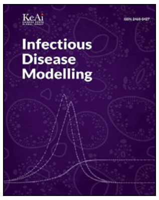
Contents lists available at ScienceDirect

Infectious Disease Modelling

journal homepage: www.keaipublishing.com/idm



Identifying the measurements required to estimate rates of COVID-19 transmission, infection, and detection, using variational data assimilation

Eve Armstrong [a], [b], [*], Manuela Runge [c], Jaline Gerardin [c]

[a] *Department of Physics, New York Institute of Technology, 1855 Broadway, New York, NY, 10023, USA*
[b] *Department of Astrophysics, American Museum of Natural History, New York, NY, 10024, USA*
[c] *Department of Preventive Medicine, Northwestern University, 710 N Lake Shore Drive Suite 800, Chicago, IL, 60611, USA*

ARTICLE INFO

Article history:
Received 5 August 2020
Received in revised form 26 October 2020
Accepted 27 October 2020
Available online 2 November 2020
Handling editor: Dr. J Wu

Keywords:
COVID-19
Epidemiology
Inference
Measurement noise

ABSTRACT

We demonstrate the ability of statistical data assimilation (SDA) to identify the measurements required for accurate state and parameter estimation in an epidemiological model for the novel coronavirus disease COVID-19. Our context is an effort to inform policy regarding social behavior, to mitigate strain on hospital capacity. The model unknowns are taken to be: the time-varying transmission rate, the fraction of exposed cases that require hospitalization, and the time-varying detection probabilities of new asymptomatic and symptomatic cases. In simulations, we obtain estimates of undetected (that is, unmeasured) infectious populations, by measuring the detected cases together with the recovered and dead - and without assumed knowledge of the detection rates. Given a noiseless measurement of the recovered population, excellent estimates of all quantities are obtained using a temporal baseline of 101 days, with the exception of the time-varying transmission rate at times prior to the implementation of social distancing. With low noise added to the recovered population, accurate state estimates require a lengthening of the temporal baseline of measurements. Estimates of all parameters are sensitive to the contamination, highlighting the need for accurate and uniform methods of reporting. The aim of this paper is to exemplify the power of SDA to determine what properties of measurements will yield estimates of unknown parameters to a desired precision, in a model with the complexity required to capture important features of the COVID-19 pandemic.

1. Introduction

The coronavirus disease 2019 (COVID-19) is burdening health care systems worldwide, threatening physical and psychological health, and devastating the global economy. Individual countries and states are tasked with balancing population-level mitigation measures with maintaining economic activity. Mathematical modeling has been used to aid policymakers'

[*] Corresponding author. Department of Physics, New York Institute of Technology, 1855 Broadway, New York, NY, 10023, USA.
E-mail address: evearmstrong.physics@gmail.com (E. Armstrong).
Peer review under responsibility of KeAi Communications Co., Ltd.

https://doi.org/10.1016/j.idm.2020.10.010

plans for hospital capacity needs, and to understand the minimum criteria for effective contact tracing (Murray et al. 2020). Both state-level decision-making and accurate modeling benefit from quality surveillance data. Insufficient testing capacity, however, especially at the beginning of the epidemic in the United States, and other data reporting issues have meant that surveillance data on COVID-19 is biased and incomplete (Heggeness, 2020; Li et al., 2020a; Weinberger et al., 2020). Models intended to guide intervention policy must be able to handle imperfect data.

Within this context, we seek a means to quantify what data must be recorded in order to estimate specific unknown quantities in an epidemiological model of COVID-19 transmission. These unknown quantities are: i) the transmission rate, ii) the fraction of the exposed population that acquires symptoms sufficiently severe to require hospitalization, and iii) time-varying detection probabilities of asymptomatic and symptomatic cases. In this paper, we demonstrate the ability of statistical data assimilation (SDA) to quantify the accuracy to which these parameters can be estimated, given certain properties of the data including noise level.

SDA is an inverse formulation (Tarantola, 2005): a machine learning approach designed to optimally combine a model with data. Invented for numerical weather prediction (An et al. 2017; Betts, 2010; Evensen, 2009; Kalnay, 2003; Kimura, 2002; Whartenby et al., 2013), and more recently applied to biological neuron models (Armstrong, 2020; Meliza et al., 2014; Hamilton et al., 2013; Kostuk et al., 2012; Nogaret et al., 2016; Schiff, 2009; Toth et al., 2011), SDA offers a systematic means to identify the measurements required to estimate unknown model parameters to a desired precision.

