Coalescent dynamics of planktonic communities

Paula Villa Martín,[*] Anzhelika Koldaeva,[*] and Simone Pigolotti[†]
Biological Complexity Unit, Okinawa Institute of Science and Technology Graduate University, Onna, Okinawa 904-0495, Japan

(Received 7 April 2022; accepted 19 September 2022; published 17 October 2022)

Planktonic communities are extremely diverse and include a vast number of rare species. The dynamics of these rare species is best described by individual-based models. However, individual-based approaches to planktonic diversity face substantial difficulties, due to the large number of individuals required to make realistic predictions. In this paper, we study the diversity of planktonic communities by means of a spatial coalescence model that incorporates transport by oceanic currents. As a main advantage, our approach requires simulating a number of individuals equal to the size of the sample one is interested in, rather than the size of the entire community. By theoretical analysis and simulations, we explore the conditions upon which our coalescence model is equivalent to individual-based dynamics. As an application, we use our model to predict the impact of chaotic advection by oceanic currents on biodiversity. We conclude that the coalescent approach permits one to simulate marine microbial communities much more efficiently than with individual-based models.

DOI: 10.1103/PhysRevE.106.044408

I. INTRODUCTION

Ecological communities are made up of a large number of species. Their diversity varies in space and time as a result of the ecological forces they are subject to [1]. In diverse ecological communities, one typically encounters a few very abundant species and many rare species, some of which are represented by just a few individuals. Because of these rare species, the diversity of ecological communities is best described by spatially explicit individual-based models (IBMs) [2,3], rather than by models based on species concentration or densities. IBMs have contributed to rationalizing fundamental biodiversity patterns in terrestrial ecosystems, such as the scaling of the average number of encountered species with the size of the sampled area [2–6].

In prototypical spatially explicit IBMs such as the multispecies voter model [3,4,6,7], individuals are placed on a two-dimensional lattice and stochastically reproduce, die, and disperse. Even simple spatial IBMs are challenging to solve analytically [3]. A significant advancement in their understanding originates in the concept of *duality* [4,7]. In this context, "duality" is a mapping between the IBM and a different model, whose dynamics proceeds backward in time. For this reason, we often refer to the original model as the "forward" model and its dual as the "backward" model. The backward model considers a sample of the population of the forward model at a very long time and seeks to reconstruct its species composition. Individuals in the sample are represented as particles that evolve backward in time. If two particles happen to be on the same site, they can coalesce, signaling that the two corresponding individuals have a common ancestor and are, therefore, conspecific. Due to these events, the backward model is also termed the "coalescence" model. By tracking coalescence events, the backward dynamics reconstructs the species composition of the original sample.

In short, duality maps a spatial IBM into a system of coalescing random walkers. The advantage of this mapping is twofold. First, mathematical results have been obtained for systems of coalescing random walkers [8,9], leading to exact predictions of biodiversity patterns [3]. Second, backward models are much more efficient than forward models to simulate on a computer [2,3,5,6].

IBMs have also been used to study microbial planktonic communities [10–14]. These models have been used to predict, for example, how fluid flows affect the fate of mutants characterized by a reproductive advantage [13–17] or by a different diffusivity [18–21]. However, when used to predict biodiversity patterns, these models face severe computational limitations. Even state-of-the-art approaches are usually limited to communities of tens of thousands of individuals. For comparison, a liter of oceanic water can contain tens or hundreds of millions of planktonic cells [22]. To encompass this problem, it has been suggested that each individual in a IBM can be considered as a representative of an entire subpopulation [23]. It is, however, unclear whether this interpretation can account for the dynamics of very rare species.

With this motivation in mind, we recently proposed a coalescence model for the dynamics of microbial planktonic communities [24], which encompasses the limitation of IBMs. In this paper, we study the dynamics of this coalescence model and argue that, under certain conditions, it is dual to an IBM. We support this mathematical prediction with numerical simulations of both models. We then apply our model to understand observational metabarcoding data from protist communities [24].

The paper is organized as follows. In Sec. II, we introduce the (forward) IBM and the (backward) coalescence model. We prove that, in the weak-noise limit, these two models are dual. In Sec. III, we briefly introduce the main observables that are commonly employed in ecology to quantify biodiversity. In

[*] These authors contributed equally to this work.
[†] simone.pigolotti@oist.jp

Sec. IV, we test our theory by extensive numerical simulations of the forward and backward models, both in the presence and in the absence of chaotic advection. We also compare the model prediction with observational data. Section V is devoted to conclusions and perspectives.

II. MODELS

In this section, we introduce two approaches for modeling the diversity of planktonic populations.

The first approach is via an IBM, in which an initial population stochastically evolves forward in time as a result of reproduction events, competition among individuals, advection-diffusion of individuals in space, and speciation, i.e., events that give rise to new species. The forward model can be seen as the multispecies version of a two-species competition model [13,14] that does not include speciation. The second approach is based on a backward (coalescence) model [24]. The backward model considers a sample of N_s individuals and seeks to reconstruct its species composition by tracing the ancestry of the individuals backwards in time.

We conclude the section by defining two regimes (weak and strong noise) characterizing the forward model. We then demonstrate that duality between the forward and the backward models rigorously holds in the weak-noise regime.

A. Forward model

A microbial population inhabits a two-dimensional square area $A = L \times L$, representing an aquatic environment. Initially, individuals are homogeneously distributed. They can belong to different species, but for simplicity we neglect their species identity for the time being. Individuals can stochastically die, reproduce, and displace. As a result of these events, the total number of individuals $N(t)$ fluctuates over time. Each individual asexually reproduces at rate λ. When a reproduction occurs, the daughter individual is placed at a random position in a square of side l centered on the mother position. From now on, we refer to this square as the "neighborhood" of an individual. Individuals die in a density-dependent way with rate $\lambda \hat{n}$, where $\hat{n}$ is the number of other individuals in their neighborhood. The dependence of the death rate on the local density represents competition. In this respect, the length scale l can be interpreted as the characteristic distance below which individuals are in direct competition with each other (see Ref. [24]).

Each individual moves in space according to the advection-diffusion equations

$$\begin{aligned} \frac{d}{dt}x &= v_x(x, y, t) + \sqrt{2D}\xi_x(t), \\ \frac{d}{dt}y &= v_y(x, y, t) + \sqrt{2D}\xi_y(t), \end{aligned} \tag{1}$$

where x and y are the coordinates of the given individual. The terms proportional to $\sqrt{2D}$ represent effective diffusion. Besides molecular diffusion, this term can incorporate the effect of flows at length scales shorter than those resolved by the fluid flow [24]. The functions $\xi_x(t)$ and $\xi_y(t)$ are white noise sources satisfying $\langle \xi_i(t) \rangle = 0$, $\langle \xi_i(t)\xi_j(t') \rangle = \delta_{ij}\delta(t - t')$, where $i \in \{x, y\}$ and $\langle \cdots \rangle$ denotes an average over realizations. For the time being, we impose periodic boundary conditions. The functions $v_x(x, y, t)$ and $v_y(x, y, t)$ represent an advecting fluid flow. In this paper we only consider incompressible velocity fields: $\vec{\nabla} \cdot \vec{v} = 0$, where $\vec{v} = (v_x, v_y)$, and $\vec{\nabla} = (\partial/\partial x, \partial/\partial y)$. Indeed, two-dimensional velocity fields in the oceans are incompressible to a very good approximation, although in some local regions vertical currents can alter this picture [25], potentially impacting population dynamics [13,26,27].

