Modeling the velocity of evolving lineages and predicting dispersal patterns

Paul Bastide[a], Pauline Rocu[b], Johannes Wirtz[c], Gabriel W. Hassler[d], François Chevenet[e], Denis Fargette[f], Marc A. Suchard[g,h,i], Simon Dellicour[j,k], Philippe Lemey[k], Stéphane Guindon[b,*]

[a]*IMAG, Université de Montpellier, CNRS, Montpellier, France*

[b]*Laboratoire d'Informatique, de Robotique et de Microélectronique de Montpellier. CNRS - UMR 5506. Montpellier, France*

[c]*CEFE, Université de Montpellier, CNRS, EPHE, IRD, Montpellier, France*

[d]*Department of Economics, Sociology, and Statistics, RAND, Santa Monica, CA, USA*

[e]*MIVEGEC, IRD, CNRS, Université de Montpellier, Montpellier, France*

[f]*PHIM, IRD, INRAE, CIRAD, Université de Montpellier, Montpellier, France*

[g]*Department of Biostatistics, Jonathan and Karin Fielding School of Public Health, University of California, Los Angeles, CA, USA*

[h]*Department of Human Genetics, David Geffen School of Medicine at UCLA, University of California, Los Angeles, CA, USA*

[i]*Department of Computational Medicine, David Geffen School of Medicine at UCLA, University of California, Los Angeles, CA, USA*

[j]*Spatial Epidemiology Lab (SpELL), Université Libre de Bruxelles, Brussels, Belgium*

[k]*Department of Microbiology, Immunology and Transplantation, Rega Institute, Laboratory for Clinical and Epidemiological Virology, KU Leuven, Leuven, Belgium*

Abstract

Accurate estimation of the dispersal velocity or speed of evolving organisms is no mean feat. In fact, existing probabilistic models in phylogeography or spatial population genetics generally do not provide an adequate framework to define velocity in a relevant manner. For instance, the very concept of instan-

*Corresponding author

taneous speed simply does not exist under one of the most popular approaches that models the evolution of spatial coordinates as Brownian trajectories running along a phylogeny [30]. Here, we introduce a new family of models – the so-called "Phylogenetic Integrated Velocity" (PIV) models – that use Gaussian processes to explicitly model the velocity of evolving lineages instead of focusing on the fluctuation of spatial coordinates over time. We describe the properties of these models and show an increased accuracy of velocity estimates compared to previous approaches. Analyses of West Nile virus data in the U.S.A. indicate that PIV models provide sensible predictions of the dispersal of evolving pathogens at a one-year time horizon. These results demonstrate the feasibility and relevance of predictive phylogeography in monitoring epidemics in time and space.

Keywords: phylogeography, integrated velocity models, Bayesian inference, West Nile virus, PhyREX, BEAST

---

1. Introduction

Evaluating the pace at which organisms move in space during the course of evolution is an important endeavor in biology. When considering deep evolutionary time scales, understanding past dispersal events is key to explaining the spatial diversity of contemporaneous species. Over shorter time frames, making sense of the migration patterns of closely related organisms is crucial in building a detailed picture of a population's demographic past, present and future dynamics. Tracking the spatial dynamics of pathogens during a pandemic, in particular, is of utmost interest as it conveys useful information about the means and the rapidity at which a disease is spreading in a population. Epidemiological data generally consists in records of incidence of the disease at various points in time and space. Yet, estimating the speed at which an organism spreads at the onset of an epidemic from count data is challenging [40, 38]. Similarly, characterizing the migration process at equilibrium, i.e., in situations where the organism under scrutiny is well-established in a region, is simply not feasible from occurrence data only. These difficulties mainly stem from the fact that count or occurrence data do not convey information about the non-independence between observations due their shared evolutionary paths.

Genomes carry useful information about the relationships between pathogens. Observed differences between homologous genetic sequences is at the core of phylogenetic and population genetics approaches which provide a sound framework to account for the non-independence between data points in downstream analyses. This framework also accommodates for the situation where nucleotide (or protein) sequences are sampled at various points in time [13]. Heterochronous samples combined with the molecular clock hypothesis [44] may then serve as

a basis to infer the rate at which substitutions accumulate and reconstruct the time scale of past demographic trajectories of the population under scrutiny (see e.g., [22] for a review).

Designing models for the joint analysis of genetic sequences and their locations of collection was initiated in the middle of the last century by Fisher and Malécot who brought forward the isolation by distance model [43, 31]. The rise of statistical phylogeography over the last decade proposed alternatives that are less mechanistic but still aim at capturing the main features of the spatial diffusion process. These approaches are also well suited to deal with heterochronous data and handle cases where the population of interest is scattered along a spatial continuum rather than structured into discrete demes. Lemey et al. [30], in particular, described a hierarchical model whereby spatial coordinates evolve along a phylogenetic tree according to a Brownian diffusion process with branch specific diffusion rates. The so-called Relaxed Random Walk (RRW) model has since then been used to characterize the spatial dynamics of several pathogens of high public health and societal impacts, including Ebola [9] and the Rice yellow mottle [26] viruses for instance.

One of the key objectives of the RRW model is to infer the pace at which organisms disperse. Pybus et al. [32] suggested using a diffusion coefficient which derives from the ratio of the estimated squared displacement between the start and the end of a branch and the corresponding elapsed time. The branch-level ratios are then averaged over the edges in the phylogeny. Pybus et al. [32] and then Dellicour et al. [8] later introduced wavefront through time plots, deriving from the displacement between the estimated root location and the most distant tip locations at various points in time. Trovão et al. [39] considered instead dispersal rates which are defined as ratios of estimated displacements (using great-circle distances) by the elapsed time.

These statistics generally provide a rough characterization of the dispersal process. The limitations of the dispersal statistics mainly stem from the very nature of the RRW model: because Brownian trajectories are nowhere differentiable, the concept of instantaneous velocity is simply not defined under that family of models. Also, the sum of displacements deriving from the observation of a Brownian particle at various points in time grows with the square root of the number of (equally spaced in time) observations, making the estimation of an average speed sampling inconsistent. Finally, the analysis of spatial data simulated under the Brownian motion model along birth-death trees show that the standard dispersal statistics often fail to provide accurate estimates of the speed at which lineages move in space during the course of evolution [12].

The present study tackles the issue of dispersal velocity and speed estimation by introducing a new approach that models the instantaneous velocity of lineages explicitly. Under these models, the spatial coordinates of lineages derive from integrating their velocities so that we refer to "Phylogenetic Integrated

Velocity" (PIV) models throughout this article. To our knowledge, this study is the first to use the integrated velocity models in a phylogenetic context. Integrated processes are common in the ecology literature however. They model animal movement with smooth stochastic processes [25]. Integrated processes have also been used in virology, e.g. to model longitudinal CD4 T-cells and beta-2-microglobulin measurements [36]. Some of the analytical formulas used in the present study are already known and developed, e.g. in [7], to model population growth.

Although velocity is not directly observable from heterochronous and geo-referenced genetic sequences, our results indicate that this quantity can be estimated reliably. Using simulations under a realistic spatial population genetics model, we show that the velocity estimates obtained from the new models are more accurate than those deriving from the RRW approach. Velocity estimates obtained from the analysis of multiple West Nile virus data sets were also used to predict the spatial distribution of the pathogen over a one-year time horizon in the U.S.A. Comparison of these predictions to incidence data at the county-level suggests that important features of the spatial distribution of that pathogen are indeed amenable to reasonably accurate predictions.

Our ability to efficiently monitor and anticipate the spread of emerging epidemics depends on the accuracy with which the pace of dispersal can be quantified. The family of new models introduced in this study provides a relevant tool to achieve this objective. While important aspects of viral evolution may escape prediction indefinitely [23] and predicting when or where the next virus outbreak will appear remains out of reach [41, 24], the present study shows how predictive phylogeography may complement classical approaches in epidemiology.

2. Results

2.1. PIV models: main rationale

The main attributes of models that belong to the PIV family are presented first. We focus on the process of interest along a given time interval $[0, t]$, corresponding to the length (in calendar time units) of a given branch in the phylogeny of a sample of the organism of interest. Let $X(s)$ be the random variable representing the location (i.e., the coordinates) of a lineage at time $0 \leq s \leq t$. $Y(s)$ is its velocity, i.e., the instantaneous speed and direction at time s. Both X and Y are typically vectors of length two, corresponding to latitude and longitude. The location $X(t)$ at the end of the branch may then be expressed as follows:

$$X(t) = x(0) + \int_0^t Y(s)\mathrm{d}s,$$

where $x(0)$, the location at the time of origin, is fixed. The Brownian Motion (BM) and the RRW models focus on $\{X(s), 0 \leq s \leq t\}$, i.e., the process describing the evolution of the location during a time interval. While, in one dimension, BM models have a single dispersal parameter that applies to all edges in the phylogeny, RRW have branch-specific dispersal parameters in a manner similar to the relaxed clock model [14].

Instead of modeling the fluctuation of coordinates, PIV models deal with $\{Y(s), 0 \leq s \leq t\}$, i.e., the process describing the variation of velocity in that interval. The dynamics of spatial coordinates then derive from the integration over the velocity as stated in the equation above, hence the name "phylogenetic integrated velocity". In the following, we introduce two stochastic processes for $\{Y(s), 0 \leq s \leq t\}$ and characterize the corresponding distributions of $X(t)$. In order to ease the presentation, we provide formulas for univariate processes only in the main text. Formulas for bi-variate (and, more generally, multivariate) processes are given in Supplementary Information (sections 5.3 and 5.4).

2.2. Behavior of PIV models

Velocities. The Integrated Brownian Motion (IBM) model relies on a Wiener process with shift and scale parameters $y(0)$ and σ respectively to model $\{Y(t); t > 0\}$. That process is Gaussian and we have:

$$\begin{aligned} \mathrm{E}(Y(t) \mid y(0)) &= y(0), \\ \mathrm{Cov}(Y(u), Y(v)) &= \sigma^2 u \quad \text{with } 0 < u \leq v. \end{aligned}$$

The Integrated Ornstein-Uhlenbeck (IOU) model uses instead a Ornstein-Uhlenbeck (OU) process to describe the evolution of velocity. The mean and variance of velocity at time t are given below:

$$\begin{aligned} \mathrm{E}(Y(t) \mid y(0)) &= y(0)e^{-\theta t} + \mu(1 - e^{-\theta t}), \\ \mathrm{Cov}(Y(u), Y(v)) &= \frac{\sigma^2 e^{-\theta v}}{\theta} \sinh(\theta u) \quad \text{with } 0 < u \leq v. \end{aligned}$$

The parameter θ in the OU model governs the strength with which $Y(t)$ is pulled towards the trend μ.

Spatial coordinates. We now examine the evolution of spatial coordinates under the PIV models. The stochastic processes modeling the fluctuation of velocity being Gaussian, the coordinates also follow a Gaussian process [7]. We give below the mean and variance of the distribution of $X(t)$ given $x(0)$ and $y(0)$, the coordinates and velocity at time 0. Characterizing the process governing the evolution of spatial coordinates will shed light on the biological relevance of the proposed approach and exhibit the main difference in behavior in comparison with the BM and, by extension, the RRW models.

When velocity follows a Brownian process (IBM process), we have:

$$\begin{aligned} \mathrm{E}(X(t) \mid x(0), y(0)) &= x(0) + y(0)t, \\ \mathrm{V}(X(t) \mid x(0), y(0)) &= \frac{\sigma^2 t^3}{3}. \end{aligned}$$

A linear increase of the spatial coordinates is thus expected with a direction that is determined by the initial velocity. Because of the inertia deriving from their velocity, spatial coordinates of lineages evolving under IBM thus tend to resist changes in their direction of motion. This mean drift is similar to the directional random walk, used e.g. in [17] to model the spatial spread of HIV-1. The BM has a distinct behavior and authorizes sudden changes of direction and the RRW can even lead to large discontinuous "jumps" from one place to another [3]. In contrast, the IBM is smooth by design, yet, as suggested by the last expression above, the variance of coordinates grows cubically in time thereby allowing the PIV model to accommodate for dispersal events over long distances in short time periods.

