Bulletin of Mathematical Biology (2024) 86:101
https://doi.org/10.1007/s11538-024-01337-6

Society for Mathematical Biology

ORIGINAL ARTICLE

A Diffusion-Based Approach for Simulating Forward-in-Time State-Dependent Speciation and Extinction Dynamics

Albert C. Soewongsono[1] · **Michael J. Landis**[1]

Received: 5 February 2024 / Accepted: 27 June 2024
© The Author(s), under exclusive licence to the Society for Mathematical Biology 2024

Abstract
We establish a general framework using a diffusion approximation to simulate forward-in-time state counts or frequencies for cladogenetic state-dependent speciation-extinction (ClaSSE) models. We apply the framework to various two- and three-region geographic-state speciation-extinction (GeoSSE) models. We show that the species range state dynamics simulated under tree-based and diffusion-based processes are comparable. We derive a method to infer rate parameters that are compatible with given observed stationary state frequencies and obtain an analytical result to compute stationary state frequencies for a given set of rate parameters. We also describe a procedure to find the time to reach the stationary frequencies of a ClaSSE model using our diffusion-based approach, which we demonstrate using a worked example for a two-region GeoSSE model. Finally, we discuss how the diffusion framework can be applied to formalize relationships between evolutionary patterns and processes under state-dependent diversification scenarios.

Keywords Evolution · Speciation · Extinction · Diffusion processes · Branching processes · Stationary frequencies

1 Introduction

The branching events of a phylogenetic tree exhibit a pattern that stores information about the underlying speciation and extinction processes (Nee et al. 1994). In Nee et al. (1994), they first considered a model where both speciation and extinction are

Albert C. Soewongsono and Michael J. Landis have contributed equally to this work.

✉ Albert C. Soewongsono
soewongsono@wustl.edu

Michael J. Landis
michael.landis@wustl.edu

1 Department of Biology, Washington University in St. Louis, Rebstock Hall, St. Louis, MO 63130, USA

Published online: 06 July 2024

treated as a constant-rate birth-death process by which lineages give birth to new lineages (speciation) at a rate λ and lineages die (extinction) at a rate μ. Speciation and extinction rates, however, are expected to vary idiosyncratically among phylogenetic lineages and over geological timescales. For example, Nee et al. (1994) also considered another model in which speciation and extinction rates vary over time. Workers have designed birth-death models to study a variety of intrinsic and extrinsic factors that might shape diversification rates. Species age (Hagen et al. 2015; Alexander et al. 2016; Soewongsono et al. 2022) and inherited traits (Kontoleon 2006; Maddison et al. 2007; FitzJohn 2010, 2012; Soewongsono et al. 2023) are two types of intrinsic factors thought to drive diversification rates, whereas environment (Condamine et al. 2013; Quintero et al. 2023) and geography (Goldberg et al. 2011; Landis et al. 2022; Swiston and Landis 2023) are common extrinsic factors of interest. In the end, a common goal of these models is to infer the underlying event rates given an observed phylogenetic pattern either through likelihood-based (Morlon et al. 2010; Stadler 2013; Louca and Pennell 2020b) or likelihood-free approaches (Nee et al. 1994; Voznica et al. 2022; He et al. 2023; Lambert et al. 2023; Thompson et al. 2023).

Fundamentally, birth-death processes model the random arrival times of discrete events that generate or "build" a phylogenetic tree over time (Nee et al. 1994; Maddison et al. 2007). As an alternative to this tree-based representation of the process, recent work (Chevin 2016) introduced an equivalent diffusion-based representation for a class of birth-death models with state-dependent rates, known as state-dependent speciation-extinction (SSE) models (Maddison et al. 2007). As noted by Chevin (2016), population genetics theory has benefited immensely from diffusion-based approximations to population-based models of allele frequency change, yet diffusion-based approximations of birth-death models remain underexplored in the phylogenetics literature. Despite the widespread popularity of birth-death models among evolutionary biologists, these models recently entered a phase of intense but overdue scrutiny to better understand what the models can and cannot estimate reliably when fitted to real biological datasets (Louca and Pennell 2020a; Morlon et al. 2022; Vasconcelos et al. 2022; Dragomir et al. 2023; Kopperud et al. 2023; Legried and Terhorst 2023; Truman et al. 2024; Celentano et al. 2024; Tarasov and Uyeda 2024). This has created demand for new frameworks to understand the mathematical properties of these complex stochastic processes to guide biological research programs.

As mentioned above, applying diffusion processes in the macroevolutionary context is not new, and was recently applied by Chevin (2016) to study the properties of the BiSSE (Maddison et al. 2007) and QuaSSE (FitzJohn 2010) models. Our work begins by extending the diffusion-based BiSSE representation of Chevin (2016) to a general multi-state SSE model that allows for both cladogenetic and anagenetic state changes, known as the ClaSSE model (Goldberg and Igić 2012). We then show how our formulation may be used to determine the relationship between a set of SSE rates and their implied stationary state frequencies. Inverting this perspective, we show that our framework correctly delimits classes of SSE rate values that yield a given set of stationary frequencies. This establishes a many-to-one mapping of SSE rates on to stationary frequencies. After introducing our general framework for ClaSSE models, we apply it to a special geographical case of the ClaSSE model, known as the GeoSSE model (Goldberg et al. 2011). We choose the GeoSSE model because

it possesses a complex but structured relationship among its parameters and its constituent events—i.e. dispersal, within-region speciation, between-region speciation, and local extinction—that impact lineages over evolutionary time. We then validate our theoretical results by simulating state frequency trajectories using both tree-based and diffusion-based simulators.

The rest of the paper is organized as follows. Firstly, in Sect. 2.1, we give a brief overview of SSE models in general. In Sect. 2.2 we visit relevant results in the theory of stochastic process, then in Sect. 2.3 we apply our framework to analyze the ClaSSE model, and later for the GeoSSE model with arbitrary number of regions in Sect. 2.4. Following these, in Sects. 2.5 and 2.6 we present a method for simulating state dynamics under our framework and deriving rate parameters given stationary state frequencies. In Sect. 2.7, we derive a result to compute theoretical stationary state frequencies given rate parameters. Moreover, in Sect. 2.8, we describe a procedure to compute time to reach stationary frequencies in a 2-region GeoSSE system using results derived in Sect. 2.7. Furthermore, in Sect. 3.1, we show, through simulation examples, that our diffusion-based framework offers a good approximation for simulating range state dynamics when comparing to tree-based approach. In Sect. 3.2, using an example, we show the existence of alternative rate scenarios that lead to the same stationary state frequencies. Additionally, we apply results derived in Sect. 2.7 and Sect. 2.8 to that example in Sect. 3.2. Lastly, in Sect. 4, we summarize our results and discuss promising ways to study pattern-process relationships for data generated by SSE models, and ideas for future work using our framework.

2 Methods

This section describes the framework for how construct our diffusion approximation for a ClaSSE model to analyze the dynamics of states through time. Key results include derivations of the transition probabilities and the infinitesimal mean and variance parameters of the diffusion equation. We describe and implement the methods for simulating the evolution of state frequencies, and derive relevant results for the stationary conditions, focusing on two- and three-region GeoSSE models, which are special cases of the ClaSSE model.

2.1 Overview of State-Dependent Speciation and Extinction Models

In this section, we give a brief overview of SSE models by highlighting the key assumptions and different events occurring along lineages. Then, we briefly re-visit a particular SSE model type, the GeoSSE model (Goldberg et al. 2011). Then, we guide towards how to shift from tree-based perspective to non-tree-based perspective to derive our object of interest.

In general, SSE models are stochastic branching processes with state-dependent birth (speciation) and death (extinction) rates. The states can either be discrete or continuous (Maddison et al. 2007; FitzJohn 2010, 2012) and can represent various things, ranging from phenotypic traits to geographical ranges (Goldberg et al. 2011).

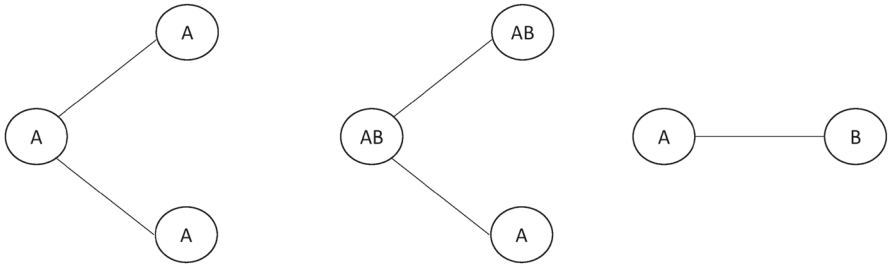

Fig. 1 From left to right: a speciation event without cladogenetic state changes, a speciation event with cladogenetic state changes, an anagenetic state change

Some SSE models have processes that are only defined by anagenetic process and state-dependent diversification process (Maddison et al. 2007), while others have processes that are defined by both anagenetic and cladogenetic processes (Goldberg et al. 2011; Goldberg and Igić 2012) shown in Fig. 1. An anagenetic process is defined as a process of trait evolution within lineages, between branching events. In the BiSSE model (Maddison et al. 2007), this corresponds to trait transition events of going from a discrete trait A to another discrete trait B or vice versa. These trait-dependent transition rates are encoded in the infinitesimal rate matrix $\boldsymbol{Q}$, for which the off-diagonal entry q_{ij} defines the rate of transitioning from state i to j. A cladogenetic process is defined as a process in which state transition occurs in conjunction with a branching event (with speciation) of a lineage. SSE models with anagenetic and cladogenetic events are referred to as ClaSSE models.

Part of this paper will consider a special case of the ClaSSE model, the GeoSSE model (Goldberg et al. 2011). A GeoSSE model describes how species move and evolve among a sets of discrete geographical regions, called species ranges. Species that occur in just one region are said to be endemic to that region. Species occurring in two or more regions are said to be widespread.

GeoSSE events can be classified as anagenetic or cladogenetic events. Anagenetic events in GeoSSE include dispersal events and local extinction (sometimes called extirpation) events. Dispersal events add one region to a species range. Local extinction remove one region from a species range. A species experiences complete extinction (i.e. it is removed from the species pool) when it goes locally extinct in the last region in its range. Note that widespread species cannot experience complete extinction through a single event under a GeoSSE model; their widespread ranges must first be reduced to a single region before complete extinction is a possibility.

Cladogenetic events under GeoSSE include within-region speciation and between-region speciation events. Each within-region speciation event creates a new species within any single region of the parental species range. Each between-region speciation event causes a widespread parental species and its range to split, such that all regions in the parental range are distributed among the two new daughter lineages. Section 2.4 defines how GeoSSE assigns rates to different events.

Given a phylogeny with range state information as seen in Fig. 2, one can observe the dynamics of range states accumulated by species though time. In Sect. 2.2, we

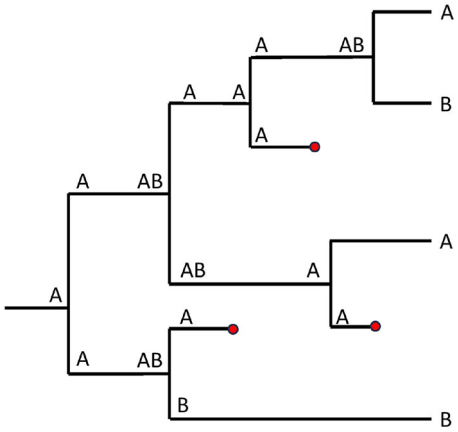

Fig. 2 An illustration of GeoSSE events on a phylogeny with range state information

present the necessary theory that will later be used to allow us transitioning from a tree-based process to an alternative, diffusion-based process to simulate the dynamics.

2.2 Transforming a Stochastic Process

In this section, we briefly describe the relevant results in the theory of stochastic processes that enable us to transform one stochastic process into another stochastic process. In the context of the ClaSSE model described in Sect. 2.1, we want to define a process that simulates the (discrete) count of species with state i through time. This process can then be used to define a second process that simulates the (continuous) frequency of species with state i over time.

Theorem 1 *Itô's transformation formula Consider a stochastic process $\{Z(t)\}$ with infinitesimal parameters $\mu(z)$ and $\sigma^2(z)$. Define a new stochastic process $\{Y(t)\}$ with $Y(t) = g(Z(t))$ where g is a strictly monotone continuous and twice-differentiable function. Then, the new process $\{Y(t)\}$ has infinitesimal parameters given by,*

- $\mu_Y(y) = \mu(z)g'(z) + \frac{1}{2}\sigma^2(z)g''(z)$,
- $\sigma_Y^2(y) = \sigma^2(z)\left[g'(z)\right]^2$.

Proof This theorem is also known as Itô's formula or Itô's lemma. The proof is given in Ito (1951), Karlin and Taylor (1981). □

Lemma 1 *Given a stochastic process $\{N_i(t) := n_i(t)\}$ with infinitesimal mean and variance parameters $\mu_i = \mathbb{E}(dn_i/dt)$ and $\sigma_i^2 = var(dn_i/dt)$, respectively. Define a stochastic process $\{X(t)\}$ derived using the following transformation.*

$$X(t) = g(N) = g\left(\sum_i n_i\right) = \sum_i h(n_i), \tag{1}$$

where $\{N(t) := \sum_i n_i(t)\}$ *is a stochastic process with infinitesimal parameters defined as follows,*

$$\begin{aligned}
\mu(N) &= \mu\left(\sum_i n_i(t)\right) \\
&= \sum_i \mu\left(n_i(t)\right) \\
&= \sum_i \mu_i. \\
\sigma^2(N) &= \sigma^2\left(\sum_i n_i(t)\right) \\
&= \sum_i \sigma^2(n_i(t)) + \sum_{\substack{i,j \\ i \neq j}} \sigma_{ij} \\
&= \sum_i \sigma^2(n_i(t)) \\
&= \sum_i \sigma_i^2.
\end{aligned}$$

Note here we have used the fact that $\sigma_{ij} = 0$ *for* $i \neq j$ *to account for independent birth-death processes. The infinitesimal mean and variance parameters for* $\{X(t)\}$ *are given by,*

$$\mu_X = \sum_i \frac{\partial X}{\partial n_i}\mu_i + \frac{1}{2}\sum_i \frac{\partial^2 X}{\partial n_i^2}\sigma_i^2 \tag{2}$$

$$\sigma_X^2 = \sum_i \left(\frac{\partial X}{\partial n_i}\right)^2 \sigma_i^2. \tag{3}$$

Proof Proof of Lemma 1 is given in "Appendix 5.1". □

2.3 Diffusion-Based Framework for State-Dependent Diversification Model

In this section, we establish the framework for simulating state dynamics for state-dependent speciation and extinction models using diffusion processes. We show how to implement the framework in the ClaSSE model introduced in Goldberg and Igić (2012). Then, we relate our framework to earlier research (Chevin 2016) using a diffusion process for the BiSSE model (Maddison et al. 2007) and, later on, for the GeoSSE model (Goldberg et al. 2011).

Our first goal is to define the stochastic process $\{N_i(t)\}$, which describes the number of species with state $i \in S$ at time t, where S is the state space of the model. Then, using the method presented in Sect. 2.2, we can obtain the stochastic process $\{\Pi_i(t)\}$, which

describes the frequency of species with state i at time t. Using these two processes, we then derive results that directly link model parameters with stationary state frequency patterns that the model generates.

To proceed, we define the following probabilities:

$$\begin{aligned} Prob(\{N_i \to N_i + 1 \text{ in } \Delta t\}) &= Prob(N_i(t+\Delta t) = n_i + 1 \mid N_i(t) = n_i) \\ &:= \mathbb{P}_i^+ \Delta t, \\ Prob(\{N_i \to N_i - 1 \text{ in } \Delta t\}) &:= \mathbb{P}_i^- \Delta t, \\ Prob(\{N_i \to N_i \text{ in } \Delta t\}) &:= \mathbb{P}_i \Delta t. \end{aligned} \tag{4}$$

These probabilities correspond to gaining a new species in state i $\left(\mathbb{P}_i^+\right)$, losing a species in state i $\left(\mathbb{P}_i^-\right)$, and neither losing nor gaining a new species in state i $(\mathbb{P}_i)$ within an infinitesimal time step Δt.

For the ClaSSE model, we can write those probabilities as follows,

$$\begin{aligned} \mathbb{P}_i^+ \Delta t &= S_i^+ + E_i^+ + Q_i^+, \\ \mathbb{P}_i^- \Delta t &= S_i^- + E_i^- + Q_i^-, \\ \mathbb{P}_i \Delta t &= 1 - \left(\mathbb{P}_i^+ + \mathbb{P}_i^-\right) \Delta t, \end{aligned} \tag{5}$$

where

$$\begin{aligned} S_i^+ &= \text{Probability of events that lead to an increase in the number of species in state } i \text{ through} \\ &\quad \text{state-dependent speciation and speciation in conjunction with } \textit{cladogenetic} \text{ state change.} \\ E_i^+ &= \text{Probability of events that lead to an increase in the number of species in state } i \text{ through} \\ &\quad \text{extinction.} \\ Q_i^+ &= \text{Probability of events that lead to an increase in the number of species in state } i \text{ through} \\ &\quad \textit{anageneticstatechange}. \\ S_i^- &= \text{Probability of events that lead to a decrease in the number of species in state } i \text{ through} \\ &\quad \text{state-dependent speciation and speciation in conjunction with } \textit{cladogenetic} \text{ state change.} \\ E_i^- &= \text{Probability of events that lead to a decrease in the number of species in state } i \text{ through} \\ &\quad \text{extinction.} \\ Q_i^- &= \text{Probability of events that lead to a decrease in the number of species in state } i \text{ through} \\ &\quad \textit{anagenetic} \text{ state change.} \end{aligned} \tag{6}$$

Next, we define the infinitesimal mean $\mu_i = \mathbb{E}\left(dN_i/dt\right)$ and variance $\sigma_i^2 = var\left(dN_i/dt\right)$ for the stochastic process $\{N_i(t) : t > 0\}$.

