22/05/2024

Dear Recommender,

We are submitting here a revised version for the manuscript entitled "*Hierarchizing multi-scale environmental effects on agricultural pest population dynamics: a case study on the annual onset of Bactrocera dorsalis population growth in Senegalese orchards*", which received a request for major revisions in the previous round of reviews. We are grateful for the constructive suggestions made by the reviewers and the recommender and provide detailed answers to all comments below. We hope that these revisions have improved the MS in a satisfactory manner and that you can consider our work for publication in PCI Ecology.

Sincerely yours,

Cécile Caumette on behalf of the authors,

Recommender's comments:

1/ My main concern is related to the resampling 500 times each estimate of t0. The distributions are rather narrow (see Figure 4), with an average range of 0.7 weeks, which is inferior to the unit of temporal resolution. I would think intuitively that it does create pseudo-replication within the dataset. One of the reviewer was surprised with this approach as well, so please try to clarify this point.

⇨ We apologize for the lack of clarity on the definition and use of the 500 random draws in the posterior estimates of the t_0 parameter (i.e. POPFIT method). Indeed, we note that there is significant overlap between this comment, the next comment and the 2nd comment of reviewer Jianqiang Sun. First of all, we would like to make it clear that each of the 500 sample sets contained a single value of t_0 for each of the 65 orchards and 3 years (i.e. a total of 195 values in each different sample set). In other words, as we have Bayesian posterior distributions of t_0 (instead of single values for each orchard and year), each random sample set (that we named "replicate" or "resampling" in the previous version of the MS) represents one realisation of plausible t_0 parameter values, given the data, for all 195 combinations of orchards and years. Thus, there is no replication here, rather resampling, and we believe that the confusion might be partly explained by the inappropriate use of the term "replicates", at different lines in the previous version of the MS, which we removed in the revised version and replaced by "sample sets".

The replication of the machine learning (ML) algorithm (i.e., GPBoost) on the 500 different sample sets of t_0 values drawn randomly in the Bayesian posterior distributions for each orchard and year allowed to take into account the (rather low) uncertainty in the procedure of environmental variable selection. This is to say that the GPBoost analysis to identify the best predictors is performed 500 times independently on each of the 500 random sample sets, with an independent optimisation of the hyperparameter values. Thus, we are not considering a single ML model, but 500 independent models, run on 500 sample sets of plausible values for the input parameter t_0 of the model. Altogether, the replicated GPboost analyses allowed us to assess how the selection of the best environmental predictors is affected by the uncertainty around the 'measure' of the input t_0 parameter. Although the narrow Bayesian posterior distributions of the t_0 parameter suggested a high precision in the estimation for this parameter (as mentioned by the recommender), the results showed that even small differences in this parameter values can lead to different GPBoost outputs, in terms of SHAP values for each environmental predictor (i.e.

importance of the predictors in the model; see boxplots of Figure 4). With the aim of ranking the predictors in order of importance, we then considered for each of them the median of the distribution of the 500 S_{mean} (mean of the SHAP values) (still shown in Figure 4). To conclude, our procedure allows to select the best predictors given the Bayesian estimation uncertainty on the t_0 parameter, as opposed to the more classical and simpler approach, which would consist in running a single ML model on a single summary statistic of the t_0 input parameter distributions (either mean, median, or mode). We believe that the use of the latter approach would have masked the possible impact of the uncertainty of the Bayesian estimation of the annual onset of *Bactrocera dorsalis* population growth, i.e. t_0 used as an input parameter of the machine learning model.

With this in mind, we formulated more explicitly, in the section "Estimation of the starting date of BD population growth within orchards" in the "Materials and Methods" section, the method and aim for this resampling procedure (see lines 185-191): "*For each remaining orchard and year, we randomly resampled 500 values of the t_0 parameter from the posterior distributions. These values were then associated to build 500 sample sets of t_0, each of them including a single value of t_0 for each combination of orchard and year. This procedure, which contrasts with a more classical approach consisting in extracting a single point value of the posterior distributions (either mean, median or mode), allows to consider the range of plausible values of t_0 given the data and, then, to account for the uncertainty of the estimation in further analyses in which this parameter is the input response variable.*". We also made some changes in the section "GPBoost" of the "Materials and methods" to clarify the procedure carried out independently on each of the 500 sample sets.

Finally, to further help the understanding, we proposed a visual representation of the procedure (in a new figure in the Supplementary material, section 5A, Fig. S5.1). We hope that this is now clearer.

2/ I am also puzzled by the partitioning of the 500 replicates are partitioned into building and training datasets, as they are not independent and values are all really close: basically you use more or less the same date for both steps, which does not really count as validation. You could rather use different orchards, picked from the different biogeographic areas.