In the absence of birth and death dynamics and thanks to incompressibility, Eqs. (1) predict a homogeneous stationary distribution of individuals. We tentatively assume that this distribution remains homogeneous in the presence of birth and death processes. We call N_0 the average population size under this hypothesis. By imposing that birth events statistically balance death events, we find that

$$N_0 = L^2/l^2, \tag{2}$$

i.e., the average population size is equal to the ratio between the system area and the area of the interaction neighborhood (see Appendix A). In practice, this means that each interaction neighborhood contains one individual, on average. In the following, we take this value as the initial population size, $N(0) = N_0$.

We want to understand whether the assumption of homogeneous density holds. To this aim, we study a macroscopic description of our IBM (see Appendix B). We find that the populations remain homogeneous when the stochastic fluctuations induced by birth and death processes are relatively small. In this case, the average number of individuals does not significantly deviate from N_0.

In two dimensions, the relative strength of fluctuations is controlled by the dimensionless parameter

$$\tilde{D} = \frac{DN_0}{\lambda L^2} \tag{3}$$

(see Ref. [14]). For $\tilde{D} \gg 1$, stochastic fluctuations are small. In contrast, for $\tilde{D} \ll 1$, fluctuations dominate the dynamics. In the following, we refer to the $\tilde{D} > 1$ and $\tilde{D} < 1$ cases as the "weak-noise" regime and the "strong-noise" regime, respectively.

Simulations of the model confirm that, in the strong-noise regime, the average number of individuals significantly differ from N_0 (see Fig. 1). This means that our assumption that the birth-death dynamics does not affect the average number of individuals breaks down.

In particular, in the absence of advection, our simulations show that the average number of individuals exceeds N_0 [see Fig. 1(a)]. This result contrasts with that in Ref. [14], where a reduction of the average number of individuals was observed in the strong-noise regime. In general, in the strong-noise regime, one can expect that the results depend on details of the microscopic rules. In our particular case, this discrepancy could be caused by a difference in the way birth is

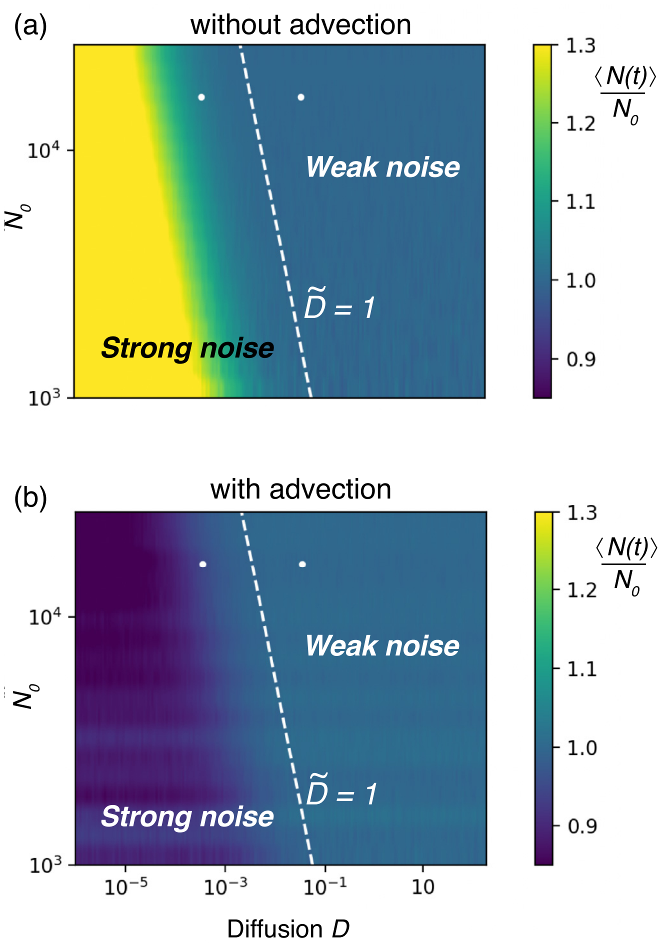

FIG. 1. Normalized average number of individuals $\langle N(t)\rangle/N_0$ in (a) the absence and (b) the presence of advection. The stationary population size N_0 is varied by tuning the neighborhood linear size l. The dashed line marks the theoretical condition $\tilde{D} = 1$ separating the weak-noise regime from the strong-noise regime. In all simulations, we fixed $L = \lambda = 1$. In both panels, two dots at $\tilde{D} = 0.1$ and $\tilde{D} = 10$ (with $N_0 = 16384$) mark the set of parameters chosen for our further analysis.

implemented in the two models: here, the daughter cell is placed at a random position in the neighborhood of her mother, whereas in Ref. [14] the daughter cell is placed at the same position as the mother.

In contrast, in the presence of advection, the average population size decreases in the strong-noise regime [see Fig. 1(b)]. A qualitative explanation is that advection effectively precludes individuals to visit some regions, thereby increasing effective competition. Our simulations show that the transition from the weak-noise regime to the strong-noise regime occurs for $\tilde{D} \approx 1$ in the presence of advection as well, suggesting that the velocity field does not play a dominant role in determining the boundary between these two regimes. However, this fact might be due to our choice of velocity field and might not hold in general.

Hereafter, we fix $\lambda = 1$, $L = 7.5$, and $N_0 = 16384$. The value of l corresponding to these choices is obtained from Eq. (2).

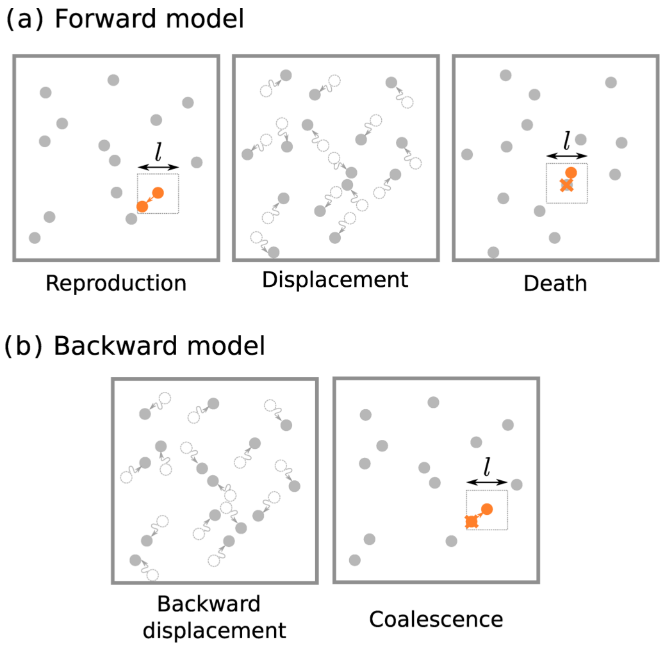

FIG. 2. Events in the forward and backward models. (a) In the forward model, individuals reproduce at rate λ within a neighborhood, displace, and die with rate $\lambda\,\hat{n}$, where $\hat{n}$ is the number of other individuals in their square neighborhood of linear size l. (b) In the backward model individuals displace and coalesce with rate $\bar{\lambda}$ with individuals within their neighborhood. In the weak-noise regime, the two models are dual under the condition $\lambda = \bar{\lambda}$.

