data was obtained from the Center for Tropical Forest Science website (https://forestgeo.si.edu). Abundance data collected at 5 year intervals, in years 1990, 1995, 2000, and 2005 was used. Only trees that were alive and had a diameter at breast height > 10 cm were counted. The trees were grouped into four height clusters: shrubs, understory treelets, midstory trees, and canopy trees based on Ref.[29]. There were 87, 75, 60, and 63 species in the four height clusters. Relative abundance of a species in a particular height cluster was calculated as the absolute abundance of the trees of the species at that time point divided by the absolute abundance of all tree trees within that specific height cluster. We assume that all species within a height cluster and thus being highly similar in trait values, are described the same set of SLRM parameters; this provides us the data required to fit the data with the model predictions. Different height clusters are fit separately, like different employment sectors. Time interval for calculating LFD, Δt, was 5 years.

Fitting and sampling procedures

The data was fit and parameters were estimated by Maximum Likelihood Estimation from the Scipy package in Python. In addition to the probability distribution P_{fluc}, the corresponding cumulative distribution (see SI) was defined to make sampling from the distribution more efficient.

In the city data set, we had substantially more data points in the each LFD ($\sim 30,000$ points) than for the empirical abundance distribution (~ 300 points). This made performing the goodness of fit test computationally harder for the LFD than the abundance distribution, since the test required generationg samples from the distribution of the same size as the data. Hence for each fit of the abundance distribution, we obtained 1000 samples to compare with the data likelihood, whereas for each fit of the LFD, we obtained only 100 samples.

The coefficient of variation of tabulated in Fig. 6A, used inferred parameters values of ENR, $\sigma_i\tau_i$, and τ_i to construct the various parameter combinations. τ_i was estimated from fitting the LFD, $\sigma_i\tau_i$ was estimated from fitting the abundance distribution, and the ENR was the geometric mean of the estimates from the LFD and abundance distribution.

Estimating generation times in each system

The generation times used for scaling the inferred time scale in Fig. 3 were calculated based on estimates in the literature. Since a direct estimate of generation time of microbes in the human gut is unavailable, we used the measured generation time in mice on a regular diet of 5.7 divisions per day(4.2 hours)[41]. For trees, the generation time used was 55.5 years[42]. For urban employment, we measured generation time as the time required for people to change jobs between industrial sectors. Using the median job duration in US of 5 years[43] and the fact that roughly half of the job changes are between sectors[44], we obtained a generation time of 10 years. The estimated generation times neglect possible variation between species and sector due to data limitations.

Environmental stochasticity produces normally distributed fluctuations

The dynamics of a population driven purely by environmental noise is given by

$$\frac{dx_i}{dt} = E_i x\, \eta_i(t), \tag{5}$$

where E_i quantifies the strength of environmental noise and η_i is delta correlated gaussian noise[34,45]. We can rewrite this equation for $\log x_i$ instead as

$$\frac{d\log x_i}{dt} = E_i\, \eta_i(t). \tag{6}$$

Clearly, the fluctuations of $\log x_i$ are now gaussian distributed. Thus environmental noise can produce a normally distributed LFD. Note, however, that this equation does not have a steady-state. Additional terms are required in the dynamical equation to stabilize the population and ensure a steady-state. One such case, of a logistic equation with environmental noise has been examined in the literature[8]. Interestingly, this model also has a gamma distributed steady-state distribution. However, no analytical solution for the distribution of fluctuations exists because of the nonlinear nature of the equation.

Although a subset of microbes have an LFD that is well-explained by a normal distribution, we are able to fit the majority of the data without using environmental noise. While adding environmental noise would affect model behavior, the success of the model without environmental noise suggests that the populations could be in a parameter-range where the effect of environmental noise is unimportant or that the quantities we examine are not sensitive to the addition of environmental noise on top of demographic noise.

