February 16, 2024

Re: PCOMPBIOL-D-23-01609. "Informing policy via dynamic models: Cholera in Haiti"

Dear PLOS Computational Biology editorial board,

[TODO: Edit this letter] Thank you for arranging the review of our manuscript, and for the invitation to submit a revision. We are grateful for the thoughtful comments from the referees. Our revised manuscript has various edits for clarity, following the referee suggestions, as well as some small corrections in the text. The numerical results are unchanged.

To make space for some additional sentences, we cut a few sentences that seemed least essential. We also placed some of the new material in the supplement.

Below are point-by-point responses to the referee reports. The reviewer comments are shown in green, italic type. Our responses are in black, and material copied verbatim from the revision is in blue.

Sincerely,

Jesse Wheeler
PhD Student, Statistics Department
University of Michigan
jeswheel@umich.edu

Referee 1

> *The authors' manuscript aims to explain how public health decisions can best be informed by combining an understanding of the underlying pathogen transmission dynamics with epidemiological data. In particular, the authors make a number of recommendations of best practices that modellers can adopt when using mechanistic models to make statistical inferences from time series data. To make their recommendations more concrete, they perform a re-analysis of cholera incidence data from Haiti spanning the years 2014-2019. These data were previously analysed in a modelling study performed by a separate team of scientists, published in Lee et al. (2020), see Ref. 1. Using refinements of three models from Lee et al., the authors demonstrate a number of ways in which the analysis could have been improved. Furthermore, the authors present differences between the findings of their reanalysis and the original analysis of Lee et al., for instance in the predicted probability of eliminating cholera from Haiti (see Figure 8).*
>
> *My most substantial concern is with the framing of the paper, and whether that might mislead potential readers. To me, the paper serves more of an educational purpose, presenting ideas on how to best use mathematical models to gain insight from time series data. While the case study into cholera in Haiti is methodologically sound, it does not seem to be the main focus of the manuscript. To such an extent that, in the discussion, the only comment to on the Haiti analysis is that the authors "used the same data and models, and even much of the same code, as Lee et al. [1], and yet ended up with drastically different conclusions."*
>
> *This major concern can be resolved in two ways, both of which seem acceptable to me, either a) reframe the paper as a more general series of recommendations for how to better perform statistical inference on time series data using mechanistic models (including a longer introduction and discussion of existing methodological deficiencies) or b) perform a more in-depth analysis and commentary on cholera in Haiti (some of which has already been done but is buried in supplementary information files that are only briefly mentioned in the main text, e.g. S5).*

Thanks for the feedback on the framing of our contribution. The original goal of our manuscript was to combine the two foci identified by the referee: we use an in-depth case study to motivate general procedures for fitting mathematical models to time series data, while explaining how these procedures help to address limitations of previous approaches. The case study was intended both to help motivate the advice and to make sure that the advice is pertinent and practical on at least one suite of models. Since our approach is predicated on the possibility of carrying out some nontrivial computational procedures, we are led to address computational issues in parallel with scientific and statistical issues.

The referee would prefer us to emphasize one of these foci, and this feedback has helped us see how readers might find themselves struggling to orient themselves while reading the manuscript. We agree that the "educational" purpose of the paper is primary to the specific conclusions about the Haiti cholera epidemic. However, the validity of our reanalysis of the Haiti data and models provides critical support to our methodological points. For the revision, we have edited the abstract and introduction to make explicit how the case study fits into the overall goals of the paper. In the discussion, we return from the concrete consideration of the case study, placing it back in the more general setting of inferring complex dynamics from time series data. In the revision, we have added more signposts to help the reader see how the points that arise in the discussion generalize the corresponding points arising in the case study.

> *I found the methodology used to be well-suited to addressing the aims of the paper.*
>
> *Overall, I found the paper interesting and informative, but a little unfocused. I expect with some heavy rewriting in places (especially the introduction and discussion), the paper will be acceptable for publication in PLoS Computational Biology. I do not anticipate the authors having to perform additional computational analysis.*

Thank you for this feedback. We have made substantial changes to the abstract, author summary, and introduction in order to make the focus of the paper more clear. We have also done a nearly complete rewritting of the discussion section. Specifically, though the content and arguments made remain the same, we specifically point out ways that our case study demonstrates the usefulness of our general suggestions for model based inference of dynamic systems.

> ***Abstract***
> *The authors state they "develop data analysis strategies leading to improved statistical fit." It's not clear to me exactly what novel strategies have been developed, it seems the authors exclusively use pre-existing techniques from the literature. This is fine, if the focus of the paper is either a) pedagogical or b) on understanding a specific disease system (cholera in Haiti). If any methods are novel, this should be more clearly stated.*

Fitting models 1 and 2 does not require any methodological innovation. Model 3 has a stochastic spatiotemporal partially-observed structure that leads to methodological challenges. [1] made some approximations to avoid carrying out inference for the full coupled system. We show some consequences of those approximations, taking advantage of a newly developed method which has not yet been applied in the context of scientific investigations. So, although the method is not entirely new, its applicability to this task is novel. Additional commentary to this point has been added in the abstract, introduction, and the subsection on fitting Model 3.

We have further modified the statement in the abstract that led to this confusion. Rather than stating that we "develop data analysis strategies leading to improved statistical fit", we now say:

> *We assess current methodological approaches to these issues via a case study of the 2010-2019 cholera epidemic in Haiti. We consider three dynamic models developed by expert teams to advise on vaccination policies. We evaluate previous methods used for fitting these models, and we demonstrate modified data analysis strategies leading to improved statistical fit. Speficially, we present appraoches of diagnosing model misspecification and the consequent development of improved models. We futher demonstrate the utility of recent advances in likelihood maximization for high-dimensional nonlinear dynamic models, enabling likelihood-based inference for high-dimensional models for spatiotemporal incidence data.*

> ***Introduction***
> *1. To make clear that the issues being addressed by this paper go beyond one modelling study of Haiti, the authors should provide examples and references to support the (accurate in my opinion) statement about, "common modeling decisions that may not provide an adequate statistical explanation of the data".*

That statement arose in the author summary, where references are not expected. However, we appreciate that the point should be expanded on, and we have done so in the revision. Specifically, we cite papers from the literature review we performed on the use of dynamic models for analyzing the Haiti cholera oubreak. The literature review was previously mentioned, but citations from the review were not made. Furthermore, we modified the quoted statement in the author summary to be the following:

> *we use an in-depth case study to motivate general procedures for fitting mathematical models to time series data, while explaining how these procedures help to address limitations of previous approaches.*

Among the most common mistakes in mechanistic modeling, in our opinion, is not properly assessing / evaluating a fitted model. We originally tried to make this point by having an entire section dedicated to model diagnostics, though we admit that the connection between the statement in the author summary and a section later in the article was not clear. We have addressed this by adding sign-posts to references and suggestions back to this statement, and other similar statements in the abtract / introduction. Our modified discussion also includes detailed examples of principles that were very useful in our case study but are not universal practice.

> *2. Given the focus of the paper on how to better perform inference, explicitly commenting on what are the "existing guidelines for creating models to inform policy [4, 5]". It would then also be helpful in the discussion to comment on how the recommendations of this study extend these pre-existing guidelines.*

We have added two paragraphs in the introduction discussing how recommendations we make in this article fit in the broader category of papers that make recommendations on fitting mechanistic models to dynamic systems. This includes discussing these two papers that we originally cited in more detail, as well as other articles on similar topics:

> *There are many guidelines available regarding the use of mechanistic models for policy influence. For instance, Behrend et al. [2] conducted an extensive literature review on modeling principles and standards. They used this review to develop general recommendations for policy-guiding modeling. This review is particularly relevant given that their recommendations are specifically intended for the modeling of neglected tropical diseases, which includes cholera. Their resulting recommendations for communicating model results align with other manifestos on the proper use of models intended to inform policy [3, 4], each emphasizing the importance of involving stakeholders in the modeling process, transparency, and communicating uncertainty. Other recommendations that focus on issues of model calibration and assessment also exist [5–7], demonstrating the considerable amount of published guidance on how to effectively model dynamic systems.*

> *Based on our assessment and literature review, [1] was among the studies that best followed these general recommendations for policy-driven modeling. Concretly, [1] follows at least four of the five principles outlined in [2]:* complete model documentation, complete description of data used, communicating uncertainty, *and* testable model outcomes. *Determining the level of adherence to the first principle,* stakeholder engagement, *is difficult based solely on the article. Despite this, the inconsistency between their forecasts and the cholera incidence from 2019 to 2022 suggests that existing recommendations and standards related to policy-driven modeling may be insufficient. [8] suggests that improvements in model model-based outputs may be obtained by developing structures and standards based on statistical principles. Our general recommendations therefore complement and extend existing guidelines by focusing on the methodological tasks of calibrating and evaluating dynamic models in a rigorous statistical framework. We specifically emphasize principles that proved essential in our case study; complementary methodological suggestions arising from a spatio-temporal analysis of COVID-19 are given in [9].*

We have also added some material in the discussion connecting our results to the stated goals of the paper from the abstract and introduction. We could have gone further and discussed in more detail the relationship to the references [2,3], following your recommendation. However, our newly added reference [9] more directly addresses [2,3] and so we have referred the reader to that paper (currently available on arXiv) for additional discussion of this point.