Lemma 2 *The infinitesimal mean μ_i and variance σ_i^2 for the stochastic process $\{N_i(t) : t > 0\}$ is given by*

$$\mu_i = \mathbb{P}_i^+ - \mathbb{P}_i^-, \tag{7}$$

$$\sigma_i^2 = \mathbb{P}_i^+ + \mathbb{P}_i^-. \tag{8}$$

Proof Proof of Lemma is given in "Appendix 5.2". □

Next, we define a stochastic process $\{\Pi_i(t) : t > 0\}$ where

$$\Pi_i = \frac{N_i}{\sum_{j \in S} N_j} = \frac{N_i}{N}.$$

$\Pi_i(t)$ denotes the frequency of species being in state i at time t. We define the infinitesimal mean and variance for the process in Lemma 3.

Lemma 3 *The infinitesimal mean μ_{Π_i} and variance $\sigma^2_{\Pi_i}$ for the stochastic process $\{\Pi_i(t) : t > 0\}$ is given by*

$$\mu_{\Pi_i} = \frac{1}{N}\left(\mu_i - \frac{\sigma_i^2}{N}\right) + \frac{\Pi_i}{N}\sum_{j \in S}\left(-\mu_j + \frac{\sigma_j^2}{N}\right), \tag{9}$$

$$\sigma^2_{\Pi_i} = \left(\frac{\sigma_i}{N}\right)^2 (1 - 2\Pi_i) + \left(\frac{\Pi_i}{N}\right)^2 \sum_{j \in S} \sigma_j^2. \tag{10}$$

Proof Proof of Lemma 3 is given in "Appendix 5.3". □

From Eqs. (9)–(10), it is clear that the diffusion parameters $\left(\text{i.e. } \mu_{\Pi_i}, \sigma^2_{\Pi_i}\right)$ are undefined under a total extinction scenario of a tree (i.e. where $N = 0$ appears in multiple denominators).

To demonstrate the generality of the framework, we show the BiSSE model (Maddison et al. 2007) (and similarly for the MuSSE model (FitzJohn 2012)) can be represented as a diffusion process as follows. Under the BiSSE model, species possess binary traits with values in the state space $S = \{1, 2\}$. BiSSE is a special case of the ClaSSE model that, while it allows anagenetic trait transition and extinction events, its speciation events do not cause cladogenetic trait changes. That is, daughter lineages identically inherit the parent lineage state following speciation. Readers can refer to the supplementary material from Goldberg and Igić (2012) for its derivation. For the BiSSE model, we have

$$S_1^+ = \lambda_1 N_1 \Delta t,\ E_1^+ = 0,\ Q_1^+ = q_{21} N_2 \Delta t,$$
$$S_1^- = 0,\ E_1^- = \mu_1 N_1 \Delta t,\ Q_1^- = q_{12} N_1 \Delta t,$$

where λ_1 and μ_1 (b_1 and d_1 in Chevin (2016)) are speciation and extinction rates for trait 1, respectively. q_{12} and q_{21} (τ_{12} and τ_{21} in Chevin (2016)) are anagenetic trait transition from 1 to 2 and from 2 to 1, respectively. Similarly, following the definitions in Eq. (6), we also have

$$S_2^+ = \lambda_2 N_2 \Delta t,\ E_2^+ = 0,\ Q_2^+ = q_{12} N_1 \Delta t,$$
$$S_2^- = 0,\ E_2^- = \mu_2 N_2 \Delta t,\ Q_2^- = q_{21} N_2 \Delta t,$$

Using Eq. (7) and Eq. (8) we have the infinitesimal mean and variance of N_1,

$$\mu_1 = (\lambda_1 - \mu_1 - q_{12})\, N_1 + q_{21} N_2, \tag{11}$$
$$\sigma_1^2 = (\lambda_1 + \mu_1 + q_{12})\, N_1 + q_{21} N_2, \tag{12}$$

and similarly for N_2 with indices changed accordingly. These are the same μ_1 and σ_1^2 as described in Eq. (2) in Chevin (2016). □

2.4 Diffusion-Based Framework for the GeoSSE Model

In this section, we use the framework established in Sect. 2.3 for general ClaSSE models to the GeoSSE model. The procedure we apply here is also compatible with any model from the ClaSSE family. For the GeoSSE model, unlike the BiSSE model described in Sect. 2.3, some speciation events also cause *cladogenetic* state changes. Thus, following the notation used in the previous section we have,

$$\begin{aligned}
S_i^+ &= W_i^+ + B_i^+ \\
E_i^+ &= E_i^+ \\
Q_i^+ &= D_i^+ + E_i^+ \\
S_i^- &= W_i^- + B_i^- \\
E_i^- &= E_i^- \\
Q_i^- &= D_i^- + E_i^-,
\end{aligned}$$

where

W_i^+ = Probability of events that lead to an increase in the number of species in range state i through within-region speciation for either widespread or endemic species.

B_i^+ = Probability of events that lead to an increase in the number of species in range state i through between-region speciation for widespread species.

E_i^+ = Probability of events that lead to an increase in the number of species in range state i through extinction for either widespread species (local extinction) or endemic species (species extinction).

D_i^+ = Probability of events that lead to an increase in the number of species in range state i through range dispersal event for endemic species.

W_i^- = Probability of events that lead to a decrease in the number of species in range state i through within-region speciation for either widespread or endemic species.

B_i^- = Probability of events that lead to a decrease in the number of pecies in range state i through between-region speciation for widespread species.

E_i^- = Probability of events that lead to a decrease in the number of species in range state i through extinction for either widespread species (local extinction) or endemic species (species extinction).

$D_i^- =$ Probability of events that lead to a decrease in the number of species in range state i through range dispersal event for endemic species.

Next, consider an n-region GeoSSE model where $n \in \mathbb{Z}^+$, we define the following state space and variable,

$$\begin{aligned} R &= \text{state space for regions e.g., } R = \{A, B\}. \\ S &= \text{state space for species ranges e.g., } S = \{\{A\}, \{B\}, \{A, B\}\} \\ N_i &= \text{number of species with range state} i \text{where} i \in S. \end{aligned}$$

Then, we define the following rate parameters,

$$\begin{aligned} d_{k\ell} &= \text{per lineage dispersal rate of any species in region} k \text{to colonize region} \ell. \\ w_\ell &= \text{ per lineage within-region speciation rate of any species in region } \ell. \\ b_j^i &= \text{per lineage between-region speciation rate of a widespread species into.} \\ &\quad \text{two daughter species with ranges} i \text{and} j, \text{ respectively. Note that } b_j^i \equiv b_i^j. \\ e_\ell &= \text{ local extinction rate of any species in region } \ell. \end{aligned}$$

Thus, both w_ℓ and b_j^i determine state-dependent speciation rate, e_ℓ determines state-dependent extinction rate, and $d_{k\ell}$ and (among widespread species) e_ℓ determine the anagenetic state transition rate.

We define a stochastic process $\{N_i(t)\}$ with infinitesimal mean $\mu_i = \mathbb{E}(dN_i/dt)$ and variance $\sigma_i^2 = var(dN_i/dt)$. Here, $N_i(t)$ represents the number of species with range state i at time t. The infinitesimal mean μ_i and variance σ_i^2 follow directly from Lemma 2. We derive the transition probabilities described in Eq. (4) in the context of the GeoSSE model, as shown in Eqs. (13)–(15).

Each of these probabilities describe possible events in a GeoSSE model occurring within an infinitesimal time step that result in gaining a new species with range state i ($\mathbb{P}_i^+$), losing a species with range state i ($\mathbb{P}_i^-$), and neither losing nor gaining a species with range state i ($\mathbb{P}_i$).

$$\begin{aligned} \mathbb{P}_i^+ \Delta t &= W_i^+ + D_i^+ + B_i^+ + E_i^+ \\ &= \underbrace{\sum_{j \in S} \sum_{\substack{\ell \in j \\ \{\ell\} = i}} N_j w_\ell \Delta t}_{W_i^+} + \underbrace{\sum_{k \in i} \sum_{\substack{\ell \in i \\ \ell \neq k}} N_{i \setminus \{\ell\}} d_{k\ell} \Delta t}_{D_i^+} + \underbrace{\sum_{\substack{j \in S \\ i \subset j}} N_j b_{j \setminus i}^i \Delta t}_{B_i^+} + \underbrace{\sum_{\substack{j \in S \\ |j \setminus i| = 1}} \sum_{\ell \in j \setminus i} N_j e_\ell \Delta t}_{E_i^+} \end{aligned} \tag{13}$$

$$\begin{aligned} \mathbb{P}_i^- \Delta t &= W_i^- + D_i^- + B_i^- + E_i^- \\ &= \underbrace{0}_{W_i^-} + \underbrace{\sum_{\substack{k \in i \\ |i| < R}} \sum_{\ell \in R \setminus \{k\}} N_i d_{k\ell} \Delta t}_{D_i^-} + \underbrace{\sum_{\substack{j \in S \\ j \subset i}} \frac{1}{2} N_i b_{i \setminus j}^j \Delta t}_{B_i^-} + \underbrace{\sum_{\substack{\ell \in R \\ \ell \in i}} N_i e_\ell \Delta t}_{E_i^-} \end{aligned} \tag{14}$$

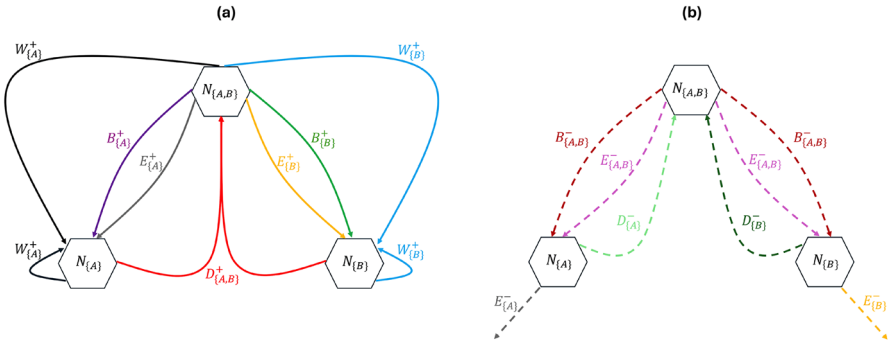

Fig. 3 Graphical illustrations of probabilities of events following Eq. (13) shown in (**a**), and Eq. (14) shown in (**b**) for a 2-region GeoSSE system with state space $S = \{\{A\}, \{B\}, \{A, B\}\}$. N_i represents the number of species with range state $i \in S$. An incoming arrow into N_i compartment means there is an increase in species count with range state i and an outgoing arrow from N_i means there is a decrease in species count with range state i. All the events and arrows are color-coded accordingly (Color figure online)

$$\mathbb{P}_i \Delta t = 1 - \left(\mathbb{P}_i^+ + \mathbb{P}_i^-\right) \Delta t. \tag{15}$$

For clarity, we provide the biogeographic interpretation on how each term in Eqs. 13–15 is derived and a graphical illustration of the events in Fig. 3.

1. $\boldsymbol{W_i^+}$. To gain a new species with range state i through a within-region speciation event, the new species range i must contain only region ℓ ($\ell \in i$ and $|i| = 1$). This endemic species can undergo a speciation event with probability $w_\ell N_i$. Any species with range state j that also occupies region ℓ can undergo a within-region speciation event with probability $w_\ell \sum_{j \in S} \mathbb{1}_{i \subseteq j} N_j$. The total probability of this event occurring within Δt is,

$$\sum_{j \in S} \sum_{\substack{\ell \in j \\ \{\ell\} = i}} N_j w_\ell \Delta t.$$

As an example, in a 2-region GeoSSE system with state space $S = \{\{A\}, \{B\}, \{A, B\}\}$ we have,

$$\begin{aligned} W_{\{A\}}^+ &= \left(N_{\{A,B\}} + N_{\{A\}}\right) w_A \Delta t \\ W_{\{B\}}^+ &= \left(N_{\{A,B\}} + N_{\{B\}}\right) w_B \Delta t \\ W_{\{A,B\}}^+ &= 0. \end{aligned}$$

2. $\boldsymbol{D_i^+}$. To gain a new species with range state i through a dispersal event, the species adds the new region ℓ to its ancestral range. Species are always widespread immediately following dispersal. The total probability of this event occurring within Δt is,

$$\sum_{k \in i} \sum_{\substack{\ell \in i \\ \ell \neq k}} N_{i \setminus \{\ell\}} d_{k\ell} \Delta t.$$

As an example, in a 2-region GeoSSE system with state space $S = \{\{A\}, \{B\}, \{A, B\}\}$ we have,

$$\begin{aligned} D^{+}_{\{A\}} &= 0 \\ D^{+}_{\{B\}} &= 0 \\ D^{+}_{\{A,B\}} &= N_{\{B\}} d_{BA} \Delta t + N_{\{A\}} d_{AB} \Delta t. \end{aligned}$$

3. $\boldsymbol{B_i^+}$. To gain a new species with range state i through a between-region speciation event, the new species can be either endemic or widespread $|i| > 0$ that originated from a widespread ancestral species with larger range state j ($i \subset j$). In general, we have no information of whether the new species occurs in left or right lineage following a speciation event, so we do not consider the orientation. The total probability of this event occurring within Δt is,

$$\sum_{\substack{j \in S \\ i \subset j}} N_j b^{i}_{j \setminus i} \Delta t.$$

As an example, in a 2-region GeoSSE system with state space $S = \{\{A\}, \{B\}, \{A, B\}\}$ we have,

$$\begin{aligned} B^{+}_{\{A\}} &= N_{\{A,B\}} b^{A}_{B} \Delta t \\ B^{+}_{\{B\}} &= N_{\{A,B\}} b^{B}_{A} \Delta t \\ B^{+}_{\{A,B\}} &= 0. \end{aligned}$$

4. $\boldsymbol{E_i^+}$. To gain a new species with range state i through a local extinction event, the ancestral species must have a larger range state j with size that differs by 1 from the new species' range state i such that $|j \setminus i| = 1$. The total probability of this event occurring within Δt is,

$$\sum_{\substack{j \in S \\ |j \setminus i| = 1}} \sum_{\ell \in j \setminus i} N_j e_\ell \Delta t.$$

As an example, in a 2-region GeoSSE system with state space $S = \{\{A\}, \{B\}, \{A, B\}\}$ we have,

$$\begin{aligned} E^{+}_{\{A\}} &= N_{\{A,B\}} e_B \Delta t \\ E^{+}_{\{B\}} &= N_{\{A,B\}} e_A \Delta t \\ E^{+}_{\{A,B\}} &= 0. \end{aligned}$$

5. $\boldsymbol{W_i^-}$. The probability of losing a either endemic or widespread species with range state i through a within-region speciation event is 0. This is because the event will only increase the local abundance in a region and causes the widespread abundance to remain unchanged.
6. $\boldsymbol{D_i^-}$. To lose a species with range state i through a dispersal event, the species must disperse to a new region. The species count remains unchanged if the species already occupies all regions ($|i| = |R|$). The total probability of this event occurring within Δt is,

$$\sum_{\substack{k \in i \\ |i| < R}} \sum_{\ell \in R \setminus \{k\}} N_i d_{k\ell} \Delta t.$$

As an example, in a 2-region GeoSSE system with state space $S = \{\{A\}, \{B\}, \{A, B\}\}$ we have,

$$\begin{aligned} D^-_{\{A\}} &= N_{\{A\}} d_{AB} \Delta t \\ D^-_{\{B\}} &= N_{\{B\}} d_{BA} \Delta t \\ D^-_{\{A,B\}} &= 0. \end{aligned}$$

7. $\boldsymbol{B_i^-}$. To lose a species with range state i through a between-region speciation event, the species must be widespread and undergo a speciation event that gives rise to a new species in state j with smaller range state size ($|j| < |i|$). The factor of 1/2 corrects for double-counting the new species with range j being either the left daughter or right daughter lineage. The total probability of this event occurring within Δt is,

$$\sum_{\substack{j \in S \\ |j| < |i|}} \frac{1}{2} N_i b^j_{i \setminus j} \Delta t.$$

As an example, in a 2-region GeoSSE system with state space $S = \{\{A\}, \{B\}, \{A, B\}\}$ we have,

$$\begin{aligned} B^-_{\{A\}} &= 0 \\ B^-_{\{B\}} &= 0 \\ B^-_{\{A,B\}} &= \frac{1}{2} N_{\{A,B\}} \left(b^A_B + b^B_A \right) \Delta t. \end{aligned}$$

8. $\boldsymbol{E_i^-}$. To lose a species with range state i through a local extinction event, a species must undergo an extinction event in one of its regions. If the species is endemic, this event leads to total extinction of the species. The total probability of this event

occurring within Δt is,

$$\sum_{\substack{\ell \in R \\ \ell \in i}} N_i e_\ell \Delta t.$$

As an example, in a 2-region GeoSSE system with state space $S = \{\{A\}, \{B\}, \{A, B\}\}$ we have,

$$\begin{aligned} E^-_{\{A\}} &= N_{\{A\}} e_A \Delta t \\ E^-_{\{B\}} &= N_{\{B\}} e_B \Delta t \\ E^-_{\{A,B\}} &= N_{\{A,B\}} \left(e_A + e_B\right) \Delta t. \end{aligned}$$

The next section uses Eqs. 13–15 to define the stochastic process $\{\Pi_i(t) : t > 0\}$ that models the frequency of species in range state i at time t. The infinitesimal mean μ_{Π_i} and variance $\sigma^2_{\Pi_i}$ follow directly from Lemma 3.

