# Embedded model discrepancy: A case study of Zika modeling

Rebecca E. Morrison (iD), and Americo Cunha (iD)

View Online          Export Citation          CrossMark

# Embedded model discrepancy: A case study of Zika modeling

View Online     Export Citation     CrossMark

Rebecca E. Morrison[1,a] (iD) and Americo Cunha, Jr.[2,b] (iD)

## AFFILIATIONS

[1] Department of Computer Science, University of Colorado Boulder, Boulder, Colorado 80309, USA
[2] Department of Applied Mathematics, Rio de Janeiro State University (UERJ), Rio de Janeiro 20550, Brazil

[a] **Author to whom correspondence should be addressed:** rebeccam@colorado.edu. **URL:** rebeccaem.github.io
[b] americo@ime.uerj.br. **URL:** www.americocunha.org

## ABSTRACT

Mathematical models of epidemiological systems enable investigation of and predictions about potential disease outbreaks. However, commonly used models are often highly simplified representations of incredibly complex systems. Because of these simplifications, the model output, of, say, new cases of a disease over time or when an epidemic will occur, may be inconsistent with the available data. In this case, we must improve the model, especially if we plan to make decisions based on it that could affect human health and safety, but direct improvements are often beyond our reach. In this work, we explore this problem through a case study of the Zika outbreak in Brazil in 2016. We propose an embedded discrepancy operator—a modification to the model equations that requires modest information about the system and is calibrated by all relevant data. We show that the new enriched model demonstrates greatly increased consistency with real data. Moreover, the method is general enough to easily apply to many other mathematical models in epidemiology.

Potential epidemics of communicable diseases are a major health concern of the modern world, especially as city density, air and water pollution, and worldwide travel steadily increase. A stark example of this is the global coronavirus outbreak, already responsible for more than 2000 deaths around the world at the time of this article's submission and more than 13 000 at the time of revision, approximately one month later. When faced with a potential outbreak, decision-makers such as health officials and medical professionals rely on mathematical models to aid their decision-making processes. However, oftentimes, these models are not consistent with the dynamical system they are designed to represent. The discrepancy between the output of a model and the real system is then a serious impediment, as it may decrease our confidence in the model, or even invalidate it entirely, so that it can no longer be used to aid in decision-making. When such a discrepancy is observed, we, as modelers, must either improve the model or somehow account for the discrepancy itself. While a direct model improvement is usually the most desirable solution, how to do so may be infeasible because of computational reasons, time constraints, or lack of domain knowledge. This paper provides a systematic method to instead account for the discrepancy itself, explored via a case study of the Brazilian Zika epidemic of 2016. The method is not a correction of the model output to data, but rather a modification of the model equations themselves by a so-called embedded discrepancy operator. The operator is designed with three critical properties in mind: interpretability, domain-consistency, and robustness. We show that including the embedded discrepancy operator greatly increases the fidelity of the model so that the model output and real data are now in fact consistent.

---

## I. INTRODUCTION

Mathematical models of scientific systems necessarily include simplifications about the actual system they aim to represent. In some cases, these simplifications do not preclude the use of the model to understand, investigate, and make decisions and predictions about the system. The quintessential example of this comes from the domain of classical mechanics: Newtonian mechanics ignores quantum and relativistic effects but, over a wide domain of masses and energies, provides a completely adequate model to

describe the motion of macroscopic objects. Outside this domain, however, quantum or relativistic effects are no longer negligible, and Newtonian mechanics fails.

Just as classical mechanics is insufficient to describe a quantum system, the simplifications of a modern mathematical model may yield a discrepancy between the model and the system at hand too great to be ignored. This discrepancy is revealed during *model validation*, a process by which we check that the mathematical model is a reliable representation of reality. Without accounting for the discrepancy, one cannot trust the model output, much less use it to make predictions or decisions. In this case, there are two immediate options: (1) improve the model directly, i.e., from first principles or by including additional information and (2) represent the model discrepancy itself. While option 1 is usually desirable, it may not be feasible due to computational constraints or because we in fact lack the knowledge to directly improve the model. Then, we are left with option 2—represent the model discrepancy.

A common approach to account for a model discrepancy is through a *response discrepancy function*,[1] also called a bias function. A response discrepancy function corrects the model output (or response) to data. Typically, an additive function on the model output is calibrated to data, either point-wise or with a parametric form. An advantage of this approach is that it can be implemented even if the model is a black box; that is, one only needs access to the model output, not the model itself. There are also disadvantages. In essence, a response discrepancy function builds a better interpolation to a single dataset over the range of usable data. Thus, this approach provides no basis for extrapolation to, for example, make a prediction about the probability of an epidemic next year. Furthermore, the action of this bias function is not interpretable, as it lies outside the model equations.

Instead, in this paper, we show how to modify equations directly to account for the model error with an *embedded discrepancy operator*. The advantages of this approach are threefold:

1. *Interpretability*: As the embedded operator appears within the model equations, and acts on state variables, the action of this operator is interpretable.
2. *(Domain-)Consistency*: Information or constraints about the system can be incorporated into the discrepancy operator.
3. *Robustness*: Discrepancy parameters can, and should, be calibrated over all the available data. This can include data from multiple scenarios or initial conditions.

These three properties—interpretability, consistency, and robustness—are designed to allow for decisions or extrapolative predictions. The inclusion of the embedded discrepancy operator into the original, or reduced model, yields an *enriched model*. In essence, the enriched model takes advantage of both mechanistic and statistical modeling: it retains the reduced mechanistic model and incorporates a general, statistically calibrated discrepancy model. Of course, this intrusive approach highly depends on the context. Here, we investigate the value of an embedded discrepancy operator in the context of epidemiology modeling.