⇨ It seems to us that that the recommender's comment may result from a misunderstanding about the use of the 500 random samples from the posterior distributions of the t_0 parameter used as an input of the machine learning analysis (please see our answer to the previous comment). Indeed, the comment suggests that the partitioning of the data into a training and a test dataset has been done on the entirety of the 500 samples of input parameter values, which would mean that a given orchard and year could be included in both, the training and test sets with just a small difference in the t_0 value considering the narrow posterior distribution of this parameter. This is not the case as the partitioning was performed independently for each sample set by sampling 80% (training data) and 20% (test data) of the data among the 195 combinations of orchards and years. Thus, within a single sample set, one combination of orchard and year was included either in the training or test but neither in both. With such a procedure, the training and testing steps performed for a given sample set show how well the ML model performs on previously unseen data (i.e. the ability of the model to generalise), while the comparison of the outputs between the 500 ML models (performed independently on the 500 sample sets) shows how the ML algorithm handles the uncertainty on the Bayesian estimation of the t_0 response variable (as for the selection of the best environmental predictors; see our response to previous comment) and, to some extent, the effect of the random partitioning into training and test sets as they can differ between the independent sample sets. For the sake of clarity, the only constraint we imposed in the sampling procedure to build the training sets was that each of the 18 combinations of years (2012, 2013 and 2014) and sites (S1 to S6) was represented at least by one orchard. We did this because we believe that the numbers of years (3 only) and the number of sites (6 only) are insufficient for the ML algorithm to learn from enough situations to be able to predict a situation that has not been encountered at all. This point echoes the comment #4 of the reviewer Jianqiang Sun on the possibility to consider cross-validation for time-series (please see our response to this comment for further details on this point). Finally, we would like to emphasise that we are by no means claiming that our model can be used for prediction. Here, the predictive performance was assessed as a way of 'validating' the selection of environmental variables, i.e. that the subset of top-ranked variables is sufficient to explain much of the variation in the t_0 parameter.

3/ There are about 200 data points, nested within 69 orchards. How can you be sure of having enough power to test the combined effects of about 30 explanatory variables?

⇨ Here, we indeed have p=28 variables for n=195 observations (65 orchards x 3 years). Boosting is a technique that has proven to be highly efficient to perform variable selection for high-dimensional data (p > n) as appropriate regularization is done through various tuning parameters such as early stopping (Rosset et al. 2004; Bühlmann and Hothorn 2007). We specified this point at lines 265-267 and added the references in the method section of the revised MS.

4/ What do you use the Popfit model if you are only interested in the initial growth phase? Would a simple, exponential-type model be enough?

⇨ The main reason for using the POPFIT mechanistic model is that it is a close representation of the observed BD annual demographic kinetics within orchards (as specified at lines 266-267 of the previous version of the MS). Indeed, temporal dynamics of BD in the Niayes area is characterized by a phase of null or quasi null abundance during the mango off-season, followed by a rapid increase at the beginning of the mango season leading to the demographic peak, and then a rapid decline in abundance to return to the null abundance phase. The use of an exponential-type model requires removing the declining phase to not distort the model fit (or would require to consider a time varying growth rate). We have actually tried such a model using the following approach: we first removed the declining phase by identifying the date of the maximum abundance (the peak) in each time-series and dropping all data after that date; we then log-transformed the abundance data and finally used linear regressions to estimate t_0 as the intersection of the regression line and the baseline of the zero abundance phase. Estimates of the onset of population growth using this approach correlated relatively well with the t_0 estimates from the POPFIT model, but the quality of the fit was less satisfactory for a number of orchards.

Furthermore, although not addressed in this study and as pointed out by the recommender, the POPFIT model provides the possibility to study other demographic parameters of the annual dynamics (e.g. duration of the peak). Studying these demographic parameters may also be useful for understanding the interdependencies between the demographic processes and the environmental matrix; something we intend to address – using the same statistical analysis pipeline (including POPFIT) – in future work.

5/ Could you identify on Figure 3 which were the data points that were excluded from the analysis?

⇨ The four orchards excluded from the analysis were located within site S2. We have now specified that point in the Results section of the MS at line 325.

6/ On Fig 7, the gam-smoothed curves seem overfitted. How did you choose the fitting parameters you used? Have you tried different degrees of constraint on the fitting? Please give more details in the text on the fitting method.

⇨ We agree with the recommender about the GAM overfitting and thank the recommender for pointing this out. The basis dimension (k) was by default in the fitting procedure (bug). We have now fixed this issue by adjusting the value of the k parameter for each predictor's GAM independently. Details on the fit are now provided in the legend of Figure 6. For the sake of completeness, we also added the residuals on the plots in Figure 6.

7/ You interpret the influence of LU13 landscape class as possibly arising from re-infestation from urban micro-gardens. If this is the case, I would expect the LU4 class (orchards) to also stand out in the analysis?

⇨ The density of orchards in the Niayes area is quite high and the monitored orchards are all located in spatial clusters of orchards, which could explain that this class is not identified as a main factor to explain the spatio-temporal variation in the demographic onset of BD populations. In other words, there is not enough variation in the presence of the LU4 class around the monitored orchards to reveal its effect. As mentioned in the last paragraph of the discussion, one of the limitations of this pre-existing dataset is that the monitoring of BD abundance in orchards is clustered in six sites. In further work, a dedicated dataset with a better spatial coverage, including monitoring of orchards located in areas representing both, high and low density of orchards, should allow a better characterisation of the potential effect of this land class on BD population dynamics.

Review by Kévin Tougeron

General comments

1 - In the introduction, it would be great to have a bit more description of the ecology of the fruit fly; how many generation per year, is there a dormancy phase at any moment of the cycle, what is the phenology of the species, how long is the development time, what are the main and alternative food sources, what are the main natural enemies and competitors in Senegal, etc.