Data assimilation has been presented as a means for general epidemiological forecasting (Bettencourt, Ribeiro, Chowell, Lant, & Castillo-Chavez, 2007), and one work has examined variational data assimilation specifically - the method we employ in this paper - for estimating parameters in epidemiological models (Rhodes & Hollingsworth, 2009). Related Bayesian frameworks for estimating unknown properties of epidemiological models have also been explored (Bettencourt & Ribeiro, 2008; Cobb et al., 2014). To date, there have been two employments of SDA for COVID-19 specifically. Ref (Sesterhenn, 2020) used a simple SIR (susceptible/infected/recovered) model, and Ref (Nadler et al., 2020) expanded the SIR model to include a compartment of patients in treatment. Another study has used a Bayesian inference framework to examine a fully stochastic epidemiological model, with relevance to COVID-19 (Li et al., 2020b).

Two features of our work distinguish this paper as novel. First, we expand the model in terms of the number of compartments. The aim here is to capture key epidemiological and public health intervention features of COVID-19 such that the model structure is relevant for questions from policymakers on containing the pandemic. These features are: i) asymptomatic, presymptomatic, and symptomatic populations, ii) undetected and detected cases, and iii) two hospitalized populations: those who do and do not require critical care. For our motivations for these choices, see *Model*. Second, we employ SDA for the specific purpose of examining the sensitivity of estimates of time-varying parameters to various properties of the measurements, including the degree of noise (or error) added. Moreover, we aim to demonstrate the power and versatility of the SDA technique to explore what is required of measurements to complete a model with a dimension sufficiently high to capture the policy-relevant complexities of COVID-19 transmission and containment - an examination that has not previously been done.

To this end, we sought to estimate the parameters noted above, using simulated data representing a metropolitan-area population loosely based on New York City. We examined the sensitivity of estimations to: i) the subpopulations that were sampled, ii) the temporal baseline of sampling, and iii) uncertainty in the sampling.

Results using simulated data are threefold. First, reasonable estimations of time-varying detection probabilities require the reporting of new detected cases (asymptomatic and symptomatic), dead, and recovered. Second, given noiseless measurements, a temporal baseline of 101 days is sufficient for the SDA procedure to capture the general trends in the evolution of the model populations, the detection probabilities, and the time-varying transmission rate following the implementation of social distancing. Importantly, the information contained in the measured *detected* populations propagates successfully to the estimation of the numbers of severe *undetected* cases. Third, the state evolution - and importantly the populations requiring inpatient care - tolerates low (~ five percent) noise, given a doubling of the temporal baseline of measurements; the parameter estimates do not tolerate this contamination.

Finally, we discuss necessary modifications prior to testing with real data, including lowering the sensitivity of parameter estimates to noise in data.

2. Model

The model is written in 22 state variables, each representing a subpopulation of people; the total population is conserved. Fig. 1 shows a schematic of the model structure. Each member of a Population S that becomes Exposed (E) ultimately reaches either the Recovered (R) or Dead (D) state. *Absent additive noise, the model is deterministic.* Five variables correspond to measured quantities in the inference experiments.

As noted, the model is written with the aim to inform policy on social behavior and contact tracing so as to avoid exceeding hospital capacity. To this end, the model resolves asymptomatic-versus-symptomatic cases, undetected-versus-detected cases, and the two tiers of hospital needs: the general (inpatient, non-intensive care unit (ICU)) H versus the critical care