Mathematical modeling of disease spread and outbreaks has a long and rich history; see Refs. 2–5, to name just a few. One

of the most common classes of these models consists of coupled ordinary differential equations (ODEs), whose state variables include populations of the host (here, humans) and the disease carrier, or vector. These populations are further specified as either **S**usceptible, **E**xposed, **I**nfected, or **R**ecovered, leading to thus-named SEIR models.[6] These models are relatively simple to implement and understand. Model parameters allow for the specification of transmission rates, incubation times, etc.

In particular, we investigate the model discrepancy of a well-studied SEIR-SEI model of the Zika outbreak in Brazil in 2016.[7] In the previous work, after calibration of model parameters, the reduced model captured major tendencies of the outbreak. This was a major improvement compared to the reduced model with parameter values as suggested by current literature, which bore almost no usable resemblance to the real epidemic data. However, the calibrated model was still insufficient to precisely capture the dynamical behavior of the Zika outbreak. The current work extends previous works of the embedded model discrepancy, used in the contexts of combustion[8] and ecological models,[9] to the current domain of epidemiology. The discrepancy model is embedded within the coupled model differential equations, and the introduced discrepancy parameters are calibrated to the available data. The enriched model is shown to greatly outperform the original model.

To differentiate the current article from Ref. 9, note that the study was primarily a numerical study over a constrained set of scenarios. The interaction matrices (determining reduced and true model coefficients) follow a number of assumptions, such as negative-definiteness, yielding highly well-behaved models. In addition, the actual discrepancy was known exactly between each reduced model and the corresponding data-generating model; some of that information was used to further constrain the discrepancy model parameters. Thus, the previous paper provided no guarantee that this method would work in a highly applied, real-world model scenario without those strong assumptions.

Although the current modeling scheme only describes a single outbreak,[10] and thus is not suitable to describe multiple incidences of the disease, this type of model is useful for guiding decision and policy makers. For example, Ref. 11 describes common questions faced by decision-makers, such as how many total people will be infected or even how to slow or prevent the outbreak from occurring. The objective of the present article is to reproduce with reasonable precision the data of a real epidemic. An appropriate enrichment and calibration of the model can then yield useful predictions about the dynamical system.

A final complication of this field is that new disease cases are often not reported, causing the outbreak numbers to appear artificially low. A study from 2018 estimates that as much as 90% of the cases are not reported.[12] However, as we will see shortly, the issue of faulty data is insufficient to account for the discrepancy between the original model and observations. At the same time, the issue of under-reported cases does obviously play an important role during the model validation process. We try to disentangle the two problems—observational errors and model errors—by first considering only model errors and later allowing for both model errors and significant under-reporting. We consider how the enriched model

performs in different possible under-reporting scenarios, such as 10% or 50%.

The rest of the paper is organized as follows. Section II describes the specific model of the 2016 Brazilian Zika outbreak and reconstructs previous results as a reference. Section III presents the formulation and calibration of the embedded discrepancy operator and also corresponding numerical results. Section IV explores the issue of under-reporting and how well the enriched model might perform given some sample under-reporting scenarios. A brief concluding discussion is given in Sec. V.

## II. ZIKA DISEASE MODELING

As mentioned in Sec. I, a typical approach to model the spread of an infectious disease is with a set of coupled ordinary differential equations. Here, we consider the well-known SEIR-SEI model, which describes coupled growth rates of species of interest, namely, susceptible, exposed, infected, and recovered humans, as well as susceptible, exposed, and infected vectors. In the case of Zika (also dengue and yellow fever) in Brazil, this vector is the *Aedes aegypti* mosquito.

In this section and in Sec. III, we assume that the data represent the actual truth. That is, the modeling objective is to achieve a model consistent with the given data.

### A. Model specification

We follow the SEIR-SEI model discussed by Dantas *et al.*,[7] which included species $S_h, E_h, I_h, R_h, S_v, E_v$, and $I_v$. Subscripts $h$ and $v$ indicate humans and vectors, respectively. The model also includes a state variable $C(t)$, which counts cumulative new cases over time. Then, the eight coupled equations are

$$\frac{dS_h}{dt} = -\beta_h S_h I_v / N_v, \tag{1a}$$

$$\frac{dE_h}{dt} = \beta_h S_h I_v / N_v - \alpha_h E_h, \tag{1b}$$

$$\frac{dI_h}{dt} = \alpha_h E_h - \gamma I_h, \tag{1c}$$

$$\frac{dR_h}{dt} = \gamma I_h, \tag{1d}$$

$$\frac{dS_v}{dt} = \delta N_v - \beta_v S_v I_h / N_h - \delta S_v, \tag{1e}$$

$$\frac{dE_v}{dt} = \beta_v S_v I_h / N_h - (\alpha_v + \delta) E_v, \tag{1f}$$

$$\frac{dI_v}{dt} = \alpha_v E_v - \delta I_v, \tag{1g}$$

$$\frac{dC}{dt} = \alpha_h E_h, \tag{1h}$$

where $N_h$ represents Brazil's human population and $N_v$ represents the vector population. Nominal values of the interaction rates were determined by a careful literature study. These rates are

$$\text{Extrinsic incubation period:} \quad \frac{1}{\alpha_v} = 9.1, \tag{2a}$$

$$\text{Intrinsic incubation period:} \quad \frac{1}{\alpha_h} = 5.9, \tag{2b}$$

$$\text{Human infectious period:} \quad \frac{1}{\gamma} = 7.9, \tag{2c}$$

$$\text{Vector lifespan:} \quad \frac{1}{\delta} = 11, \tag{2d}$$

$$\text{Mosquito to human infection time:} \quad \frac{1}{\beta_h} = 11.3, \tag{2e}$$

$$\text{Human to mosquito infection time:} \quad \frac{1}{\beta_v} = 8.6. \tag{2f}$$

Let us collect these model parameters into the vector $\theta$, and let $\theta_n$ refer to the nominal values given above.