B. Backward model and duality

The backward model (or coalescence model) describes the dynamics of N_s Lagrangian tracers. These tracers represent a sample of individuals from a final population, i.e., from a population that evolved according to the forward model at a time $t_f \gg 0$. We assume that the sample is homogeneously distributed in a sample area $L_s \times L_s$, with $L_s \leqslant L$. We trace back in time the evolution of the tracers in the sample. The coordinates of the tracers evolve according to Eq. (1), which we integrate with negative time increments.

While evolving back in time the coordinate of a given tracer, we might reach the time at which the individual represented by the tracer was born. From that instant, the tracer represents the position of the mother of the chosen individual. This means that, at a given time, each tracer represents either an individual in the final population or one of its ancestors. If the mother of the chosen individual (or one of her ancestors) is alive in the final population, this implies that the two tracers must be in the neighborhood of each other at the time in which the birth event occurs. If this is the case, we say that a "coalescence" has occurred, and the two tracers are merged into one.

The events occurring in the forward and backward models are summarized in Figs. 2(a) and 2(b), respectively.

To fully specify the coalescence model, we need to determine the rate $\bar{\lambda}$ at which two individuals coalesce if they are in the neighborhood of each other. We do so by considering a situation in which $N_s = N_0$, i.e., the number of tracers is equal to the average number of individuals in the forward model. In this scenario we are tracing all individuals; therefore, the total rate of birth λN_0 in the forward model must match the total rate of coalescence in the backward model. This latter rate is equal to $\bar{\lambda} N_s$ by the same argument made in Appendix A. Imposing that the two total rates must be equal and using $N_s = N_0$ leads to $\bar{\lambda} = \lambda$; i.e., the coalescence rate for individuals in the same neighborhood should match their birth rate in the forward model.

We remark that this argument relies on the assumption that the dynamics of the forward model is in the weak-noise regime. In the strong-noise regime, some of the assumptions underlying duality do not hold. First, we cannot assume that a sample of individuals in the final population is homogeneously distributed. Second, since in the strong-noise regime $\langle N(t)\rangle \neq N_0$, we cannot easily draw a correspondence between the rates of the forward model and those of the backward model.

C. Speciation and multispecies dynamics

We now add to the forward and backward models the notion of species identity. In the forward model, we assume that a newborn individual belongs to a new species with probability μ. In ecological terms, the probability μ can be interpreted either as a speciation probability or as a probability for individuals to be replaced by individuals belonging to new species immigrating from outside the community. In the well-mixed case, this process is known as the infinite-allele model of population genetics [28]. If the model is used to interpret metabarcoding studies, μ represents the probability for individuals to accumulate sufficient mutations to be considered as a new operational taxonomic unit (see discussion in Sec. IV E). When simulating the forward model, one can keep track of species identity during the dynamics and assign each newborn individual to a new species with probability μ. We, however, adopt an equivalent, but more efficient strategy [see Refs. [3,29]). We simulate the dynamics without keeping track of species identity, until all individuals descend from a single individual in the original population [see Fig. 3(a)]. The collection of descendants of this individual constitutes the ancestry tree. Individuals that do not belong to the ancestry tree [light gray in Fig. 3(a)] do not affect the diversity of the final population and can therefore be ignored. A further simplification is to remove individuals that have only one descendant in the ancestry tree [yellow in Fig. 3(a)]. This can be done by keeping track of the number of duplications d_i occurred in each branch i separating the remaining individuals [see Fig. 3(b)]. Since at each duplication event the probability of a speciation is equal to μ, the total probability that at least a speciation has occurred in a branch i is equal to $p_i = 1 - (1-\mu)^{d_i}$. In this way, we assign speciation events *a posteriori* by drawing from the probabilities that a speciation has occurred in each branch [see Fig. 3(c)].

In the backward model, speciation events occur continuously at a stochastic rate $\mu\lambda dt$. The reason is that, in the backward model, we do not consider individual birth events, unless they lead to coalescence. As anticipated, the advantage of the coalescence model is that it is possible to reconstruct the identity of a sample of N_s individuals in the final population being a subset of the total population. The corresponding ancestry tree is a subset of the corresponding tree in the forward model which includes all ancestors of the individuals in the sample up to their most recent common ancestor [see Fig. 3(d)]. In practice, speciation events can be assigned on the branches of this tree in a way similar to that for the forward model. However, this is not necessarily an efficient procedure in this case. The reason is that, for large system size and in the presence of advection, the time it takes for all tracers to coalesce can be exceedingly long. Instead, it is enough to track tracers backward in time until their first speciation event, as their preceding history does affect the diversity of the final sample. For this reason, in the backward dynamics we eliminate tracers from the system as soon as they speciate for the first time.

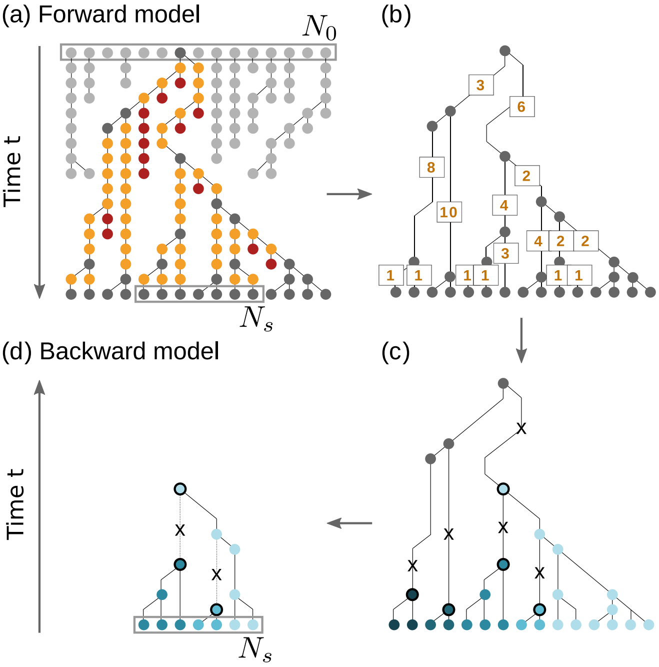

FIG. 3. Ancestry trees. (a) Dynamics of the forward model represented as a tree. An initial set of N_0 individuals reproduce and die over time until all remaining individuals share the same common ancestor. Individuals that die before the final time (dark red dots) are not relevant for the final population and are therefore removed. Individuals that do not have a descendant in the final population (light gray dots) are removed as well. (b) Individuals that have one direct descendant in the remaining tree (orange dots) are removed; their numbers are saved as variables d_i for each branch i. (c) Speciation events (marked with an X) are introduced at each tree branch i with a probability $p_i = 1 - (1-\mu)^{d_i}$ (see Ref. [3]). (d) Ancestry tree of the backward model. In this case, we consider a sample of N_s individuals from the final population and reconstruct its ancestry. Individuals can coalesce or speciate until one individual remains. This approach directly ignores branches corresponding to individuals that do not belong to the sample.