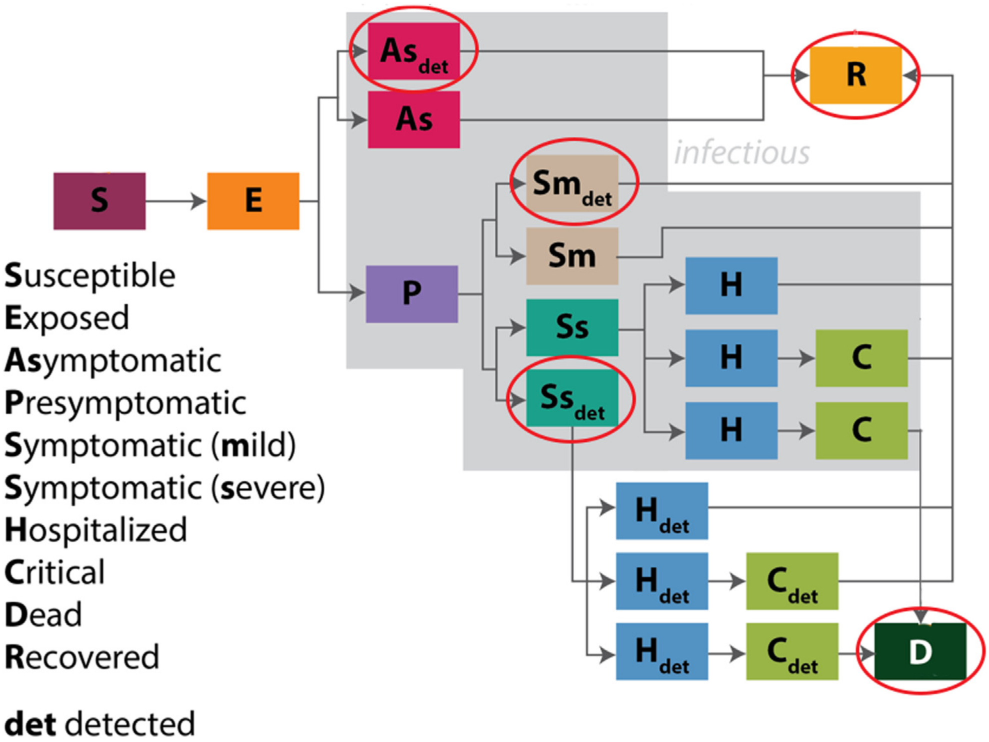


Fig. 1. Schematic of the model. Each rectangle represents a population. Note the distinction of asymptomatic cases, undetected cases, and the two tiers of hospitalized care: H and C. The aim of including this degree of resolution is to inform policy on social behavior so as to minimize strain on hospital capacity. The red ovals indicate the variables that correspond to measured quantities in the inference experiments.

(ICU) C populations. The resolution of asymptomatic versus symptomatic cases was motivated by an interest in what interventions are necessary to control the epidemic. For example, is it sufficient to focus only on symptomatic individuals, or must we also target and address asymptomatic individuals who may not even realize they are infected?

The detected and undetected populations exist for two reasons. First, we seek to account for underreporting of cases and deaths. Second, we desire a model structure that can simulate the impact of increasing detection rates on disease transmission, including the impact of contact tracing. Thus the model was structured from the beginning so that we might examine the effects of interventions that were imposed later on. The ultimate aim here is to inform policy on the requirements for containing the epidemic.

We included both H and C populations because hospital inpatient and ICU bed capacities are the key health system metrics that we aim to avoid straining. Any policy that we consider must include predictions on inpatient and ICU bed needs. Preparing for those needs is a key response if or when the epidemic grows uncontrolled.

For details of the model, including the differential equations describing the mass action between susceptible and infectious individuals and the disease progression through different sub-populations, see Appendix A.

3. Method

3.1. General inference formulation

SDA is an inference procedure, or a type of machine learning, in which a model dynamical system is assumed to underlie any measured quantities. This model F can be written as a set of *D* ordinary differential equations that evolve in some parameterization t as:

$$\frac{dx_a(t)}{dt} = F_a(\boldsymbol{x}(t), \boldsymbol{p}(t)); \quad a = 1, 2, \ldots, D,$$

where the components xa of the vector x are the model state variables, and unknown parameters to be estimated are contained in p(t). A subset L of the D state variables is associated with measured quantities. One seeks to estimate the p unknown parameters and the evolution of all state variables that is consistent with the L measurements. A prerequisite for estimation using real data is the design of simulated experiments, wherein the true values of parameters are known. In addition to providing a consistency check, simulated experiments offer the opportunity to ascertain *which* and *how few* experimental measurements, in principle, are necessary and sufficient to complete a model.

3.2. Optimization framework

SDA can be formulated as an optimization, wherein a cost function is extremized. We take this approach, and write the cost function in two terms: 1) one term representing the difference between state estimate and measurement (measurement error), and 2) a term representing model error. It will be shown below in this Section that treating the model error as finite offers a means to identify whether a solution has been found within a particular region of parameter space. This is a non-trivial problem, as any nonlinear model will render the cost function non-convex. We search the surface of the cost function via the variational method, and we employ a method of annealing to identify a lowest minimum - a procedure that has been referred to loosely in the literature as variational annealing (VA).

