Check for updates

MAIN ARTICLE

A Bayesian approach to calibrate system dynamics models using Hamiltonian Monte Carlo

Jair Andrade* and Jim Duggan

Abstract

Model calibration is an essential test that dynamic hypotheses must pass in order to serve as tools for decision-making. In short, it is the search for a match between actual and simulated behaviours using parameter inference. Here, we approach such an inference process from a Bayesian perspective. Under this paradigm, we provide statements about the parameters (viewed as random variables) and data in probabilistic terms. These statements stem from a posterior distribution whose solution is often found via statistical simulation. However, the uptake of these methods within the system dynamics field has been somewhat limited, and state-of-the-art algorithms have not been explored. Therefore, we introduce Hamiltonian Monte Carlo (HMC), an efficient algorithm that outperforms random-walk methods in exploring complex parameter spaces. We apply HMC to calibrate an SEIR model and frame the process within a practical workflow. In doing so, we also recommend visualisation tools that facilitate the communication of results.
Copyright © 2021 The Authors. *System Dynamics Review* published by John Wiley & Sons Ltd on behalf of System Dynamics Society.

Syst. Dyn. Rev. **37**, 283–309 (2021)

Additional Supporting Information may be found online in the supporting information tab for this article.

Introduction

From its beginnings in the mid-1950s to its modern practice, system dynamics (SD) has been a purpose-driven approach. Namely, it is a field interested in problems to be solved, situations that need to be better understood, or undesirable behaviours that need to be corrected or avoided (Forrester, 1993). To meet these goals, SD practitioners develop simulation models, formal representations often via ordinary differential equations (ODE) that capture the dynamic complexity of the problem situation and from which behavioural inferences can be made (Saleh *et al.*, 2010). The validity of these inferences hinges on the ability of the model's internal structure to adequately represent the aspects of the system that are relevant to the problem behaviour (Barlas, 1996). Adequacy, nevertheless, is not merely related to the appropriateness of the equations. In order

Data Science Institute and School of Computer Science, National University of Ireland Galway, Galway, Ireland

* Correspondence to: Jair Andrade, National University of Ireland Galway, University Rd, Galway, Ireland.
E-mail: jair.andrade@nuigalway.ie

Accepted by Markus Schwaninger, Received 22 December 2020; Revised 7 July 2021; 4 September 2021 and 20 September 2021; Accepted 23 September 2021

System Dynamics Review
System Dynamics Review vol 37, No 4 (October/December 2021): 283–309
Published online in Wiley Online Library
(wileyonlinelibrary.com) DOI: 10.1002/sdr.1693

This is an open access article under the terms of the Creative Commons Attribution License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited.

for a model to be useful, it must provide an assessment of future behaviour (Duggan, 2016). That is, models ought to reasonably estimate the likely impact of interventions in a system, a process that cannot be achieved without a plausible quantification of the model's parameters. Given the continuous nature of ODE representations, there are infinite possibilities for parameter quantification. To address this uncertainty, analysts fit models to available data in order to obtain estimates for the unknown quantities. This procedure, referred to as *model calibration*, serves a dual purpose. In addition to reducing uncertainty around the parameters, model calibration is a stringent validity test that assesses the causal claim that a particular structure accounts for an observed behaviour (Oliva, 2003). It thus follows that one should reject models that fail this test.

Generally speaking, model calibration is the process of finding a match between observed and simulated behaviours via *statistical inference* (Oliva, 2003). In other words, we search for plausible parameter values or *model configurations* that *accurately* account for the available data. Traditionally, within the SD field, practitioners have followed a *frequentist* approach. Following this paradigm, one employs nonlinear optimisation algorithms to maximise a statistical function (often a likelihood function), which expresses how well the model fits a time series of data pertaining to an important model variable (Dangerfield and Duggan, 2020). However, such optimisation routines can be inefficient for finding a match in nontrivial and high-dimensional parameter spaces (Andrade and Duggan, 2020). To deal with this difficulty, SD practitioners adopt the strategy of running the optimisation algorithm from multiple starts and select the result with the largest likelihood. Unfortunately, as the number of parameters increases, so does the risk of exhausting computational resources before finding the optimal start. To further complicate matters, the maximum likelihood estimate (MLE) may not even be located in regions of high-probability mass in high-dimensional spaces (Betancourt, 2017a).

Furthermore, around the MLE, one can construct uncertainty bounds using frequentist approaches such as the likelihood ratio method (Pawitan, 2013), a technique offered by SD software (*Vensim* and *Stella*). ODE models, nevertheless, often violate the assumptions implicit in the likelihood ratio method, such as identically and independently distributed (IID) normal error terms (Dogan, 2007). Fortunately, the advent of powerful computational resources has been a catalyst that enabled the development of methods based on repeated random sampling to obtain numerical results (Robert and Casella, 2010). These statistical simulation algorithms can be oriented to explore complex parameter spaces and lift stringent restrictions on the shape of the uncertainty bounds. Although the SD community has not ignored these advances, a search on the SD literature suggests that the adoption of these methods has been gradual. Particularly, Dogan (2007) and Struben *et al.* (2015) propose bootstrapping as a robust frequentist method

© 2021 The Authors. *System Dynamics Review* published by John Wiley & Sons Ltd on behalf of System Dynamics Society.
DOI: 10.1002/sdr

10991727, 2021, 4, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/sdr.1693 by Health Research Board, Wiley Online Library on [04/08/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

for confidence interval estimation and demonstrate its application on a service-quality model (Oliva and Sterman, 2001). In essence, this method creates *new* data sets by resampling the original data, and then parameter values are estimated from each of these *new* bootstrap samples (Dogan, 2007). Similarly, Ansah *et al*. (2017) use bootstrapping to quantify the uncertainty in the parameter estimates of a model that predicts the number of Chinese elderly with some degree of cognitive impairment by 2060.

On the other hand, Pierson and Sterman (2013, p. 129) report the first use of a Markov chain Monte Carlo (MCMC) algorithm to perform inference on a system dynamics model: "We estimate model parameters by maximum likelihood methods during both partial model tests and full model estimation using Markov chain Monte Carlo methods to establish confidence intervals." Specifically, these authors estimated uncertainty bounds for 21 parameters of an industry-level model of airline profits. Likewise, Keith *et al*. (2017) estimated the parameters of seven alternative models that account for product diffusion in the hybrid electric-vehicle market. In these two case studies, the Markov chains are started from a point estimate obtained from nonlinear optimisation routines. More recently, Ghaffarzadegan and Rahmandad (2020) inferred the value of a nine-parameter epidemiological model that describes the early infectious process of COVID-19 in Iran. All of these authors employed enhanced versions (Osgood and Liu, 2015; Vrugt *et al*., 2009) of the Metropolis algorithm (Metropolis *et al*., 1953). Osgood and Liu (2015) provides a technical overview of the method, accompanied by a practical example.

Despite the benefits that statistical simulation offers for parameter inference, bottom-up implementations require from the practitioner a new mathematical and programming skill set. Therefore, a more viable strategy is the use of predefined routines provided by statistical packages. Even though these tools automate the process, their use requires the practitioner to understand what the method is trying to solve, why it works, and when and why it fails. However, the literature of parameter inference on ODE models via statistical simulation is sparse, and the notation can be challenging for practitioners with nonmathematical backgrounds, which impedes adoption within the SD community. This observation serves as the motivation for writing this article. Thus, the contribution of this work is twofold. First, we introduce to the SD field a state-of-the-art MCMC algorithm, known as Hamiltonian Monte Carlo (Neal, 2011) or HMC, oriented to explore nontrivial parameter spaces such as those common to SD models. As model size and complexity grow, this method outperforms other MCMC implementations (Beraha *et al*., 2021; Monnahan *et al*., 2017) and, in some instances, nonlinear routines (Andrade and Duggan, 2020). Second, we frame the article in the context of *Bayesian* statistics, in which statements about parameters and data are given in terms of *probability* (Gelman *et al*., 2017). Specifically, we draw on a

© 2021 The Authors. *System Dynamics Review* published by John Wiley & Sons Ltd on behalf of System Dynamics Society.
DOI: 10.1002/sdr

10991727, 2021, 4, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/sdr.1693 by Health Research Board, Wiley Online Library on [04/08/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

practical workflow to illustrate how one can think of model calibration as the result of knowledge update in the light of new information. This workflow facilitates the interpretation and communication of results in an intuitive fashion. It should be noted that although the workflow is intuitive, model calibration is essentially a statistical procedure, and as such, key concepts like *random variables*, *continuous probability distributions*, and *conditional expectations* are necessary (Blitzstein and Hwang, 2019). We demonstrate this workflow's application by fitting an epidemiological model to data using HMC. In doing so, we describe the logical process followed in each step. The article concludes with an overview of the insights obtained from the inference process. The model is built in Stella, and all the analysis is performed in R and Stan. The code is made freely available at https://github.com/jandraor/SDR_Bayes.

Context

As mentioned in the introduction, an SD endeavour starts with a problem. For didactic purposes, we follow a widely analysed case study (Vynnycky and White, 2010). In 1918, the H1N1 virus led to an influenza pandemic that spread over the entire world in less than 6 months and killed tens of millions of people (Patterson and Pyle, 1991). This pandemic occurred in three distinct waves, the second wave being the deadliest. Having learned from data-collection difficulties in the first wave, the U.S. Public Health Service organised special surveys in several localities to determine as accurately as possible the proportion of the population infected (Frost and Sydenstricker, 1919). From this information, we extract the report of new cases detected in the city of Cumberland (Maryland) during the autumn of 1918 (Figure 1). These case counts will serve as the basis to ascertain an estimate of the disease's degree of contagiousness, a feature commonly measured by the basic reproduction number (R_0). Simply put, this metric is the average number of secondary infections produced when one infected individual is introduced into a totally susceptible population (Anderson and May, 1992). In addition to other techniques (Farrington *et al.*, 2001), one can estimate R_0 by means of compartmental models (Vynnycky and White, 2010).

