A Distribution of coordinates for the IOU process conditioned on the initial velocity

The distribution is normal with expectation and variance given below:

$$\mathrm{E}(X(t) \mid x(0), y(0)) = x(0) + (y(0) - \mu)\left(\frac{1 - e^{-\theta t}}{\theta}\right) + \mu t \tag{10}$$

$$\mathrm{V}(X(t) \mid x(0), y(0)) = \frac{\sigma^2}{\theta^3}\left(t\theta - \frac{3}{2} - \frac{e^{-2t\theta}}{2} + 2e^{-t\theta}\right) \tag{11}$$

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.

B Distributions of velocity and coordinates for PIV models conditioned on the initial and final 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 Section "Likelihood calculation and Bayesian inference" in the "Material and Methods" of the main text).

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:

$$Y(s) = \frac{s}{t}y(t) + \left(\frac{t-s}{t}\right)y(0) + W(s) - \frac{s}{t}W(t) \tag{12}$$

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:

$$\mathrm{E}(Y(s) \mid y(0), y(t)) = \frac{s}{t}(y(t) - y(0)) + y(0) \tag{13}$$

$$\mathrm{Cov}(Y(u), Y(v)) = \sigma^2 \frac{u(t-v)}{t} \tag{14}$$

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

When the velocity follows a OU bridge, Lemma 1 in (Papież and Sandison, 1990) shows that:

$$Y(s) = \frac{\sinh(\theta s)}{\sinh(\theta t)}y(t) + \frac{\sinh(\theta(t-s))}{\sinh(\theta t)}y(0) + W(s) - \frac{\sinh(\theta s)}{\sinh(\theta t)}W(t) \tag{15}$$

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

$$\mathrm{E}(Y(s) \mid y(0), y(t)) = \frac{\sinh(\theta s)}{\sinh(\theta t)}y(t) + \frac{\sinh(\theta(t-s))}{\sinh(\theta t)}y(0) + \mu\left(1 - e^{-\theta s}\right) - \frac{\sinh(\theta s)}{\sinh(\theta t)}\mu\left(1 - e^{-\theta t}\right) \tag{16}$$

$$\mathrm{Cov}(Y(u), Y(v)) = \frac{\sigma^2}{\theta}\frac{\sinh(\theta u)\sinh(\theta(t-v))}{\sinh(\theta t)} \tag{17}$$

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 $\int_0^t \mathrm{E}(Y(s) \mid y(0), y(t))\mathrm{d}s$ and $\int_0^t \int_0^t \mathrm{Cov}(Y(u), Y(v))\mathrm{d}u\mathrm{d}v$ and obtain the following expressions:

$$\mathrm{E}(X(t) \mid x(0), y(0), y(t)) = x(0) + \frac{t}{2}(y(t) + y(0)) \tag{18}$$

$$\mathrm{V}(X(t) \mid x(0), y(0), y(t)) = \frac{\sigma^2 t^3}{12}. \tag{19}$$

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:

$$\mathrm{E}(X(t) \mid x(0), y(0), y(t)) = x(0) + \left(\frac{\cosh(\theta t) - 1}{\theta \sinh(\theta t)}\right)(y(t) + y(0)) + \mu t - \frac{\mu}{\theta}\left(1 - e^{-\theta t}\right)\left(1 + \frac{\cosh(\theta t) - 1}{\sinh(\theta t)}\right) \tag{20}$$

$$\mathrm{V}(X(t) \mid x(0), y(0), y(t)) = \frac{\sigma^2}{\theta^3}\left(\theta t - 2\left(\frac{\cosh(\theta t) - 1}{\sinh(\theta t)}\right)\right) \tag{21}$$

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 thus 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.

C The pruning approach for the IBM and IOU processes

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 $\mathrm{pa}(i)$ (Mitov *et al.*, 2020; Bastide *et al.*, 2021):

$$\mathbf{Z}_i \mid \mathbf{Z}_{\mathrm{pa}(i)} \sim \mathcal{N}\left(\mathbf{q}_i \mathbf{Z}_{\mathrm{pa}(i)} + \mathbf{r}_i, \mathbf{\Sigma}_i\right), \tag{22}$$