We compare predictions of the two models for a sample of N_s individuals taken at a final large time t_f. We fix $N_s < N_0$. In the absence of advection, since we might have $\langle N(t)\rangle \gtrsim N_0$, we fix $N_s = 16\,000$. In the presence of advection, population sizes are $\langle N(t)\rangle \lesssim N_0$, and we consider $N_s = 8192$ to ensure $N_s < N_0$.

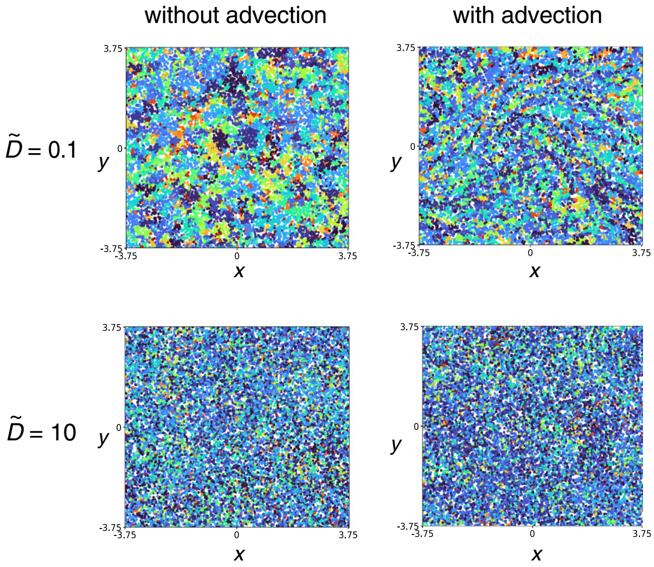

FIG. 4. Spatial species-distribution. The four panels present distributions of individuals in space in four different simulations of the backward model. Each dot represents an individual and colors represent species identity. The four panels represent cases without and with advection and strong ($\tilde{D} = 0.1$) and weak ($\tilde{D} = 10$) noise as indicated in the figure.

Spatial species distributions predicted by the backward model are shown in Fig. 4. In the strong-noise regime, $\tilde{D} = 0.1$, we observe spatial patterns generated by the velocity field.

Source code in C++ for the forward and backward models are available on GitHub [30].

III. DIVERSITY MEASURES

Quantifying biodiversity by simply counting the number of competing species is not always appropriate. The reason is that most ecological communities, including planktonic ones, are composed of relatively few abundant species and a large number of rare species. The definition of biodiversity in terms of number of species is insensitive to this distinction. This definition also relies on the possibility of observing all the rare species, which is often impossible in practice. To address this shortcoming, other measures of biodiversity have been proposed in the literature [31].

These additional measures take into account more explicitly spatial distributions and compositional heterogeneity. Specifically, at equal numbers of species, a population can be equally distributed among species or be dominated by a few of them. Moreover, the manner in which species distribute in space, i.e., whether individuals of a species are grouped within specific areas or highly dispersed in space, is an important characterization of biodiversity. Here, we review the most common measures of biodiversity.

(i) α diversity. The α diversity is the first and simplest measure. It is defined as the total average number of species in the community [32]. We measure the α diversity by simply averaging the total number of species over multiple realizations of the forward and backward models.

(ii) β diversity. The β diversity quantifies spatial correlations within species. It describes how species composition changes from local to larger scales. There exist slightly different definitions of β diversity in the literature [33,34]. We define β diversity as the probability that two random individuals at a given distance $r = \sqrt{x^2 + y^2}$ belong to the same species:

$$\beta(r) = \frac{\sum_i N_{s_i,s_i}(r)}{\sum_{i,j} N_{s_i,s_j}(r)}, \tag{4}$$

where $N_{s_i,s_j}(r)$ is the number of pairs i, j of species s_i, s_j at distance r [3]. At equal numbers of species, highly clustered communities are characterized by a steeper β diversity than more dispersed ones.

(iii) Species-area relation. The species-area relation is defined as the average number of species S found in an area A [35]. The species-area relation has been fitted by mathematical relations of the form $S(A) \propto A$ for small and large scales, and $S(A) \propto A^z$ for intermediate scales [35–38]. To compute the species-area relation, we average the final number of species S over several realizations of our models for increasing sampling areas of size $A = L_s \times L_s$ located at the center of our system.

(iv) Species-abundance distribution. The species-abundance distribution (SAD) quantifies the compositional heterogeneity of a population in terms of the relative abundance of species. It is defined as the frequency $P(n)$ of species with abundance n in a sample. In a well-mixed system of population size N_0 and with speciation probability μ, the species-abundance distribution has the form

$$P(n) = \frac{N_0 \mu e^{-\mu n}}{n} \tag{5}$$

(see Refs. [37,39]). We expect our spatially explicit model to generate a similar species-abundance distribution, at least in the high diffusion limit. Apart from this limiting case, analytical predictions for the species-abundance distribution in spatially explicit models are rather hard to obtain [3]. The species-abundance distribution is normalized to the α diversity, i.e., the average total number of species in the community:

$$S \approx \int_1^\infty dn \, \frac{N_0 \mu e^{-\mu n}}{n}. \tag{6}$$

IV. RESULTS

In this section, we numerically verify that the forward and backward models are equivalent in the weak-noise regime, and we test whether this equivalence extends in the strong-noise regime. We also study whether this equivalence is affected by the value of the speciation probability μ. To these aims, we extensively simulate both systems for a broad parameter range and compute diversity measures of the final population averaged over 10^3 realizations. We perform these comparisons both in the presence and in the absence of advection.