The corresponding expectation and variance for the IOU model are given in Supplementary Information (section 5.1). Here again, the average coordinates at the end of the branch of focus are determined by the coordinates at the start of that branch ($x(0)$) plus the expected displacement ($y(0)t$) along that same edge. While for small values of θ the IOU has a behavior that is similar to the IBM, for larger values of that parameter, its variance grows linearly in time and the process behaves like a directional BM [17].

Figure 1 illustrates the behavior of the classical random walk and integrated models along a 5-tip tree. Trajectories of coordinates generated with the BM and Ornstein-Uhlenbeck (OU) versions of the random walk model are intricate, showing sudden changes of directions in the movements (Fig. 1d, e). The same behavior is displayed by the velocity trajectories under the IBM and IOU models (Fig. 1b, c) as the models are here identical to that used for the BM and OU models indeed. Yet, integrating over these rugged paths gives smooth trajectories of coordinates under the corresponding models (Fig. 1f, g), with particles moving swiftly away from their initial points, illustrating the cubic variance pointed above. The IOU model presented here converges to a (1,1) velocity so that the coordinates of the five lineages show a clear directionality, stronger than that obtained with the OU model (Fig. 1g vs. e).

2.3. Accuracy of speed estimation

100 data sets simulated under the spatial Lambda-Fleming-Viot model [15, 1] were analyzed with the IBM and IOU models along with the standard RRW model. Velocities inferred at each node in the phylogeny under the PIV models were averaged across nodes in order to derive a tree-level speed statistic. The

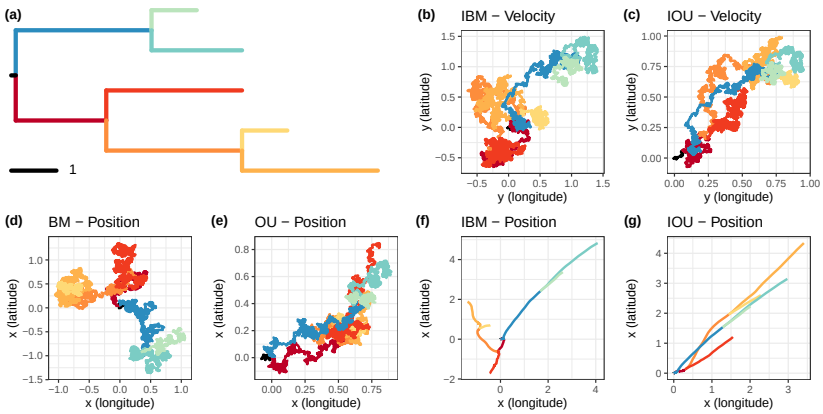

Figure 1: **Simulated trajectories of classical random walk and PIV models on a simple tree.** Each process was simulated on the tree **(a)**, with branches of matching colors. The latitude and longitudes were simulated independently. **(d)** random walk using Brownian motion (BM) with variance $\sigma^2 = 0.1$, and starting point $x(0) = (0, 0)$. **(e)** Ornstein-Uhlenbeck process (OU) with stationary variance of $\sigma^2/(2\theta) = 0.1$, selection strength $\theta = 0.17$, starting at $x(0) = (0, 0)$, and converging to its optimal value $\mu = (1, 1)$. **(b)** and **(f)** velocity y and position x of an Integrated Brownian Motion (IBM) with variance $\sigma^2 = 0.1$, starting point $x(0) = (0, 0)$, and starting velocity $y(0) = (0, 0)$. **(c)** and **(g)** velocity y and position x of an Integrated Ornstein-Uhlenbeck (IOU) with stationary variance $\sigma^2/(2\theta) = 0.1$, selection strength $\theta = 0.17$, speed optimal value of $\mu = (1, 1)$, starting point $x(0) = (0, 0)$, and starting velocity $y(0) = (0, 0)$.

classical weighted lineage dispersal velocity [11] was used instead for all analyses performed under the RRW model.

Figure 2c displays the mean squared error of speed estimates deriving from the different models. For RRW, the error in the speed estimates grows in an exponential-like manner with true speed. Examination of the estimated vs. true speed relationship (Fig. 2d) indicates that the RRW model systematically underestimates speed and the bias worsens with increasing speed. The spatial Lambda-Fleming-Viot model assumes a finite-size habitat (a square here) and boundary effects, which occur for large and small dispersal parameters (see Figure 2a and b), are expected to impact the estimation of speed under models that ignore this constraint. Yet, PIV models seem to be largely immune to this issue. While the IOU model underestimates speed for large dispersal data sets and overestimates it for small dispersal, its estimates are clearly superior to those deriving from the dispersal velocity estimated under RRW. Similar performance is obtained with the IBM model, even though this approach overestimates speed for large dispersal.

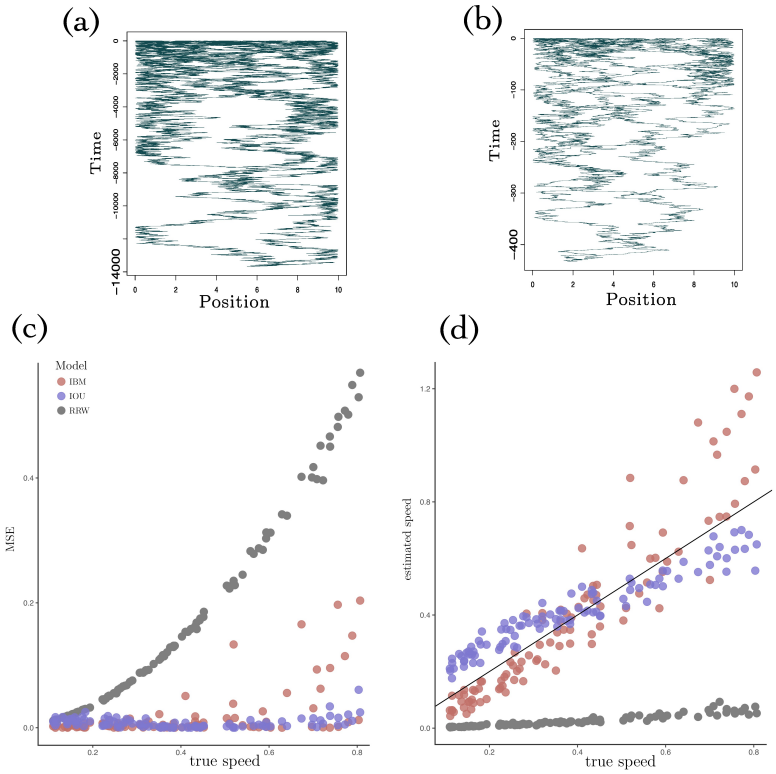

Figure 2: **Accuracy of speed estimation under the RRW and PIV models.** 100 data sets were simulated using the spatial Lambda-Fleming-Viot process on a 10 by 10 square. **(a)** shows one simulation using a small dispersal parameter. Time 0 corresponds to the start of the simulation and all lineages coalesced around 14000 time units in the past. The position of lineages along one dimension is reported along the x axis. **(b)** realization of another simulation with a larger dispersal parameter. All lineages coalesced here about 400 time units in the past. **(c)** mean squared error of speed estimates vs. true speed values. **(d)** estimated vs. true speed values. Speed is expressed here in terms of Euclidean distance per time unit.

2.4. Dispersal dynamics of the West Nile virus in the U.S.A.

The phylogeography of WNV in the U.S.A. has been studied extensively (see, e.g., [11]). The origin of this epidemic took place in New York City during the summer 1999 [29, 5]. By 2004, human infections, veterinary disease cases or infections in mosquitoes, birds, or sentinel animals had been reported to the Centers for Disease Control and Prevention (CDC) in most counties.

We fitted the PIV and RRW models to several subsets of the 801 geo-referenced sequence data set analyzed in [11]. PIV models are less flexible than the RRW approach as they do not authorize sudden changes of direction (see Supplementary Information, section 5.2). Therefore, ensuring that both approaches nonetheless provide comparable fit to the data is a prerequisite to further analyses. We then used the IBM model to predict the dispersal patterns and evaluate these predictions through the comparison with incidence data for the 2000-2007 time period.

2.4.1. Model comparison

We compared the fit of the RRW and PIV models to the WNV data using cross-validation of location information. Cross-validation is a powerful model comparison technique in the context of phylogenetic factor analysis [21]. Using a subset of 150 data points chosen uniformly at random among the 801 available observations, a leave-one-out procedure was applied to the sample coordinates. Each tip location was first hidden and its posterior density was estimated using MCMC from the remaining 149 locations and all 150 sequences (see Supplementary Information 5.7).

Figure 3 shows the distribution of the great circle distances between the observed and reconstructed tip locations as inferred under the PIV and RRW models. These distributions indicate that the three models display similar behavior overall with a majority of distances between true and reconstructed tip locations ranging between 222 km (25% quantile of distribution from MCMC output pooled across models) and 831 km (75% quantile) and a median of 427 km. In contrast, if inferred locations are uniform at random within the range of sampled locations, the median distance is 1,332 km, i.e., about three times that estimated using our models. This result demonstrates the ability of the phylogeographic models considered here to extract meaningful signal from the data. Examination of the posterior distribution deriving from each model taken separately indicates that the median distances obtained under the IBM, IOU and RRW models are 437, 496 and 383 km respectively. While the fit of the RRW model as measured here is superior to that of the PIV models, the performance of the three models are nonetheless qualitatively similar.

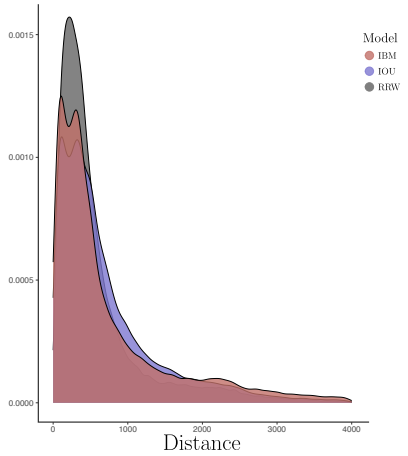

Figure 3: **Distribution of great circle distances between true and estimated tip coordinates under the PIV and RRW models.** A subset of 150 data points taken uniformly at random among 801 observations was analyzed. Each tip coordinates was considered as parameter and the posterior distribution of the distance to the observed coordinates reconstructed.

2.4.2. Predicting dispersal using PIV models

PIV models enable the estimation of dispersal velocity of sampled lineages. These velocities may then serve as a basis to predict the spatial distribution of the underlying population in the near future. Here, we tested the ability of the IBM model to anticipate the dynamics of dispersal of the WNV in the early and later stages of the epidemic.

Sequences collected earlier than December of year Y were randomly subsampled from the complete data set with exponentially increasing weights given to recent samples. Data sets with 150 sequences were obtained except for years 2000-2002 where smaller sample sizes were considered due to a lack of data in this time period. Estimated posterior distributions of velocities at the tips of the obtained phylogeny under the IBM model were then used as predictors of the spatial distribution of the virus in year $Y + 1$ (see Material and Methods, section 4.3). The predicted occurrences were compared to yearly incidence data collected at the county level.

Figure 4 shows the incidence and the predicted occurrence of the WNV in the early stages in the epidemic. Samples for years 2000, 2001 and 2002 included only 7, 19 and 68 geo-referenced sequences, thereby making any prediction inherently challenging. For instance, while the virus had reached Florida by 2001, our model failed to predict its presence south of North Carolina. Predictions for subsequent years rely on larger numbers of observations and demonstrate the relevance of our approach. Indeed, the PIV model successfully predicted the

arrival of the pathogen along the west coast of the U.S.A. by the end of 2002. It also correctly predicted that the north west corner of the country would remain largely virus-free until the end of 2003.