The cost function A_0 used in this paper is written as:

$$A_0(\boldsymbol{x}(n),\boldsymbol{p}) \quad = \sum_{j=1}^{J}\sum_{l=1}^{L}\frac{R_m^l}{2}(y_l(n)-x_l(n))^2 + \sum_{n=1}^{N-1}\sum_{a=1}^{D}\frac{R_f^a}{2}(x_a(n+1)-f_a(\boldsymbol{x}(n),\boldsymbol{p}(n)))^2. \tag{1}$$

One seeks the path $X_0 = x\ (0), \ldots, x(N), p\ (0), \ldots p(N)$ in state space on which A_0 attains a minimum value[1]. Note that Equation (1) is shorthand; for the full form, see Appendix A of Ref (Armstrong, 2020). For a derivation - beginning with the physical Action of a particle in state space - see Ref (Abarbanel, 2013).

The first squared term of Equation (1) governs the transfer of information from measurements y_l to model states x_l. The summation on j runs over all discretized timepoints J at which measurements are made, which may be a subset of all integrated model timepoints. The summation on l is taken over all L measured quantities.

The second squared term of Equation (1) incorporates the model evolution of all D state variables x_a. The term $f_a\ (x(n))$ is defined, for discretization, as: $\frac{1}{2}[F_a(\boldsymbol{x}(n)) + F_a(\boldsymbol{x}(n+1))]$. The outer sum on n is taken over all discretized timepoints of the model equations of motion. The sum on a is taken over all D state variables.

R_m and R_f are inverse covariance matrices for the measurement and model errors, respectively. We take each matrix to be diagonal and treat them as relative weighting terms, whose utility will be described below in this Section.

The procedure searches a $(D\,(N+1)+p\,(N+1))$-dimensional state space, where D is the number of state variables, N is the number of discretized steps, and p is the number of unknown parameters. To perform simulated experiments, the equations of motion are integrated forward to yield simulated data, and the VA procedure is challenged to infer the parameters and the evolution of all state variables - measured and unmeasured - that generated the simulated data.

This specific formulation has been tested with chaotic models (Abarbanel et al., 2011; Rey et al., 2014; Ye et al., 2014, 2015), and used to estimate parameters in models of biological neurons (Armstrong, 2020; Meliza et al., 2014; Kadakia et al., 2016; Kostuk et al., 2012; Toth et al., 2011; Wang, Breen, Abraham, Abarbanel, & Cauwenberghs, 2016), as well as astrophysical scenarios (Armstrong et al. 2017).

3.3. Annealing to identify a solution on a non-convex cost function surface

Our model is nonlinear, and thus the cost function surface is non-convex. For this reason, we iterate - or anneal - in terms of the ratio of model and measurement error, with the aim to gradually freeze out a lowest minimum. This procedure was introduced in Ref (Ye et al., 2015), and has since been used in combination with variational optimization on nonlinear models in Refs (An et al. 2017; Armstrong, 2020; Armstrong et al. 2017; Kadakia et al., 2016) above. The annealing works as follows.

We first define the coefficient of measurement error R_m to be 1.0, and write the coefficient of model error R_f as: $R_f \ = R_{f,0}\alpha^{\beta}$, where $R_{f,0}$ is a small number near zero, α is a small number greater than 1.0, and β is initialized at zero. Parameter β is our annealing parameter. For the case in which $\beta = 0$, relatively free from model constraints the cost function surface is smooth and there exists one minimum of the variational problem that is consistent with the measurements. We obtain an estimate of that minimum. Then we increase the weight of the model term slightly, via an integer increment in β, and recalculate the cost. We do this recursively, toward the deterministic limit of $R_f \gg R_m$. The aim is to remain sufficiently near to the lowest minimum to not become trapped in a local minimum as the surface becomes resolved. We will show in Results that a plot of the cost as a function of β reveals whether a solution has been found that is consistent with both measurements and model.

[1] It may interest the reader that one can derive this cost function by considering the classical physical Action on a path in a state space, where the path of lowest Action corresponds to the correct solution (Abarbanel, 2013)

4. The experiments

4.1. Simulated experiments

We based our simulated locality loosely on New York City, with a population of 9 million. For simplicity, we assume a closed population. Simulations ran from an initial time t_0 of four days prior to 2020 March 1, the date of the first reported COVID-19 case in New York City (New York Times, 2020). At time t_0, there existed one detected symptomatic case within the population of 9 million. In addition to that one initial detected case, we took as our initial conditions on the populations to be: 50 undetected asymptomatics, 10 undetected mild symptomatics, and one undetected severe symptomatic.[2]

We chose five quantities as unknown parameters to be estimated (Table 1): 1) the time-varying transmission rate $K_i(t)$; 2) the detection probability of mild symptomatic cases $d_{Sym}(t)$, 3) the detection probability of severe symptomatic cases $d_{Sys}(t)$, 4) the fraction of cases that become symptomatic fsympt, and 5) the fraction of symptomatic cases that become severe enough to require hospitalization fsevere. Here we summarize the key features that we sought to capture in modeling these parameters; for their mathematical formulations, see Appendix B.

