

A hybrid framework for compartmental models enabling simulation-based inference

Domenic P.J. Germano ^{a,b,*}, Alexander E. Zarebski ^{a,c,*}, Sophie Hautphenne ^d, Robert Moss ^d, Jennifer A. Flegg ^a, and Mark B. Flegg ^e

^a*The School of Mathematics and Statistics, The University of Melbourne, Parkville, VIC, Australia*

^b*The School of Mathematics and Statistics, The University of Sydney, Camperdown, NSW, Australia*

^c*Pandemic Sciences Institute, University of Oxford, Oxford, United Kingdom*

^d*Melbourne School of Population and Global Health, The University of Melbourne, Parkville, Vic, Australia*

^e*School of Mathematics, Monash University, Clayton, VIC, Australia*

Keywords— Hybrid simulation — stochastic modelling — viral clearance — extinction — compartmental modelling

Abstract

Multi-scale systems often exhibit stochastic and deterministic dynamics. Capturing these aspects in a compartmental model is challenging. Notably, low occupancy compartments exhibit stochastic dynamics and high occupancy compartments exhibit deterministic dynamics. Failing to account for stochasticity in small populations can produce ‘atto-foxes’, e.g. in the Lotka-Volterra ordinary differential equation (ODE) model. This limitation becomes problematic when studying extinction of species or the clearance of infection, but it can be resolved by using discrete stochastic models e.g. continuous time Markov chains (CTMCs). Unfortunately, simulating CTMCs is infeasible for most realistic populations.

We develop a novel mathematical framework, to couple continuous ODEs and discrete CTMCs: ‘Jump-Switch-Flow’ (JSF). In this framework populations can reach extinct states (“absorbing states”), thereby resolving atto-fox-type problems. JSF has the desired behaviours of exact CTMC simulation, but is substantially faster than existing alternatives.

JSF’s utility for simulation-based inference, particularly multi-scale problems, is demonstrated by several case-studies. In a simulation study, we demonstrate how JSF can enable a more nuanced analysis of the efficacy of public health interventions. We also carry out a novel analysis of longitudinal within-host data from SARS-CoV-2 infections to quantify the timing of viral clearance. JSF offers a novel approach to compartmental model development and simulation.

Dynamical systems represent a powerful method for describing the world and have successfully been applied in many fields. However, modelling populations which change in size across multiple scales remains a challenge. Population sizes in biological processes often change over orders of magnitude, e.g. the spread of infectious disease (Anderson et al., 1991), the boom-and-bust of insect populations (Ludwig et al., 1978), and the immune response to infections including HIV (Perelson, 2002), and influenza virus (Baccam et al., 2006).

The dynamics of small populations can be heavily influenced by stochastic effects, while for larger populations these fluctuations often average out, justifying the use of continuum models. When the population under consideration does not change in size over orders of magnitude, there is usually a natural (and obvious) choice between a stochastic or deterministic model. However, for multi-scale models, this remains a challenge. Developing modelling methods that can bridge these scales has been a long-standing goal of applied mathematics (Cotter et al., 2016; Flegg et al., 2014; Isaacson, 2013).

Compartmental models describe how quantities (e.g. the number of cells, people, or molecules) change in a dynamical system. Two popular ways to represent compartmental models are ordinary differential equations (ODEs) and continuous time Markov chains (CTMCs). Many ODEs one encounters are actually “ensemble averages” of CTMCs, although the mathematical justification of this is not always straightforward (Kurtz, 1970; Kurtz, 1971; Kurtz, 1972).

In ODEs, the state of the system changes continuously. This can lead to “atto-fox problems” where populations shrink to infeasibly small sizes where a real population would have gone extinct (Fowler, 2021; Mollison, 1991; Lobry et al., 2015). In CTMCs, the discrete state changes stochastically and there may be absorbing states e.g. extinction of a population. CTMCs offer a more natural description of small populations sizes, however applying them to large populations can be challenging. Moreover, a range of powerful techniques can be brought to bear on ODEs, enabling more thorough analysis. This raises a natural question about what to do when compartment occupancy changes across orders of magnitude.

There are exact and approximate algorithms to simulate CTMCs Simoni et al., 2019. Exact methods (e.g. Doob-Gillespie, first/next reaction, and rejection based methods) are computationally expensive when state transitions occur at a high rate (Sanft et al., 2015). Approximate methods (e.g. Tau-leaping (Gillespie, 2001) and chemical Langevin equations (Rao et al., 2003; Cao et al., 2006; Gillespie, 2000; Gibson et al., 2000)) scale to high transition rates but can have unacceptable

*These authors contributed equally to this work

approximation error. To overcome the limitations of classical approximate methods, over the past two decades, there has been a concerted effort to develop *hybrid* stochastic-deterministic approaches (Simoni et al., 2019).

Some of the approaches to hybrid modelling involve partitioning the transitions in CTMC into fast and slow transitions which can be sampled individually and subsequently synchronised is efficient but complex to implement Simoni et al., 2019. Jump-diffusion differential equations partition the model compartments into *fluid* and *discrete* to construct *hybrid switching jump diffusion* processes Buckwar et al., 2011; Angius et al., 2015. Recently, Kynaston et al., 2023 considered extended systems that represent each compartment with both a continuous and a discrete version and allow the for conversion between them.

Essential for the real-world utility of any mathematical model of a dynamical system is its ability to be calibrated to data. For deterministic compartment models, standard parameter inference methods such as maximum likelihood and posterior sampling via Markov chain Monte Carlo (MCMC) can readily be applied (Ionides et al., 2006). For stochastic compartment models these inference problems are usually much harder. Simulation-based inference has emerged as a way to handle these problems, e.g. *particle filter* (Arulampalam et al., 2002; Kitagawa, 1996) and *approximate Bayesian computation* (Alahmadi et al., 2020).

In this paper, we present a simple and efficient hybrid simulation method: *Jump-Switch-Flow* (JSF). This method enables adaptive transitions between stochastic and deterministic regimes across compartments, ensuring conservation principles are maintained. In this approach, compartments can individually switch regimes, allowing for the compartments to be split into stochastic and deterministic subsets, while still being coupled. We demonstrate the utility and properties of the JSF method through simulation studies and a case study in which we analyse existing longitudinal SARS-CoV-2 viral load dataset.

Two simulation studies are presented to explore the properties of JSF and demonstrate its use in an inference setting. By comparing simulations from an epidemic model, we contrast the computational efficiency and accuracy of JSF with the exact Doob-Gillespie method and Tau-leaping. A simulation study demonstrates the types of insight available when using an approach supporting absorbing (extinction) states.

The significance of being able to do inference with models with absorbing states, i.e. models in which one of the compartments can go extinct, is demonstrated by a case study in which we analyse longitudinal SARS-CoV-2 viral load data Ke et al., 2022 using a TEIRV (Target cells – Eclipsed cell – Infectious cells – Refractory cell – Virions) model to infer the state of host and virus across the infection. Understanding how a viral infection is cleared has important ramifications for treatment and prevention. For example, the initial exposure may fail to initiate a systematic infection, understanding the conditions under which this happens is important for infection prevention Pearson et al., 2011; the infection may reach low levels, potentially escaping detection, but the virus may rebound and cause disease, understanding when the virus has been cleared is important for understanding how long treatment must be maintained. Inferring viral clearance can be computationally challenging Yan et al., 2016, this case study demonstrates how JSF enables the tractable estimation of viral clearance.

Methods

Jump-Switch-Flow mathematical framework

Consider a compartmental model with n compartments $\vec{V} = \{V_i\}_{i=1}^n$ where $V_i(t)$ represents the value of the i th compartment at time t . For example, V_i could be the number of people infected with a pathogen, or the copy number of a molecule in a cell. The state variables V_i may take values from different domains depending upon the resolution needed for the model. For example, in an ODE, \vec{V} will have real values and in a CTMC \vec{V} might have integer values.

Typically, discrete values are used to represent small populations, while larger populations will be represented with a continuum. To accommodate both scales, we model the domain of V_i as $\mathcal{V}_{\Omega_i} = \{0, 1, \dots, \Omega_i\} \cup (\Omega_i, \infty)$. The switching threshold parameter, $\Omega_i \in \mathbb{Z}_{\geq 0}$, is where the i th compartment transitions from discrete to continuous dynamics. If a compartment V_i has a value in $\{0, 1, \dots, \Omega_i\}$, we call it *discrete* (or *jumping*), and if it has a value in (Ω_i, ∞) , we call it *continuous* (or *flowing*). While the switching threshold can be compartment specific, for ease of exposition, we will only consider a single threshold shared between all compartments $\Omega = \Omega_i$. At any moment in time let us assume q of the n compartments are flowing. We use the notation $\vec{V}_F = \{V_i : V_i > \Omega\} \in (\Omega, \infty)^q$ and $\vec{V}_J = \{V_i : V_i \leq \Omega\} \in \{0, 1, \dots, \Omega\}^{(n-q)}$ to represent the compartments in each of the flowing and jumping states, respectively.

The dynamics of each compartment V_i are described by a set of m reactions $\mathcal{R} = \{\mathcal{R}_k\}_{k=1}^m$. Each reaction \mathcal{R}_k is defined by two properties: the rate (per unit time) at which it occurs, λ_k , which may be (and usually is) a function of the state \vec{V} ; and the effect on the state, i.e. the change η_{ik} to the size of compartment V_i when reaction \mathcal{R}_k occurs. As a matrix, $\eta \in \mathbb{Z}^{n,m}$ is referred to as the *stoichiometric matrix*. For ODE models, these reactions occur continuously and are written in the form

$$\frac{d\vec{V}}{dt} = \eta \vec{\lambda}(\vec{V}), \quad (1)$$

while for CTMC models, reactions in the system \mathcal{R} occur as discrete events. In the later case, each reaction \mathcal{R}_k has a separate propensity described by $\lambda_k(\vec{V})$, this propensity remains constant between events but when an event \mathcal{R}_k occurs, there is a change in \vec{V} (as specified by the elements of η_{ik}) and therefore in $\vec{\lambda}(\vec{V})$.

Consider for example the SIR model. The state of this model is $\vec{V} = (S, I, R)^\top$, and there are two “reactions”: infections (\mathcal{R}_1) and recoveries (\mathcal{R}_2). For infections, the rate of reaction is $\lambda_1 = \beta V_1 V_2 / (V_1 + V_2 + V_3)$, and the entries of the

stoichiometric matrix are $\eta_{1,1} = -1$, $\eta_{2,1} = 1$ and $\eta_{3,1} = 0$. For recoveries, the rate of reaction is $\lambda_2 = \gamma V_2$, and the entries of the stoichiometric matrix are, $\eta_{1,2} = 0$, $\eta_{2,2} = -1$, and $\eta_{3,2} = 1$.

We define the subset of reactions $\mathcal{S} \subseteq \mathcal{R}$ to contain those treated as stochastic events. The set \mathcal{S} can be defined in one of two ways. We only use the second in this manuscript, but describe both to assist the presentation. The first way to define \mathcal{S} is as follows; $\mathcal{S} = \{\mathcal{R}_k : \exists i \text{ s.t. } V_i \in \vec{V}_J \text{ and } \eta_{ik} \neq 0\}$. In this definition, a reaction \mathcal{R}_k is included in \mathcal{S} if the reaction has an effect on a jumping (discrete) population $V_i \in \vec{V}_J$. This is a minimum requirement; a reaction should not be permitted to evoke a continuous change in a discrete population. However, we do want to allow reactions to make discrete changes to flowing (continuous) populations. The second way to define \mathcal{S} , which we use throughout this manuscript, captures a larger set of reactions; $\mathcal{S} = \{\mathcal{R}_k : \exists i \text{ s.t. } V_i \in \vec{V}_J \text{ and } (\eta_{ik} \neq 0 \text{ or } \partial_{V_i} \lambda_k \neq 0)\}$. In this definition, a reaction is included in \mathcal{S} if either (1) it causes a change in jumping (discrete) populations *or* (2) it is influenced by a discrete population (perhaps as reactants for example).

Reactions in \mathcal{S} are simulated using stochastically sampled times similar to CTMC models (SI Section 1A). It is important to note that, unlike time homogeneous CTMC models, the propensities are *not* constant because the state \vec{V} (and therefore $\vec{\lambda}$) are continuously varying. When any reaction $\mathcal{R}_k \in \mathcal{S}$ occurs, we say the system has *jumped* and an instantaneous change of η_{ik} for each compartment V_i occurs (irrespective of whether $V_i \in \vec{V}_J$ or $V_i \in \vec{V}_F$ to ensure mass conservation is observed). We will refer therefore to reactions in \mathcal{S} as *jumps*. The reactions in $\mathcal{S}' = \mathcal{R} \setminus \mathcal{S}$ are not stochastic, we call these *flows* because they represent the continual change of value of the relevant compartments, all of which are continuous by definition of \mathcal{S}' (SI Section 1B). At any moment in time, we denote $|\mathcal{S}'| = p = m - |\mathcal{S}|$ to be the number of reactions which are flowing.

Finally, the hybrid model that we propose is capable of *switching*. Switch events are defined as a compartment between \vec{V}_F and \vec{V}_J . These events occur when a compartment's value crosses the switching threshold Ω (SI Sections 1C and 1D). Importantly, switch events can change \mathcal{S} and are paradigm defining events which should occur infrequently compared to jumps (frequent) and flows (continuous).

Due to the way that \mathcal{R} is partitioned, it is possible to order the rows and columns of η at any moment into the upper-triangular block form

$$\eta = \left(\begin{array}{c|c} \eta_{\mathcal{S}'} & \bar{\eta}_{\mathcal{S}} \\ \hline 0 & \eta_{\mathcal{S}} \end{array} \right),$$

where $\eta_{\mathcal{S}'} \in \mathbb{Z}^{q \times p}$, $\eta_{\mathcal{S}} \in \mathbb{Z}^{(n-q) \times (m-p)}$ and $\bar{\eta}_{\mathcal{S}} \in \mathbb{Z}^{q \times (m-p)}$ refer to stoichiometric coefficients for changes in flowing compartments under flows, jumping compartments under jumps, and flowing compartments under jumps, respectively. Written as a system of equations analogous to (1), the hybrid JSF model we propose formally takes the following form. For any time interval $t_0 < t < t_1$ between switching events,

$$\begin{aligned} \frac{d\vec{V}_F}{dt} &= \eta_{\mathcal{S}'} \vec{\lambda}_{\mathcal{S}'}(\vec{V}) + \bar{\eta}_{\mathcal{S}} \vec{\Lambda}_{\mathcal{S}}(\vec{V}), \\ \vec{V}_J(t) &= \vec{V}_J(t_0) + \eta_{\mathcal{S}} \int_{t_0}^t \vec{\Lambda}_{\mathcal{S}}(\vec{V}) \, ds, \end{aligned}$$

where $\vec{\lambda}_{\mathcal{S}'} \in \mathbb{R}^p$ are the reaction rates of flows and $\vec{\Lambda}_{\mathcal{S}}$ is a stochastic vector of $m-p$ delta-function spike trains that are derived from the realisations of $m-p$ different jumps sampled at rates which are dependent on the dynamic changes in the propensities $\vec{\lambda}_{\mathcal{S}} \in \mathbb{R}^{m-p}$ for these jumps (SI Section 1A).