Methods

3. Fig 2. Model parameters. I think this should be "Table 1"?

While Fig 2 really is a table, here we are following PLOS Computational Biology Submission guidelines (`https://journals.plos.org/ploscompbiol/s/tables`):

> "If the table has a very complex structure or contains graphics, the safest solution is to make it into a figure. Export the table as a TIFF, and cite and re-label it as a figure."

The Table in this case is necessarily complex, as we are trying to put all comparable model parameters from three seperate models into a single table. If the journal requests that we make this into a table rather than a figure we are happy to comply.

4. "and $z \in 1 : Z$ describes hypothetical vaccination programs" It would help to give an example of the different vaccine programs here. E.g. does this mean the number of doses the individual has received, or differences in the vaccine administered?

It could be either of the above, though in the analysis we only consider compartments based on the vaccination scenarios outlined in the Forecasts section. For Model 1, the assumption made by Lee et al. [1], and reused in our analysis, is that the added dynamics of the vaccination compartments is an increased probability that an infection is asymptomatic, as described in a paragraph that follows the quoted text. a clause has been added to this sentence in order to clarify this point:

> *Here, $z = 0$ corresponds to unvaccinated individuals, and $z \in 1 : Z$ describes hypothetical vaccination programs, where each program z indexes differences in both the number of doses administered (one versus two doses per individual) and the round of vaccine administration.*

> *5. Eq 6 and line 1: I think the notation for the asymptomatic fraction might be clearer and more consistent with the other models if you used $f_z(t)$ instead of $\vartheta_z(t)$. In fact, this is the notation used in the table in Figure 2.*

Thank you for this suggestion, we have made this change.

> *6. Line 198: Reference for where value of v_{rate} comes from.*

This parameter comes from Lee et al. [1], and the only details that they provide is a subscript comment in one of their tables (S13) in their supplement: "v_{rate} was chosen to mimic the distribution of infections in Haiti". They do not, however, list it as a parameter that was "fit" in the same table, making it unclear where the value comes from. Presumably, it was chosen somewhat subjectively based on their understanding of cholera dynamics in Haiti, because in order to obtain a fit for their model they had to have fixed the value at something without actually estimating it, and perhaps only considered a very small set of possible values. Notably, this parameter can mathematically be estimated like all other parameters in the model, but their specific implementation of Model 2 does not permit its estimation. Parameters that are partially estimated via subjective measures is outside the scope of our article, so to avoid this the revised manuscript simply reads:

> $v_{\mathrm{rate}} = 10^{-12}\, km^2 yr^{-1}$ *was fixed at the value used by [1].*

> *7. Usually, vaccine efficacy = 1 implies the vaccine completely blocks transmission and 0 implies no effect. In model 2 the definition of vaccine efficacy (lines 200-205 and Eq 14) implies the opposite. I suggest re-parameterising the model, or renaming the variable. For what its worth, models 1 and 3 seem to follow the usual convention.*

This change has been made.

> *8. I found the commentary on comparing fitted mechanistic models to a (statistical) benchmark model to be informative and helpful. I was wondering if the authors had any comments on whether (or not) to directly compare the AIC values of the different fitted models directly. Am I right in thinking this should be possible for models 2 and 3 given (I think) they are fitted to the same data? Model 1 is fitted to a different data set, which I understand makes such a comparison impossible.*

It is correct that models fit to the same datasets can be directly compared using the AIC values. Models 2 and 3 are not fit to the same datasets, however, due to the way Model 3 is initialized. To initialize the latent states of Model 3, Lee et al [1] used approximately 60% of the available data as "burn-in" observations to initialize latent states. According to their supplement, this was done primarily because they believed the dynamics from the start of the outbreak were significantly different enough from the dynamics at the end of the outbreak to suggest that two seperate models would be needed to appropriately describe the data. We feel that a properly designed and calibrated model should be able to explain the entire dynamics of the outbreak, including the change from what Lee et al [1] referred to as the "epidemic" and "endemic" periods of the outbreak.

One possible approach to re-analyzing their model would then be to fit all available observations; this would make Models 2 and 3 directly comparable, but would require a significant change to Model 3 itself, as the initialization model used by Lee et al [1] requires a "burn-in" period. One of our goals, however, was to demonstrate that the models provided by Lee et al. [1] were not the issue, and a complete refactorization of the initialization could potentially contradict this objective.

Therefore instead we decided to substantially shrink the “burn-in” period to only the first three available observations, which corresponds to a natural change in the observations themselves [10]. These details are available in the Initial Values section in the supplement material.

In short, though models that are fit to the same dataset are directly comparable using AIC, this is not the case for Models 2 and 3, as the first few observations in the dataset were not considered likelihood evalution of Model 3. This also explains the slight discrepency in AIC for the benchmark models of these two models (Table 2). Without going into too much detail in the main article, we added the following sentence to clarify this point:

> *Models that are fit to the same datasets can be directly compared using AIC values, making it a useful tool to compare to benchmark models. Though Models 2 and 3 are both fit to department level incidence reports, their AIC values are not directly comparable due to the way Model 3 initializes latent states, which is described in detail in Sec. S5 of the supplement.*

> *9. Lines 308-309: Isn’t this a log-linear trend in the transmission rate?*

Yes, that is correct. This has been fixed in the revision. In other instances where the linear trend was previously mentioned, we simply refer to it now as a trend in transmision rate, unless the additional detail of the trend being log-linear is needed for clarity.

> *10. While the data support a linear trend at the 95% confidence level, it’s worth commenting on the magnitude of the trend, e.g. the total reduction in $\beta(t)$ over the period of study.*

Thank you for this suggestion, a sentence on the overall effect of the parameter estimate has been included. For reference, the maxmum likelihood estimate of the parameter corresponds with a 7.3% reduction to the transmission rate over the course of the outbreak, with a 95% confidence interval $(1.8\%, 17.9\%)$ for the overall reduction in transmission.

> *11. I might have misunderstood it slightly, but I don’t entirely follow the reasoning of lines 340-346. It seems that two different claims are being made: A) the model can’t identify which mechanism underlies a trend in $\beta(t)$ and B) the model can’t definitively state there is a trend in $\beta(t)$ compared with some other time-varying parameter (e.g. reporting probability). For model 1, I agree with the authors that claims A and B are valid. What I take issue with is the final statement, which seems to be stronger and apply to both claims and mechanistic models more generally: “we argue that a decreasing transmission rate is a plausible way to explain this, but the incidence data themselves do not provide enough information to pin down the mechanism.” I agree that the “incidence data themselves” (which I take to imply “without additional covariates data”) make claim A valid regardless of the model. However, I don’t think claim B holds regardless of the model. By changing model 1 to include time variation in the reporting rate I don’t see* a priori *why the “the incidence data themselves” might not provide enough information to distinguish between time variation in transmission compared with time variation in reporting without the need for additional data. The deficiency is in model 1 rather than the information content of the data. I suggest the authors rework the paragraph to make the reasoning clearer.*

Our views align with the referee’s comments, and we have rewritten the article to make this clearer. The revised paragraph is now:

If a mechanistic model including a feature (such as a representation of a mechanism, or the inclusion of a covariate) fits better than mechanistic models without that feature, and also has competitive fit compared to associative benchmarks, this may be taken as evidence supporting the scientific relevance of the feature. As for any analysis of observational data, we must be alert to the possibility of confounding. For a covariate, this shows up in a similar way to regression analysis: the covariate under investigation could be a proxy for some other unmodeled or unmeasured covariate. For a mechanism, the model feature could in principle explain the data by helping to account for some different unmodeled phenomenon.