A common choice for modelling the transmission dynamics of influenza is the Susceptible–Exposed–Infectious–Removed (SEIR) framework (Chowell *et al.*, 2007; Mills *et al.*, 2004). In this formulation, $S(t)$ denotes the number of susceptible individuals at time t. Likewise, $E(t)$, $I(t)$, and $R(t)$ denote the number of exposed, infectious, and recovered individuals at time t, respectively. $C(t)$ represents the number of cumulative cases at time t. Here, we assume that the outbreak's time scale is much faster than the characteristic times for demographic processes (births and deaths) so that their effects are not included. Hence, it follows that the population is constant and whose

© 2021 The Authors. *System Dynamics Review* published by John Wiley & Sons Ltd on behalf of System Dynamics Society.
DOI: 10.1002/sdr

10991727, 2021, 4, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/sdr.1693 by Health Research Board, Wiley Online Library on [04/08/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

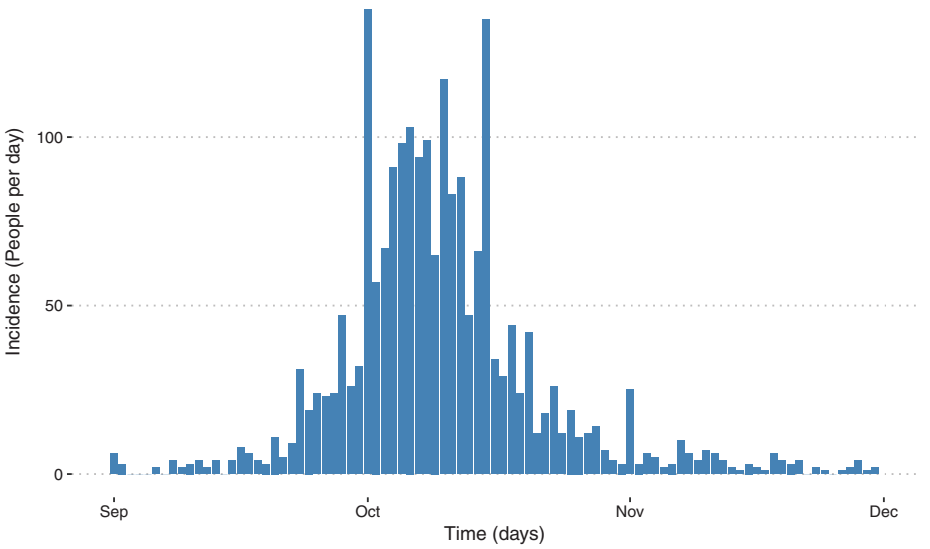

Fig. 1. Daily number of influenza cases detected by the U.S. Public Health Service in Cumberland (Maryland) during the 1918 influenza pandemic, from 22 September 1918 to 30 November 1918 [Color figure can be viewed at wileyonlinelibrary.com]

size N is determined by $S(t)+E(t)+I(t)+R(t)$. Furthermore, β represents the rate of effective contacts per infected individual, σ the rate of onset of infectiousness, and γ the recovery rate. To reconcile the discrete nature of the data and the continuous nature of compartmental models, we define the expected reported incidence (x) by Eq. (6), where τ is restricted to nonnegative discrete values. We assume that the rate of reporting ρ scales the true incidence ($C(\tau+1)-C(\tau)$). This rate reflects the fact that asymptomatic and paucisymptomatic (mild symptoms) individuals may not be *captured* by surveillance systems (Gamado *et al.*, 2014). From this model, R_0 can be estimated from the number of new infections caused by one infected individual in the period in which the individual is contagious (γ^{-1}), namely: $\beta\gamma^{-1}$.

$$\dot{S}=\frac{-\beta S(t)I(t)}{N} \tag{1}$$

$$\dot{E}=\frac{-\beta S(t)I(t)}{N}-\sigma E(t) \tag{2}$$

$$\dot{I}=\sigma E(t)-\gamma I(t) \tag{3}$$

$$\dot{R}=\gamma I(t) \tag{4}$$

$$\dot{C}=\sigma E(t) \tag{5}$$

$$x(\tau+1)=\rho(C(\tau+1)-C(\tau)) \tag{6}$$

© 2021 The Authors. *System Dynamics Review* published by John Wiley & Sons Ltd on behalf of System Dynamics Society.
DOI: 10.1002/sdr

10991727, 2021, 4, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/sdr.1693 by Health Research Board, Wiley Online Library on [04/08/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

Bayesian inference workflow

In an SD model, the parameters to be estimated correspond to time-independent variables and also initial conditions for stocks. To illustrate the estimation of such quantities, we follow a simplified adaptation (Figure 2) of a workflow from the statistical literature (Gabry *et al.*, 2019; Gelman *et al.*, 2020). Under this approach, we view the quantities of interest as random variables that describe our uncertainty about the actual values in the face of incomplete knowledge (McElreath, 2020). Following this approach, it is possible to apply concepts of statistical inference. That is, we update our knowledge about the underlying properties that generate the problem behaviour in light of the evidence, thus adopting a Bayesian learning perspective. To make such a process intuitive, we also draw upon data visualisation. This technique is an important tool that complements the process of model calibration, as demonstrated below.

Prior information

The first step in an inference process consists of the identification of the unknown parameters in the model. In other words, whenever possible, parameters that can be directly observed or estimated from sources at the individual level should be treated as part of the known structure (Graham, 1980) and removed from the process as *unmodelled predictors*. In doing so, we mitigate the risk of model misspecification by preventing a match between actual and simulated behaviour based on unrealistic corrections to known parameter values that mask errors in the model formulation (Oliva, 2003). For the remaining unknown parameters, an initial plausibility assignment should be estimated based on domain expertise, such as educated guesses, as recommended in the early days of the SD field (Graham, 1980).

In the study of infectious diseases, it is common to find observational studies at the individual level that report epidemiological quantities such as the latent (σ^{-1}) and infectious period (γ^{-1}). In fact, Anderson and May (1992) provide estimates of such quantities for 10 viral and bacterial infections, including influenza. However, measuring the average number of effective contacts by an infected person (β) remains a challenging task inasmuch as this variable encompasses individuals' social nature, the propensity of infected individuals to transmit the pathogen, and the propensity of susceptible individuals to being infected. In the same vein, continuous and exhaustive measuring of the infected population's true proportion was clearly not a viable option for the United States in the early 20th century. Consequently, we incorporate the parameters σ and γ into the model's structure, whereas β and ρ are considered the unknown time-independent variables.

© 2021 The Authors. *System Dynamics Review* published by John Wiley & Sons Ltd on behalf of System Dynamics Society.
DOI: 10.1002/sdr

10991727, 2021, 4, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/sdr.1693 by Health Research Board, Wiley Online Library on [04/08/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

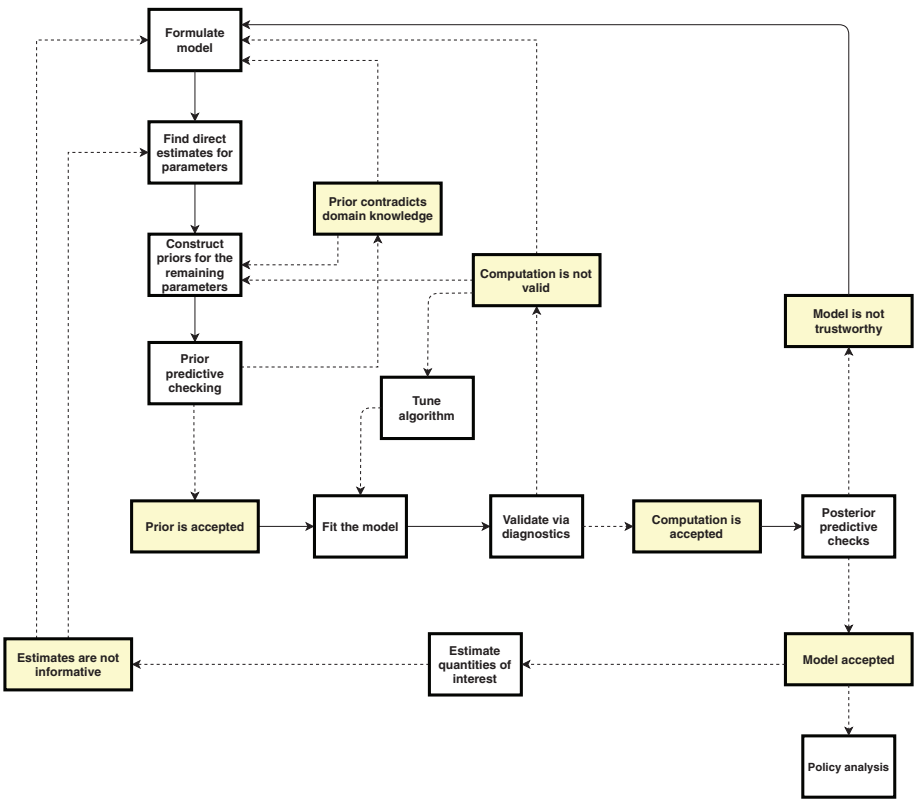

Fig. 2. Adaptation of a Bayesian workflow (Gelman *et al.*, 2020) to calibrate system dynamics models. Shaded boxes indicate modeller decisions. Dotted lines indicate alternative pathways [Color figure can be viewed at wileyonlinelibrary.com]

In relation to the initial conditions for stocks, researchers have estimated, from serological studies in similar settings, that 70 percent of individuals were still susceptible to infection after the first wave (Vynnycky *et al.*, 2007). This implies $S(0) = 0.7N - I(0)$ and $R(0) = 0.3N$. For simplicity purposes, we assumed there were no exposed individuals at the beginning of the second wave. Namely, $E(0) = 0$. These assumptions leave as the only unknown the number of individuals that trigger the outbreak ($I(0)$), which is assumed as the initial value for the stock that tracks the number of cumulative cases.