To fully specify this model, it remains to provide initial conditions. These are[13]

$$S_h(0) = N_h - E_h(0) - I_h(0) - R_h(0), \tag{3a}$$

$$E_h(0) = I_h(0), \tag{3b}$$

$$I_h(0) = C(0), \tag{3c}$$

$$R_h(0) = 29\,639, \tag{3d}$$

$$S_v(0) = N_v - E_v(0) - I_v(0), \tag{3e}$$

$$E_v(0) = I_v(0), \tag{3f}$$

$$I_v(0) = 2.2 \times 10^{-4}, \tag{3g}$$

$$C(0) = 8\,201. \tag{3h}$$

Finally, let us call the above model $\mathscr{Z}$ and the state vector $x$, where $x$ is ordered in the same way as Eqs. (1a)–(1h) ($x_1 = S_h$, $x_2 = E_h$, and so on). Then, we may refer to the above model as

$$\frac{dx}{dt} = \mathscr{Z}(x; \theta). \tag{4}$$

We may also refer to this model as the original model or a *reduced* model.

### B. Previous results compared to data

The work in Ref. 7 presents a detailed approach to calibrate this model to data. The data are made available by the Brazilian Ministry of Health[14] and are available as a supplementary material in Ref. 7. Each data point $d_i$, $i = 1, \ldots, 52$, gives the recorded cumulative number of Zika cases at epidemiological week $i$ of the year 2016. In this section, we re-plot the results from that paper to serve as an immediate reference and comparison.

First, Fig. 1 compares the model output to data, using the above model with nominal parameters. The model with nominal parameter values, $\mathscr{Z}(x; \theta = \theta_n)$, severely underestimates the outbreak. Note that under-reporting cannot explain the observed discrepancy: higher reporting rates would only increase this discrepancy.

Clearly, the reduced model, given $\theta_n$ parameter values, is a poor representation of reality. After observing such a discrepancy, the
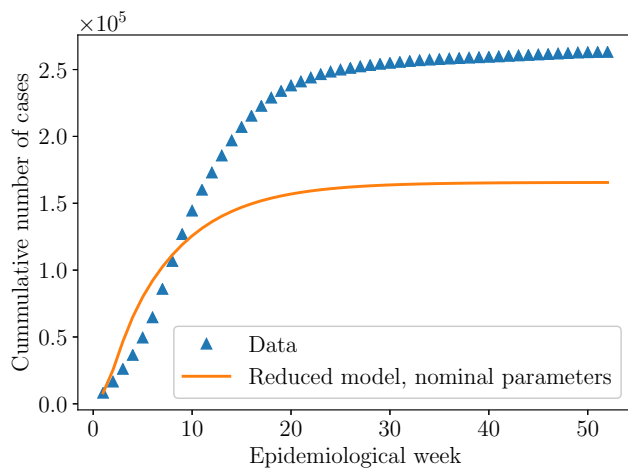
FIG. 1. Outbreak data and a reduced model response using nominal parameter values.

authors of Ref. 7 performed a sophisticated calibration of the model parameters $\theta$, using the Trust-Region-Reflective (TRR) method.[15,16] Following that method, and using the public data to calibrate, two slightly different results are obtained by imposing a different set of constraints on the possible parameter values. The model outputs after both calibrations are shown in Fig. 2. Although the model output is now much closer to the data, still a detectable inconsistency persists. To be specific, note that after about week 30, the difference is tens of thousands of new cases of humans infected by Zika.

From a modeling perspective, the salient point is that the model is unable to capture the dynamical behavior of the outbreak, even after calibration. Assuming (as we are for now) that the given data are correct, this suggests that the problem lies
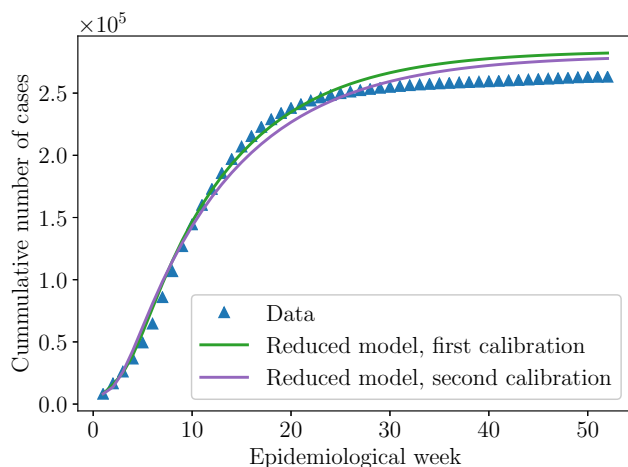


FIG. 2. Outbreak data and a reduced model response using TRR calibrated parameter values.
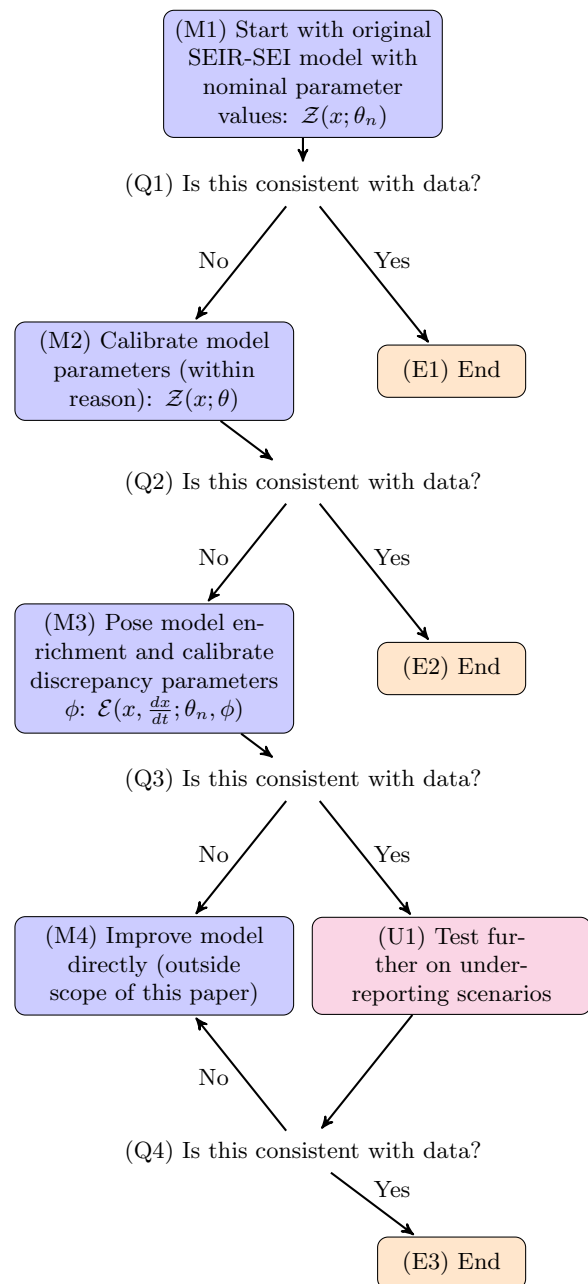