⇨ More information on the biology of *B. dorsalis* (development, reproduction, diapause, hosts and shelters) is now included in the introduction, at lines 91-98, lines 104-105 and lines 110-114 of the new version of the MS. However, we have not addressed all the points mentioned (such as natural enemies, competitors…) because we feel that this would distract the introduction away from the main focus of our study. In addition, our work does not bring any prominent results and conclusions on these aspects that we cannot investigate with the data at hand. We would also like to point out that the phenology of BD, in relation to environmental conditions, was already largely addressed in the discussion section (see in particular lines 438-445 and lines 490-511 of the new version of the MS).

2 - As regards to the analyses carried out, I do not have the necessary expertise to judge the relevance of the methods used. I do, however, find that the part concerning the methods is well described and makes it easy to understand what was done.

⇨ We thank the reviewer for his support. This was an objective of the study to share a repeatable, clear and sound methodology. We are very pleased to read that this objective has been achieved.

3 - The capture method of BD flies is the use of pheromones to attract males. Yet, females and eggs are the main issue for orchards. I wonder if there is any proof that male and female abundances are well correlated. If not, it could be an issue regarding the selected method.

⇨ Field monitoring of BD abundance has largely been conducted using pheromone-based attractants for males. This strategy has been justified by the fact that BD sex ratio has been estimated to be 1:1 and by the difficulty of trapping females, which relies on food bait that are far less efficient as well as nonspecific (catching other non-target species). Manrakhan et al. (2017 - DOI: 10.1111/jen.12373; 2019 – Citrus International Research) conducted a more rigorous field study to compare the number of catches of BD males (using male-lure methyl eugenol baited traps) and females (using food baited traps). Their results show that if the temporal dynamics of both sex in orchards is quite similar, catches of males are generally higher and earlier. This difference is attributed to the higher efficiency of pheromone-based attractants for males rather than a significant difference in abundance between both sexes (including an attraction range of 500m for methyl eugenol compared to 30m for food bait). The authors actually advised to use male-lure traps to obtain indication on the local population size and for early detection and, food-based attractants, as an indication on the potential threat of female flies as the fruits ripen. We have now added information on this point at lines 156-164 of the new version of the MS.

4/5 - I find the results section quite hard to follow, especially as the methods used and the metrics calculated are not the most common. Care should be taken to avoid statistical jargon as much as possible, and to explain the results obtained in biological terms. For example, without going into interpretation, what does it mean that SHAP values correlate with PCA Axis 3 values? This also applies to the discussion, which is—for some parts—still quite complex to understand. It is interesting to discuss the new methods put forward in this article, but the authors should also

take care to make their article accessible to a wider audience of ecologists and entomologists, as the subject covered by the article is also highly applied to biological control.

- ⇨ We understand that statistical terms and interpretations may somewhat hinder the flow for readers who are more interested in going straight to the biological interpretation. However, we believe that avoiding appropriate vocabulary from the statistics field when referring to analyses and metrics (the so called "statistical jargon") or avoiding to discuss the novelty, conditions of use, advantages and/or limits of the statistical methods is not a satisfactory mean to ease the reading. To us, and we believe for the audience of PCI Ecology as well, it is critical to make as clear and precise as possible, what has been done and why, since it ensures the quality of the reviewing process, the replicability of the results, and the reusability of our statistical methodology on other datasets by other researchers. We have made efforts in this direction in the first version of the MS, and more generally to follow the FAIR (findable, accessible, interoperable, and reusable) principles, not only for the (raw) data, but also for the workflow (including different used algorithms) that led to the generation of the (processed) data (and results). Several considerations from the recommender and reviewers have helped us to further clarify some aspects of the methodological process in the new version of the MS.

  In the revised version, we have done our best to further facilitate the reading and understanding of the results in terms of biological meaning. In particular, as we were aware that interpreting the importance (expressed with the SHAP value, which is defined in the method section) of composite variables, such as the principal components summarizing the variability of physical variables, may not be straightforward, we have been careful to give insights on the biological meaning of the relationships between the statistical metrics used. For instance, in the example pointed out in the reviewer comment ("what does it mean that SHAP values correlate with PCA Axis 3 values?") we provided insights on the biological meaning of this correlation in the results section - without going too deep in the interpretation - in the following sentences (lines 352-355 of the revised MS): *"The SHAP-dependence plot (Figure 6) showed a positive relationship between PC3 values and the individual SHAP values (i.e. negative SHAP values for the lowest PC3 values and positive SHAP values for highest PC3 values), meaning that earlier start dates of BD population growth (t_0) in orchards were associated with higher values of NDWI, i.e. humidity."*

  We hope that these modifications will make it easier to read the MS and understand the biological significance of the results.

Specific comments:

L37: It is not clear in the abstract what "Gradient boosting" is. I am not sure the term should appear here.

- ⇨ This term has been removed from the abstract.

L57-58: It may be an issue but is has been largely addressed in agroecology studies in the past decade. I would suggest to temper a bit this statement.

- ⇨ The point here was not to elude decades of researches in agroecology but rather to point out that despite this effort, ecological processes are still often overlooked in IPM, as detailed in Deguine et al. (2021). We apologize to the reviewer if this sentence appeared the wrong way and we changed it in order to better acknowledge past agroecological research: "*However, and despite decades of research in agroecology, IPM implementation often still lacks careful* *consideration of the spatio-temporal heterogeneity of ecological processes occurring in agroecosystems (Deguine et al., 2021)*” (lines 53 to 55 of the MS).