Consequently, in this example, we focus on the estimation of the three parameters identified above to which we denote by the vector θ. For each one, we outline their plausibility before assessing the evidence (Figure 3). For β and $I(0)$, our domain knowledge indicates that they should be nonnegative. Further, we suppose that these two quantities concentrate at low values considering the slow progression at the outbreak's start (Figure 1). This formulation does not discard values away from such concentration. However, we assign them small plausibilities, measured by the height of the function. This height is known as the probability density, and it is often

© 2021 The Authors. *System Dynamics Review* published by John Wiley & Sons Ltd on behalf of System Dynamics Society.
DOI: 10.1002/sdr

10991727, 2021, 4, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/sdr.1693 by Health Research Board, Wiley Online Library on [04/08/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

modelled by standard statistical distributions that we denote by the Greek letter π. In this case, $\pi(\beta)$ and $\pi(I(0))$ are distributed according to the $lognormal(0,1)$. We choose this distribution to reflect our belief that these parameters should be positive and relatively small. Regarding the reporting rate ρ, although we are unsure of its magnitude, we know that it should be between 0 and 1, and by including it in the model, we tacitly assumed that it should be far from the boundaries. If we had thought that the parameter was close to 0, we would have discarded the reported cases (C) stock. If we had thought the value was close to 1, we would not have needed the reporting rate parameter. We model this assumption by $\pi(\rho) \sim beta(2,2)$. Taking into account that no evidence suggests otherwise, we assume independence in the parameters. That is, having information about one parameter does not provide knowledge about the others. Mathematically, $\pi(\theta) = \pi(\beta)\pi(I(0))\pi(\rho)$. In statistical language, this is known as the prior distribution. Interested readers are referred to Gelman *et al.* (2017) for philosophical and practical considerations about the prior distribution.

Observational model and probability of the data

$$\pi(y|\theta) = Pois(y|x(\tau)) \tag{7}$$

$$\pi(y) = \int \pi(y|\theta)\pi(\theta)d\theta \tag{8}$$

Naturally, it is expected that prior knowledge of the parameters leads to dynamics that capture the essence of the problem being studied. In other words, if we plug the vector θ into the simulation model, the latter ought to produce outbreak-like trajectories, including the observed behaviour. Thus far, the simulation model (Eqs. (1)–(6)) is deterministic. That is, the model always produces the same smooth output from a given configuration. Unsurprisingly, even the most *perfect* configuration (or any other) will yield values that differ from the measurements as the model only approximates the studied phenomenon. Therefore, these differences must also be formally accounted for by a formulation, $\pi(y|\theta)$, which we refer to as the *measurement or observational model*. Since daily reported cases are nonnegative discrete quantities, we formulate the observational model (Eq. (7)) in terms of a Poisson distribution (see Appendix S1 section 3.2.1 for a discussion on the distribution choice). Note that with the addition of Eq. (7), a single model configuration can yield different reported incidences. Thus, when θ is fixed to a single set (θ_i), the observational model is known as the *sampling distribution*, $\pi(y|\theta_i)$.

© 2021 The Authors. *System Dynamics Review* published by John Wiley & Sons Ltd on behalf of System Dynamics Society.
DOI: 10.1002/sdr

10991727, 2021, 4, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/sdr.1693 by Health Research Board, Wiley Online Library on [04/08/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

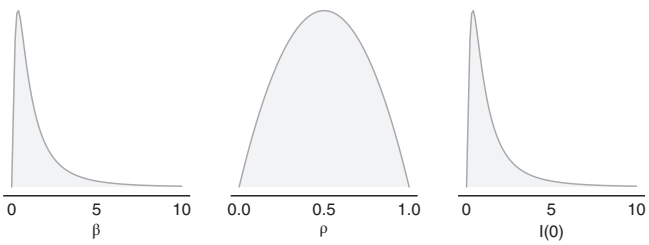

Fig. 3. Prior distribution. Shaded areas indicate probability mass

In an ideal and unrealistic scenario, one would generate infinite samples (θ_{sim}) from the prior distribution $\pi(\theta)$, then feed the simulation model with those samples to produce the entire universe of possible trajectories of reported incidences y_{sim}. Once the complete set of y_{sim} has been sampled, one aggregates similar trajectories to establish which behaviours are more likely to be observed than others. Formally, this is expressed by Eq. (8), where $\pi(y)$ denotes the *average probability of the data* (McElreath, 2020) or *prior predictive distribution* (Gelman *et al.*, 2013). Although generating infinite samples is infeasible, one can draw a finite number of samples to reason about the model's behaviour conditioned on current knowledge. Accordingly, we draw 500 random samples ($\theta_1, \theta_2 \ldots, \theta_{500}$) from the prior distribution to generate an equal amount of trajectories ($y_1, y_2 \ldots, y_{500}$). We present the results in Figure 4. Here, we notice that large swathes of samples generate outbreak-like behaviours, possibly the observed data (solid points). Overall, this process is referred to as *prior predictive checking*, and it is a powerful tool for understanding the structure of models (Gabry *et al.*, 2019). Prior predictive checking aims to answer the question "Could this prior generate the type of data we expect to see?" (Gelman *et al.*, 2017). Should none of the simulations resulted in outbreak-like behaviours, or should these behaviours not captured the observed data, it would have been an indication for reassessing the validity of the prior distribution or the model itself (Gelman *et al.*, 2020).

Expectation

$$\pi(\theta|y_c) = \frac{\pi(\theta)\pi(y_c|\theta)}{\pi(y_c)} \tag{9}$$

$$\mathbb{E}[f(\theta)] = \int f(\theta)\pi(\theta|y_c)\, d\theta \tag{10}$$

Thus far, we have considered the likely behaviours over time that we could have observed. Nonetheless, the chief interest is performing parameter

© 2021 The Authors. *System Dynamics Review* published by John Wiley & Sons Ltd on behalf of System Dynamics Society.
DOI: 10.1002/sdr

10991727, 2021, 4, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/sdr.1693 by Health Research Board, Wiley Online Library on [04/08/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

inference based on the available data set rather than on the infinite set of possible observations. We thus shift the focus from the universe of measurements (y_{sims}) to the observed behaviour, Cumberland's incidence data (y_c). In doing so, the density function $\pi(y_{sims})$ becomes the constant $\pi(y_c)$. Furthermore, when the observational model is regarded as a function of θ, for a fixed y, it is called the *likelihood function* (Gelman *et al.*, 2013). This mathematical construct, $\pi(y_c|\theta)$, is the target of optimisation algorithms and a statement about the data, which quantifies the relative consistency of each model configuration with the observed data. Simply put, if θ_1 produces a larger likelihood value than θ_2, then y_c is more likely to have occurred from θ_1. However, in a Bayesian setting, the plausibility of a trajectory is not the desired outcome. On the contrary, the interest lies in establishing which values of the vector θ are more plausible than others given the observed trajectory (y_c), or in more formal terms, the posterior distribution of the estimated parameters, $\pi(\theta|y_c)$. Conveniently, by Bayes theorem (Eq. (9)), we can express the posterior distribution in terms of the prior distribution, the probability of the data, and the likelihood function. Given that the posterior encodes all the information learned by our model, one could extract inferences about the data and the parameters using expectations (Eq. (10)). As a matter of fact, prior predictive checking is an expectation, where f is the observational model averaged over the prior distribution (instead of the posterior). In the sections below, we will see that, should a solution for Eq. (9) be available, obtaining a model fit is nothing more than the application of Eq. (10). Regardless of the query, answers are always given in probabilistic terms. Indeed, this is the main feature of the Bayesian approach, where uncertainty is quantified with probability distributions.

Markov chain Monte Carlo

Consequently, from a Bayesian perspective, one approaches *parameter inference* as the process of finding the posterior distribution. It is often the case that *closed-form* solutions do not exist for such a type of formulation (Robert and Casella, 2005). To address this difficulty, in the late 1940s, researchers at Los Alamos developed stochastic simulation techniques, known as Monte Carlo methods (Robert and Casella, 2011). Early conceptualisations employed *exact sampling* (Robert and Casella, 2010), the generation of independent and identically distributed (IID) samples to explore the extent of the parameter space unconditionally. Regions of high probability, however, are concentrated on specific locations rather than being scattered around (Betancourt, 2017a). Thus, IID sampling would squander finite computational resources on low-probability regions until they are eventually exhausted before reaching the target location. Aware of this, this same group of researchers enhanced the method by generating *correlated samples* from a

© 2021 The Authors. *System Dynamics Review* published by John Wiley & Sons Ltd on behalf of System Dynamics Society.
DOI: 10.1002/sdr

10991727, 2021, 4, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/sdr.1693 by Health Research Board, Wiley Online Library on [04/08/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

Markov chain to approximate the equilibrium distribution of a liquid (Metropolis *et al.*, 1953). Hence, the term Markov chain Monte Carlo (Geyer, 2011). Even though further improvements in subsequent decades, such as the Metropolis-Hastings algorithm (Hastings, 1970) and the Gibbs sampler, (Geman and Geman, 1984) broadened the method's scope, it was only until the early 1990s (Gelfand and Smith, 1990), and partly due to the growth in computational power, that the mainstream statistical community widely noticed the method (Robert and Casella, 2011). Since then, there has been significant growth in the number of applications to a wide range of fields, including epidemiology (Chatzilena *et al.*, 2019; Davies *et al.*, 2020).

A Markov chain is a sequence of random variables $\Theta_1, \Theta_2, \ldots, \Theta_n$, in which each variable depends only on the previous one (Blitzstein and Hwang, 2019). To iteratively draw samples or realisations, we apply a conditional probability distribution denoted by $T(\Theta_{i+1}|\Theta_i)$, also referred to as the *transition kernel*. The strength of this approach lies in the improvement achieved at the generation of each new sample—improvement in the sense of converging to the target distribution $\pi(\theta|y_c)$. If run long enough, the Markov chain is expected to reach an equilibrium state—or stationary state—where the samples describe the posterior distribution. This approximation has the advantage that it does not impose constraints in the shape of the posterior. Under ideal conditions, the Markov chain starts from any place in the parameter space and gradually moves towards the target distribution. This initial phase is known as *warm-up*. Once the target distribution has been found, the Markov chain explores high-probability regions (*sampling phase*), namely parameter values that have larger contributions to the observed behaviour. To obtain unbiased estimators, one discards the samples from the warm-up phase.

Although theoretically, the Markov chain will eventually reach the stationary state; in practice, this result is not guaranteed, especially for high-dimensional target distributions and distributions that exhibit nontrivial dependencies among the parameters (Betancourt, 2017b). Early implementations of MCMC, such as the Metropolis-Hastings and Gibbs samplers, become slow at exploring complex parameter spaces to the extent that computational resources are depleted before providing accurate estimates. This inefficiency occurs due to these algorithms' random-walk behaviour to generate new samples, resulting in zig-zag movements across the parameter space (Gelman *et al.*, 2013). For instance, Pierson and Sterman (2013, p.143) report that in the calibration of an industry-level model of airline profits, "over 1 million MCMC runs were needed to arrive at stable estimates for the confidence bounds."