The transmission rate K_i (often referred to as the effective contact rate) in a given population for a given infectious disease is measured in effective contacts per unit time. This may be expressed as the total contact rate multiplied by the risk of infection, given contact between an infectious and a susceptible individual. The contact rate, in turn, can be impacted by amendments to social behavior[3].

As a first step in applying SDA to a high-dimensional epidemiological model, we chose to condense the significance of K_i into a relatively simple mathematical form. We assumed that K_i was constant prior to the implementation of a social-distancing mandate, which then effected a rapid transition of K_i to a lower constant value. Specifically, we modeled K_i as a smooth approximation to a Heaviside function that begins its decline on March 22, the date that the stay-at-home order took effect in New York City (NY Governor's Office, 2020): 25 days after time t_0. For further simplicity, we took K_i to reflect a single implementation of a social distancing protocol, and adherence to that protocol throughout the remaining temporal baseline of estimation.

Detection rates impact the sizes of the subpopulations entering hospitals, and their values are highly uncertain (Li et al., 2020a; Weinberger et al., 2020). Thus we took these quantities to be unknown, and - as detection methods will evolve - time-varying. We also optimistically assumed that the methods will improve, and thus we described them as increasing functions of time. We used smoothly-varying forms, the first linear and the second quadratic, to preclude symmetries in the model equations. Meanwhile, we took the detection probability for asymptomatic cases (d_{As}) to be known and zero, a reasonable reflection of the state of testing in that population during our study period.

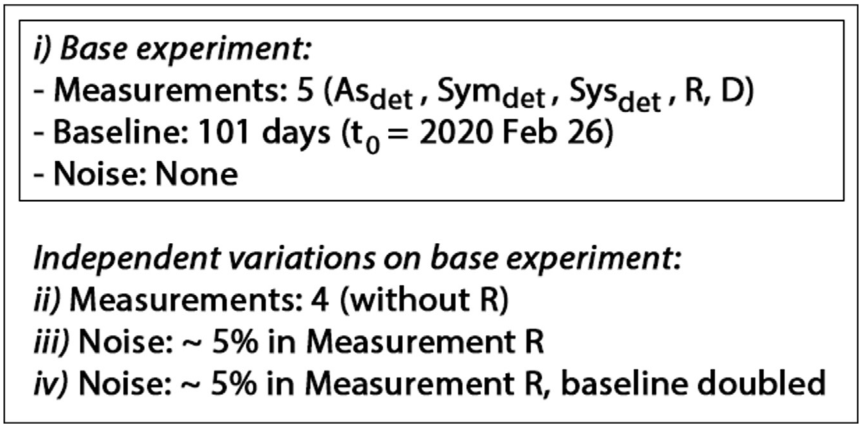
Finally, we assigned as unknowns the fraction of cases that become symptomatic (f_{sympt}) and fraction of symptomatic cases that become sufficiently severe to require hospitalization (f_{severe}), as these fractions possess high uncertainties (Refs (Oran and Topol, 2020) and (Salje et al., 2020), respectively). As they reflect an intrinsic property of the disease, we took them to be constants. All other model parameters were taken to be known and constant (Appendix A); however, the values of many other model parameters also possess significant uncertainties given the reported data, including, for example, the fraction of those hospitalized that require ICU care. Future VA experiments can treat these quantities as unknowns as well.

Table 1
Unknown parameters to be estimated. K_i, d_{Sym}, and d_{Sys} are taken to be time-varying. Parameters fsympt and fsevere are constant numbers, as they are assumed to reflect an intrinsic property of the disease. The detection probability of asymptomatic cases is taken to be known and zero.

Parameter	Description
$K_i(t)$	Time-varying transmission rate
$d_{Sym}(t)$	Time-varying detection probability of mild symptomatics
$d_{Sys}(t)$	Time-varying detection probability of symptomatics requiring hospitalization
f_{sympt}	Fraction of positive cases that produce symptoms
f_{severe}	Fraction of symptomatics that are severe

[2] Our choices for these numbers were difficult to make, given the scant data and poor understanding of the evolutionary history of the disease that existed at the time that this study was conducted. Ultimately, we chose numbers that represent neither lower nor upper limits, but rather plausible scenarios. The general temporal evolution of the model sub-populations was not sensitive to changes in these numbers of several tens of percentage points, but it will be important in future work to conduct a more detailed study on model sensitivity to initial conditions.