The statistical evidence of a trend in transmission rate in this model could be explained by any trending variable (such as hygiene improvements, or changes in population behavior), resulting in confounding from collinear covariates. Because we only have access to incidence data, the decrease in observed cases over the course of the outbreak may alternatively be attributed to a trend in other model parameters, such as a decrease in the reporting rate over time. This results in confounded mechanisms, as one may not formally test all possible ways to include a trending parameter. In such a case, the robust statistical conclusion is that a model which allows for change fits better than one which does not. Because the incidence data themselves may not provide enough information to pin down the mechanism, a strongly supported conclusion should avoid ruling out untested hypotheses. In this case study, we did not explicitly test if the incidence data favor a trend in transmission over a trend in reporting rate. We argue that a decreasing transmission rate is a plausible way to explain the decrease in cases over time, as there is alternative evidence that supports this model [11–13]. Additionally, there is little scientific evidence to support a decrease in reporting rate (for example, see Figure 1 of [12]).

12. Lines 553-556: I push back against this point. The models assume an exponential distribution for vaccine-derived immunity. Assuming a mean duration of 10 years, the proportion of individuals who remain immune 9 years after vaccination is . Even for much larger durations of immunity a non-negligible fraction of the population will lose immunity after 9 years, e.g. $e^{-9/10} \approx 0.41$*. I therefore don't expect values for the duration of immunity around 10 years to "effectively result in the same model dynamics".*

The qualitative comparison we had in mind was the regime with short immunity where many people become infected multiple times during the ten year epidemic, allowing multiple epidemic peaks due to susceptible replenishment. We have reworded the sentence to avoid making too strong a claim about 10yr immunity being a precise point where the dynamics transition to a regime where reinfection events become negligible. The revised paragraph is provided below.

Our case study provides an example of this in the parameter estimate for the duration of natural immunity due to cholera infection, μ_{RS}^{-1}. Under the framework of Model 2, the best estimate for this parameter is 1.4×10^{11} yr, suggesting that individuals have effectively permanent immunity to cholera once infected. Rather than interpreting this as scientific evidence that individuals have permanent immunity from cholera, this result suggests that Model 2 favors a regime where reinfection events are a negligble part of the dynamics. The depletion of susceptible individuals may be attributed to confounding mechanisms—such as localized vaccination programs and non-pharmaceutical interventions that reduce cholera transmission [12, 14]—that were not accounted for in the model. Perhaps the best interpretation of the estimated parameter, then, is that under the modeling framework that was used, the model most adequately describes the observed data by having a steady decrease in the number of susceptible individuals. The weak statistical fit of Model 2 compared to a log-linear benchmark (see Table 2) cautions us against drawing quantitative conclusions from this model. A model that has a poor statistical fit may nevertheless provide a useful conceptual framework for thinking about the system under investigation. However, a claim that the model has been validated against data should be reserved for situations where the model provides a statistical fit that is competitive against alternative explanations.

Discussion

13. In contrast with the introduction, which almost exclusively focuses on cholera in Haiti, the discussion only mentions it once, briefly: "We used the same data and models, and even much of the same code, as Lee et al. [1], and yet ended up with drastically different conclusions." It would help the reader to have those discrepancies summarised. It would be helpful to summarise the recommendations the authors make attempting to fit mechanistic models to time series data (even as a list). Along similar lines, it would be helpful to present a more concise summary of what the "more accurate policy evaluations" found by using the approaches outlined in this study.

The revised version of the paper contains significant edits to the introduction and discussion sections that we feel address this concern.

Typos

Eq. 1: Is there a missing star on $y_{1:N}$ on the right-hand-side?

Yes, there should be an asterisk. This has been fixed.

Line 146: delete "the" before describing

Fixed.

Referee 2

> *This well-written article is a strong piece of work that will be useful to the readers of this journal and, in particular, to researchers who perform statistical inference on compartmental models of infectious diseases. Furthermore, this paper is enhanced by its transparency and reproducibility, as the provided code is well-documented and organised in an R package. Additionally, the supplementary information is a valuable resource for researchers in this field.*
>
> *In a nutshell, the authors argue that existing criteria to evaluate the validity of a disease model are insufficient. Therefore, they propose more stringent standards for evaluating models' ability to fit the available data in order to obtain more reliable forecasts. The authors use a Cholera case study to outline their suggestions. To highlight, a key contribution from this work is the recommendation of employing inductive (associative) models as a goodness-of-fit benchmark, as evidenced by this sentence: "It should be universal practice to present measures of goodness of fit for published models, and mechanistic models should be compared against benchmarks". Undoubtedly, this approach provides an objective measure to judge the ability of mechanistic models to fit the data. In the following sections, I express my opinion on how this paper may be improved.*
>
> ***Major Concerns***
>
> ***1. Literature***
> *While the arguments provided throughout the article are well-articulated, there needs to be more supporting literature at the beginning of major sections. For instance, I don't need to be convinced that the structure of a model should be based on a realistic theory about the observed phenomenon; namely, models should be a white box. However, not everyone is on board with this premise, and supporting literature that argues in favour of this approach should be mentioned. In short, more citations should be added at the beginning of each major section.*

The referee's suggestions have led to the addition of various references in the revision. For the particular issue of black box versus white box models, we have included the following paragraph

> *Mechanistic models representing biological phenomena are valuable for epidemiology and consequently for public health policy [15, 16]. More broadly, they have useful roles throughout biology, expecially when combined with statistical methods that properly account for stochasticity and nonlinearity [17]. In some situations, modern machine learning methods can outperform mechanistic models on epidemiological forecasting tasks [18, 19]. The predictive skill of non-mechansitic models can reveal limitations in mechanistic models, but cannot readily replace the scientific understanding obtained by describing the biological dynamics of the system in a mathematical model [19].*

This addition is intended to compliment the discussion about mechanistic (white-box) vs associative (black-box) models that was provided at the start of the "Materials and methods" section.

2. Limitations

In various passages of this paper, it is hinted that modellers should refrain from deterministic models and instead opt for more realistic stochastic representations. While the critique of ODE models is valid, the shift to stochastic structures is not a free choice. For example, introducing extra-demographic variability adds one additional parameter (infinitesimal variance). In ODE models, one extra parameter can lead to unidentifiability, and there's no apparent reason why this would differ in a stochastic version. Moreover, transitioning to stochastic models involves abandoning well-established MCMC algorithms in favour of methods still in development. Therefore, modellers should not assume that more realistic models with additional parameters are necessarily better without proper caution. In my experience, diagnosing unidentifiability is easier in ODE models than in POMP structures. Hence, there is a trade-off between benefits and costs.

The referee raises several points in this paragraph, and we consider them one at a time.

We agree that adding to model complexity has costs, including loss of identifabiility. AIC and other formal statistical methods can assess whether the additional complexity is justified by a sufficiently large improvement of statistical fit. Identifiability and model simplicity have practical scientific value which goes beyond their statistical role in reducing generalization error and hence improving out-of-fit predictions. This is a variation on the issue of "statistical significance versus practical significance" familiar from introductory statistics textbooks.

Moving from an ODE model to a stochastic model may not involve any additional parameters, if only demographic stochasticity is included. Generally, it is advisable to consider the possibility of over-dispersed variation on the transmission rate [20, 21] leading to one extra parameter compared to a deterministic model. In the context of cholera dynamics, [22] found a substantial improvement in model fit by including stochasticity, involving one over-dispersion parameter. [23] found similar results, favoring a stochastic dynamic model with over-dispersion, when investigating transmission dynamics of a different enteric pathogen, rotavirus. In other contexts, [24] noted various scientific advantages for including stochasticity in the model; it is undoubtedly part of the biological dynamics, and its omission can lead to over-confident predictions.

In our study, we find that the deterministic model (model 2) has poor statistical fit. This can lead to a risk of identifying spurious scientific phenomena: additional deterministic complexity added to the model may help to describe unmodeled stochastic phenomena present in the data, and may therefore spuriously be assessed as a highly statistically significant model improvement. We see some evidence of this in our analysis. Aside from not being stochastic, the mechanistic structure of Model 2 is far more complex than the other two models considered in our study; we also see the most clear cases of spurious scientific findings, as some estimated parameters favor regions of the parameter space that we would not expect scientifically (see the added supplement material for confidence intervals of model parameters). This may partially explain the choice of Lee et al [1] to consider a scientifically unjustifiable jump in latent state values from their epidemic to endemic periods: without this additional flexibility, Model 2 cannot adequately describe cholera dynamics in Haiti while keeping parameter values in scientifically plausible ranges. From our experience, this type of a result is quite common: for a deterministic model to adequately account for unmodeled features of a dynamic system, parameters that are estimated are often forced to unexpected regions of the parameter space. Adding stochasticity to the model helps address this issue by capturing unmodeled features in the inherent stochasticity of the model rather than in estimated parameters.