FIG. 3. A schematic diagram of the different models and their relationships considered in this paper.

with the model itself. Indeed, there are several possible sources of model errors that may impact not only this specific Zika model but many other epidemiological models. First, these SEIR models are not built from first principles, but rather from assumptions about interactive behavior, empirical information, and domain

scientists' intuition and experience. Second, these models provide a continuous, deterministic description of discrete interactions, which naturally involve some stochasticity.[17] With large enough populations, though, this should not be a problem. Third, diseases do not spread in a closed system of host and vector. Rather, the spread of a disease involves other species such as livestock and non-human primates.[18] Fourth, other modes of transmission are possible, such as sexual interactions[19,20] and blood transfusions.[21] Finally, there could certainly be additional time- or spatially-dependent effects, such as migrations and local dynamics,[22,23] collective behavior, and time-delayed synchronization dynamics.[24] Some modelers assume power-law dynamics of the networks,[25] while others use fractional derivatives to describe relevant dynamics.[26] Instead, this model assumes time-independent parameters and only model populations over time, not space. In summary, the spread of a contagious disease is an incredibly complex problem, and it remains unclear what is critically missing from the model or how to best improve it directly from epidemiological information. Here, then, we can turn to the field of model discrepancy to help.

## III. EMBEDDED DISCREPANCY OPERATOR

Before describing the embedded discrepancy operator, we illustrate the overall relationship between the different models considered in this paper. A schematic diagram is shown in Fig. 3. As seen in Sec. II, after calibrating the model parameters to data, there is still a significant discrepancy between the model output and the data. That is, the answer to Q2 (and Q1) in Fig. 3 is "No," and therefore, we move to state (M3): we model this discrepancy with the goal of reaching states (U1) and ultimately (E3).

### A. Proposed approach: Embedded discrepancy operator

Previous work has shown that missing dynamics on the right-hand side (RHS) of differential equations can be approximated with "extra" information about the existing state variables,[27–29] such as memory or derivative information. Exploiting this, we pose the following enriched model:

$$\frac{dx}{dt} = \mathscr{Z}(x;\theta) + \Delta\left(x, \frac{dx}{dt}, \phi\right), \tag{5}$$

where

$$\Delta_{Sh} = \kappa_1 S_h + \lambda_1 \frac{dS_h}{dt}, \tag{6a}$$

$$\Delta_{Eh} = \kappa_2 E_h + \lambda_2 \frac{dE_h}{dt}, \tag{6b}$$

$$\Delta_{Ih} = \kappa_3 I_h + \lambda_3 \frac{dI_h}{dt}, \tag{6c}$$

$$\Delta_{Rh} = \kappa_4 R_h + \lambda_4 \frac{dR_h}{dt}, \tag{6d}$$

$$\Delta_{Sv} = \kappa_5 S_v + \lambda_5 \frac{dS_v}{dt}, \tag{6e}$$

$$\Delta_{Ev} = \kappa_6 E_v + \lambda_6 \frac{dE_v}{dt}, \tag{6f}$$

$$\Delta_{Iv} = \kappa_7 I_v + \lambda_7 \frac{dI_v}{dt}, \tag{6g}$$

$$\Delta_C = 0, \tag{6h}$$

with $\kappa = (\kappa_1,\ldots,\kappa_7)$ and $\lambda = (\lambda_1,\ldots,\lambda_7)$. That is, the differential equation for $x_i, i = 1,\ldots,7$ in the reduced model is modified by two additional terms, one linear in $x_i$ and the other linear in $dx_i/dt$. The discrepancy parameters are collected into the vector $\phi$,

$$\phi = (\kappa,\lambda) = (\kappa_1,\ldots,\kappa_7,\lambda_1,\ldots,\lambda_7). \tag{7}$$

Note, the RHS for $dC(t)/dt$ is not modified because this function simply counts the exposed cases as given by the model. A change here would be analogous to modifying the model output itself, not interpretable, and not reliable for any type of decision or prediction.

As mentioned in Sec. I, this type of discrepancy model can be constrained to the available information about the system. For example, the discrepancy operator for a combustion reaction in Ref. 8 is constrained to satisfy conservation of atoms and conservation of energy. In this scenario, we do not have such strict constraints; see Ref. 9 for constrained operators in similar Lotka–Volterra models.