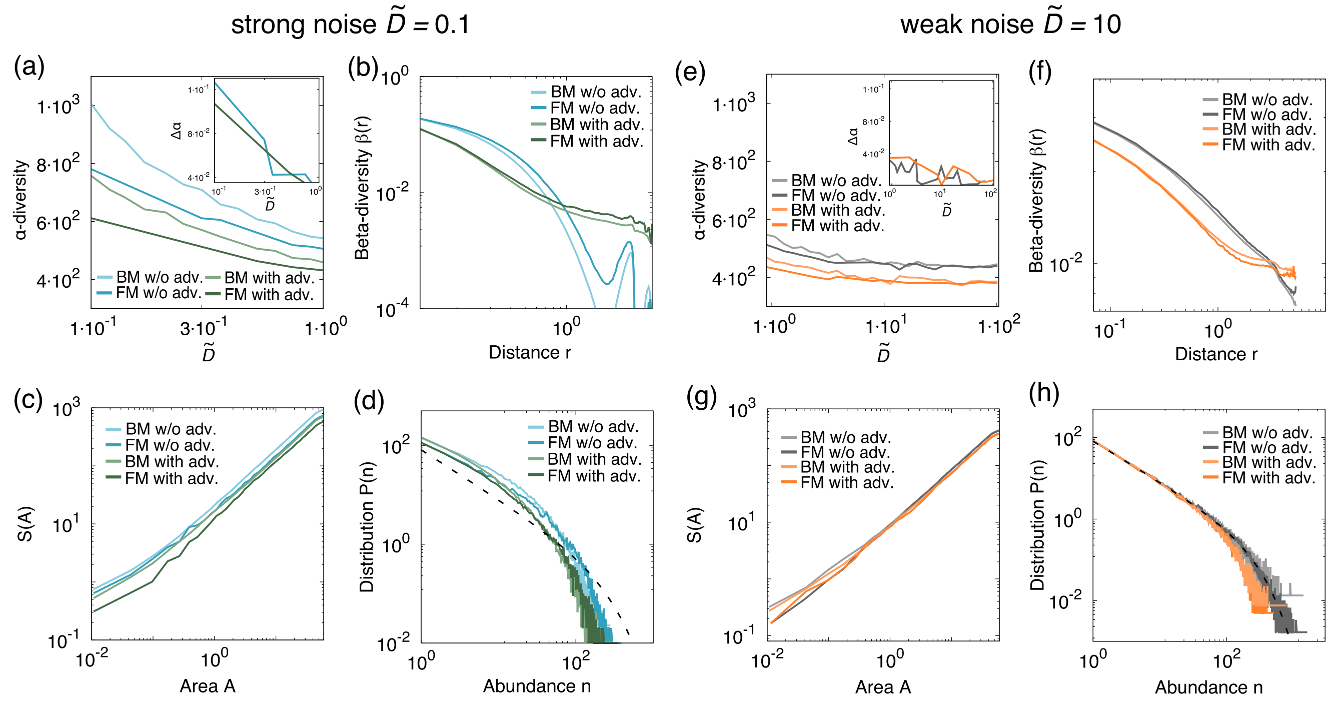

FIG. 5. Diversity measures predicted by the models with and without advection in the strong (a)–(d) and weak (e)–(h) noise regimes. (a), (e) Value of α diversity as a function of the parameter $\tilde{D}$. Measures for forward model (FM) (darker colors) and backward model (BM) (lighter colors) are shown. (b), (f) β diversity, (c), (g) species-area $S(A)$, and (d), (h) species abundance distribution for $\tilde{D} = 0.1$ and 10. In these panels, dark and light colors show curves for the forward and backward models, respectively, for $\tilde{D} = 10$; these curves nearly coincide. Dashed lines of panels (d) and (h) show the analytical prediction for a well-mixed system, Eq. (5). In all panels, the speciation probability is equal to $\mu = 5 \times 10^{-3}$.

A. Biodiversity measures in the absence of advection

We consider the two models in the weak- and strong-noise regimes and first focus on the α diversity for different diffusion rates D. As expected, the predictions of the forward and backward models present a small but significative discrepancy in the strong-noise regime [see Fig. 5(a)]. We quantify this discrepancy by the absolute relative difference

$$\Delta\alpha = \frac{|\alpha_f - \alpha_b|}{\alpha_f + \alpha_b}, \tag{7}$$

where α_f and α_b are the α diversities measured in the forward and backward dynamics, respectively. In the weak-noise regime, we observe compatible values of the α diversity [see Fig. 5(e)]. The same does not hold in the strong-noise regime. For example, for $\tilde{D} \approx 0.1$, the discrepancy is on the order of 10% [see inset of Fig. 5(a)]. In particular, the backward model predicts a higher number of species than the forward model. As discussed in Sec. II for the number of individuals, the differences between the forward and backward models in the strong-noise limit can depend on model details, such as the microscopic implementation of speciation.

Comparisons of the β diversity [Figs. 5(b) and 5(f)], the species-area relation [Figs. 5(c) and 5(g)], and the species-abundance distribution [Figs. 5(d) and 5(h)] lead to similar conclusions. In the weak-noise regime, the forward and backward models yield nearly identical predictions for all these quantities, as expected. In the strong-noise regime, we observe discrepancies of comparable magnitude as for the α diversity. We note that the peak of the β diversity at large r in Fig. 5(b) [and, less pronounced, in Fig. 5(f)] is caused by the periodic boundary conditions.

B. Biodiversity measures in the presence of chaotic advection

We now move to the case with advection. We define a two-dimensional incompressible advecting field in terms of the stream function $\phi(x, y)$. The components of the field are related with the stream function by

$$v_x(x, y; t) = -\frac{\partial \phi(x, y; t)}{\partial y},$$
$$v_y(x, y; t) = \frac{\partial \phi(x, y; t)}{\partial x}. \tag{8}$$

This definition automatically guarantees the incompressibility condition $\vec{\nabla} \cdot \vec{v} = 0$. We choose a dimensionless stream function that generates a chaotic vortex in proximity to the meandering jet [40]:

$$\phi(x, y) = -\tanh\left(\frac{y - B(t)\cos(kx)}{\sqrt{1 + k^2 B^2(t)\sin^2(kx)}}\right) + cy, \tag{9}$$

where $B(t) = B_0 + \epsilon\cos(wt + \Phi)$. We fix $k = 2\pi/L$, $c = 0.12$, $L = 7.5$, $B_0 = 1.2$, $\epsilon = 0.3$, $w = 0.4$, and $\Phi = \frac{\pi}{2}$. We impose periodic boundary conditions. We numerically solve

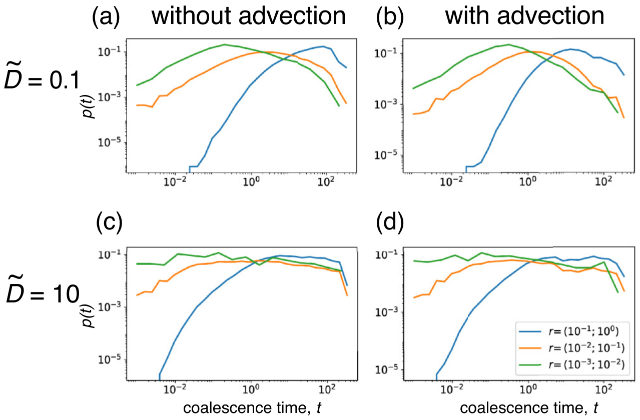

FIG. 6. Coalescence time distribution. The four panels present distributions of individuals in space in simulations of the backward model in four different scenarios (strong noise, weak noise, with advection, and without advection). Each curve represents a distribution for pairs of individuals chosen within an initial distance range r as shown in the figure legend. The median coalescent times are (a) 57.7, 3.2, and 0.4; (b) 20.8, 1.6, and 0.4; (c) 14.4, 1.3, and 0.2; (d) 14.7, 0.8, and 0.2. The three values for each panel correspond to values of r in the ranges $(10^{-1}; 10^{0})$, $(10^{-2}; 10^{-1})$, and $(10^{-3}; 10^{-2})$, respectively.

the differential equations (1) with the advecting field (8) using a fourth-order Runge-Kutta method.