2.5 Comparison on Diffusion-Based and Tree-Based Models Using Simulation

In this section we show that our diffusion-based approach correctly models the temporal behaviour of range state frequencies in a GeoSSE model. To validate, we compare our results with a tree-based approach that explicitly simulates phylogenetic trees under the same GeoSSE parameter values using the MASTER package (Vaughan and Drummond 2013) implemented in BEAST2 (Bouckaert et al. 2014). When simulating given a large number of species initially, $N(0) >> 0$, both diffusion-based and MASTER-based simulations are conditioned only for the process to run until a specific elapsed time T. Later in "Appendix 5.6", when we simulate using both approaches starting with a single species in random state, $N(0) = 1$, we condition the process under both elapsed time and survival until the present. Details for setting up reaction equations for the MASTER simulation can be found in "Appendix 5.5".

For simulations under a diffusion, we generate sample paths on $[0, T]$, where T is the simulation running time. Each simulation yields a time-series of state frequencies for the provided SSE rate values. Simulations were generated as follows:

1. Given the following Itô stochastic differential equation (SDE) and the initial number of species in each range state, $N_i(0), \forall i \in S$,

$$dN_i = \mu_i(t)dt + \sigma_i(t)dW_t, \tag{16}$$

where dW_t is a Wiener process, we draw a sample path by using the following approximation,

$$N_i(t + \Delta t) = N_i(t) + \mu_i(t)\Delta t + \sigma_i(t)\sqrt{\Delta t}U_t, \tag{17}$$

where $\sqrt{\Delta t}U_t \sim \sqrt{\Delta t}N(0,1)$ is a (discretized) standard Wiener process, and $\mu_i(t)$ and $\sigma_i(t)$ are computed using Eqs (7)–(8), respectively.

2. Given $N_i(t+\Delta t)$ for each $i \in S$ from step 1, we compute the total number of species at $t+\Delta t \in [0, T]$

$$N(t+\Delta t) = \sum_{i \in S} N_i(t+\Delta t).$$

3. Next, using $N_i(t)$ and $N(t)$ from steps 1-2, we compute the infinitesimal mean, $\mu_{\Pi_i}(t)$, and infinitesimal variance, $\sigma_{\Pi_i}(t)$ using Eqs. (9)–(10), respectively. Given $\mu_{\Pi_i}(t)$, $\sigma_{\Pi_i}(t)$, and the following Itô SDE with the initial frequency of species of range state i, $\Pi_i(0) = \frac{N_i(0)}{N(0)}$,

$$d\Pi_i = \mu_{\Pi_i}(t)dt + \sigma_{\Pi_i}(t)dW_t, \tag{18}$$

where dW_t is a Wiener process, we draw a sample path by using the following approximation,

$$\Pi_i(t+\Delta t) = \Pi_i(t) + \mu_{\Pi_i}(t)\Delta t + \sigma_{\Pi_i}(t)\sqrt{\Delta t}U_t, \tag{19}$$

where $\sqrt{\Delta t}U_t \sim \sqrt{\Delta t}N(0,1)$ is a (discretized) standard Wiener process.

In Sect. 3.1, we show that the dynamic of the range state frequencies can be well-approximated using the diffusion-based framework. We provide different examples through numerical simulations under a variety of GeoSSE scenarios to visualize this result. Specifically, we apply the following procedure,

1. We consider a 3-region GeoSSE model, then we simulate range state dynamics using tree-based approach (via the MASTER package in BEAST2) and the diffusion-based approach over 1000 replicates on [0, 10] time interval with 1000 time steps. Note that if one simulates over a longer time interval, then one needs to choose larger time steps to reduce the chance that multiple events occur within Δt for the diffusion-based approach. For diffusion-based approach, at each time step, we assign a zero value to any state with a count less than zero since the number of species in any range states cannot be negative. This is reasonable because if $N_i(t) = 0$, then some events are not permitted such as a local extinction. Note that although some N_i's might be equal to 0, it is very unlikely for the whole clade to become extinct, i.e., $N(t) = 0$, given a relatively large clade size at the beginning of each process (Fig. 4) and value of each parameter we pick for the simulations (Figs. 5, 6, 7, 8). We consider the following scenarios for the GeoSSE model,

Example 1 GeoSSE model with only within-region speciation and between-region speciation events (Fig. 5).

Example 2 GeoSSE model with only within-region speciation and range dispersal events (Fig. 6).

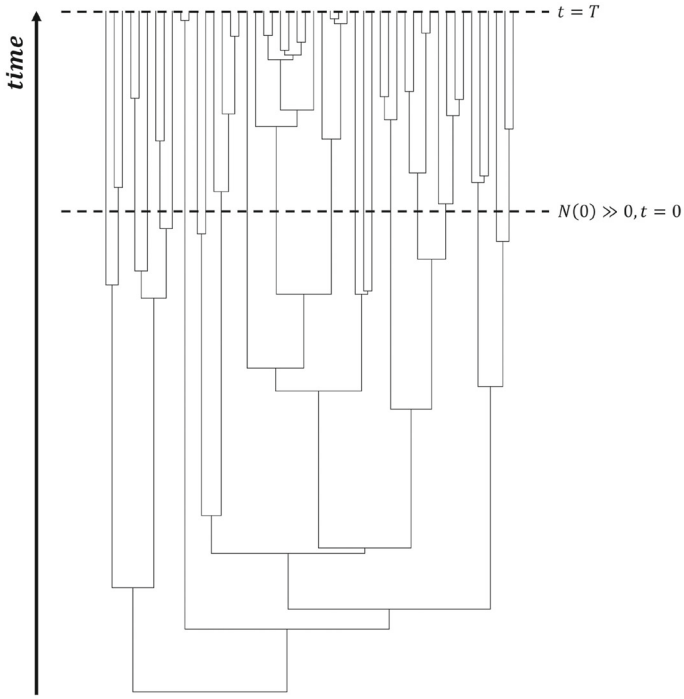

Fig. 4 For each diffusion-based and MASTER-based simulation, we assume that we start each $N_i(t)$ simulation, given a relatively large clade size at the beginning, $N(0) >> 0$

Example 3 GeoSSE model with only within-region speciation and local extinction events (Fig. 7).

Example 4 GeoSSE model with all the events included (Fig. 8).

2. We visualize the trajectory of mean state counts for each range state from both diffusion and tree-based approaches. For each simulation, we start the forward-in-time simulation given relatively large clade size for diffusion-based approach to accurately predict the dynamics given by tree-based approach from MASTER simulations. We also visualize stacked bar charts of expected state frequencies for both approaches. To compute the state frequencies under the tree-based approach across replicates, we use the following analytical formula

$$\Pi_i(t) = \frac{N_i(t)}{\sum_{i \in S} N_i(t)}.$$

We simulate frequency trajectories under the diffusion-based approach using Eq. (19). Also for diffusion-based approach, we normalize $\Pi_i(t)$ at each time step for each $i \in S$. Thus, keeping $\Pi_i(t) \leq 1$ at any time.

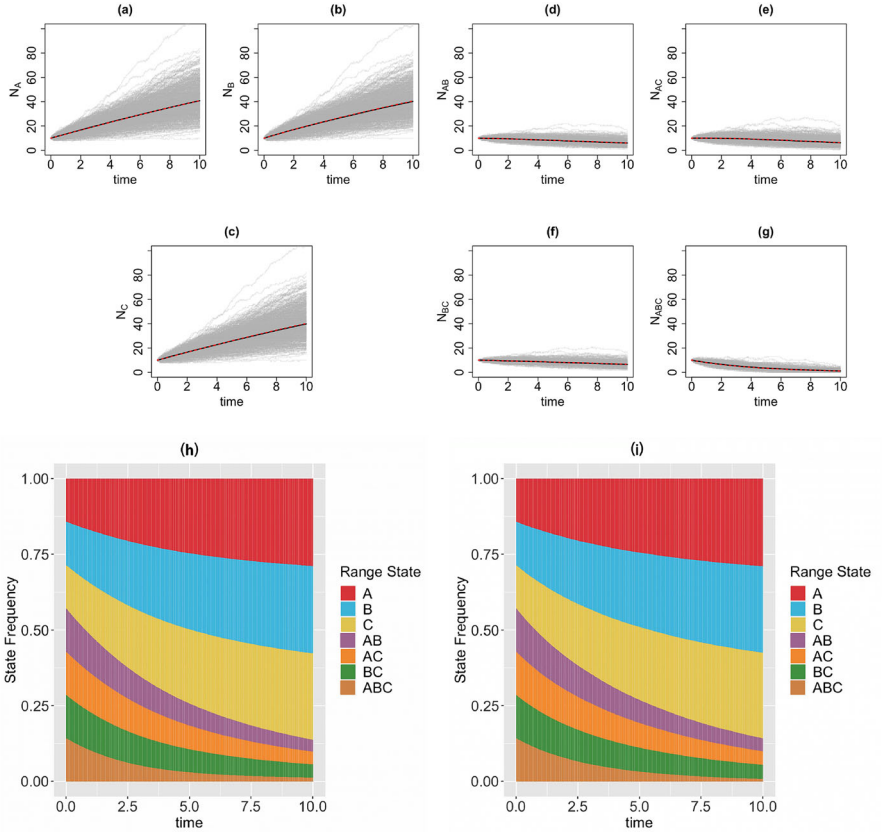

Fig. 5 Top & middle panels: the trajectories of average count of range states for endemic species (**b**–**c**) and widespread species (**d**–**g**) over [0, 10] time interval and over 1000 simulations runs for the three-region GeoSSE model as described in Example 1 each simulated under both diffusion-based process (red line) and tree-based process (black line). The gray trajectories show the dynamics across 1000 replicates simulated under diffusion-based process. Bottom panel: stacked bar chart showing the state frequencies over time using diffusion-based approach (**h**) and tree-based approach (**i**). In both approaches, we start the process with $N(0) = 40$ and the following initial state frequencies: $\Pi_{\{A\}}(0) = \Pi_{\{B\}}(0) = \Pi_{\{C\}}(0) = \Pi_{\{A,B\}}(0) = \Pi_{\{A,C\}}(0) = \Pi_{\{B,C\}}(0) = \Pi_{\{A,B,C\}}(0) = \frac{1}{7}$. At $t = 10$, the mean frequencies for each range state from both diffusion-based and tree-based simulations are as follows: $\bar{\Pi}^{diffusion}_{\{A\}} = 0.29$, $\bar{\Pi}^{tree}_{\{A\}} = 0.29$; $\bar{\Pi}^{diffusion}_{\{B\}} = 0.29$, $\bar{\Pi}^{tree}_{\{B\}} = 0.29$; $\bar{\Pi}^{diffusion}_{\{C\}} = 0.29$, $\bar{\Pi}^{tree}_{\{C\}} = 0.28$; $\bar{\Pi}^{diffusion}_{\{A,B\}} = 0.04$, $\bar{\Pi}^{tree}_{\{A,B\}} = 0.04$; $\bar{\Pi}^{diffusion}_{\{A,C\}} = 0.04$, $\bar{\Pi}^{tree}_{\{A,C\}} = 0.04$; $\bar{\Pi}^{diffusion}_{\{B,C\}} = 0.04$, $\bar{\Pi}^{tree}_{\{B,C\}} = 0.05$; $\bar{\Pi}^{diffusion}_{\{A,B,C\}} = 0.01$, $\bar{\Pi}^{tree}_{\{A,B,C\}} = 0.01$. Simulations are conducted using the following parameter values: $w_A = w_B = w_C = 0.03$, $b^A_B = 0.08$, $b^A_C = 0.10$, $b^B_C = 0.06$, $b^A_{BC} = 0.04$, $b^B_{AC} = 0.12$, $b^C_{AB} = 0.06$, $e_A = e_B = e_C = 0$, $d_{AB} = d_{BA} = d_{AC} = d_{CA} = d_{BC} = d_{CB} = 0$ (Color figure online)

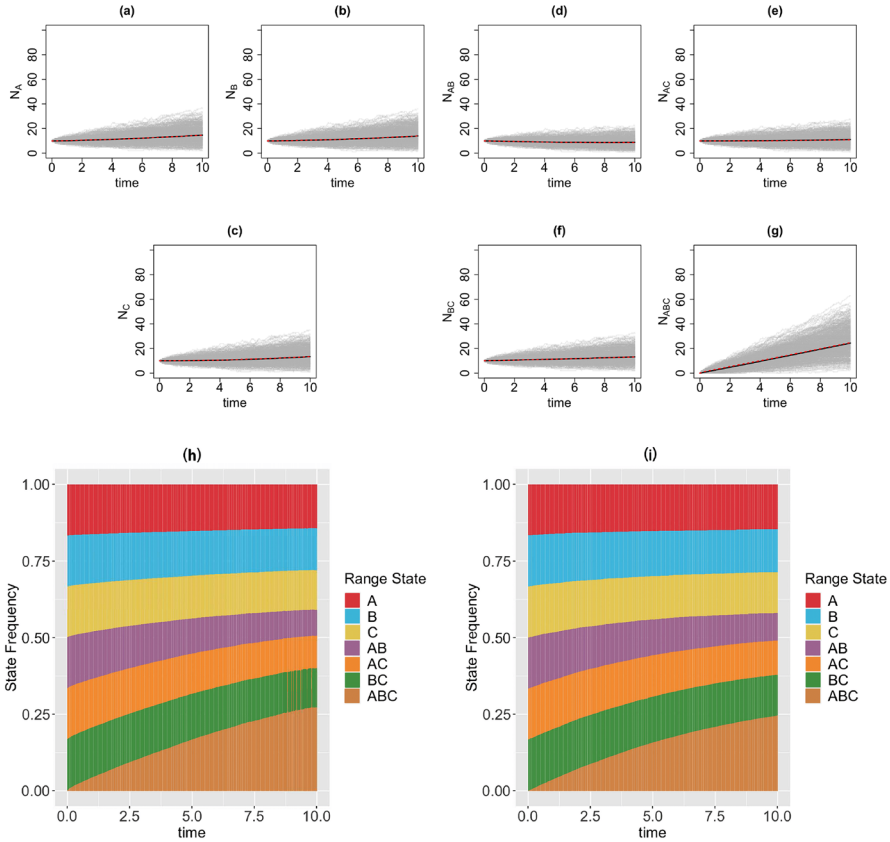

Fig. 6 Top & middle panels: the trajectories of average count of range states for endemic species (**a**–**c**) and widespread species (**d**–**g**) over [0, 10] time interval and over 1000 simulations runs for the three-region GeoSSE model as described in Example 2 simulated under both diffusion-based process (red line) and tree-based process (black line). The gray trajectories show the dynamics across 1000 replicates simulated under diffusion-based process. Bottom panel: stacked bar chart showing the state frequencies over time using diffusion-based approach (**h**) and tree-based approach (**i**). In both approaches, we start the process with $N(0) = 40$ and the following initial state frequencies: $\Pi_{\{A\}}(0) = \Pi_{\{B\}}(0) = \Pi_{\{C\}}(0) = \Pi_{\{A,B\}}(0) = \Pi_{\{A,C\}}(0) = \Pi_{\{B,C\}}(0) = \frac{1}{6}$, $\Pi_{\{A,B,C\}}(0) = 0$. At $t = 10$, the mean frequencies for each range state from both diffusion-based and tree-based simulations are as follows: $\bar{\Pi}^{diffusion}_{\{A\}} = 0.14$, $\bar{\Pi}^{tree}_{\{A\}} = 0.15$; $\bar{\Pi}^{diffusion}_{\{B\}} = 0.14$, $\bar{\Pi}^{tree}_{\{B\}} = 0.14$; $\bar{\Pi}^{diffusion}_{\{C\}} = 0.13$, $\bar{\Pi}^{tree}_{\{C\}} = 0.13$; $\bar{\Pi}^{diffusion}_{\{A,B\}} = 0.08$, $\bar{\Pi}^{tree}_{\{A,B\}} = 0.09$; $\bar{\Pi}^{diffusion}_{\{A,C\}} = 0.11$, $\bar{\Pi}^{tree}_{\{A,C\}} = 0.11$; $\bar{\Pi}^{diffusion}_{\{B,C\}} = 0.13$, $\bar{\Pi}^{tree}_{\{B,C\}} = 0.13$; $\bar{\Pi}^{diffusion}_{\{A,B,C\}} = 0.27$, $\bar{\Pi}^{tree}_{\{A,B,C\}} = 0.25$. Simulations are conducted using the following parameter values: $w_A = w_B = w_C = 0.03$, $b^A_B = b^A_C = b^B_C = b^A_{BC} = b^B_{AC} = b^C_{AB} = 0$, $e_A = e_B = e_C = 0$, $d_{AB} = d_{BA} = 0.03$, $d_{AC} = d_{CA} = 0.04$, $d_{BC} = d_{CB} = 0.05$ (Color figure online)