Hamiltonian Monte Carlo

According to Betancourt (2017b), MCMC has benefited from an evolving interplay between statistics and physics from its inception to present

© 2021 The Authors. *System Dynamics Review* published by John Wiley & Sons Ltd on behalf of System Dynamics Society.
DOI: 10.1002/sdr

10991727, 2021, 4, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/sdr.1693 by Health Research Board, Wiley Online Library on [04/08/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

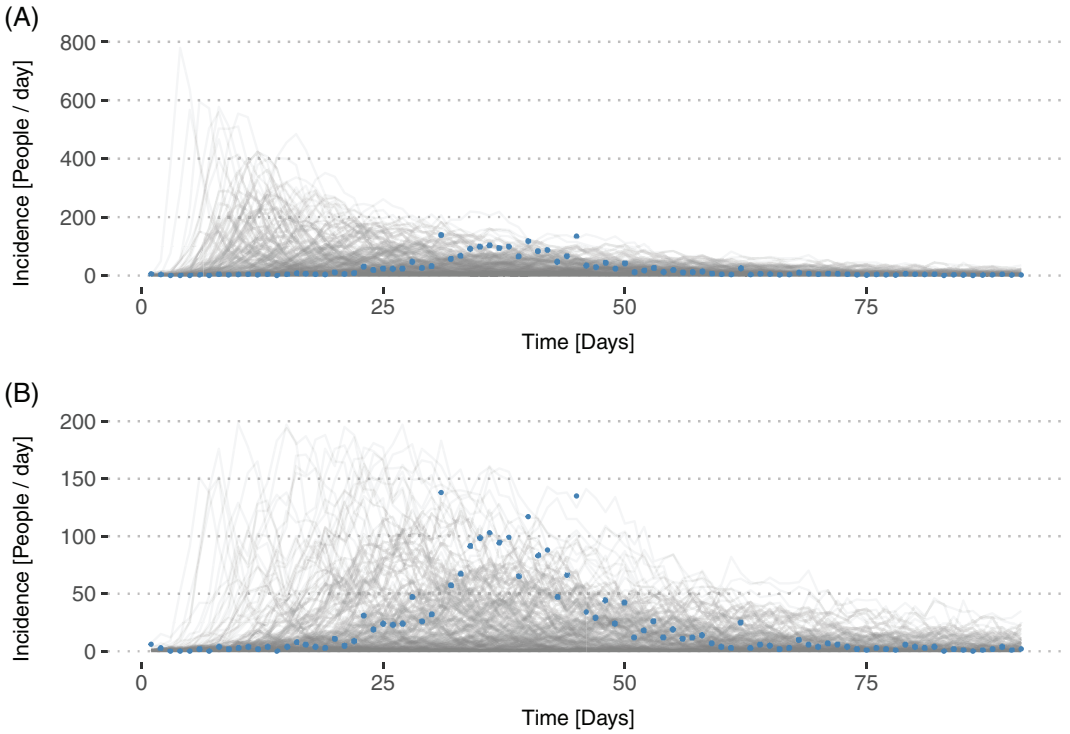

Fig. 4. Prior predictive checks. (a) Simulation of 500 predicted incidence measurements (grey lines) from the proposed dynamic hypothesis and the prior distribution. Dots denote the actual data. (b) Zoom to trajectories that may resemble the actual data. We show predicted measured incidences whose peak is lower than 200 new cases in a day [Color figure can be viewed at wileyonlinelibrary.com]

developments. Conceiving a statistical system as a physical one provides an innovative way to improve computational methods. Indeed, the realisation that molecular simulation methods—in which the motion of molecules was deterministic, following Newton's laws of motion—and MCMC could be combined, yielded a technique of wide applicable potential. In such a framework, the description of molecular motion has an elegant formalisation as *Hamiltonian dynamics*, hence the term *Hamiltonian Monte Carlo* (Neal, 2011). In particular, the HMC algorithm simulates the movement of a fictitious and frictionless particle (McElreath, 2020) over a surface whose ruggedness is determined by the likelihood function and the prior distribution. Formally, the Hamiltonian function—the sum of potential and kinetic energies—describes such mechanics (Neal, 2011). In turn, this function depends on the characterisation of each parameter in terms of *position* and *momentum*. The former is straightforward, considering that it corresponds to the fictitious particle's location in the parameter space; for the latter, the algorithm adds an artificial

© 2021 The Authors. *System Dynamics Review* published by John Wiley & Sons Ltd on behalf of System Dynamics Society.
DOI: 10.1002/sdr

10991727, 2021, 4, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/sdr.1693 by Health Research Board, Wiley Online Library on [04/08/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

variable per parameter. As a result of this conceptualisation, the random-walk behaviour from early MCMC implementations is suppressed (Gelman *et al.*, 2013), resulting in a tool that becomes efficient at traversing the complex parameter spaces. For an analytical treatment of the method, the reader is referred to Neal (2011). Likewise, Betancourt (2017a) offers an intuitive description. In Appendix S2 section 1, we provide an example where we compare how HMC and the Metropolis algorithm explore parameter spaces.

Stan

Although HMC is a powerful method, its geometrical foundations (Betancourt *et al.*, 2014) render ad hoc implementations onerous. To address this challenge, a group of researchers developed Stan (Carpenter *et al.*, 2017), a statistical modelling platform. This tool provides an interface to perform Bayesian inference via the No-U-Turn-Sampler or NUTS (Hoffman and Gelman, 2011), an HMC algorithm. NUTS takes advantage of the *warm-up phase* to identify the algorithm's configuration that best adapts to the user-supplied model for efficient parameter space explorations, resulting in significant gains in sampling speed. Furthermore, Stan supports gradient evaluation (Carpenter *et al.*, 2015) to a broad range of distribution families and ODE solvers. In spite of this support, in some cases, SD practitioners will not be able to avail themselves of familiar built-ins (such as those offered by Vensim and Stella). As a result, they will have to formulate equations explicitly, or in the case of table functions, the practitioner will have to devise parametric formulations.

In practice, Stan only requires, from the practitioner, the specification (code) of the model's equations, the prior distribution, the likelihood function, the data, and the number of draws. To these specifications, Stan runs the NUTS algorithm internally and returns a set of samples for each parameter. To run the simulation, one can directly interact with Stan through a command interface or popular statistical software such as Python and R. In this case, we choose the latter to draw upon the package *readsdr* (Andrade, 2021), which automatically converts XMILE files from Stella and Vensim to Stan code.

Diagnostics

Bayesian inference via iterative simulation is performed by extracting insights from the entire collection of simulated draws from the sampling phase. Specifically, we estimate posterior probability densities and compute quantities of interest that describe the calibrated parameters, such as expected values (mean) and credible intervals. However, if the chains are not run long enough, predicted convergence to the stationary distribution may not be achieved. The resulting draws may partially or inaccurately describe the posterior distribution, thereby producing unreliable estimates. To address this challenge, an effective strategy consists of running at least four chains that start from different locations in the parameter space and verify that all of them converge to the

© 2021 The Authors. *System Dynamics Review* published by John Wiley & Sons Ltd on behalf of System Dynamics Society.
DOI: 10.1002/sdr

10991727, 2021, 4, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/sdr.1693 by Health Research Board, Wiley Online Library on [04/08/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

same region. Accordingly, we run four chains of 2000 iterations in this example: 1000 allocated to warm-up and 1000 to sampling.

Graphically, one can inspect convergence via trace plots. These visualisation tools are time series of the draws for a particular parameter. Here, *time* refers to the order in which the draws were sampled. In Figure 5a, we present the sequence of the first 100 draws for each of the three calibrated parameters. Initially, the chains traverse the parameter space before settling on a unique location. By augmenting the time frame to the complete set of samples (Figure 5b), it can be seen that the chains mixed; that is to say, the draws trace out a common distribution. Additionally, there is no obvious trend or change in the spread in the chains. In other words, they are stationary. These two properties suggest that the sampling procedure reached the predicted convergence. Quantitatively, the potential scale-reduction factor (Gelman and Rubin, 1992) denoted by $\widehat{R}$ is a useful metric to validate this assessment. This statistic compares within-chain variance (stationarity) to between-chain variance (mixing). At convergence, $\widehat{R}$ should be <1.01 (Vehtari *et al.*, 2021), whereas higher values indicate that the chains describe different locations of the parameter space or different trends within a single chain. In this example, all chains exhibit potential scale-reduction factors below this threshold (see Appendix S1 section 3.3.1.2 for a technical description and results).

Other diagnostics to gain confidence in the results include the effective sample size (*ESS*). In general, simulation inference from correlated samples is less precise than from the same number of independent samples (Gelman *et al.*, 2013). If the correlation among samples is strong, chains must be run for longer periods in order to obtain accurate estimations. To measure this correlation, we employ *ESS* to determine the number of independent simulation draws from the MCMC process. For reliability, this metric should be above 400 (100 per chain) per parameter (Vehtari *et al.*, 2021), as in this example (see Appendix S1 section 3.3.1.2 for a technical description and results).

To conclude with MCMC diagnostics, the Hamiltonian approach also allows us to evaluate the robustness of the results. A key feature of the Hamiltonian function (sum of potential energy and kinetic energy) in the sampling phase is that it remains invariant along the trajectory in which the particle moves. Any divergence from its initial value indicates pathological behaviour (abnormal movements) in the chains to the extent that they cannot be trusted, and the calibration setup (SD model, prior, likelihood, algorithm's parameter values) must be reformulated. By default, Stan reports divergences and provides ways to access which iterations encountered divergences.[i]

[i]See Stan Manual for more details.

© 2021 The Authors. *System Dynamics Review* published by John Wiley & Sons Ltd on behalf of System Dynamics Society.
DOI: 10.1002/sdr

10991727, 2021, 4, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/sdr.1693 by Health Research Board, Wiley Online Library on [04/08/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

Given this example's didactic scope, all diagnostics unsurprisingly return favourable results. In practical applications, however, the path to these results can be significantly less straightforward. From code bugs to structural problems, such as nonidentifiability in the model (two or more parameterisations that are observationally equivalent), achieving convergence can take several iterations. To complicate matters, exploring the target distribution of differential equation models involves expensive gradient evaluations that slow down the HMC algorithm, limiting the number of debugging runs. Thus, the analyst must efficiently identify the problem's source. To this end, Gelman *et al.* (2020) recommend the process of *fake data* (also known as *synthetic data* or *simulated data*). That is, feeding the simulation model with known and plausible parameter values to obtain behaviours over time similar to the real data being analysed. Then, we should check whether the same model and the inference method can recover the known values. In doing so, it is possible to identify strategies to address computational issues. These strategies range from model simplification and more data collection to recognising the method's inappropriateness for the studied problem. For instance, HMC works correctly under well-defined posterior densities, and it is restricted to continuous parameters. Conversely, challenging geometries with sharp corners or multiple modes (Betancourt, 2015) in the posterior distribution render the algorithm impractical, and other types of methods should be employed (Valderrama-Bahamóndez and Fröhlich, 2019). We refer the reader to Gelman *et al.* (2020) for a comprehensive treatment of these methodological issues. Furthermore, in Appendix S3, we draw upon *synthetic data* to illustrate the Bayesian workflow presented above (Figure 2) in the context of *wrong* assumptions, complex parameter spaces, and the necessity for data collection.