Example: Lotka-Volterra model

The Lotka-Volterra model describes the populations of two species: prey, V_1 , and predators, V_2 . The prey reproduce at a rate α , the predators die at a rate γ and new predators are produced through predation at rate β . When modelled with ODEs, this gives us the following system:

$$\begin{aligned} \frac{dV_1}{dt} &= \alpha V_1 - \beta V_1 V_2, \\ \frac{dV_2}{dt} &= \beta V_1 V_2 - \gamma V_2. \end{aligned} \tag{2}$$

The periodic solutions to (2) persist even when the predator species reaches infeasible population sizes: populations of order 10^{-18} giving us “atto-foxes”.

Fig. 1A shows a compartmental diagram of the Lotka-Volterra model with the reaction: \mathcal{R}_1 , birth of V_1 ; \mathcal{R}_2 , death of V_2 ; and \mathcal{R}_3 , conversion of V_1 into V_2 . Fig. 1B shows a representation of the reactions and their representation with a stoichiometric matrix.

Fig. 1C shows a cartoon trajectory of this process when represented with JSF. There are three switching events (at t_1 , t_2 and t_3), and each configuration of discrete and continuously changing state features:

1. During $[0, t_1]$, the representation is $\vec{V}_F = V_1$ and $\vec{V}_J = V_2$. Reactions: $\mathcal{S} = \{\mathcal{R}_2, \mathcal{R}_3\}$.
2. During $[t_1, t_2]$, the representation is $\vec{V}_F = (V_1, V_2)^\top$ and $\vec{V}_J = \emptyset$. Reactions: $\mathcal{S} = \emptyset$ (ODE regime).

3. During $[t_2, t_3]$, the representation is $\vec{V}_F = V_2$ and $\vec{V}_J = V_1$. Reactions: $\mathcal{S} = \{\mathcal{R}_1, \mathcal{R}_3\}$.

4. From t_3 onwards the representation is $\vec{V}_F = \emptyset$ and $\vec{V}_J = (V_1, V_2)^\top$. Reactions: $\mathcal{S} = \mathcal{R}$ (CTMC regime).

Figs. 1D and E show random trajectories sampled from the Lotka-Volterra process as represented with JSF. Fig. 1D shows several cycles of the two populations and Fig. 1E demonstrates how, with the inclusion of discrete stochastic behaviour, the predator species can go extinct, allowing the prey to grow exponentially. Full details, including description of the algorithms used to sample from the JSF process are given in SI Sections 1 and 2.

Data and Software Availability

All data and code to reproduce our results is available at <https://github.com/DGermano8/JSFGermano2024>. A Python package implementing the Jump-Switch-Flow method is available at <https://dgermano8.github.io/JSF>.

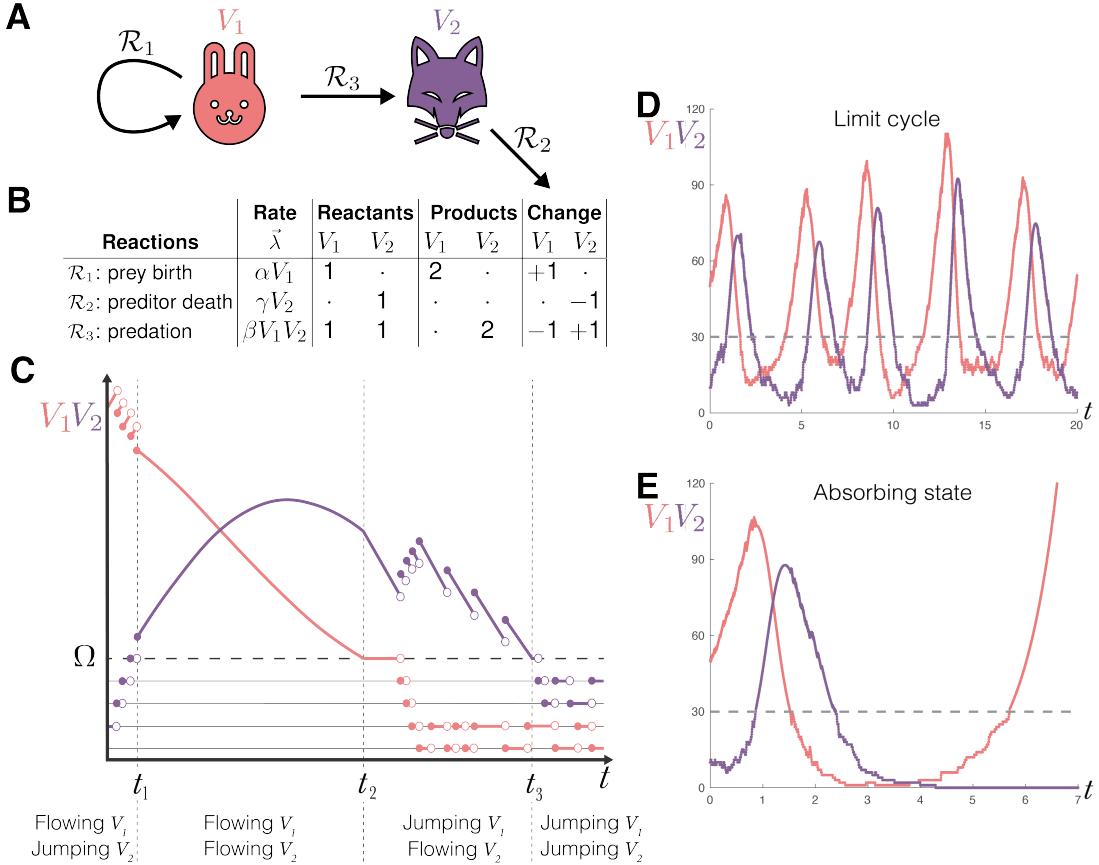


Figure 1: JSF provides a way to capture both the continuous deterministic dynamics of large populations and the discrete stochastic dynamics of small populations. **A** A compartmental model representation of the Lotka-Volterra system ((2)): prey (which reproduce at rate α), predators (which die at rate γ), and their interaction (predating at a rate proportional to β). **B** The compartmental model can be formalised as a reaction network via a stoichiometry matrix η which details each species role in the reactions. The reactants consumed η^- and the products produced η^+ can be written explicitly with $\eta = \eta^+ - \eta^-$. **C** A cartoon example demonstrating how the representation captures the continuous variation of some compartment-reaction pairs (flow) and the discrete stochastic changes (jumps) in others. Transitioning (switching) between a continuous and discrete state occurs at a threshold Ω which is indicated with a horizontal dashed line. **D** An example trajectory from the Lotka-Volterra model exhibiting the characteristic cyclical behaviour with additional domesticity influencing the dynamics at low population sizes. **E** An example trajectory showing the possibility for the predator species to go extinct and subsequent growth of the prey species.

Results

Effects of interventions to reduce transmission

Simulation based inference (e.g. the particle filter) relies on being able to efficiently simulate from the generative process. To demonstrate how JSF can be used in this setting, we first present a simulation study of forecasting the elimination of a infectious pathogen.

SIRS model with demography

The SIRS model with demography is an extension of the classic susceptible-infectious-removed (SIR) model. In the SIR model *susceptible* individuals may be infected by contact with *infectious* individuals; infected individuals eventually cease to be infectious and transition to the *removed* compartment. The SIRS model with demography extends the SIR model by allowing removed individuals to transition back to being susceptible to infection and for individuals to give birth to new (susceptible individuals) and die (at equal rates). Fig. 2A shows a compartmental diagram of the SIRS model with demography.

Fig. 2B shows our simulated (via the Doob-Gillespie algorithm) time series of daily noisy measurements of the prevalence of infection. The parameters used in the simulation are shown in Fig. 2C; these values are broadly consistent with existing estimates for influenza or SARS-CoV-2. We assume that these measurements are drawn from a negative binomial distribution where the expected value is equal to the true prevalence of infection and there is a constant (known) dispersion parameter, $k = 100$.

Estimating elimination probabilities

Combining the SIRS model with a particle filter (Moss, 2024; Kitagawa, 1996) we used the first 100 days of the time series to forecast the remaining 150 days. Fig. 2B shows the estimated true prevalence across the first 100 days and forecasts the prevalence for the subsequent 150 days. Included in the forecast is a daily estimate of the probability that the pathogen has been eliminated (in the simulation the pathogen was eliminated on day 250). It is important to note that there is an endemic equilibrium in this model, so it is reasonable for the probability to plateau as it does.

To demonstrate the utility of the posterior distribution (i.e. the capacity of the particle filter to forecast the epidemic after the 100 days of observed data), we considered the impact of a possible intervention. This intervention, introduced on day 100, reduces the force of infection (and hence the effective reproduction number) by a factor of α . Fig. 2D shows how the probability of eliminating the pathogen increases substantially as we decrease α from 1 (no intervention) to 0.7 (a 30% reduction in transmission).

The full details of the configuration of the particle filter, and the marginal posterior distributions are given in SI Section 3E.

Computational properties

To further investigate the computational properties of JSF and to compare it with exact (Doob-Gillespie) simulation and (approximate) tau-leaping simulation we used repeated simulations from the SIRS model with demography. We partitioned the simulations into three classes: early stochastic extinction, fade-out after a single epidemic, and sustained transmission (SI Section 3.) As seen in Table S3, JSF and exact simulation generate very similar proportions of samples in each of these classes. Two aspects of this system of practical importance are the timing and magnitude of peak infections, and the total number of infections. As can be seen in Fig. S4 the distribution of peak timing and total number of infections are in good agreement between JSF and exact simulation. For the magnitude of peak infections, there is substantially less variability in the JSF samples than the exact samples, but the proportional difference is small.

With regards to computational efficiency, Fig. S7 shows the average time required for simulations using each of JSF, exact simulation and the tau-leaping approximation. We see that for populations of size above about 10^5 , (i.e. approximately the size of a small city) JSF is substantially faster than both exact simulation and the tau-leaping approximation.

SARS-CoV-2 virus clearance informed by longitudinal data

Understanding viral clearance is important when deciding how long treatment must be maintained. This case study demonstrates how JSF enables inference of virus clearance, which hitherto has been computationally challenging Yan et al., 2016; Farrukee et al., 2018. In particular, we use a mathematical model — the TEIRV model described below — to study viral clearance using longitudinal data from 60 individuals infected with SARS-CoV-2 Ke et al., 2022.

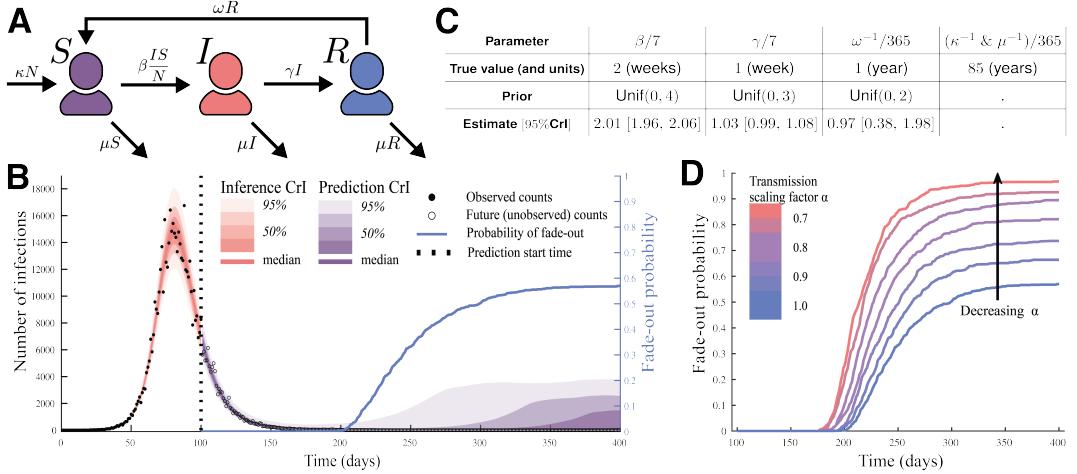


Figure 2: JSF enables us to forecast when a pathogen will be eliminated from a population, i.e. when we can claim that an epidemic has “ended”, and quantify the likely impact of varying levels of intervention. **A** A compartmental model representation of the SIRS with demography. **B** A simulated time series of noisy measurements of prevalence along with inferred prevalence trajectories and forecasts of future prevalence with associated uncertainty bands. The dashed vertical line indicates the date up until which we have data. The solid blue line indicates an estimate of the probability that the pathogen has been eliminated from the population (as opposed to persisting at very low levels). **C** The true parameters used in the simulation along with their posterior estimates and the associated priors used. Each of the marginal posterior distributions peaks tightly about the true value. **D** The probability of pathogen elimination increases with the strength of the intervention against transmission. Near certainty of elimination is achieved with a 30% reduction of transmission.

TEIRV model of within-host viral dynamics

The TEIRV model is an extension of the classic target-infected-virus (TIV) model Perelson, 2002. In the TIV model *target* cells may be *infected* by the virus; before dying, infected cells produce *virus*; and the virus can degrade or infect remaining target cells. The TEIRV extends the TIV model through the inclusion of an *eclipsed* compartment, to model the delay between infection of a target cell and the subsequent production of virus, and a *refractory* compartment, to model heightened antiviral defences of target cells, e.g. through the effects of interferons Ke et al., 2022; Blanco-Melo et al., 2020. To simplify the use of this model a quasi-steady-state approximation for interferon production is employed. This approximation allows us to avoid the need to explicitly model the amount of interferon present.

Fig. 3A shows a compartmental diagram of the TEIRV model. Target cells becoming infected by virions at rate β , which then enter the eclipsed phase. These eclipsed cells then become infectious at rate k . Infectious cells are cleared from the population at rate δ , and produce virions at rate π . Virions are themselves cleared at rate c . The infectious cells recruit interferon, which cause the target cells to become refractory (and hence protected against infection) at rate Φ . The refractory cells return to a naive state as target cells at rate ρ . These assumptions are represented with the following ODE system:

$$\begin{aligned} \frac{dT}{dt} &= -\beta VT - \Phi IT + \rho R, \\ \frac{dE}{dt} &= \beta VT - kE, \\ \frac{dI}{dt} &= kE - \delta I, \\ \frac{dV}{dt} &= \pi I - cV, \\ \frac{dR}{dt} &= \Phi IT - \rho R. \end{aligned} \tag{3}$$

The stoichiometric matrix corresponding to these assumptions is shown in Fig. 3B. Fig. 3C, which shows a trajectory sampled from the TEIRV model when represented with JSF; the classic boom-bust dynamics of the viral populations can be seen in the exponential growth and decline of the V compartment. Unlike solutions of the ODE model in (3), we see that by time $t = 14$, the populations of eclipsed and infected cells, and the virus have gone extinct, indicating a definitive end to the infection.