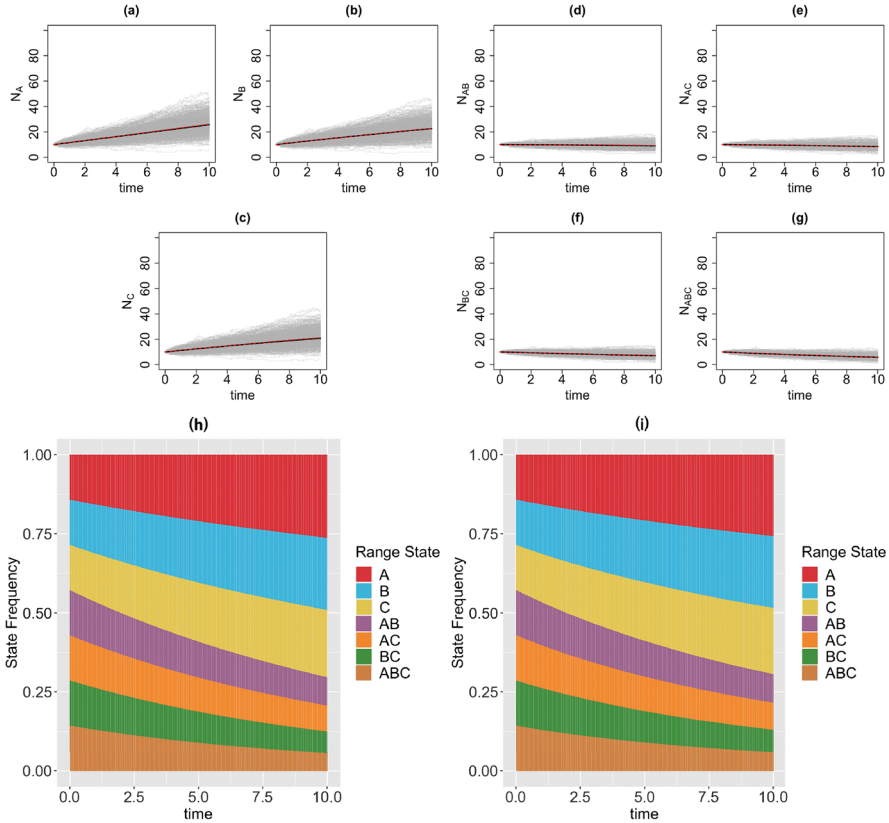

Fig. 7 Top & middle panels: the trajectories of average count of range states for endemic species (**a**–**c**) and widespread species (**d**–**g**) over [0, 10] time interval and over 1000 simulations runs for the three-region GeoSSE model as described in Example 3 simulated under both diffusion-based process (red line) and tree-based process (black line). The gray trajectories show the dynamics across 1000 replicates simulated under diffusion-based process. Bottom panel: stacked bar chart showing the state frequencies over time using diffusion-based approach (**h**) and tree-based approach (**i**). In both approaches, we start the process with $N(0) = 40$ and the following initial state frequencies: $\Pi_{\{A\}}(0) = \Pi_{\{B\}}(0) = \Pi_{\{C\}}(0) = \Pi_{\{A,B\}}(0) = \Pi_{\{A,C\}}(0) = \Pi_{\{B,C\}}(0) = \Pi_{\{A,B,C\}}(0) = \frac{1}{7}$, the mean frequencies for each range state from both diffusion-based and tree-based simulations are as follows: $\bar{\Pi}^{diffusion}_{\{A\}} = 0.26$, $\bar{\Pi}^{tree}_{\{A\}} = 0.26$; $\bar{\Pi}^{diffusion}_{\{B\}} = 0.23$, $\bar{\Pi}^{tree}_{\{B\}} = 0.23$; $\bar{\Pi}^{diffusion}_{\{C\}} = 0.21$, $\bar{\Pi}^{tree}_{\{C\}} = 0.21$; $\bar{\Pi}^{diffusion}_{\{A,B\}} = 0.09$, $\bar{\Pi}^{tree}_{\{A,B\}} = 0.09$; $\bar{\Pi}^{diffusion}_{\{A,C\}} = 0.08$, $\bar{\Pi}^{tree}_{\{A,C\}} = 0.08$; $\bar{\Pi}^{diffusion}_{\{B,C\}} = 0.07$, $\bar{\Pi}^{tree}_{\{B,C\}} = 0.07$; $\bar{\Pi}^{diffusion}_{\{A,B,C\}} = 0.06$, $\bar{\Pi}^{tree}_{\{A,B,C\}} = 0.06$. Simulations are conducted using the following parameter values: $w_A = w_B = w_C = 0.03$, $b^A_B = b^A_C = b^B_C = b^A_{BC} = b^B_{AC} = b^C_{AB} = 0$, $e_A = 0.01$, $e_B = 0.02$, $e_C = 0.025$, $d_{AB} = d_{BA} = d_{AC} = d_{CA} = d_{BC} = d_{CB} = 0$ (Color figure online)

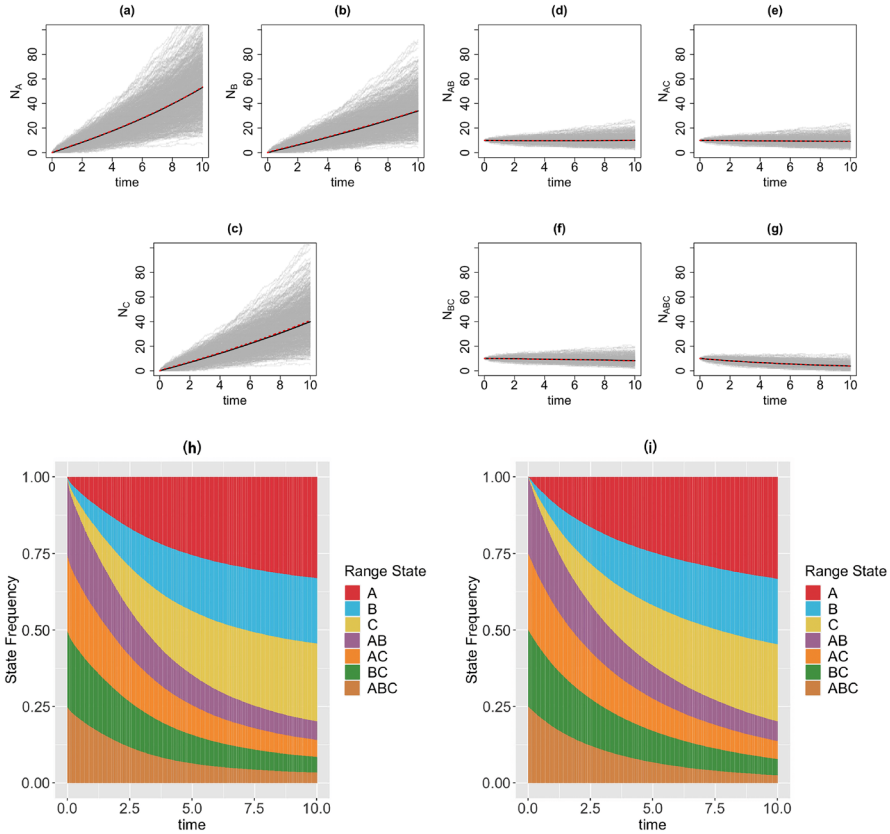

Fig. 8 Top & middle panels: the trajectories of average count of range states for endemic species (**a**–**c**) and widespread species (**d**–**g**) over [0, 10] time interval and over 1000 simulations runs for the three-region GeoSSE model as described in Example 4 simulated under the diffusion-based process (red line) and tree-based process (black line). The gray trajectories show the dynamics across 1000 replicates simulated under diffusion-based process. Bottom panel: stacked bar chart showing the state frequencies over time using diffusion-based approach (**h**) and tree-based approach (**i**). In both approaches, we start the process with $N(0) = 40$ and the following initial state frequencies: $\Pi_{\{A,B\}}(0) = \Pi_{\{A,C\}}(0) = \Pi_{\{B,C\}}(0) = \Pi_{\{A,B,C\}}(0) = \frac{1}{4}$ and $\Pi_{\{A\}}(0) = \Pi_{\{B\}}(0) = \Pi_{\{C\}}(0) = 0$. At $t = 10$, the mean frequencies for each range state from both diffusion-based and tree-based simulations are as follows: $\bar{\Pi}^{diffusion}_{\{A\}} = 0.33$, $\bar{\Pi}^{tree}_{\{A\}} = 0.33$; $\bar{\Pi}^{diffusion}_{\{B\}} = 0.21$, $\bar{\Pi}^{tree}_{\{B\}} = 0.21$; $\bar{\Pi}^{diffusion}_{\{C\}} = 0.25$, $\bar{\Pi}^{tree}_{\{C\}} = 0.25$; $\bar{\Pi}^{diffusion}_{\{A,B\}} = 0.06$, $\bar{\Pi}^{tree}_{\{A,B\}} = 0.06$; $\bar{\Pi}^{diffusion}_{\{A,C\}} = 0.06$, $\bar{\Pi}^{tree}_{\{A,C\}} = 0.06$; $\bar{\Pi}^{diffusion}_{\{B,C\}} = 0.05$, $\bar{\Pi}^{tree}_{\{B,C\}} = 0.05$; $\bar{\Pi}^{diffusion}_{\{A,B,C\}} = 0.03$, $\bar{\Pi}^{tree}_{\{A,B,C\}} = 0.03$. Simulations are conducted using the following parameter values: $w_A = 0.09$, $w_B = 0.06$, $w_C = 0.07$, $b^A_B = b^A_C = b^B_C = b^A_{BC} = b^B_{AC} = b^C_{AB} = 0.04$, $e_A = 0.002$, $e_B = 0.003$, $e_C = 0.001$, $d_{AB} = d_{BA} = 0.006$, $d_{AC} = d_{CA} = 0.003$, $d_{BC} = d_{CB} = 0.001$ (Color figure online)

3. We find the 95% confidence intervals of expected state counts at the end time for both diffusion and tree-based simulations for each GeoSSE scenario described above. Then, we apply the Welch's unequal variances t-test (Welch 1947) for testing the following hypothesis

$$H_0 : \bar{\mu}_{N_i,\text{tree}} = \bar{\mu}_{N_i,\text{diffusion}}$$
$$H_1 : \bar{\mu}_{N_i,\text{tree}} \neq \bar{\mu}_{N_i,\text{diffusion}},$$

where $\bar{\mu}_{N_i,\text{tree}}$ and $\bar{\mu}_{N_i,\text{diffusion}}$ are population means of state counts for range i at the end time from tree and diffusion-based approaches, respectively.

2. We also conduct the F test for testing the following hypothesis

$$H_0 : \bar{\sigma}^2_{N_i,\text{tree}} = \bar{\sigma}^2_{N_i,\text{diffusion}}$$
$$H_1 : \bar{\sigma}^2_{N_i,\text{tree}} \neq \bar{\sigma}^2_{N_i,\text{diffusion}},$$

where $\bar{\sigma}^2_{N_i,\text{tree}}$ and $\bar{\sigma}^2_{N_i,\text{diffusion}}$ are population variances of state counts for range i at the end time from tree and diffusion-based approaches, respectively.

5. We compute ratio of two sample variances for range state i as

$$r_{i,\text{var}} = \frac{s^2_{i,\text{diffusion}}}{s^2_{i,\text{tree}}},$$

where $s^2_{i,\text{diffusion}}$ and $s^2_{i,\text{tree}}$ are sample variances from diffusion- and tree-based simulations for range state i, respectively. Then, we construct the 95% confidence interval for $r_{i,\text{var}}$.
If the diffusion-based and tree-based simulation methods are statistically indistinguishable, we should fail to reject all null hypotheses and that the confidence intervals of the ratios of variances include the value 1 at the appropriate significance levels.
While all the diffusion-based simulations presented in the main text assume that we always start with a relatively large clade size, this is not how phylogenetic trees are normally simulated. Instead, most simulations generate the entire clade, beginning with one stem or two sister lineages to represent the origin of the process. However, the diffusion approximation assumes the number of species is large. Therefore, to adapt our diffusion-based model for clade-generation scenarios where the initial number of species is small, we adapted our diffusion-based simulation method to start the process with a single species in a random state (see "Appendix 5.6"). We show that the difference between diffusion-based and tree-based simulations is reduced after applying the correction.

2.6 Deriving Rate Parameters that Lead to Stationary State Frequencies When *N* is Large

In this section, we derive conditions for the rate parameters such that there is no change in state frequency, Π_i, over time for a given a range state $i \in S$, assuming large N. That is, we derive the conditions when $\frac{d\Pi_i}{dt} = 0, \forall i \in S$.

Knowing that $\Pi_i = \frac{N_i}{N}$, we re-write Eqs. (13)–(14) as follows,

$$\begin{aligned}\mathbb{P}_i^+ &= N\left[\underbrace{\sum_{j\in S}\sum_{\substack{\ell\in j\\ \{\ell\}=i}}\Pi_j w_\ell}_{\hat{W}_i^+} + \underbrace{\sum_{k\in i}\sum_{\substack{\ell\in i\\ \ell\neq k}}\Pi_{i\setminus\{\ell\}}d_{k\ell}}_{\hat{D}_i^+} + \underbrace{\sum_{\substack{j\in S\\ i\subset j}}\Pi_j b^i_{j\setminus i}}_{\hat{B}_i^+} + \underbrace{\sum_{\substack{j\in S\\ |j\setminus i|=1}}\sum_{\ell\in j\setminus i}\Pi_j e_\ell}_{\hat{E}_i^+}\right]\\ &= N\hat{\mathbb{P}}_i^+ \end{aligned} \tag{20}$$

$$\begin{aligned}\mathbb{P}_i^- &= N\left[\underbrace{0}_{\hat{W}_i^-} + \underbrace{\sum_{\substack{k\in i\\ |i|<R}}\sum_{\ell\in R\setminus\{k\}}\Pi_i d_{k\ell}}_{\hat{D}_i^-} + \underbrace{\sum_{\substack{j\in S\\ j\subset i}}\frac{1}{2}\Pi_i b^j_{i\setminus j}}_{\hat{B}_i^-} + \underbrace{\sum_{\substack{\ell\in R\\ \ell\in i}}\Pi_i e_\ell}_{\hat{E}_i^-}\right]\\ &= N\hat{\mathbb{P}}_i^- \end{aligned} \tag{21}$$

Then, Eqs. (7)–(8) can be re-written as follows

$$\mu_i = N\left(\hat{\mathbb{P}}_i^+ - \hat{\mathbb{P}}_i^-\right), \tag{22}$$

$$\sigma_i^2 = N\left(\hat{\mathbb{P}}_i^+ + \hat{\mathbb{P}}_i^-\right). \tag{23}$$

Given Eqs. (22)–(23), as $N \to \infty$, Eqs. (9)–(10) become

$$\begin{aligned}\hat{\mu}_{\Pi_i} &= \lim_{N\to\infty}\mu_{\Pi_i}\\ &= \hat{\mathbb{P}}_i^+ - \hat{\mathbb{P}}_i^-, \end{aligned} \tag{24}$$

$$\begin{aligned}\hat{\sigma}^2_{\Pi_i} &= \lim_{N\to\infty}\sigma^2_{\Pi_i}\\ &= 0. \end{aligned} \tag{25}$$

Moreover, we no longer have the stochastic component from the SDE given in Eq. (18). Instead, we solve the following ordinary differential equation

$$d\Pi_i = \hat{\mu}_{\Pi_i}dt$$

$$\frac{d\Pi_i}{dt} = \hat{\mu}_{\Pi_i}. \tag{26}$$

Given stationary frequency of each range state, $\hat{\Pi}_i$, where $\sum_i \hat{\Pi}_i = 1$, the rate parameters must satisfy

$$\hat{\mu}_{\Pi_i} = 0 \iff \hat{\mathbb{P}}_i^+ = \hat{\mathbb{P}}_i^-.$$

Furthermore, we assume all rate parameters must be positive, as all modeled events have some non-zero probability of occurring. That is,

$$w_i > 0, e_i > 0, d_{ij} > 0, \forall i, j \in R. \text{ and } b_t^s > 0, \forall s, t \in S$$

Next, we define total rates of all events occurring in each range state i, $\Phi_{total,i}$, as follows

$$\Phi_{total,i} = \left(r_{W_i^+} + r_{D_i^+} + r_{B_i^+} + r_{E_i^+}\right) - \left(r_{D_i^-} + r_{B_i^-} + r_{E_i^-}\right),$$

where $r_{W_i^+}, r_{D_i^+}, r_{B_i^+}, r_{E_i^+}, r_{D_i^-}, r_{B_i^-}, r_{E_i^-}$ consist of sums of rates across all adjacent states that correspond to the events $W_i^+, D_i^+, B_i^+, E_i^+, D_i^-, B_i^-, E_i^-$, respectively. $\Phi_{total,i}$ can also be thought as a flux for range state i. That is, it is a difference between total incoming rates and outgoing rates. For example, in a two-region GeoSSE model, we can define $\Phi_{total,\{A\}}$ as follows,

$$\Phi_{total,\{A\}} = \left(2w_A + 0 + b_B^A + e_B\right) - (d_{AB} + 0 + e_A),$$

where we have $r_{W_{\{A\}}^+} = 2w_A$ because within-region speciation rate w_A is acting on both endemic species with state $\{A\}$ and widespread species with state $\{A, B\}$.