By 2004 the virus reached an endemic state and the spatial dynamics of the epidemic diverged from that of the early stages. Figure 5 shows the results for the 2004-2007 time period. Prediction at local spatial scales has limited accuracy. For instance, a high probability of occurrence was systematically estimated for the states in the North East corner of the country and the south of Texas while incidence was generally mild in these areas. Note that the difference between the predicted and observed occurrence could reflect a relatively lower ecological suitability of these regions to host local WNV circulation, thereby serving a useful purpose. Moreover, the IBM model correctly predicts the expansion of the epidemic north of California and Nevada between 2005 and 2006. Also, according to our predictions, the pathogen covered limited distances in the 2004-2007 period compared to the early stages of the epidemic. This quasi stasis is confirmed by the largely similar distributions of yearly incidences. Hence, here again, our approach manages to capture changes in the spatial dynamics of the pandemic that are central in the context of pathogen surveillance.

3. Discussion

The present study addresses shortcomings in the estimation of the velocity of lineages using popular models in phylogeography. These approaches rest on the probabilistic modeling of the coordinates of lineages along their phylogeny. Yet, the central concept of instantaneous speed simply does not exist under the most popular Relaxed Random Walk (RRW) model. As a consequence, measuring speed as a ratio between a displacement and the corresponding elapsed time leads to difficulties in the estimation of velocity. In order to circumvent these limitations, we introduce Phylogenetic Integrated Velocity (PIV) models. The originality of this family of models lies in their modeling of the velocity of evolving lineages instead of their coordinates. This approach enables a proper definition of instantaneous speed, which can be inferred anywhere along the tree, including at its tips.

Data sets were simulated under a mechanistic model of evolution that, unlike the RRW and PIV models, derive from first principles in biology. Results show that speed estimates obtained with PIV models are generally more accurate than those deriving from the RRW approach, especially in cases where the pace of dispersal is high. Yet, PIV models are less flexible than RRW in their description of the movement of lineages during the course of evolution. In particular, sudden changes in the direction of dispersal are not well accounted for by PIV models. These changes would indeed require "breaks" in the trajectory of velocities, which the underlying Gaussian processes do not allow. However, our analysis

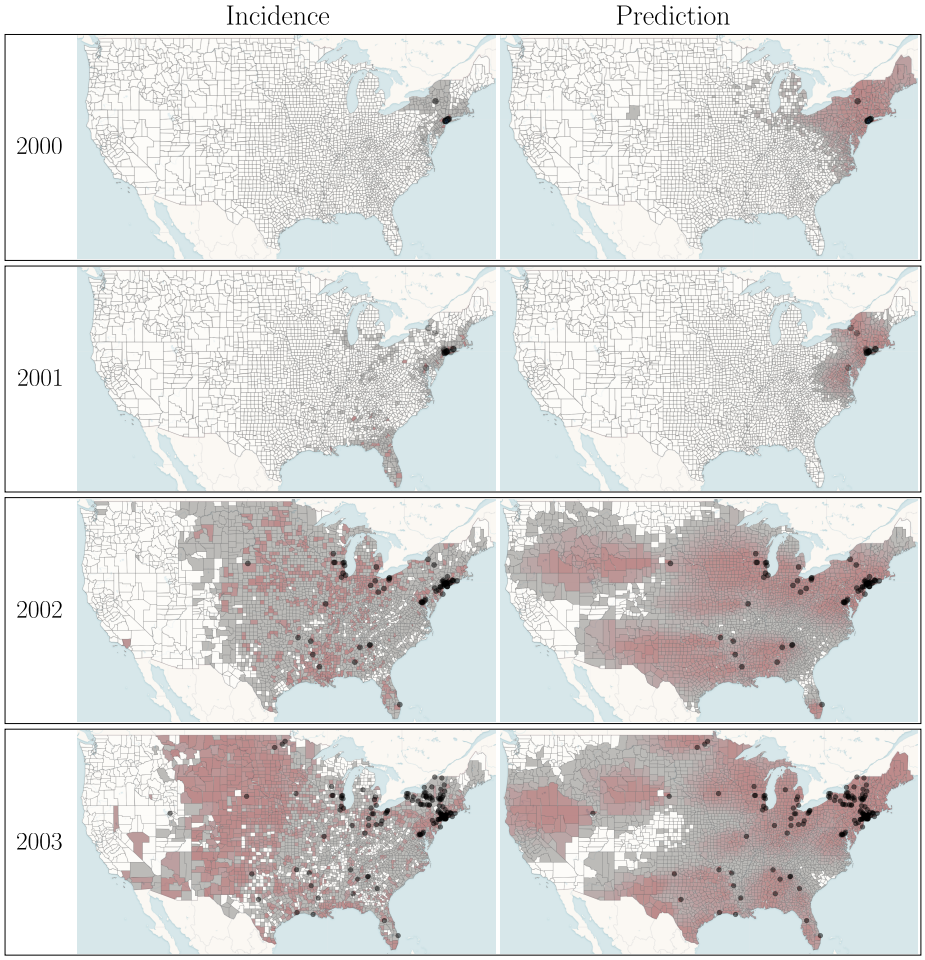

Figure 4: **Incidence and predicted occurrence of WNV in the early phase of the epidemic.** Black dots correspond to sampled locations. Incidence data (left) for each year and each county was obtained from the CDC. For year Y, predicted occurrence of the WNV (right) was inferred using data collected earlier than the end of December of year $Y-1$. The maps were generated with EvoLaps2 [6]

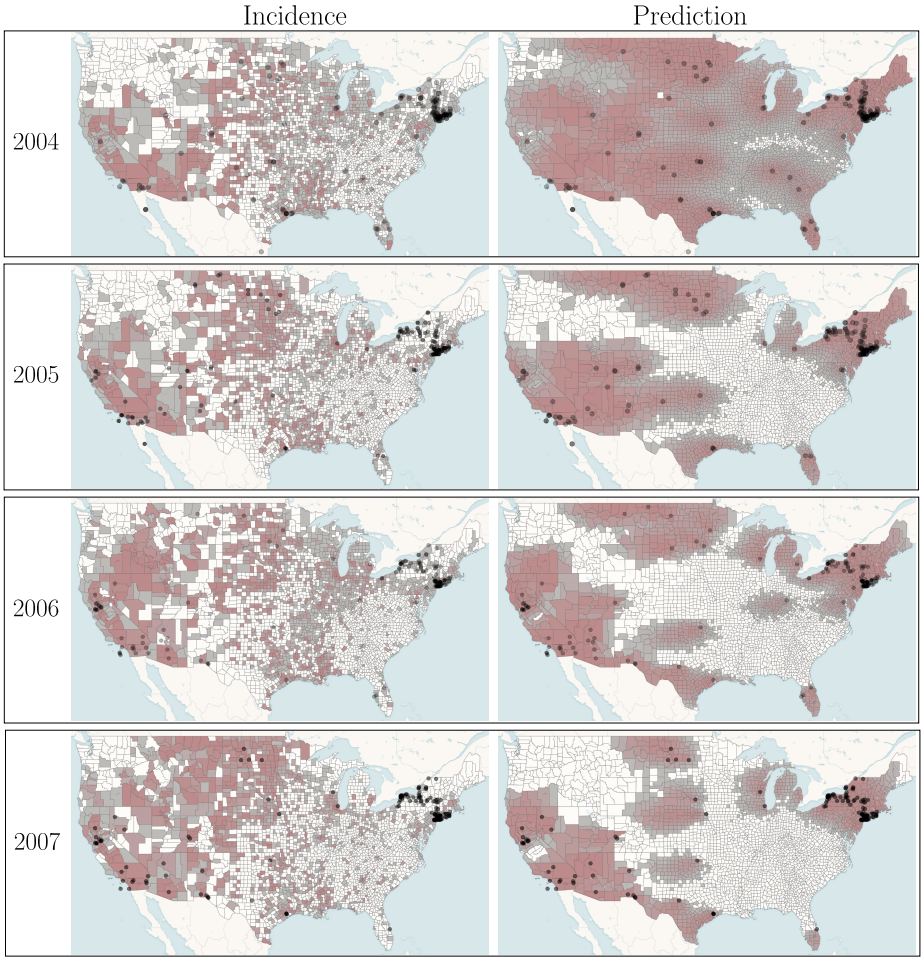

Figure 5: **Incidence and predicted occurrence of WNV in an endemic regime.** See caption of Figure 4.

of West Nile virus data in the U.S.A. indicates that the movements of lineages display here enough inertia so that abrupt changes in the spatial trajectories are seldom observed. Cross-validation suggests in fact that the two PIV models tested here provide a fit to the data similar to that obtained with the RRW. Moreover, the analytical expressions of the variance of coordinates under the IBM model grows with time in a superlinear manner, thereby allowing large displacements in short amounts of time.

Velocities at the tips of phylogenetic trees can be inferred using the new PIV models. These approaches are thus well suited to model future dispersal events. Here, we evaluate the accuracy of predicted movements through the analyses of subsets of a large data set of West Nile virus geo-referenced sequences and county-level yearly incidence data. Our predictions focus on deciding whether the pathogen will occupy (or be absent from) a given county at a given time interval in the future, i.e., a modest endeavor compared to predicting future incidence. The proposed approach accurately predicted the arrival of the virus along the west coast of the USA in 2002 from the analysis of data collected before the end of December 2001. Furthermore, the predictions clearly point to a change of dispersal dynamics around 2004-2005 with a transition from an expansion phase to an endemic regime whereby rapid east-to-west dispersal events are replaced with short-distance migrations. While the proposed predictions have limited accuracy in the early stages of the pandemic where data is scarce and sampling likely to be biased, the PIV models successfully anticipate dispersal events in many instances. Altogether, our results indicate that the predictive phylogeography approach put forward in the present study could indeed serve a useful purpose in real time forecasting of the spread of an epidemic. Future work could aim at incorporating data on the ecological suitability of the investigated areas in order to improve predictions, in a manner similar to that used in "landscape phylogeography" [10].

In addition to prediction, the PIV models are also expected to prove useful in many cases where the RRW model has been applied to quantify and compare dispersal velocity. These applications range from animal and human viruses to plant viruses. For instance, lower rates of dengue virus dispersal in urban as opposed to rural settings has implicated a major role for mosquito-mediated dispersal [33]. Dispersal velocity has often been estimated for rabies lineages with dogs as the main host species, resulting in hypotheses of their spread being impacted by human activities [37]. More recently, a slow dispersal has been estimated for Lassa Virus in its rodent reservoir, which could in part explain the restricted distributions of the virus [28]. Also, increasing dispersal rates of Rice Yellow Mottle virus in Africa has led to the suggestion that intensification of rice cultivation could have been favorable to virus spread [34]. Applications of the PIV models could increase the credibility of these and many more hypotheses of viral spread.

The proposed new models and predictions have limitations however. Sam-

pling is likely to impact the results in case it is driven by practical aspects (e.g., the distribution of genomic surveillance facilities is not uniform throughout the habitat) and does not reflect the underlying spatial distribution of the population under scrutiny. Recent work [18] shows how different sampling strategies can be incorporated in the RRW model. A similar framework could apply to PIV models and mitigate the impact of sampling. Additionally, when available, incidence data conveys information about the demographic dynamics of an epidemic. Hence, increased accuracy of the predictions may be achievable through the incorporation of past incidence data in the new models presented in this work.

4. Material and Methods

4.1. Likelihood calculation and Bayesian inference

Let $\mathbf{X}^*$ and $\mathbf{X}$ correspond to random variables denoting the vectors of positions at the tips and the internal nodes respectively. $\mathbf{x}^* = \{x_1, \ldots, x_n\}$ and $\mathbf{x} = \{x_{n+1}, \ldots, x_{2n-1}\}$ are realizations of the corresponding random variables, where n is the number of tips and $2n-1$ is the index of the root node. $\mathbf{Y}^*$ and $\mathbf{Y}$ are the vectors of velocities at tip and ancestral nodes respectively.

