February 7, 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).

WE CAN TRIM DOWN THE RESPONSE LATER. HERE ARE SOME INITIAL BRAINSTORMING THOUGHTS.

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. [TODO: 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.] [TODO: In the methods and results sections, we have made explicit how the decisions about what to present in the main text and in the supplement relate to these goals.] 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. [TODO: 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.

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. [TODO: We have clarified this in the manuscript WHERE?]

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 [TODO:]. We have cited [2–4] for examples demonstrating situations where overdispersed dynamic noise is present, and [5] for an investigation of the statistical limitations of fitting deterministic models to data.

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 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 [4, 5], following this recommendation. However, a subsequent manuscript addresses [4,5] more directly [4] and so we have referred the reader to that paper

(currently available on arXiv) for additional discussion of this point.

[IF THE P0INTS OF 4, 5 CAN BE CONNECTED DIRECTLY TO THIS DATA ANALYSIS, WE OUGHT TO BE ABLE TO DISCUSS THAT WITHOUT CONFLICTING WITH LI23. WE SHOULD PROBABLY RE-READ THESE REFERENCES AND FIGURE IT OUT.]

[ANOTHER IDEA: WE CAN VIEW OUR PAPER AND LI23 AS SYNERGISTIC RATHER THAN IN COMPETITION. WE COULD EVEN TALK ABOUT HOW THAT DIFFERENT SETTING (COVID, WITH AN EMPHASIS ON SPATIOTEMPORAL MODELING AND COMPARING PRIMARILY WITH ENSEMBLE FILTER METHODS) LEADS TO SIMILAR CONCLUSIONS TO HAITI CHOLERA. FROM THAT PERSPECTIVE, THE PAPERS STRENGTHEN EACH OTHER. AS LONG AS BOTH LISTS OF RECOMMENDATIONS ARE FIRMLY TIED TO THE RESPECTIVE DATA ANALYSES, IT SHOULD BE A STRENGTH NOT A WEAKNESS IF THEY OVERLAP. BUT IF WE GO THAT ROUTE WE'LL HAVE TO EXPLAIN THIS CONVINCINGLY.]

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?

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.

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.

8. I found the commentary on comparing fitted mechanistic models to a (statistical) bench-mark 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.

[TODO: This is right, as long as there are not any initialization issues...]

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.

12. [I THINK THIS MEANS LINES 553-556, ASSUMING THE NUMBERING MATCHES] Lines 53-56: 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. [TODO: 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.]

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.

underlineTypos

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 [TODO: where?] [How does the Materials and Methods: Mechanistic models for disease modeling sub-section sound?]

Mechanistic models representing biological phenomena are valuable for epidemiology and consequently for public health policy [6, 7]. More broadly, they have useful roles throughout biology, expecially when combined with statistical methods that properly account for stochasticity and nonlinearity [8]. In some situations, modern machine learning methods can outperform mechanistic models on epidemiological forecasting tasks [9]. 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.

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 identifability. 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. [HOW DOES THIS PLAY OUT IN THE MS? HOW DO WE CLARIFY IT?]

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 [2, 10] leading to one extra parameter compared to a deterministic model. In the context of cholera dynamics, [11] found a substantial improvement in model fit by including stochasticity, involving one over-dispersion parameter. [3] found similar results, favoring a stochastic dynamic model with over-dispersion, when investigating transmission dynamics of a different enteric pathogen, rotavirus. In other contexts, [5] 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.

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 9 previous Haiti cholera investigations including stochastic dynamic models [1, 12–19], four of them carried out inference via MCMC [14, 16–18]. [WHAT SORTS OF MCMC ALGORITHMS DID THEY USE?] 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, 19]. In principle, MCMC methods can be applied for inference in over-dispersed stochastic dynamic models via so-called particle MCMC (PMCMC) [20]. In practice, these models have been fitted by maximum likelihood, some examples including [3, 11, 21–24]. 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. 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 [25]. 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 nowaways 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 diagnostic 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 [26] 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.

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 pionts provide a consensus on the maximum log-likelihood, and the Monte Carlo variance is small, we can have reasonable confidence in our

L-9

maximization. Monte Carlo adusted profile (MCAP) methods [27, 28] 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 [REFS].

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. 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. [TODO:]

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.

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.

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.

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.

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.

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

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

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.

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.

7. In line 159, it is stated that "$v^\star(t)$ 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.

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]. Judging by the referee's comment, they will not be suprised to learn that this structure led to difficulties in some situations. [TODO:]

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?

The cholera pathogen was recently introduced to Haiti, which facilitates the specification of initial conditions. 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. Our approaches to addressing this issue involve: (i) looking at residual diagnostic plots 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.

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

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

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 is similar to point 11 of referee 1. LET'S DISCUSS THIS. Our point is that there are many other ways one might include time-varying effects in the model, and a strongly supported conclusion should avoid ruling out untested hypotheses. We can say that the best available model with trending parameters is statistically superior to the best available model with time-constant parameters. This frames the evidence in a way that avoids asserting clear support for any one particular way of modeling this effect.

R1 mentioned the possibility of reporting rate decreasing instead. I'M TRYING TO RECALL OUR REASONING THAT MODEL 3 IN LEE ET AL OVER-EMPHASIZED THIS. APART FROM THE POSSIBILITY OF A CASE DEFINITION CHANGE, ONE MIGHT EXPECT REPORTING RATE TO INCREASE?

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

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

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.

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 [29, 30] then fits can be compared against an independent benchmark for each block, and then summed to compare on the entire dataset.

Since the value of benchmarks is a major point in this paper, we have added these extra details in a supplementary section [TODO:].

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.

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.

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

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. 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.

3. 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.

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

5. 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.

6. 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.

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

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

9. 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.

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

11. 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.

12. Tuite AR, Tien J, Eisenberg M, Earn DJ, Ma J, Fisman DN. Cholera Epidemic in Haiti, 2010: Using a Transmission Model to Explain Spatial Spread of Disease and Identify Optimal Control Interventions. Annals of internal medicine. 2011;154(9):593–601.

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

14. 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.

15. 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.

16. 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.

17. 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.

18. Kühn J, Finger F, Bertuzzo E, Borgeaud S, Gatto M, Rinaldo A, et al. Glucose-but not rice-based oral rehydration therapy enhances the production of virulence determinants in the human pathogen Vibrio cholerae. PLoS neglected tropical diseases. 2014;8(12):e3347.

19. 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.

20. 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.

21. 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.

22. 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.

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

24. 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.

25. 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.

26. 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.

27. 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.

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

29. 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.

30. 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.