Stochasticity in the model alone may not fully solve this problem as evidence suggests that the failure to account for extra-demographic variablility may lead to biases in parameter estimates [21]. Our profile likelihood results for this case study also provide strong evidence for the inclusion of extra-demographic stochasticity in Models 1 and 3 (see supplement material section S8).

The decision on whether to choose Markov chain Monte Carlo (MCMC) methods is not fully aligned with the decision on whether to include stochasticity in the model. In our literature review, of the 12 previous Haiti cholera investigations including stochastic dynamic models [1,25–34], four of them carried out inference via MCMC [26, 29, 32, 33]. We agree with the referee that MCMC is more routine for determinstic models than for stochastic models, and advanced MCMC methods such as Hamiltonian Monte Carlo become more readily applicable. Sequential Monte Carlo (SMC) was used for only two of the investigations in our literature review [1, 34]. In principle, MCMC methods can be applied for inference in over-dispersed stochastic dynamic models via so-called particle MCMC (PMCMC) [35]. In practice, these models have been fitted by maximum likelihood, some examples including [22, 23, 36–39]. In this case, identifiability is generally addressed via profile likelihood methods.

We acknowledge the existence of useful mathematical tools for studying ODE models, their identifiability and stability and limiting properties. Perhaps one example demonstrating the usefulness of ODE models is the Lee et al. [1] study: of the four models considered, the forecasts from Model 2 were most consistent with the actual cholera situation in Haiti. Despite their inability to describe incidence data, deterministic models can do a good job at detecting overal trends, like the gradual decline of cholera cases in Haiti. This, coupled with the ease in fitting ODEs to data, makes ODEs a useful tool to gain preliminary insight on a dynamic system. Furthermore, stochastic dynamic models can be reduced to a deterministic model (a so-called determinsitic skeleton) which can usefully be analyzed to help understand the stochastic dynamics [40]. We therefore encourage the use of ODE analysis tools whether or not a stochastic model is adopted. We have not followed that approach in the current manuscript.

In summary, maximum likelihood (ML) inference via SMC nowadays provides a fairly well estabilished alternative to Bayesian MCMC inference. ML appears to be methodologically more convenient for stochastic dynamic models with overdispersion. When analyzing long epidemiological datasets, this stochasticity appears to be important for statistical model fit. That alone is not enough to show that models with overdispersed stochastic dynamics are scientifically preferable. However, discovery of a model feature which generates a large improvement in statistcal fit from adding one more parameter is typically treated as scientific evidence favoring the inclusion of that feature in a scientific model for the phenomenon.

> *Moreover, Monte Carlo methods, such as the Particle Filter and, by extension, Iterated Filtering, aim to approximate integrals (the posterior or filtering distributions). However, one cannot take for granted that these methods provide accurate descriptions of these targets without proper validation. In more 'traditional' MCMC methods, diagnostics like the potential scale reduction factor and effective sample size play a crucial role in the inference process. Unsatisfactory values of these diagnostics render inferences unreliable, often necessitating model reformulation. In contrast, in the literature on POMP models (including this paper), diagnostics are tangentially mentioned. Users (like me) sometimes face uncertainty about whether the lack of fit is due to model misspecification or problems with the Monte Carlo algorithm exploring the parameter space.*

We agree with the referee that diagnostics are an important aspect of careful data analysis. The most widely used diagnostic tools for particle filter and iterated particle filter methods probably match those provided by default in the `pomp` R package [41] when objects generated by `pfilter()` or `mif2()` are plotted. For filtering, one tracks the effective sample size, the conditional log likelihood for each observation, and the values of the filtered states. For iterated filtering, one checks the filtering diagnostics for the last iteration and also how the parameter estimates and the log-likelihood estimate evolve through iterations, typically superimposing the results of 10-100 independent searches.

The particle filter provides access to the likelihood of the data, and conditional likelihood of each individual data point. By contrast, obtaining reliable likelihood values from MCMC is not straightforward. Thus, particle filter methods facilitate the use of various likelihood-based methods for diagnostics and model selection. We demonstrate this in Figs S4, S5, S6, S7. [TODO: Check all figure numbers, especially to the supplement.]

As a proper scoring rule [42], log-likelihood can be used as an effective way to evaluate filters as well as models. The particle filter is unbiased for the likelihood, and therefore negatively biased for the log-likelihood with bias approaching zero as the number of particles increases and the Monte Carlo variance decreases. If optimization searches from diverse starting points provide a consensus on the maximum log-likelihood, and the Monte Carlo variance is small, we can have reasonable confidence in our maximization. Monte Carlo adjusted profile (MCAP) methods [43,44] provide methodology which reduces and evaluates the consequences of this Monte Carlo error. MCAP methods have been used in various recent investigations using particle filters [38,45,46].

We conclude that suitable diagnostic methods are available for the POMP methods we employ, and some of them are demonstrated and discussed in our paper. We would like to do what we can, within the constraints of this paper, to help build awareness of these diagnostic methods and ways in which they can be useful. A full exposition of this topic is out of the scope of this paper.

[This text below is something that Ed added. I think it's a good idea. However, simply because of the overal length of the paper and total number of citations, I think it's probably okay to just leave this out:] Howerver, we can alert the reader to the topic, and provide some additional references. Not infrequently, diagnostic issues are relegated to supplementary material and so we have indicated to interested readers some papers with extensive supplements demonstrating these diagnostic issues in action.

> *For example, in Figure 5, department Ouest exhibits substantial uncertainty from 2014 onwards, and this figure is on a log scale. Essentially, the inference suggests that 'anything can happen'. I would like to pinpoint the nature of this collapse in uncertainty. Identifiability issues might be at play, given the possibility of more estimated parameters than the incidence data can inform. Sometimes, we ask too much from the data. Observe the discrepancy in trends between the average behaviour and the uncertainty ribbons.*
>
> *In summary, the authors should elaborate on the limitations of the proposed approach.*

We agree with the referee that uncertainty quantification and evaluation of identifiability are fundamental tasks for model-based data analysis. In the revision, we have included profile confidence intervals on all parameters, except those taking separate values on each unit. For these unit-specific parameters, the spread of values across units provides an alternative measure of uncertainty. We find that all parameters are identifiable, though some are more tightly estimable than others.

We had not done this in the original submission since it was only tangentially relevant to the comparison with [1] that lies at the center of our paper. However, these additional results are worthwhile extension to our analysis and we are grateful for the encouragement to include it.

Wide prediction intervals are not necessarily a sign of a problem. For example, nobody knew for sure whether cholera was banished from Haiti after it was declared eliminated in February 2022. Cholera reemerged in September 2022, but a reasonable model might be expected to cover the possiblity of this reemergence either happening or not happening, leading to wide prediction intervals. It is also critical to recognize that the simulations in Figure 5 are obtained from initial conditions, effectively making simulations at the end of the time series (January 2019) 8 year forecasts from the starting point in November 2010. The 95% confidence interval for Ouest in this figure is from $0 - 1094$ reported cases. This range may be considered small given the inherent demographic and environmental stochasticity present in the dynamic system and all of the many possible outcomes of the outbreak from its initial state in 2010.

> ***3. Conclusions***
>
> *I find that the conclusions are somewhat disconnected from the introduction and abstract, which state that the paper presents a methodology to diagnose model misspecification, develop alternative models, and make computational improvements. It would improve readability to include a summary in this section. Specifically, link each contribution to a particular example. For instance, in the case of model 1, computational improvements increased the log-likelihood. In short, connect the findings more explicitly to the research question and stated goals.*

We have added several paragraphs in the discussion to address this point. These paragraphs are listed below for reference.