Lemma 4 *Given a GeoSSE with state space S, set of stationary frequencies,* $\{\hat{\Pi}_i, \forall i \in S\}$*, and initial state frequencies* $\Pi_i(0)$*, the rate parameters satisfy the following system of equations*

$$\begin{aligned}
&\hat{\mathbb{P}}_i^+ = \hat{\mathbb{P}}_i^- \\
&\Phi_{total,i} \begin{cases} = \Phi_{total,j}, \ \textit{if } \hat{\Pi}_i = \hat{\Pi}_j \\ > \Phi_{total,j}, \ \textit{if } \hat{\Pi}_i > \hat{\Pi}_j \\ < \Phi_{total,j}, \ \textit{if } \hat{\Pi}_i < \hat{\Pi}_j \end{cases} \\
&\sum_{i \in S} \Pi_i(0) = 1 \\
&w_i > 0, e_i > 0, d_{ij} > 0, b_t^s > 0, \Pi_i(0) \geq 0, \ \forall i, j \in R \textit{ and } \forall s, t \in S.
\end{aligned} \tag{27}$$

In Sect. 3.2, we demonstrate the application of Lemma 4 for a 2-region GeoSSE model.

2.7 Deriving Stationary State Frequencies Given Rate Parameters in a GeoSSE Model

In this section, we use our framework to find the stationary state frequencies that result from a given set of rate parameters. This result links the configuration of a data-generating process to its expected pattern, which complements results from Sect. 2.6 that link expected patterns to data-generating processes. We present the result in Lemma 5 for the case of a 2-region GeoSSE model for simplicity.

Lemma 5 *Consider a 2-region GeoSSE model with state space* $S = \{\{A\}, \{B\}, \{A, B\}\}$. *Given the rate parameters from the model and initial state frequencies,* $\Pi_{\{A\}}(0) = \Pi^0_A$, $\Pi_{\{B\}}(0) = \Pi^0_B$, $\Pi_{\{A,B\}}(0) = \Pi^0_{AB}$, *the general solution to Eq.* (26) *is given by,*

$$\begin{aligned}\boldsymbol{\Pi} &= \begin{bmatrix}\Pi_{\{A\}}(t)\\ \Pi_{\{B\}}(t)\end{bmatrix}\\ &= C_1\boldsymbol{v}_1 e^{\lambda_1 t} + C_2\boldsymbol{v}_2 e^{\lambda_2 t} + \boldsymbol{K},\end{aligned} \tag{28}$$

and $\Pi_{\{A,B\}}(t) = 1 - \Pi_{\{A\}}(t) - \Pi_{\{B\}}(t)$, *provided that* $\Pi_{\{A\}}(t) + \Pi_{\{B\}}(t) \leq 1$.

Furthermore, the stationary frequencies are given by

$$\hat{\Pi}_{\{A\}} = \frac{num_A}{denom_A}, \tag{29}$$

$$\hat{\Pi}_{\{B\}} = 1 - \left(\frac{e_A + d_{AB} + b^A_B + e_B}{w_A + b^A_B + e_B}\right)\left(\frac{num_A}{denom_A}\right), \tag{30}$$

$$\hat{\Pi}_{\{A,B\}} = 1 - \hat{\Pi}_{\{A\}} - \hat{\Pi}_{\{B\}}, \tag{31}$$

where

$$\begin{aligned}num_A &= \left(w_A + b^A_B + e_B\right)(e_B + d_{BA} - w_B),\\ denom_A &= \left(e_A + d_{AB} + b^A_B + e_B\right)\left(e_B + d_{BA} + b^A_B + e_A\right)\\ &\quad - \left(w_B + b^A_B + e_A\right)\left(w_A + b^A_B + e_B\right),\\ R &= \sqrt{R_1 + R_2},\\ R_1 &= 4\left(b^A_B\right)^2 + 4\left(b^A_B e_A + b^A_B e_B + b^A_B w_A + b^A_B w_B\right)\\ &\quad + 4\left(e_A e_B + e_A w_A + e_B w_B + w_A w_B\right),\\ R_2 &= -2d_{AB}d_{BA} + \left(d^2_{AB} + d^2_{BA}\right),\\ \lambda_1 &= \frac{1}{2}\left(-2b^A_B - d_{AB} - d_{BA} - 2e_A - 2e_B - R\right),\\ \lambda_2 &= \frac{1}{2}\left(-2b^A_B - d_{AB} - d_{BA} - 2e_A - 2e_B + R\right),\end{aligned}$$

$$\boldsymbol{v}_1 = \begin{bmatrix} -\frac{1}{2\left(b_B^A + e_A + w_B\right)} \left(-d_{AB} + d_{BA} - R\right) \\ 1 \end{bmatrix},$$
$$\boldsymbol{v}_2 = \begin{bmatrix} -\frac{1}{2\left(b_B^A + e_A + w_B\right)} \left(-d_{AB} + d_{BA} + R\right) \\ 1 \end{bmatrix},$$
$$\boldsymbol{K} = \begin{bmatrix} \hat{\Pi}_{\{A\}} \\ \hat{\Pi}_{\{B\}} \end{bmatrix},$$

$$C_1 = \frac{\left(\Pi_A^0 - K_1\right)\left(b_B^A + e_A + w_B\right)}{R} - \frac{\left(\Pi_B^0 - K_2\right)\left(d_{AB} - d_{BA} - R\right)}{2R},$$
$$C_2 = \frac{\left(K_1 - \Pi_A^0\right)\left(b_B^A + e_A + w_B\right)}{R} + \left(\Pi_B^0 - K_2\right)\left(1 + \frac{d_{AB} - d_{BA} - R}{2R}\right).$$

Proof Proof of Lemma 5 is given in "Appendix 5.4". □

We note that this strategy can be generalized to accommodate arbitrary models within the ClaSSE family. Specifically, as seen in the proof of Lemma 5 in "Appendix 5.4", for a ClaSSE model with $|S|$ states, one only needs to find eigenvalues (either numerically or analytically) and eigenvectors that correspond to a $(|S| - 1) \times (|S| - 1)$ matrix to obtain a general solution. The resulting solution for the stationary frequencies would then reflect the parameterization of the particular ClaSSE model variant being studied. Note that this approach of solving a matrix with one dimension lower than the state space only holds providing that the sum of the remaining frequencies is less than or equal to 1. This assumption, however, can be ignored if one is to solve the full system by finding eigenvalues and eigenvectors that correspond to a $|S| \times |S|$ matrix, and normalize the resulting stationary frequencies.

In Sect. 3.2, we use Lemma 5 using rates obtained from Lemma 4 to verify that the system, indeed, converges to the true stationary frequencies that we observe through simulations.

2.8 Deriving Time to Reach Stationary State Frequencies in a GeoSSE Model

In this section, we describe a method for deriving time to reach stationary state frequencies in a 2-region GeoSSE model. Note that we have assumed a relatively large clade size at the start of the process for simulating $\Pi_i(t)$. Thus, the following is time to stationary frequencies since some relatively large clade size (Fig. 4).

From Lemma 5 in Sect. 2.7, we have derived an analytical expression to compute state frequencies over time, given large N. In order to find the time to stationarity for each range state, we define the following procedure, as follows

1. Given the initial state frequencies, Π_A^0, Π_B^0, Π_{AB}^0, and that the system runs from $[0, T]$, we find the mixing time t_i^* for all $i \in S$ such that,

$$\left|\Pi_i\left(t_i^*\right) - \Pi_i\left(t_i^* - \Delta t\right)\right| < \epsilon, \tag{32}$$

for some $\Delta t > 0$ and $\epsilon > 0$. t_i^* is the stationary time for the range state i, given the ϵ value.

2. We visually check that t_i^* derived from the theory reconciles with what we observe from simulations.

We apply this procedure to an example in Sect. 3.2.

3 Results

3.1 Diffusion-Based Approach is a Good Approximation to Tree-Based Approach for Describing State Dynamics

In this section, we visualize the range state dynamics using tree-based and diffusion-based approaches under several GeoSSE scenarios described in Sect. 2.5 (Figs. 5, 6, 7, 8). In all these scenarios, we show that the null hypothesis that the average counts of the ranges states at the end of the simulation time between these approaches are equal cannot be rejected (Table 1). This shows that the diffusion-based approach is a good approximation for means to the tree-based approach.

In most cases, we observe that data (state counts and frequencies) simulated under diffusion-based approach relatively have higher variances compared to data simulated under tree-based approach (Table 1). The 95% confidence interval for the ratio of two variances, shown in Table 1, gives an interval estimate on how much variation one would expect to get for generating state patterns under the diffusion process. Moreover, assuming that data simulated using the MASTER package (Vaughan and Drummond 2013) represent the true distribution of range state counts, this observation implies that diffusion process is not a good approximation for the second moment of the sampled state state frequencies. While this is not ideal, this is to be expected since diffusion is an approximation method to a generative model. Therefore, we should not expect state counts from both approaches to be drawn from the same distribution.

3.2 Multiple Rate Scenarios Lead to the Same Stationary State Frequencies

We apply the theoretical results from Sects. 2.6–2.8 for a 2-region GeoSSE model. The different sets of relationships between rate parameters given stationary frequencies in Example 5 are derived using Mathematica (Wolfram Research Inc 2023). In this example, we show that there exist alternative rate scenarios leading to the same stationary frequencies. Furthermore, using Lemma 5, we confirm that the stationary frequencies observed from simulations converge to the theoretical frequencies given the rate parameters, which are derived using Lemma 4. Using the procedure described in Sect. 2.8, we compute time to stationary frequencies in Example 5 for each rate scenario and different sets of initial frequencies.

Example 5 We consider a 2-region GeoSSE model with range state space $S = \{\{A\}, \{B\}, \{A, B\}\}$. We find a set of rate parameters and initial state frequencies that give the following stationary range state frequencies,

Table 1 The sample mean count for each range state at the end of simulation time, $\bar{N}_{i,end}$, computed under tree-based and diffusion-based simulations across different GeoSSE scenarios described in Sect. 2.5

Range state	$\bar{N}_{i,end}$		Lower bound		Upper bound		p_{mean}	p_{var}	95% CI $r_{i,var}$
	Tree	Diffusion	Tree	Diffusion	Tree	Diffusion			
Example 1: GeoSSE with within-region and between-region speciation events									
$\{A\}$	40.835	40.897	40.517	40.156	41.153	41.638	0.880	$\ll 0.001$	[4.794, 6.144]
$\{B\}$	40.240	40.544	39.906	39.842	40.574	41.246	0.444	$\ll 0.001$	[3.908, 5.008]
$\{C\}$	39.875	40.234	39.555	39.518	40.195	40.950	0.370	$\ll 0.001$	[4.409, 5.651]
$\{A, B\}$	5.980	5.981	5.858	5.837	6.102	6.125	0.992	$\ll 0.001$	[1.219, 1.562]
$\{A, C\}$	6.239	6.305	6.107	6.128	6.371	6.481	0.558	$\ll 0.001$	[1.587, 2.033]
$\{B, C\}$	6.506	6.625	6.391	6.494	6.621	6.756	0.182	$\ll 0.001$	[1.139, 1.459]
$\{A, B, C\}$	1.185	1.112	1.121	1.048	1.25	1.176	0.115	0.782	[0.899, 1.152]
Example 2: GeoSSE with within-region speciation and dispersal events									
$\{A\}$	14.689	14.642	14.450	14.282	14.928	15.002	0.831	$\ll 0.001$	[2.002, 2.566]
$\{B\}$	13.960	13.993	13.728	13.637	14.192	14.349	0.879	$\ll 0.001$	[2.076, 2.661]
$\{C\}$	13.339	13.387	13.109	13.035	13.568	13.739	0.823	$\ll 0.001$	[2.082, 2.669]
$\{A, B\}$	8.870	8.740	8.707	8.512	9.033	8.968	0.363	$\ll 0.001$	[1.732, 2.220]
$\{A, C\}$	11.107	10.870	10.930	10.596	11.284	11.144	0.155	$\ll 0.001$	[2.113, 2.709]
$\{B, C\}$	13.175	13.172	12.985	12.853	13.365	13.491	0.987	$\ll 0.001$	[2.482, 3.182]

Table 1 continued

Range state	$\bar{N}_{i,end}$		Lower bound		Upper bound		p_{mean}	p_{var}	95% CI $r_{i,var}$
	Tree	Diffusion	Tree	Diffusion	Tree	Diffusion			
$\{A, B, C\}$	24.427	24.790	24.189	24.172	24.665	25.408	0.283	$\ll 0.001$	[5.968, 7.649]
Example 3: GeoSSE with within-region speciation and local extinction events									
$\{A\}$	25.672	25.950	25.385	25.550	25.959	26.350	0.269	$\ll 0.001$	[1.714, 2.196]
$\{B\}$	22.540	22.630	22.266	22.262	22.814	22.998	0.701	$\ll 0.001$	[1.592, 2.040]
$\{C\}$	20.804	21.179	20.536	20.825	21.072	21.533	0.098	$\ll 0.001$	[1.547, 1.983]
$\{A, B\}$	8.960	9.105	8.843	8.973	9.077	9.237	0.108	< 0.001	[1.111, 1.425]
$\{A, C\}$	8.467	8.370	8.355	8.244	8.579	8.496	0.260	< 0.001	[1.109, 1.421]
$\{B, C\}$	7.007	7.024	6.902	6.911	7.112	7.137	0.829	0.027	[1.016, 1.302]
$\{A, B, C\}$	5.805	5.811	5.703	5.711	5.907	5.911	0.934	0.469	[0.844, 1.081]
Example 4: GeoSSE with full events									
$\{A\}$	53.067	53.494	52.420	52.235	53.714	54.753	0.555	$\ll 0.001$	[3.347, 4.290]
$\{B\}$	33.919	34.425	33.472	33.575	34.366	35.275	0.302	$\ll 0.001$	[3.193, 4.092]
$\{C\}$	39.981	41.044	39.476	40.060	40.486	42.028	0.060	$\ll 0.001$	[3.353, 4.297]
$\{A, B\}$	10.096	10.193	9.942	9.968	10.250	10.418	0.486	$\ll 0.001$	[1.880, 2.409]
$\{A, C\}$	9.229	9.224	9.091	9.028	9.367	9.420	0.967	$\ll 0.001$	[1.772, 2.271]
$\{B, C\}$	8.309	8.138	8.181	7.969	8.437	8.307	0.115	$\ll 0.001$	[1.526, 1.956]
$\{A, B, C\}$	3.897	3.890	3.789	3.767	4.005	4.013	0.933	$\ll 0.001$	[1.149, 1.472]

The “Lower bound" and “Upper bound" represent the 95% confidence interval of the average count for each range state using diffusion and tree based approaches. The “95% CI $r_{i,var}$" correspond to the 95% confidence interval of the ratio of two sample variances from diffusion and tree based approaches for range state i. p_{var} and p_{mean} correspond to p value from the F test and the Welch’s unequal variances t-test, respectively

$$\hat{\Pi}_{\{A\}} = \frac{1}{3}, \hat{\Pi}_{\{B\}} = \frac{1}{3}, \hat{\Pi}_{\{A,B\}} = \frac{1}{3}.$$

That is, by Eq. (27), we have,

$$\begin{aligned}
\frac{2}{3}w_A + \frac{1}{3}b_B^A + \frac{1}{3}e_B &= \frac{1}{3}(d_{AB} + e_A) \\
\frac{2}{3}w_B + \frac{1}{3}b_B^A + \frac{1}{3}e_A &= \frac{1}{3}(d_{BA} + e_B) \\
\frac{1}{3}(d_{AB} + d_{BA}) &= \frac{1}{3}(b_B^A + e_A + e_B) \\
2w_A + b_B^A + e_B - e_A - d_{AB} &= 2w_B + b_B^A + e_A - e_B - d_{BA} \\
2w_B + b_B^A + e_A - e_B - d_{BA} &= d_{AB} + d_{BA} - b_B^A - (e_A + e_B) \\
\sum_{i \in S} \Pi_i(0) &= 1, \ \Pi_{\{A\}}(0), \Pi_{\{B\}}(0), \Pi_{\{A,B\}}(0) \geq 0 \\
w_A, w_B, e_A, e_B, d_{AB}, d_{BA}, b_B^A &> 0.
\end{aligned} \tag{33}$$

We found a set of solutions to Eq. (33). That is,

$$\begin{aligned}
w_A &= \frac{1}{2}\left(-2b_B^A + 2d_{AB} + d_{BA} - 2e_B\right) \\
w_B &= \frac{1}{2}\left(-d_{AB} + 2e_B\right) \\
e_A &= -b_B^A + d_{AB} + d_{BA} - e_B
\end{aligned}$$

$$\begin{aligned}
&0 < b_B^A \leq d_{AB} - e_B, \ e_B < d_{AB} < 2e_B \\
&d_{BA} > 0, \ e_B > 0.
\end{aligned} \tag{34}$$

Another set of solutions is given by,

$$\begin{aligned}
w_A &= \frac{1}{2}\left(-2b_B^A + 2d_{AB} + d_{BA} - 2e_B\right) \\
w_B &= \frac{1}{2}\left(-d_{AB} + 2e_B\right) \\
e_A &= -b_B^A + d_{AB} + d_{BA} - e_B
\end{aligned}$$

$$\begin{aligned}
&b_B^A > 0, \ 0 < d_{AB} \leq e_B \\
&d_{BA} > 2\left(b_B^A - d_{AB} + e_B\right), \ e_B > 0.
\end{aligned} \tag{35}$$