Posterior distribution

Bayesian inference is concerned with updating knowledge in the light of new evidence (McElreath, 2020). Once we have gained confidence in the sampling procedure, we take the draws returned by Stan and construct probability densities. By restricting the analysis to a single parameter (marginal posterior distribution), it is possible to determine which values are plausible for the parameters after seeing the data. We can visually portray such a knowledge update process by comparing marginal prior and posterior distributions (Figure 6). In this graph, we observe that the concentration of probability shifted for each parameter. For instance, before the calibration, we assumed *ignorance* for the reporting fraction (ρ). On the contrary, the marginal posterior distribution indicates that 95 percent of the samples—or 95 percent credible interval—concentrates on the region [0.74, 0.81]. In relation to R_0, we estimate its 95 percent credible interval between 2.53 and 2.63, a value consistent with the estimate ([2.56–2.59]) reported by Vynnycky and White (2010). Furthermore, Mills *et al.* (2004) estimate that R_0 for the 1918

© 2021 The Authors. *System Dynamics Review* published by John Wiley & Sons Ltd on behalf of System Dynamics Society.
DOI: 10.1002/sdr

10991727, 2021, 4, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/sdr.1693 by Health Research Board, Wiley Online Library on [04/08/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

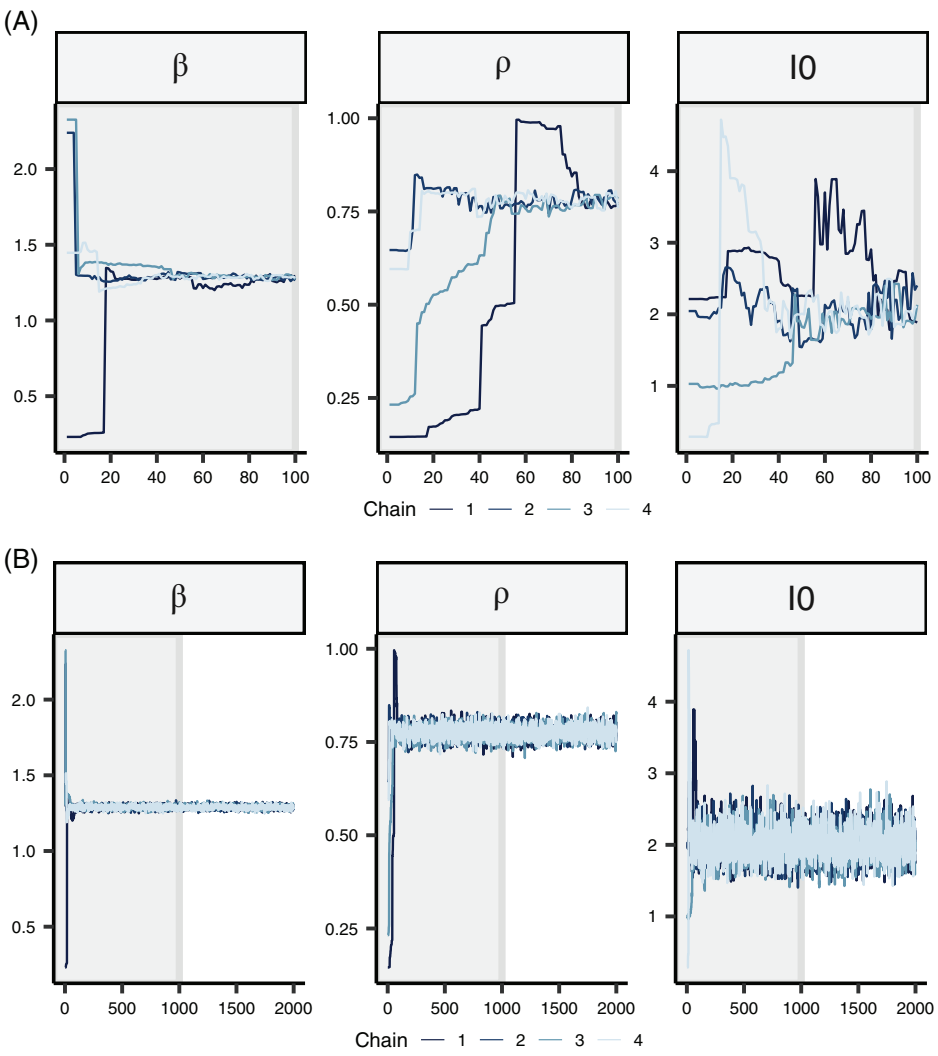

Fig. 5. (a) Early warm-up phase for our three parameters (first 100 iterations). (b) Complete sequence of iterations (2000) in the warm-up and sampling phases. The shaded area indicates the warm-up phase. In both figures, we present four Markov chains per calibrated parameter through sequences of points (samples obtained from Stan) joined by lines [Color figure can be viewed at wileyonlinelibrary.com]

pandemic influenza was approximately between 2 and 3. Similarly, Vynnycky *et al.* (2007) concluded that this value was in the range of 2.4–4.3 in community-based settings. Thus, during this pandemic, one infected individual could potentially infect, on average, almost three susceptible ones.

In the SD literature, it is not unusual that researchers restrict model calibration results to reports of the mean and standard deviation of the fitted parameters. Although useful for descriptive purposes, we instead have focused our interest on the complete set of samples. The reader should bear in mind that in Bayesian inference, we quantify uncertainty by an entire

© 2021 The Authors. *System Dynamics Review* published by John Wiley & Sons Ltd on behalf of System Dynamics Society.
DOI: 10.1002/sdr

10991727, 2021, 4, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/sdr.1693 by Health Research Board, Wiley Online Library on [04/08/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

probability distribution, which cannot be characterised by a single-point estimate (in Appendix S1 section 4, we elaborate on this issue). Through the samples obtained from MCMC methods, for instance, we can extract rich information about the parameter interactions. Initially, given the lack of evidence, we assumed independence among the calibrated parameters. Nevertheless, system dynamics models, by definition, depict problems as an interconnected confluence of factors. Based on this logic, it would be surprising that a parameter does not interact with another. To explore this property, we draw upon *pair plots*. This graphical tool displays all possible pairwise combinations (joint) of probability distributions. Moreover, we include the correlation for each combination along with the marginal distributions to gain a global perspective (Figure 7). This plot shows a strong interaction among the parameters to the extent of an almost perfect correlation between two parameters ($I(0)$ and β), indicating that, relatively speaking, *large* values of β are solely compatible with *low* values of $I(0)$. Consequently, independence assumptions are unwarranted. The implications of this finding are explored in the Policy Analysis section.

Posterior predictive checks

$$\pi(y|y_c) = \int \pi(y|\theta)\pi(\theta|y_c)\,d\theta \tag{11}$$

The ultimate purpose of model calibration is to search for a match between observed and simulated behaviour that builds confidence in the proposed dynamic hypothesis. Following this Bayesian workflow, we frame this purpose as: "if a model is a good fit we should be able to use it to generate data that resemble the data that we observed" (Gabry *et al.*, 2019, p. 396). Notice that this statement is similar to Oliva's quote (Oliva, 2003, p. 554): "Confidence that a particular structure, with reasonable parameter values, is a valid representation increases if the structure is capable of generating the observed behavior." To provide an answer, we can use the posterior distribution to obtain predictions for the measured quantities and compare them to the observed data (Gelman and Hill, 2007). Thus, in this case, $f(\theta)$ (see Eq. (10)) corresponds to the observational model. This process is analogous to prior predictive checking, with the difference that we average over the posterior distribution. Unsurprisingly, this process is called *posterior predictive checking* (Eq. (11)). In consequence, obtaining a model's fit under this Bayesian paradigm is equivalent to solving Eq. (11).

Accordingly, we generate 500 draws from $\pi(\theta|y_c)$ and insert them into the observational model to obtain predictions for the measured incidences (Figure 8). Qualitatively, the simulated trajectories appear to be reasonable

© 2021 The Authors. *System Dynamics Review* published by John Wiley & Sons Ltd on behalf of System Dynamics Society.
DOI: 10.1002/sdr

10991727, 2021, 4, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/sdr.1693 by Health Research Board, Wiley Online Library on [04/08/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

approximations to the reference behaviour. To verify this appraisal, for each trajectory, we calculate the mean absolute scaled error or MASE, a metric of forecast accuracy (Hyndman and Koehler, 2006). In practice, this procedure entails to define $f(\theta)$ as the MASE of each trajectory and average the results over $\pi(\theta|y_c)$. Considering that values lower than one indicate adequate predictive performance, and all simulated behaviours concentrate below such a threshold (see Appendix S1 section 3.3.2.4), we gain support to the claim that the simulation model explored in this article is an adequate structure to account for Cumberland's incidence data.

Policy analysis

As we have seen, the usefulness of estimating $\pi(\theta|y_c)$ is not exclusively confined to find a match between observed and simulated behaviours (model calibration). Once we have gained confidence in these results, we can employ this distribution to evaluate the future dynamics in similar settings where the model is relevant. To illustrate this procedure, we simulate the model in a hypothetical situation. In particular, we are interested in predicting the dynamics of an outbreak in a city of 10,000 people under two scenarios: *unmitigated* and *intervention*. The former corresponds to the scenario where the virus is left to run unchecked until the disease runs its course. The latter describes the implementation of social-distancing measures aimed at reducing the number of contacts among the population. To do so, we consider the sampling procedure. Stan returns a collection of draws for each calibrated parameter. In this case, the output forms a matrix of three columns (parameters) and 4000 rows (samples). Since the parameters exhibit correlation (Figure 6), we sample entire rows $\{\beta_i, \rho_i, I(0)_i\}$, where i denotes a specific row. Should the parameters be independent, we would sample separately from each column, yielding sets $\left\{\beta_j, \rho_k, I(0)_l\right\}$. We follow this procedure in situations where we cannot infer correlations from the data (e.g. report of marginal distributions).