[3] The reproduction number R_0, in the simplest SIR form, can be written as the effective contact rate divided by the recovery rate. In practice, R_0 is a challenge to infer (Bettencourt and Ribeiro, 2008; Thompson et al., 2019; Cori et al., 2013; Wallinga and Teunis, 2004).



i) Base experiment:
- Measurements: 5 (As_{det}, Sym_{det}, Sys_{det}, R, D)
- Baseline: 101 days (t_0 = 2020 Feb 26)
- Noise: None

Independent variations on base experiment:
ii) Measurements: 4 (without R)
iii) Noise: ~ 5% in Measurement R
iv) Noise: ~ 5% in Measurement R, baseline doubled

Fig. 2. Schematic of the four simulated experiments.

The simulated experiments are summarized in the schematic of Fig. 2. They were designed to probe the effects upon estimations of three considerations: a) the number of measured subpopulations, b) the temporal baseline of measurements, and c) contamination of measurements by noise. To this end, we designed a “base” experiment sufficient to yield an excellent solution, and then four variations on this experiment.

The base experiment (denoted "*i*" in Fig. 2) possesses the following features: a) five measured populations: detected asymptomatic As_{det}, detected mild symptomatic Sym_{det}, detected severe symptomatic Sys_{det}, Recovered R, and Dead D; b) a temporal baseline of 101 days, beginning on 2020 February 26; c) no noise in measurements.

The three variations on this basic experiment (denoted "*ii*" through "*iv*" in Fig. 2), incorporate the following independent changes. In Experiment *ii*, the R population is not measured - an example designed to reflect the current situation in some localities (e.g. Refs (Li et al., 2020a; Weinberger et al., 2020)).

Experiment iii includes a ~ five percent noise level (for the form of additive noise, see Appendix C) in the simulated R data, and Experiment *iv* includes that noise level in addition to a doubled temporal baseline.

For each experiment, twenty independent calculations were initiated in parallel searches, each with a randomly-generated set of initial conditions on state variable and parameter values. For technical details of all experimental designs and implementation, see Appendix C.

5. Result

5.1. General findings

The salient results for the simulated experiments *i* through *iv* are as follows:

• (base experiment): Excellent estimate of all - measured and unmeasured - state variables, and all parameters except for $K_i(t)$ at times prior to the onset of social distancing;

• (absent a measurement of Population R): Poor estimate of all quantities;

• (~ 5% additive noise in R): Poor estimates of all quantities;

• (~ 5% additive noise in R, with a doubled baseline of 201 days): Estimates of state evolution are robust to noise, while parameter estimates are sensitive to noise.

Figures of the estimated time evolution of state variables and time-varying parameters are shown in their respective subsections, and the estimates of the static parameters are listed in Table 2.

Table 2
Estimates of static parameters f_{sympt} and f_{severe} over all simulated experiments. The established values are taken from Refs (Oran and Topol, 2020) and (Salje et al., 2020). For Experiments *i* and *iv*, the reported numbers are taken from the annealing iteration with a value of parameter β of 32 and 40, respectively: once the deterministic limit has been reached (see text). For Experiment ii, an attempt was made to retrieve parameter estimates at β = 2; that is: before the solution grows unstable exponentially (see Fig. 5). See specific subsections for details of each experiment.

Experiment	f_{sympt}	(established: 0.6)	f_{severe}	(established: 0.07)
	Mean	Variance	Mean	Variance
i	0.59	2×10^{-4}	0.07	4×10^{-6}
ii	–			
iii	–			
iv	0.39	0.8	0.19	0.2