Altogether, the enriched model is

$$\frac{dS_h}{dt} = -\beta_h S_h I_v/N_v + \Delta_{Sh}, \tag{8a}$$

$$\frac{dE_h}{dt} = \beta_h S_h I_v/N_v - \alpha_h E_h + \Delta_{Eh}, \tag{8b}$$

$$\frac{dI_h}{dt} = \alpha_h E_h - \gamma I_h + \Delta_{Ih}, \tag{8c}$$

$$\frac{dR_h}{dt} = \gamma I_h + \Delta_{Rh}, \tag{8d}$$

$$\frac{dS_v}{dt} = \delta N_v - \beta_v S_v I_h/N_h - \delta S_v + \Delta_{Sv}, \tag{8e}$$

$$\frac{dE_v}{dt} = \beta_v S_v I_h/N_h - (\alpha_v + \delta)E_v + \Delta_{Ev}, \tag{8f}$$

$$\frac{dI_v}{dt} = \alpha_v E_v - \delta I_v + \Delta_{Iv}, \tag{8g}$$

$$\frac{dC}{dt} = \alpha_h E_h. \tag{8h}$$

Denote the enriched model as $\mathscr{E}(x, dx/dt; \theta, \phi)$, i.e.,

$$\frac{dx}{dt} = \mathscr{E}\left(x, \frac{dx}{dt}; \theta, \phi\right) \tag{9}$$

$$= \mathscr{Z}(x;\theta) + \Delta\left(x, \frac{dx}{dt}; \phi\right). \tag{10}$$

We set $\theta = \theta_n$ and use the same initial conditions as in Eqs. (3a)–(3h). The final step to fully specify the enriched model is to calibrate the discrepancy parameters $\phi$; this step is explained in Subsection III B.

## B. Calibration details

In contrast with the calibration process of Ref. 7 described in Sec. II B, here we use a Bayesian framework[30,31] to calibrate the discrepancy parameters $\phi$ given the data $d$. This allows for the representation of uncertainty about these parameters and also how this uncertainty propagates to the model output.

Recall that the observations are cumulative cases at each epidemiological week, $d = \{d_i\}, i = 1, \ldots, 52$. For each $d_i$, let $y_i$ be the corresponding model output. We assume that the measurements are independent and that the measurement error is additive and Gaussian as such,

$$d_i = y_i + \epsilon, \quad \epsilon \sim \mathcal{N}(0, \sigma_\epsilon^2), \quad (11)$$

with standard deviation $\sigma_\epsilon = 5 \times 10^3$. This standard deviation value seems reasonable as the uncertainty in reported values is high and because the observations are on the order of tens to hundreds of thousands.

In the Bayesian framework, the conditional probability density of $\phi$ given the data $d$, $p_{\text{po}}(\phi|d)$, is called the *posterior* and given as

$$p_{\text{po}}(\phi|d) = \frac{p_{\text{li}}(d|\phi)p_{\text{pr}}(\phi)}{p_{\text{ev}}(d)}. \quad (12)$$

We specify each term on the RHS above:

- *Prior:* The prior density $p_{\text{pr}}(\phi)$ collects the prior knowledge we have about the parameters. Specifically, these parameters are assumed independent and uniform in the prior, where each $p_{\text{pr}}(\phi_i) = \mathcal{U}(-0.3, 0.15)$, and therefore,

$$p_{\text{pr}}(\phi) = \prod_{i=1}^{14} p_{\text{pr}}(\phi_i). \quad (13)$$

- *Likelihood:* The likelihood $p_{\text{li}}(d|\phi)$ tells us how likely it is to observe $d$, given a particular value of $\phi$. The measurement error model in Eq. (11) yields the likelihood function

$$p_{\text{li}}(d|\theta) = \frac{1}{\sqrt{2\pi|\Sigma|}} \exp\left(-\frac{1}{2}(d-y)^T \Sigma^{-1}(d-y)\right), \quad (14)$$

  where $\Sigma = \sigma_\epsilon^2 I$.

- *Evidence:* The evidence $p_{\text{ev}}(d) = \int p_{\text{li}}(d|\phi)p_{\text{pr}}(\phi)d\phi$ gives the probability of observing the data $d$. This is typically difficult to compute, but note that it is not a function of $\phi$. With a Markov chain Monte Carlo (McMC) approach, the posterior is found by computing ratios of the RHS in Eq. (12) (for different values of $\phi$), and therefore, fortunately, this term cancels.

Under this framework, the discrepancy parameters $\phi$ are calibrated using the DRAM method, developed by Ref. 32 and implemented through the library QUESO.[33] The complete code for this project is available at https://github.com/rebeccaem/zika.[34]

Table I presents posterior means and standard deviations for each of the 14 marginal posterior densities of the discrepancy parameters.

## C. Numerical results

Figure 4 shows the enriched model response compared to the data. Uncertainty in discrepancy parameters $\phi$ is propagated

**TABLE I.** Information from the parameter posterior $p_{\text{po}}(\phi|d)$.

| Parameter | Posterior mean | Posterior standard deviation |
|---|---|---|
| $\kappa_1$ | −0.04 | 0.005 |
| $\kappa_2$ | −0.26 | 0.02 |
| $\kappa_3$ | 0.10 | 0.02 |
| $\kappa_4$ | −0.04 | 0.13 |
| $\kappa_5$ | 0.07 | 0.01 |
| $\kappa_6$ | 0.11 | 0.02 |
| $\kappa_7$ | 0.10 | 0.01 |
| $\lambda_1$ | 0.00 | 0.09 |
| $\lambda_2$ | −0.15 | 0.08 |
| $\lambda_3$ | −0.18 | 0.08 |
| $\lambda_4$ | −0.15 | 0.09 |
| $\lambda_5$ | −0.02 | 0.10 |
| $\lambda_6$ | −0.15 | 0.10 |
| $\lambda_7$ | −0.06 | 0.09 |

through to the model output: the thick centerline shows the median response, the darker band shows the 50% confidence interval, and the lighter band the 95% confidence interval (CI). Importantly, all observations are in fact captured by the 95% CI.

For comparison's sake, Fig. 5(a) presents at once all model responses considered in this paper, and Fig. 5(b) shows the same but zoomed into weeks 12–30. (For visualization purposes, only the median line is shown for the enriched model.) The enriched model is clearly an improvement.