Here, we describe two different approaches for the Bayesian inference of PIV model parameters. The first method, implemented in PhyREX [18], relies on data augmentation. It starts with the computation of $p(\mathbf{x}^*, \mathbf{x}, \mathbf{y}^*, \mathbf{y})$, i.e., the joint density of all (i.e., ancestral and tip) locations and velocities. This density is also conditioned on the phylogeny, i.e., a rooted tree topology with node heights, which is not included in the formula below for the sake of conciseness. Given the locations and velocities at all nodes in the tree, the evolutionary processes taking place along every branch is independent from that happening along the other edges. The likelihood is then evaluated as follows:

$$\begin{aligned} p(\mathbf{x}^*, \mathbf{x}, \mathbf{y}^*, \mathbf{y}) &= \pi(y_\rho, x_\rho) \prod_{i=1}^{2n-2} p(x_i \mid x_{\mathrm{pa}(i)}, y_i, y_{\mathrm{pa}(i)}) p(y_i \mid y_{\mathrm{pa}(i)}, x_{\mathrm{pa}(i)}) \\ &= \pi(y_\rho, x_\rho) \prod_{i=1}^{2n-2} p(x_i \mid x_{\mathrm{pa}(i)}, y_i, y_{\mathrm{pa}(i)}) \prod_{i=1}^{2n-2} p(y_i \mid y_{\mathrm{pa}(i)}) \quad (1) \end{aligned}$$

where the subscript $\mathrm{pa}(i)$ corresponds to the direct parent of node i. Also, $\pi(y_\rho, x_\rho)$ is the velocity and location density at the root node. In the present work, we use a normal density for the corresponding distribution. Since $(X_i \mid x_{\mathrm{pa}(i)}, y_i, y_{\mathrm{pa}(i)})$ is normally distributed, we can use the pruning algorithm as described in [32] to integrate over $\mathbf{X}$, giving the following likelihood:

$$p(\mathbf{x}^*, \mathbf{y}^*, \mathbf{y}) = \pi(y_\rho, x_\rho) \phi(\mathbf{x}^*; \mathbf{y}^*, \mathbf{y}) \prod_{i=1}^{2n-2} p(y_i \mid y_{\mathrm{pa}(i)}) \quad (2)$$

where $\phi(\mathbf{x}^*; \mathbf{y}^*, \mathbf{y})$ is obtained through a post-order tree traversal, assuming the movements along both spatial axes are independent from one another and using the means and variances for either the IBM or the IOU model (see Supplementary Information, section 5.2). The calculation just described relies on augmented data since velocities at all nodes in the tree are considered as known. Uncertainty around these latent variables is non-negligible. Samples from the joint posterior distribution of all model parameters, including ancestral and contemporaneous velocities, were obtained through Markov Chain Monte Carlo integration.

The second method for evaluating the likelihood of PIV models is implemented in BEAST [35]. It relies on the direct computation of $p(\mathbf{x}^*)$, the likelihood of the observed positions at the tips conditionally on the tree. It uses the fact that the stochastic process $\mathbf{Z}(t) = (\mathbf{Y}(t), \mathbf{X}(t))$ that describes the joint evolution of both the velocity and position is a multivariate Markov process, that can be framed as linear Gaussian as in [4] (see Supplementary Information, section 5.3), with all the velocities at the tips considered as missing (as these are not observed). In [4], a general pruning algorithm is described to deal with this kind of process (with missing values), that provides not only the likelihood (one post-order traversal) but also the conditional distribution of non-observed traits conditioned on observed traits at the tips (on additional pre-order traversal). This algorithm hence readily gives the posterior distribution of velocities without the need to sample from them using MCMC. Moreover, it does not need to assume that the spatial axes are independent from one another.

In both approaches, standard operators were used to update the topology of the phylogenetic tree, the node ages along with the parameters of a HKY [20] nucleotide substitution model. The diffusion parameters of the Brownian process were also updated using standard Metropolis-Hastings steps. Most results in this study were derived with PhyREX even though BEAST outperformed PhyREX in terms of speed of parameter inference (see Supplementary Information, section 5.5). The two independent implementations of Bayesian samplers under the same models provide a robust validation of most results presented in this study.

4.2. Simulations

We conducted simulations that aimed at assessing the ability of the different models to estimate the average speed at which lineages move in space during the course of their evolution. Genealogies and the accompanying spatial coordinates were generated according to the "individual-based" spatial Lambda-Fleming-Viot model using a normal dispersal kernel [15, 1]. We selected this model as it describes the evolution of a population of related individuals along a spatial continuum as opposed to discrete demes. The spatial Lambda-Fleming-Viot model is not subject to the shortcomings that hinder other popular spatial

population genetics models such as sampling inconsistency [2] or Felsenstein's infamous "pain in the torus" [16]. As a coalescent model, it allows simulating the evolution of a sample of individuals backward in time in an efficient manner. Finally and most importantly, because lineages' coordinates evolve here according to a jump process, the exact spatial coordinates of each lineage at each point in time can be monitored. This information may then serve as a basis to evaluate the total distance covered by all lineages in the genealogy. The ratio of this distance by the corresponding elapsed time gives an (average) speed that genuinely reflects the dispersal ability of the organisms under scrutiny.

50 individuals were sampled on a 10 by 10 square defining the habitat of the corresponding population. The rate of events where lineages die and/or give birth to descendants (the so-called REX events in [19]) was set to 10^3 events per unit of time per unit area and the variance of the normal density that defines the radius parameter in the spatial Lambda-Fleming-Viot model was chosen uniformly at random in $[0.1, 0.3]$. These parameter values are such that lineage jumps are short and frequent, thereby mimicking the behavior of a Brownian process [42]. Edges in the obtained genealogy were rescaled so that the average length of an edge after scaling was 0.05 nucleotide substitutions per site. Nucleotide sequences were then generated under a strict clock model according to the HKY model of evolution [20] with transition/transversion ratio set to 4.0. 100 genealogies along with the corresponding spatial coordinates and homologous nucleotide sequences were generated this way.

Each simulated data set was processed using the RRW, IBM and IOU models with independent coordinates. When considering their spatial components only, these models have have 3, 2 and 6 parameters respectively. The RRW model used a log-normal distribution of branch-specific dispersal rates, which is the standard parametrization for that model. The nucleotide substitution rate was set to its simulated value by taking the ratio of the tree length as expressed in molecular and calendar units. The tree-generating process was assumed to be Kingman's coalescent [27] with constant effective population size and a flat (improper) prior distribution on that parameter. Although sequences evolved according to a strict clock model, we used an uncorrelated relaxed clock model [14] with a log-normal distribution of edge-specific substitution rate multipliers. An exponential prior with rate set to 100 was used for the variance of this log-normal density.

For each data set, the true average speed was taken as the actual euclidean distance covered by every lineage divided by the tree length in calendar time unit (see above). For RRW, euclidean distances between the coordinates at each end of every branch in the tree were inferred and used to derive the dispersal rate. The posterior median of that statistic was used as our speed estimate. For PIV models, speed at the tree level was obtained by averaging the speed estimated at each node, the latter deriving from the corresponding velocities. Here again, we obtained the posterior distribution of the tree-level speed and use

the median as our estimate. Note that none of the processes used for inference is the "true" process used for simulation, but simplified versions of it.

4.3. Predictive phylogeography

The PIV models provide an adequate framework to estimate velocities at the tips of the inferred phylogenies. It thus makes sense to apply them to predicting dispersal patterns. Here, we designed a prediction technique which goal is to assess whether the organism under scrutiny may be found in a given region at a given point in time after the most recent sample was collected. Our approach utilizes the posterior distribution of the velocities estimates at each tip of the phylogeny to build a predictor. The latter is obtained by linear extrapolation of the estimated velocity at each tip in the tree that assumes a constant speed of lineages after their sampling. Survival of these linear trajectories is taken into account so that older samples are less likely than recent ones to survive to a given time point in the future. This approach therefore puts more weight on recent samples to predict dispersal patterns (see Supplementary Information, section 5.6). Incidence data used for comparison was extracted from `https://www.cdc.gov/west-nile-virus/data-maps/historic-data.html`.

Acknowledgments

The research leading to these results has received funding from the European Research Council under the European Union's Horizon 2020 research and innovation programme (grant agreement no. 725422 - ReservoirDOCS) and the US National Institutes of Health (R01 AI153044, F31 AI154824). PR's internship at the University of Montpellier was founded by the I-SITE MUSE through the Key Initiative "Data and Life Sciences". PB thanks Pierre Gloaguen for useful discussions on the integrated models. SD acknowledges support from the Fonds National de la Recherche Scientifique (F.R.S.-FNRS, Belgium; grant n°F.4515.22) and the Research Foundation — Flanders (Fonds voor Wetenschappelijk Onderzoek — Vlaanderen, FWO, Belgium; grant n°G098321N). PL acknowledges support by the Research Foundation – Flanders ('Fonds voor Wetenschappelijk Onderzoek – Vlaanderen', G051322N and G005323N).

Data and code availability

The data and code to reproduce all analyses and figures displayed in this study are available at `https://github.com/pbastide/integrated_phylogenetic_`

`models`. The PhyREX and BEAST programs are open source and freely available from `https://github.com/stephaneguindon/phyml` and `https://github.com/beast-dev/beast-mcmc`.

References

[1] N. Barton, A. Etheridge, and A. Véber. A new model for evolution in a spatial continuum. *Electronic Journal of Probability*, 15, 2010.

[2] N. Barton, A. Etheridge, and A. Véber. Modelling evolution in a spatial continuum. *Journal of Statistical Mechanics: Theory and Experiment*, 2013 (01):P01002, 2013.

[3] P. Bastide and G. Didier. The Cauchy process on phylogenies: A tractable model for pulsed evolution. *Systematic Biology*, 72:1296–1315, 2023.

[4] P. Bastide, Lam Si T. Ho, G. Baele, P. Lemey, and M. Suchard. Efficient Bayesian Inference of General Gaussian Models on Large Phylogenetic Trees. *The Annals of Applied Statistics*, 15:971–997, 2021. doi: 10.1214/20-aoas1419.

[5] G. Campbell, A. Marfin, R. Lanciotti, and D. Gubler. West nile virus. *The Lancet Infectious Diseases*, 2:519–529, 2002.

[6] F. Chevenet, D. Fargette, P. Bastide, T. Vitré, and S. Guindon. EvoLaps 2: Advanced phylogeographic visualization. *Virus Evolution*, 10(1):vead078, 2024.

[7] W. Cumberland and C. Rohde. A multivariate model for growth of populations. *Theoretical Population Biology*, 11:127–139, 1977.

[8] S. Dellicour, R. Rose, N. Faria, L. Vieira, H. Bourhy, M. Gilbert, P. Lemey, and O. Pybus. Using viral gene sequences to compare and explain the heterogeneous spatial dynamics of virus epidemics. *Molecular Biology and Evolution*, 34:2563–2571, 2017.

[9] S. Dellicour, G. Baele, G. Dudas, N. Faria, O. Pybus, M. Suchard, A. Rambaut, and P. Lemey. Phylodynamic assessment of intervention strategies for the West African Ebola virus outbreak. *Nature Communications*, 9: 1–9, 2018.

[10] S. Dellicour, B. Vrancken, N. Trovão, D. Fargette, and P. Lemey. On the importance of negative controls in viral landscape phylogeography. *Virus Evolution*, 4:vey023, 2018.

[11] S. Dellicour, S. Lequime, B. Vrancken, M. Gill, P. Bastide, K. Gangavarapu, N. Matteson, Y. Tan, L. Du Plessis, A. Fisher, et al. Epidemiological hypothesis testing using a phylogeographic and phylodynamic framework. *Nature Communications*, 11:5620, 2020.

[12] S. Dellicour, P. Bastide, P. Rocu, D. Fargette, O. Hardy, M. Suchard, S. Guindon, and P. Lemey. How fast are viruses spreading in the wild? *bioRxiv*, pages 2024–04, 2024.

[13] A. Drummond, O. Pybus, A. Rambaut, R. Forsberg, and A. Rodrigo. Measurably evolving populations. *Trends in Ecology and Evolution*, 18:481–488, 2003.

[14] A. Drummond, S. Ho, M. Phillips, and A. Rambaut. Relaxed phylogenetics and dating with confidence. *PLOS Biology*, 4:e88, 2006.

[15] A. Etheridge. Drift, draft and structure: some mathematical models of evolution. *Banach center publications*, 80:121–144, 2008.

[16] J. Felsenstein. A pain in the torus: some difficulties with models of isolation by distance. *American Naturalist*, 109:359–368, 1975.

