Dear Recommender,

Following your appreciation, we have revised our manuscript, entitled: ‘’A mechanistic-statistical approach to infer dispersal and demography from invasion dynamics, applied to a plant pathogen”.
These suggestions helped us to improve our manuscript greatly. A detailed response to the reviewers’ comments can be found below. Our replies are indicated in green. We provide you with two types of files highlighting revision marks or not in the main document and the Appendix. For the following responses to reviewers’ comments, lines refer to the version of the manuscript highlighting revision marks.
We hope that in its revised state, the manuscript will be suitable for recommendation in PCI Mathematical and Computational Biology.

Yours sincerely,

Méline Saubin

Dear Dr. Méline Saubin,

We received two reviews of your preprint entitled ' A mechanistic-statistical approach to infer dispersal and demography from invasion dynamics, applied to a plant pathogen'. Both appreciate the work and provide a few comments, which may be answered appropriately. I would also encourage the authors to add a set of figures which show the predicted values by the four models. By contrasting it with Figure 3 and describing the cause of insufficient fit of the other models, the manuscript may become even more persuasive for general readers.

We warmly thank you and the referees for all the constructive comments and suggestions on our manuscript. We replied point by point to all reviewers' concerns below.

Action: *We added in Appendix S4.4 two figures (Figures S6, S7) to supplement the results depicted in Figures 3 and 5 by presenting the predicted values and coverage rates for models J_{Exp} and J_{Gauss}. Unfortunately, due to a poor fit to the data, we were not able to draw this graphical representation for model R.D., which is now explained in Appendix S4.4.*

These new graphics and the associated coverage rates (0.69 and 0.67 for models J_{Exp} and J_{Gauss}, respectively, compared to 0.75 for the selected model J_{ExpP}) help to visualise the differences in dispersal between the models and our data, and the lover coverage rates for models J_{Exp} and J_{Gauss} compared to J_{ExpP}. This is especially true at the first sampling date when the epidemic intensity is underestimated upstream and overestimated downstream.
Action: *A sentence was added in section 4.3 (lines 363-366) to highlight this point.*

Sincerely yours,

Hirohisa Kishino

Comments by Reviewer 1

I declare that I have no conflict of interest with the authors or the content of the article

Anonymously

Review as text

This article proposes a mechanistic-statistical approach to model dispersal events. The article is original, well-written and interesting.

I have few comments to help to clarify few points:

You propose two models: RD is based on diffusion while ID is based on kernel. Can you discuss the pro and the con of both models (especially from an applied point of view)?
We discuss the advantages and disadvantages of R.D. and I.D. models (in terms of realism of the underlying assumptions, and the modelling of individual movements) in the introduction (lines…) and in the discussion (line…). From a more practical point of view, R.D. is known to be numerically easier to compute and faster to simulate than I.D. models, which are numerically more complex. Developping new I.D. models is thus a strength and originality of our approach.
***Action:** A sentence was added in the introduction to complement this point (lines 94-95).*

Your kernel is described for $\tau<1$, $\tau=1$ and $\tau=2$. What happens when $\tau>1$ (and not equal to 2)?
We expected similar dynamics among all dispersal kernels with $\tau > 1$, (kernels with constant speed of propagation). We have therefore chosen to explore in more details the kernels for which $\tau < 1$ and to take two other particular cases classically studied: the Exponential kernel ($\tau = 1$) and the Gaussian kernel ($\tau = 2$).