The basic reproduction number, \mathcal{R}_0 is a fundamental quantity of epidemiological models. For the TEIRV model, \mathcal{R}_0 can be calculated from the next-generation matrix Diekmann et al., 2010:

$$\mathcal{R}_0 = \frac{\pi \beta T(0)}{\delta c}. \tag{4}$$

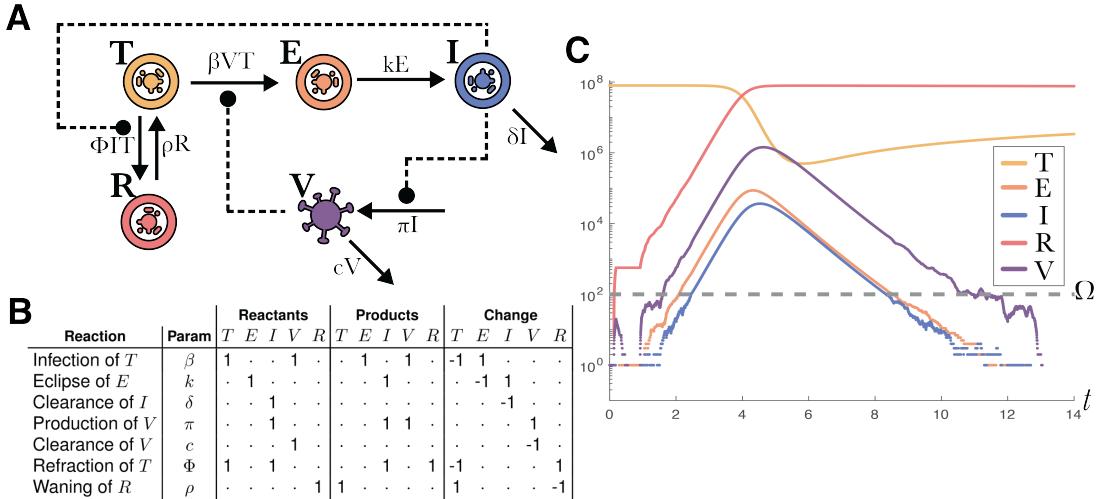


Figure 3: The TEIRV model describes the within-host dynamics of host cells and virus during infection. The model accounts for delays in virus production post cellular infection and the capability of target cells to enter a refractory state as a defence against infection. **A** A compartmental diagram of the TEIRV model showing the interactions between: *Target cells (T)*, *Refractory cells (R)*, cells in the *Eclipsed* phase of infection (*E*) and *Infectious* cells (*I*); and the *Virions (V)*. Solid arrows represent the exchange of mass (cells or virions) through compartments, while dashed lines represent the influence of a compartment on a reaction's rate of occurrence. **B** The representation of this model as a reaction network with a stoichiometric matrix. This formalises the dependency on different variables indicated by the dashed arrows in the compartmental diagram. **C** An example trajectory showing the exponential growth and decline in the amount of free virus across the duration of the infection, this process terminates in the virus, eclipse and infected cell populations reaching zero (shortly before time 12). Reaching zero which is possible in this hybrid continuous/discrete stochastic model but does not occur in a purely ODE based representation.

Observation model linking virus to time series

The longitudinal data contains cycle number (CN) values from nasal samples. The CN values are inversely proportional to the number of virions present. In our analysis we used an existing (empirical) model to link the cycle numbers to the (logarithmic) viral genome load Ke et al., 2022: $\log_{10} V = 11.35 - 0.25\text{CN}$. We assume the observed values are drawn from a normal distribution with mean $\log_{10} V$ and unit standard deviation, and that the values are truncated at the detection limit of -0.65 .

Estimating virus reproduction and clearance

Recall that the particle filter is capable of combining a mechanistic model of the within-host dynamics, such as the TEIRV, and an observation model, such as the viral load measurements, and will return a (Bayesian) posterior sample of both the parameters of the process and the trajectory of virus and cell populations through time. We assume that the infection begins with a single exposed target cell (in a population of 8×10^7 target cells) Ke et al., 2022. This gives us an initial condition for the process: $T(0) = 8 \times 10^7$, $E(0) = 1$, $I(0) = 0$, $R(0) = 0$. We leave the initial viral load, $V(0)$, as a parameter to be fit to the data.

We selected six patient time series Ke et al., 2022, choosing ones that contained a full 14 data points, for both consistency and simplicity. Two of the model parameters were fixed as in previous analysis Ke et al., 2022: $c = 10$, $k = 4$. See SI Section 4 for full details of the particle filter and JSF configuration.

Figure 4A shows our model fits to the first 10 days of viral load data for the six selected patients. After day 10, using the estimated parameters, we generate and predict the distribution of subsequent viral load until day 20. The estimated viral peak coincides with the data for each of the patients and the predicted viral trajectories closely match subsequent observations.

We estimate the probability of viral clearance (right blue axis), over time, given by the blue lines, where viral clearance is said to occur when all of the virion, eclipsed, and infected cell populations become extinct.

There is clear heterogeneity in the amount of time required to clear the virus (Figure 4A). We estimate that with high certainty patients 423192, 443108 and 445602 clear the virus within 20 days. In contrast, for patients 444332, 444391 and 451152 we infer that they have not cleared the virus within that time frame. Within those who did clear the virus, we observe that different patients require different amounts of time to clear the virus. Specifically, patient 432192 (4A) is inferred to have cleared the virus first. Patients 443108 and 445602 both require an estimated 16 days to obtain viral clearance.

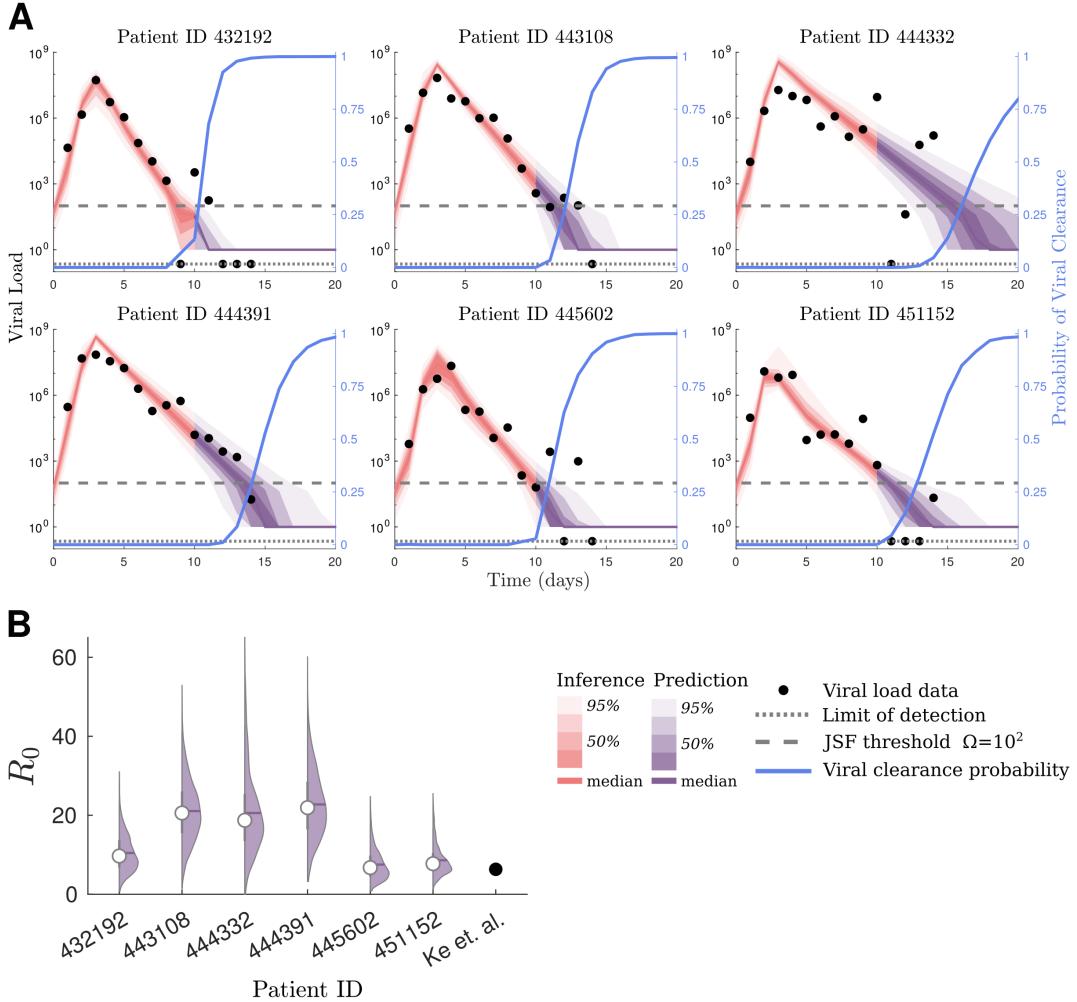


Figure 4: JSF enables the inference of the viral load through time and the probability that the virus has been cleared from longitudinal data of SARS-CoV-2 viral load measurements. **A** Model fits to viral load data based on nasal swabs of six patients from Ke et al., 2022 using a refractory cell within-host model, using our Jump-Switch-Flow method. Nasal viral load data is shown as solid dots, and the limit of detection is shown as a dotted line. We show the switching threshold, $\Omega = 10^2$ by a dashed line. The red and purple bands show mean (0%), 25%, 50%, 75% and 95% credible intervals (from dark to light) for the inference and prediction, respectively. We also estimate the probability of viral clearance at a given point in time, based on the previous data to that point, in blue (light). **B** The posterior distribution of viral R_0 for each patient, along with the median (white circle), These patient specific estimates are contrasted with the single point estimate from previous analysis Ke et al., 2022.

The parameter estimates and priors, are given in SI Section 4. Because of potential identifiability issues with the rate parameters, we instead compare the estimated reproduction number, R_0 , to results from previous analysis Ke et al., 2022. Figure 4B shows the posterior distributions of R_0 , as well as the point estimate of Ke et al., 2022. Since the previous analysis, Ke et al., 2022, reports identical estimates for π , β and δ across these patients, we only include a single point estimate for their work. Our estimate is consistent with previous results Ke et al., 2022 for patients 432192, 445602 and 451152. However, patients 443108, 444332 and 444391 all have higher R_0 estimates in our results. Our R_0 estimates are consistent with similar within-host viral infection analyses of other respiratory pathogens (Baccam et al., 2006; Hernandez-Vargas et al., 2020; Gubbins, 2024).

Discussion

We have presented a simple efficient hybrid simulation method for compartmental models: *Jump-Switch-Flow* (JSF). JSF allows compartments to dynamically change between stochastic and deterministic behaviour. Combining JSF with a simulation based inference method, e.g. a particle filter, enables inference for multi-scale models, in particular when there are atto-type problems where the discrete nature of small populations is important. We demonstrated the properties of JSF in extensive simulation studies and applied it to the analysis of SARS-CoV-2 viral load data.

Simulations shows JSF produces trajectories largely consistent with gold standard exact simulation techniques, e.g. Doob-

Gillespie algorithm (see Figs. S4 and S6). However, JSF is much faster for realistic population sizes (see Fig. S7). We demonstrate how our approach can be incorporated into a particle filter (Moss, 2024), to perform parameter and state estimation and prediction (see Fig. 2). A strength of JSF is the capacity to infer epidemic fade-out, which is important for calibrating intervention measures.

We analysed viral load for SARS-CoV-2 infections using the JSF within a particle filter with a TEIRV, refractory cell model (Ke et al., 2022). In the subset of the data considered, we find consistent parameter values between patients (SI Section 5). In a novel analysis, we estimated the probability of viral clearance through time, finding substantial heterogeneity in the time until viral clearance (see Fig. 4A). This is important as an accurate quantification of when the virus has been cleared is crucial for determining appropriate treatment regimes.

A key advantage of JSF is the ability to combine stochasticity of discrete model for small populations with convenient deterministic models for large populations. The point at which a compartment will transition between these descriptions is specified by the threshold parameters Ω_i (one for each compartment). Setting these parameters to low values speeds up computation, while setting them higher captures more of the stochasticity in the process. As demonstrated in the simulation studies, an appropriate value can be determined through some preliminary simulations. An important consideration in selecting this parameter is ensuring that the absorbing states can still be reached (e.g. compartment extinction), and that stable and steady states can be perturbed by random fluctuations. Furthermore, developing theory to better understand when a CTMC is well approximated by an ODE is at the core to determine how the switching threshold should be chosen. However, this lies outside the scope of this work.

Describing a model via stoichiometric matrices forces systematic consideration of exchange between compartments, encouraging realistic descriptions of physical systems. This formalism can be assisted with Petri nets, which more formally describe such systems (Gilbert et al., 2006) and are an area for future work. Additional extensions include the incorporation of an intermittent Stochastic Differential Equation approximation between the CTMC and ODE regimes, and extension to spatial processes which exhibit both stochastic and deterministic behaviours.

The modelling framework presented in this paper has the potential to change the way compartmental models are developed and calibrated, moving us towards more accurate and more efficient hybrid methods. These models will help to form the basis of informed decision making, based on realistic and accurate descriptions of the system. This has broad applicability, from ecological models, chemical systems and single cell models, to infectious diseases and within-host models.

Acknowledgement

We thank Ada Yan for helpful discussions regarding within-host modelling. J.A.F.’s research is supported by the Australian Research Council (DP200100747, FT210100034) and the National Health and Medical Research Council (APP2019093). S.H.’s research is supported by the Australian Research Council (DP200101281).

Author Contributions

J.A.F and M.B.F conceptualized research; D.P.J.G, A.E.Z, S.H., R.M, J.A.F and M.B.F designed research; D.P.J.G and A.E.Z performed research; D.P.J.G, A.E.Z and R.M contributed software tools; D.P.J.G and A.E.Z analyzed data; J.A.F and M.B.F supervised research; D.P.J.G and A.E.Z wrote the paper; D.P.J.G, A.E.Z, S.H., R.M, J.A.F and M.B.F reviewed the paper; J.A.F acquired funding.