[17] M. Gill, Lam S. Tung H., G. Baele, P. Lemey, and M. Suchard. A relaxed directional random walk model for phylogenetic trait evolution. *Systematic Biology*, 66:299–319, 2016.

[18] S. Guindon and N. De Maio. Accounting for spatial sampling patterns in Bayesian phylogeography. *Proceedings of the National Academy of Sciences*, 118:e2105273118, 2021.

[19] S. Guindon, H. Guo, and D. Welch. Demographic inference under the coalescent in a spatial continuum. *Theoretical Population Biology*, 111: 43–50, 2016.

[20] M. Hasegawa, H. Kishino, and T. Yano. Dating of the Human-Ape splitting by a molecular clock of mitochondrial-DNA. *Journal of Molecular Evolution*, 22:160–174, 1985.

[21] G. Hassler, B. Gallone, L. Aristide, W. Allen, M. Tolkoff, A. Holbrook, G. Baele, P. Lemey, and M. Suchard. Principled, practical, flexible, fast: A new approach to phylogenetic factor analysis. *Methods in Ecology and Evolution*, 13:2181–2197, 2022.

[22] S. Ho and B. Shapiro. Skyline-plot methods for estimating demographic history from nucleotide sequences. *Molecular Ecology Resources*, 11:423–434, 2011.

[23] E. Holmes. What can we predict about viral evolution and emergence? *Current Opinion in Virology*, 3:180–184, 2013.

[24] E. Holmes, A. Rambaut, and K. Andersen. Pandemics: spend on surveillance, not prediction. *Nature*, 2018.

[25] M. Hooten and D. Johnson. Basis function models for animal movement. *Journal of the American Statistical Association*, 112:578–589, 2017.

[26] S. Issaka, O. Traoré, R. Dimitri S. Longué, A. Pinel-Galzi, M. Gill, S. Dellicour, P. Bastide, S. Guindon, E. Hébrard, M-J. Dugué, et al. Rivers and landscape ecology of a plant virus, rice yellow mottle virus along the niger valley. *Virus Evolution*, 7:veab072, 2021.

[27] J. Kingman. The coalescent. *Stochastic Processes and their Applications*, 13:235–248, 1982.

[28] R. Klitting, L. Kafetzopoulou, W. Thiery, G. Dudas, S. Gryseels, A. Kotamarthi, B. Vrancken, K. Gangavarapu, M. Momoh, J. Sandi, et al. Predicting the evolution of the Lassa virus endemic area and population at risk over the next decades. *Nature communications*, 13:5596, 2022.

[29] R Lanciotti, J Roehrig, V Deubel, J Smith, M Parker, K Steele, B Crise, K Volpe, M Crabtree, J Scherret, et al. Origin of the West Nile virus responsible for an outbreak of encephalitis in the northeastern United States. *Science*, 286(5448):2333–2337, 1999.

[30] P. Lemey, A. Rambaut, J. Welch, and M. Suchard. Phylogeography takes a relaxed random walk in continuous space and time. *Molecular Biology and Evolution*, 27:1877–1885, 2010.

[31] G. Malécot. *Mathematics of heredity*. Paris: Masson et Cie., 1948.

[32] O. Pybus, M. Suchard, P. Lemey, F. Bernardin, A. Rambaut, F. Crawford, R. Gray, N. Arinaminpathy, S. Stramer, M. Busch, et al. Unifying the spatial epidemiology and molecular evolution of emerging epidemics. *Proceedings of the National Academy of Sciences*, 109:15066–15071, 2012.

[33] J. Raghwani, A. Rambaut, E. Holmes, Vu Ty Hang, Tran Tinh Hien, J. Farrar, B. Wills, N. Lennon, B. Birren, M. Henn, et al. Endemic dengue associated with the co-circulation of multiple viral lineages and localized density-dependent transmission. *PLoS Pathogens*, 7:e1002064, 2011.

[34] M. Rakotomalala, B. Vrancken, A. Pinel-Galzi, P. Ramavovololona, E. Hebrard, J. Randrianangaly, S. Dellicour, P. Lemey, and D. Fargette. Comparing patterns and scales of plant virus phylogeography: Rice yellow mottle virus in Madagascar and in continental Africa. *Virus Evolution*, 5:vez023, 2019.

[35] M. Suchard, P. Lemey, G. Baele, D. L Ayres, A. Drummond, and A. Rambaut. Bayesian phylogenetic and phylodynamic data integration using BEAST 1.10. *Virus Evolution*, 4:vey016, 2018.

[36] J. Sy, J. Taylor, and W. Cumberland. A stochastic model for the analysis of bivariate longitudinal AIDS data. *Biometrics*, 53:542–555, 1997.

[37] C. Talbi, P. Lemey, M. Suchard, E. Abdelatif, M. Elharrak, N. Jalal, A. Faouzi, J. Echevarría, Vazquez S., A. Rambaut, et al. Phylodynamics and human-mediated dispersal of a zoonotic virus. *PLoS pathogens*, 6: e1001166, 2010.

[38] C. Tisseuil, A. Gryspeirt, R. Lancelot, M. Pioz, A. Liebhold, and M. Gilbert. Evaluating methods to quantify spatial variation in the velocity of biological invasions. *Ecography*, 39:409–418, 2016.

[39] N. Trovão, G. Baele, B. Vrancken, F. Bielejec, M. Suchard, D. Fargette, and P. Lemey. Host ecology determines the dispersal patterns of a plant virus. *Virus evolution*, 1:vev016, 2015.

[40] F. van den Bosch, R. Hengeveld, and J. Metz. Analysing the velocity of animal range expansion. *Journal of Biogeography*, 19:135–150, 1992.

[41] M. Wille, J. Geoghegan, and E. Holmes. How accurately can we assess zoonotic risk? *PLoS biology*, 19:e3001135, 2021.

[42] J. Wirtz and S. Guindon. On the connections between the spatial Lambda-Fleming-Viot model and other processes for analysing geo-referenced genetic data. *arXiv preprint arXiv:2307.02431*, 2023.

[43] S. Wright. Isolation by distance. *Genetics*, 28:114–138, 1943.

[44] E. Zuckerkandl and L. Pauling. Molecules as documents of evolutionary history. *Journal of Theoretical Biology*, 8:357–366, 1965.

5. Supplementary Information

5.1. Distribution of coordinates for the IOU process conditioned on start velocity

The distribution is normal with expectation and variance given below:

$$\mathrm{E}(X(t) \mid x(0), y(0)) = x(0) + (y(0) - \mu)\left(\frac{1 - e^{-\theta t}}{\theta}\right) + \mu t$$

$$\mathrm{V}(X(t) \mid x(0), y(0)) = \frac{\sigma^2}{\theta^3}\left(t\theta - \frac{3}{2} - \frac{e^{-2t\theta}}{2} + 2e^{-t\theta}\right)$$

On expectation, the coordinates thus grow linearly with time, with a rate given by the trend μ of the underlying OU, when the process has reached its equilibrium, which happens when $e^{-\theta t} \simeq 0$. The variance of $X(t)$ grows exponentially for values of θt smaller than ~ 2 and linearly beyond that. Note that when θ goes to 0, the variance converges to $\sigma^2 t^3/3$, so that the IOU converges to an IBM.

5.2. Distributions of velocity and coordinates for PIV models conditioned on start and end velocities

Velocities. We consider the situation where velocities at the start and the end of a branch of length t, $y(0)$ and $y(t)$, are given, and aim at deriving the distribution of the location at the end of the branch, $X(t)$, given the location at the start, $x(0)$. Characterizing this distribution is required in the calculation of the likelihood of the PIV models in the PhyREX approach (see main text, section 4.1).

In case the velocity evolves according to a Brownian bridge starting at $y(0)$ and stopping at $y(t)$, the corresponding process is defined as follows:

$$Y(s) = \frac{s}{t}y(t) + \left(\frac{t-s}{t}\right)y(0) + W(s) - \frac{s}{t}W(t) \tag{3}$$

where W denotes the Wiener process (with $W(0) = 0$). The expected value and variance of velocity at time s (with $0 \leq s \leq t$) are thus as given below:

$$\begin{aligned} \mathrm{E}(Y(s) \mid y(0), y(t)) &= \frac{s}{t}(y(t) - y(0)) + y(0) \\ \mathrm{Cov}(Y(u), Y(v)) &= \sigma^2 \frac{u(t-v)}{t} \end{aligned}$$

with $u \leq v \leq t$ and $s \leq t$.

When the velocity follows a OU bridge, Lemma 1 in [16] shows that:

$$Y(s) = \frac{\sinh(\theta s)}{\sinh(\theta t)}y(t) + \frac{\sinh(\theta(t-s))}{\sinh(\theta t)}y(0) + W(s) - \frac{\sinh(\theta s)}{\sinh(\theta t)}W(t) \tag{4}$$

where W is the non-tied OU process with $W(0) = 0$. The distribution of velocity is thus normal here with expectation and covariance as follows:

$$\begin{aligned} \mathrm{E}(Y(s) \mid y(0), y(t)) &= \frac{\sinh(\theta s)}{\sinh(\theta t)}y(t) + \frac{\sinh(\theta(t-s))}{\sinh(\theta t)}y(0) \\ &\quad \left[+ \mu\left(1 - e^{-\theta s}\right) - \frac{\sinh(\theta s)}{\sinh(\theta t)}\mu\left(1 - e^{-\theta t}\right) \right. \\ \mathrm{Cov}(Y(u), Y(v)) &= \frac{\sigma^2}{\theta}\frac{\sinh(\theta u)\sinh(\theta(t-v))}{\sinh(\theta t)} \end{aligned}$$

with $u \leq v \leq t$ and $s \leq t$.

Spatial coordinates. We use the mean and covariance of the constrained velocity derived in the previous section in order to derive that of the spatial coordinates at the end of an edge of length t, given the coordinates at the start of

that branch along with the velocity at both extremities of the same edge. We thus focus on $\mathrm{E}(X(t) \mid x(0), y(0), y(t)) = x(0) + \int_0^t \mathrm{E}(Y(s) \mid y(0), y(t))\mathrm{d}s$ and $\int_0^t \int_0^t \mathrm{Cov}(Y(u), Y(v))\mathrm{d}u\mathrm{d}v$ and obtain the following expressions:

$$\begin{aligned}
\mathrm{E}(X(t) \mid x(0), y(0), y(t)) &= x(0) + \frac{t}{2}(y(t) + y(0)) \\
\mathrm{V}(X(t) \mid x(0), y(0), y(t)) &= \frac{\sigma^2 t^3}{12}.
\end{aligned}$$

The expectation of $X(t)$ therefore grows linearly with t at a pace determined by the velocity averaged over the two nodes at the extremity of the branch under scrutiny.

Using an approach equivalent to that applied to the IBM process, the expectation and variance of the coordinates given the velocities at both extremities of an edge are given below for the IOU model:

$$\begin{aligned}
\mathrm{E}(X(t) \mid x(0), y(0), y(t)) = x(0) + \left(\frac{\cosh(\theta t) - 1}{\theta \sinh(\theta t)}\right)(y(t) + y(0)) \\
\left[+ \mu t - \frac{\mu}{\theta}\left(1 - e^{-\theta t}\right)\left(1 + \frac{\cosh(\theta t) - 1}{\sinh(\theta t)}\right)\right. \\
\mathrm{V}(X(t) \mid x(0), y(0), y(t)) = \frac{\sigma^2}{\theta^3}\left(\theta t - 2\left(\frac{\cosh(\theta t) - 1}{\sinh(\theta t)}\right)\right)
\end{aligned}$$

When $\theta \ll 1$, i.e., the pace to reach the equilibrium is slow with respect to t, we have $(1 - e^{-\theta t})/\theta \to t$ and $(\cosh(\theta t) - 1)/\sinh(\theta t) \sim \theta t/2$ so that $\mathrm{E}(X(t) \mid x(0), y(0), y(t)) \simeq x(0) + (t/2)(y(t) + y(0))$. In addition, developing the variance term up to order 4 at the numerator and order 3 at the denominator, we get that $(\cosh(\theta t) - 1)/\sinh(\theta t) \sim (\theta t)/2 - (\theta t)^3/24$, so that $\mathrm{V}(X(t) \mid x(0), y(0), y(t)) \simeq \sigma^2 t^3/12$. As expected, when θ goes to 0, the IOU and the IBM processes behave similarly. Also, note that the function $f : x \mapsto x - 2(\cosh(x) - 1)/\sinh(x)$ converges to 0 when $x \to 0^+$ and the derivative of f with respect to x is non negative, so that f is non negative for $x \geq 0$, and $\mathrm{V}(X(t) \mid x(0), y(0), y(t))$ as expressed above is non negative for all values of $\theta t \geq 0$. If θ is large, then $\mathrm{E}(X(t) \mid x(0), y(0), y(t)) \simeq x(0) + ((y(t) - \mu) + (y(0) - \mu))/\theta + \mu t) \simeq x(0) + \mu t$, and $\mathrm{V}(X(t) \mid x(0), y(0), y(t)) \simeq \frac{\sigma^2}{\theta^3}(\theta t - 2)$.