In your raw sampling, you consider a tree as a group of independent leaves. This assumption seems strong. Did I miss something?
Each tree is considered a group of independent leaves, but only regarding habitat suitability. This assumption can indeed seem strong but holds if the leaves observed on the same tree are sufficiently far from each other and represent a large variety of environmental conditions, and therefore habitat suitabilities (for example, leaves observed within a tree will not have the same light nor moisture level depending on their positions and sun exposition). Varying environmental conditions are expected to highly influence infection efficiency of poplar rust, hence habitat suitability. This was the case for our case study, where environmental conditions differed strongly across leaves from the same tree, as the observed leaves were selected to be far away from each other. This contrasts with leaves sampled on the same twig for the refined sampling: these leaves share local environmental conditions and therefore can be considered as having the same habitat suitability.
***Action:** The text was amended in section 2.3 (lines 188-192) to clarify this point.*

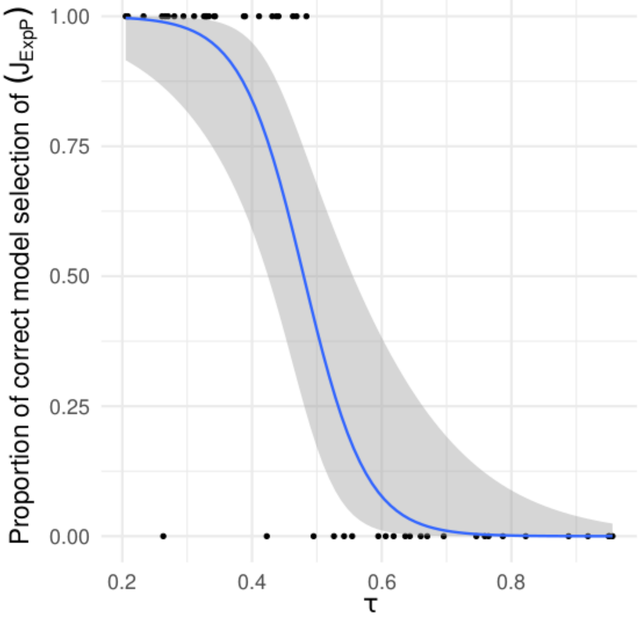
AIC is often conservative in terms of model selection. Did you try some alternative such as BIC?
We used the alternative criterion BIC for the model selection and we obtained similar results

for models JGauss and R.D. (Table R1). Our results were slightly different for JExp and JExP: BIC tends to attribute more cases to JExp and to discriminate less accurately JExp and JExpP, especially for high values of τ (Figure R1). However, we find similarly high values of dBIC$_{true}$ for correct JExpP selections. As for the AIC, BIC also selects for JExpP in our application, with a high value of dBIC$_{true}$, which comforts the model choice of our case study.

Table R1. Efficiency of model selection using Bayesian Information Criterion (BIC). The four first columns indicate the proportion of cases, among 50 replicates, where each tested model was selected using BIC, given that data sets were generated under a particular model (i.e. true model). Column dBICtrue (resp. dBICwrong) indicates the mean difference between the BIC of the model selected when the model selected is the true one (resp. when the model selected is not the true model) and the second best model (resp. being the true model or not).

	Selected Model					
True Model	J_{Exp}	J_{Gauss}	J_{ExpP}	R.D.	dBIC$_{true}$	dBIC$_{wrong}$
J_{Exp}	**0.68**	0.22	0.00	0.10	1.06	0.92
J_{Gauss}	0.34	**0.26**	0.00	0.40	1.08	0.88
J_{ExpP}	0.38	0.04	**0.52**	0.06	**114.17**	1.16
R.D.	0.18	0.24	0.00	**0.58**	0.71	0.30

Figure R1. Logistic regression of the proportion of correct model selection of dispersal JExpP ***based on the BIC****, as a function of τ. Dots represent the values of τ used for the 50 replicates of simulated dispersal model JExpP, and the estimated dispersal model (1 for a correct model selection of JExpP and 0 for a wrong model selection). The blue line corresponds to the predicted value of the proportion of correct model selection JExpP as a function of τ, and the grey area corresponds to the confidence envelope at 95%.*



Based on your simulations, I have the impression that your model is particularly efficient for fat-tail exponential power kernel and has a low power for the other cases. Fortunately, it corresponds to your application and the expected value of τ is low enough. It makes sense since you will not have sufficient data for long range dispersal for thin-tail kernel. Can you discuss this point?