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:

$$\mathbf{q}_i = \begin{pmatrix} \mathbf{I}_p & \mathbf{0}_{p,p} \\ \mathbf{I}_p t_i & \mathbf{I}_p \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ t_i & 1 \end{pmatrix} \otimes \mathbf{I}_p \tag{23}$$

$$\mathbf{r}_i = \begin{pmatrix} \boldsymbol{\delta} t_i \\ \boldsymbol{\delta} t_i^2/2 \end{pmatrix} = \begin{pmatrix} t_i \\ t_i^2/2 \end{pmatrix} \otimes \boldsymbol{\delta} \tag{24}$$

$$\mathbf{\Sigma}_i = \begin{pmatrix} \mathbf{\Sigma} t_i & \mathbf{\Sigma} t_i^2/2 \\ \mathbf{\Sigma} t_i^2/2 & \mathbf{\Sigma} t_i^3/3 \end{pmatrix} = \begin{pmatrix} t_i & t_i^2/2 \\ t_i^2/2 & t_i^3/3 \end{pmatrix} \otimes \mathbf{\Sigma}, \tag{25}$$

where t_i is the length of the branch going from pa(i) to i, and $\otimes$ denotes the Kronecker product. For the IOU, assuming that the velocity vector follows a OU with actualization matrix $\boldsymbol{\Theta}$, central vector $\boldsymbol{\mu}$ and variance $\boldsymbol{\Sigma}$, we get (Cumberland and Rohde, 1977):

$$\mathbf{q}_i = \begin{pmatrix} e^{-\boldsymbol{\Theta} t_i} & \mathbf{0}_{p,p} \\ \boldsymbol{\Theta}^{-1}(\mathbf{I}_p - e^{-\boldsymbol{\Theta} t_i}) & \mathbf{I}_p \end{pmatrix} \tag{26}$$

$$\mathbf{r}_i = \begin{pmatrix} (\mathbf{I}_p - e^{-\boldsymbol{\Theta} t_i})\boldsymbol{\mu} \\ \boldsymbol{\mu} t_i - \boldsymbol{\Theta}^{-1}(\mathbf{I}_p - e^{-\boldsymbol{\Theta} t_i})\boldsymbol{\mu} \end{pmatrix} \tag{27}$$

$$\boldsymbol{\Sigma}_i = \begin{pmatrix} \mathbf{K}_{\mathbf{Y}}(t_i) & \mathbf{K}_{\mathbf{X},\mathbf{Y}}(t_i)^T \\ \mathbf{K}_{\mathbf{X},\mathbf{Y}}(t_i) & \mathbf{K}_{\mathbf{X}}(t_i) \end{pmatrix}, \tag{28}$$

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:

$$\mathbf{K}_{\mathbf{Y}}(t) = \mathbf{S} - e^{-\boldsymbol{\Theta} t}\mathbf{S}e^{-\boldsymbol{\Theta}^T t} \tag{29}$$

$$\mathbf{K}_{\mathbf{X}}(t) = \boldsymbol{\Delta} t - (\mathbf{I}_p - e^{-\boldsymbol{\Theta} t})\boldsymbol{\Theta}^{-1}\boldsymbol{\Delta} - \boldsymbol{\Delta}\boldsymbol{\Theta}^{-T}(\mathbf{I}_p - e^{-\boldsymbol{\Theta}^T t}) + \boldsymbol{\Theta}^{-1}\mathbf{S}\boldsymbol{\Theta}^{-T} - e^{-\boldsymbol{\Theta} t}\boldsymbol{\Theta}^{-1}\mathbf{S}\boldsymbol{\Theta}^{-T}e^{-\boldsymbol{\Theta}^T t} \tag{30}$$

$$\mathbf{K}_{\mathbf{X},\mathbf{Y}}(t) = \mathbf{S}\boldsymbol{\Theta}^{-1}[\mathbf{I}_p - e^{-\boldsymbol{\Theta} t}] - [\mathbf{I}_p - e^{-\boldsymbol{\Theta} t}]\boldsymbol{\Theta}^{-1}\mathbf{S}e^{-\boldsymbol{\Theta}^T t} \tag{31}$$

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

Pruning algorithm

As showed in (Mitov *et al.*, 2020; Bastide *et al.*, 2021), for a general linear Gaussian process of the form of Eq. 22 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 (Bastide *et al.*, 2021), 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 (Suchard *et al.*, 2018), and allows for the Bayesian fit of the model without resorting to stochastic integration of the internal node velocities as is done in PhyREX. Note that, although we used a standard Metropolis-Hastings update for the parameters of the processes, from (Bastide *et al.*, 2021) we could also get derivative with respect to the parameters of the IBM or IOU, allowing for efficient Hamiltonian Monte Carlo sampling schemes (Neal, 2011).