## D. Interpretation

The embedded discrepancy operator can be interpreted from two different points of view. The first is a more mathematical lens, although not disconnected from the physics: we interpret the
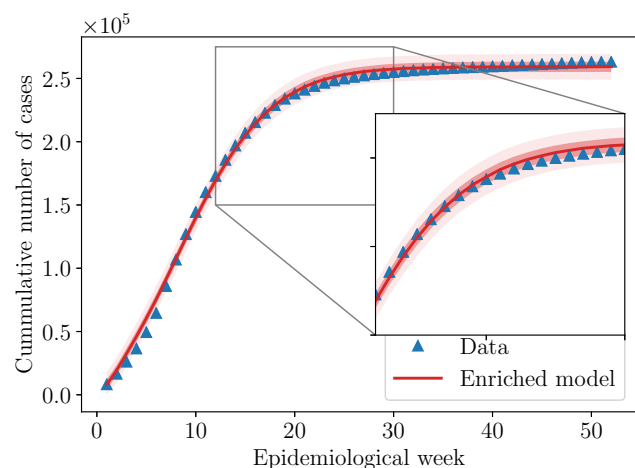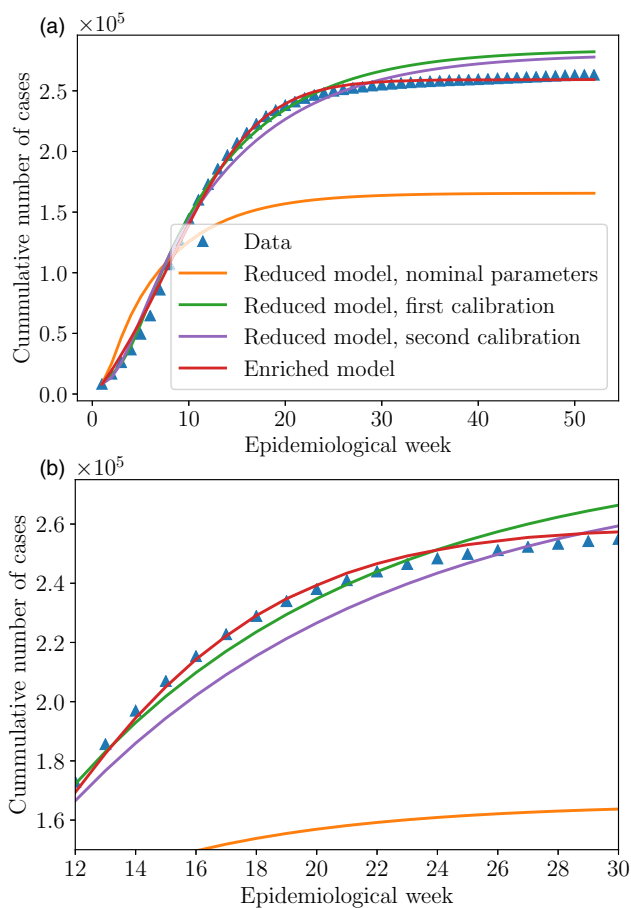


**FIG. 4.** Outbreak data and an enriched model response.

**FIG. 5.** Outbreak data compared to all model responses. Figure (b) is zoomed into weeks 12–30.



**FIG. 6.** Modified outbreak data, assuming 10% under-reporting, and an enriched model response.

discrepancy operator as a linear feedback signal. The second, in contrast, relies on an epidemiological basis: here, we interpret the corrections made by the discrepancy operator as effects due to causes of a biological origin.

This second point of view is especially interesting for the goal of elucidating potential deficiencies in the baseline model. As a wide range of issues must yet be explored to obtain a consistent epidemiological interpretation, this line will not be addressed in this paper but will be the topic of future work. Instead, the first perspective is explored further below.

In light of the theory of systems with a linear feedback, the discrepancy operator is a linear combination of the system state and its first order time-derivative, thus defining a signal that feeds the original nonlinear system with information from the present state and its rate of change. Roughly speaking, the parameters of the enrichment can be seen as "gains" that adjust to drive the epidemic curve generated by the model toward the real observational curve. These parameters are identified via Bayesian inference, with prior distributions that admit negative and positive values: thus, the gains
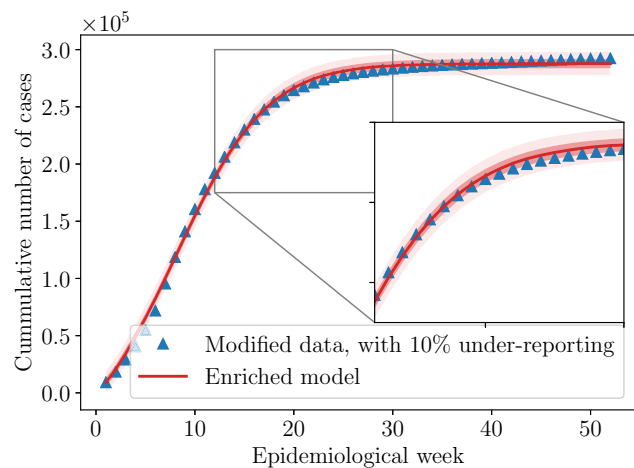
can define both negative and positive feedbacks. The realizations of the enriched dynamical system may then admit a superposition between negative and positive feedbacks, generating a kind of competition between the model's stimuli signals. This competition stabilizes some coordinates of the state vector and destabilizes others, while the global effect materializes in the corrected (enriched) epidemic curve.