Having established the sampling procedure, for the no-intervention scenario, we feed the SEIR model with the samples and run the simulation; whereas for the intervention scenario, we multiply all β_i by a factor of 40 percent to describe the effect of social-distancing measures implemented before the occurrence of the first case. As expected, decreasing the population's contact rate (β) translates into a slower transmission process with fewer cases. Undoubtedly, the added value of performing policy analysis from this Bayesian perspective stems from the fact that we simultaneously gauge the uncertainty in the predicted behaviours, offering a broader picture to decision-makers. Nevertheless, in this case, such uncertainty is tempered by the correlation among parameters. To visualise this, we also run the model with independently sampled draws. In Figure 9, it can be seen the extra

© 2021 The Authors. *System Dynamics Review* published by John Wiley & Sons Ltd on behalf of System Dynamics Society.
DOI: 10.1002/sdr

10991727, 2021, 4, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/sdr.1693 by Health Research Board, Wiley Online Library on [04/08/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

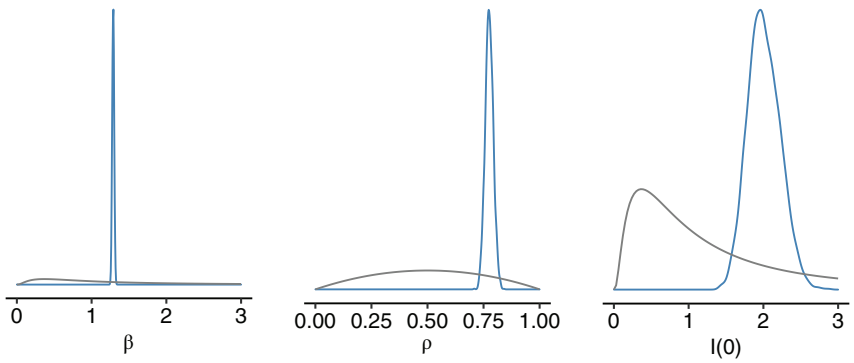

Fig. 6. Comparison between marginal prior and marginal posterior distributions for our three parameters. Grey lines denote prior distributions. Blue lines denote posterior distributions [Color figure can be viewed at wileyonlinelibrary.com]

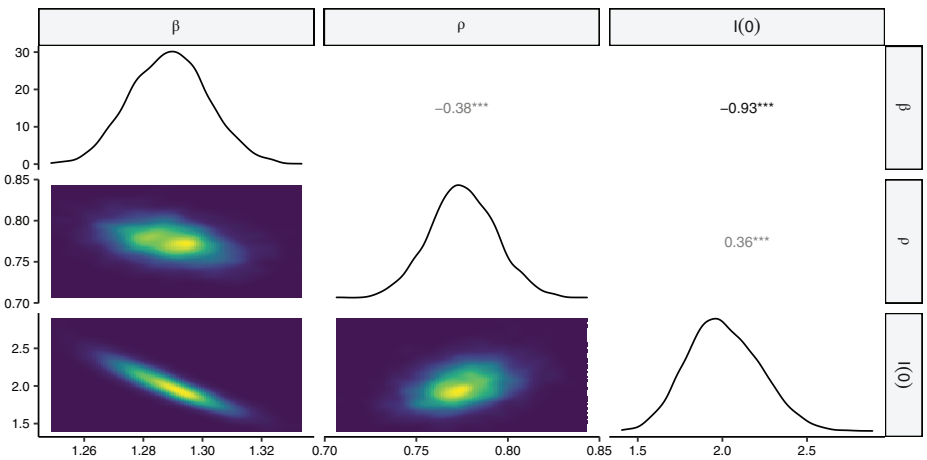

Fig. 7. Posterior distribution for our three parameters. The diagonal shows the posterior marginal distributions. In the lower triangular part, the joint posterior distribution of each possible combination of two parameters is displayed. The upper triangular part shows the correlation among parameters [Color figure can be viewed at wileyonlinelibrary.com]

uncertainty added by the independence assumption, evidenced by the excess of blue contour in comparison with that of the grey one. Notice that this application is also an instance of Eq. (10), namely an expectation.

Performance comparison

SD practitioners certainly require a clear indication that investing time and resources in applying a novel method is worth the effort. In the SD literature, we find two enhanced versions of the Metropolis algorithm: the *DREAM* sampler (Vrugt *et al.*, 2009) implemented in *Vensim*, and a Random-Walk Metropolis -RWM- algorithm (*MCMCmetrop1R*) offered by the *MCMCPack*, the method used by Osgood and Liu (2015). We select the latter for the

© 2021 The Authors. *System Dynamics Review* published by John Wiley & Sons Ltd on behalf of System Dynamics Society.
DOI: 10.1002/sdr

10991727, 2021, 4, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/sdr.1693 by Health Research Board, Wiley Online Library on [04/08/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

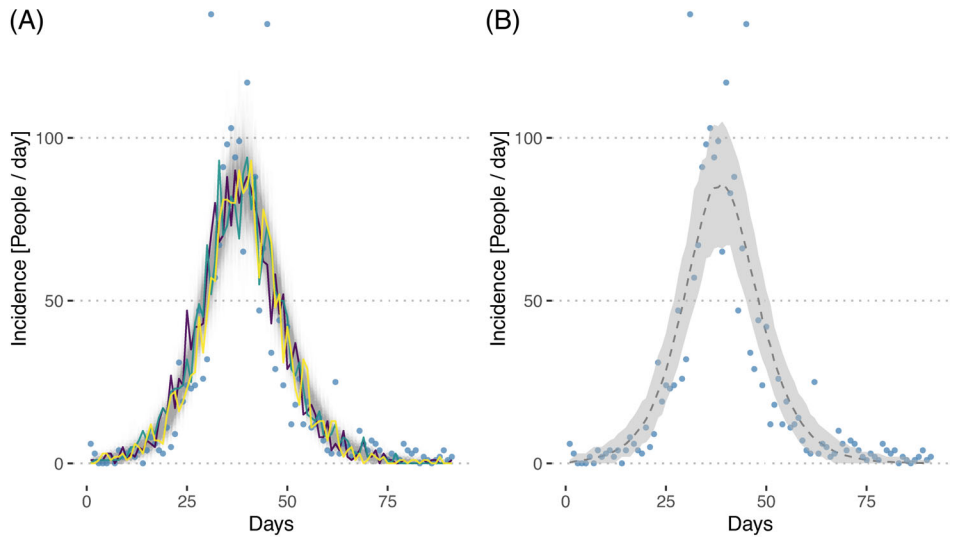

Fig. 8. (a) 500 predicted incidence measurements (grey lines) from the posterior predictive distribution (model's fit). To obtain a single predicted measurement, we draw a sample from the posterior distribution and use it to generate a trajectory from the observational model (Eq. (7)). Three different predicted incidence measurements are highlighted in Viridis colours. Dots denote the actual data. (b) Posterior predictive distribution described in terms of the mean (dashed line) and 95 percent credible intervals (contour) [Color figure can be viewed at wileyonlinelibrary.com]

comparison analysis, given that it is open source and a similar approach adopted throughout this work.

Specifically, we fit the SEIR model (presented in the *Context* section), under the conditions (priors and unknowns) described in the *Prior information* section, to Cumberland's incidence data. This calibration is performed in six different scenarios, which differ in the number of iterations (100, 200, 500, 1000, 1500, 2000) allocated to both MCMC algorithms. The reader can find the complete analysis in Appendix S2 section 2. The results show that HMC is computationally faster than RWM for obtaining an equal amount of samples. However, given technological implementations, it is not possible to definitively determine whether the performance differences are due to the algorithms themselves. For this reason, we compare technologically independent metrics of convergence ($\widehat{R}$) and efficiency (*ESS*). In Figure 10a, we observe that RWM requires at least 2000 burn-in samples so that all parameters reach convergence ($\widehat{R} < 1.01$), a value significantly higher than the equivalent number of samples (500) required by HMC. On the other hand, the effective sample size (*ESS*) is a measure of efficiency. This metric helps us answer: are X samples from RWM equivalent to X samples from HMC? The reader should recall that the *ESS* approximates the number of independent samples. We present two types of *ESS*: *bulk* and *tail* (see definitions in

© 2021 The Authors. *System Dynamics Review* published by John Wiley & Sons Ltd on behalf of System Dynamics Society.
DOI: 10.1002/sdr

10991727, 2021, 4, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/sdr.1693 by Health Research Board, Wiley Online Library on [04/08/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

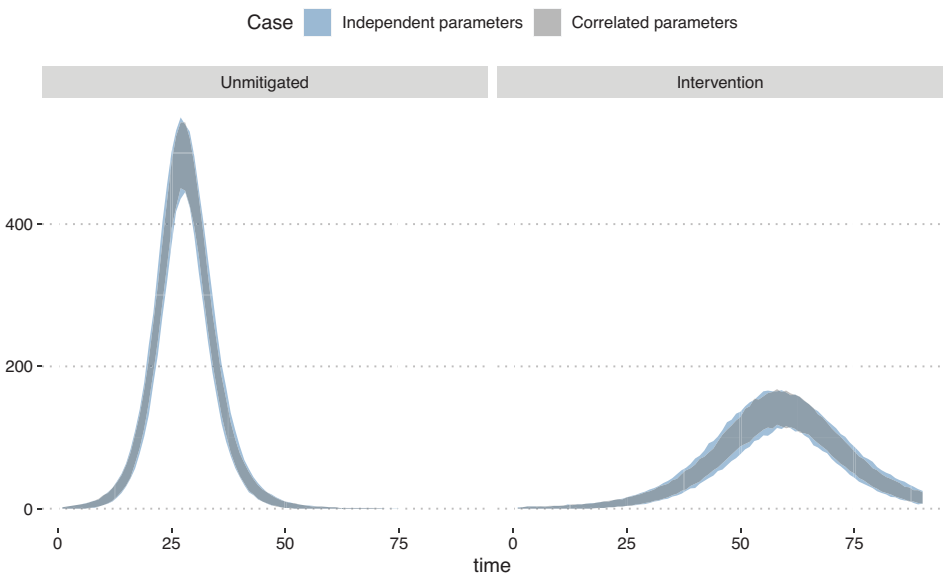

Fig. 9. Forecast of scenarios (unmitigated and intervention) under two parameter interaction assumptions. These assumptions correspond to the correlations revealed by the calibration process and perfect independence. We simulated 500 trajectories per experiment in this two-by-two design. For each experiment, we plot a silhouette of the predicted measured incidence. The width of these silhouettes corresponds to 95 percent credible intervals. Two silhouettes are superimposed per scenario [Color figure can be viewed at wileyonlinelibrary.com]