D 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 $\boldsymbol{\Sigma}$ and root position parameter $\mathbf{X}_\rho \sim \mathcal{N}(\boldsymbol{\mu}^{\mathbf{X}}, \boldsymbol{\Gamma}^{\mathbf{X}})$, 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:

$$\mathrm{E}[\mathbf{X}_{\mathrm{t}}^i] = \boldsymbol{\mu}^{\mathbf{X}} \quad \text{and} \quad \mathrm{V}[\mathbf{X}_{\mathrm{t}}^i; \mathbf{X}_{\mathrm{t}}^j] = \boldsymbol{\Sigma}\tau_{ij} + \boldsymbol{\Gamma}^{\mathbf{X}} \tag{32}$$

with τ_{ij} the time between the root and the most recent common ancestor of i and j (see e.g. (Felsenstein, 1973)).

Marginal distribution under an IBM

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

$$\mathrm{E}[\mathbf{Y}_{\mathrm{t}}^i] = \boldsymbol{\mu}^{\mathbf{Y}} \quad \text{and} \quad \mathrm{V}[\mathbf{Y}_{\mathrm{t}}^i; \mathbf{Y}_{\mathrm{t}}^j] = \boldsymbol{\Sigma}\tau_{ij} + \boldsymbol{\Gamma}^{\mathbf{X}}. \tag{33}$$

For the positions, we get:

$$\mathrm{E}[\mathbf{X}_{\mathrm{t}}^i] = \boldsymbol{\mu}^{\mathbf{Y}}\tau_i + \boldsymbol{\mu}^{\mathbf{X}} \tag{34}$$

$$\mathrm{V}[\mathbf{X}_{\mathrm{t}}^i; \mathbf{X}_{\mathrm{t}}^j] = \boldsymbol{\Gamma}^{\mathbf{X}} + \boldsymbol{\Gamma}^{\mathbf{Y}}\tau_i\tau_j + \boldsymbol{\Sigma}\tau_{ij}\left[\tau_i\tau_j + \tau_{ij}\left(\frac{\tau_{ij}}{3} - \frac{\tau_i + \tau_j}{2}\right)\right]. \tag{35}$$

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:

$$\mathrm{V}[\mathbf{X}_{\mathrm{t}}^i; \mathbf{Y}_{\mathrm{t}}^j] = \boldsymbol{\Gamma}^{\mathbf{Y}}\tau_i + \boldsymbol{\Sigma}\tau_{ij}\left[\tau_i - \frac{\tau_{ij}}{2}\right]. \tag{36}$$

The proof of these formulas rely on the following equality:

$$\mathrm{V}[\mathbf{X}_{\mathrm{t}}^i; \mathbf{X}_{\mathrm{t}}^j] = \int_0^{\tau_i}\int_0^{\tau_j} \mathrm{V}[\mathbf{Y}^i(t); \mathbf{Y}^j(s)]dtds \tag{37}$$

$$\mathrm{V}[\mathbf{X}_{\mathrm{t}}^i; \mathbf{Y}_{\mathrm{t}}^j] = \int_0^{\tau_i} \mathrm{V}[\mathbf{Y}^i(t); \mathbf{Y}_{\mathrm{t}}^j]dt, \tag{38}$$

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:

$$\mathrm{V}[\mathbf{Y}^i(t); \mathbf{Y}^j(s)] = \begin{cases} \boldsymbol{\Sigma}\min(s,t) + \boldsymbol{\Gamma}^{\mathbf{Y}} & \text{if } s \leq \tau_{ij} \text{ or } t \leq \tau_{ij} \\ \boldsymbol{\Sigma}\tau_{ij} + \boldsymbol{\Gamma}^{\mathbf{Y}} & \text{otherwise.} \end{cases} \tag{39}$$

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 $\boldsymbol{\Theta}$, central 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. (Clavel *et al.*, 2015)):

$$\mathrm{E}[\mathbf{Y}_{\mathrm{t}}^i] = e^{-\boldsymbol{\Theta}t}\boldsymbol{\mu}^{\mathbf{Y}} + (\mathbf{I}_p - e^{-\boldsymbol{\Theta}t})\boldsymbol{\mu} \quad \text{and} \tag{40}$$

$$\mathrm{V}[\mathbf{Y}_{\mathrm{t}}^i; \mathbf{Y}_{\mathrm{t}}^j] = e^{-\boldsymbol{\Theta}(\tau_i - \tau_{ij})}\mathbf{S}e^{-\boldsymbol{\Theta}^T(\tau_j - \tau_{ij})} - e^{-\boldsymbol{\Theta}\tau_i}\mathbf{S}e^{-\boldsymbol{\Theta}^T\tau_i} + e^{-\boldsymbol{\Theta}\tau_i}\boldsymbol{\Gamma}^{\mathbf{Y}}e^{-\boldsymbol{\Theta}^T\tau_i}. \tag{41}$$