> *In our cholera case study, we highlight the importance of model diagnostics; they are instrumental in enhancing model fits, thus leading to more trustworthy conclusions. We compared the Models 1–3 of [1] to simple statistical benchmarks, enabling us to identify deficiencies in the models that could be improved. In the case of Model 3, comparing a fitted model to a benchmark and then following up by questioning why the model did not describe the data as efficiently as its benchmark helped us discern that the model fell short in explaining the surge in cholera transmission following Hurricane Matthew. This led us to suggest a model modification. Because Model 1 outperformed its benchmark model, we were able to confidently scrutinize the implications of the fitted model. We found that the seasonality term of Model 1 closely mirrors the seasonal rainfall patterns in Haiti, reaffirming alternative evidence of the importance of rainfall as a driver of cholera infection and bolstering our confidence in Model 1's ability to capture the critical dynamics of the cholera outbreak.*
>
> *Likelihood based inference enables researchers to choose between any model that is evaluated on the same dataset using a metric like AIC. Nested model variations are particularly useful as they enable formal statistical testing of the nested features. The cholera case study demonstrated the advantage of such a model variation in all three models. For instance, we tested the inclusion of a trend in transmission rate over time in Model 1; in Model 2, we considered adding an additional phase parameter, which enables a shift in the seasonal peaks of cholera infections; in Model 3, we included additional parameters such as those associated with Hurricane Matthew. In all these cases, we found strong statistical evidence in favor of these model variations over their simpler counterparts.*

Another important consideration when fitting a mechanistic model to a dynamic system is the complexity of the model. Mathematical models necessarily simplify complex dynamical systems in order to make inference possible. Decision-making about which features of a system need to be incorporated in a model and which should be excluded can significantly influence the model-based conclusions. However, these decisions can be tested using a likelihood-based framework. Standard model decisions cover deterministic versus stochastic modeling and the decision to include or exclude a meta-population structure. Model benchmarks can assist in determining the level of complexity to be used in the model, enabling researchers to gauge a model's relative performance against a better-understood statistical model.

Unmodeled features of a dynamic system can lead to spurious or misleading parameter estimates if the feature greatly impacts observed data. This is because the model must compensate for these unaccounted-for features using its inherent flexibility, particularly in the calibration of parameters. In deterministic models, parameter estimation is the sole source of such flexibility. When the influence of unmodeled dynamics is significant, it can skew parameter estimates toward scientifically implausible values, a phenomenon observed with Model 2 in our case study. Incorporating demographic and environmental stochasticity into models can mitigate the impact of these unmodeled features. These stochastic processes are not only arguably present in most dynamic systems, but their inclusion in a model also allows observed data variations to be attributed to inherent randomness rather than to distorted parameter values. Furthermore, incorporating extra-demographic stochasticity in a model has also been shown to be beneficial [9, 21, 23], which aligns with our findings: Models 1 and 3 suggest the presence of extra-demographic stochsticity, as evidenced by the confidence intervals provided in Suplement S8.

The three models in this case study offer a unique combination of decisions of model complexity. Model 3 provides an interesting example in that it is both stochastic and has a meta-population structure, making it challenging to draw likelihood-based inferences. In this paper, we demonstrated how this model class could be calibrated to incidence data using the innovative IBPF algorithm. One of only a few examples of fitting a model with this level of complexity, this case study exemplifies the algorithm's potential benefits and provides an example for future researchers on a possible approach to fitting a high-dimensional non-linear model.

If forecasts are an important component of a modeling task, the most recent information on the dynamic system should be considered with increased importance. This is particularly true for long time series, as the latent states from simulations from initial conditions may be more likely to diverge from the truth as time increases. One way to address this issue is by simulating forward from the filtering distribution at the last available time point, as this proceedure considers available evidence about the latent states of a system under the framework of the dynamic model. This type of forecasting, however, is not possible using a deterministic model. Deterministic simulations may also result in over-confidence in model forecasts, as they can only account for uncertainty due to the parameter estimation proceedure [24]. Despite these limitations, forecasts from deterministic models can provide valuable insights on potential future trends, though researchers should be transparent about these limitations when presenting forecasts from deterministic models.

Minor Comments:

1. In the author summary, this part is hard to follow: "and provides careful justification of valid conclusions from the fitted model. Objective measures are used to benchmark model fit; when these are combined with reproducibility, a framework emerges for continual improvement when revisiting the data and models." Please rephrase.

The previous statement has been changed to the following:

> *Our analysis presents methodology for diagnosing how well a model describes observed data. Objective measures are used to evaluate model fit; when objective measures of the goodness-of-fit of a model are combined with reproducibility, a framework emerges for continual improvement to model based inference when revisiting the data. This framework promotes scientific discovery and bolsters model-based policy recommendations by simplifying the process for researchers to build off of previous results.*

2. Lines 1-8. Please add more citations.

Citations have been added.

3. Line 36. Can you be explicit about what the forecasts predicted? Did the models predict a rise in cases?

This statement has been modified to include more details. It now reads:

> *Despite their autonomy, the four independent teams obtained a consensus that an extensive nationwide vaccination campaign would be necessary to eliminate cholera from Haiti. They also provided a range for the forecasted number of cumulative cases from February 2019 to Februaruy 2024 based on model simulations, with median estimates from each model measured in hundreds of thousands. These forecasts are inconsistent with the prolonged period with no confirmed cholera cases between February, 2019 and September, 2022 [14].*

4. Line 42. Add hyphen: "Model-based conclusions".

A hyphen has been added for all instances of the phrase: "model-based".

5. Lines 67-78. Add more citations.

It is difficult to directly add citations as suggested, as the text that was contained in lines 67-78 was original thought or argued in [47], which was already cited. However, we appreciate the point that an improved discussion about the role of mechanistic vs associative modeling would be beneficial, and would be strengthened by other opinions / evidence on the topic.

To that effect, we have modified the original text to be more clear that the definitions of *associative* versus *mechanistic* models are our own, as the precise difference between these may be different in alternative contexts. Furthermore, we provide references from our literature review that demonstrate the usefulness of associative models in some cases, whereas we typically focus on the benefits of mechanistic models in the article. Finally, we added a paragraph that includes a discussion and references on mechanistic / associative modeling, which also addresses the concerns of Reviewer 1. This paragraph contains several citations, and is included below for reference.

> *Mechanistic models representing biological phenomena are valuable for epidemiology and consequently for public health policy [15, 16]. More broadly, they have useful roles throughout biology, expecially when combined with statistical methods that properly account for stochasticity and nonlinearity [17]. In some situations, modern machine learning methods can outperform mechanistic models on epidemiological forecasting tasks [18, 19]. The predictive skill of non-mechansitic models can reveal limitations in mechanistic models, but cannot readily replace the scientific understanding obtained by describing the biological dynamics of the system in a mathematical model [19].*

> *6. Please add uncertainty intervals to the estimated parameters in Table 1. If necessary, consider splitting the table into two or including this additional information in the supplementary material. I suggest this update because there is an indication of unidentifiability when parameter estimates are fairly broad.*

Due to the size and complexity of the table we have elected to put uncertainty estimates elsewhere. Specifically, have created an additional supplement material document that outlines how confidence intervals are obtained, along with tables and figures that show parameter uncertainty estimates. In the caption for Table 1, we include a reference to this supplement material.

> *7. In line 159, it is stated that "$\upsilon^\star(t)$ is efficacy at time t since vaccination for adults" and then "single and double vaccine doses were modeled by changing the waning of protection; protection was modeled as equal between single and double dose until 52 weeks after vaccination, at which point the single dose becomes ineffective". I examined the reference from which this function is based but found only a numeric table. Please provide the equation of this function or a detailed description in the supplementary information. It would be beneficial for readers to understand how to model this complex feature.*

The table you found is the function $\upsilon^\star(t)$, as that is what is used by each of their models. Lee et al [1] provide a smoothed version of this function in Figure 2 (B), but the actual function implemented by each of their models (and consequently, our models) is the numeric table itself; in this sense, the smoothed function they show in Figure 2 (B) is an approximation of the true vaccination function that they use, which is provided in Table S4 in their supplement material.

According to Lee et al [1], this function was obtained by "fitting a log-linear weighted regression model to the raw data from a published meta-analysis on killed OCV efficacy against medically attended culter-confirmed cholera (figure 2, appendix 3 p 6) [48]". Aside from this statement, they provided limited details on how they came up with this function. Because of this, we unfortunately can't provide much more additional information other than the statement that was already included in our article:

> *$\upsilon^\star(t)$ is efficacy at time t since vaccination for adults, taken from [1], Table S4, ...*

> *8. Lines 209-211. In model 2, an incidence measurement is employed to configure a prevalence compartment. As the authors may know, initial values severely condition the dynamics of a model. Please explain this decision. Is it because that was the approach followed in the original formulation (Lee at al's paper)?*

Our starting point was the formulation of [1]. We did consider the effect of not calibrating initial values. Unsurprisingly, fitting initial values results in a quantitative improvement to model-fit, measured by AIC, but we also found that this had negligable qualitative effects on model-based conclusions. On a nationally aggregated scale, the total number of initial infectious individuals in the fixed scenario was 36500, and 33480 when the initial states were estimated. We have included additional material in the supplement for Initialization Models that demonstrates this.