Acknowledgements: The authors would like to thank Zachary Miller, Alice Doucet Beaupré, and members of the O'Dwyer group for helpful feedback and comments. The authors acknowledge funding support from Simons Foundation Grant #376199 and McDonnell Foundation Grant #220020439 to J.O.D.

[1] MacArthur, R. Species packing and competitive equilibrium for many species. *Theoretical Population Biology* **1**, 1–11 (1970). URL https://linkinghub.elsevier.com/retrieve/pii/0040580970900390.

[2] Goldford, J. E. *et al.* Emergent simplicity in microbial community assembly. *Science* **361**, 469–474 (2018). URL http://www.sciencemag.org/lookup/doi/10.1126/science.aat1168.

[3] Hubbell, S. P. *The unified neutral theory of biodiversity and biogeography*. No. 32 in Monographs in population biology (Princeton University Press, Princeton, 2001).

[4] Simini, F., González, M. C., Maritan, A. & Barabási, A.-L. A universal model for mobility and migration patterns. *Nature* **484**, 96–100 (2012). URL http://www.nature.com/articles/nature10856.

[5] Gabaix, X. Zipf's Law for Cities: An Explanation. *The Quarterly Journal of Economics* **114**, 739–767 (1999). URL https://academic.oup.com/qje/article-lookup/doi/10.1162/003355399556133.

[6] Acs, Z. J. MEASURES OF JOB FLOW DYNAMICS IN THE U.S. *Economic Studies* 47 (1999).

[7] Ji, B. W., Sheth, R. U., Dixit, P. D., Tchourine, K. & Vitkup, D. Macroecological dynamics of gut microbiota. *Nature Microbiology* **5**, 768–775 (2020). URL http://www.nature.com/articles/s41564-020-0685-1.

[8] Grilli, J. Macroecological laws describe variation and diversity in microbial communities. *Nature Communications* **11**, 4743 (2020). URL http://www.nature.com/articles/s41467-020-18529-y.

[9] Bettencourt, L. M. A., Samaniego, H. & Youn, H. Professional diversity and the productivity of cities. *Scientific Reports* **4**, 5393 (2015). URL http://www.nature.com/articles/srep05393.

[10] Hong, I., Frank, M. R., Rahwan, I., Jung, W.-S. & Youn, H. The universal pathway to innovative urban economies. *Science Advances* **6**, eaba4934 (2020). URL https://advances.sciencemag.org/lookup/doi/10.1126/sciadv.aba4934.

[11] Uhlenbeck, G. E. & Ornstein, L. S. On the Theory of the Brownian Motion. *Physical Review* **36**, 823–841 (1930). URL https://link.aps.org/doi/10.1103/PhysRev.36.823.

[12] Stein, E. M. & Stein, J. C. Stock Price Distributions with Stochastic Volatility: An Analytic Approach. *Review of Financial Studies* **4**, 727–752 (1991). URL https://academic.oup.com/rfs/article-lookup/doi/10.1093/rfs/4.4.727.

[13] Vasicek, O. An equilibrium characterization of the term structure. *Journal of Financial Economics* **5**, 177–188 (1977). URL https://linkinghub.elsevier.com/retrieve/pii/0304405X77900162.

[14] Verbavatz, V. & Barthelemy, M. The growth equation of cities. *Nature* **587**, 397–401 (2020). URL http://www.nature.com/articles/s41586-020-2900-x.

[15] Castellano, C., Fortunato, S. & Loreto, V. Statistical physics of social dynamics. *Reviews of Modern Physics* **81**, 591–646 (2009). URL https://link.aps.org/doi/10.1103/RevModPhys.81.591.

[16] Mantegna, R. N. & Stanley, H. E. *An introduction to econophysics: correlations and complexity in finance* (Cambridge Univ. Press, Cambridge, 2007), digitally printed version edn.

[17] Vandeputte, D. *et al.* Temporal variability in quantitative human gut microbiome profiles and implications for clinical research. *Nature Communications* **12**, 6740 (2021). URL https://www.nature.com/articles/s41467-021-27098-7.