To understand more deeply why this competition between corrective signals produces such an effective correction, consider the injection and removal of information (i.e., energy) in the system, as well as its flow between the different coordinates of the state (groups of human and mosquito populations). To make an analogy with
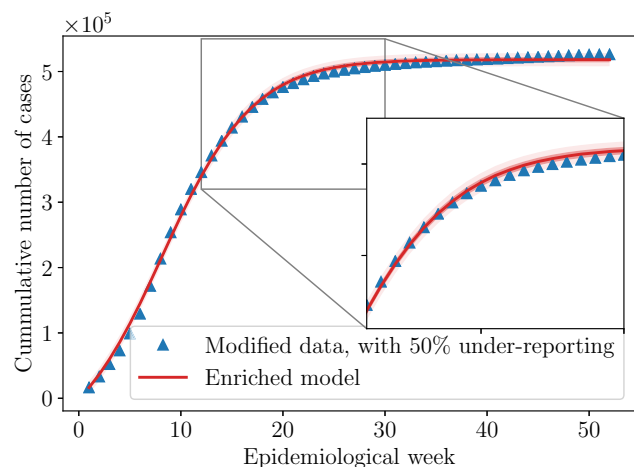


**FIG. 7.** Modified outbreak data, assuming 50% under-reporting, and an enriched model response.

the dynamics of mechanical oscillators, the feedback effects proportional to the state derivative produce a kind of "viscous force," which introduces (via a positive feedback) or removes (via a negative feedback) information into or from the epidemiological state variables. Furthermore, the terms proportional to the system state correspond to a kind of "restoring force," which redistributes information among the different population groups. The intensity of this additional information flow between the different coordinates of the system state is controlled by the new time scales induced by the feedback signals, which are nonlinear functions of the gains $\kappa_i$ and $\lambda_i$, $i = 1, \ldots, 7$.

It should be noted that the approach presented in the paper is neither control theory in the literal sense, since no information from the biological system is obtained in real-time, nor is an action signal sent to the real system to adjust its epidemic curve trajectory. Therefore, the observability and controllability issues related to the real system are not taken into account. The notion of duality between parameters and gains is explored above only as a way to pave an initial reasonable interpretation of how the discrepancy operator acts to correct the model's response, by promoting additional information flows between the different compartments of the populations.

## IV. EFFECTS OF UNDER-REPORTING

Now, let us also consider the scenario that the data are in fact under-reported. First, suppose 10% of cases are not reported so that $d_i = 0.9d_i^*$, where $d_i^*$ represents the value of observations we would expect without under-reporting. (This is not claiming that $d_i^*$ is the exact true value, as we still expect an unbiased measurement error.) The discrepancy parameters are re-calibrated, and the corresponding model response is shown in Fig. 6. Again, all (modified) observations are captured by the enriched model response.

Finally, we suppose that only 50% of cases are reported so that $d_i = 0.5d_i^*$. These results are shown in Fig. 7. Even here, the enriched model adapts to this scenario and covers the dynamical behavior of the outbreak in this highly under-reported scenario.

## V. CONCLUSION

This work presents an initial endeavor to represent the model discrepancy of an epidemiological system, namely, the 2016 Brazilian Zika outbreak. Preliminary results are promising—compared to the original model, the embedded discrepancy operator greatly improves the consistency between the model output and available observations.

The general applicability of this method to other epidemiological models is best understood in two parts. One one hand, the formulation of the enriched model equations is immediately applicable to another model, if that model is also comprised of a set of ODEs. That is, nothing prevents a modeler from testing the proposed model enrichment framework in that case. On the other hand, the particular details of the calibration and whether or not this approach is in fact able to capture the discrepancy between the original model and the data may depend on domain specific information. Future studies will test this approach in other outbreak data sets.

Many other open questions remain. In Sec. III D, we discuss two possible interpretations of the embedded discrepancy operator.

First, the linear terms added into the differential equations resemble a type of a linear feedback, with one term proportional to the state variable and one a differential "control" term. In this case, the discrepancy operator is not actually controlling the real system itself but driving the model system to the target (epidemic data). Second, and perhaps more importantly, would be a physiological interpretation of the discrepancy terms. However, beyond the scope of this paper, a deep exploration of interpretability, connections to the linear feedback theory, and an explanation of these discrepancy terms in a physiological sense will be the subject of immediate future work.

Related to the point above, we would also like to understand what the calibrated discrepancy operator implies about the missing dynamics of the reduced model. That is, can we use the learned discrepancy model to infer what the reduced model is most critically lacking? This question is currently under study, also in the context of ecological models (which have a similar structure, as sets of coupled ODEs). Doing so would allow the use of these embedded operators to function as a type of modeling tool, as opposed to only a model correction.

Finally, this study would be perhaps more convincing with more trustworthy data. How to achieve this, though, is just as complex a problem as the epidemiological system itself, as it involves accessibility to healthcare in remote regions, public awareness of mandatory reporting policies, and incentives and rewards for timely reporting of a communicable disease.

## DATA AVAILABILITY

The data are available in a database maintained by the Brazilian Ministry of Health (Ref. 14) and also available as supplementary material in Ref. 7. All data (and code) needed to reproduce results in this paper are also included in the code base at https://doi.org/10.5281/ZENODO.3666845, Ref. 34.

## REFERENCES

[1] M. C. Kennedy and A. O'Hagan, "Bayesian calibration of computer models," J. R. Stat. Soc.: Ser. B **63**, 425–464 (2001).
[2] M. Martcheva, *An Introduction to Mathematical Epidemiology* (Springer, 2015), Vol. 61.
[3] J. C. Frauenthal, *Mathematical Modeling in Epidemiology* (Springer Science & Business Media, 2012).
[4] D. Pfeiffer, T. P. Robinson, M. Stevenson, K. B. Stevens, D. J. Rogers, A. C. Clements *et al.*, *Spatial Analysis in Epidemiology* (Oxford University Press, Oxford, 2008), 142.
[5] K. E. Nelson and C. M. Williams, *Infectious Disease Epidemiology: Theory and Practice* (Jones & Bartlett Publishers, 2014).
[6] Commonly, the model name will specify which sub-populations are included for both species. For example, an SEIR-SEI model includes host sub-populations of

the susceptible, exposed, infected, and recovered and vector sub-populations of susceptible, exposed, and infected.

[7]E. Dantas, M. Tosin, and A. Cunha, Jr., "Calibration of a SEIR–SEI epidemic model to describe the Zika virus outbreak in Brazil," Appl. Math. Comput. **338**, 249–259 (2018).