L53-59: The term “spatio-temporal heterogeneity” should appear in this paragraph.

- ⇨ We added the “spatio-temporal heterogeneity” as follow in line 55: **“***However, and despite decades of research in agroecology, IPM implementation often still lacks careful consideration of the spatio-temporal heterogeneity of ecological processes occurring in agroecosystems (Deguine et al., 2021)***”**

L89: Please give full taxonomic description and authority of B. dorsalis

- ⇨ This information has been added at lines 86-87 of the new version of the MS.

L96-98: Has this hypothesis been formulated anywhere else in a research paper or technical report before?

- ⇨ To our knowledge, this hypothesis has not been formulated for BD under the Sahelian environment from the Niayes, Senegal. However, our hypothesis echoes other observations or explanations proposed for other fly species in tropical environments, such as in Bateman 1972 (*“Overwintering in the more tropical species is normally accomplished by adults. They tend to congregate in locations which provide shelter and food…”*) or in Clarke et al. (2022).

  We have now specified this point in the Introduction section and acknowledged similar formulated hypotheses or previous observations in line with our hypothesis. Note that Clarke et al. (2022) was initially detailed only in the discussion section in order to address the possible reproductive arrest in *Bactrocera spp.* during the dry season when breeding hosts are scarce.

L107: Precise what you mean by “environmental features”

- ⇨ The type of environmental features is explained in the previous sentence *“...environmental data on a large number of candidate predictors at different spatial scales, including cropping systems (Diame et al., 2015; Grechi et al., 2013), landscape structure (Jolivot 2021), and weather variability (Karger et al., 2021; Didan 2015)”*. The full description of the 28 variables is provided in the method section, and a synthetic table is also presented in the Supplementary material, Section 5A, Table S5.1. Since we feel that the misunderstanding can come from the alternative use of the term “features” and “variables”, we have homogenized with the more conventional “variables” term.

L128: In Fig 2, does the “orchard” land cover only represent mango orchards?

- ⇨ The “orchard” class of the map of Jolivot (2021) includes all orchards with no difference between orchard types. In the Niayes area, a few citrus-only orchards are present, but otherwise mango orchards remain largely dominant. These orchards can, however, be mixed with other fruit trees, especially citrus. We added this information at line 203-205 of the new version of the MS *(“In the Niayes, orchards can consist entirely of mango trees or be mixed with other fruit trees, such as citrus, papaya or guava, which are potential alternative hosts for BD (Grechi et al. 2013)”*). In addition, we have specified the information for our 65 studied orchards by adding the following sentence in the Supplementary material, Section 2: *“Among the 65 studied orchards, 11 orchards (17%) were strictly composed of mango trees, 1 strictly of citrus trees (1.5%) and 53 were mixed, with both mango trees and alternative host tree species (Table S2.1B)”*. Other species that can be found in much lower density in Niayes orchards are detailed for our 65 studied orchards in the Supplementary material, Section 2, Tab S2.1.

L143: So the time series is at a daily scale, right?

- ⇨ Time series are at a weekly scale, i.e. traps were collected once a week. The starting date is expressed in number of weeks since the first of January for each orchard and year. We have tried to make this point more explicit in MS. Please see lines 165-169 of the new version of the MS.

L154: What threshold did you chose to assess if the model was poorly fitted to the data?

- ⇨ The assessment of the quality of the model was done in two steps:

1) We checked for the good convergence of the MCMC chains to ensure that the parameter t_0 was correctly inferred by looking at the trace plot of the chains and by computing the Gelman-Rubin statistics: a Rhat value above the threshold of 1.1 indicates that the convergence has not been achieved (see details in the Supplementary material, Section 1). The convergence was not satisfactory for one orchard (even when the length of the run and the number of chains were increased, i.e. 500000 iterations, 300000 burnin, 200 thin and 40 chains). This orchard was then removed from the dataset.

2) We also did a visual check of the fit (i.e. there is no threshold since it is not quantitative), from which we decided to remove 3 additional orchards that had suspicious abundance time series with a very sharp change in abundance in just one week.

Examples of poor fits have been now added in the Figure S1.1 in the Supplementary material. Thus, this figure now provides examples of times series with both, good or poor fits. In addition, those points have been clarified in both the main text (see lines 180-185 of the MS) and in the Supplementary material, Section 1.

L168: It is not clear; are the orchards all mixed-species orchards?

- ⇨ No, some orchards, especially when dedicated to export, are 100% mango trees, most time even mono-cultivar, while more traditional orchards can be mixed with other fruit trees (mostly citrus) that have been listed as potential alternative hosts for BD. In addition, vegetable crops can be grown under trees. Table S2.1 in Supplementary material summarize the different mango cultivars and fruit trees observed in the studied orchards and the way they have been classified based on their phenology to be integrated as predictors in the GPBoost model. We added information in the MS and the Supplementary material to clarify this point (please see our answer to the previous comment: “L128 - does the “orchard” land cover only represent mango orchards?”).

L178-179: At some point it would be necessary to mention the name of the species other than mango that can host the fly and their respective phenology.

- ⇨ Other species suspected to host BD in the studied orchards and their phenological classes are presented Table S2.1 in Supplementary material, Section 2, which is referenced at the end of the paragraph “Multi-scale environmental predictors” of the section “Materials and methods” in the MS.