For the positions, we get:

$$\mathrm{E}[\mathbf{X}_{\mathrm{t}}^i] = \boldsymbol{\mu}^{\mathbf{X}} + \boldsymbol{\Theta}^{-1}(\mathbf{I}_p - e^{-\boldsymbol{\Theta}t})(\boldsymbol{\mu}^{\mathbf{Y}} - \boldsymbol{\mu}) + \boldsymbol{\mu}t \quad \text{and} \tag{42}$$

$$\begin{aligned}\mathrm{V}[\mathbf{Y}_{\mathrm{t}}^i; \mathbf{Y}_{\mathrm{t}}^j] = \boldsymbol{\Gamma}^{\mathbf{X}} &+ [\mathbf{I}_p - e^{-\boldsymbol{\Theta}\tau_i}]\boldsymbol{\Theta}^{-1}\boldsymbol{\Gamma}^{\mathbf{Y}}\boldsymbol{\Theta}^{-T}[\mathbf{I}_p - e^{-\boldsymbol{\Theta}^T\tau_j}] + \boldsymbol{\Delta}\tau_{ij} + (\mathbf{I}_p - e^{\boldsymbol{\Theta}\tau_{ij}})e^{-\boldsymbol{\Theta}\tau_i}\boldsymbol{\Theta}^{-1}\boldsymbol{\Delta} \\ &+ \boldsymbol{\Delta}\boldsymbol{\Theta}^{-T}e^{-\boldsymbol{\Theta}^T\tau_j}(\mathbf{I}_p - e^{\boldsymbol{\Theta}^T\tau_{ij}})e^{-\boldsymbol{\Theta}\tau_i}\left[e^{\boldsymbol{\Theta}\tau_{ij}}\boldsymbol{\Theta}^{-1}\mathbf{S}\boldsymbol{\Theta}^{-T}e^{\boldsymbol{\Theta}^T\tau_{ij}} - \boldsymbol{\Theta}^{-1}\mathbf{S}\boldsymbol{\Theta}^{-T}\right]e^{-\boldsymbol{\Theta}^T\tau_j}\end{aligned} \tag{43}$$

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

The proof of the formulas rely on the integration of the following covariance function (see (Cumberland and Rohde, 1977)), defined, for any $0 \leq s \leq t$, as:

$$\mathrm{V}[\mathbf{Y}^i(t); \mathbf{Y}^j(s)] = \mathbf{S}e^{-\boldsymbol{\Theta}(t-s)} - e^{-\boldsymbol{\Theta}s}\mathbf{S}e^{-\boldsymbol{\Theta}^Tt} + e^{-\boldsymbol{\Theta}s}\boldsymbol{\Gamma}^{\mathbf{Y}}e^{-\boldsymbol{\Theta}^Tt}, \tag{44}$$

if $s \leq \tau_{ij}$ or $t \leq \tau_{ij}$, and:

$$\mathrm{V}[\mathbf{Y}^i(t); \mathbf{Y}^j(s)] = e^{-\boldsymbol{\Theta}(s-\tau_{ij})}\mathbf{S}e^{-\boldsymbol{\Theta}^T(t-\tau_{ij})} - e^{-\boldsymbol{\Theta}s}\mathbf{S}e^{-\boldsymbol{\Theta}^Tt} + e^{-\boldsymbol{\Theta}s}\boldsymbol{\Gamma}^{\mathbf{Y}}e^{-\boldsymbol{\Theta}^Tt} \tag{45}$$

otherwise.

	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 S1: Comparison of PhyREX and BEAST estimations of variance and root parameters for a bi-variate IBM on a fixed tree (WNV data of (Pybus *et al.*, 2012) with 104 tips).

Likelihood computation checks

The formulas above give the full distribution of the observed positions at the tips of the tree, that is matrix normal. They are therefore 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 C is to be preferred, as it is linear in the number of tips (see e.g. (Mitov *et al.*, 2020; Bastide *et al.*, 2021)). 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` (R Core Team, 2024), and checked on small examples that the two software gave matching likelihoods.