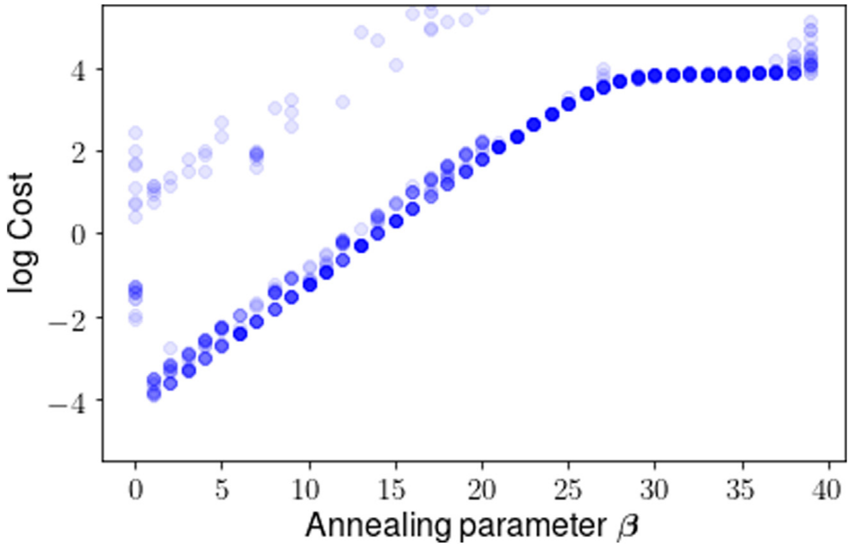


Fig. 3. Cost function plotted at each annealing step β for the base experiment i, for twenty paths in state space, where β scales the rigidity of the imposed model constraint. At low β the procedure endeavours to fit the measured variables to the simulated measurements. As β increases, the cost increases until it approaches a plateau (around β = 30), indicating that a solution has been found that is consistent with both measurements and model.

5.2. Base experiment i

The base experiment that employed five noiseless measured populations over 101 days yielded an excellent solution in terms of model evolution and parameter estimates. Prior to examining the solution, we shall first show the cost function versus the annealing parameter β, as this distribution can serve as a tool for assessing the significance of a solution.

Fig. 3 shows the evolution of the cost throughout annealing, for the ten distinct independent paths that were initiated; the x-axis shows the value of Annealing Parameter β, or: the increasing rigidity of the model constraint. At the start of iterations, the cost function is mainly fitting the measurements to data, and its value begins to climb as the model penalty is gradually imposed. If the procedure finds a solution that is consistent not only with the measurements, but also with the model, then the cost will plateau. In Fig. 4, we see this happen, around β = 30, with some scatter across paths. The reported estimates in this Subsection are taken at a value of β of 32: on the plateau. The significance of this plateau will become clearer upon examining the contrasting case of Experiment ii.

We now examine the state and parameter estimates for the base experiment *i*. For all experiments, each solution shown is representative of the solution for all twenty paths. Fig. 4 shows an excellent estimate of all state variables during the temporal window in which the measured variables were sampled. For consistency in illustrating the time evolution of all state variables, we use the state estimates for the Recovered (R) and Dead (D) populations, which are cumulative, rather than follow standard epidemiological practice of showing incident R or D. The time-varying parameters are also estimated well, excepting $K_i(t)$ at times prior to its steep decline. We noted no improvement in this estimate for $K_i(t)$, following a tenfold increase in the temporal resolution of measurements (not shown). The procedure does appear to recognize that a fast transition in the value of K_i occurred at early times, and that value was previously higher. It will be important to investigate the reason for this failure in the estimation of K_i at early times, to rule out numerical issues involved with the quickly-changing derivative[4].

5.3. Experiment ii: no measurement of R

Fig. 5 shows the cost as a function of annealing for the case with no measurement of Recovered Population R. Without examining the estimates, we know from the Cost(β) plot that no solution has been found that is consistent with both measurements and model: no plateau is reached. Rather, as the model constraint strengthens, the cost increases exponentially.

Indeed, Fig. 6 shows the estimation, taken at β = 2, prior to the runaway behavior. Note the excellent fit to the measured states and simultaneous poor fit to the unmeasured states. As no stable solution is found at high β, we conclude that there exists insufficient information in As_{det}, Sym_{det}, Sys_{det}, and D alone to corral the procedure into a region of state-and-parameter space in which a model solution is possible. We repeated this experiment with a doubled baseline of 201 days, and noted no improvement (not shown).

[4] As noted in Experiments, we chose K_i to reflect a rapid adherence to social distancing at Day 25 following time t_0, which then remained in place through to Day 101. For the form of K_i, see Appendix B.)