L192-199: This method is interesting and well-described.

- ⇨ Thanks

L268: How so, exactly?

- ⇨ Please see our answer to the previous comment, i.e. “What threshold did you chose to assess if the model was poorly fitted to the data?”

L269-273: I am not convinced that Figure 4 is essential to present here in the main document. It could be put in supplementary material. The figure is quite difficult to understand and does not bring much added value to the text.

- ⇨ We agree with the reviewer and the figure has now been moved in the Supplementary material as Figure S1.2 in Section 1.

L344 and onwards: What is the potential role of temperature variations along with humidity? In phenological models that are usually constructed for temperate regions, temperature is usually the main factor affecting insect pullulation and early arrival in the fields. It's understandable that humidity is a major factor here, but how is it linked or correlated to temperature?

- ⇨ We added a graph in the Supplementary material, Section 4 which illustrates the average annual variations of the monthly temperatures (mean, minimal and maximal) and precipitations in the Niayes area. Mean and min temperatures tend to increase from April and are maximal during the rainy season (June to October). As we detail in the discussion, “*temperatures for optimal immature development ranged around 25-30°C, with development time (or mortality) increasing at lower (or higher) temperatures, preventing from any adult emergence above 35°C (and below 9-10°C)*” and “*females can only lay eggs between 15 and 35°C, with the optimal conditions for a higher number of eggs being between 20 and 25°C*”. So, the continuous increase in the min and mean temperatures between April and July may promote immature development and female laying, which can then boost the pullulation. It is worth noting that minimal temperature conditions in Senegal never reach critical level for BD, contrary to maximal temperatures, which easily go higher than 35°C. If the temporal variation in temperature may indeed have an effect on early population development, in this study we rather found an effect of the spatial variation in temperature on the onset of BD population growth. As we state in the discussion “*the most favourable temperature range for early population development in orchards lies between oceanic conditions in the coastal part and inland conditions where the maximum daily temperature easily exceeds 35°C during the dry season*”. We further address the relationship between temperature and humidity as follow*: “As temperatures above 35°C challenge all components of BD life history, spatial and inter-annual weather variability in the Niayes region is likely to interact with local factors providing higher levels of humidity and shading (e.g. water bodies and groundwater, vegetation and soil moisture, canopy structure) to create favourable microhabitats allowing BD to mitigate hydric and thermal stress during the dry season”* (lines 447-451 of the revised MS).

L365 and onwards: NDWI is generally related to canopy density, but also to the semi-natural elements present around the orchard. The buffering effect of certain elements on temperature or humidity has already been demonstrated, as has the effect on crop pests and their natural enemies, and would certainly merit further consideration in this discussion.

- ⇨ In our study, NDWI, like all other physical variables, is considered at a resolution of 1 km, so that it integrates the canopy effect of orchards and the potential buffering effect of their surroundings (semi-natural elements or otherwise).

Is the NDWI calculation a reliable source for measuring a microclimatic effect on the scale of an orchard plot?

- ⇨ Precise measures of humidity at the orchard scale would probably require dedicated data loggers. Here, we consider NDWI, like other physical variables, at a resolution of 1km, which we believe can still be considered a microclimatic effect but at the scale of the orchards and their surroundings rather than the orchard per se.

L390: Could a link be made here with other insect models, such as Drosophila suzukii for instance?

- ⇨ A comparison could be made with species such as *Drosophila suzukii*, for which the presence of urban habitats has been shown to enable winter survival in northern climates (e.g. Dalton et al. 2011 - DOI 10.1002/ps.2280; Rossi-Stacconi et al. 2016 - DOI 10.1007/s10340-016-0753-8). We have added a sentence specifying that point as well as a reference to a very recent study on the Mediterranean fruit fly, which points the role of nearby urban area as a source of infestation in orchards (Broadley et al. 2024 - DOI 10.1016/j.ecoinf.2024.102536) (lines 458-462 of the revised MS).

L397-401: What about the potential role of refuges or suitable habitats for natural enemies of the fly?

- ⇨ Habitat such as savannah may indeed constitute refuge for parasitoids and predators of BD. We added the following sentence to specify this point (lines 472-473 of the revised MS): *"Habitats such as shrub savannah may also shelter natural enemies that could impact BD abundance and dispersal (Vayssières et al., 2016)."*

L405: Is there any mango variety known to better resist BD?

- ⇨ Although there are more or less susceptible varieties in Africa (Diatta et al., 2013; Isabirye et al., 2016 – DOI 10.1080/15538362.2015.1042821; Mokam et al., 2024 – DOI 10.1093/jisesa/ieae027), the main resistance trait of mango varieties is their earliness, enabling them to be harvested before the demographic peak of BD and subsequent damages on fruits (Grechi et al., 2021 – DOI 10.1016/j.cropro.2021.105663).

L454-464: Those limitations are worth being mentioned, but maybe not at the very end of the discussion. The previous paragraph (L442-453) would fit better as a conclusion paragraph.

- ⇨ We agree and the two paragraphs have been swapped in the revised version of the MS.

Review by Jianqiang Sun:

1 - Authors developed a flexible analysis pipeline to hierarchize the effects of multiscale env variables on the timing of annual BD population growth. However, there is a lack of validation of the methodology. The authors should indicate how much better the developed method is compared to existing methods/pipelines. Is the performance of conventional methods (e.g., random forests, LASSO) definitely lower than the proposed method?