Based on your remark, we checked whether using a higher resolution for solving the equations may facilitate the distinction between thin-tail kernels. However, we obtained the same results than those obtained with a lower resolution. The limits to distinguish between thin-tail kernel is not inherent to our modelling algorithm. However, the requirement for improving the capacity to distinguish between thin-tail kernels may lie in the sampling scheme. Here our sampling sites are regularly spaced, over a large sampling domain; which is indeed ideal to monitor long-distance dispersal. Designing another sampling scheme, with more frequent data in both time and space (or nested spatial sampling), might be the solution to more finely estimating the shape of the kernel at shorter distances. This would deserve a dedicated study.

Action: *Sentences were added in section 5.3 (lines 444-450) to discuss this point.*

Comments by Reviewer 2

I declare that I have no conflict of interest with the authors or the content of the article

Anonymously

Review as text

This is a well written paper, and the topic of the paper is clearly important. A strength of the paper is that the code and data are available, so that the reader can reproduce (I have not tried) or modify the analysis.

Comments and questions:

What are the assumptions about the random effects Ri(t)? Are they independent between two timepoints t1 and t2, but what if t1 and t2 are close?
The random effects Ri(t) only intervene in the observation processes. They are indeed independent between time points because we consider that trees are never observed twice in different samplings. Therefore, no matter how close the samplings are in time, Ri(t) are always independent.
***Action:** A sentence was added in section 2.3 (lines 185-187) to clarify this point.*

One weakness of the approach (seen from my perspective), is that the initial condition u0(x) needs to be specified. Multiple initial conditions were used to test sensitivity, and the results were found to be insensitive to the choice of u0. However, in other situations, the conclusion may be different, and hence making the approach less attractive.
The initial vector of population densities u(0,x) for x over [-R,R] was estimated from the data of the first sampling date, by fitting a general model for analysis of dose-response data (package Drc on R, Ritz et al., 2015). Therefore, this vector of initial infection is fixed among all simulations (See Appendix S4) and represents the starting point of all the simulated epidemics.
We referred to the initial conditions of the optimisation algorithm to talk about the initial vector θ of parameters to estimate, which can be confusing with regard to the initial conditions of infections represented by u(0,x).
***Action:** To avoid any confusion between θ and u(0,x), we corrected this vocabulary throughout the article, the appendices, and the Git and Zenodo repositories. We now refer to θ as the vector of initial parameter values, and u(0,x) as the vector of initial population densities, or initial conditions.*

Through our simulations, we tested for multiple initial parameter values (θ) but not for multiple initial population densities (u(0,x)). This represents indeed a weakness of our approach as the results may vary for different initial population densities. One solution may be to estimate u(0,x) along with the five other parameters in the vector θ, especially if the studied organism does not allow a simple estimation of u(0,x).
***Action:** Sentences were added in section 5.4 (lines 492-494) to discuss this point.*

The authors use a derivative free optimizer to maximize the likelihood function, but I wonder why they did not use automatic differentiation to calculate derivatives. This would have had

several benefits: 1) speed up, and make more numerically stable, the optimization process, 2) yield exact sensitivities wrt to u0, 3) allow inclusion of (parts of) u0 among the parameters that are estimated. A software package such as TMB (Kristensen et al., 2016) can calculate the first and second order derivatives of the log-likelihood. The requirement for this to work is that the likelihood is differentiable (not containing if-statements that depend on parameter values), and I have not verified if this is the case here.
In our case, preliminary tests revealed that classical optimisation algorithms were not accurate enough to provide satisfactory rates of convergence due to local optimum problems. Thus, we adopted an hybrid strategy combining first a Nelder-Mead algorithm (improving global search ability) and then a Nlminb algorithm (that converges quite fastly) and starting from 20 initial values. Although we did not formally test this, our understanding is that the Nealder-Mead algorithm (a derivative free method) combined with 20 initial values allows us to identify a candidate region for the optimum parameter's value. In our view, the TMB algorithm could have been interesting to replace Nlminb but not prevent us to use the hybrid approach with its first Nelder-Mead step. Although this point could deserve further studies, it is in our view another question overlapping statistics and algorithmic while we wanted here to tackle a biological question applied to a specific case study.
Action: *A paragraph was added in Appendix S4.1, Step 3. to clarify this point and we discussed in section 5.4 (lines 466-469) that a hybrid strategy was needed in our case study.*