E 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.*, 2012) (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. Since the analyses are computationally intensive, we 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 (Hasegawa *et al.*, 1985b), and an uncorrelated relaxed random local clock (Drummond *et al.*, 2006) 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.

We used `R` (R Core Team, 2024) to run analyses, using packages `ape` (Paradis and Schliep, 2019) and `treeio` (Wang *et al.*, 2019) for tree manipulation, `tracerer` (Bilderbeek and Etienne, 2018) for reading and summarizing log files, `ggplot2` (Wickham, 2016) and `cowplot` (Wilke, 2024) for plotting results, `here` (Müller, 2020) for file manipulation, and `kableExtra` (Zhu, 2024) 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 S1). Unsurprisingly, as BEAST only needs to sample two parameters

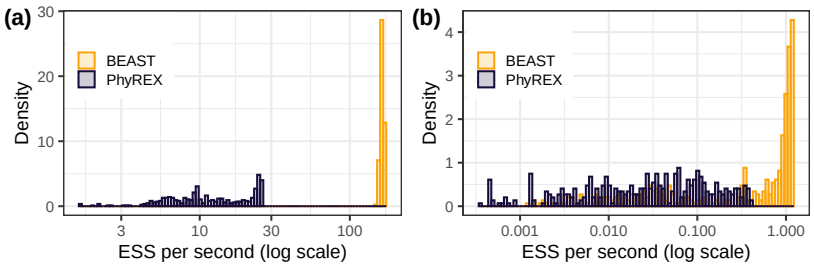

Figure S1: Effective sample size per seconds of all node velocities (internal and tips) of BEAST versus PhyREX for a bi-variate IBM on a fixed tree (a) or with an inferred tree (b). WNV data of (Pybus *et al.*, 2012) with 104 tips, i.e. 208 velocity estimates per software program.

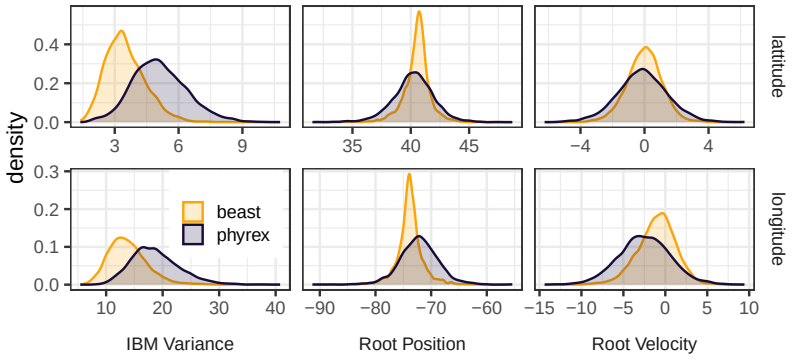

Figure S2: Comparison of PhyREX and BEAST estimations of variance and root parameters for a bi-variate IBM when the tree is inferred using WNV data of (Pybus *et al.*, 2012) with 104 tips.

(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 S1). 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. S1-a). Note that each PhyREX iteration was about 155.0 times faster than each BEAST iteration, as it requires less computations.

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. S1-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 (Ayres *et al.*, 2012) 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. S2). Reconstructed velocities at tips were highly similar in both approaches (not shown).

F 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:

$$\mathcal{P}_{\mathcal{A},t} = 1 - \prod_{i=1}^{n} \Pr(\alpha_i(t) \notin \mathcal{A}) \tag{46}$$

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:

$$\mathcal{P}_{\mathcal{A},[t,t+s]} = \frac{1}{s} \int_t^{t+s} \mathcal{P}_{\mathcal{A},x} \mathrm{d}x \tag{47}$$

$$\simeq \frac{1}{K+1} \sum_{i=0}^{K} \mathcal{P}_{\mathcal{A},t+i\frac{s}{K}} \tag{48}$$

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 (Mooers *et al.*, 2012) states that the rate of death of lineage is given by the inverse of the average length of an external edge. Note that PIV models allow us to 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 C. We could therefore derive the distribution of the particle starting from this tip after a time t, assuming that the velocity is constant. 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.