[18] Clark, T. E. Employment Fluctuations in U.S. Regions and Industries: The Roles of National, Region-Specific, and Industry-Specific Shocks. *Journal of Labor Economics* **16**, 202–229 (1998). URL https://www.journals.uchicago.edu/doi/10.1086/209887.

[19] Dai, L., Vorselen, D., Korolev, K. S. & Gore, J. Generic Indicators for Loss of Resilience Before a Tipping Point Leading to Population Collapse. *Science* **336**, 1175–1177 (2012). URL https://www.science.org/doi/10.1126/science.1219805.

[20] Holling, C. S. Resilience and Stability of Ecological Systems. *Annual Review of Ecology and Systematics* 24 (1973).

[21] West, M. & Harrison, J. *Bayesian forecasting and dynamic models*. Springer series in statistics (Springer, New York, 1997), 2nd ed edn.

[22] Rosenberg, K. V. *et al.* Decline of the North American avifauna. *Science* **366**, 120–124 (2019). URL https://www.science.org/doi/10.1126/science.aaw1313.

[23] Cao, H.-T., Gibson, T. E., Bashan, A. & Liu, Y.-Y. Inferring human microbial dynamics from temporal metagenomics data: Pitfalls and lessons. *BioEssays* **39**, 1600188 (2017). URL http://doi.wiley.com/10.1002/bies.201600188.

[24] George, A. B. & Korolev, K. S. Ecological landscapes guide the assembly of optimal microbial communities. preprint, Ecology (2021). URL http://biorxiv.org/lookup/doi/10.1101/2021.12.14.472701.

[25] Samuels, C. L. & Drake, J. A. Divergent perspectives on community convergence. *Trends in Ecology & Evolution* **12**, 427–432 (1997). URL https://linkinghub.elsevier.com/retrieve/pii/S0169534797011828.

[26] Fujita, M. *Urban economic theory: land use and city size* (Cambridge University Press, Cambridge [Cambridgeshire] ; New York, 1989).

[27] Henderson, J. V. The Sizes and Types of Cities. *The American Economic Review* **64**, 18 (1974).

[28] Caporaso, J. G. *et al.* Moving pictures of the human microbiome. *Genome Biology* **12**, R50 (2011). URL http://genomebiology.biomedcentral.com/articles/10.1186/gb-2011-12-5-r50.

[29] D'Andrea, R. *et al.* Counting niches: Abundance-by-trait patterns reveal niche partitioning in a Neotropical forest. *Ecology* **101** (2020). URL https://onlinelibrary.wiley.com/doi/10.1002/ecy.3019.

[30] Horvath, M. Sectoral shocks and aggregate fluctuations. *Journal of Monetary Economics* 38 (2000).

[31] Mangan, S. A. *et al.* Negative plant–soil feedback predicts tree-species relative abundance in a tropical forest. *Nature* **466**, 752–755 (2010). URL http://www.nature.com/articles/nature09273.

[32] Sardans, J. *et al.* Empirical support for the biogeochemical niche hypothesis in forest trees. *Nature Ecology & Evolution* **5**, 184–194 (2021). URL http://www.nature.com/articles/s41559-020-01348-1.

[33] Donado, F., Moctezuma, R. E., López-Flores, L., Medina-Noyola, M. & Arauz-Lara, J. L. Brownian motion in non-equilibrium systems and the Ornstein-Uhlenbeck stochastic process. *Scientific Reports* **7**, 12614 (2017). URL http://www.nature.com/articles/s41598-017-12737-1.

[34] Lande, R., Engen, S. & Sæther, B.-E. *Stochastic population dynamics in ecology and conservation*. Oxford series in ecology and evolution (Oxford University Press, Oxford ; New York, 2003).

[35] Azaele, S., Pigolotti, S., Banavar, J. R. & Maritan, A. Dynamical evolution of ecosystems. *Nature* **444**, 926–928 (2006). URL http://www.nature.com/articles/nature05320.