To put the approach in a broader context it would have been nice to include a comparison with space-time latent Gaussian random field (LGRF) (Lindgren et al., 2011) as an alternative to the mechanistic model. I am not in a position to request that the author actually do this in their paper, but it would be nice to have a discussion of the merits of the two approaches. An advantage of the LGRF is that it would automatically estimate u0.
Fitting spatio-temporal statistical models properly accounting for dependence over space and time in repeated observations is a classical approach to understand the dynamics of processes of interest. During the last decade, the use of latent Gaussian random field (Lindgren et al., 2011) allowed to efficiently implement statistical inference tools for a large variety of statistical models and, in particular, of space-time modelling approaches (Opitz, 2017). These methods have been popularised through the R-INLA package (Rue et al., 2009), relying on solving Stochastic Partial Differential Equation (SPDE). Among the possibilities offered, the spatial dependence among observations in a continuous space is represented through a Matérn covariance function with one of its estimated parameters defining the spatial range of the spatial process considered.
Fitting such models to our dataset would indeed be possible, all the more that R-INLA can manage in the same model several different likelihoods, i.e. different types of observations. However, the estimated parameters defining the strength of the temporal and spatial dependencies will not allow one to distinguish between the different shapes of dispersal kernels, which was the main goal of our work. Currently, using the SPDE approach for spatio-temporal datasets to model advection and diffusion processes is a current forefront of research (Clarotto et al., 2023).
Action: *Sentences were added in section 5 (lines 379-383) to discuss this point in relation to our objectives.*

References:

Clarotto, L., Allard, D., Romary, T., Desassis, N. (2023). The SPDE approach for spatio-temporal datasets with advection and diffusion. arXiv, DOI: https://doi.org/10.48550/arXiv.2208.14015

Lindgren, F., Rue, H., and Lindstrm, J. (2011). An explicit link between Gaussian fields and Gaussian Markov random fields: the stochastic partial differential equation approach. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 73(4):423–498. eprint:https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1467-9868.2011.00777.x

Opitz, T. (2017). Latent Gaussian modeling and INLA: A review with focus on space-time applications. Journal de la Socit Franaise de Statistique, 158(3):62–85

Rue, H., Martino, S., and Chopin, N. (2009). Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 71(2):319–392.

Minor comments:

1. I was able to extract this material from the Zenodo repository., but not from the gitlab repository, which at the time of trying was “private”.
Action: *We checked the GitLab repository, and the code is now publicly available.*

2. Line 88: should be “the true organism’s dispersal process”
Action: *This sentence was amended (line 88).*

3. Line 101: Missing space before reference
Action: *This sentence was amended (line 101).*

4. Line 624, 662: “Mechanistic-Statistical”
Action: *These references were amended (lines 654, 691).*

References:

Kristensen, Kasper, et al. "TMB: Automatic Differentiation and Laplace Approximation." Journal of Statistical Software 70 (2016): 1-21.

Lindgren, Finn, Håvard Rue, and Johan Lindström. "An explicit link between Gaussian fields and Gaussian Markov random fields: the stochastic partial differential equation approach." Journal of the Royal Statistical Society Series B: Statistical Methodology 73.4 (2011): 423-498.

Journal of the Royal Statistical Society Series B: Statistical Methodology 73.4 (2011): 423-498.