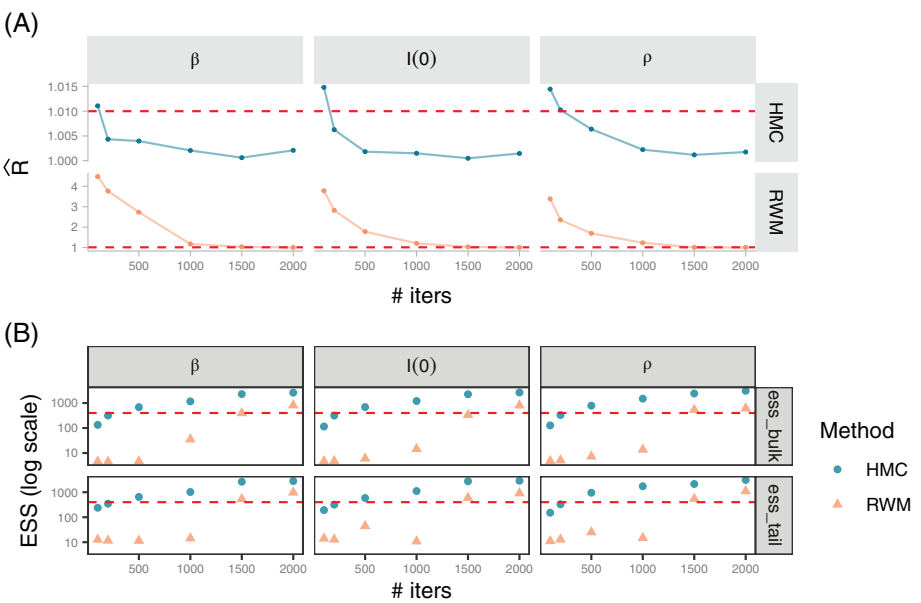

Fig. 10. (a) Potential scale reduction factor by scenario and calibration algorithm for our three parameters. Red dashed line denotes the acceptance threshold (1.01). (b) Effective sample size (bulk and tail) by scenario and calibration algorithm for our three parameters. Red dashed line denote the acceptance threshold (400) [Color figure can be viewed at wileyonlinelibrary.com]

© 2021 The Authors. *System Dynamics Review* published by John Wiley & Sons Ltd on behalf of System Dynamics Society.
DOI: 10.1002/sdr

10991727, 2021, 4, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/sdr.1693 by Health Research Board, Wiley Online Library on [04/08/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

Appendices S1 and S2). Both metrics should be at least 400 (Vehtari *et al.*, 2021). In Figure 10b, we see that HMC produces a higher number of ESS than RWM in all scenarios. Furthermore, HMC exceeds the 400-threshold from 500 iterations per chain, a third of the iterations required by RWM. In a nutshell, HMC converges faster, and its samples provide more information than those of RMW.

In terms of performance, we corroborate previous theoretical (Betancourt, 2017a) and practical studies that suggest HMC as the method of choice. For instance, Monnahan *et al.* (2017) found that HMC outperforms the Gibbs sampler (a random-walk MCMC algorithm) in estimating the parameters of population-ecology models (hierarchical and state-space) across a range of dimensions and complexity. This performance gap grew to the extent that HMC was 63 times more efficient when fitting a logistic model. Likewise, Beraha *et al.* (2021) conducted a systematic study on probabilistic programming languages (PPL). The authors evaluated three PPLs on four classes of models: linear, logistic regression, mixture models, and accelerated failure time. The results from this study indicate that Stan (using the NUTS algorithm) is the "default go-to software" over the other two random-walk-based platforms. To the best of our knowledge, the most recent benchmark analysis on ODE models was carried out a decade ago (Girolami and Calderhead, 2011), a time where the NUTS algorithm had not been developed. This observation suggests that future research endeavours should systematically explore the variations in the performance of MCMC algorithms across various ODE models.

Conclusion

In his reflection on the 60-year history of the SD field, Sterman (2018, p.40) encouraged practitioners to "master the state of the art and modern methods to develop, test, communicate, and implement rigorous, reliable and effective insights into the dynamics of complex systems." In that context, we introduce Hamiltonian Monte Carlo to the SD community to perform robust model calibration using a state-of-the-art statistical package (Stan). In doing so, we notice that valuable information about the results and the method itself is often missing in model calibration reports. Due to the established tradition of using nonlinear optimisation techniques, we often find that authors limit the calibration report to the mean and standard deviation of parameter estimates. This information is complete only in the case of symmetric and independent distributions, a set of assumptions that do not hold for this simple case study. On the contrary, MCMC-based methods provide richer information about the calibrated parameters and allow the practitioner to evaluate the robustness of the estimates. In the same fashion that a model's behaviour must be obtained from the right reasons, parameter estimates must

© 2021 The Authors. *System Dynamics Review* published by John Wiley & Sons Ltd on behalf of System Dynamics Society.
DOI: 10.1002/sdr

10991727, 2021, 4, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/sdr.1693 by Health Research Board, Wiley Online Library on [04/08/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

be obtained for the right reasons as well. To communicate this process, we suggest a workflow grounded on logic and visualisation. Such a workflow is possible due to the combination of SD Software, R, and Stan. This synergy produces robust results, facilitates reproducibility, and, more importantly, enhances the comprehension of the process undertaken.

Acknowledgement

This publication has emanated from research supported in part by a research grant from Science Foundation Ireland (SFI) under Grant Number SFI/12/RC/2289_P2, co-funded by the European Regional Development Fund. For the purpose of Open Access, the author has applied a CC BY public copyright licence to any Author Accepted Manuscript version arising from this submission. The project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 883285. The material presented and views expressed here are the responsibility of the author(s) only. The EU Commission takes no responsibility for any use made of the information set out.

Biographies

Jair Andrade is a PhD candidate at the National University of Ireland Galway. His research interests focus on applying simulation-based statistical methods to perform rigorous inference on non-linear complex models for the study of infectious diseases.

Jim Duggan is a personal professor in Computer Science at the National University of Ireland Galway, and his research interests centres on the application of system dynamics modelling and data science to enhance public health policy. He has recently applied system dynamics modelling with the Irish Epidemiological Modelling Advisory Group (IEMAG), where his role is to work as part of a team to develop and deliver epidemiological models to inform operational and policy decision making. Jim is a managing editor of the System Dynamics Review, and is a member of the World Health Organisation's Global Outbreak Alert and Response Network (GOARN).

References

Anderson RM, May RM. 1992. *Infectious Diseases of Humans: Dynamics and Control*. Oxford University Press: Oxford.

Andrade, J., 2021. Readsdr: translate models from system dynamics software into 'R'.

© 2021 The Authors. *System Dynamics Review* published by John Wiley & Sons Ltd on behalf of System Dynamics Society.
DOI: 10.1002/sdr

10991727, 2021, 4, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/sdr.1693 by Health Research Board, Wiley Online Library on [04/08/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

Andrade J, Duggan J. 2020. An evaluation of Hamiltonian Monte Carlo performance to calibrate age-structured compartmental SEIR models to incidence data. *Epidemics* **33**: 100415. https://doi.org/10.1016/j.epidem.2020.100415.

Ansah JP, Koh V, Chiu C-T, Chei C-L, Zeng Y, Yin Z-X, Shi X-M, Matchar DB. 2017. Projecting the number of elderly with cognitive impairment in China using a multi-state dynamic population model. *System Dynamics Review* **33**: 89–111. https://doi.org/10.1002/sdr.1581.

Barlas Y. 1996. Formal aspects of model validity and validation in system dynamics. *System Dynamics Review* **12**: 183–210. https://doi.org/10.1002/(SICI)1099-1727(199623)12:3<183::AID-SDR103>3.0.CO;2-4.

Beraha, M., Falco, D., Guglielmi, A. 2021. JAGS, NIMBLE, stan: a detailed comparison among Bayesian MCMC software.

Betancourt, M., 2017a. A conceptual introduction to Hamiltonian Monte Carlo.

Betancourt, M., 2017b. The convergence of Markov chain Monte Carlo methods: from the metropolis method to Hamiltonian Monte Carlo.

Betancourt, M.J. 2015. Adiabatic Monte Carlo.

Betancourt, M.J., Byrne, S., Livingstone, S., Girolami, M., 2014. The geometric foundations of Hamiltonian Monte Carlo.

Blitzstein JK, Hwang J. 2019. *Introduction to Probability, Second edition, Chapman & Hall/CRC Texts in Statistical Science*. Second edition, Boca Raton: CRC Press.

Carpenter B, Gelman A, Hoffman MD, Lee D, Goodrich B, Betancourt M, Brubaker M, Guo J, Li P, Riddell A. 2017. Stan: a probabilistic programming language. *Journal of Statistical Software* **76**: 1–32. https://doi.org/10.18637/jss.v076.i01

Carpenter, B., Hoffman, M.D., Brubaker, M., Lee, D., Li, P., Betancourt, M., 2015. The Stan math library: reverse-mode automatic differentiation in C++.

Chatzilena A, van Leeuwen E, Ratmann O, Baguelin M, Demiris N. 2019. Contemporary statistical inference for infectious disease models using Stan. *Epidemics* **29**: 100367. https://doi.org/10.1016/j.epidem.2019.100367.

Chowell G, Nishiura H, Bettencourt LMA. 2007. Comparative estimation of the reproduction number for pandemic influenza from daily case notification data. *Journal of the Royal Society Interface* **4**: 155–166. https://doi.org/10.1098/rsif.2006.0161.

Dangerfield B, Duggan J. 2020. Optimization of system dynamics models. In *System Dynamics: Theory and Applications*, Dangerfield B (ed). Springer US: New York, NY; 139–152. https://doi.org/10.1007/978-1-4939-8790-0_542.