References

- Alahmadi, Amani A., Jennifer A. Flegg, Davis G. Cochrane, Christopher C. Drovandi, and Jonathan M. Keith (2020). “A comparison of approximate versus exact techniques for Bayesian parameter inference in nonlinear ordinary differential equation models”. In: *Royal Society Open Science* 7.3, p. 191315. doi: [10.1098/rsos.191315](https://doi.org/10.1098/rsos.191315).
- Anderson, R.M and R.M. May (1991). *Infectious diseases of humans: dynamics and control*. Oxford University Press.
- Angius, A., G. Balbo, M. Beccuti, E. Bibbona, A. Horvath, and R. Sirovich (2015). “Approximate analysis of biological systems by hybrid switching jump diffusion”. In: *Theoretical Computer Science* 587, pp. 49–72. doi: [10.1016/j.tcs.2015.03.015](https://doi.org/10.1016/j.tcs.2015.03.015).
- Arulampalam, M.S., S. Maskell, N. Gordon, and T. Clapp (2002). “A tutorial on particle filters for online nonlinear/non-Gaussian Bayesian tracking”. In: *IEEE Transactions on Signal Processing* 50.2, pp. 174–188. doi: [10.1109/78.978374](https://doi.org/10.1109/78.978374).
- Baccam, P., C. Beauchemin, C.A. Macken, F.G. Hayden, and A.S. Perelson (2006). “Kinetics of Influenza A Virus Infection in Humans”. In: *Journal of Virology* 80.15, pp. 7590–7599. doi: [10.1128/jvi.01623-05](https://doi.org/10.1128/jvi.01623-05).
- Blanco-Melo, Daniel, Benjamin E. Nilsson-Payant, Wen-Chun Liu, Skyler Uhl, Daisy Hoagland, Rasmus Møller, Tristan X. Jordan, Kohei Oishi, Maryline Panis, David Sachs, Taia T. Wang, Robert E. Schwartz, Jean K. Lim, Randy A. Albrecht, and Benjamin R. tenOever (May 2020). “Imbalanced Host Response to SARS-CoV-2 Drives Development of COVID-19”. In: *Cell* 181.5, 1036–1045.e9. doi: [10.1016/j.cell.2020.04.026](https://doi.org/10.1016/j.cell.2020.04.026).

- Buckwar, E. and M.G. Riedler (2011). “Runge–Kutta methods for jump-diffusion differential equations”. In: *Journal of Computational and Applied Mathematics* 236.6, pp. 1155–1182. DOI: 10.1016/j.cam.2011.08.001.
- Cao, Y., D.T. Gillespie, and L.R. Petzold (Jan. 2006). “Efficient step size selection for the tau-leaping simulation method”. In: *The Journal of Chemical Physics* 124.4, p. 044109. DOI: 10.1063/1.2159468.
- Cotter, S.L. and R. Erban (2016). “Error Analysis of Diffusion Approximation Methods for Multiscale Systems in Reaction Kinetics”. In: *SIAM Journal on Scientific Computing* 38.1, B144–B163. DOI: 10.1137/14100052X.
- Diekmann, O., J. A. P. Heesterbeek, and M. G. Roberts (2010). “The construction of next-generation matrices for compartmental epidemic models”. In: *Journal of The Royal Society Interface* 7.47, pp. 873–885. DOI: 10.1098/rsif.2009.0386.
- Farrukee, R., A. E. Zarebski, J. M. McCaw, J. D. Bloom, P. C. Reading, and A. C. Hurt (2018). “Characterization of Influenza B Virus Variants with Reduced Neuraminidase Inhibitor Susceptibility”. In: *Antimicrobial Agents and Chemotherapy* 62.11, 10.1128/aac.01081-18. DOI: 10.1128/aac.01081-18.
- Flegg, M.B., S.J. Chapman, L. Zheng, and R. Erban (2014). “Analysis of the Two-Regime Method on Square Meshes”. In: *SIAM Journal on Scientific Computing* 36.3, B561–B588. DOI: 10.1137/130915844.
- Fowler, A.C. (Aug. 2021). “Atto-Foxes and Other Minutiae”. In: *Bulletin of Mathematical Biology* 83.10, p. 104. DOI: 10.1007/s11538-021-00936-x.
- Gibson, M.A. and J. Bruck (2000). “Efficient Exact Stochastic Simulation of Chemical Systems with Many Species and Many Channels”. In: *The Journal of Physical Chemistry A* 104.9, pp. 1876–1889. DOI: 10.1021/jp993732q.
- Gilbert, D. and M. Heiner (2006). “From Petri Nets to Differential Equations – An Integrative Approach for Biochemical Network Analysis”. In: *Petri Nets and Other Models of Concurrency - ICATPN 2006*. Ed. by Susanna Donatelli and P. S. Thiagarajan. Berlin, Heidelberg: Springer Berlin Heidelberg, pp. 181–200.
- Gillespie, D.T. (1976). “A general method for numerically simulating the stochastic time evolution of coupled chemical reactions”. In: *Journal of Computational Physics* 22.4, pp. 403–434. DOI: 10.1016/0021-9991(76)90041-3.
- (July 2000). “The chemical Langevin equation”. In: *The Journal of Chemical Physics* 113.1, pp. 297–306. DOI: 10.1063/1.481811.
- (2001). “Approximate accelerated stochastic simulation of chemically reacting systems”. In: *The Journal of chemical physics* 115.4, pp. 1716–1733.
- Gubbins, S. (2024). “Quantifying the relationship between within-host dynamics and transmission for viral diseases of livestock”. In: *Journal of the Royal Society Interface* 21.211, p. 20230445. DOI: 10.1098/rsif.2023.0445.
- Hernandez-Vargas, E.A. and J.X. Velasco-Hernandez (2020). “In-host Mathematical Modelling of COVID-19 in Humans”. In: *Annual Reviews in Control* 50, pp. 448–456. DOI: 10.1016/j.arcontrol.2020.09.006.
- Ionides, E.L., C. Bretó, and A.A. King (2006). “Inference for nonlinear dynamical systems”. In: *Proceedings of the National Academy of Sciences* 103.49, pp. 18438–18443. DOI: 10.1073/pnas.0603181103.
- Isaacson, S.A. (Aug. 2013). “A convergent reaction-diffusion master equation”. In: *The Journal of Chemical Physics* 139.5, p. 054101. DOI: 10.1063/1.4816377.
- Ke, R., P.P. Martinez, R.L. Smith, L.L. Gibson, A. Mirza, M. Conte, N. Gallagher, C.H. Luo, J. Jarrett, R. Zhou, A. Conte, T. Liu, M. Farjo, K.K.O. Walden, G. Rendon, C.J. Fields, L. Wang, R. Fredrickson, D.C. Edmonson, M.E. Baughman, K.K. Chiu, H. Choi, K.R. Scardina, S. Bradley, S.L. Gloss, C. Reinhart, J. Yedetore, J. Quicksall, A.N. Owens, J. Broach, B. Barton, P. Lazar, W.J. Heetderks, M.L. Robinson, H.H. Mostafa, Y.C. Manabe, A. Pekosz, D.D. McManus, and C.B. Brooke (May 2022). “Daily longitudinal sampling of SARS-CoV-2 infection reveals substantial heterogeneity in infectiousness”. In: *Nature Microbiology* 7.5, pp. 640–652. DOI: 10.1038/s41564-022-01105-z.
- Kitagawa, Genshiro (1996). “Monte Carlo Filter and Smoother for Non-Gaussian Nonlinear State Space Models”. In: *Journal of Computational and Graphical Statistics* 5.1, pp. 1–25. DOI: 10.1080/10618600.1996.10474692.
- Klein, R.W. and S.D. Roberts (1984). “A time-varying Poisson arrival process generator”. In: *Simulation* 43.4, pp. 193–195. DOI: 10.1177/003754978404300406.
- Kurtz, T.G. (1970). “Solutions of ordinary differential equations as limits of pure jump Markov processes”. In: *Journal of Applied Probability* 7.1, pp. 49–58. DOI: 10.2307/3212147.
- (1971). “Limit theorems for sequences of jump Markov processes approximating ordinary differential processes”. In: *Journal of Applied Probability* 8.2, pp. 344–356. DOI: 10.2307/3211904.
- (Oct. 1972). “The Relationship between Stochastic and Deterministic Models for Chemical Reactions”. In: *The Journal of Chemical Physics* 57.7, pp. 2976–2978. DOI: 10.1063/1.1678692.
- Kynaston, Joshua C., Christian A. Yates, Anna V. F. Hekkink, and Chris Guiver (2023). “The regime-conversion method: a hybrid technique for simulating well-mixed chemical reaction networks”. In: *Frontiers in Applied Mathematics and Statistics* 9. DOI: 10.3389/fams.2023.1107441.
- Lobry, C. and T. Sari (2015). “Migrations in the Rosenzweig-MacArthur model and the “atto-fox” problem”. In: *Revue Africaine de la Recherche en Informatique et Mathématiques Appliquées* 20, pp. 95–125.
- Ludwig, D., D.D. Jones, and C.S. Holling (1978). “Qualitative analysis of insect outbreak systems: the spruce budworm and forest”. In: *Journal of Animal Ecology* 47.1, pp. 315–332.
- Mollison, D. (1991). “Dependence of epidemic and population velocities on basic parameters”. In: *Mathematical Biosciences* 107.2, pp. 255–287. DOI: 10.1016/0025-5564(91)90009-8.
- Moss, R. (2024). “pypfilt: a particle filter for Python”. In: *Journal of Open Source Software* 9.96, p. 6276. DOI: 10.21105/joss.06276. URL: 10.21105/joss.06276.

- Pearson, John E., Paul Krapivsky, and Alan S. Perelson (Feb. 2011). “Stochastic Theory of Early Viral Infection: Continuous versus Burst Production of Virions”. In: *PLOS Computational Biology* 7.2, pp. 1–17. doi: 10.1371/journal.pcbi.1001058.
- Perelson, Alan S. (Jan. 2002). “Modelling viral and immune system dynamics”. In: *Nature Reviews Immunology* 2.1, pp. 28–36. doi: 10.1038/nri700.
- Rao, C.V. and A.P. Arkin (Mar. 2003). “Stochastic chemical kinetics and the quasi-steady-state assumption: Application to the Gillespie algorithm”. In: *The Journal of Chemical Physics* 118.11, pp. 4999–5010. doi: 10.1063/1.1545446.
- Reboli, N.P., N.G. Bean, and J.V. Ross (2017). “Hybrid Markov chain models of S–I–R disease dynamics”. In: *Journal of Mathematical Biology* 75, pp. 521–541.
- Sanft, K.R. and H.G. Othmer (Aug. 2015). “Constant-complexity stochastic simulation algorithm with optimal binning”. In: *The Journal of Chemical Physics* 143.7, p. 074108. doi: 10.1063/1.4928635.
- Simoni, G., F. Reali, C. Priami, and L. Marchetti (2019). “Stochastic simulation algorithms for computational systems biology: Exact, approximate, and hybrid methods”. In: *WIREs Systems Biology and Medicine* 11.6, e1459. doi: 10.1002/wsbm.1459.
- Yan, Ada W. C., Pengxing Cao, and James M. McCaw (Oct. 2016). “On the extinction probability in models of within-host infection: the role of latency and immunity”. In: *Journal of Mathematical Biology* 73.4, pp. 787–813. doi: 10.1007/s00285-015-0961-5.

Supplementary Information: Jump-Switch-Flow: hybrid stochastic-deterministic solutions of compartmental models

Domenic P.J. Germano ^{a,b,*}, Alexander E. Zarebski ^{a,c,*}, Sophie Hautphenne ^a, Robert Moss ^d, Jennifer A. Flegg ^a, and Mark B. Flegg ^e

^aThe School of Mathematics and Statistics, The University of Melbourne, Parkville, VIC, Australia

^bThe School of Mathematics and Statistics, The University of Sydney, Camperdown, NSW, Australia

^cPandemic Sciences Institute, University of Oxford, Oxford, United Kingdom

^dMelbourne School of Population and Global Health, The University of Melbourne, Parkville, Vic, Australia

^eSchool of Mathematics, Monash University, Clayton, VIC, Australia

S1 Jump-Switch-Flow: mathematical details

In this section, we shall address the mathematical details for how jumps, flows and switches are computed. For completeness, we redefine equations from the main text here. In particular, we recall that the hybrid JSF model we propose formally takes the following form, for any time interval $t_0 < t < t_1$ between switching events,

$$\frac{d\vec{V}_F}{dt} = \eta_{S'} \vec{\lambda}_{S'}(\vec{V}) + \bar{\eta}_S \vec{\Lambda}_S(\vec{V}), \quad (\text{S1})$$

$$\vec{V}_J(t) = \vec{V}_J(t_0) + \eta_S \int_{t_0}^t \vec{\Lambda}_S(\vec{V}) \, ds, \quad (\text{S2})$$

where $\vec{\lambda}_{S'} \in \mathbb{R}^p$ are the reaction rates of flows and $\vec{\Lambda}_S$ is a stochastic vector of $m - p$ delta-function spike trains that are derived from the realisations of $m - p$ different jumps sampled at rates which are dependent on the dynamic changes in the propensities $\vec{\lambda}_S \in \mathbb{R}^{m-p}$ for these jumps. We recall that $\eta \in \mathbb{Z}^{n,m}$ is the stoichiometric matrix, written as:

$$\eta = \left(\begin{array}{c|c} \eta_{S'} & \bar{\eta}_S \\ \hline 0 & \eta_S \end{array} \right), \quad (\text{S3})$$

where $\eta_{S'} \in \mathbb{Z}^{q \times p}$, $\eta_S \in \mathbb{Z}^{(n-q) \times (m-p)}$ and $\bar{\eta}_S \in \mathbb{Z}^{q \times (m-p)}$ refer to stoichiometric coefficients for changes in flowing compartments under flows, jumping compartments under jumps, flowing compartments under jumps, respectively, and $q \leq n$ is the number of flowing compartments.

S1.1 Jump events

The reactions that are defined in \mathcal{S} are stochastic events that produce discontinuous jumps in the state vector \vec{V} . In (S1)-(S2), we denote the jumps using the notation $\vec{\Lambda}_S(\vec{V})$. Each element in the vector $\vec{\Lambda}_S$ corresponds to a reaction in \mathcal{S} . Consider, for example, \mathcal{S}_K the K th reaction in \mathcal{S} and we shall suppose that this reaction corresponds to the k th reaction in the full model $\mathcal{S}_K = \mathcal{R}_k$. The K th element of $\vec{\Lambda}_S$ is

$$\Lambda_{S,K} = \sum_e \delta(t - t_k^{(e)}),$$

where δ is the Dirac measure and $t_k^{(e)}$ is the e th time at which the jump event \mathcal{R}_k takes place. In this way, the term $\vec{\Lambda}_S$ in (S1)-(S2) manifests as discrete jumps in both \vec{V}_F and \vec{V}_J at the moments of each jump event. For the sake of the simulation, the computation of the jump times $t_k^{(e)}$ for each instance e of each reaction $\mathcal{R}_k \in \mathcal{S}$ is all that is required. The stoichiometric coefficients present in (S1)-(S2) then indicate the amplitude of the jumps in each compartment.