Hence, in both regimes of θ, the expected value of the coordinates grows linearly with time. It is doing so in a manner that is proportional to the velocities at both extremities of the branch (if θ is small) or proportional to the drift term (if θ is large). In terms of variance, as previously, the IOU behaves similarly to the IBM when θ is small, and grows linearly in t when θ is large.

Joint process. We assume here a general multivariate integrated process of dimension p (typically, $p = 2$ for phylogeography), and denote by $\mathbf{Z}(t) = (\mathbf{Y}^T(t), \mathbf{X}^T(t))^T$ the joint process of dimension $2p$ describing the evolution of both the velocity and position vectors. This joint process is then a linear Gaussian process, and it can be fully described by the Gaussian distribution of the trait at any node i given the trait at its parent $\mathrm{pa}(i)$ [13, 2]:

$$\mathbf{Z}_i \mid \mathbf{Z}_{\mathrm{pa}(i)} \sim \mathcal{N}\big(\mathbf{q}_i \mathbf{Z}_{\mathrm{pa}(i)} + \mathbf{r}_i, \mathbf{\Sigma}_i\big)\,, \tag{5}$$

with $\mathbf{q}_i$ an actualization matrix and $\mathbf{\Sigma}_i$ a variance matrix both of size $2p \times 2p$, and $\mathbf{r}_i$ a vector or size $2p$, that are all independent from the data, and depend only on the tree and its branch lengths. For the IBM, assuming that the velocity vector follows a BM with constant directional drift $\boldsymbol{\delta}$ and variance $\mathbf{\Sigma}$, we get:

$$\begin{aligned}
\mathbf{q}_i &= \begin{pmatrix} \mathbf{I}_p & \mathbf{0}_{p,p} \\ \mathbf{I}_p t_i & \mathbf{I}_p \end{pmatrix} &&= \begin{pmatrix} 1 & 0 \\ t_i & 1 \end{pmatrix} \otimes \mathbf{I}_p \\
\mathbf{r}_i &= \begin{pmatrix} \boldsymbol{\delta} t_i \\ \boldsymbol{\delta} t_i^2/2 \end{pmatrix} &&= \begin{pmatrix} t_i \\ t_i^2/2 \end{pmatrix} \otimes \boldsymbol{\delta} \\
\mathbf{\Sigma}_i &= \begin{pmatrix} \mathbf{\Sigma} t_i & \mathbf{\Sigma} t_i^2/2 \\ \mathbf{\Sigma} t_i^2/2 & \mathbf{\Sigma} t_i^3/3 \end{pmatrix} &&= \begin{pmatrix} t_i & t_i^2/2 \\ t_i^2/2 & t_i^3/3 \end{pmatrix} \otimes \mathbf{\Sigma},
\end{aligned}$$

where t_i is the length of the branch going from $\mathrm{pa}(i)$ to i, and $\otimes$ denotes the Kronecker product. For the IOU, assuming that the velocity vector follows a OU with actualization matrix $\mathbf{\Theta}$, optimal vector $\boldsymbol{\mu}$ and variance $\mathbf{\Sigma}$, we get [4]:

$$\begin{aligned}
\mathbf{q}_i &= \begin{pmatrix} e^{-\mathbf{\Theta} t_i} & \mathbf{0}_{p,p} \\ \mathbf{\Theta}^{-1}(\mathbf{I}_p - e^{-\mathbf{\Theta} t_i}) & \mathbf{I}_p \end{pmatrix} \\
\mathbf{r}_i &= \begin{pmatrix} (\mathbf{I}_p - e^{-\mathbf{\Theta} t_i})\boldsymbol{\mu} \\ \boldsymbol{\mu} t_i - \mathbf{\Theta}^{-1}(\mathbf{I}_p - e^{-\mathbf{\Theta} t_i})\boldsymbol{\mu} \end{pmatrix} \\
\mathbf{\Sigma}_i &= \begin{pmatrix} \mathbf{K}_{\mathbf{Y}}(t_i) & \mathbf{K}_{\mathbf{X},\mathbf{Y}}(t_i)^T \\ \mathbf{K}_{\mathbf{X},\mathbf{Y}}(t_i) & \mathbf{K}_{\mathbf{X}}(t_i) \end{pmatrix},
\end{aligned}$$

with, for any t, $\mathbf{K}_{\mathbf{Y}}(t)$, $\mathbf{K}_{\mathbf{X}}(t)$, and $\mathbf{K}_{\mathbf{X},\mathbf{Y}}(t)$, the variance and covariance of the velocity and position vectors at time t, given by:

$$\begin{aligned}
\mathbf{K}_{\mathbf{Y}}(t) &= \mathbf{S} - e^{-\mathbf{\Theta} t} \mathbf{S} e^{-\mathbf{\Theta}^T t} \\
\mathbf{K}_{\mathbf{X}}(t) &= \mathbf{\Delta} t - (\mathbf{I}_p - e^{-\mathbf{\Theta} t})\mathbf{\Theta}^{-1}\mathbf{\Delta} - \mathbf{\Delta}\mathbf{\Theta}^{-T}(\mathbf{I}_p - e^{-\mathbf{\Theta}^T t}) \\
&\quad + \mathbf{\Theta}^{-1}\mathbf{S}\mathbf{\Theta}^{-T} - e^{-\mathbf{\Theta} t}\mathbf{\Theta}^{-1}\mathbf{S}\mathbf{\Theta}^{-T} e^{-\mathbf{\Theta}^T t} \\
\mathbf{K}_{\mathbf{X},\mathbf{Y}}(t) &= \mathbf{S}\mathbf{\Theta}^{-1}[\mathbf{I}_p - e^{-\mathbf{\Theta} t}] - [\mathbf{I}_p - e^{-\mathbf{\Theta} t}]\mathbf{\Theta}^{-1}\mathbf{S} e^{-\mathbf{\Theta}^T t}
\end{aligned}$$

with $\mathbf{S}$ the stationary variance of the OU, and $\mathbf{\Delta} = \mathbf{\Theta}^{-1}\mathbf{S} + \mathbf{S}\mathbf{\Theta}^{-T}$.

Pruning algorithm. As showed in [13, 2], for a general linear Gaussian process of the form of Equation 5 we can compute the likelihood of the traits at the tips of the tree in an efficient way by integrating over all the internal states using a pruning algorithm. This yields an algorithm in $O(np^3)$, that is linear in the number of tips. In the case of an integrated process, the velocity is never observed at the tips, only the spatial coordinates are. As shown in [2], missing data can be accounted for in this algorithm, and because the velocity and position traits are correlated, we can still get information on the velocity, even though it is not observed. This approach is implemented in BEAST [20], and allows for the Bayesian fit of the model without resorting to stochastic integration of the internal nodes as is done in PhyREX. Note that, although we used a standard Metropolis-Hastings update for the parameters of the processes, from [2] we could also get derivative with respect to the parameters of the IBM or IOU, allowing for efficient Hamiltonian Monte Carlo sampling schemes [15].

5.4. Marginal tip position distribution under the IBM and IOU processes

In the previous sections, we used the conditional distribution of a node given its parent to derive efficient pruning algorithm for the computation of the likelihood. However, as the IBM and IOU are Gaussian processes, it is also possible to directly derive the marginal distribution of the observed positions at the tip of the process. Denote by $\mathbf{X}_\mathrm{t}$ the $n \times p$ matrix of observations at the tips of the tree, and by $\mathbf{X}_\mathrm{t}^i$ the vector of p observations at tip i ($\mathbf{X}_\mathrm{t} = (\mathbf{X}_\mathrm{t}^1 \ \cdots \ \mathbf{X}_\mathrm{t}^n)^T$).

Marginal distribution under a BM. Recall that for a simple multivariate BM on a tree with rate variance $\mathbf{\Sigma}$ and root position parameter $\mathbf{X}_\rho \sim \mathcal{N}(\boldsymbol{\mu}^\mathbf{X}, \mathbf{\Gamma}^\mathbf{X})$, then we get that $\mathbf{X}_\mathrm{t}$ has a matrix normal distribution, with, for any two vectors of observations at tips i and j, an expectation vector and variance covariance matrix given by:

$$\mathrm{E}[\mathbf{X}_\mathrm{t}^i] = \boldsymbol{\mu}^\mathbf{X} \quad \text{and} \quad \mathrm{V}[\mathbf{X}_\mathrm{t}^i; \mathbf{X}_\mathrm{t}^j] = \mathbf{\Sigma}\tau_{ij} + \mathbf{\Gamma}^\mathbf{X}$$

with τ_{ij} the time between the root and the most recent common ancestor of i and j (see e.g. [7]).

Marginal distribution under an IBM. We now assume that the process is an IBM, with a root position parameter $\mathbf{X}_\rho \sim \mathcal{N}(\boldsymbol{\mu}^\mathbf{X}, \mathbf{\Gamma}^\mathbf{X})$, a root velocity parameter $\mathbf{Y}_\rho \sim \mathcal{N}(\boldsymbol{\mu}^\mathbf{Y}, \mathbf{\Gamma}^\mathbf{Y})$, and rate variance matrix of the velocities equal to $\mathbf{\Sigma}$. Then the distribution of the traits at the tips is still matrix normal. More precisely, as the velocities are simply Brownian, we get that:

$$\mathrm{E}[\mathbf{Y}_\mathrm{t}^i] = \boldsymbol{\mu}^\mathbf{Y} \quad \text{and} \quad \mathrm{V}[\mathbf{Y}_\mathrm{t}^i; \mathbf{Y}_\mathrm{t}^j] = \mathbf{\Sigma}\tau_{ij} + \mathbf{\Gamma}^\mathbf{X}.$$

For the positions, we get:

$$\begin{aligned}
\mathrm{E}[\mathbf{X}_{\mathrm{t}}^i] &= \boldsymbol{\mu}^{\mathbf{Y}}\tau_i + \boldsymbol{\mu}^{\mathbf{X}} \\
\mathrm{V}[\mathbf{X}_{\mathrm{t}}^i;\mathbf{X}_{\mathrm{t}}^j] &= \boldsymbol{\Gamma}^{\mathbf{X}} + \boldsymbol{\Gamma}^{\mathbf{Y}}\tau_i\tau_j + \boldsymbol{\Sigma}\tau_{ij}\left[\tau_i\tau_j + \tau_{ij}\left(\frac{\tau_{ij}}{3} - \frac{\tau_i+\tau_j}{2}\right)\right].
\end{aligned}$$

Note that, if $i = j$, then $\tau_{ii} = \tau_i$, and we recover that the variance of a tip is cubic in the time of evolution. We can also get the covariances between velocities and positions:

$$\mathrm{V}[\mathbf{X}_{\mathrm{t}}^i;\mathbf{Y}_{\mathrm{t}}^j] = \boldsymbol{\Gamma}^{\mathbf{Y}}\tau_i + \boldsymbol{\Sigma}\tau_{ij}\left[\tau_i - \frac{\tau_{ij}}{2}\right].$$