- ⇨ Although our aim was primarily to highlight the potential and flexibility of GPBoost when dealing with spatio-temporal data using our case study, we do agree with the reviewer that the performance of GPBoost compared to other methods should be addressed to some extent. To summarize, first, GPBoost is a tree-boosting method, which generally shows the highest prediction accuracy among ML methods on a wide range of datasets (Johnson and Zhang, 2013; Nielsen, 2016; Grinsztajn et al., 2022). Second, not all machine learning methods can be combined with mixed effect models and even less so with a Gaussian process. When possible, for example with random forests, the integration of a Gaussian process is only possible via a two-step approach (Saha et al., 2023 – DOI 10.1080/01621459.2021.1950003), and such approaches are known to be much less efficient than the joint estimation of the Gaussian process and the mean function (see e.g. Sigrist, 2022). Finally, Sigrist (2022) who compared a wide range of statistical and machine learning methods, such as linear models, gradient boosting or random forests, in combination with mixed effects models, found that the GPBoost algorithm gave the highest prediction accuracy.

  Although other methods could be used instead of GPBoost in the pipeline we have described, we believe that this new method can provide an efficient alternative for research in ecology. In particular, to address issues of spatial and temporal dependencies that arise when studying spatio-temporal population dynamics. We now provide more details in the main text, at lines 266-276, to better highlight the seven models we tested in this study (which are detailed in Supplementary material, Section 5B), as well as references to previous work highlighting the performance of tree-boosting methods in general, and GPBoost in particular, compared to other methods.

2 - [L154]: I couldn't catch "500 values of the estimated t0 (L154)". The data consists of 69 orchards over 3 years. To my understanding, at most 69*3=207 models can be obtained from all data. Please explain simply here.

- ⇨ Please see our detailed response to the recommender's first comment. In brief, the (POPFIT) Bayesian inference does not provide a single value of t_0 but posterior distributions of plausible parameter values given the data, for each combination of orchard and year. We sampled 500 values of t_0 in the posterior distributions of this parameter, for each of the 65 retained orchards and 3 years (i.e. a total of 195 combinations), in order to replicate the GPBoost analysis on 500 plausible but different sample sets of data. Thus, each of the 500 demographic sample sets is built by combining only one value of t_0 from each combination of orchard and year, for a total of 195 values (see the new Figure S5.1 in Supplementary material, section 5A, which illustrates this procedure). This strategy allows us to take into account the uncertainty around the t_0 estimates, used as the input response variable for the GPBoost model for the selection variable procedure (i.e. effect of the variability – even small – of the input dataset on the output of the machine learning model).

3 - results clearly suggest that humidity conditions are the primary driver ... [L345]: Is the humidity the primary factor? Is it possible that pests start to increase in line with the time of year when fruit starts to ripen? Is there a pseudo-correlation between the time of fruit ripening and the time of increased humidity, with increased humidity leading to an increase in pests? What happens when humidity is removed from the model? Does prediction performance deteriorate significantly?

- ⇨ This is a relevant and tricky question to unravel in the case of real world (field) data. We below develop some arguments, which altogether rather suggest that our results are robust to temporal correlations and possible interaction effects between humidity (the top predictor in our results) and mango fruit ripening. Admittedly, there is a strong annual correlation between fruit ripening and humidity, as the rainy season in the Niayes area is roughly from May to October and the mango production season mostly from June to August.