The instantaneous propensity for a jump associated with reaction \mathcal{R}_k is $\lambda_k(\vec{V})$. We note that this propensity $\lambda_k(\vec{V})$ is likely to change with time even between jumps due to the continuous change caused by flows. There are a number of ways to sample the jump times $t_k^{(e)}$, see Klein et al., 1984 for a detailed discussion of sampling strategies. We use a variant of the Next Reaction Method (NRM) to sample jump times. We first note that the propensity for a jump is dependent only on the instantaneous state \vec{V} , and therefore if at a current time t_0 there has been $e - 1$ jumps associated with reaction \mathcal{R}_k , it has no bearing on the distribution of the time $t_k^{(e)}$, and therefore we shall simply denote $t_k^{(e)} = t_k$ as the *next* jump time for reaction \mathcal{R}_k . The cumulative probability function from which t_k is sampled is dependent on the current time t_0 and the evolution of the state variables in time $\vec{V}(t)$. In particular, $\text{CDF}(t; k) = 1 - \exp \left\{ - \int_{t_0}^t \lambda_k(\vec{V}(s)) ds \right\}$. To sample t_k , inverse

*These authors contributed equally to this work

transform sampling is used. Specifically, we first sample $u_k \sim \text{Unif}(0,1)$ and then solve $\text{CDF}(t_k; k) = u_k$ for t_k . We define $J_k(t)$ as the *jump clock* for reaction \mathcal{R}_k , noting that u_k and $1 - u_k$ have the same distribution:

$$J_k(t_k) := \log(u_k^{-1}) - \int_{t_0}^{t_k} \lambda_k(\vec{V}(s)) \, ds = 0. \quad (\text{S4})$$

In general, we cannot solve directly for t_k . Instead, we solve for it numerically by tracking the value of $J_k(t)$ as \vec{V} evolves through flows, jumps and switches. For each reaction \mathcal{R}_k , at some initial time (for example $t_k^{(e-1)}$) we sample u_k and initialise $t_0 = t_k^{(e-1)}$. The initial value of $J_k(t)$ is therefore equal to the positive number $\log(u_k^{-1})$. As time progresses, $J_k(t)$ decreases according to (S4) since $\lambda_k \geq 0$. Its value ‘ticks’ down to zero over time and when J_k reaches 0, a jump associated with \mathcal{R}_k is triggered (hence the name jump clock). As a jump clock reaches 0 and a jump is triggered, the clock is reset by sampling a new random number, $u_k \sim \text{Unif}(0,1)$. To update the jump clock, we require numerical integration of $\lambda_k(\vec{V}(t))$ forward in time. Fortunately, we also have piece-wise polynomial approximations for $\vec{V}_F(t)$ as a result of our numerical treatment of the continuous flows (see Subsection S1.2 ‘Flow events’) combined with piece-wise constant values for $\vec{V}_J(t)$ which only update once jumps occur. We will discuss further the numerical integration and jump clock updates in Subsection S1.3 ‘Jump clock updates’.

S1.2 Flow events

Between jumps, \vec{V}_J remains constant and \vec{V}_F evolves continuously according to (S1). In particular,

$$\frac{d\vec{V}_F}{dt} = \eta_{\mathcal{S}'} \vec{\lambda}_{\mathcal{S}'}(\vec{V}). \quad (\text{S5})$$

This is a standard dynamical system of ODEs. We shall numerically integrate (S5) forward in time over discrete time steps Δt using a simple Forward Euler method. However, higher order forward methods may be substituted. In particular,

$$\vec{V}_F(t + \Delta t) = \vec{V}_F(t) + \Delta t \Delta \vec{V}_F(t) = \vec{V}_F(t) + \Delta t \eta_{\mathcal{S}'} \vec{\lambda}_{\mathcal{S}'}(\vec{V}(t)), \quad (\text{S6})$$

noting that $\vec{V}_J(t + \Delta t) = \vec{V}_J(t)$.

S1.3 Jump clock updates

We return to (S4). For a given jump reaction \mathcal{R}_k , a clock is initialised at time t_0 ; and $u_k \sim \text{Unif}(0,1)$, giving $J_k = \log(u_k^{-1})$. Over the course of a time step from t to $t + \Delta t$, it is observed from (S4) that the clock ticks down from J_k to $J_k - \Delta J_k$ where, since Δt is small

$$\Delta J_k = \int_0^{\Delta t} \lambda_k(\vec{V}(t+s)) \, ds = \int_0^{\Delta t} (\alpha + \beta s + O(s^2)) \, ds, \quad (\text{S7})$$

$$\approx \frac{\Delta t}{2} (2\alpha + \beta \Delta t), \quad (\text{S8})$$

$$= \frac{\Delta t}{2} (2\alpha + (\Delta t \Delta \vec{V}_F^\top)(\nabla_{\vec{V}_F} \lambda_k)), \quad (\text{S9})$$

where $\alpha = \lambda_k(\vec{V}(t))$ is simply the propensity of reaction \mathcal{R}_k at time t and $\beta = \frac{d\vec{V}^\top(t)}{dt} \nabla_{\vec{V}} \lambda_k$. Importantly, we know that between jumps \vec{V}'_F is given by (S5) whilst $\vec{V}'_J = 0$. Thus, $\beta = \vec{\lambda}_{\mathcal{S}'}^\top(\vec{V}) \eta_{\mathcal{S}'} \nabla_{\vec{V}_F} \lambda_k$ (evaluated at t). Equation (S9) then follows from (S6).

To calculate the updated jump clock, we subtract ΔJ_k from $J_k(t)$ to get a provisional $J_k(t + \Delta t)$. If $J_k(t) - \Delta J_k > 0$, then no jump occurred during the interval $(t, t + \Delta t)$ and we have the jump clock $J_k(t + \Delta t) := J_k(t) - \Delta J_k$. If instead $J_k(t) - \Delta J_k < 0$ then there is a jump (i.e. a \mathcal{R}_k reaction) during the interval $(t, t + \Delta t)$ (where $0 < \Delta \tau < \Delta t$) which we need to account for in the updated jump clock. Let $t + \Delta \tau$ denote the time at which this jump occurs. We can find $\Delta \tau$ by solving the equation $2\Delta J_k - \Delta \tau(2\alpha + \beta \Delta \tau) = 0$ where ΔJ_k is the residual of the jump clock from t to $t + \Delta \tau$. However, the time $t + \Delta \tau$ ($0 < \Delta \tau < \Delta t$) where the jump occurs can be found by solving $2\Delta J_k - \Delta \tau(2\alpha + \beta \Delta \tau) = 0$ where ΔJ_k is the residual of the jump clock from t to $t + \Delta \tau$.

$$\Delta \tau = \begin{cases} \frac{\sqrt{\alpha^2 + 2\beta \Delta J_k} - \alpha}{\beta}, & \beta \neq 0, \\ \frac{\alpha}{\Delta J_k}, & \beta = 0. \end{cases} \quad (\text{S10})$$

In the case that a jump clock runs out, instead of using Δt to push forward the flows in Equation (S6) we instead use $\Delta \tau$ and implement the jump after the flow to time $t + \Delta \tau$. Subsequently, we reinitialise the jump clock J_k . For a summary of this procedure see Figure S1.

S1.4 Switching events

Switching events describe instances where compartment membership of \vec{V}_J and \vec{V}_F can suddenly change as well as reaction membership in \mathcal{S} and \mathcal{S}' . There are two types of switching events. The first involves the transitioning of a compartment

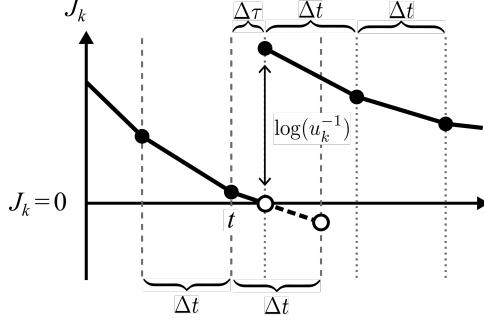


Figure S1: The jump clock J_k counts down until there is a jump change in the state. The jump event occurs between t and $t + \Delta\tau$. To implement, we solve for the time $t + \Delta\tau$ using (S10), we then adjust the flows by using $\Delta\tau$ as a time step instead of Δt in (S6). Once the jump has occurred, we reinitialise $J_k = \log(u_k^{-1})$ where $u_k \sim \text{Unif}(0, 1)$.

from \vec{V}_J to \vec{V}_F . This transition is straightforward as a new equation is added to (S5) the state \vec{V}_F is initialised at the switching time by continuation and initialisation of the new flowing compartment at Ω .

The second type of switching event involves the transition of a compartment from \vec{V}_F to \vec{V}_J . Let V_i be the compartment switching from \vec{V}_F to \vec{V}_J due to a jump event, such that $V_i \leq \Omega$. In general, these types of jump events result in V_i being non-integer, i.e. $V_i \notin \mathcal{V}_\Omega$. To ensure the values of V_i stay in \mathcal{V}_Ω we add another constraint to the process proposed by (Rebuli et al., 2017). Let \hat{V}_i be the value of the flowing compartment V_i after jumping down across the threshold Ω but before being initialised into \mathcal{V}_Ω . We apply the following rule to reinitialise V_i after the switch. We take $V_i = \lceil \hat{V}_i \rceil$ with probability $\hat{V}_i - \lfloor \hat{V}_i \rfloor$, and otherwise we round down by setting $V_i = \lfloor \hat{V}_i \rfloor$. This ensures the expected state of the process after the switch is \hat{V}_i as described under the flowing paradigm from which this compartment has come and that the variable remains in the domain \mathcal{V}_Ω .

S2 Jump-Switch-Flow: implementation details

To draw exact samples from the JSF process requires solving the differential equations for the flowing variables \vec{V}_F . For nonlinear systems these differential equations are usually intractable, so they are numerically integrated. Here we describe an algorithm for approximately sampling trajectories, assuming we have a numerical solver for the differential equations (when viewed as an initial value problem (IVP)), to approximate the solutions.

S2.1 Jump-clock

We implement the JSF algorithm with a slightly different jump clock as defined by J_k in Section S1.1. In particular, equate $u_k = \text{CDF}(t_k; k)$ but do not take the log of this expression before defining J_k in Equation (S4). As a result, in our codes we consider the jump clock \tilde{J}_k :

$$\tilde{J}_k(t) = u_k - \left(1 - \exp\left\{-\int_{t_0}^t \lambda_k(\vec{V}(s)) \, ds\right\}\right), \quad (\text{S11})$$

where $u_k \sim \text{Unif}(0, 1)$ and the time of the previous jump event t_0 . Computing the times for the jump events follows a similar process as above. We compute the jump clock on the regular ODE mesh used to solve \vec{V}_F in Equation (S5). From Equation (S11) we can write $\tilde{J}_k(t + \Delta t)$ in terms of $\tilde{J}_k(t)$ and an integral of the reaction rate:

$$\tilde{J}_k(t + \Delta t) = \tilde{J}_k(t) + \left(\exp\left\{-\int_t^{t+\Delta t} \lambda_k(\vec{V}(s)) \, ds\right\} - 1\right) (\tilde{J}_k(t) + 1 - u_k). \quad (\text{S12})$$

For steps where $\tilde{J}_k(t + \Delta t) > 0$ we can do a single step of this process to get the updated jump clock values. However, if $\tilde{J}_k(t + \Delta t) < 0$ a jump has occurred at some time $t + \Delta\tau \in (t, t + \Delta t)$, and we need to solve $\tilde{J}_k(t + \Delta\tau) = 0$ to find when the jump occurred. Therefore, we require a method to find $\Delta\tau$, where $0 < \Delta\tau < \Delta t$. To do this, we start by letting the right hand side of Equation (S12) equal zero and rearrange to get:

$$\int_t^{t+\Delta\tau} \lambda_k(\vec{V}(s)) \, ds = \ln\left\{\frac{\tilde{J}_k(t) + 1 - u_k}{1 - u_k}\right\}. \quad (\text{S13})$$

If λ_k is a positive constant then:

$$\Delta\tau = \frac{1}{\lambda_k} \ln\left\{\frac{\tilde{J}_k(t) + 1 - u_k}{1 - u_k}\right\}. \quad (\text{S14})$$

If λ_k is not constant, we approximate λ_k , which we call $\hat{\lambda}_k$, using the integration of $\vec{V}(t)$, as per Equation (S5), since we know $\vec{V}(t)$ and $\vec{V}(t + \Delta t)$. Therefore, we also know λ_k at t and $t + \Delta t$. We then write $\hat{\lambda}_k$, at $t + \Delta\tau$ between times t and $t + \Delta t$ via linear interpolation:

$$\hat{\lambda}_k(t + \Delta\tau) \approx \left(1 - \frac{\Delta\tau}{\Delta t}\right) \lambda_k(\vec{V}(t)) + \frac{\Delta\tau}{\Delta t} \lambda_k(\vec{V}(t + \Delta t)). \quad (\text{S15})$$

We can now substitute this approximation into Equation (S13), solve it, and rearrange for $\Delta\tau$. This gives a quadratic in $\Delta\tau$ with the following solution:

$$\Delta\tau = \frac{\sqrt{\left(\Delta t \hat{\lambda}_k(t)\right)^2 + 2\alpha \Delta t \Delta \lambda_k} - \Delta t \hat{\lambda}_k(t)}{\Delta \lambda_k}, \quad (\text{S16})$$

where we have introduced the shorthand $\Delta \lambda_k = \lambda_k(\vec{V}(t + \Delta t)) - \lambda_k(\vec{V}(t))$.

S2.2 Pseudocode details

As with a typical IVP, we require the following data:

- the initial condition of the system, $\vec{V}(0)$;
- the final time $T_{\max} > 0$;
- the stoichiometric Reactant, η^- , and Product, η^+ , matrices, from which we derive the Exchange matrix, $\eta = \eta^+ - \eta^-$;
- a set of the associated propensities of the reactions, \mathcal{R} ;
- a time step size for the numerical integrator, Δt ;
- a vector of switching thresholds, Ω ;

To sample the Jump-Switch-Flow process, we have identified both an *exact* sampler, that samples the process exactly, and an *operator-splitting* sampler, that uses some simplifying approximations that makes sampling substantially faster.

Exact Jump-Switch-Flow sampler

Here, we describe an *exact* algorithm for sampling trajectories from Jump-Switch-Flow. The algorithm to sample *exactly* from Jump-Switch-Flow is given in Algorithm S1. First, initialise the jump clocks (line 3). On each iteration through the ODE loop, we first partition the reactions based on inclusion in \mathcal{S} (line 6). We use the IVP solver to compute $\Delta \vec{V}_F(t)$ based on the flowing reactions (line 7). For each step of the ODE loop, we update the jump clocks (line 9). On lines 10–12, we determine if a jump event has occurred based on the sign of the the Jump-clock. If a jump event has occurred, compute the corresponding event time. On line 13 we then rewind the state (and jump clocks) to when the jump occurred. Resample the appropriate jump clock (line 14), and update the state to account for the jump (line 15). Repeat until the time has reached the maximum desired time.