The proof of these formulas rely on the following equality:

$$\begin{aligned}
\mathrm{V}[\mathbf{X}_{\mathrm{t}}^i;\mathbf{X}_{\mathrm{t}}^j] &= \int_0^{\tau_i}\int_0^{\tau_j} \mathrm{V}[\mathbf{Y}^i(t);\mathbf{Y}^j(s)]dtds \\
\mathrm{V}[\mathbf{X}_{\mathrm{t}}^i;\mathbf{Y}_{\mathrm{t}}^j] &= \int_0^{\tau_i} \mathrm{V}[\mathbf{Y}^i(t);\mathbf{Y}_{\mathrm{t}}^j]dt,
\end{aligned}$$

where $\mathbf{Y}^i(t)$ denotes the value of the velocity process on lineage leading to tip i at time t, and the covariance function is equal to:

$$\mathrm{V}[\mathbf{Y}^i(t);\mathbf{Y}^j(s)] = \begin{cases} \boldsymbol{\Sigma}\min(s,t) + \boldsymbol{\Gamma}^{\mathbf{Y}} & \text{if } s \le \tau_{ij} \text{ or } t \le \tau_{ij} \\ \boldsymbol{\Sigma}\tau_{ij} + \boldsymbol{\Gamma}^{\mathbf{Y}} & \text{otherwise.} \end{cases}$$

Marginal distribution under an IOU. We now assume that the process is an IOU, with a root position parameter $\mathbf{X}_\rho \sim \mathcal{N}(\boldsymbol{\mu}^{\mathbf{X}}, \boldsymbol{\Gamma}^{\mathbf{X}})$, a root velocity parameter $\mathbf{Y}_\rho \sim \mathcal{N}(\boldsymbol{\mu}^{\mathbf{Y}}, \boldsymbol{\Gamma}^{\mathbf{Y}})$, rate variance matrix of the velocities equal to $\boldsymbol{\Sigma}$, actualization matrix $\boldsymbol{\Theta}$, optimal vector $\boldsymbol{\mu}$ and stationary variance $\mathbf{S}$. Then the distribution of the traits at the tips is still matrix normal, and we get that the velocities are OU distributed as (see e.g. [3]):

$$\begin{aligned}
\mathrm{E}[\mathbf{Y}_{\mathrm{t}}^i] &= e^{-\boldsymbol{\Theta}t}\boldsymbol{\mu}^{\mathbf{Y}} + (\mathbf{I}_p - e^{-\boldsymbol{\Theta}t})\boldsymbol{\mu} \quad \text{and} \\
\mathrm{V}[\mathbf{Y}_{\mathrm{t}}^i;\mathbf{Y}_{\mathrm{t}}^j] &= e^{-\boldsymbol{\Theta}(\tau_i-\tau_{ij})}\mathbf{S}e^{-\boldsymbol{\Theta}^T(\tau_j-\tau_{ij})} - e^{-\boldsymbol{\Theta}\tau_i}\mathbf{S}e^{-\boldsymbol{\Theta}^T\tau_i} + e^{-\boldsymbol{\Theta}\tau_i}\boldsymbol{\Gamma}^{\mathbf{Y}}e^{-\boldsymbol{\Theta}^T\tau_i}.
\end{aligned}$$

For the positions, we get:

$$\begin{aligned}
\mathrm{E}[\mathbf{X}_{\mathrm{t}}^i] &= \boldsymbol{\mu}^{\mathbf{X}} + \boldsymbol{\Theta}^{-1}(\mathbf{I}_p - e^{-\boldsymbol{\Theta}t})(\boldsymbol{\mu}^{\mathbf{Y}} - \boldsymbol{\mu}) + \boldsymbol{\mu}t \quad \text{and} \\
\mathrm{V}[\mathbf{Y}_{\mathrm{t}}^i;\mathbf{Y}_{\mathrm{t}}^j] &= \boldsymbol{\Gamma}^{\mathbf{X}} + [\mathbf{I}_p - e^{-\boldsymbol{\Theta}\tau_i}]\boldsymbol{\Theta}^{-1}\boldsymbol{\Gamma}^{\mathbf{Y}}\boldsymbol{\Theta}^{-T}[\mathbf{I}_p - e^{-\boldsymbol{\Theta}^T\tau_j}] \\
&\quad + \boldsymbol{\Delta}\tau_{ij} + (\mathbf{I}_p - e^{\boldsymbol{\Theta}\tau_{ij}})e^{-\boldsymbol{\Theta}\tau_i}\boldsymbol{\Theta}^{-1}\boldsymbol{\Delta} + \boldsymbol{\Delta}\boldsymbol{\Theta}^{-T}e^{-\boldsymbol{\Theta}^T\tau_j}(\mathbf{I}_p - e^{\boldsymbol{\Theta}^T\tau_{ij}}) \\
&\quad + e^{-\boldsymbol{\Theta}\tau_i}\left[e^{\boldsymbol{\Theta}\tau_{ij}}\boldsymbol{\Theta}^{-1}\mathbf{S}\boldsymbol{\Theta}^{-T}e^{\boldsymbol{\Theta}^T\tau_{ij}} - \boldsymbol{\Theta}^{-1}\mathbf{S}\boldsymbol{\Theta}^{-T}\right]e^{-\boldsymbol{\Theta}^T\tau_j}
\end{aligned}$$

using the same notations as in section 5.3 for $\boldsymbol{\Delta}$.

The proof of the formulas rely on the integration of the following covariance function (see [4]), valid for any $0 \leq s \leq t$:

$$\mathrm{V}[\mathbf{Y}^i(t); \mathbf{Y}^j(s)] = \begin{cases} \mathbf{S}e^{-\mathbf{\Theta}(t-s)} - e^{-\mathbf{\Theta}s}\mathbf{S}e^{-\mathbf{\Theta}^T t} + e^{-\mathbf{\Theta}s}\mathbf{\Gamma}^{\mathbf{Y}}e^{-\mathbf{\Theta}^T t} & \text{if } s \leq \tau_{ij} \text{ or } t \leq \tau_{ij} \\ e^{-\mathbf{\Theta}(s-\tau_{ij})}\mathbf{S}e^{-\mathbf{\Theta}^T(t-\tau_{ij})} - e^{-\mathbf{\Theta}s}\mathbf{S}e^{-\mathbf{\Theta}^T t} + e^{-\mathbf{\Theta}s}\mathbf{\Gamma}^{\mathbf{Y}}e^{-\mathbf{\Theta}^T t} & \text{otherwise.} \end{cases}$$

Likelihood computation checks. The formulas above give the full distribution of the observed positions at the tips of the tree, that is matrix normal, and can hence be used to compute the likelihood of the data directly. As in the standard BM case, this direct computation involves the inversion of large matrices, and the pruning approach of section 5.3 is to be preferred, as it is linear in the number of tips (see e.g. [13, 2]). However, these formulas can be used to test that the likelihoods obtained by BEAST and PhyREX do match with the direct "naïve" formulas. We implemented these direct formulas in `R` [19], and checked on small examples that the two software gave matching likelihoods.

5.5. Comparison of PhyREX and BEAST implementations

We compared two independent implementations of the IBM model in PhyREX and BEAST, with the WNV data from Pybus et al. [17] (104 tips), independent IBM models for the latitudes and longitudes, and a vague Gaussian prior centered at 0 and with variance 1000 for the root trait. We compared the two implementations first using a common fixed tree, and then starting from sequences and inferring the tree.

Settings. For the fixed tree analysis, as PhyREX needs to sample interval velocities, we ran a longer MCMC chain for this software, with 10 million iterations sampled every 10 thousands steps, while we used a chain with only 10 thousands iterations sampled every 10 steps for BEAST. We reproduced this analysis 10 times on a 13-inch M2 2022 MacBook Pro, and took the mean times and estimates.

For the inferred tree analysis, we ran a MCMC chain with 50 million iterations sampled every 10 thousands steps for both software, and as the analyses are computationally intensive, only ran the chain once with the same set up and compared running times. For both analyses, we used a constant population coalescent prior, an HKY substitution model [10], and an uncorrelated relaxed random local clock [5] with a log-normal distribution of rates with a strong prior on a small standard error to speed up convergence. Tip heights were fixed at their sampling times.

	PhyREX			BEAST		
parameter	ESS/s	mean	95% HPDI	ESS/s	mean	95% HPDI
σ^2 lat	9.73	2.37	(1.7, 3.1)	100.15	2.29	(1.7, 3.0)
σ^2 lon	10.82	11.81	(8.7, 15.3)	111.90	11.32	(8.4, 14.5)
root lat	5.74	39.88	(32.9, 47.1)	164.66	40.65	(33.7, 47.6)
root lon	6.78	-71.07	(-86.2, -55.7)	164.11	-74.62	(-90.3, -59.3)
root veloc lat	17.63	-0.13	(-3.4, 3.3)	166.60	-0.26	(-3.5, 3.1)
root veloc lon	17.81	-3.80	(-11.0, 3.7)	161.19	-2.32	(-9.6, 4.9)

Table 1: **Comparison of PhyREX and BEAST estimations of variance and root parameters** for a bi-variate IBM on a fixed tree (WNV data of [17] with 104 tips).

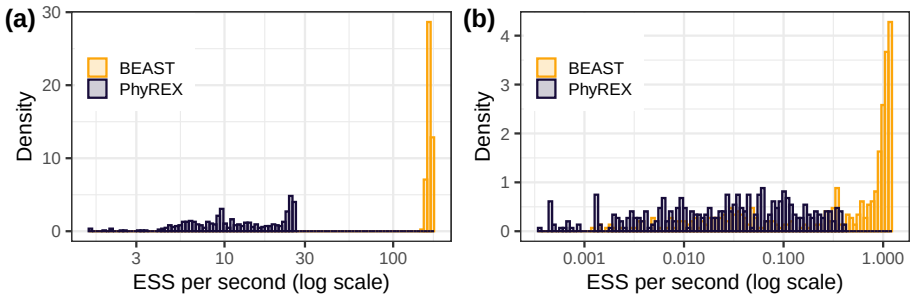

Figure 6: **Effective sample size per seconds of all node velocities (internal and tips) of BEAST versus PhyREX for a bi-variate IBM on a fixed tree (a) or with an inferred tree (b).** WNV data of [17] with 104 tips, i.e. 208 velocity estimates per software program.

We used `R` [19] to run analyses, using packages `ape` [6] and `treeio` [21] for tree manipulation, `tracerer` [18] for reading and summarizing log files, `ggplot2` [8] and `cowplot` [22] for plotting results, `here` [12] for file manipulation, and `kableExtra` [9] for extracting tables automatically for display.

Results for the fixed tree analysis. We found that the two approaches gave very similar results for the estimation of the variance and root parameters (see Table 1). Unsurprisingly, as BEAST only needs to sample two parameters (the variance parameters), it was faster in this setting, with analyses taking around 5.2 seconds, versus 33.6 seconds with PhyREX. This led to a mean approximate 15.3 times factor increase of the BEAST versus PhyREX implementation in terms of effective sample size per seconds (see Table 1). Because velocities are sampled during the MCMC in PhyREX, but directly sampled in their posterior distribution in BEAST, their ESS in BEAST was more reliably high and less spread out than PhyREX estimations (see Fig. 6-a). Note that each PhyREX iteration was about 155.0 times faster than each BEAST iteration, as it requires less computations.

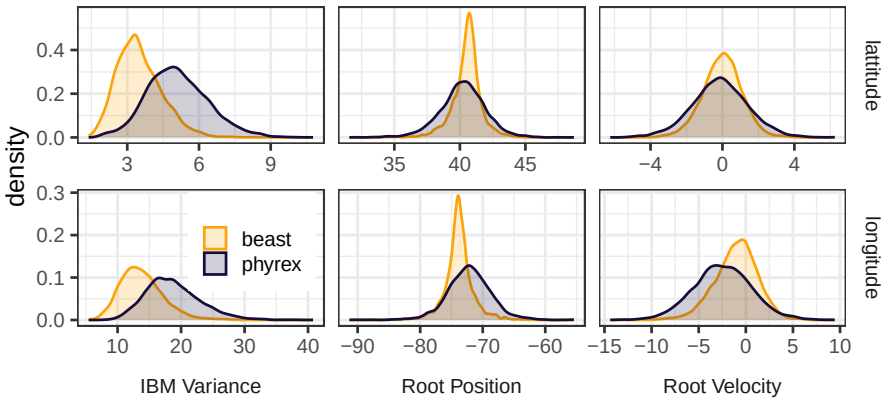

Figure 7: **Comparison of PhyREX and BEAST estimations of variance and root parameters for a bi-variate IBM when the tree is inferred** using WNV data of [17] with 104 tips.