Next, we simulate the range state dynamics, shown in Fig. 9, using the method

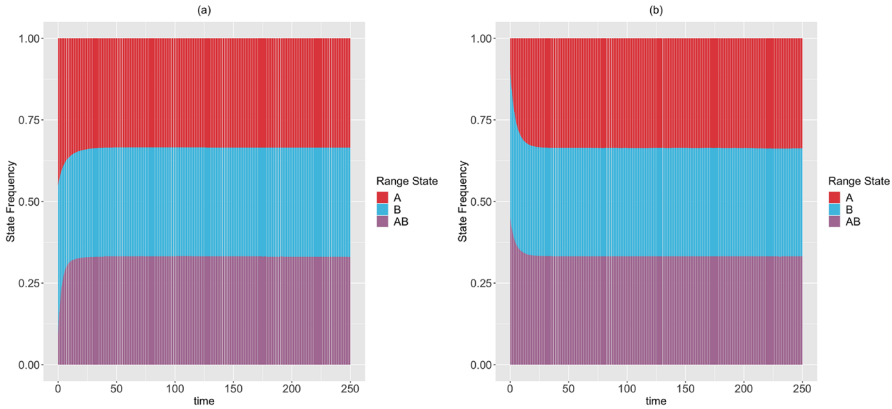

Fig. 9 The expected range state dynamics over [0, 250] time interval and over 100 trajectories for the two-region GeoSSE model as described in Example. 5. Each process is simulated under the following initial state frequencies and rate parameters according to Eq. (34): (Left panel) $\Pi_{\{A\}}(0) = \Pi_{\{B\}}(0) = 0.45, \Pi_{\{A,B\}}(0) = 0.1, w_A = 0.090, e_A = 0.176, w_B \approx 0, e_B = 0.008, d_{AB} = 0.015, d_{BA} = 0.173, b_B^A = 0.004$; (Right panel) $\Pi_{\{A\}}(0) = 0.1, \Pi_{\{B\}}(0) = \Pi_{\{A,B\}}(0) = 0.45, w_A = 0.160, e_A = 0.315, w_B = 0.002, e_B = 0.009, d_{AB} = 0.014, d_{BA} = 0.310, b_B^A = 0.001$. In both panels, $\mathbb{E}(\hat{\Pi}_{\{A\}}) \to \frac{1}{3}, \mathbb{E}(\hat{\Pi}_{\{B\}}) \to \frac{1}{3}, \mathbb{E}(\hat{\Pi}_{\{A,B\}}) \to \frac{1}{3}$. Using Lemma 5, we confirm that these expected stationary frequencies from simulations converge to the theoretical, and true stationary frequencies given these sets of rates. Furthermore, using the procedure described in Sect. 2.8 with $\epsilon = 10^{-9}$, we found that the stationary frequencies are reached at: $t_A^* = 114.114$, $t_B^* = 111.862, t_{AB}^* = 102.603$ (Left panel); $t_A^* = 76.827$, $t_B^* = 75.576, t_{AB}^* = 70.320$ (Right panel) (Color figure online)

described in Sect. 2.5 and rate parameters chosen according to Eq. (34).

To show that there are multiple rate scenarios that lead to the same stationary distribution, we simulate the range state dynamics, shown in Fig. 10, using rate parameters that satisfy the alternative set of solutions described in Eq. (35), but do not satisfy Eq. (34).

3.3 Comparing Our Method of Computing Stationary State Frequencies with Existing Literature

In this section, we compare our method for computing stationary state frequencies from rate parameters introduced in Sect. 2.7 with another method used in diversitree package (FitzJohn 2012) for the ClaSSE (Goldberg and Igić 2012) and GeoSSE (Goldberg et al. 2011) models. Although the technique used in diversitree has not been discussed in any SSE papers, such as the papers introducing the MuSSE (FitzJohn 2012), ClaSSE (Goldberg and Igić 2012), and GeoSSE (Goldberg et al. 2011) models, the technique applies projection matrix models that are widely used in the context of population biology to obtain ClaSSE and GeoSSE stationary frequencies (pers. comm. E. E. Goldberg and R. FitzJohn). Originally developed for applications in discrete-time models with either size-structured or age-structured population (Van Groenendael et al. 1988), this approach has also been adapted for continuous-time models with

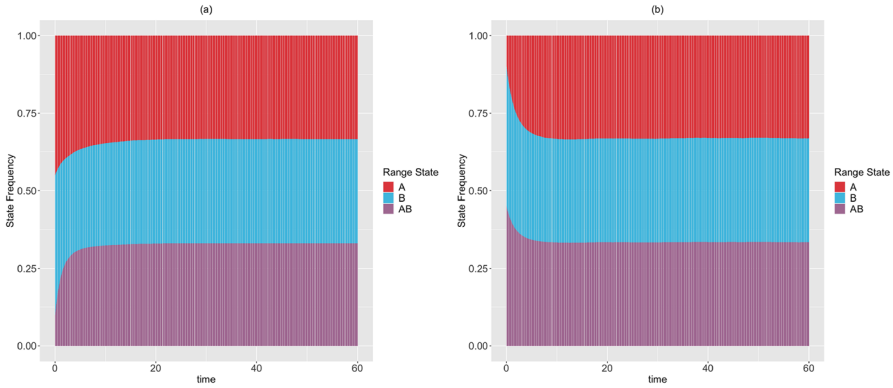

Fig. 10 The expected range state dynamics over [0, 60] time interval and over 100 trajectories for the two-region GeoSSE model as described in Example. 5. Each process is simulated under the following initial state frequencies and rate parameters according to Eq. (35): (Left panel) $\Pi_{\{A\}}(0) = \Pi_{\{B\}}(0) = 0.45$, $\Pi_{\{A,B\}}(0) = 0.1$, $w_A = 0.107$, $e_A = 0.309$, $w_B = 0.008$, $e_B = 0.008$, $d_{AB} = 0.001$, $d_{BA} = 0.405$, $b_B^A = 0.089$; (Right panel) $\Pi_{\{A\}}(0) = 0.1$, $\Pi_{\{B\}}(0) = \Pi_{\{A,B\}}(0) = 0.45$, $w_A = 0.049$, $e_A = 0.470$, $w_B = 0.005$, $e_B = 0.008$, $d_{AB} = 0.006$, $d_{BA} = 0.843$, $b_B^A = 0.371$. In both panels, $\mathbb{E}(\hat{\Pi}_{\{A\}}) \to \frac{1}{3}$, $\mathbb{E}(\hat{\Pi}_{\{B\}}) \to \frac{1}{3}$, $\mathbb{E}(\hat{\Pi}_{\{A,B\}}) \to \frac{1}{3}$. Using Lemma 5, we confirm that these expected stationary frequencies from simulations converge to the theoretical, and true stationary frequencies given these sets of rates. Furthermore, using the procedure described in Sect. 2.8 with $\epsilon = 10^{-9}$, we found that the stationary frequencies are reached at: $t_A^* = 53.153$, $t_B^* = 51.952$, $t_{AB}^* = 48.048$ (Left panel); $t_A^* = 30.781$, $t_B^* = 30.330$, $t_{AB}^* = 28.378$ (Right panel) (Color figure online)

the latter structured population (Kapur 1979). Under this approach, one would create a square matrix with entries that map the state of a structured population from one time to the next. Then, the dominant eigenvalue of such matrix represents the overall population growth rate with its eigenvector represents the stable stage distribution (Van Groenendael et al. 1988).

Through examples below we find that our method returns similar state frequencies to those computed under the projection matrix model in diversitree package (FitzJohn 2012). For example, under the following rate parameters in a two-region GeoSSE model,

$$w_A = 0.01,\ w_B = 0.02,\ b_B^A = 0.003,\ e_A = 0.169,\ e_B = 0.008,$$
$$d_{AB} = 0.002,\ d_{BA} = 0.178,$$

our method gives $\hat{\Pi}_{\{A\}} \approx 0.057$, $\hat{\Pi}_{\{B\}} \approx 0.506$, $\hat{\Pi}_{\{A,B\}} \approx 0.437$ while the projection matrix approach implemented in diversitree returns $\hat{\Pi}_{\{A\}} \approx 0.055$, $\hat{\Pi}_{\{B\}} \approx 0.490$, $\hat{\Pi}_{\{A,B\}} \approx 0.455$. Another example using the following rate parameters,

$$w_A \approx 0.0006,\ w_B \approx 0.0003,\ b_B^A \approx 0,\ e_A \approx 0.0048,$$
$$e_B \approx 0.0045,\ d_{AB} \approx 0.0370,\ d_{BA} \approx 0.03703,$$

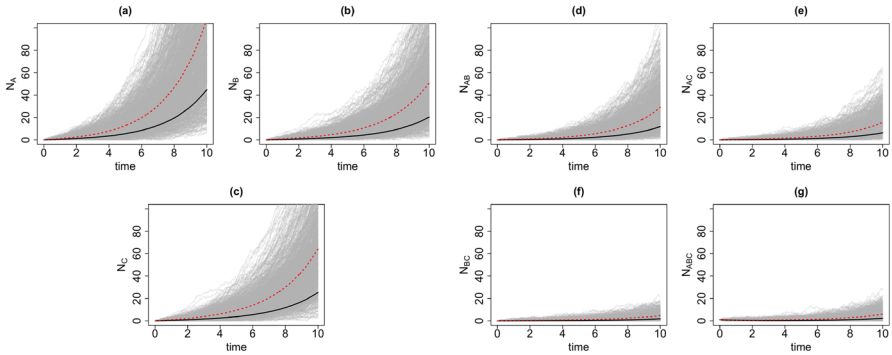

Fig. 11 The trajectories of mean counts of range states for endemic species (**a**–**c**) and widespread species (**d**–**g**) over the [0, 10] time interval. Trajectories were simulated under the diffusion-based process (red line) and tree-based process (black line), starting with 1 species in a random state. For each starting state, we simulate 150 trajectories (1050 trajectories in total) for each approach. The gray trajectories show the dynamics across 1050 replicates simulated under diffusion-based process. Simulations are conducted using the following parameter values: $w_A = 0.36, w_B = 0.24, w_C = 0.28, b_B^A = b_C^A = b_C^B = b_{BC}^A = b_{AC}^B = b_{AB}^C = 0.16, e_A = 0.02, e_B = 0.03, e_C = 0.01, d_{AB} = d_{BA} = 0.12, d_{AC} = d_{CA} = 0.06, d_{BC} = d_{CB} = 0.02$ (Color figure online)

we have $\hat{\Pi}_{\{A\}} \approx 0.0996$, $\hat{\Pi}_{\{B\}} \approx 0.0996$, $\hat{\Pi}_{\{A,B\}} \approx 0.8008$ while the other method produces $\hat{\Pi}_{\{A\}} \approx 0.0997$, $\hat{\Pi}_{\{B\}} \approx 0.0997$, $\hat{\Pi}_{\{A,B\}} \approx 0.8006$.

4 Discussion and Conclusion

In our paper, we have constructed a general framework using diffusion processes to study state dynamics over time from a general state-dependent speciation and extinction model with both anagenetic and cladogenetic state transitions, making it suitable for studying members of the ClaSSE model family (Goldberg and Igić 2012; Magnuson-Ford and Otto 2012; Goldberg et al. 2011; Freyman and Höhna 2018). We have applied this framework under various diversification scenarios for the GeoSSE model (Goldberg et al. 2011), a special case of the ClaSSE model, as described in Sects. 2.4–2.5. Our framework can easily be applied to other discrete state-dependent diversification models, such as simpler BiSSE and MuSSE models (Maddison et al. 2007; FitzJohn 2012) and Markovian Binary Tree (MBT) models (Kontoleon 2006; Hautphenne et al. 2009; Soewongsono et al. 2023). Through simulations and statistical analyses, we have shown that state dynamics simulated under diffusion-based approach and tree-based approach are comparable, given that we start the simulations with relative large clade size (Figs. 5, 6, 7, 8, Table 1). We also obtain good agreement between diffusion-based and tree-based simulations when beginning the process with a single species in random state, after applying a model-based correction procedure ("Appendix 5.6", Fig. 12). We also show, using a statistical test, that our diffusion framework offers a good approximation for the mean of state counts. This result allows one to understand how data generating process i.e. rate parameters

from a diversification model can explain observed state patterns without using phylogenetic information. For inferring rates using empirical state data at present, this diffusion-based approach to simulate state dynamics could be treated as a way to validate whether rates estimated from biological datasets using phylogenetic methods are sensible.

Moreover, in Sects. 2.6–2.7, we have derived theoretical results to deduce the expected state frequencies generated by a set of rates, and what possible rates will generate a given set of expected state frequencies. These results are generalizable to accommodate a system having more states, and provide an alternative way to validate the correctness of SSE simulation and inference methods. Additionally, in Sect. 2.8, we described a procedure to compute the minimum time for an SSE process to reach stationarity in its state frequencies. We have applied these results for a 2-region GeoSSE model. As seen in Figs. 9 and 10, we showed that there exist multiple different rate scenarios that can lead to the same stationary behaviour of state pattern. Our framework also creates an alternative mathematical approach to tree-based models that could help establish conditions for which SSE model parameters are and are not identifiable.

We next plan to study the time for perturbed SSE models to reach stationarity. This would help biologists understand how evolutionary systems re-equilibrate and how long that re-equilibration takes following perturbation. In particular, we plan to apply this framework to study scenarios where SSE rates shift across time (Condamine et al. 2013; Quintero et al. 2023). Scenarios with time-heterogeneous rates are particularly interesting for GeoSSE model variants, mainly because regions experience changes in their features (e.g., region size, distance with nearby regions, separation types) over time. As studied in Landis et al. (2022) and Swiston and Landis (2023), paleogeographically-changing regional features should influence rates of speciation, extinction, and dispersal over time. Mathematical knowledge of expected state (range) frequencies for arbitrary biogeographical systems could help biodiversity researchers assess whether certain clades of regions are within or between states of equilibrium.

5 Appendix

5.1 Proof of Lemma 1

From Eq. (1), we compute $g'(N)$ and $g''(N)$,

$$\begin{aligned} g'(N) &= \frac{\partial g}{\partial N} = \frac{\partial \left(\sum_i h(n_i)\right)}{\partial N} \\ &= \frac{\partial \left(\sum_i h(n_i)\right)}{\partial n_i} \text{ (For each i, the other partial derivatives w.r.t} j \neq i \text{equals 0)} \\ &= \frac{\partial g}{\partial n_i} \\ &= \frac{\partial X}{\partial n_i}. \end{aligned} \tag{36}$$

Similarly, we have,

$$g''(N) = \frac{\partial^2 X}{\partial n_i^2}. \tag{37}$$

Now applying Theorem (1) to Eq. (1), we have,

$$\mu_X = \sum_i \frac{\partial X}{\partial n_i}\mu_i + \frac{1}{2}\sum_i \frac{\partial^2 X}{\partial n_i^2}\sigma_i^2, \tag{38}$$

$$\sigma_X^2 = \sum_i \left(\frac{\partial X}{\partial n_i}\right)^2 \sigma_i^2. \tag{39}$$

□

5.2 Proof of Lemma 2

By definition, following Eqs. 1.2–1.3 of Chapter 5 in Karlin and Taylor (1981), we define the infinitesimal mean μ_i and infinitesimal variance σ_i^2 of the stochastic process $\{N_i(t) : t > 0\}$ as follows,

$$\begin{aligned}
\mu_i &= \lim_{\Delta t \to 0} \frac{1}{\Delta t}\mathbb{E}\left(N_i(t+\Delta t) - N_i(t) | N_i(t) = N_i\right) \\
&= \lim_{\Delta t \to 0} \frac{1}{\Delta t}\{\mathbb{E}(N_i(t+\Delta t)) - \mathbb{E}(N_i(t)|N_i(t) = N_i)\} \\
&= \lim_{\Delta t \to 0} \frac{1}{\Delta t}\{\mathbb{E}(N_i(t+\Delta t)) - N_i\}, \\
\sigma_i^2 &= \lim_{\Delta t \to 0} \frac{1}{\Delta t}\mathbb{E}\left((N_i(t+\Delta t) - N_i(t))^2 | N_i(t) = N_i\right) \\
&= \lim_{\Delta t \to 0} \frac{1}{\Delta t}\mathbb{E}\left(N_i^2(t+\Delta t) - 2N_i(t)N_i(t+\Delta t) + N_i^2(t)|N_i(t)\right) \\
&= \lim_{\Delta t \to 0} \frac{1}{\Delta t}\{\mathbb{E}\left(N_i^2(t+\Delta t)\right) - 2N_i\mathbb{E}\left(N_i(t+\Delta t)\right) + N_i^2\}.
\end{aligned}$$

By definition of the first-order and second-order moments we have,

$$\begin{aligned}
\mathbb{E}(N_i(t+\Delta t)) &= (N_i+1)\mathbb{P}_i^+\Delta t + (N_i-1)\mathbb{P}_i^-\Delta t + (N_i)\mathbb{P}_i\Delta t \\
&= N_i\left(\mathbb{P}_i^+ + \mathbb{P}_i^- + \mathbb{P}_i\right)\Delta t + \left(\mathbb{P}_i^+ - \mathbb{P}_i^-\right)\Delta t \\
&= N_i + \left(\mathbb{P}_i^+ - \mathbb{P}_i^-\right)\Delta t, \\
\mathbb{E}(N_i^2(t+\Delta t) &= (N_i+1)^2\mathbb{P}_i^+\Delta t + (N_i-1)^2\mathbb{P}_i^-\Delta t + (N_i)^2\mathbb{P}_i\Delta t.
\end{aligned}$$

Thus,

$$\mu_i = \mathbb{P}_i^+ - \mathbb{P}_i^-, \tag{40}$$
$$\sigma_i^2 = \mathbb{P}_i^+ + \mathbb{P}_i^-. \tag{41}$$