Operator Splitting Jump-Switch-Flow sampler

Here, we describe the algorithm for the *operator-splitting* algorithm, for sampling trajectories from Jump-Switch-Flow, which utilises some simplifying approximations to improve sampling performance.

The Operator Splitting Jump-Switch-Flow sampler is described in Algorithm S2. Unlike the Exact JSF sampler, when a jump event occurs, we continue to integrate the system with the same flow event, $\Delta \vec{V}_F(t)$, instead of resampling it as before. This, in turn, ensures that the ODE is not continually being resampled. While this adds complexity at the implementation stage, significant computational time saving is achieved.

We first initialise the Jump-Switch-Flow process by initialising all the jump clocks (line 3). On each iteration through the ODE loop, we first partition the reactions based on inclusion in \mathcal{S} (line 6). We use the IVP solver to compute $\Delta \vec{V}_F(t)$ based on the flowing reactions (line 7). We now iterate through this current time-step, Δt , checking if any jump events occur (lines 10 – 24), named the *jump loop*. First, we track how much “relative time” (δt) passes through this jump loop (line 9). We then update the jump clocks to the end of the jump loop, accounting for any relative time that has occurred (line 11). As with the Exact JSF sampler, on lines 12 – 14, we determine if a jump event has occurred using the Jump-clock. This gives us which Reaction has occurred, and at what time, $t + \Delta\tau$. On line 15 we then reverse the system to the time the jump event occurs, noting that we must reverse the excess between when the jump event occurred and the remainder of the time-step. We then resample the associated jump clock that just fired (line 16), and update the whole state (i.e. both flowing and Jumping variables) to the current time (line 17). We update the current time to time $t + \Delta\tau$ (line 18) and track

Algorithm S1 An algorithmic description of the Exact Jump-Switch-Flow sampler.

Require: $\vec{V}(0)$, $\Delta t > 0$ and $T_{\max} > 0$

- 1: Initialise model state
- 2: $t \leftarrow 0$
- 3: Initialise jump clocks $u_i \sim \text{Unif}(0, 1)$ for $i = 1, \dots, N$
- 4: (\triangleright Start ODE loop)
- 5: **while** $t < T_{\max}$ **do**
- 6: Compute Jumping reactions $\mathcal{S}(t)$
- 7: Compute flow event, $\Delta\vec{V}_F(t)$
- 8: (\triangleright Perform jump loop for ODE mesh step)
- 9: Update jump clocks, \tilde{J}_k , to $t + \Delta t$ (\triangleright As per Equation (S9))
- 10: **if** any $\tilde{J}_k \leq 0$ (\triangleright Reaction Occurred) **then**
- 11: Identify first fired Reaction, j
- 12: Compute time of jump event, $\Delta\tau$ (\triangleright As per Equation (S10))
- 13: Reverse jump clocks to time $t + \Delta\tau$
- 14: Resample $u_j \sim \text{Unif}(0, 1)$
- 15: Update both: $\vec{V}_J(t + \Delta\tau) = \vec{V}_J(t) + \eta_j$ **and** $\vec{V}_F(t + \Delta\tau) = \vec{V}_F(t) + \Delta\tau \Delta\vec{V}_F(t) + \eta_j$
- 16: $t \leftarrow t + \Delta\tau$
- 17: **else if** $\tilde{J}_k > 0, \forall k$ **then**
- 18: Update: $\vec{V}_F(t + \Delta t) = \vec{V}_F(t) + \Delta t \Delta\vec{V}_F(t)$
- 19: $t \leftarrow t + \Delta t$
- 20: **end if**
- 21: **end while**

the relative time in the jump loop (line 19). We then continue through the jump look. If no jump events have occurred, we leave the loop (line 21). We then simply update the whole state according to the flow event calculated, accounting for any time spent inside the jump loop (line 24).

Algorithm S2 An algorithmic description of the Operator Splitting Jump-Switch-Flow sampler .

Require: $\vec{V}(0)$, $\Delta t > 0$ and $T_{\max} > 0$

- 1: Initialise model state
- 2: $t \leftarrow 0$
- 3: Initialise jump clocks $u_i \sim \text{Unif}(0, 1)$ for $i = 1, \dots, N$
- 4: (\triangleright Start ODE loop)
- 5: **while** $t < T_{\max}$ **do**
- 6: Compute jumping reactions $\mathcal{S}(t)$
- 7: Compute flow event, $\Delta\vec{V}_F(t)$
- 8: (\triangleright Perform jump loop for ODE mesh step)
- 9: Set $\delta t \leftarrow 0$
- 10: **while** $\Delta t > \delta t$ **do**
- 11: Update jump clocks, \tilde{J}_k , to $t + (\Delta t - \delta t)$
- 12: **if** any $\tilde{J}_k \leq 0$ (\triangleright Reaction Occurred) **then**
- 13: Identify first fired Reaction, j
- 14: Compute time of jump event, $t + \Delta\tau$
- 15: Reverse jump clocks to time $t + \Delta\tau$
- 16: Resample $u_j \sim \text{Unif}(0, 1)$
- 17: Update both: $\vec{V}_J(t + \Delta\tau) = \vec{V}_J(t) + \eta_j$ **and** $\vec{V}_F(t + \Delta\tau) = \vec{V}_F(t) + \Delta\tau \Delta\vec{V}_F(t) + \eta_j$
- 18: $t \leftarrow t + \Delta\tau$
- 19: $\delta t \leftarrow \delta t + \Delta\tau$
- 20: **else if** $\tilde{J}_k > 0, \forall k$ **then**
- 21: Leave while loop
- 22: **end if**
- 23: **end while**
- 24: Update: $\vec{V}_F(t + (\Delta t - \delta t)) = \vec{V}_F(t) + (\Delta t - \delta t) \Delta\vec{V}_F(t)$
- 25: $t \leftarrow t + (\Delta t - \delta t)$
- 26: **end while**

S3 Simulation study: SIRS with demography

We first present how the SIRS model with demography can be expressed with the Jump-Switch-Flow method. For the Jump-Switch-Flow method to be a useful method, it must reproduce behaviour that is representative of the exact process (i.e. the CTMC). That is, given some summary statistic, distributions produced via sampling the JSF process should be comparable to those obtained via sampling the CTMC. Moreover, we require that it does so with acceptable computational efficiency. We show, through simulation experiments, how the Jump-Switch-Flow method compares to the Doob-Gillespie method (Gillespie, 1976) for accuracy, and how our approach can exhibit sensitivity with respect to model inputs. We present how our Jump-Switch-Flow method compares to both the Doob-Gillespie and Tau-Leaping methods.

Figure S2a shows the main reactions of the SIRS model with demography: individuals fall into one of three compartments: susceptible (S), infectious (I) and recovered (R). Here, arrows represent how individuals move and progress through compartments. Specifically, individuals are born into the susceptible compartment. Births occur from individuals in the S , I and R compartments, and therefore depend explicitly on the populations (discrete or continuous) in each of these compartments. Individuals can die within each of the compartments. Finally, individuals may progress through compartments via susceptible individuals experiencing infection, infected individuals experiencing recovery, and recovered individuals experiencing waning immunity.

In order to represent the SIRS model with demography as a Jump-Switch-Flow process, we need to associate rates, reactants and products to each of the arrows (see Figure S2b). Note that the births arrow is split into three arrows depending upon the compartment of the parent. Moreover, one of the key assumptions of any SIR-type model is that the population is homogeneously mixed, which enables us to describe how individuals interact with one another. This enables us to say that susceptible individuals interact with infectious individuals with a rate inversely proportional to the whole population. However, there are likewise infectious individuals interacting with other infectious individuals, and also with recovered individuals. However, these last two interactions (as shown in blue in Figure S2b) do not result in any exchange of individuals between compartments.

Figure S2c shows the SIRS model with demography as a Jump-Switch-Flow process. In this figure, we assume that the S and R compartments are continuous (flowing), while the I compartment is discrete (jumping). To decide which of the arrows/reactions are jumping, i.e. in S , and which are flowing, i.e. in S' , we consider the reactants and products of the arrows: if an arrow has any discrete reactant or product, then that arrow is also jumping, otherwise it is flowing. Table S1 shows the reactants and products of the SIRS model with demography.

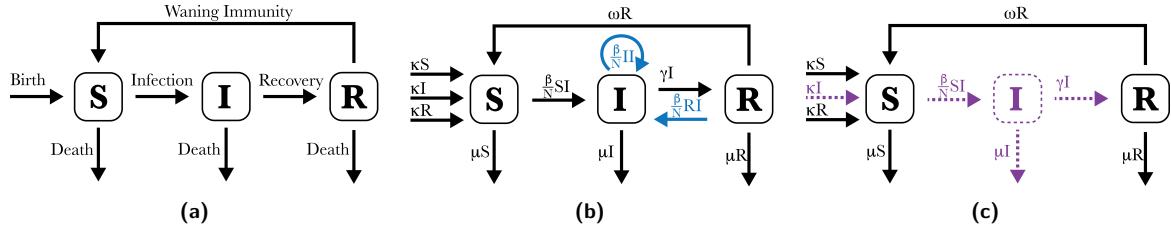


Figure S2: (a) The labelled compartmental SIRS model with demography. (b) The SIRS model with demography where the arrows describe the interactions between the reactants and products. Here, blue arrows do not result in any flow through the system, and can be ignored. (c) The model as a Jump-Switch-Flow system with continuous (flowing) S and R (black), and discrete (jumping) I (purple). Since I is jumping, the reactions involving I are modelled as discrete, jumping reactions.

Reaction	Rate	Reactants, η^-			Products, η^+			η		
		S	I	R	S	I	R	S	I	R
Birth by S	κ	1	.	.	2	.	.	1	.	.
*Birth by I	κ	.	1	.	1	1	.	1	.	.
Birth by R	κ	.	.	1	1	.	1	1	.	.
*Infection of S	β_{SI}/N	1	1	.	.	2	.	-1	1	.
*Infection of I	β_{II}/N	.	2	.	.	2
*Infection of R	β_{RI}/N	.	1	1	.	1	1	.	.	.
*Recovery of I	γ	.	1	.	.	.	1	.	-1	1
Waning of R	ω	.	.	1	1	.	.	1	.	-1
Death of S	μ	1	-1	.	.
*Death of I	μ	.	1	-1	.
Death of R	μ	.	.	1	-1

Table S1: The stoichiometric matrices for the SIRS model with demography. The (*) indicates reactions that are jumping when I is discrete and S and R are continuous. The rows with blue coloured entries do not alter the exchange through the system, and so need not be represented.

For the remainder of this section, we suppose parameter values for the SIRS model with demography that would be typical for a newly introduced, yearly seasonal communicable disease, as presented in Table S2. With these parameter values, on

average, we expect an infected individual to infect 2 other people per week (during the initial outbreak), and be infectious for 1 week, with immunity to the disease lasting for 1 year. For the population turnover, we assume individuals live for, on average, 85 years. Initially there are two infectious individuals (i.e. $I(0) = 2$), no recovered individuals (i.e. $R(0) = 0$), and the remainder of the population is susceptible (i.e. $S(0) = N(0) - I(0)$).

Parameter	β	γ	ω	κ	μ
Rate description	Infection	Recovery	Immunity Waning	Birth	Death
Value	2/7	1/7	1/365	1/(85 × 365)	1/(85 × 365)

Table S2: Parameter values used for SIRS model with demography.

As well as being computationally efficient, our Jump-Switch-Flow method is capable of capturing the inherent stochastic nature of the sampled process. For the SIRS model with demography, this is realised by three key different scenarios:

1. *Extinction*: all of the individuals recover before the disease manages to spread significantly throughout the population (see Figure S3a and S3d);
2. *Fade-out*: the disease spreads throughout the population, however it fades-out in the trough proceeding the first wave (see Figure S3b and S3e);
3. *Endemic*: the disease persists indefinitely within the population (see Figure S3c and S3f).

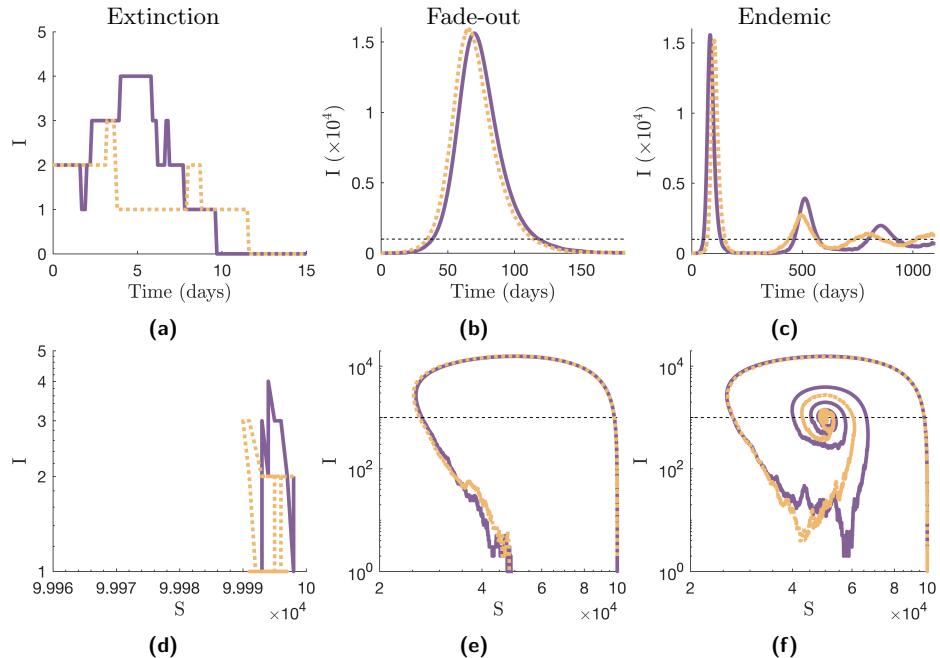


Figure S3: Possible simulated trajectories of SIRS model with demography for an initial population size $N(0) = 10^5$, from Doob-Gillespie (yellow) and Jump-Switch-Flow (purple, $\Omega = 10^3$) methods: (a) two examples of extinction trajectories of the infectious compartment, with associated $S - I$ phase plane in (d); (b) two examples of fade-out trajectories of the infectious compartment, with associated $S - I$ phase plane in (e); (c) two examples of endemic trajectories of the infectious compartment, with associated $S - I$ phase plane in (f).