G 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:

$$p(x_i^*, \mathbf{y}^*, \mathbf{y}, \theta \mid \mathbf{x}_{-i}^*, \mathbf{s}^*) \propto p(x_i^*, \mathbf{x}_{-i}^*, \mathbf{s}^*, \mathbf{y}^*, \mathbf{y}, \theta) \tag{49}$$

$$\propto p(\mathbf{s}^* \mid \theta) p(\mathbf{x}^*, \mathbf{y} \mid \theta) \tag{50}$$

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

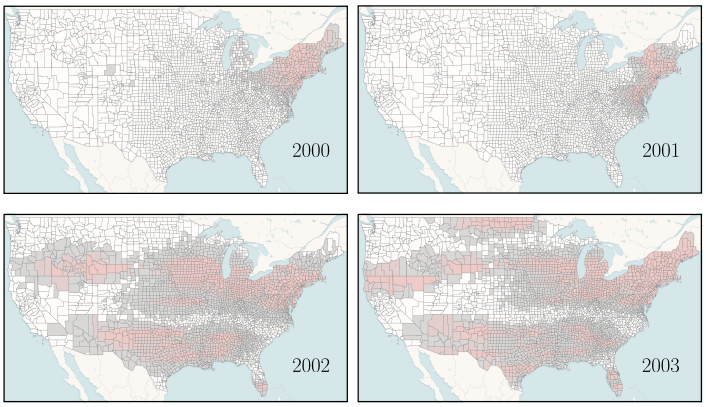

Figure S3: Predicted occurrence of WNV in the early phase of the epidemic (model for prediction: IBM). Here, the prior distribution for the velocity at the root had a variance set to $(10^{-2}, 10^{-2})$ (mean set to $(0, 0)$) as opposed to a virtually flat prior (variance set to $(10^2, 10^2)$) by default (see Fig. 4 of the main text).

full data set so as to reach the stationary distribution of the Markov chain that generates correlated 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 C, 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 (Hassler *et al.*, 2022).

H Prediction of WNV incidence deriving from alternative models

By default, predictions were performed using the IBM model with a flat prior on each of the two parameters making up the variance of the velocity vector. A normal distribution centered on $(0, 0)$ and variance $(10^2, 10^2)$ was used here. Figure S3 gives the predicted occurrence of WNV in the early stages of the epidemic using a more informative prior with variance vector $(10^{-2}, 10^{-2})$. While the predictions for years 2001-2003 are similar to that obtained with a flat prior, the occupied area for year 2000 is smaller when using the informative prior compared to that obtained with a non-informative one. The sensitivity to priors observed here is a likely consequence of the lack of signal conveyed by the limited amount of data (only seven sequences with coordinates are available for this time point).

We also performed prediction analyses using the RRW model. Velocity at each tip was estimated during the MCMC analysis as follows: (1) ancestral location were sampled from their joint posterior density; (2) great-circle distances between each tip location and that sampled for its direct ancestor were evaluated; (3) the obtained distance was divided by the time elapsed along the corresponding (external) edge. Predictions were then made using the same approach as that used with the IBM model (see SI, section F). Figures S4 and S5 show the incidence from the CDC data (see main text) and the corresponding predictions using the RRW model. We note that the predictions of the RRW are much more spread out than the ones of the IBM (see main text and Figure S3), which is consistent with the fact that the RRW can allow for jumps in the process, i.e. large dispersion events in small time scales, so that, according to this model, a spread far away from the origin is not unlikely, even in the early phases of the epidemics.

We compared the IBM and RRW models by evaluating their sensitivity (true positive rate) and specificity (true negative rate) corresponding to the predicted occurrence of the virus in each county. More specifically, for each year in the 2000-2007 time period, a county was said to be predicted as "infected" as soon as at least one lineage was predicted in this county. This county was then labeled as a "true positive" if the corresponding incidence from the CDC data (see main text) showed at least one case. Figure S6 shows that the RRW has a greater sensitivity, but a lower sensitivity. This is consistent with the spatial patterns observed on Fig. S5 and S4, that showed a wider dispersion for the RRW.