Results for the inferred tree analysis. As inferring the tree is much more computationally intensive, these analyses took longer to run, taking, respectively, 1.05 hours for BEAST and 2.85 hours for PhyREX. PhyREX convergence was slower, taking around 10 million iterations to warm up, while BEAST reached a reasonable sampling area within less than 1 million iterations. This led to a mean approximate 13.70 times factor increase of the BEAST versus PhyREX implementation in term of effective sample size per seconds. As in the fixed tree case, ESS in BEAST was more reliably high and less spread out than PhyREX estimations (see Fig. 6-b). Because many operators are involved in the tree search, it is difficult to pinpoints the exact reasons for these different convergence behaviors, and it may not be entirely due to the different IBM implementations. Contrary to the fixed tree case, each BEAST iteration was faster than a PhyREX iteration by a factor of about 2.70. This time difference is likely due to the use of BEAGLE [1] in the BEAST analyse, which allows for efficient parallelization and GPU use. The two approaches gave very similar results for the estimation of the root time and position, but gave slightly different variance parameter estimations, although with intersecting HPD intervals (see Fig. 7). Reconstructed velocities at tips were highly similar in both approaches (data not shown).

5.6. Dispersal prediction via linear extrapolation of tip velocities

Let $\alpha_i(t)$ design the coordinates of tip i at time t that occurs after the sample corresponding to tip i was observed, which is noted as t_i (i.e. $t \geq t_i$). $\alpha_i(t)$ then corresponds to the position of lineage i, should it survive up to time t. Also,

we assume that, after time t_i, lineage i dies at rate λ so that the probability of surviving up to time t is $\exp(-\lambda(t - t_i))$. Finally, $\mathcal{A}$ represents a particular region, e.g., a state or county. Below is the probability that one or more lineage occupies $\mathcal{A}$ at time t, where $t \geq t_i$ for all tips $i = 1, \ldots, n$. Let $\mathcal{P}_{\mathcal{A},t}$ be that probability. We have:

$$\mathcal{P}_{\mathcal{A},t} = 1 - \prod_{i=1}^{n} \Pr(\alpha_i(t) \notin \mathcal{A})$$

where $\Pr(\alpha_i(t) \notin \mathcal{A})$ is the probability that lineage i is not within $\mathcal{A}$ at time t. We have $\Pr(\alpha_i(t) \notin \mathcal{A}) = 1 - \Pr(\alpha_i(t) \in \mathcal{A})$ and $\Pr(\alpha_i(t) \in \mathcal{A})$ is the probability that (1) lineage i survives up to time t and (2) the linear extrapolation of its position from time t_i to time t given its velocity at time t_i, falls within $\mathcal{A}$. In practice, we are interested in the probability of occupation for a given time interval $[t, t + s]$, which we approximate as follows:

$$\begin{aligned}\mathcal{P}_{\mathcal{A},[t,t+s]} =& \frac{1}{s} \int_{t}^{t+s} \mathcal{P}_{\mathcal{A},x} \mathrm{d}x \\ \simeq& \frac{1}{K+1} \sum_{i=0}^{K} \mathcal{P}_{\mathcal{A},t+i\frac{s}{K}}\end{aligned}$$

The time unit considered in this study is the year and we used $K = 4$ so that one year is split up into four parts of equal lengths. Also, we fixed the value of λ to 1.0 so that the probability of a lineage to survive for a period of one year is 0.37. This value of λ derived from the observation that the length of external branches in WNV phylogenies are generally close to 1.0. Assuming a critical birth-death model approximates the branching process here, Theorem 3 in [14] states that the rate of death of lineage is given by the inverse of the average length of an external edge. Note that, as we can derive the joint distribution of the velocity and position of each tip conditionally on all other tips thanks to the pruning algorithm described in Section 5.3, we could, assuming that the velocity is constant, get the distribution of the particle starting from this tip after a time t. Probabilities $\mathcal{P}_{\mathcal{A},x}$ would then be obtained by integrating this distribution over on domain $\mathcal{A}$. Such an approach could be computationally expensive, and would need to be carefully examined for possible use in future work.

5.7. Cross-validation of tip coordinates

In an attempt to compare the fit of the PIV and the RRW models to the WNV data sets, we assessed the ability of these two models to recover coordinates at tips where only sequence data is made available. A MCMC analysis was first performed on the full data set. Then, for each tip taken in a sequential manner, coordinates were hidden and considered as parameters of the model.

The posterior distribution of the standard model parameters (including the tree topology, age of internal nodes plus the dispersal parameters of the model considered) along with that of the missing tip location, were obtained and the posterior distribution of the great circle distance between the true and estimated tip locations was recorded.

Let x_i^* be the tip coordinates at tip i and $\mathbf{x}_{-i}^*$ the set of coordinates observed at all tips except i. Also, let $\mathbf{s}^*$ be the set of observed sequences at the tips. θ is a generic parameter that encompasses all the parameters of the model excluding the tip and ancestral velocities, noted as $\mathbf{y}^*$ and $\mathbf{y}$ respectively. Our objective is to draw samples from the distribution of $x_i^*, \mathbf{y}^*, \mathbf{y}, \theta \mid \mathbf{x}_{-i}^*, \mathbf{s}^*$. Given samples from this joint distribution, one marginalizes over $\mathbf{y}^*, \mathbf{y}$ and θ in order to recover the posterior distribution of interest. We have:

$$\begin{aligned} p(x_i^*, \mathbf{y}^*, \mathbf{y}, \theta \mid \mathbf{x}_{-i}^*, \mathbf{s}^*) &\propto p(x_i^*, \mathbf{x}_{-i}^*, \mathbf{s}^*, \mathbf{y}^*, \mathbf{y}, \theta) \\ &\propto p(\mathbf{s}^* \mid \theta) p(\mathbf{x}^*, \mathbf{y} \mid \theta) \end{aligned}$$

so that samples from the target distribution can be obtained by standard MCMC with x_i considered as a parameter of the model along $\mathbf{y}^*, \mathbf{y}$ and θ. As explained above, we ran a first MCMC on the full data set so as to reach the stationary distribution of the Markov chain that correlated generates samples from $p(\mathbf{y}^*, \mathbf{y}, \theta \mid \mathbf{x}^*, \mathbf{s}^*)$. Each tip coordinates is then hidden sequentially. For each hidden tip coordinates x_i^*, a shorter MCMC analysis is ran with $p(x_i^*, \mathbf{y}^*, \mathbf{y}, \theta \mid \mathbf{x}_{-i}^*, \mathbf{s}^*)$ as its target distribution.

Note that, thanks to the pruning algorithm described in Section 5.3, one could directly get the distribution of left out tips conditionally on observed ones, to carry out a cross validation relying on the expected log predictive density similar to [11].

References

[1] D. Ayres, A. Darling, D. Zwickl, P. Beerli, M. Holder, P. Lewis, J. Huelsenbeck, F. Ronquist, D. Swofford, M. Cummings, et al. BEAGLE: an application programming interface and high-performance computing library for statistical phylogenetics. *Systematic Biology*, 61:170–173, 2012.

[2] P. Bastide, Lam Si T. Ho, G. Baele, P. Lemey, and M. Suchard. Efficient Bayesian Inference of General Gaussian Models on Large Phylogenetic Trees. *The Annals of Applied Statistics*, 15:971–997, 2021. doi: 10.1214/20-aoas1419.

[3] J. Clavel, G. Escarguel, and G. Merceron. mvmorph : an R package for fitting multivariate evolutionary models to morphometric data. *Methods in Ecology and Evolution*, 6:1311–1319, 2015.

[4] W. Cumberland and C. Rohde. A multivariate model for growth of populations. *Theoretical Population Biology*, 11:127–139, 1977.

[5] A. Drummond, S. Ho, M. Phillips, and A. Rambaut. Relaxed phylogenetics and dating with confidence. *PLOS Biology*, 4:e88, 2006.

[6] Paradis E. and Schliep K. ape 5.0: an environment for modern phylogenetics and evolutionary analyses in R. *Bioinformatics*, 35:526–528, 2019. doi: 10.1093/bioinformatics/bty633.

[7] J. Felsenstein. Maximum-likelihood estimation of evolutionary trees from continuous characters. *The American Journal of Human Genetics*, 25:471–492, 1973.

[8] Wickham H. *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York, 2016. ISBN 978-3-319-24277-4. URL `https://ggplot2.tidyverse.org`.

[9] Zhu H. *kableExtra: Construct Complex Table with 'kable' and Pipe Syntax*, 2024. URL `https://CRAN.R-project.org/package=kableExtra`. R package version 1.4.0.

[10] M. Hasegawa, H. Kishino, and T. Yano. Dating of the Human-Ape splitting by a molecular clock of mitochondrial-DNA. *Journal of Molecular Evolution*, 22:160–174, 1985.

[11] G. Hassler, B. Gallone, L. Aristide, W. Allen, M. Tolkoff, A. Holbrook, G. Baele, P. Lemey, and M. Suchard. Principled, practical, flexible, fast: A new approach to phylogenetic factor analysis. *Methods in Ecology and Evolution*, 13:2181–2197, 2022.

[12] Müller K. *here: A simpler way to find your files*, 2020. URL `https://CRAN.R-project.org/package=here`. R package version 1.0.1.

[13] V. Mitov, K. Bartoszek, G. Asimomitis, and T. Stadler. Fast likelihood calculation for multivariate Gaussian phylogenetic models with shifts. *Theoretical Population Biology*, 131:66–78, 2020.

[14] A. Mooers, O. Gascuel, T. Stadler, H. Li, and M. Steel. Branch lengths on birth–death trees and the expected loss of phylogenetic diversity. *Systematic Biology*, 61:195–203, 2012.

[15] R. Neal. MCMC using Hamiltonian dynamics. In S. Brooks, Andrew Gelman, G. L. Jones, and X. L. Meng, editors, *Handbook of Markov Chain Monte Carlo*, pages 113–162. CRC Press, New York, NY, 2011.

[16] L. Papież and G. Sandison. A diffusion model with loss of particles. *Advances in Applied Probability*, 22:533–547, 1990.

[17] O. Pybus, M. Suchard, P. Lemey, F. Bernardin, A. Rambaut, F. Crawford, R. Gray, N. Arinaminpathy, S. Stramer, M. Busch, et al. Unifying the spatial epidemiology and molecular evolution of emerging epidemics. *Proceedings of the National Academy of Sciences*, 109:15066–15071, 2012.

[18] Bilderbeek R. and Etienne R. babette: BEAUti 2, BEAST 2 and Tracer for R. *Methods in Ecology and Evolution*, 2008. URL `https://doi.org/10.1111/2041-210X.13032`.

[19] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2024. URL `https://www.R-project.org/`.

[20] M. Suchard, P. Lemey, G. Baele, D. L Ayres, A. Drummond, and A. Rambaut. Bayesian phylogenetic and phylodynamic data integration using BEAST 1.10. *Virus Evolution*, 4:vey016, 2018.

[21] Li-Gen Wang, Tommy Tsan-Yuk Lam, Shuangbin Xu, Zehan Dai, Lang Zhou, Tingze Feng, Pingfan Guo, Casey W. Dunn, Bradley R. Jones, Tyler Bradley, Huachen Zhu, Yi Guan, Yong Jiang, and Guangchuang Yu. Treeio: An R Package for Phylogenetic Tree Input and Output with Richly Annotated and Associated Data. *Molecular Biology and Evolution*, 37(2):599–603, October 2019. ISSN 0737-4038. doi: 10.1093/molbev/msz240.

[22] Claus O. Wilke. *cowplot: Streamlined Plot Theme and Plot Annotations for 'ggplot2'*, 2024. URL `https://wilkelab.org/cowplot/`. R package version 1.1.3.