> *9. Lines 245-247. Same comment as before. What's the justification for assuming incidence measurements as the basis for prevalence states? What are the risks?*

Our experience with mechanistic modeling reflects that of the reviewer: calibrating initial values often results in improved dynamics. As such, we would recommend calibrating initial values in most cases; here, doing so may constitute a major change from the initialization model used by Lee et al. [1]. Because one of our goals was to demonstrate that the models of [1] were alone not responsible for their poor forecasts, we elected to use the same initialization approach to the extent possible.

The goals of inferring initial prevalence is to reduce the number of parameters. The risk is that the form of this assumption could have substantial effects on the dynamics, and hence on the consequences of the analysis. In our supplement material, we described our approach to addressing this issue: (i) looking at conditional log-likelihoods to assess time points at which the model is misspecified; (ii) if the initial values seem problematic, calculate the room for improvement by estimating the initial values as free parameters. We found that enforcing model dynamics for initial states worked well for many of the departments, but was problematic for departments where this resulted in zero counts in Infected, Asymptomatic, and Bacterial compartments, leading to the estimation of initial counts of Infected individuals in two departments (Grand'Anse and Nippes). Initial values for Asymptomatic and Bacterial compartments were then obtained by enforcing the dynamics of the model using the estimates for Infected individuals, as described in the supplement material.

> *10. Line 262. "deterministic Model 2 is a degenerate case of a stochastic model". Please explain why or provide a reference.*

The text has been modified to add additional clarity. The text now reads:

> *deterministic Model 2 is a special case of a Markov processes solving a stochastic differential equation in the limit as the noise parameter goes to zero.*

> *11. Lines 260-274. Please add more citations.*

We have added a citation to [41], which discusses POMP models and their implementation.

> *12. Lines 343-345. This sentence is key for this paper, but it's hard to follow: "The robust statistical conclusion is that a model which allows for change fits better than one which does not—we argue that a decreasing transmission rate is a plausible way to explain this, but the incidence data themselves do not provide enough information to pin down the mechanism". Please rephrase.*

This suggestion is similar to one posed by Reviewer 1. We have made substantial edits to the paragraph that included this statement that clarifies our point:

> *The statistical evidence of a trend in transmission rate in this model could be explained by any trending variable (such as hygiene improvements, or changes in population behavior), resulting in confounding from collinear covariates. Because we only have access to incidence data, the decrease in observed cases over the course of the outbreak may alternatively be attributed to a trend in other model parameters, such as a decrease in the reporting rate over time. This results in confounded mechanisms, as one may not formally test all possible ways to include a trending parameter. In such a case, the robust statistical conclusion is that a model which allows for change fits better than one which does not. Because the incidence data themselves may not provide enough information to pin down the mechanism, a strongly supported conclusion should avoid ruling out untested hypotheses. In this case study, we did not explicitly test if the incidence data favor a trend in transmission over a trend in reporting rate. We argue that a decreasing transmission rate is a plausible way to explain the decrease in cases over time, as there is alternative evidence that supports this model [11–13]. Additionally, there is little scientific evidence to support a decrease in reporting rate (for example, see Figure 1 of [12]).*

> *13. Line 357. "Determining the necessary computational effort needed to maximize model likelihoods and acting accordingly" How do we determine the necessary computational effort?*

We have added the following text to clarify this point:

> *The large increase in the log-likelihood for Model 1 (see Table 2) can primarily be attributed to increasing the computational effort in fitting the model. This result highlights the importance of carefully determining the necessary computational effort needed to maximize model likelihoods and acting accordingly. In this case study, this was done by performing standard diagnostics for the IF2 and particle filter algorithms [41]. Given the considerable computational costs of simulation-based algorithms, we find it useful to perform an initial assessment using hyperparameter values that enable relatively quick calculations. The insights obtained from this preliminary analysis help in accurately determining the amount of computation that is required to achieve reliable outcomes.*

> *14. Please add the predicted intervals to Fig 4. I would like to see the effect of the log-normal measurement model.*

We intentionally do not include predictive intervals in Fig 4. There are several reasons for this, one of which is mentioned in the article:

> *In this analysis, we do not construct forecasts accounting for parameter uncertainty as our focus is on the estimation and diagnosis of mechanistic models. Furthermore, we use the projections from a single point estimate to highlight the deficiency of deterministic models that the only variability in model projections is a result of parameter uncertainty, which can lead to over-confidence in forecasts [24].*

[Below are some thoughts I had. We don't have to include all / any of this, but I thought it was something to think about on the topic of parameter uncertainty]

Parameter uncertainty in model forecasts is an important topic, one that we would have liked to have addressed in our article but felt that we could not since the article is already becoming longer than one would expect in this journal. In our literature review of dynamic models for cholera in Haiti, and our experience with literature with dynamic modeling outside of this problem, the most common approaches to accounting for parameter uncertainty are the following:

- Maximum likelihood etimation theory states that under certain conditions, the difference between the MLE and the generating model parameter is asymptoticly normally distributed with mean zero and variance determined by the inverse Fisher information matrix. Many numeric optimization techniques often approximate the curvature of the likelihood surface, thereby Fisher information matrix. This can then be used to obtain approximate confidence intervals using the asymptotic distribution of the MLE and the approximation to Fisher's information matrix. While mathematically and computationally convinient, the validity of these confidence intervals depends on asymptotics and assumptions that are often not appropriate for non-linear models. In fact, there is evidence that these types of confidence intervals are often far too narrow even in the most idealistic cases for dynamic models: linear Gaussian models fit with a large number of observations [49]. Because of this, confidence regions created in this way likely too narrow, resulting in over-confidence in parameter estimates.

- Alternatively, it is common to account for parameter uncertainty by performing a sensitivity analysis. This is often done when parameters were selected via non-likelihood based approaches, but our review contained instances (e.g. [1]) where this was done even if parameters were calibrated by maximizing likelihoods. To perform a sensitivity anaylsis, researchers typically pick ranges of parameter estimates, creating a hypercube from which they obtain sample parameter values. The bounds of this hypercube, however, are often obtained using approximations of Fisher's Information Matrix (the deficiencies of which are mentioned above), or by some non-data driven approach. For example, bounds are often selected by arbitrarily adding / subtracting a percentage of the estimated parameter (where the bounds can be either larger or smaller than what should be used), or by using parameter estimates from other studies (in which case the parameter estimates may not make sense in the context of the model and data being considered, again resulting in intervals that may be arbitrarily wide or narrow). A persistent theme with all of the above approaches is that they enable researchers to effectively pick the level of uncertainty to present in their reports.

- Sampling parameters from a Bayesian posterior. For many researchers, this may be one of the primary reasons they prefer Baysian methods over maximum likelihood estimates, as these methods do have an advantage in accounting for parameter uncertainty. Bayesian methods to estimate model parameters may not be possible for all scientific models of interest. We have been made aware of recent work that demonstrates cases where the authors would like to have a posterior distribution for their model, but the scientific model they propose is unable to be calibrated using state-of-the-art Bayesian approaches; they therefore use IF2 to obtain parameter estimates that result in good behaviour of particle filters, and then arbitrarily select a uniform distribution centered around the MLE as a prior distribution to Bayesian methods (e.g. [50, 51]). This approach is related to Empirical Bayes, where the prior distribution is selected using information in the available data. We propose something similar in our supplement material, but our proposition does not require additional runs of pMCMC which can be computationally expensive. One drawback of our proposed Empirical Bayes approach is that when the Monte Carlo variance in likelihood evaluation is high, it is possible that the posterior mass function has the majority of its mass assigned to only a few parameter sets.

A full investigation / discussion on parameter uncertainty is outside of the scope of this article.

> *15. Lines 513-515. Please clarify the comparison between disaggregated models and a benchmark. Let's say I have spatial units 1 and 2, for which I have observations y1 and y2. Should I fit the disaggregated mechanistic model to y1 and y2 simultaneously (as usual) but keep a record of the individual log-likelihoods (log-lik y1 and log-lik y2). In parallel, fit the benchmark independently to y1 and y2, and then compare by log-likelihoods or information criteria by spatial unit.*

Thank you for this suggestion. We have added the following details to help clarify how the benchmarks for the spatially explicit models were created, and how to appropriately compare model likelihoods to these benchmarks:

> *To obtain a benchmark for models with a meta-population structure, we fit independent auto-regressive negative binomial models to each spatial unit. Under the assumption of independence, the log-likelihood of the benchmark on the entire collection of data can be obtained by summing up the log-likelihood for each independent model. In general, a spatially explicit model may not have well-defined individual log-likelihoods, and therefore comparisons to benchmarks should be made at the level of the joint model.*

The benchmarks we have looked at operate independently on each unit, and in this case the joint log-likelihood benchmark for units 1 and 2 is the sum of the marginal log-likelihoods. A simple statistical benchmark does not need to have this structure; for example, one could consider an autoregressive model including spatial dependence.