Davies NG, Klepac P, Liu Y, Prem K, Jit M, Eggo RM, CMMID COVID-19 working group. 2020. Age-dependent effects in the transmission and control of COVID-19 epidemics. *Nature Medicine* **26**: 1205–1211. https://doi.org/10.1038/s41591-020-0962-9.

Dogan G. 2007. Bootstrapping for confidence interval estimation and hypothesis testing for parameters of system dynamics models. *System Dynamics Review* **23**: 415–436. https://doi.org/10.1002/sdr.362.

Duggan J. 2016. *System Dynamics Modeling with R, Lecture Notes in Social Networks*. Basel: Springer International Publishing.

Farrington CP, Kanaan MN, Gay NJ. 2001. Estimation of the basic reproduction number for infectious diseases from age-stratified serological survey data. *Journal of the Royal Statistical Society: Series C (Applied Statistics)* **50**: 251–292. https://doi.org/10.1111/1467-9876.00233.

© 2021 The Authors. *System Dynamics Review* published by John Wiley & Sons Ltd on behalf of System Dynamics Society.
DOI: 10.1002/sdr

10991727, 2021, 4, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/sdr.1693 by Health Research Board, Wiley Online Library on [04/08/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

Forrester JW. 1993. System dynamics and the lessons of 35 years. In *A Systems-Based Approach to Policymaking*, De Greene KB (ed). Springer US: Boston, MA; 199–240. https://doi.org/10.1007/978-1-4615-3226-2_7.

Frost WH, Sydenstricker E. 1919. Influenza in Maryland: preliminary statistics of certain localities. *Public Health Reports (1896-1970)* **34**: 491–504.

Gabry J, Simpson D, Vehtari A, Betancourt M, Gelman A. 2019. Visualization in bayesian workflow. *Journal of the Royal Statistical Society: Series A (Statistics in Society)* **182**: 389–402. https://doi.org/10.1111/rssa.12378.

Gamado KM, Streftaris G, Zachary S. 2014. Modelling under-reporting in epidemics. *Journal of Mathematical Biology* **69**: 737–765. https://doi.org/10.1007/s00285-013-0717-z.

Gelfand AE, Smith AFM. 1990. Sampling-based approaches to calculating marginal densities. *Journal of the American Statistical Association* **85**: 398–409.

Gelman A, Carlin JB, Stern HS, Dunson DB, Vehtari A, Rubin DB. 2013. *Bayesian Data Analysis*. Third edition, Boca Raton: CRC Press.

Gelman A, Hill J. 2007. *Data Analysis Using Regression and Multilevel/Hierarchical Models, Analytical Methods for Social Research*. Cambridge: Cambridge University Press.

Gelman A, Rubin DB. 1992. Inference from iterative simulation using multiple sequences. *Statistical Science* **7**: 457–472.

Gelman A, Simpson D, Betancourt M. 2017. The prior can often only be understood in the context of the likelihood. *Entropy* **19**(10): 555. https://doi.org/10.3390/e19100555.

Gelman, A., Vehtari, A., Simpson, D., Margossian, C.C., Carpenter, B., Yao, Y., Kennedy, L., Gabry, J., Bürkner, P.-C., Modrák, M., 2020. Bayesian workflow.

Geman S, Geman D. 1984. Stochastic relaxation, gibbs distributions, and the Bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **6**: 721–741. https://doi.org/10.1109/TPAMI.1984.4767596.

Geyer CJ. 2011. Introduction to markov chain Monte Carlo. In *Handbook of Markov Chain Monte Carlo*. Boca Raton: CRC Press; 3–48.

Ghaffarzadegan N, Rahmandad H. 2020. Simulation-based estimation of the early spread of COVID-19 in Iran: actual versus confirmed cases. *System Dynamics Review* **36**: 101–129. https://doi.org/10.1002/sdr.1655.

Girolami M, Calderhead B. 2011. Riemann manifold Langevin and Hamiltonian Monte Carlo methods. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **73**: 123–214. https://doi.org/10.1111/j.1467-9868.2010.00765.x.

Graham A. 1980. Parameter estimation in system dynamics modeling. In *Elements of the System Dynamics Method*. New York: Productivity Press; 143–161.

Hastings WK. 1970. Monte Carlo sampling methods using Markov chains and their applications. *Biometrika* **57**: 97–109.

Hoffman, M.D., Gelman, A., 2011. The no-u-turn sampler: adaptively setting path lengths in Hamiltonian Monte Carlo.

Hyndman RJ, Koehler AB. 2006. Another look at measures of forecast accuracy. *International Journal of Forecasting* **22**: 679–688. https://doi.org/10.1016/j.ijforecast.2006.03.001.

Keith DR, Sterman JD, Struben J. 2017. Supply constraints and waitlists in new product diffusion. *System Dynamics Review* **33**: 254–279. https://doi.org/10.1002/sdr.1588.

© 2021 The Authors. *System Dynamics Review* published by John Wiley & Sons Ltd on behalf of System Dynamics Society.
DOI: 10.1002/sdr

10991727, 2021, 4, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/sdr.1693 by Health Research Board, Wiley Online Library on [04/08/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

McElreath R. 2020. *Statistical Rethinking: A Bayesian Course with Examples in R and Stan*. Second edition, Boca Raton: CRC Press.

Metropolis N, Rosenbluth AW, Rosenbluth MN, Teller AH, Teller E. 1953. Equation of state calculations by fast computing machines. *The Journal of Chemical Physics* **21**: 1087–1092. https://doi.org/10.1063/1.1699114.

Mills CE, Robins JM, Lipsitch M. 2004. Transmissibility of 1918 pandemic influenza. *Nature* **432**: 904–906. https://doi.org/10.1038/nature03063.

Monnahan CC, Thorson JT, Branch TA. 2017. Faster estimation of Bayesian models in ecology using Hamiltonian Monte Carlo. *Methods in Ecology and Evolution* **8**: 339–348. https://doi.org/10.1111/2041-210X.12681.

Neal R. 2011. MCMC using Hamiltonian dynamics. In *Handbook of Markov Chain Monte Carlo*. Boca Raton: CRC Press; 116–162.

Oliva R. 2003. Model calibration as a testing strategy for system dynamics models. *European Journal of Operational Research* **151**: 552–568. https://doi.org/10.1016/S0377-2217(02)00622-7.

Oliva R, Sterman JD. 2001. Cutting corners and working overtime: quality erosion in the service industry. *Management Science* **47**: 894–914. https://doi.org/10.1287/mnsc.47.7.894.9807.

Osgood N, Liu J. 2015. Combining markov chain Monte Carlo approaches and dynamic modeling. In Rahmandad, H, Oliva, R & Osgood, N (Eds.), *Analytical Methods for Dynamic Modelers*. Cambridge: MIT Press; 125–170.

Patterson K, Pyle G. 1991. The geography and mortality of the 1918 influenza pandemic. *Bulletin of the History of Medicine* **65**: 4–21.

Pawitan Y. 2013. *In all Likelihood: Statistical Modelling and Inference Using Likelihood, in all Likelihood: Statistical Modelling and Inference Using Likelihood*. Oxford: Oxford University Press.

Pierson K, Sterman JD. 2013. Cyclical dynamics of airline industry earnings. *System Dynamics Review* **29**: 129–156. https://doi.org/10.1002/sdr.1501.

Robert C, Casella G. 2005. *Monte Carlo Statistical Methods, Springer texts in statistics*. Springer: New York.

Robert C, Casella G. 2010. *Introducing Monte Carlo Methods with R, Use R!*. New York: Springer.

Robert C, Casella G. 2011. A short history of MCMC: subjective recollections from incomplete data. In *Handbook of Markov Chain Monte Carlo*. Boca Raton: CRC Press; 49–66.

Saleh M, Oliva R, Kampmann CE, Davidsen PI. 2010. A comprehensive analytical approach for policy analysis of system dynamics models. *European Journal of Operational Research* **203**: 673–683. https://doi.org/10.1016/j.ejor.2009.09.016.

Sterman J. 2018. System dynamics at sixty: the path forward. *System Dynamics Review* **34**: 5–47. https://doi.org/10.1002/sdr.1601.

Struben J, Sterman J, Keith D. 2015. Parameter estimation through maximum likelihood and bootstrapping methods. In Rahmandad, H, Oliva, R & Osgood, N (Eds.), *Analytical Methods for Dynamic Modelers*. Cambridge: MIT Press; 3–38.

Valderrama-Bahamóndez GI, Fröhlich H. 2019. MCMC techniques for parameter estimation of ODE based models in systems biology. *Frontiers in Applied Mathematics and Statistics* **5**: 55. https://doi.org/10.3389/fams.2019.00055.

© 2021 The Authors. *System Dynamics Review* published by John Wiley & Sons Ltd on behalf of System Dynamics Society.
DOI: 10.1002/sdr

10991727, 2021, 4, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/sdr.1693 by Health Research Board, Wiley Online Library on [04/08/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

Vehtari A, Gelman A, Simpson D, Carpenter B, Bürkner P-C. 2021. Rank-normalization, folding, and localization: an improved $\hat{R}$ for assessing convergence of MCMC. *Bayesian Analysis* **16**: 1–28. https://doi.org/10.1214/20-BA1221.

Vrugt JA, ter Braak CJF, Diks CGH, Robinson BA, Hyman JM, Higdon D. 2009. Accelerating Markov chain Monte Carlo simulation by differential evolution with self-adaptive randomized subspace sampling. *International Journal of Nonlinear Sciences and Numerical Simulation* **10**: 273–290. https://doi.org/10.1515/IJNSNS.2009.10.3.273.

Vynnycky E, Trindall A, Mangtani P. 2007. Estimates of the reproduction numbers of Spanish influenza using morbidity data. *International Journal of Epidemiology* **36**: 881–889. https://doi.org/10.1093/ije/dym071.

Vynnycky E, White R. 2010. *An Introduction to Infectious Disease Modelling*. Oxford: Oxford University Press.

Supporting information

Additional supporting information may be found in the online version of this article at the publisher's website.

Appendix S1. Supporting Information.

Appendix S2. Supporting Information.

Appendix S3. Supporting Information.

© 2021 The Authors. *System Dynamics Review* published by John Wiley & Sons Ltd on behalf of System Dynamics Society.
DOI: 10.1002/sdr

10991727, 2021, 4, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/sdr.1693 by Health Research Board, Wiley Online Library on [04/08/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License