□

5.3 Proof of Lemma 3

We compute the following and substitute to Eqs. (2)–(3).

$$\begin{aligned}
\frac{\partial \Pi_i}{\partial N_i} &= \frac{\partial}{\partial N_i}\left(\frac{N_i}{\sum_{\substack{k\in s\\ k\neq i}} N_k + N_i}\right)\\
&= \frac{\sum_{\substack{k\in S\\ k\neq i}} N_k}{\left(\sum_{k\in S} N_k\right)^2}\\
&= \frac{1-\Pi_i}{N},\\
\frac{\partial \Pi_i}{\partial N_j} &= \frac{\partial}{\partial N_j}\left(\frac{N_i}{\sum_{\substack{k\in s\\ k\neq i}} N_k + N_i}\right),\ j\neq i\\
&= -\frac{N_i}{N^2}\\
&= -\frac{\Pi_i}{N},\\
\frac{\partial^2 \Pi_i}{\partial N_i^2} &= \frac{\partial}{\partial N_i}\left(\frac{1-\Pi_i}{N}\right)\\
&= \frac{-\frac{\partial \Pi_i}{\partial N_i} N - (1-\Pi_i)\frac{\partial N}{\partial N_i})}{N^2}\\
&= \frac{-(1-\Pi_i)-(1-\Pi_i)}{N^2}\\
&= -\frac{2(1-\Pi_i)}{N^2},\\
\frac{\partial^2 \Pi_i}{\partial N_j^2} &= \frac{\partial}{\partial N_j}\left(-\frac{N_i}{N^2}\right),\ j\neq i\\
&= \frac{N_i(2N)}{N^4}\\
&= \frac{2\Pi_i N}{N^3}\\
&= \frac{2\Pi_i}{N^2}.
\end{aligned}$$

Thus,

$$
\begin{aligned}
\mu_{\Pi_i} &= \sum_{j\in S} \frac{\partial \Pi_i}{\partial N_j}\mu_j + \frac{1}{2}\sum_{j\in S}\frac{\partial^2 \Pi_i}{\partial N_j^2}\sigma_j^2 \\
&= \frac{\partial \Pi_i}{\partial N_i}\mu_i + \sum_{\substack{j\in S\\ j\neq i}} \frac{\partial \Pi_i}{\partial N_j}\mu_j + \frac{1}{2}\frac{\partial^2 \Pi_i}{\partial N_i^2}\sigma_i^2 + \frac{1}{2}\sum_{\substack{j\in S\\ j\neq i}}\frac{\partial^2 \Pi_i}{\partial N_j^2}\sigma_j^2 \\
&= \frac{1-\Pi_i}{N}\mu_i + \sum_{\substack{j\in S\\ j\neq i}} \frac{-\Pi_i}{N}\mu_j - \frac{1-\Pi_i}{N^2}\sigma_i^2 + \sum_{\substack{j\in S\\ j\neq i}}\frac{\Pi_i}{N^2}\sigma_j^2 \\
&= \left(\frac{1-\Pi_i}{N}\right)\left(\mu_i - \frac{\sigma_i^2}{N}\right) + \sum_{\substack{j\in S\\ j\neq i}} \frac{\Pi_i}{N}\left(-\mu_j + \frac{\sigma_j^2}{N}\right) \\
&= \frac{1}{N}\left(\mu_i - \frac{\sigma_i^2}{N}\right) + \frac{\Pi_i}{N}\sum_{j\in S}\left(-\mu_j + \frac{\sigma_j^2}{N}\right), \qquad (42)
\end{aligned}
$$

$$
\begin{aligned}
\sigma_{\Pi_i}^2 &= \sum_{j\in S}\left(\frac{\partial \Pi_i}{\partial N_j}\right)^2 \sigma_j^2 \\
&= \left(\frac{\partial \Pi_i}{\partial N_i}\right)^2 \sigma_i^2 + \sum_{\substack{j\in S\\ j\neq i}}\left(\frac{\partial \Pi_i}{\partial N_j}\right)^2 \sigma_j^2 \\
&= \left(\frac{1-\Pi_i}{N}\right)^2 \sigma_i^2 + \sum_{\substack{j\in S\\ j\neq i}}\left(\frac{\Pi_i}{N}\right)^2 \sigma_j^2 \\
&= \left(\frac{\sigma_i}{N}\right)^2 (1-2\Pi_i) + \left(\frac{\Pi_i}{N}\right)^2 \sum_{j\in S}\sigma_j^2, \qquad (43)
\end{aligned}
$$

where μ_i and σ_i^2 follow Eqs. (7)–(8), respectively. □

5.4 Proof of Lemma 5

We find $\hat{\Pi}_i$ such that $\lim_{t\to\infty}\Pi_i(t) = \hat{\Pi}_i$ for all $i \in S,\ S = \{\{A\}, \{B\}, \{A, B\}\}$.

For $i = \{A\}$ we have,

$$
\begin{aligned}
\hat{\mu}_{\Pi_A} &= \hat{\mathbb{P}}_A^+ - \hat{\mathbb{P}}_A^- \\
&= \left[w_A\left(\Pi_{\{A\}} + \Pi_{\{A,B\}}\right) + \Pi_{\{A,B\}} b_B^A + e_B \Pi_{\{A,B\}}\right] - \left[d_{AB}\Pi_{\{A\}} + e_A \Pi_{\{A\}}\right] \\
&= \Pi_{\{A\}}\left(w_A - e_A - d_{AB}\right) + \Pi_{\{A,B\}}\left(w_A + b_B^A + e_B\right). \qquad (44)
\end{aligned}
$$

For $i = \{B\}$ we have,

$$\begin{aligned}\hat{\mu}_{\Pi_B} &= \hat{\mathbb{P}}_B^+ - \hat{\mathbb{P}}_B^- \\ &= \left[w_B\left(\Pi_{\{B\}} + \Pi_{\{A,B\}}\right) + \Pi_{\{A,B\}} b_B^A + e_A \Pi_{\{A,B\}}\right] - \left[d_{BA}\Pi_{\{B\}} + e_B \Pi_{\{B\}}\right] \\ &= \Pi_{\{B\}}\left(w_B - e_B - d_{BA}\right) + \Pi_{\{A,B\}}\left(w_B + b_B^A + e_A\right). \end{aligned} \tag{45}$$

For $i = \{A, B\}$ we have,

$$\begin{aligned}\hat{\mu}_{\Pi_{AB}} &= \hat{\mathbb{P}}_{AB}^+ - \hat{\mathbb{P}}_{AB}^- \\ &= \left[d_{AB}\Pi_{\{A\}} + d_{BA}\Pi_{\{B\}}\right] - \left[b_B^A \Pi_{\{A,B\}} + (e_A + e_B)\Pi_{\{A,B\}}\right] \\ &= \left(d_{AB}\Pi_{\{A\}} + d_{BA}\Pi_{\{B\}}\right) - \Pi_{\{A,B\}}\left(b_B^A + e_A + e_B\right). \end{aligned} \tag{46}$$

Thus, we want to find the general solution for the following system of differential equations

$$\frac{d\Pi_{\{A\}}}{dt} = \Pi_{\{A\}}\left(w_A - e_A - d_{AB}\right) + \Pi_{\{A,B\}}\left(w_A + b_B^A + e_B\right), \tag{47}$$

$$\frac{d\Pi_{\{B\}}}{dt} = \Pi_{\{B\}}\left(w_B - e_B - d_{BA}\right) + \Pi_{\{A,B\}}\left(w_B + b_B^A + e_A\right), \tag{48}$$

$$\frac{d\Pi_{\{A,B\}}}{dt} = \left(d_{AB}\Pi_{\{A\}} + d_{BA}\Pi_{\{B\}}\right) - \Pi_{\{A,B\}}\left(b_B^A + e_A + e_B\right). \tag{49}$$

given initial state frequencies $\Pi_{\{A\}}(0) = \Pi_A^0$, $\Pi_{\{B\}}(0) = \Pi_B^0$, $\Pi_{\{A,B\}}(0) = \Pi_{AB}^0$.

However, since $\hat{\Pi}_{\{A\}} + \hat{\Pi}_{\{B\}} + \hat{\Pi}_{\{A,B\}} = 1$, we can always derive $\hat{\Pi}_{\{A,B\}}$ using $\hat{\Pi}_{\{A\}}$ and $\hat{\Pi}_{\{B\}}$. Therefore, we want to solve the following system instead.

$$\frac{d\Pi_{\{A\}}}{dt} = \Pi_{\{A\}}\left(w_A - e_A - d_{AB}\right) + \Pi_{\{A,B\}}\left(w_A + b_B^A + e_B\right), \tag{50}$$

$$\frac{d\Pi_{\{B\}}}{dt} = \Pi_{\{B\}}\left(w_B - e_B - d_{BA}\right) + \Pi_{\{A,B\}}\left(w_B + b_B^A + e_A\right). \tag{51}$$

Since $\Pi_{\{A\}}(t) + \Pi_{\{B\}}(t) + \Pi_{\{A,B\}}(t) = 1$, we have,

$$\begin{aligned}\frac{d\Pi_{\{A\}}}{dt} &= \Pi_{\{A\}}\left(-e_A - d_{AB} - b_B^A - e_B\right) - \Pi_{\{B\}}\left(w_A + b_B^A + e_B\right) \\ &\quad + \left(w_A + b_B^A + e_B\right), \end{aligned} \tag{52}$$

$$\begin{aligned}\frac{d\Pi_{\{B\}}}{dt} &= \Pi_{\{B\}}\left(-e_B - d_{BA} - b_B^A - e_A\right) - \Pi_{\{A\}}\left(w_B + b_B^A + e_A\right) \\ &\quad + \left(w_B + b_B^A + e_A\right). \end{aligned} \tag{53}$$

We write the above system in a matrix form as follows,

$$\mathbf{\Pi}' = M\mathbf{\Pi} + \boldsymbol{r}, \tag{54}$$

where

$$\begin{aligned}
\mathbf{\Pi}' &= \begin{bmatrix} \Pi'_{\{A\}} \\ \Pi'_{\{B\}} \end{bmatrix}, \ \mathbf{\Pi} = \begin{bmatrix} \Pi_{\{A\}} \\ \Pi_{\{B\}} \end{bmatrix}, \\
M &= \begin{bmatrix} -\left(e_A + d_{AB} + b_B^A + e_B\right) & -\left(w_A + b_B^A + e_B\right) \\ -\left(w_B + b_B^A + e_A\right) & -\left(e_B + d_{BA} + b_B^A + e_A\right) \end{bmatrix}, \\
\boldsymbol{r} &= \begin{bmatrix} w_A + b_B^A + e_B \\ w_B + b_B^A + e_A \end{bmatrix}.
\end{aligned} \tag{55}$$

First, we find the complimentary solution to the following equation using eigenvalues and eigenvectors of matrix M,

$$\mathbf{\Pi}' = M\mathbf{\Pi}. \tag{56}$$

Using Mathematica, the eigenvalues (λ_1, λ_2) and eigenvectors $(\boldsymbol{v}_1, \boldsymbol{v}_2)$ of M are given by,

$$\boldsymbol{\lambda} = \begin{bmatrix} \lambda_1 \\ \lambda_2 \end{bmatrix} \tag{57}$$

$$= \begin{bmatrix} \frac{1}{2}\left(-2b_B^A - d_{AB} - d_{BA} - 2e_A - 2e_B - R\right) \\ \frac{1}{2}\left(-2b_B^A - d_{AB} - d_{BA} - 2e_A - 2e_B + R\right) \end{bmatrix}, \tag{58}$$

$$\boldsymbol{v}_1 = \begin{bmatrix} -\frac{1}{2\left(b_B^A + e_A + w_B\right)}\left(-d_{AB} + d_{BA} - R\right) \\ 1 \end{bmatrix}, \tag{59}$$

$$\boldsymbol{v}_2 = \begin{bmatrix} -\frac{1}{2\left(b_B^A + e_A + w_B\right)}\left(-d_{AB} + d_{BA} + R\right) \\ 1 \end{bmatrix}, \tag{60}$$

where

$$\begin{aligned}
R &= \sqrt{R_1 + R_2}, \\
R_1 &= 4\left(b_B^A\right)^2 + 4\left(b_B^A e_A + b_B^A e_B + b_B^A w_A + b_B^A w_B\right) \\
&\quad + 4\left(e_A e_B + e_A w_A + e_B w_B + w_A w_B\right), \\
R_2 &= -2d_{AB}d_{BA} + \left(d_{AB}^2 + d_{BA}^2\right).
\end{aligned}$$

The complimentary solution for Eq. (56) is given by,

$$\mathbf{\Pi}_C = C_1 \boldsymbol{v}_1 e^{\lambda_1 t} + C_2 \boldsymbol{v}_2 e^{\lambda_2 t}, \tag{61}$$

where C_1 and C_2 are arbitrary constants.

Next, we find the particular solution $\mathbf{\Pi}_P$ for Eq. (54) using the method of undetermined coefficients. Suppose the solution $\mathbf{\Pi}_P$ is of the form

$$\mathbf{\Pi}_P = \begin{bmatrix} K_1 \\ K_2 \end{bmatrix}, \ \mathbf{\Pi}_P^{'} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}. \tag{62}$$

Substitute Eq. (62) to Eq. (54) we have,

$$\mathbf{\Pi}^{'}{}_P = M\mathbf{\Pi}_P + \boldsymbol{r}. \tag{63}$$

That is, we want to solve the following system of linear equations,

$$\left(e_A + d_{AB} + b_B^A + e_B\right) K_1 + \left(w_A + b_B^A + e_B\right) K_2 = w_A + b_B^A + e_B \tag{64}$$

$$\left(w_B + b_B^A + e_A\right) K_1 + \left(e_B + d_{BA} + b_B^A + e_A\right) K_2 = w_B + b_B^A + e_A. \tag{65}$$

Thus,

$$\begin{aligned} K_1 &= \frac{\left(w_A + b_B^A + e_B\right)(e_B + d_{BA} - w_B)}{\left(e_A + d_{AB} + b_B^A + e_B\right)\left(e_B + d_{BA} + b_B^A + e_A\right) - \left(w_B + b_B^A + e_A\right)\left(w_A + b_B^A + e_B\right)} \\ &= \frac{num_A}{denom_A}. \end{aligned} \tag{66}$$

Substitute Eq. (66) to Eq. (64) to get,

$$K_2 = 1 - \left(\frac{e_A + d_{AB} + b_B^A + e_B}{w_A + b_B^A + e_B}\right)\left(\frac{num_A}{denom_A}\right). \tag{67}$$

Therefore, the general solution is given by,

$$\begin{aligned} \mathbf{\Pi} &= \mathbf{\Pi}_C + \mathbf{\Pi}_P \\ &= C_1 \boldsymbol{v}_1 e^{\lambda_1 t} + C_2 \boldsymbol{v}_2 e^{\lambda_2 t} + \boldsymbol{K}, \end{aligned} \tag{68}$$

where

$$\boldsymbol{K} = \begin{bmatrix} K_1 \\ K_2 \end{bmatrix}. \tag{69}$$

By taking a limit of Eq. (68) as $t \to \infty$, the exponential terms in Eq. (68) will approach 0. Therefore, we have,

$$\hat{\mathbf{\Pi}} = \lim_{t \to \infty} \mathbf{\Pi}$$

$$\begin{bmatrix} \hat{\Pi}_{\{A\}} \\ \hat{\Pi}_{\{B\}} \end{bmatrix} = \begin{bmatrix} K_1 \\ K_2 \end{bmatrix}. \tag{70}$$

$\hat{\Pi}_{\{A\}}$ and $\hat{\Pi}_{\{B\}}$ from Eq. (70) are the stationary frequencies for state $\{A\}$ and $\{B\}$, respectively, as shown in Lemma 5.