Using the empirical cumulative distribution function, we transformed the predicted count data of

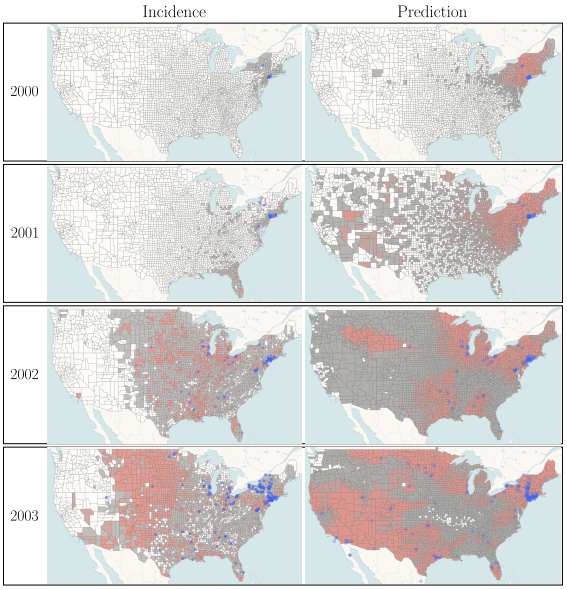

Figure S4: Incidence and predicted occurrence of WNV in the early phase of the epidemic (model for prediction: RRW). Purple dots correspond to sampled locations. Incidence data (left) for each year and each county was obtained from the CDC. For year Y, predicted occurrence of the WNV (right) was inferred using data collected earlier than the end of December of year $Y-1$. The maps were generated with EvoLaps2 (Chevenet *et al.*, 2024)

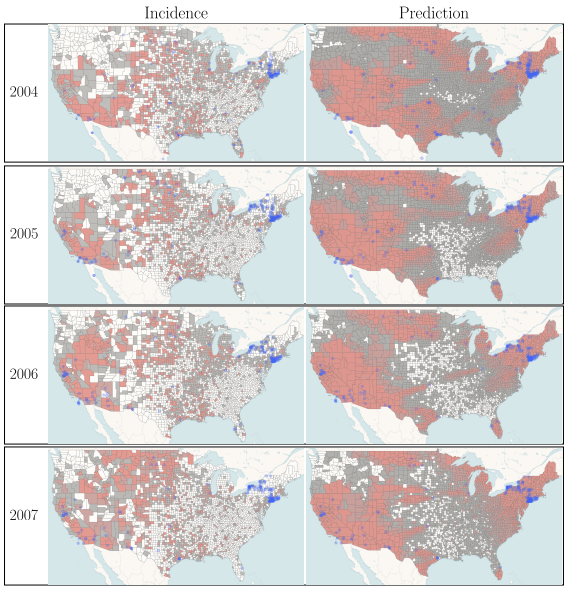

Figure S5: Incidence and predicted occurrence of WNV in an endemic regime (model for prediction: RRW). See caption of Figure S4.

both methods into probabilities. These probabilities were then used to predict the occurrence of the virus in each county. This allowed us to compute Receiver Operating Characteristic (ROC) curve for both predictor for each year, using the R package `ROCR` (Sing *et al.*, 2005). Results in Fig. S7 show that the IBM predictor is generally further from the diagonal than the RRW predictor, indicating a better overall performance of the IBM. We also note that the predictor are more accurate for the early stages of the epidemics, when the spatial distribution of the virus is limited.

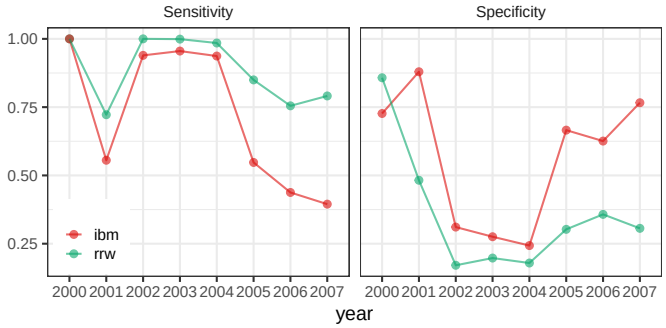

Figure S6: Sensitivity and specificity of the predicted county level occurrences by the IBM and RRW models.

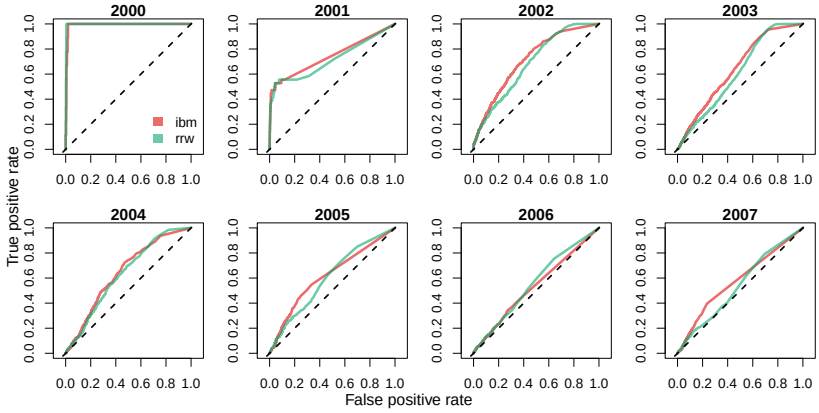

Figure S7: ROC curves for the predicted county level occurrences by the IBM and RRW models.