In general, a joint model for units 1 and 2 may not have well-defined individual log-likelihoods log-lik y1 and log-lik y2 that sum to the full likelihood. An exception arises in the case of likelihood evaluation by the block particle filter, in which conditinal log likelihoods for each block of units do sum to the total log likelihood estimate. If the block approximation is found to be appropriate [52, 53] then fits can be compared against an independent benchmark for each block, and then summed to compare on the entire dataset.

[I think it may be easier to just add an additional sentence to the manuscript.] Since the value of benchmarks is a major point in this paper, we have added these extra details in a supplementary section.

> *16. Line 528. Please add hyphen: "Model-based inference".*

A hyphen has been added for all instances of the phrase: "model-based".

> *17. Lines 576-577. "We notice that the calibrated model favors higher levels of cholera transmission than what was typically observed in the incidence data (S5 Text)". After this fragment, please summarise in one or two sentences what it will be found in S5 Text.*

[I'm not exactly sure how to respond to this point. The relevant information about what can be found in S5 text is precisly what is already stated: "by reconstructing the latent states of Model 3, we notice that the calibrated model favors higher levels of cholera transmission than what was typically observed in the incidence data (S5 Text)".]

[Maybe we could edit this part to the following:]

~~In the context of our case study, by reconstructing the latent states of Model 3, we notice that the~~

~~calibrated model favors higher levels of cholera transmission than what was typically observed in the incidence data (S5 Text).~~

In the supplement material (S5 Text), we explore in more detail the process of model fitting and diagnostics. Here we demonstrate that although Model 3 does outperform its benchmark model on the aggregate scale, the model performance compared to the benchmark on units with the most cholera cases is poor. Furthermore, by comparing simulations from the fitted model to the filtering distribution, we see that the reconstructed latent states of the model favor higher levels of cholera transmission than what is typically observed in the incidence data. These result hints at the possibility of model mispecification, and should prompt us to be skeptical of the reliability of conclusions drawn from this model.

> *18. Lines 599-601. "The decision not to do this partially explains the unsuccessful forecasts of Lee et al. [1]: their Table S7 shows that the subset of their simulations which were consistent with observing zero cases in 2019 also accurately predicted the prolonged absence of detected cholera". The fragment before the colon says that Lee was unsuccessful. However, the fragment after the colon says that was in part successful. Please clarify.*

In their article, Lee et al. [1] only presented forecasts that were based on simulations from initial conditions. These forecasts were generally inconsistent with the absence of cholera cases from 2019-2022, and do not account for recent evidence of cholera dynamics. In their supplement material, however, they provide a table (S7) in which they show that if they condition the simulations on most recent data (0 cases from 2019-2020), their forecasts predict cholera elimination. This second effort that was only presented in the supplement material is more consistent with what we suggest is correct practice: forecasts of a system should start using latent states that are consistent with available data. We have expanded this argument in order to clarify this point:

> *Forecasts of a dynamic system should ideally be based on the most recent information available as it is likely to be more relevant than older data. While this assertion may seem self-evident, it is not the case for deterministic models, for which the initial conditions together with the parameters are sufficient for forecasting, and so recent data do not have special importance. Epidemiological forecasts based on deterministic models are not uncommon in practice, despite their limitations [24]. Lee et al. [1] chose to obtain forecasts from all of their models by simulating forward from initial conditions. This decision is possibly as a result of using a deterministic model: forecasts from different models may only be considered comparable only if they are obtained in the same way, which must be by simulating from initial conditions because Model 2 is deterministic.*
>
> *In contrast, for non-deterministic Models 1 and 3, we obtain forecasts by simulating future values using latent states that are harmonious with the most recent data. This is done by drawing latent states at the last observation time (t_N) from the filtering distribution $f_{\boldsymbol{X}_N|\boldsymbol{Y}_{1:N}}(\boldsymbol{x}_N|\boldsymbol{y}^*_{1:N};\hat{\theta})$. The decision to obtain model forecasts from initial conditions partially explains the unsuccessful forecasts of Lee et al. [1]. Table S7 in their supplement material, which contains results that were not discussed in their main article, shows that the subset of their simulations which were consistent with observing zero cases from 2019-2020 were also more consistent with the disappearance of cholera from Haiti from 2019-2022. These results support our argument: forecasts of a dynamic system should start with latent states that are consistent with available data.*

19. Line 631. Since Model 1 accounts for infections at the national level, how are scenarios V1 and V2 handled?

We followed the same approach as Lee et al. [1], which was requiring that the same number of vaccinations are administered at the national level as they would have been in the spatially-explicit vaccination campaigns. The following sentence has been added to clarify this:

> *Because Model 1 only accounts for national level disease dynamics, the pre-determined department-specific vaccination campaigns are carried out by assuming the vaccines are administered in one week to the same number of individuals that would have obtained vaccines if explicitly administered to the specific departments. We refer readers to [1] and the accompanying supplement material for more details.*

References

1. Lee EC, Chao DL, Lemaitre JC, Matrajt L, Pasetto D, Perez-Saez J, et al. Achieving Coordinated National Immunity and Cholera Elimination in Haiti Through Vaccination: A Modelling Study. The Lancet Global Health. 2020;8(8):e1081–e1089.

2. Behrend MR, Basáẽz MG, Hamley JID, Porco TC, Stolk WA, Walker M, et al. Modelling for Policy: The Five Principles of the Neglected Tropical Diseases Modelling Consortium. PLoS Neglected Tropical Diseases. 2020;14(4):1–17.

3. Saltelli A, Bammer G, Bruno I, Charters E, Di Fiore M, Didier E, et al. Five Ways to Ensure that Models Serve Society: a Manifesto. Nature. 2020;582:428–484.

4. Donnelly CA, Boyd I, Campbell P, Craig C, Vallance P, Walport M, et al.. Four principles to make evidence synthesis more useful for policy; 2018.

5. Dahabreh IJ, Chan JA, Earley A, Moorthy D, Avendano EE, Trikalinos TA, et al.. Modeling and Simulation in the Context of Health Technology Assessment: Review of Existing Guidance, Future Research Needs, and Validity Assessment; 2017.

6. Egger M, Johnson L, Althaus C, Schöni A, Salanti G, Low N, et al. Developing WHO guidelines: time to formally include evidence from mathematical modelling studies. F1000Research. 2017;6.

7. Peñaloza Ramos MC, Barton P, Jowett S, Sutton AJ. A Systematic Review of Research Guidelines in Decision-Analytic Modeling. Value in Health. 2015;18(4):512–529.

8. Saltelli A. A short comment on statistical versus mathematical modelling. Nature communications. 2019;10(1):3870.

9. Li J, Ionides EL, King AA, Pascual M, Ning N. Machine Learning for Mechanistic Models of Metapopulation Dynamics. arxiv:231106702. 2023;.

10. Barzilay EJ, Schaad N, Magloire R, Mung KS, Boncy J, Dahourou GA, et al. Cholera Surveillance During the Haiti Epidemic—the First 2 Years. New England Journal of Medicine. 2013;368(7):599–609.

11. Rebaudet S, Bulit G, Gaudart J, Michel E, Gazin P, Evers C, et al. The Case-Area Targeted Rapid Response Strategy to Control Cholera in Haiti: A Four-Year Implementation Study. PLoS Neglected Tropical Diseases. 2019;13(4):e0007263.

12. Rebaudet S, Dély P, Boncy J, Henrys JH, Piarroux R. Toward Cholera Elimination, Haiti. Emerging Infectious Diseases. 2021;27(11):2932.

13. Michel E, Gaudart J, Beaulieu S, Bulit G, Piarroux M, Boncy J, et al. Estimating Effectiveness of Case-Area Targeted Response Interventions Against Cholera in Haiti. Elife. 2019;8:e50243.

14. Trevisin C, Lemaitre JC, Mari L, Pasetto D, Gatto M, Rinaldo A. Epidemicity of Cholera Spread and the Fate of Infection Control Measures. Journal of the Royal Society Interface. 2022;19(188):20210844.