S3.1 Simulation experiments: Comparing Jump-Switch-Flow and Doob-Gillespie

We compare our Jump-Switch-Flow method to the gold standard approach for simulating exact solutions to CTMCs, the Doob-Gillespie method. To do so, we first specified an initial population size of $N(0) = 10^5$, and a switching threshold of each compartment of $\Omega = 10^3$. We then generated 5,000 simulations using the parameters in Table S2.

To capture the dynamics of the model, we track the infectious compartment as a key compartment of interest. In doing so, we then obtain distributions for the peak number of infectious individuals and also the time for the infection to peak, for the fade-out scenario. We also track the cumulative number of infected individuals after a certain amount of time for the endemic scenario. These results are presented in Figure S5.

Comparing Figures S4a and S4f, we can immediately observe that the distributions for time to peak for Jump-Switch-Flow method appears to be representative of that for the Doob-Gillespie method. This is because the delay in the epidemic take off at low population is fully captured with this sufficiently large choice of switching threshold, $\Omega = 10^3$, enabling the full stochastic effects to be accurately represented.

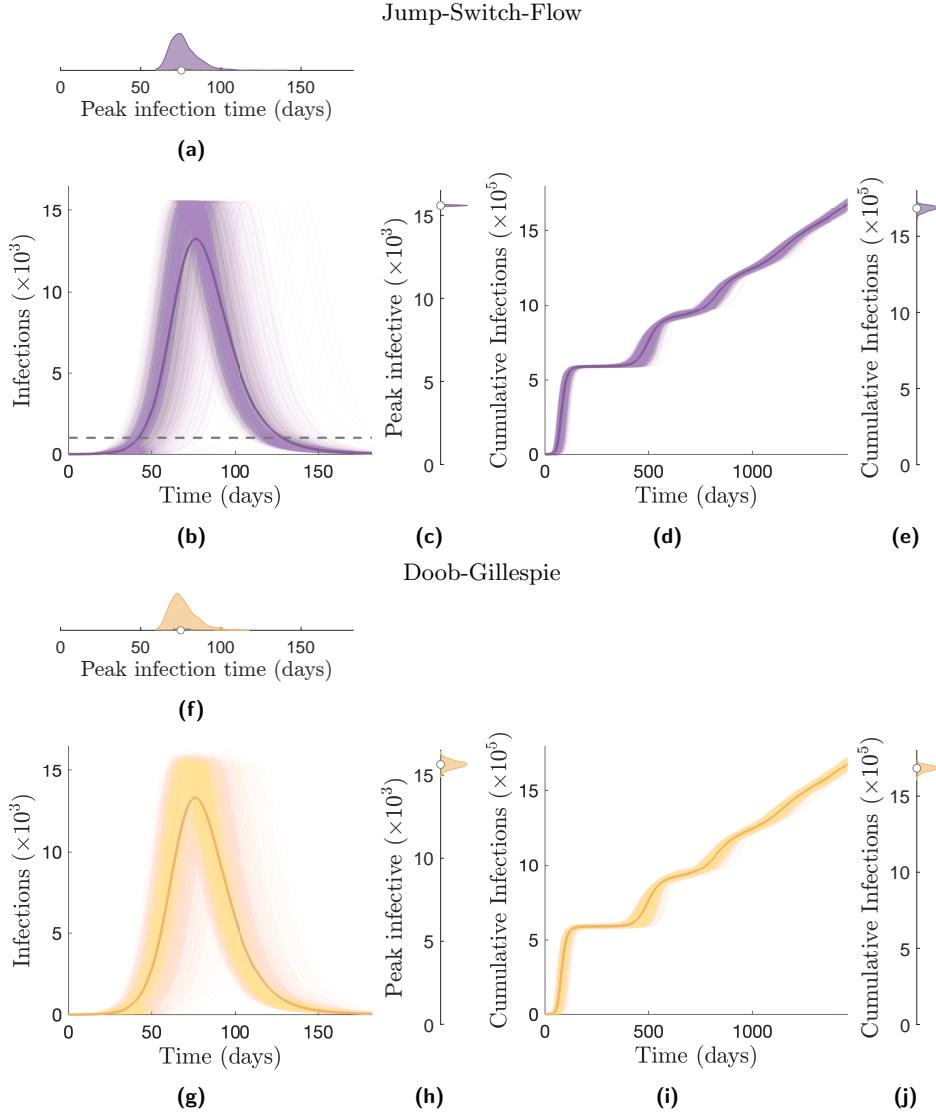


Figure S4: We compare the infectious compartment from realisations for Jump-Switch-Flow (b) and Doob-Gillespie (g) solutions, for a population size of $N(0) = 10^5$. We use a switching threshold of $\Omega = 10^3$. We also show the distributions of the time to infectious peak ((a) and (f)), and infectious peak values ((c) and (h)). We compare the cumulative infections over the simulations for Jump-Switch-Flow (d) and Doob-Gillespie (i) solutions and the final distributions ((e) and (j)). Point averages are also presented as a solid line.

However, if we compare the distributions for the peak number of infectious individuals (S4c and S4h), we can see that they differ significantly, for this chosen switching threshold. This difference can be understood if we consider how the process is behaving. Once the infectious compartment is sufficiently large for the Jump-Switch-Flow method, the method switches from a stochastic representation to a deterministic representation. Therefore, once the infectious compartment becomes deterministic, given that the susceptible population is also deterministic, its peak value is determined by the number of both infectious and susceptible individuals at this time. Since the number of infectious people at this time will always be the same, $I = \Omega = 10^3$ people, any variation in peak value is due to the number of susceptible individuals, S , at the time of switching.

If the summary statistic we instead compare is the distribution of the cumulative number of infected individuals after four years of epidemic circulation, we observe that, even for a relatively low switching threshold, the Jump-Switch-Flow method performs well comparatively to the Doob-Gillespie method, see Figures S4e and S4j. We also note that for these parameter choices and switching threshold, the infectious compartment switches between jumping and flowing states multiple times, yet the Jump-Switch-Flow method produces a cumulative distribution comparable to that of Doob-Gillespie. This indicates that the Jump-Switch-Flow method is robust for long term simulations.

The final summary statistics we are interested in are the probabilities of extinction, fade-out and endemic scenarios. Table S3 shows the probability of each scenario computed from 5,000 simulated samples. Here, we see that both methods produce comparatively similar results, which indicates that the Jump-Switch-Flow method is capable of capturing the inherent stochasticity in the exact simulation of the system via the Doob-Gillespie method.

Method	Extinction	Fade-out	Endemic
Jump-Switch-Flow	0.2476 ± 0.0120	0.4774 ± 0.0138	0.2750 ± 0.0124
Doob-Gillespie	0.2520 ± 0.0120	0.4668 ± 0.0138	0.2812 ± 0.0125

Table S3: Probabilities (with associated 95% confidence interval) of scenario outcomes for Jump-Switch-Flow and Doob-Gillespie, from 5,000 simulations.

We also compare the Jump-Switch-Flow and Doob-Gillespie methods where the switching threshold has been set to the population size, resulting in no compartments switching into the flowing state. These results are further discussed in S3.2, where we show the two methods are indistinguishable.

S3.2 Simulation experiments: Comparing Jump-Switch-Flow (with no switching) and Doob-Gillespie

We compare the Jump-Switch-Flow method to the Doob-Gillespie method. To implement the latter, we set the switching threshold to be the total population size, to ensure no compartments switch into a flowing regime. We can therefore investigate how our Jumping process compares to exact solutions of the Continuous Time Markov Chains.

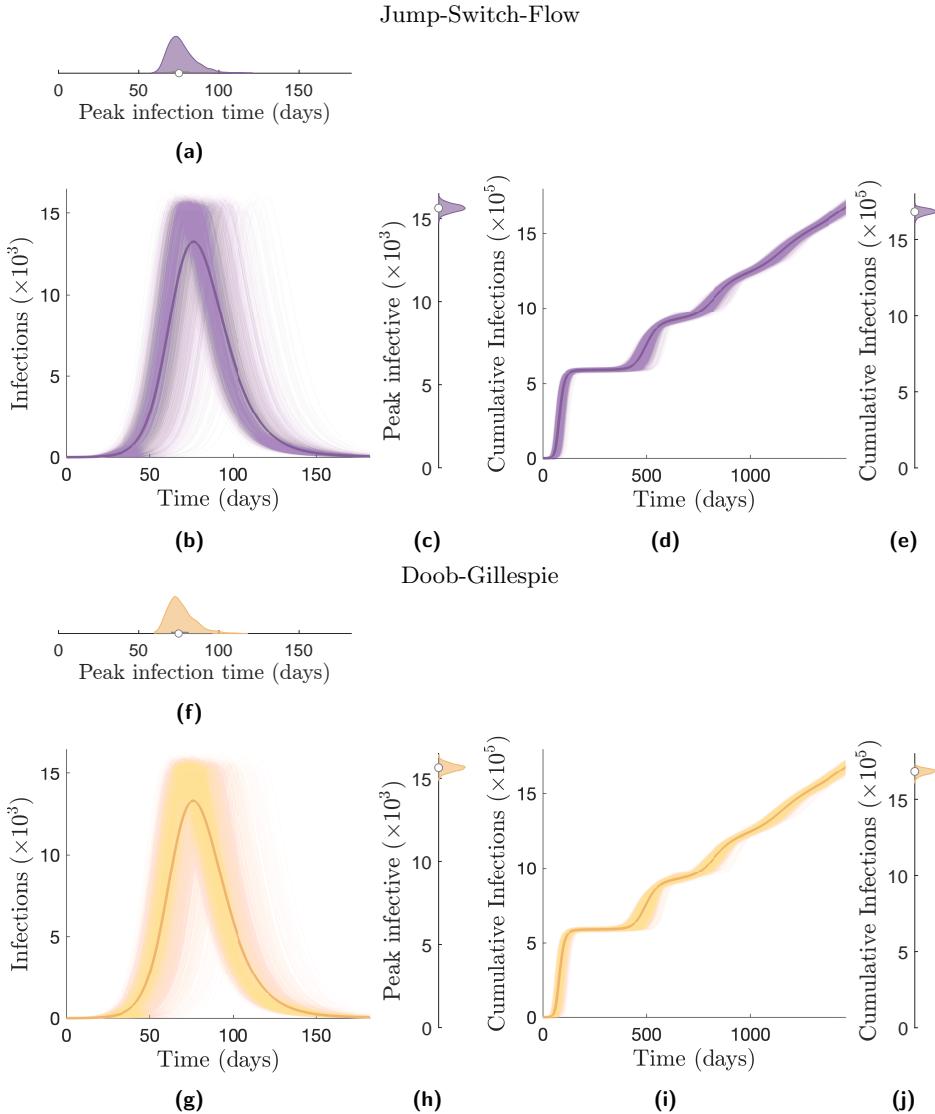


Figure S5: We compare the infectious compartment from realisations for Jump-Switch-Flow (b) and Doob-Gillespie (g) solutions, for a population size of $N_0 = 10^5$. We use a switching threshold of $\Omega = 10^3$. We also show the distributions of the time to infectious peak ((a) and (f)), and infectious peak values ((c) and (h)). We compare the cumulative infections over the simulations for Jump-Switch-Flow (d) and Doob-Gillespie (i) solutions and the final distributions ((e) and (j)). Point averages are also presented as a solid line.

Using parameter values in Table 3 of the main text, we generated 5,000 simulations, tracking the same quantities of interest. As with the case with compartment switching, the peak infection time distribution of JSF (Figure S5a) and Doob-Gillespie (Figure S5f) are comparable. However, this time, as well as a similar median peak infective individuals, the distributions of both JSF (Figure S5c) and Doob-Gillespie (Figure S5h) are also comparable. Lastly, as we saw with compartment switching, the distributions of cumulative infections after four years are comparable for both JSF (Figure S5e) and Doob-Gillespie (Figure S5j).

Table S4 shows that the probabilities of scenario outcomes are also comparable for both Jump-Switch-Flow (with no switching) and Doob-Gillespie.

Method	Extinction	Fade-out	Endemic
Jump-Switch-Flow	0.2512 ± 0.0120	0.4700 ± 0.0139	0.2788 ± 0.0123
Doob-Gillespie	0.2520 ± 0.0120	0.4668 ± 0.0139	0.2812 ± 0.0125

Table S4: Probabilities (with associated 95% confidence interval) of scenario outcomes for Jump-Switch-Flow (with no switching) and Doob-Gillespie, from 5,000 simulations.

S3.3 Simulation experiments: Sensitivity of Jump-Switch-Flow to switching threshold

In this section, we demonstrate how the accuracy of the Jump-Switch-Flow method can exhibit sensitivity with respect to the switching threshold (Figure S6). In Figure S6a, we can see that the distribution of peak number of infective individuals for the fade-out scenario is highly sensitive to the switching threshold. For a small switching threshold ($\Omega = 10^1$), we see that the distribution exhibits low variance, is representative of the deterministic expectation, and is not capturing the distribution obtained with the Doob-Gillespie method. As the switching threshold is increased, we see that the distribution of the peak number of infective individuals increases in variance. However, the Jump-Switch-Flow method is not representative of the Doob-Gillespie method until a large switching threshold ($\Omega = N(0) = 10^5$), where all compartments are in the jumping state.

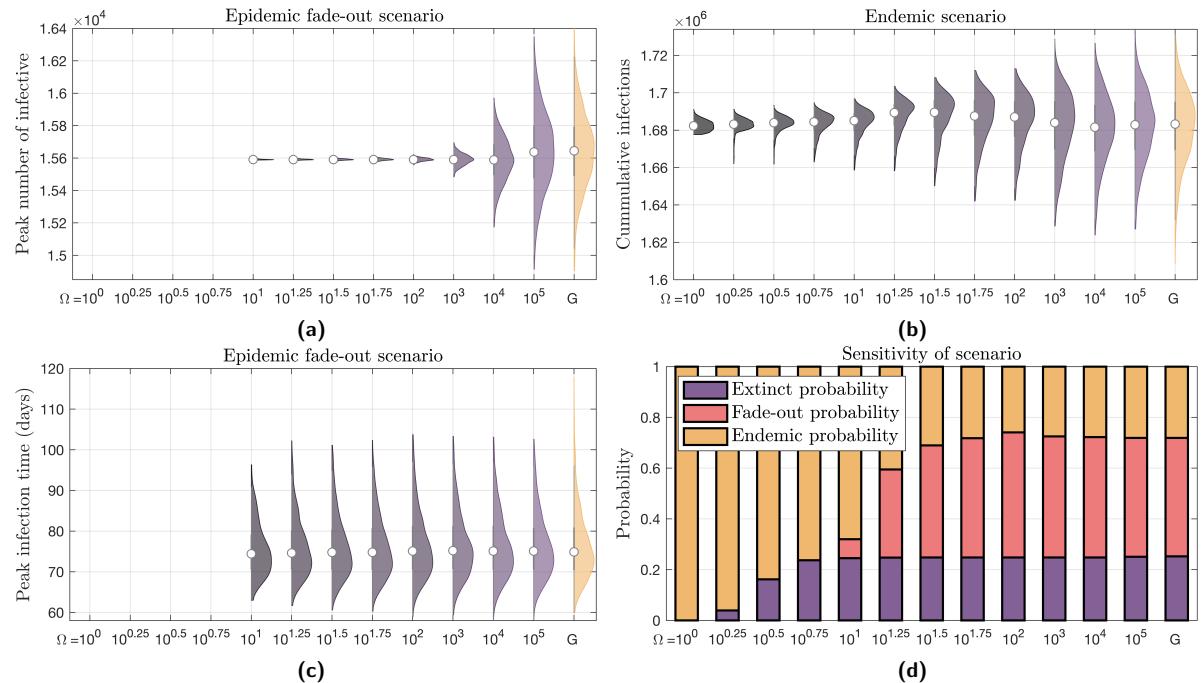


Figure S6: Summary statistics of interest to observe the sensitivity of the Jump-Switch-Flow method with respect to the switching threshold. (a) Sensitivity to the distribution of peak number of infective individuals for the fade-out scenario. (c) Sensitivity to the distribution of peak infection time for the fade-out scenario. (b) Sensitivity to the distribution of the cumulative number of infected individuals after four years for the endemic scenario. (d) Sensitivity to the probability of each scenario occurring. In each of the plots (a)-(c) grey to purple shaded distributions represent results obtained with the Jump-Switch-Flow method whilst the yellow distributions are the gold standard Doob-Gillespie algorithm (for which the placeholder ‘G’ is used to signify) which exhibit exact stochasticity but at the cost of high computational requirements.