As in the case without advection, we observe a significant discrepancy in the α diversity between the forward and backward models in the strong-noise regime [see Fig. 5(a)]. The discrepancy progressively decreases for bigger $\tilde{D}$ [see Fig. 5(e)]. This scenario is consistent with the argument made in Sec. II B, suggesting that duality might be only approximated in the strong-noise regime. In particular, the discrepancy in the strong-noise regime tends to be smaller in the presence of advection rather than in the absence of it [see inset of Fig. 5(a)]. One tentative explanation for this difference is that, at equal diffusivity, particles tend to be better mixed in the presence of advection. Other diversity measures show the expected behavior, with a close correspondence in the weak-noise regime and appreciable differences in the strong-noise regime [see Figs. 5(b)–5(d) and Figs. 5(f)–5(h)].

A comparison between Figs. 5(a) and 5(e) shows that advection tends to reduce the average number of species. The choice of periodic boundary conditions crucially impacts this result: in the presence of open boundary conditions, the model predicts that advection tends to increase the average number of species [24].

C. Coalescence time distributions

To further characterize the effect of the spatial structure of the population on its diversity, we study the dependence of the coalescent time between pairs of conspecific individuals on their initial distance. We find that closer individuals are more likely to have shorter coalescent time, i.e., to be more closely related (see Fig. 6).

The median coalescent times (reported in the caption of Fig. 6) are appreciably shorter in the weak-noise regime. Chaotic advection tends to reduce the median coalescent times in the strong-noise regime, but has little effect on them in the weak-noise regime.

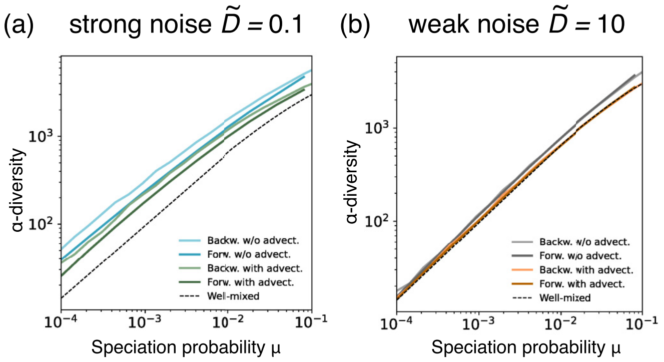

FIG. 7. α diversity as a function of the speciation probability in the strong-noise regime (a) and the weak-noise regime (b). Dashed lines show the analytical prediction for well-mixed systems [see Eq. (6)].

D. Effect of speciation probability

We now study the dependence of the α diversity on the speciation probabilities μ for both the forward and the backward model. In the strong-noise regime and for a low speciation probability, the forward model predicts an α diversity lower than that of the backward model [see Figs. 7(a) and 7(b)]. The relative difference between forward and backwards models weakly depends on μ. In the weak-noise regime, models are compatible for all values of the parameter μ [see Figs. 7(a) and 7(b)]. These behaviors are qualitatively similar in the presence and in the absence of advection.

E. Metagenomic data

In this section, we compare the SADs predicted by the coalescence model with those observed in aquatic environments.

We numerically simulate the coalescence model in the presence and the absence of advection [see Figs. 8(a) and 8(b)]. In this case, we adopt open boundary conditions as these are more relevant for the ocean, where a sample is embedded in a very large area. We do not perform a comparison with the forward dynamics, as the forward IBM cannot be easily formulated with open boundaries. Individuals are homogeneously distributed in a square at $t = 0$. In the case with advection, we set $c = 0.12$, $B_0 = 1.2$, $w = 0.5$, and $\epsilon = 4$. Our simulations show that chaotic advection leads to SAD curves characterized by a steeper decay (see Fig. 8(d) and Ref. [24]).

We compare the model prediction for the SAD with observational data. To this aim, we employ two metabarcoding datasets of protists populations: one from the TARA Oceans expedition [41], and one from freshwater lakes [42]. Metabarcoding techniques permit one to sample planktonic diversity at unprecedented resolution [43]. In metabarcoding studies, one obtains from a sample DNA fragments corresponding to a highly conserved region of the genome, in this case a

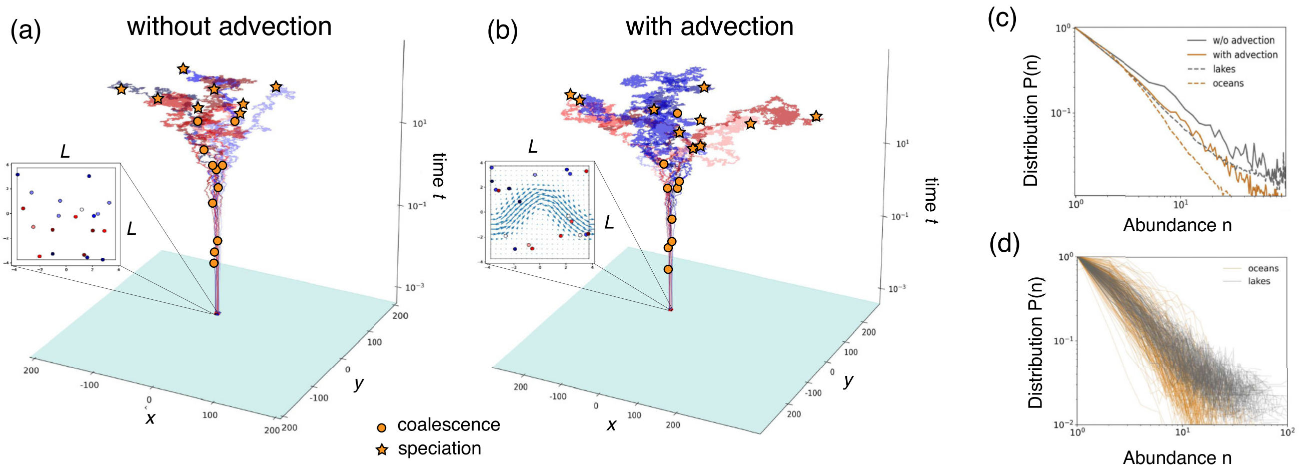

FIG. 8. Individual trajectories modeled with the coalescence model in the (a) absence and the (b) presence of advection. At the initial time $t = 0$, the populations are homogeneously distributed in a square of size $L \times L$. We simulate the coalescence model with open boundary conditions. In panel (b), we employ the velocity field given in Eqs. (8) and (9). Parameters are specified in Sec. IV B. (c) Dashed curves represent average abundance distributions of protist populations sampled in oceans and lakes; solid curves correspond to numerical simulations of the backward models with open boundary conditions without and with advection, where we fixed $\tilde{D} = 1$. (d) Species abundance distributions of individual protist populations sampled in oceans and lakes.

portion of the 18S ribosomal RNA gene. Since this region is highly conserved, sequences in the sample with a high degree of genetic similarity (at least 97%, in this case) are likely to originate from individuals within the same taxonomic group. These groups, identified by genetic similarity, are called operational taxonomic units (OTUs). We plot the abundance distributions of the observed OTUs from each dataset [see Fig. 8(d)].