15. Lofgren ET, Halloran ME, Rivers CM, Drake JM, Porco TC, Lewis B, et al. Mathematical models: A key tool for outbreak response. Proceedings of the National Academy of Sciences of the USA. 2014;111(51):18095–18096.

16. McCabe R, Donnelly CA. Disease transmission and control modelling at the science–policy interface. Interface Focus. 2021;11(6):20210013.

17. May RM. Uses and abuses of mathematics in biology. science. 2004;303(5659):790–793.

18. Lau MS, Becker A, Madden W, Waller LA, Metcalf CJE, Grenfell BT. Comparing and linking machine learning and semi-mechanistic models for the predictability of endemic measles dynamics. PLoS computational biology. 2022;18(9):e1010251.

19. Baker RE, Pena JM, Jayamohan J, Jérusalem A. Mechanistic models versus machine learning, a fight worth fighting for the biological community? Biology letters. 2018;14(5):20170660.

20. Bretó C, He D, Ionides EL, King AA. Time Series Analysis via Mechanistic Models. Annals of Applied Statistics. 2009;3:319–348.

21. He D, Ionides EL, King AA. Plug-and-Play Inference for Disease Dynamics: Measles in Large and Small Towns as a Case Study. Journal of the Royal Society Interface. 2010;7:271–283.

22. Lemaitre J, Pasetto D, Perez-Saez J, Sciarra C, Wamala JF, Rinaldo A. Rainfall as a Driver of Epidemic Cholera: Comparative Model Assessments of the Effect of Intra-Seasonal Precipitation Events. Acta Tropica. 2019;190:235–243.

23. Stocks T, Britton T, Höhle M. Model Selection and Parameter Estimation for Dynamic Epidemic Models via Iterated Filtering: Application to Rotavirus in Germany. Biostatistics. 2020;21(3):400–416.

24. King AA, Domenech de Cellès M, Magpantay FM, Rohani P. Avoidable Errors in the Modelling of Outbreaks of Emerging Pathogens, with Special Reference to Ebola. Proceedings of the Royal Society B: Biological Sciences. 2015;282(1806):20150347.

25. Kirpich A, Weppelmann TA, Yang Y, Morris Jr JG, Longini Jr IM. Controlling cholera in the Ouest Department of Haiti using oral vaccines. PLOS Neglected Tropical Diseases. 2017;11(4):e0005482.

26. Pasetto D, Finger F, Camacho A, Grandesso F, Cohuet S, Lemaitre JC, et al. Near real-time forecasting for cholera decision making in Haiti after Hurricane Matthew. PLOS Computational Biology. 2018;14(5):1–22. doi:10.1371/journal.pcbi.1006127.

27. Mukandavire Z, Smith DL, Morris Jr JG. Cholera in Haiti: reproductive numbers and vaccination coverage estimates. Scientific reports. 2013;3(1):997.

28. Kirpich A, Weppelmann TA, Yang Y, Ali A, Morris JG Jr, Longini IM. Cholera Transmission in Ouest Department of Haiti: Dynamic Modeling and the Future of the Epidemic. PLOS Neglected Tropical Diseases. 2015;9(10):1–12. doi:10.1371/journal.pntd.0004153.

29. Lewnard JA, Antillón M, Gonsalves G, Miller AM, Ko AI, Pitzer VE. Strategies to prevent cholera introduction during international personnel deployments: a computational modeling analysis based on the 2010 Haiti outbreak. PLoS medicine. 2016;13(1):e1001947.

30. Kunkel A, Lewnard JA, Pitzer VE, Cohen T. Antimicrobial resistance risks of cholera prophylaxis for United Nations peacekeepers. Antimicrobial agents and chemotherapy. 2017;61(8):10–1128.

31. Mukandavire Z, Morris Jr JG. Modeling the epidemiology of cholera to prevent disease transmission in developing countries. Microbiology spectrum. 2015;3(3):10–1128.

32. Sallah K, Giorgi R, Bengtsson L, Lu X, Wetter E, Adrien P, et al. Mathematical Models for Predicting Human Mobility in the Context of Infectious Disease Spread: Introducing the Impedance Model. International Journal of Health Geographics. 2017;16(1):1–11.

33. Azman AS, Luquero FJ, Rodrigues A, Palma PP, Grais RF, Banga CN, et al. Urban cholera transmission hotspots and their implications for reactive vaccination: evidence from Bissau city, Guinea bissau. PLoS neglected tropical diseases. 2012;6(11):e1901.

34. Azman AS, Luquero FJ, Ciglenecki I, Grais RF, Sack DA, Lessler J. The impact of a one-dose versus two-dose oral cholera vaccine regimen in outbreak settings: A modeling study. PLoS Medicine. 2015;12(8):e1001867.

35. Andrieu C, Doucet A, Holenstein R. Particle Markov chain Monte Carlo methods. Journal of the Royal Statistical Society, Series B (Statistical Methodology). 2010;72:269–342.

36. Fox SJ, Lachmann M, Tec M, Pasco R, Woody S, Du Z, et al. Real-time pandemic surveillance using hospital admissions and mobility data. Proceedings of the National Academy of Sciences of the USA. 2022;119(7):e2111870119.

37. Molodecky NA, Jafari H, Safdar RM, Ahmed JA, Mahamud A, Bandyopadhyay AS, et al. Modelling the spread of serotype-2 vaccine derived-poliovirus outbreak in Pakistan and Afghanistan to inform outbreak control strategies in the context of the COVID-19 pandemic. Vaccine. 2023;41:A93–A104.

38. Pons-Salort M, Grassly NC. Serotype-Specific Immunity Explains the Incidence of Diseases Caused by Human Enteroviruses. Science. 2018;361(6404):800–803.

39. Subramanian R, Romeo-Aznar V, Ionides E, Codeço CT, Pascual M. Predicting Re-Emergence Times of Dengue Epidemics at Low Reproductive Numbers: DENV1 in Rio de Janeiro, 1986–1990. Journal of the Royal Society Interface. 2020;17(167):20200273.

40. Coulson T, Rohani P, Pascual M. Skeletons, noise and population growth: The end of an old debate? Trends in Ecology and Evolution. 2004;19:359–364.

41. King AA, Nguyen D, Ionides EL. Statistical Inference for Partially Observed Markov Processes via the R Package pomp. Journal of Statistical Software. 2016;69:1–43.

42. Gneiting T, Raftery AE. Strictly Proper Scoring Rules, Prediction, and Estimation. Journal of the American Statistical Association. 2007;102(477):359–378.

43. Ionides EL, Breto C, Park J, Smith RA, King AA. Monte Carlo Profile Confidence Intervals for Dynamic Systems. Journal of the Royal Society Interface. 2017;14:1–10.

44. Ning N, Ionides EL, Ritov Y. Scalable Monte Carlo Inference and Rescaled Local Asymptotic Normality. Bernoulli. 2021;pre-published online.

45. Subramanian R, He Q, Pascual M. Quantifying Asymptomatic Infection and Transmission of COVID-19 in New York City using Observed Cases, Serology, and Testing Capacity. Proceedings of the National Academy of Sciences of the USA. 2021;118(9).

46. Ali ST, Lau YC, Shan S, Ryu S, Du Z, Wang L, et al. Prediction of upcoming global infection burden of influenza seasons after relaxation of public health and social measures during the COVID-19 pandemic: a modelling study. The Lancet Global Health. 2022;10(11):e1612–e1622.

47. Lucas RE, et al. Econometric Policy Evaluation: A Critique. In: Carnegie-Rochester Conference Series on Public Policy. vol. 1; 1976. p. 19–46.

48. Bi Q, Ferreras E, Pezzoli L, Legros D, Ivers LC, Date K, et al. Protection against cholera from killed whole-cell oral cholera vaccines: a systematic review and meta-analysis. The Lancet Infectious Diseases. 2017;17(10):1080–1088.

49. Wheeler J, Ionides EL. Likelihood Based Inference for ARMA Models. arXiv preprint arXiv:231001198. 2023;.

50. Smith Jr JW. Ecosystem Models in a Bayesian State Space Framework. Virginia Tech. 2022;.

51. Rivera WOM. Estimation for Disease Models Across Scales. Arizona State University. 2022;.

52. Ionides EL, Asfaw K, Park J, King AA. Bagged filters for partially observed interacting systems. Journal of the American Statistical Association. 2021;pre-published online. doi:10.1080/01621459.2021.1974867.

53. Ionides EL, Ning N, Wheeler J. An Iterated Block Particle Filter for Inference on Coupled Dynamic Systems with Shared and Unit-Specific Parameters. Statistica Sinica,. 2022; p. pre–published online.