Figure S6c presents the distribution of peak infection time for the fade-out scenario for various switching thresholds, and comparing to the Doob-Gillespie method. Here, we see that the Jump-Switch-Flow method with a relatively small switching threshold ($\Omega = 10^2$) is capable of capturing a representative distribution of the Doob-Gillespie method.

Similarly, Figure S6b presents the distribution of the cumulative number of infected individuals after four years for the endemic scenario. Here, we again see that the Jump-Switch-Flow method with a relatively small switching threshold ($\Omega = 10^3$) is also representative of the Doob-Gillespie method.

Lastly, Figure S6d shows the probability of each scenario occurring (extinction, fade-out, endemic). Here, we see that the Jump-Switch-Flow method with a switching threshold of $\Omega \leq 10^1$ over-represents the endemic scenario. However, increasing the switching threshold, we observe that the probability of each scenario occurring quickly becomes representative of those found by the Doob-Gillespie method.

S3.4 Simulation experiments: Computational efficiency of Jump-Switch-Flow

We now present the computational efficiency of the Jump-Switch-Flow method, and compare it to the Doob-Gillespie method and Tau-Leaping method (which is regarded as a computationally efficient, approximate, method).

Figure S7a shows how the computational efficiency varies with the switching thresholds, Ω , for the SIRS model with demography presented in the previous section, and compares it to the Doob-Gillespie method. We see that as we increase the switching threshold, the Jump-Switch-Flow method requires more computational time to simulate the scenario for four years, but is always more efficient than the Doob-Gillespie method.

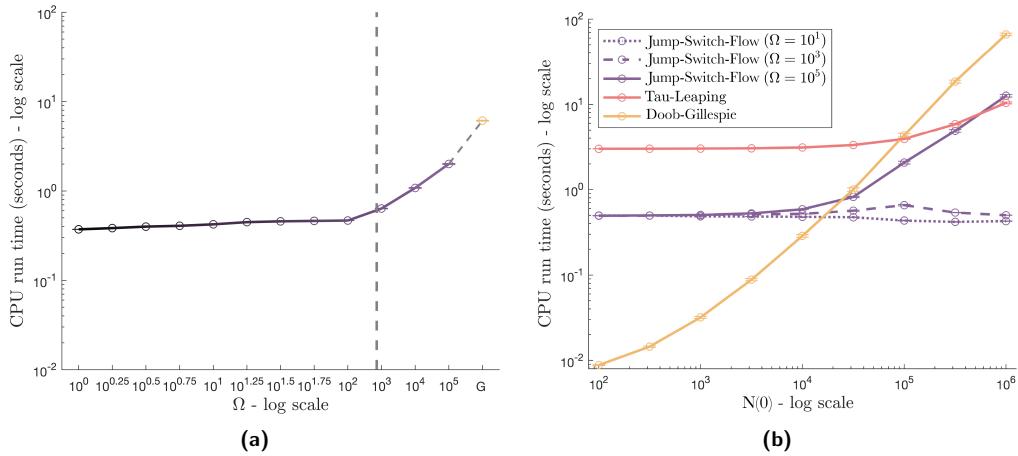


Figure S7: (a) The running time to simulate the scenario for four years for an initial population size of $N(0) = 10^5$, with varying switching threshold, Ω (grey to purple), compared with the Doob-Gillespie ('G') method (yellow). The endemic steady state number of infectious individuals is also presented (vertical dashed line). (b) The running time for the Jump-Switch-Flow method for an appropriate threshold choice is approximately constant as the population size varies (for example $\Omega = 10^3$). If a poor switching threshold is chosen, the running time begins to scale exponentially (for example $\Omega = 10^5$). For the Tau-Leaping method, the running time is constant, and then begins to grow as the rate parameter increases with population size. For the Doob-Gillespie method, the running time scales exponentially with the population size.

To observe how the computational efficiency of our Jump-Switch-Flow method varies with the total (population) size of the system, we fix the switching threshold at $\Omega = 10^3$, as this value is sufficient to capture the majority of the key dynamics we are interested in. Figure S7b compares the Jump-Switch-Flow method (purple) to the Doob-Gillespie method (yellow) and to the computationally more efficient approximation Tau-Leaping method (red), noting the log-log scale. We can see that the computational efficiency of the Doob-Gillespie method scales exponentially with system size, as expected. The Tau-Leaping method initially has constant computational time scaling, up until $N(0) = 10^4$, where it starts to grow exponentially with the system size. This change in time scaling is caused by the Tau-Leaping method computing Poisson random numbers differently, depending on the rate parameter. For sufficiently large rate parameters, this random number generation increases in computational complexity, leading to the exponential computational time scaling with system size. In contrast, our Jump-Switch-Flow method is constant in time up to an initial system size of $N(0) = 10^5$, after which it begins to slowly decrease (for $\Omega = 10^1$ and $\Omega = 10^3$). This decrease is caused by more of the scenario being simulated in the flowing state, as the system becomes highly deterministic with a larger population.

S3.5 Simulation experiments: Inference with simulated data, parameter estimates

To test how the Jump-Switch-Flow method behaves when used to perform parameter estimation, we first generated synthetic data of a known SIRS mode with demography. We simulate data for 400 days using the parameter values given in Figure 2C, with the Doob-Gillespie algorithm, and select a trajectory which experiences epidemic fade-out (at day 250), and an initial condition of two infected individuals ($I(0) = 2$), no recovered individuals ($R(0) = 0$), and a total population of $N(0) = 10^5$ individuals. We track the number of infectious individuals within the population each day, and use a truncated

binomial distribution to add noise into the data set, reflecting a realistic measurement process. We then couple our Jump-Switch-Flow method, using a switching threshold of $\Omega = 10^3$, with the open-source particle filter package pypfilt (Moss, 2024), using 2,000 particles. We first fix the demographic parameters to $\kappa = \mu = 1/(85 \times 365)$ (since these are assumed to be standard within a given population). We perform the inference on the data set for the first 100 days of the epidemic to estimate β , γ and ω , and then use our parameter estimates to predict forward in time the possible trajectories based on the obtained predictions. Using these predictions, we also estimate the probability of epidemic fade-out occurring at times $t = 100$ to $t = 400$ days.

Figure S8 shows the posterior distributions of parameters. The dotted black lines show the true value used to produce the simulated data, and the dashed red lines indicate the initial bounds on the prior estimates. We see that both β and γ are well estimated, with the posteriors centered around the true values. However ω is not well estimated. This is because ω acts on the order of 1 year, while the window of data used to fit the parameters to the data is 100 days (≈ 0.3 years).

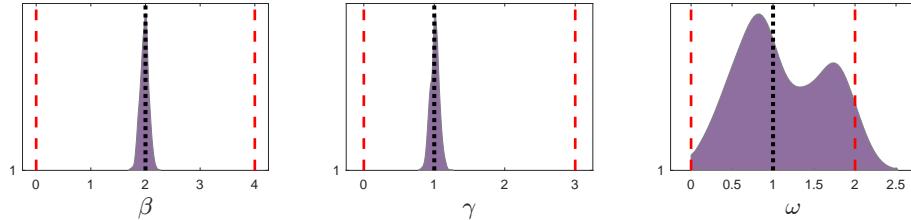


Figure S8: Posterior distributions of parameters for the SIRS model with demography, obtained via a particle filter with 2,000 particles, and the Jump-Switch-Flow sampler. The red dashed lines indicate the initial bounds on prior samples. The vertical dotted black lines indicate the true value used to produce the simulated data.

S4 Inference Study: Estimating TEIRV model parameters using a particle filter

We implemented the Refractory Cell Model described by the authors for the viral load data obtained from nasal swabs, using our Jump-Switch-Flow method with a switching threshold on all compartments $\Omega = 10^2$. We implemented this in particle filter pypfilt Moss, 2024 to estimate the model parameters, using 6,000 particles per simulation. In previous analysis Ke et al., 2022, the decay of virion was fixed at $c = 10$ virions/day, and the eclipse of cells to become infectious as $k = 4$ cells/day, leaving the remaining parameters to be estimated. Therefore, we similarly fix $c = 10$ virions/day and $k = 4$ cells/day.

S4.1 Parameter Estimation and Viral Clearance Prediction

Figure 9 of the main text highlights how the estimated viral loads closely resemble the data collected from the 6 patients. We also show how the estimated R_0 values for the patients vary. These R_0 estimates are consistent with the original study by Ke et al., 2022. However, understanding how the parameter estimates vary between patients assists in understanding how the fitting process behaves.

The posterior distributions for each of the 6 patient's parameter estimates are provided in Figure S9. These posterior distributions are found by fitting a violin envelope from the 6,000 estimates obtained from the particle filtering process. We also show the prior distributions, as red dashed lines, which show the lower and upper bounds for the uniform initial guesses.

We can see that, often, parameters ρ and Φ are not identified here, as the posterior distributions closely resemble the priors (excluding patient 451152). We also observed that for all patients, β appears to be well estimated. The parameter π is also well estimated for patients 443108, 444332, 444391 and 451152, however patients 432192 and 445602 provide less insight. The parameter δ is very well estimated by all patients, excluding patient 451152. Lastly, the initial viral loads, $\log_{10}(V_0)$, are all estimated well, and also estimated to take a large value.

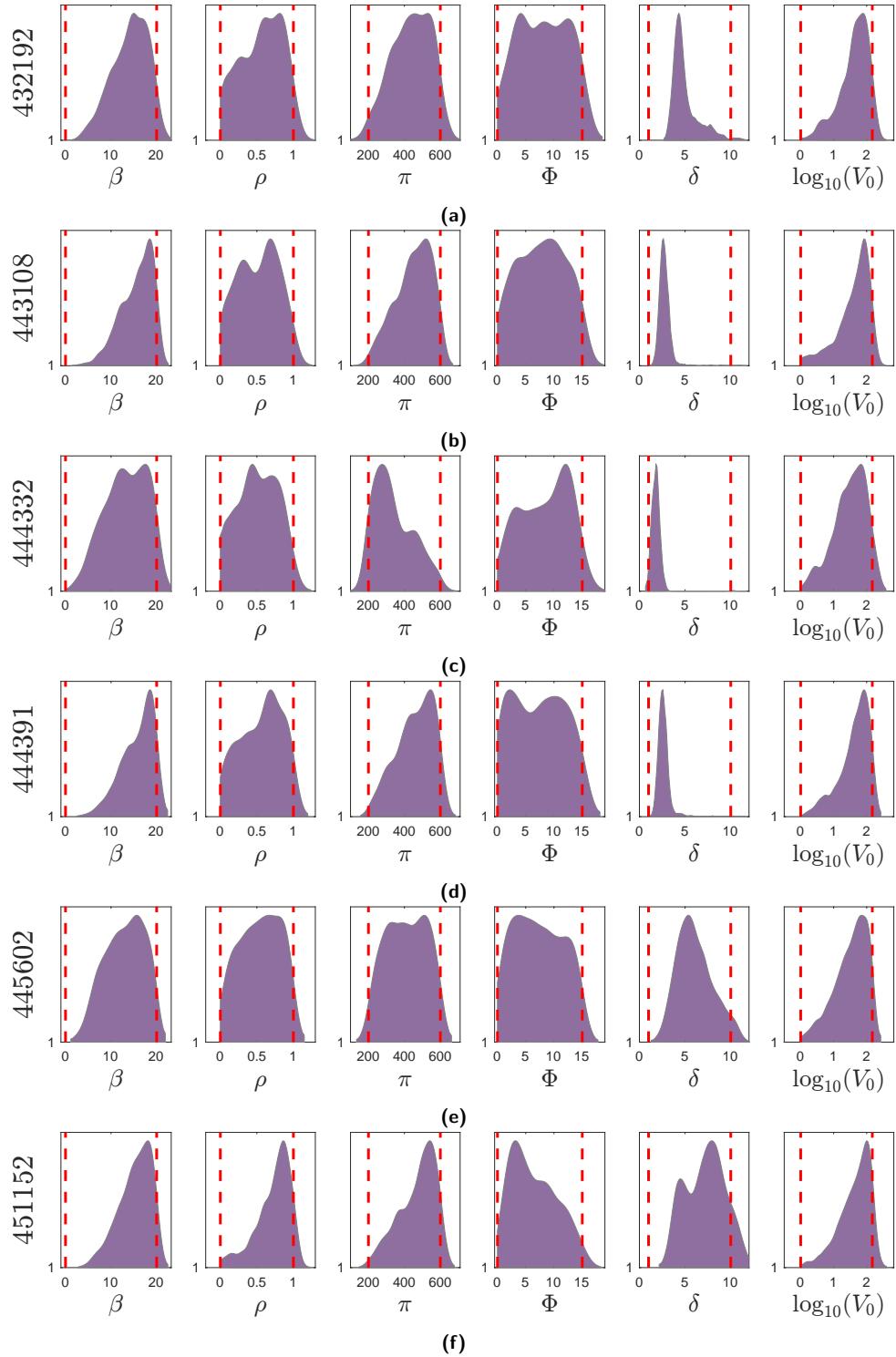


Figure S9: Posterior distributions of the model parameters and initial viral load for each of the selected SARS-CoV-2 infections. The red dashed lines indicate the initial bounds on prior samples.