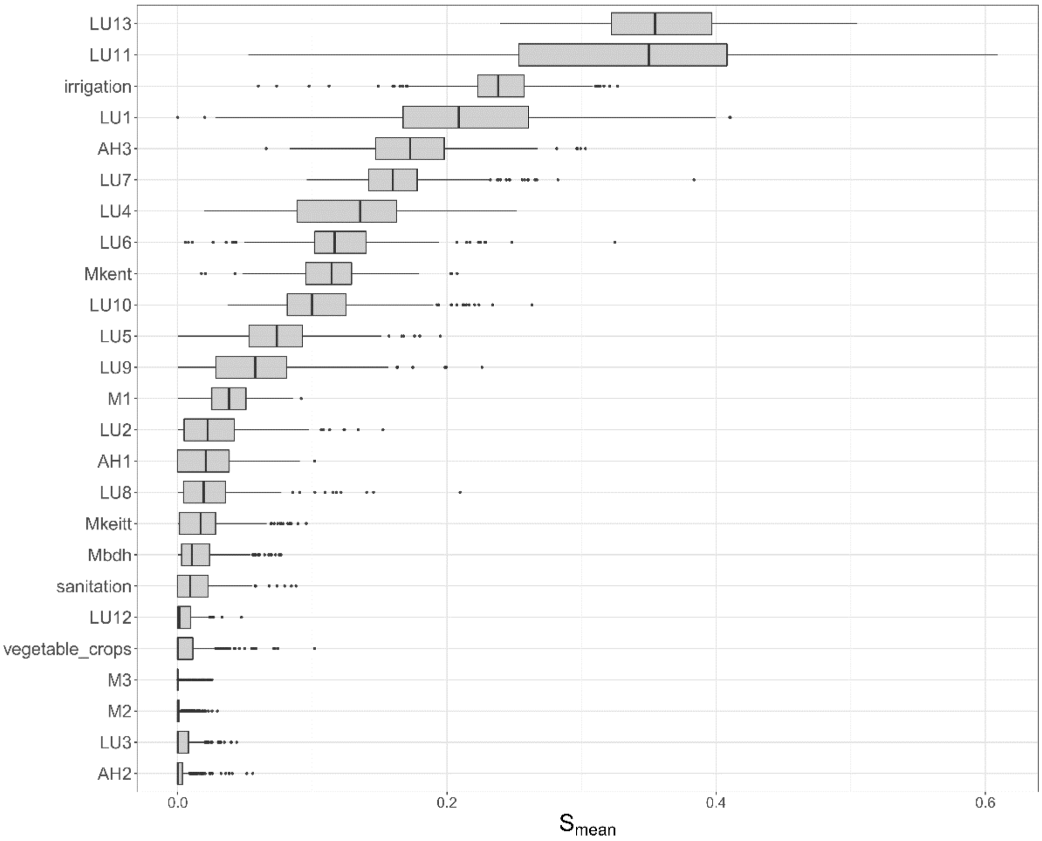
  First of all, if the proportion of mature fruits locally available for BD was the main factor, rather than humidity, we could expect to find the proportion of the earliest mango cultivars, to be among the top predictors explaining the spatio-temporal variation of the parameter t_0. Among the most common mango cultivars in the Niayes (i.e. Kent, Boukodiekhal and Keitt; see details in Table S2.1 in Supplementary material, Section 2), Boukodiekhal is the earliest, with fruits starting to mature in May, whereas fruit availability for Kent and Keitt is from June to August and July to October, respectively. We did not find any effect of the Boukodiekhal cultivar, nor of other mango cultivars, on the spatio-temporal variation of t_0.

  Moreover, and most importantly, we actually did not consider the temporal window of high correlation between humidity and fruit ripening (i.e. June to August). Indeed, all climatic variables considered were only from December to May, which corresponds to the dry and mango off-season, when precipitations are very scarce (i.e. episodic events, generally, of a few millimetres mostly between October and March (Wade 2015)). As shown by the second component of the PCA on physical variables (PC2) in Figure 5A, the earlier start dates of BD population growth are associated with the higher level of precipitation (although it remains small amounts) between February and April, before the mango fruits ripens, even for the earliest cultivars (e.g. May for Boukodiekhal).

- ⇨ To further assess the point highlighted by the reviewer, we rerun the GPBoost analysis after removing the three principal components of the PCA made on weather and NDWI data (PC1, PC2, PC3) from the list of predictors (i.e. we considered only 25 predictors instead of 28). We used the same procedure as described in the MS.

  First, we did the tuning of the hyperparameters and the model training independently on the 500 sample sets excluding PC1, PC2, PC3. The result of the ranking of the predictors (according to their importance in the model) is presented on the Figure (A) below:

(A)

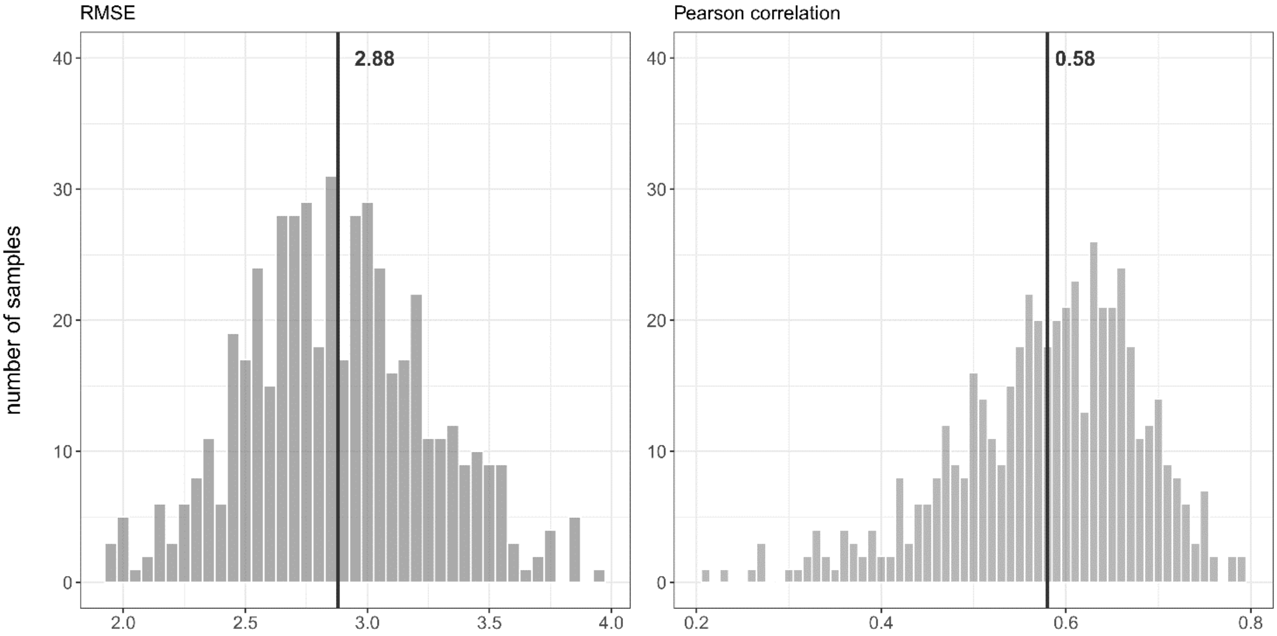


These results are broadly consistent with our results for all 28 predictors (i.e. including PC1, PC2, PC3 ; see Figure 4 of the MS), with fairly similar rankings and S_{mean} range values for the predictors.