[36] Feller, W. Two Singular Diffusion Problems. *The Annals of Mathematics* **54**, 173 (1951). URL https://www.jstor.org/stable/1969318?origin=crossref.

[37] Cox, J. C., Ingersoll, J. E. & Ross, S. A. A Theory of the Term Structure of Interest Rates. *Econometrica* **53**, 385 (1985). URL https://www.jstor.org/stable/1911242?origin=crossref.

[38] Fisher, C. K. & Mehta, P. Identifying Keystone Species in the Human Gut Microbiome from Metagenomic Timeseries Using Sparse Linear Regression. *PLoS ONE* **9**, e102451 (2014). URL https://dx.plos.org/10.1371/journal.pone.0102451.

[39] Akaike, H. Information Theory and an Extension of the Maximum Likelihood Principle. In Parzen, E., Tanabe, K. & Kitagawa, G. (eds.) *Selected Papers of Hirotugu Akaike*, 199–213 (Springer New York, New York, NY, 1998). Series Title: Springer Series in Statistics.

[40] Stanley, M. H. R. *et al.* Scaling behaviour in the growth of companies. *Nature* **379**, 804–806 (1996). URL http://www.nature.com/articles/379804a0.

[41] Gibbons, R. J. & Kapsimalis, B. Estimates of the overall rate of growth of the intestinal microflora of hamsters, guinea pigs, and mice. *Journal of Bacteriology* **93**, 510–512 (1967). URL https://journals.asm.org/doi/10.1128/jb.93.1.510-512.1967.

[42] Lewis, S. L. *et al.* Tropical forest tree mortality, recruitment and turnover rates: calculation, interpretation and comparison when census intervals vary. *Journal of Ecology* **92**, 929–944 (2004). URL https://onlinelibrary.wiley.com/doi/10.1111/j.0022-0477.2004.00923.x.

[43] U.S. Bureau of Labor Statistics News Release: Employee Tenure in 2020. Tech. Rep. USDL-20-1791 (2021). URL https://www.bls.gov/news.release/pdf/tenure.pdf.

[44] Hyatt, H. R. & McEntarfer, E. Job-to-Job Flows and the Business Cycle. *SSRN Electronic Journal* (2012). URL http://www.ssrn.com/abstract=2017672.

[45] Fung, T., O'Dwyer, J. P., Rahman, K. A., Fletcher, C. D. & Chisholm, R. A. Reproducing static and dynamic biodiversity patterns in tropical forests: the critical role of environmental variance. *Ecology* **97**, 1207–1217 (2016). URL http://doi.wiley.com/10.1890/15-0984.1.

[46] Kalyuzhny, M., Kadmon, R. & Shnerb, N. M. A neutral theory with environmental stochasticity explains static and dynamic properties of ecological communities. *Ecology Letters* **18**, 572–580 (2015). URL https://onlinelibrary.wiley.com/doi/10.1111/ele.12439.

[47] Cohen, J. E. & Xu, M. Random sampling of skewed distributions implies Taylor's power law of fluctuation scaling. *Proceedings of the National Academy of Sciences* **112**, 7749–7754 (2015). URL https://pnas.org/doi/full/10.1073/pnas.1503824112.

[48] Giometto, A., Formentin, M., Rinaldo, A., Cohen, J. E. & Maritan, A. Sample and population exponents of generalized Taylor's law. *Proceedings of the National Academy of Sciences* **112**, 7755–7760 (2015). URL https://pnas.org/doi/full/10.1073/pnas.1505882112.

[49] Fronczak, A. & Fronczak, P. Origins of Taylor's power law for fluctuation scaling in complex systems. *Physical Review E* **81**, 066112 (2010). URL https://link.aps.org/doi/10.1103/PhysRevE.81.066112.

[50] Grilli, J., Barabás, G., Michalska-Smith, M. J. & Allesina, S. Higher-order interactions stabilize dynamics in competitive network models. *Nature* **548**, 210–213 (2017). URL http://www.nature.com/articles/nature23273.