Next, to get the general solution to the system in Eq. (54), we find the constants, C_1 and C_2 by substituting the initial value condition to Eq. (68) for $t = 0$. We have,

$$C_1 \begin{bmatrix} -\frac{1}{2(b_B^A+e_A+w_B)}\left(-d_{AB}+d_{BA}-R\right) \\ 1 \end{bmatrix} + C_2 \begin{bmatrix} -\frac{1}{2(b_B^A+e_A+w_B)}\left(-d_{AB}+d_{BA}+R\right) \\ 1 \end{bmatrix} + \begin{bmatrix} K_1 \\ K_2 \end{bmatrix} = \begin{bmatrix} \Pi_A^0 \\ \Pi_B^0 \end{bmatrix}. \tag{71}$$

That is, we solve the following system of linear equations,

$$C_1 \left(\frac{R + d_{AB} - d_{BA}}{2\left(b_B^A + e_A + w_B\right)} \right) + C_2 \left(\frac{d_{AB} - d_{BA} - R}{2\left(b_B^A + e_A + w_B\right)} \right) = \Pi_A^0 - K_1, \tag{72}$$

$$C_1 + C_2 = \Pi_B^0 - K_2. \tag{73}$$

Thus,

$$C_1 = \frac{\left(\Pi_A^0 - K_1\right)\left(b_B^A + e_A + w_B\right)}{R} - \frac{\left(\Pi_B^0 - K_2\right)\left(d_{AB} - d_{BA} - R\right)}{2R}. \tag{74}$$

Then, substitute Eq. (74) to Eq. (73) we get,

$$C_2 = \frac{\left(K_1 - \Pi_A^0\right)\left(b_B^A + e_A + w_B\right)}{R} + \left(\Pi_B^0 - K_2\right)\left(1 + \frac{d_{AB} - d_{BA} - R}{2R}\right). \tag{75}$$

□

5.5 Simulating Range State Dynamics Using MASTER

We can express events in MASTER for GeoSSE using the following reaction equations.

$$\hat{S}[\{i\}] \xrightarrow{e_i} \hat{R}[i] + L[i], \ (\textit{extinction of endemic species in region } i)$$

$$\hat{S}[\{i\}] \xrightarrow{d_{ij}} \hat{S}[\{i\} \cup \{j\}] + G[j], \ (\textit{dispersal from region } i \textit{ to region } j)$$

$$\hat{S}\left[\bigcup_{i \in R} \{i\}\right] \xrightarrow{e_j} \hat{S}\left[\bigcup_{i \in R; i \neq j} \{i\}\right] + L[j], \ (\textit{local extinction in region } j)$$

$$\hat{S}\left[\bigcup_{i \in R} \{i\}\right] \xrightarrow{b_j^i} \hat{S}\left[\bigcup_{i \in R; i \neq j} \{i\}\right] + \hat{S}[\{j\}], \ (\textit{between-region speciation into ranges} \{i\} \textit{and} \{j\})$$

$$\hat{S}\left[\bigcup_{i\in R}\{i\}\right] \xrightarrow{w_j} \hat{S}\left[\bigcup_{i\in R}\{i\}\right] + S[j] + G[j], \ (\textit{within region speciation in region } j \textit{ for a widespread species})$$

$$\hat{S}[\{i\}] \xrightarrow{w_i} \hat{S}[\{i\}] + \hat{S}[\{i\}] + G[i], \ (\textit{within region speciation in region } i \textit{ for an endemic species}),$$

where $\hat{S}[\{i\}]$ indicates the number of endemic species with range $\{i\}$, $\hat{S}\left[\bigcup_{i\in R}\{i\}\right]$ indicates the number of widespread species with range $\bigcup_{i\in R}\{i\}$, $\hat{R}[i]$ indicates the number of species in region i, $L[i]$ indicates a species lost in region i, and $G[i]$ indicates a species gain in region i.

5.6 Simulation Under the Diffusion Model Starting with a Single Species

We propose a procedure to correct the diffusion-based simulation for state dynamics when starting with a single species in random states under an SSE model, as demonstrated for a GeoSSE model. Note that without this correction procedure, the bias between the tree-based and diffusion-based approaches is apparent as described in Example 6.

Example 6 Consider a 3-region GeoSSE system with state space $S = \{\{A\}, \{B\}, \{C\}, \{A, B\}, \{A, C\}, \{B, C\}, \{A, B, C\}\}$. Suppose we start with a single species in state $\{A, B\}$.
That is, at $t = 0$ we have

$$N_{\{A\}} = 0,\ N_{\{B\}} = 0, N_{\{C\}} = 0,\ N_{\{A,B\}} = 1,\ N_{\{A,C\}} = 0,\ N_{\{B,C\}} = 0,\ N_{\{A,B,C\}} = 0.$$

It is very likely, given a small total species number, for diffusion-based simulation to give a pattern at the next time step such as the one below,

$$N_{\{A\}} = 0,\ N_{\{B\}} = 0, N_{\{C\}} = 1,\ N_{\{A,B\}} = 1,\ N_{\{A,C\}} = 0,\ N_{\{B,C\}} = 0,\ N_{\{A,B,C\}} = 0. \tag{76}$$

The transition described in Eq. (76) cannot be explained using just one GeoSSE event. Furthermore, since the diffusion parameters for simulating species count across states, described in Eqs. (7)–(8), dynamically depend on the number of species in each state, the difference between diffusion-based and tree-based simulations becomes more apparent over time (Fig. 11).

In general, the correction procedure works by preventing a diffusion-based path to enter a disallowed regime when there are few species (Table 2). This procedure only applies for simulating state counts. However, we can get the state frequencies from species counts in each state using the following formula,

$$\Pi_i(t) = \frac{N_i(t)}{\sum_{i\in S} N_i(t)}.$$

1. For each simulated trajectory for state counts under the diffusion-based approach where we start with a single species in a random state, we simulate the next path according to the procedure described in Sect. 2.5.

Table 2 List of allowed differences between two consecutive paths in a 3-region GeoSSE model

Condition	$N_i(t+\Delta t) - N_i(t)$						
	$\{A\}$	$\{B\}$	$\{C\}$	$\{A, B\}$	$\{A, C\}$	$\{B, C\}$	$\{A, B, C\}$
Within region speciation							
If $N_{\{A\}}(t) \neq 0,\ N_j(t) = 0, \forall j \in S \setminus \{A\}$	1	0	0	0	0	0	0
If $N_{\{B\}}(t) \neq 0,\ N_j(t) = 0, \forall j \in S \setminus \{B\}$	0	1	0	0	0	0	0
If $N_{\{C\}}(t) \neq 0,\ N_j(t) = 0, \forall j \in S \setminus \{C\}$	0	0	1	0	0	0	0
If $N_{\{A,B\}} \neq 0,\ N_{\{A,B,C\}} = 0$	1	0	0	0	0	0	0
	0	1	0	0	0	0	0
If $N_{\{A,C\}} \neq 0,\ N_{\{A,B,C\}} = 0$	1	0	0	0	0	0	0
	0	0	1	0	0	0	0
If $N_{\{B,C\}} \neq 0,\ N_{\{A,B,C\}} = 0$	0	1	0	0	0	0	0
	0	0	1	0	0	0	0
If $N_{\{A,B,C\}} \neq 0$	1	0	0	0	0	0	0
	0	1	0	0	0	0	0
	0	0	1	0	0	0	0
Extinction in species with range size 1							
	−1	0	0	0	0	0	0
	0	−1	0	0	0	0	0
	0	0	−1	0	0	0	0
Extinction in species with range size 2							
	1	0	0	−1	0	0	0
	1	0	0	0	−1	0	0
	0	1	0	−1	0	0	0
	0	1	0	0	0	−1	0

Table 2 continued

Condition	$N_i(t+\Delta t) - N_i(t)$						
	$\{A\}$	$\{B\}$	$\{C\}$	$\{A, B\}$	$\{A, C\}$	$\{B, C\}$	$\{A, B, C\}$
	0	0	1	0	−1	0	0
	0	0	1	0	0	−1	0
Extinction in species with range size 3							
	0	0	0	1	0	0	−1
	0	0	0	0	1	0	−1
	0	0	0	0	0	1	−1
Between-region speciation							
	1	1	0	−1	0	0	0
	1	0	1	0	−1	0	0
	0	1	1	0	0	−1	0
	1	0	0	0	0	1	−1
	0	1	0	0	1	0	−1
	0	0	1	1	0	0	−1
Dispersal in species with range size 1							
	−1	0	0	1	0	0	0
	−1	0	0	0	1	0	0
	0	−1	0	1	0	0	0
	0	−1	0	0	0	1	0
	0	0	−1	0	1	0	0
	0	0	−1	0	0	1	0
Dispersal in species with range size 2							
	0	0	0	−1	0	0	1
	0	0	0	0	−1	0	1
	0	0	0	0	0	−1	1

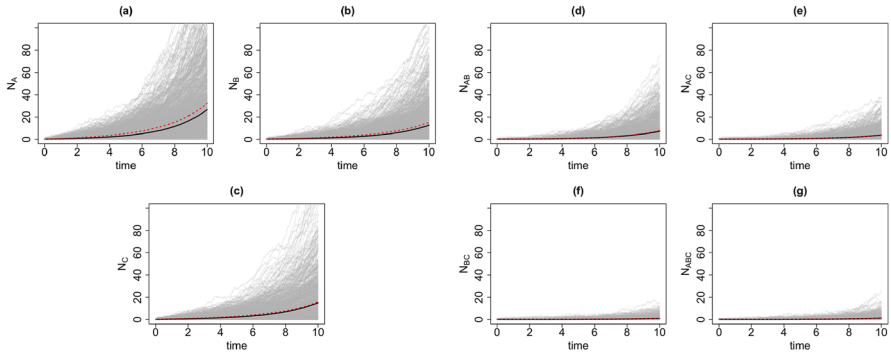

Fig. 12 The trajectories of average count of range states for endemic species (**a**–**c**) and widespread species (**d**–**g**) over the [0, 10] time interval simulated under the diffusion-based process (red line) after applying the correction procedure and tree-based process (black line), starting with 1 species in a random state. For each starting state, we simulate 150 trajectories (1050 trajectories in total) for each approach. The gray trajectories show the dynamics across 1050 replicates simulated under diffusion-based process. Simulations are conducted using the following parameter values: $w_A = 0.36$, $w_B = 0.24$, $w_C = 0.28$, $b_B^A = b_C^A = b_C^B = b_{BC}^A = b_{AC}^B = b_{AB}^C = 0.16$, $e_A = 0.02$, $e_B = 0.03$, $e_C = 0.01$, $d_{AB} = d_{BA} = 0.12$, $d_{AC} = d_{CA} = 0.06$, $d_{BC} = d_{CB} = 0.02$ (Color figure online)

2. Next, we compute the difference in paths at t and $t + \Delta t$. That is, we find

$$\boldsymbol{N}(t + \Delta t) - \boldsymbol{N}(t) = \{N_i(t + \Delta t) - N_i(t)\}_{\forall i \in S}, \tag{77}$$

   where S is the range state space of the GeoSSE model. Then, round each element of the vector in Eq. (77) to the nearest integer value.
3. We check whether this difference in paths is in the list described in Table 2. If it is not in the list, we reject this future path and re-sample until it satisfies the table. In effect, this correction approximates model-based conditional sampling for diffusion-based paths.
4. Repeat steps 1–2 until we reach a path where every state has at least 1 representative species, $\forall i : N_i(t) > 0$. Then, we generate the future paths as usual, according to the procedure described in Sect. 2.5.

As seen in Fig. 12, the difference between diffusion-based and tree-based simulation is reduced after applying the correction procedure. Note that it is possible to apply this correction procedure to any discrete-state SSE models by modifying the table (Table 2) according to events described by the model. Also, this procedure only minimizes the difference between diffusion-based and tree-based trajectories as it is still possible for a diffusion-based simulation to construct disallowed paths at a later time, despite the starting condition. However, these differences become negligible for as $N(t) >> 0$ for our example.

Acknowledgements We thank members of the Landis Lab for valuable feedback that improved the quality of this study. We also thank Emma Goldberg, Richard FitzJohn, and Mike May for directing us to literature used for computing stationary frequencies in the diversitree package.

Funding This research was funded by the National Science Foundation (NSF Award DEB-2040347), the Fogarty International Center at the National Institutes of Health (Award Number R01 TW012704) as part of the joint NIH-NSF-NIFA Ecology and Evolution of Infectious Disease program, and the Washington University Incubator for Transdisciplinary Research.

Data Availibility The datasets and all the relevant code are publicly available on https://github.com/alberts2/Diffusion_GeoSSE.git.

Declarations

Conflict of interest The authors declare that they have no Conflict of interest.

References

Alexander HK, Lambert A, Stadler T (2016) Quantifying age-dependent extinction from species phylogenies. Syst Biol 65(1):35–50

Bouckaert R, Heled J, Kühnert D et al (2014) Beast 2: a software platform for Bayesian evolutionary analysis. PLoS Comput Biol 10(4):e1003537

Celentano M, DeWitt WS, Prillo S et al (2024) Exact and efficient phylodynamic simulation from arbitrarily large populations. arXiv preprint arXiv:2402.17153

Chevin LM (2016) Species selection and random drift in macroevolution. Evolution 70(3):513–525

Condamine FL, Rolland J, Morlon H (2013) Macroevolutionary perspectives to environmental change. Ecol Lett 16:72–85

Dragomir D, Allman ES, Rhodes JA (2023) Parameter identifiability of a multitype pure-birth model of speciation. J Comput Biol 30(3):277–292

FitzJohn RG (2010) Quantitative traits and diversification. Syst Biol 59(6):619–633

FitzJohn RG (2012) Diversitree: comparative phylogenetic analyses of diversification in R. Methods Ecol Evol 3(6):1084–1092

Freyman WA, Höhna S (2018) Cladogenetic and anagenetic models of chromosome number evolution: a Bayesian model averaging approach. Syst Biol 67(2):195–215

Goldberg EE, Igić B (2012) Tempo and mode in plant breeding system evolution. Evolution 66(12):3701–3709

Goldberg EE, Lancaster LT, Ree RH (2011) Phylogenetic inference of reciprocal effects between geographic range evolution and diversification. Syst Biol 60(4):451–465

Hagen O, Hartmann K, Steel M et al (2015) Age-dependent speciation can explain the shape of empirical phylogenies. Syst Biol 64(3):432–440

Hautphenne S, Latouche G, Remiche MA (2009) Transient features for Markovian binary trees. In: Proceedings of the fourth international ICST conference on performance evaluation methodologies and tools

He M, Chan Yb, Hautphenne S (2023) Approximate Bayesian computation for Markovian binary trees in phylogenetics. arXiv preprint arXiv:2309.00194

Ito K (1951) On stochastic differential equations, vol 4. American Mathematical Society, New York

Kapur J (1979) A continuous time age structured population growth model. Appl Math Model 3(6):455–458

Karlin S, Taylor HE (1981) A second course in stochastic processes. Elsevier, London

Kontoleon N (2006) The Markovian binary tree: a model of the macroevolutionary process. Ph.D. thesis, The University of Adelaide. http://hdl.handle.net/2440/22320

Kopperud BT, Magee AF, Höhna S (2023) Rapidly changing speciation and extinction rates can be inferred in spite of nonidentifiability. Proc Natl Acad Sci 120(7):e2208851120

Lambert S, Voznica J, Morlon H (2023) Deep learning from phylogenies for diversification analyses. Syst Biol 72:1262–1279

Landis MJ, Quintero I, Muñoz MM et al (2022) Phylogenetic inference of where species spread or split across barriers. Proc Natl Acad Sci 119(13):e2116948119

Legried B, Terhorst J (2023) Identifiability and inference of phylogenetic birth-death models. J Theor Biol 568:111520

Louca S, Pennell MW (2020a) Extant timetrees are consistent with a myriad of diversification histories. Nature 580(7804):502–505
Louca S, Pennell MW (2020b) A general and efficient algorithm for the likelihood of diversification and discrete-trait evolutionary models. Syst Biol 69(3):545–556
Maddison WP, Midford PE, Otto SP (2007) Estimating a binary character's effect on speciation and extinction. Syst Biol 56(5):701–710
Magnuson-Ford K, Otto SP (2012) Linking the investigations of character evolution and species diversification. Am Nat 180(2):225–245
Morlon H, Potts MD, Plotkin JB (2010) Inferring the dynamics of diversification: a coalescent approach. PLoS Biol 8(9):e1000493
Morlon H, Robin S, Hartig F (2022) Studying speciation and extinction dynamics from phylogenies: addressing identifiability issues. Trends Ecol Evol 37:497–506
Nee S, May RM, Harvey PH (1994) The reconstructed evolutionary process. Philos Trans R Soc Lond B Biol Sci 344(1309):305–311
Quintero I, Landis MJ, Jetz W et al (2023) The build-up of the present-day tropical diversity of tetrapods. Proc Natl Acad Sci 120(20):e2220672120
Soewongsono AC, Holland BR, O'Reilly MM (2022) The shape of phylogenies under phase-type distributed times to speciation and extinction. Bull Math Biol 84(10):118
Soewongsono AC, Diao J, Stark T et al (2023) Matrix-analytic methods for the evolution of species trees, gene trees, and their reconciliation. arXiv preprint arXiv:2309.06447
Stadler T (2013) How can we improve accuracy of macroevolutionary rate estimates? Syst Biol 62(2):321–329
Swiston SK, Landis MJ (2023) Testing relationships between multiple regional features and biogeographic processes of speciation, extinction, and dispersal. bioRxiv, pp 2023–06
Tarasov S, Uyeda J (2024) Nonidentifiability of state-dependent diversification models (sses) is ubiquitous but not problematic for phylogenetics. bioRxiv. https://doi.org/10.1101/2022.07.04.498736
Thompson A, Liebeskind B, Scully EJ et al (2023) Deep learning approaches to viral phylogeography are fast and as robust as likelihood methods to model misspecification. bioRxiv, pp 2023–02
Truman K, Vaughan TG, Gavryushkin A et al (2024) The fossilised birth-death model is identifiable. bioRxiv, pp 2024–02
Van Groenendael J, de Kroon H, Caswell H (1988) Projection matrices in population biology. Trends Ecol Evol 3(10):264–269
Vasconcelos T, O'Meara BC, Beaulieu JM (2022) A flexible method for estimating tip diversification rates across a range of speciation and extinction scenarios. Evolution 76(7):1420–1433
Vaughan TG, Drummond AJ (2013) A stochastic simulator of birth-death master equations with application to phylodynamics. Mol Biol Evol 30(6):1480–1493
Voznica J, Zhukova A, Boskova V et al (2022) Deep learning from phylogenies to uncover the epidemiological dynamics of outbreaks. Nat Commun 13(1):3896
Welch BL (1947) The generalization of 'student's' problem when several different population variances are involved. Biometrika 34(1–2):28–35
Wolfram Research Inc (2023) Mathematica, Version 13.3. https://www.wolfram.com/mathematica

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Springer Nature or its licensor (e.g. a society or other partner) holds exclusive rights to this article under a publishing agreement with the author(s) or other rightsholder(s); author self-archiving of the accepted manuscript version of this article is solely governed by the terms of such publishing agreement and applicable law.