The comparison of metagenomic data between oceans and lakes shows that SAD in the oceans are characterized by a steeper slope [see Fig. 8(d)]. This comparison supports the idea that large-scale oceanic currents, which are absent in lakes, are responsible for the steeper decay of SAD curves, as predicted by our model. Notice, however, that the observed slopes are slightly steeper than predicted by the model in both cases [see Fig. 8(c)]. As extensively discussed in Ref. [24], the quantitative value of the SAD slope is affected by microscopic details of the model and other aspects of the metabarcoding analysis, such as the similarity threshold chosen to identify OTUs. Further discussion of the robustness of the numerical predictions with respect to the model assumptions and parameters can be also found in Ref. [24].

V. CONCLUSIONS

In this paper, we developed a model for the dynamics of microbial aquatic communities based on the idea of coalescence. Our coalescence model predicts the diversity of a sample of organisms embedded in a very large, spatially extended population. It encompasses the limitations of an individual-based model in describing communities made up of huge numbers of individuals. Our model has the potential to bridge the gap between ecological dynamics at the individual level and large-scale spatial dynamics.

In the context of the forward model, we have identified a weak-noise regime and a strong-noise regime. The separation between these two regimes is not much affected by chaotic advection. However, this might not be the case for different choices of the velocity field. More broadly, it will be interesting in the future to study whether it is possible to connect density fluctuations in this model with known results for passive advection [44].

We have shown that, in the weak-noise regime, the model is equivalent to an individual-based model proceeding forward in time. In the strong-noise regime, this correspondence is only approximate, but both models predict qualitatively similar biodiversity patterns. Due to its advantages, the coalescence model presented in this paper provides a versatile and powerful tool to predict biodiversity observed in metabarcoding studies of planktonic communities [24].

Although, for simplicity, we focus on a simple model of microbial competition dynamics, our approach can be extended to more general ecological settings and to other communities. For example, although neutral models predict rather well the OTU composition of microbial communities [24,45], it will be important to extend our model to non-neutral cases, where individuals belonging to different species might have different species and competition intensity depends on species similarity. Such generalizations, combined with high-throughput sequencing data, have the potential to shed light on the main ecological forces determining microbial community dynamics.

ACKNOWLEDGMENTS

We thank Massimo Cencini for comments on a preliminary version of this manuscript. We thank Florian Pflug for helping us to debug the code. We thank Ales Bucek and Tom Bourguignon for the analysis of metabarcoding data and stimulating discussions. We are grateful for the help and support provided by the Scientific Computing and Data Analysis section of the Research Support Division at OIST.

APPENDIX A: STATIONARY NUMBER OF INDIVIDUALS IN THE WEAK-NOISE LIMIT

In this Appendix, we compute the stationary number of individuals N_0 in the limit in which individuals are homogeneously distributed. The number of individuals n inside each neighborhood is a Poisson random variable with average $\mu = N_0/M$, where $M = (L/l)^2$ is the total number of neighborhoods.

To estimate N_0, we impose that the total average birth and death rates should balance. The total average birth rate is simply λN_0. Individuals die with a rate proportional to the number of other individuals in each neighborhood. Therefore, the total average death rate is equal to the following:

$$\begin{aligned}\text{total average death rate} &= \lambda M \langle n(n-1)\rangle \\ &= \lambda M \sum_n \frac{n(n-1)\mu^n e^{-\mu}}{n!} \\ &= \lambda M \sum_n \frac{\mu^n e^{-\mu}}{(n-2)!} \\ &= \lambda M \mu^2 = \frac{\lambda N_0^2}{M} \\ &= \frac{\lambda N_0^2 l^2}{L^2}. \end{aligned} \tag{A1}$$

Imposing that the total average death rate must be equal to the total average birth rate leads to the condition $L = l\sqrt{N_0}$, or equivalently $N_0 = M$.

APPENDIX B: MACROSCOPIC DESCRIPTION OF THE FORWARD MODEL

We here derive a macroscopic description of our IBM. We introduce the density of individuals $n(x, y; t)$, defined so that its integral over a given area yields the number of individuals in that area at time t. We also define the concentration $c(x, y; t) = (L^2/N_0)n(x, y; t)$. The normalization factor L^2/N_0 ensures that the average concentration is equal to 1, if the assumption of homogeneous density holds.

The dynamics of the concentration $c(x, y; t)$ can be derived in the small-noise limit, e.g., by assuming that the stochastic fluctuations induced by birth and death processes are relatively small. Under this assumption, the concentration is described by the stochastic Fisher-Kolmogorov equation

$$\frac{\partial}{\partial t} c(x, y; t) = \lambda(c - c^2) - \vec{\nabla} \cdot [\vec{v}c] + D\nabla^2 c + \sigma(c)\xi(x, y; t), \tag{B1}$$

where $\xi(x, y, t)$ is a noise field satisfying $\langle \xi(x, y, t)\rangle = 0$ and $\langle \xi(x, y, t)\xi(x', y', t')\rangle = \delta(x - x')\delta(y - y')\delta(t - t')$. The multiplicative noise is interpreted using the Ito prescription; its amplitude is equal to $\sigma(c) = \sqrt{\lambda L^2 c(1 + c)/N_0}$. Equation (B1) can be derived using a Kramers-Moyal expansion (see Ref. [14] and Chap. 13 in Ref. [46]). In the derivation, we neglected contributions to the noise coming from the diffusion operator (see Ref. [14]).

In this case without advection, $\vec{v} = 0$, the concentration $c(x, y; t)$ is subject to two competing effects: the noise term in Eq. (B1) which creates fluctuations around the average solution $\langle c(x, y; t)\rangle = 1$, and the diffusion term which smooths these fluctuations.

[1] S. A. Levin, The problem of pattern and scale in ecology: The Robert H. MacArthur award lecture, Ecology **73**, 1943 (1992).

[2] M. Cencini, S. Pigolotti, and M. A. Munoz, What ecological factors shape species-area curves in neutral models?, PLoS ONE **7**, e38232 (2012).

[3] S. Pigolotti, M. Cencini, D. Molina, and M. A. Muñoz, Stochastic spatial models in ecology: A statistical physics approach, J. Stat. Phys. **172**, 44 (2018).

[4] R. Durrett and S. Levin, Spatial models for species-area curves, J. Theor. Biol. **179**, 119 (1996).

[5] J. Rosindell and S. J. Cornell, Species–area relationships from a spatially explicit neutral model in an infinite landscape, Ecol. Lett. **10**, 586 (2007).

[6] S. Pigolotti and M. Cencini, Speciation-rate dependence in species–area relationships, J. Theor. Biol. **260**, 83 (2009).

[7] R. Durrett and S. A. Levin, Stochastic spatial models: a user's guide to ecological applications, Philos. Trans. R. Soc. London, Sect. B **343**, 329 (1994).

[8] J. T. Cox, Coalescing random walks and voter model consensus times on the torus in Z^d, Ann. Probab. **17**, 1333 (1989).

[9] M. Bramson and J. L. Lebowitz, Asymptotic behavior of densities for two-particle annihilating random walks, J. Stat. Phys. **62**, 297 (1991).