Regarding the grouped random effects values, they are also similar between the two analyses, with median and range values for the *year* and *site* of 2.54 [2.40, 2.81] and 0.17 [0, 1.03] respectively for the model including only 25 predictors, and 1.86 [0.72; 3.03] and 0.14 [0; 1.64] respectively for the model including all 28 predictors (see lines 331-332 of the MS). However, the error term increases significantly when PC1, PC2, PC3 are removed, with median and range values of 5.69 [5.23, 6.36] for the model including only 25 predictors and 0.4 [0; 1.72] for the model including 28 predictors (see lines 331-333 of the MS).

Second, we partitioned each of the 500 sample sets into a training dataset (80% of the observations) and a test dataset (20% of the observations). The tuning of the hyperparameters and the training of the model were done independently on each training dataset and the prediction was done on the corresponding test dataset. The RMSE (Root Mean Square Error) and the Pearson coefficient correlation between the observed and the predicted t_0 values, calculated on the test dataset independently for each of the 500 sample sets, are presented in the Figure (B) below (the vertical lines represent the median values over the 500 samples):

(B)



The results of the prediction including all the 28 predictors (i.e. including PC1, PC2, PC3) showed a RMSE and Pearson coefficient correlation of 2.24 and 0.77, respectively (see details in the Supplementary material, Section 5C, Figure S5.3. The predictive performance of the model shows a clear decrease when the physical variables are removed from the predictors.

4 - Please consider using cross-validation for time-series data. For example, training with 2012 and validating with 2013, training with 2012-2013 and validating with 2014.

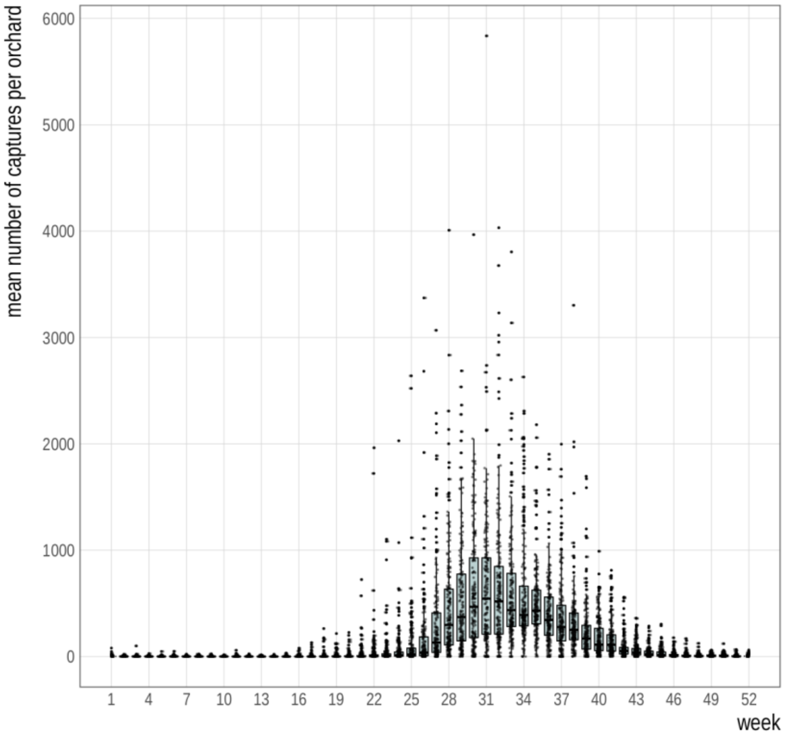
- ⇨ This point echoes the comment 2 of the recommender. We first would like to recall that we did not develop a model to be used for prediction but to select the environmental variables that best explain the variation of the start date of BD population growth for the observed set of orchards and years. But more importantly, as we only have 3 years of data, we do not think we can really do any meaningful temporal cross-validation, whatever the method. If we perform a temporal split such as 2012-2013 for training and 2014 for test, then the year random effect variance in the GPBoost model would essentially be estimated using only two observations (2012 and 2013), which would be like attempting cross validation with 3 data points. However, such an approach could technically be applied to long time series.

5 - Sahelian climate [L132]: If possible, please visualize some important meteorological data (e.g., temperature, precipitation, humidity) with charts from 2011 to 2014 or the average of three years (better to merge in Figure 3). This may help readers who are not familiar with Sahelian climate to easily understand the characters of climate variables.

- ⇨ We added the Figure S4.1 with monthly temperatures (mean, min and max) and precipitations averaged over four year (2011 to 2014), in Supplementary material, Section 4 – “Averaged weather conditions in the Niayes over the study period”.

6 - Fig 3 [L148]: Please consider using jittered points over boxplot (or use violin plot) to visualize the data density.

- ⇨ We considered these possibilities but the visuals turned out quite uninformative or even obscured, see below:


mean number of captures per orchard
week