[8]R. E. Morrison, T. A. Oliver, and R. D. Moser, "Representing model inadequacy: A stochastic operator approach," SIAM/ASA J. Uncertain. Quantif. **6**, 457–496 (2018).

[9]R. E. Morrison, "Embedded discrepancy operators in reduced models of interacting species," arXiv:1910.08191 (2019).

[10]Note that no outbreak of Zika was reported in the years following 2016, and thus, there are no useful data.

[11]M. Carrie and M. Hyman, see sinews.siam.org/Details-Page/mathematical-models-for-fighting-zika-virus for "Mathematical Models for Fighting Zika Virus" (2016); accessed February 5, 2018.

[12]M. M. Bastos and F. C. Coelho, "Estimating under-observation and the full size of the 2016 Zika epidemic in Rio de Janeiro," PLoS ONE **13**(10), e0205001 (2018).

[13]Although it might appear above as though some initial conditions are defined in terms of undefined quantities, this is just an ordering issue. The initial conditions are properly specified in the following order: $C$, $R_h$, $E_h$, $S_h$, $I_v$, $E_v$, and $S_v$.

[14]Secretaria de Vigilância em Saúde, see http://portalarquivos.saude.gov.br/images/pdf/2017/abril/06/2017-002-Monitoramento-dos-casos-de-dengue–febre-de-chikungunya-e-febre-pelo-v–rus-Zika-ate-a-Semana-Epidemiologica-52–2016.pdf for "Boletim Epidemiolólico. monitoramento dos casos de Dengue, febre de Chikungunya e febre pelo vírus Zika até a Semana Epidemiológica 52, Volume 48" (2017); accessed August 18, 2017.

[15]T. F. Coleman and Y. Li, "An interior trust region approach for nonlinear minimization subject to bounds," SIAM J. Optim. **6**, 418–445 (1996).

[16]A. R. Conn, N. I. Gould, and P. L. Toint, *Trust Region Methods* (SIAM, 2000), Vol. 1.

[17]E. Forgoston, L. Billings, and I. B. Schwartz, "Accurate noise projection for reduced stochastic epidemic models," Chaos **19**, 043110 (2009).

[18]M. L. Childs, N. Nova, J. Colvin, and E. A. Mordecai, "Mosquito and primate ecology predict human risk of yellow fever virus spillover in Brazil," Philos. Trans. R. Soc. B **374**, 20180335 (2019).

[19]F. C. Coelho, B. Durovni, V. Saraceni, C. Lemos, C. T. Codeco, S. Camargo, L. M. De Carvalho, L. Bastos, D. Arduini, D. A. Villela *et al.*, "Higher incidence of Zika in adult women than adult men in Rio de Janeiro suggests a significant contribution of sexual transmission from men to women," Int. J. Infect. Dis. **51**, 128–132 (2016).

[20]E. E. Petersen, D. Meaney-Delman, R. Neblett-Fanfair, F. Havers, T. Oduyebo, S. L. Hills, I. B. Rabe, A. Lambert, J. Abercrombie, S. W. Martin *et al.* "Update: Interim guidance for preconception counseling and prevention of sexual transmission of Zika virus for persons with possible Zika virus exposure—United States, September 2016," Morb. Mortal. Wkly. Rep. **65**, 1077–1081 (2016).

[21]I. J. Motta, B. R. Spencer, S. G. Cordeiro da Silva, M. B. Arruda, J. A. Dobbin, Y. B. Gonzaga, I. P. Arcuri, R. C. Tavares, E. H. Atta, R. F. Fernandes *et al.*, "Evidence for transmission of Zika virus by platelet transfusion," N. Engl. J. Med. **375**, 1101–1103 (2016).

[22]L. Chang, M. Duan, G. Sun, and Z. Jin, "Cross-diffusion-induced patterns in an SIR epidemic model on complex networks," Chaos **30**, 013147 (2020).

[23]Y. Gong and M. Small, "Epidemic spreading on metapopulation networks including migration and demographics," Chaos **28**, 083102 (2018).

[24]M. Sun, Y. Lou, J. Duan, and X. Fu, "Behavioral synchronization induced by epidemic spread in complex networks," Chaos **27**, 063101 (2017).

[25]D. H. Silva and S. C. Ferreira, "Activation thresholds in epidemic spreading with motile infectious agents on scale-free networks," Chaos **28**, 123112 (2018).

[26]B. Ghanbari and J. Gómez-Aguilar, "Analysis of two avian influenza epidemic models involving fractal-fractional derivatives with power and Mittag-Leffler memories," Chaos **29**, 123113 (2019).

[27]R. E. Morrison, "Exact model reduction of the generalized Lotka-Volterra equations," arXiv:1909.13837 (2019).

[28]D. Givon, R. Kupferman, and A. Stuart, "Extracting macroscopic dynamics: Model problems and algorithms," Nonlinearity **17**, R55–R127 (2004).

[29]B. Hernández-Bermejo and V. Fairén, "Algebraic decoupling of variables for systems of ODEs of quasipolynomial form," arXiv:1909.13837 (2019).

[30]C. Howson and P. Urbach, *Scientific Reasoning: The Bayesian Approach* (Open Court Publishing, 2006).

[31]E. T. Jaynes, *Probability Theory: The Logic of Science* (Cambridge University Press, 2003).

[32]H. Haario, M. Laine, A. Mira, and E. Saksman, "Dram: Efficient adaptive MCMC," Stat. Comput. **16**, 339–354 (2006).

[33]E. E. Prudencio and K. W. Schulz, "The parallel C++ statistical library 'QUESO': Quantification of uncertainty for estimation, simulation and optimization," in *Euro-Par 2011: Parallel Processing Workshops* (Springer, 2012), pp. 398–407.

[34]R. E. Morrison (2020). "rebeccaem/zika: Initial release," Zenodo. https://doi.org/10.5281/ZENODO.3666845.