[10] Z. Toroczkai, G. Károlyi, Á. Péntek, T. Tél, and C. Grebogi, Advection of Active Particles in Open Chaotic Flows, Phys. Rev. Lett. **80**, 500 (1998).

[11] G. Károlyi, Á. Péntek, I. Scheuring, T. Tél, and Z. Toroczkai, Chaotic flow: the physics of species coexistence, Proc. Natl. Acad. Sci. USA **97**, 13661 (2000).

[12] E. Hernández-García and C. López, Clustering, advection, and patterns in a model of population dynamics with neighborhood-dependent rates, Phys. Rev. E **70**, 016216 (2004).

[13] S. Pigolotti, R. Benzi, M. H. Jensen, and D. R. Nelson, Population Genetics in Compressible Flows, Phys. Rev. Lett. **108**, 128102 (2012).

[14] S. Pigolotti, R. Benzi, P. Perlekar, M. H. Jensen, F. Toschi, and D. R. Nelson, Growth, competition and cooperation in spatial population genetics, Theor. Popul. Biol. **84**, 72 (2013).

[15] F. Herrerías-Azcué, V. Pérez-Muñuzuri, and T. Galla, Stirring does not make populations well mixed, Sci. Rep. **8**, 4068 (2018).

[16] A. Plummer, R. Benzi, D. R. Nelson, and F. Toschi, Fixation probabilities in weakly compressible fluid flows, Proc. Natl. Acad. Sci. USA **116**, 373 (2019).

[17] G. Guccione, R. Benzi, and F. Toschi, Strong noise limit for population dynamics in incompressible advection, Phys. Rev. E **104**, 034421 (2021).

[18] E. Heinsalu, E. Hernández-Garcia, and C. López, Clustering Determines Who Survives for Competing Brownian and Lévy Walkers, Phys. Rev. Lett. **110**, 258101 (2013).

[19] S. Pigolotti and R. Benzi, Selective Advantage of Diffusing Faster, Phys. Rev. Lett. **112**, 188102 (2014).

[20] S. Pigolotti and R. Benzi, Competition between fast-and slow-diffusing species in non-homogeneous environments, J. Theor. Biol. **395**, 204 (2016).

[21] T. Singha, P. Perlekar, and M. Barma, Fixation in competing populations: Diffusion and strategies for survival, Phys. Rev. Res. **2**, 023412 (2020).

[22] R. Bainbridge, The size, shape and density of marine phytoplankton concentrations, Biol. Rev. **32**, 91 (1957).

[23] M. Scheffer, J. Baveco, D. DeAngelis, K. A. Rose, and E. van Nes, Super-individuals a simple solution for modelling large populations on an individual basis, Ecol. Modell. **80**, 161 (1995).

[24] P. Villa Martín, A. Bucek, T. Bourguignon, and S. Pigolotti, Ocean currents promote rare species diversity in protists, Sci. Adv. **6**, eaaz9037 (2020).

[25] L. N. Thomas, A. Tandon, and A. Mahadevan, Submesoscale processes and dynamics, Ocean Model. Eddying Regime **177**, 17 (2008).

[26] R. Benzi, M. H. Jensen, D. R. Nelson, P. Perlekar, S. Pigolotti, and F. Toschi, Population dynamics in compressible flows, Eur. Phys. J.: Spec. Top. **204**, 57 (2012).

[27] A. Plummer, M. Freilich, R. Benzi, C. J. Choi, L. Sudek, A. Z. Worden, F. Toschi, and A. Mahadevan, Oceanic frontal divergence alters phytoplankton competition and distribution, arXiv:2202.11745.

[28] M. Kimura and J. F. Crow, The number of alleles that can be maintained in a finite population, Genetics **49**, 725 (1964).

[29] J. Rosindell, Y. Wong, and R. S. Etienne, A coalescence approach to spatial neutral ecology, Ecol. Inf. **3**, 259 (2008).

[30] Code for the numerical simulations available on GitHub, https://github.com/AnzhelikaKoldaeva/Coalescent_dynamics_of_planktonic_communities.

[31] S. Xu, L. Böttcher, and T. Chou, Diversity in biology: Definitions, quantification and models, Phys. Biol. **17**, 031001 (2020).

[32] R. H. Whittaker, Vegetation of the Siskiyou Mountains, Oregon and California, Ecolo. Monogr. **30**, 279 (1960).

[33] H. Tuomisto, A diversity of beta diversities: Straightening up a concept gone awry. Part 1. Defining beta diversity as a function of alpha and gamma diversity, Ecography **33**, 2 (2010).

[34] H. Tuomisto, A diversity of beta diversities: Straightening up a concept gone awry. Part 2. Quantifying beta diversity and related phenomena, Ecography **33**, 23 (2010).

[35] M. L. Rosenzweig *et al.*, *Species Diversity in Space and Time* (Cambridge University, Cambridge, England, 1995).

[36] F. Preston, Time and space and the variation of species, Ecology **41**, 611 (1960).

[37] S. P. Hubbell, *The Unified Neutral Theory of Biodiversity and Biogeography (MPB-32)* (Princeton University, Princeton, NJ, 2001).

[38] O. Arrhenius, Species and area, J. Ecol. **9**, 95 (1921).

[39] I. Volkov, J. R. Banavar, S. P. Hubbell, and A. Maritan, Neutral theory and relative species abundance in ecology, Nature (London) **424**, 1035 (2003).

[40] M. Cencini, G. Lacorata, A. Vulpiani, and E. Zambianchi, Mixing in a meandering jet: A Markovian approximation, J. Phys. Oceanogr. **29**, 2578 (1999).

[41] E. Ser-Giacomi, L. Zinger, S. Malviya, C. De Vargas, E. Karsenti, C. Bowler, and S. De Monte, Ubiquitous abundance distribution of non-dominant plankton across the global ocean, Nat. Ecol. Evol. **2**, 1243 (2018).

[42] J. Boenigk, S. Wodniok, C. Bock, D. Beisser, C. Hempel, L. Grossmann, A. Lange, and M. Jensen, Geographic distance and mountain ranges structure freshwater protist communities on a European scale, Metab. Metagenomics **2**, e21519 (2018).

[43] C. de Vargas, S. Audic, N. Henry, J. Decelle, F. Mahé, R. Logares, E. Lara, C. Berney, N. Le Bescot, I. Probert *et al.*, Eukaryotic plankton diversity in the sunlit ocean, Science **348**, 1261605 (2015).

[44] G. Falkovich, K. Gawedzki, and M. Vergassola, Particles and fields in fluid turbulence, Rev. Mod. Phys. **73**, 913 (2001).

[45] P. Jeraldo, M. Sipos, N. Chia, J. M. Brulc, A. S. Dhillon, M. E. Konkel, C. L. Larson, K. E. Nelson, A. Qu, L. B. Schook *et al.*, Quantification of the relative roles of niche and neutral processes in structuring gastrointestinal microbiomes, Proc. Natl. Acad. Sci. USA **109**, 9692 (2012).

[46] C. W. Gardiner *et al.*, *Handbook of Stochastic Methods* (Springer, Berlin, 1985), Vol. 3.