I Tip velocity estimation under the IBM model

We performed a simulation study using the IBM both for simulation and inference in order to check that our model and implementation produced correct results when presented with datasets that matched its assumptions.

We took a fixed and dated WNV tree inferred from previous analyses, and simulated 100 datasets using an IBM with independent movements on the latitude and longitude axes, with variances respectively set to 0.1 and 1. The root location was set to the New-York region (latitude 40.65°, longitude -74.33°) with root velocity vector (-0.24, -2.48) degrees per year.

We then used the BEAST implementation of the IBM, using the true fixed tree, but inferring all other parameters from the data. For each dataset, we ran an MCMC chain for 50 000 iterations, log every 100, with a standard log-transformed random walk operator on the variance parameters, and vague half-t priors. We then extracted estimates and 95% highest posterior density intervals (HPDI) for variance parameters, root position and velocity, and all tip velocity vectors.

Fig. S8 shows that the true variance parameters and root position and velocity are correctly recovered, with unbiased estimates, as expected. Further, Fig. S9 shows that the 104 tip velocity vectors are also correctly estimated, with unbiased estimates, and HPDI reaching coverages close to their nominal values.

In this simple setting with a fixed tree and an IBM used for both simulation and inference, this experiment shows that our implementation can recover the correct velocity dynamic of the epidemic over the different regions of the tree. Further investigations, using other velocity-explicit simulation models, could be the focus of future work which goal would be to asses the robustness of PIV models and its ability to recover specific propagation dynamics.

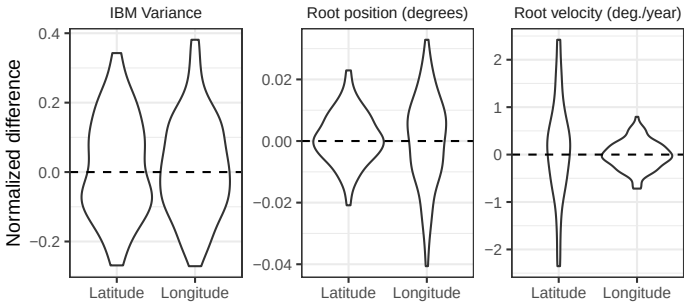

Figure S8: Variance and root parameter IBM estimate. Difference between the estimated and true value, normalized by the true value, for the IBM variance parameter (first panel) and the root position and velocity vectors (in degrees, second and last panels). Violin plot over 100 replicates. Data was simulated using an IBM on the WNV tree with 104 tips. BEAST was used for inference, using the true fixed tree, and an IBM model.

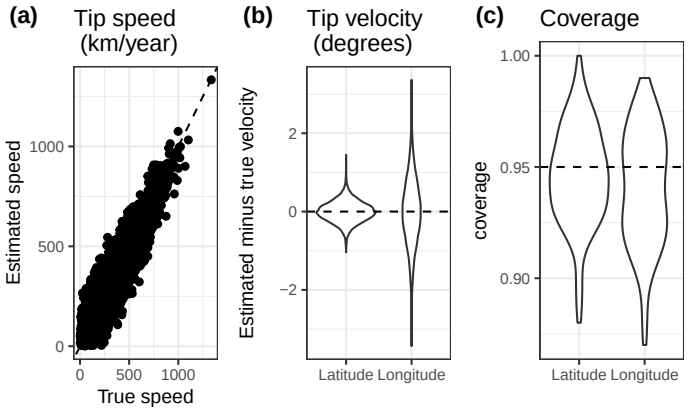

Figure S9: Tip speed and velocity estimates. **(a)** Estimated vs. true speed (in km/year) for the 104 tips and the 100 replicates. **(b)** Estimated minus true tip velocity vector (in degrees). **(c)** Realized coverage of the 95% highest posterior density intervals for the tip velocity vectors. Violin plot over 100 replicates and 104 tips. Data was simulated using an IBM on the WNV tree with 104 tips. BEAST was used for inference, using the true fixed tree, and an IBM model.