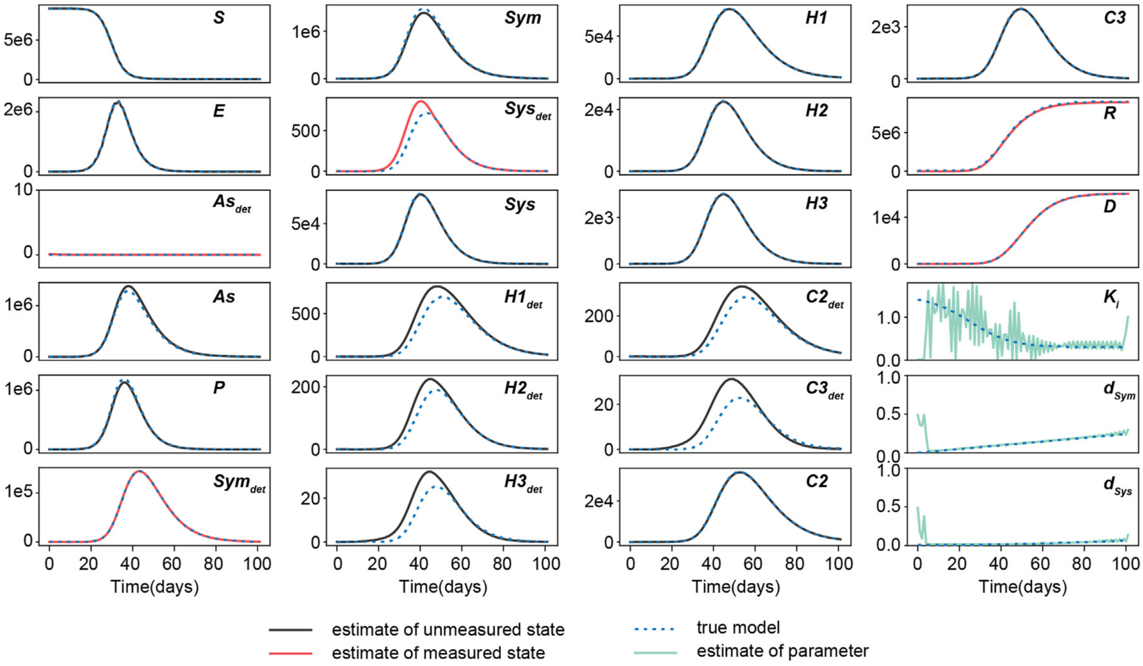


Fig. 4. Estimates of the state - measured and unmeasured - variables, and the time-varying parameters K_i, d_{Sym}, and d_{Sys}, for the base experiment *i*. Excellent estimates are obtained of all states and parameters, except early values of K_i prior to the implementation of social distancing; see text. The dotted blue lines are the simulated data. Solid red, black, and green lines are SDA estimates of measured variables, unmeasured variables, and parameters, respectively. These conventions also hold for Fig. 6 and 7. Results are taken at a value for annealing parameter β of 32.

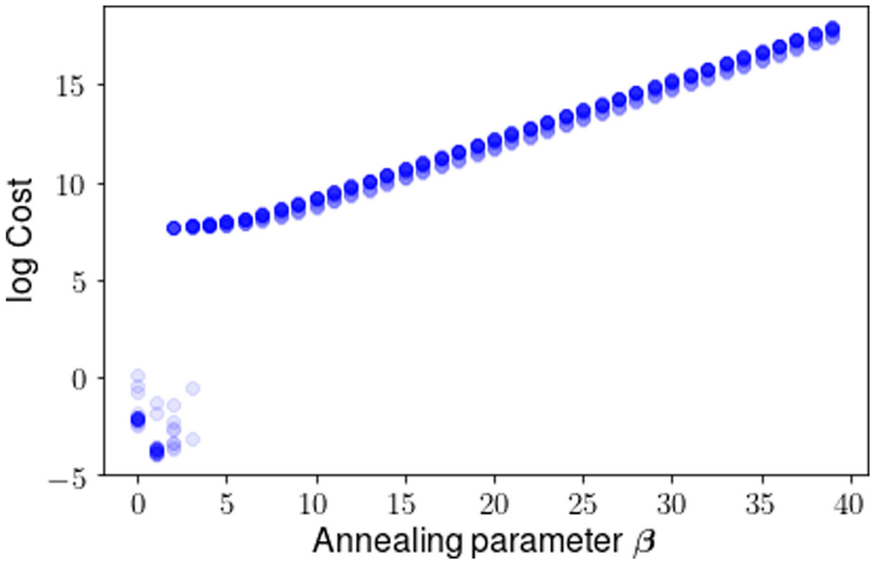


Fig. 5. Cost versus β for Experiment ii: R is not measured. As β increases, the cost increases indefinitely, indicating that no solution has been found that is consistent with both measurements and model dynamics.

5.4. Experiments iii and iv: low noise added

In Experiment *ii*, the low noise added to R yielded a poor state and parameter estimate (not shown). With a doubled temporal baseline of measurements (Experiment *iv*), however, the state estimate became robust to the contamination. Fig. 7 shows this estimate. While the ~ five percent noise added to Population R propagates to the unmeasured States S, E, and P, the general state evolution is still captured well. Importantly, the populations entering the hospital are well estimated. Note that

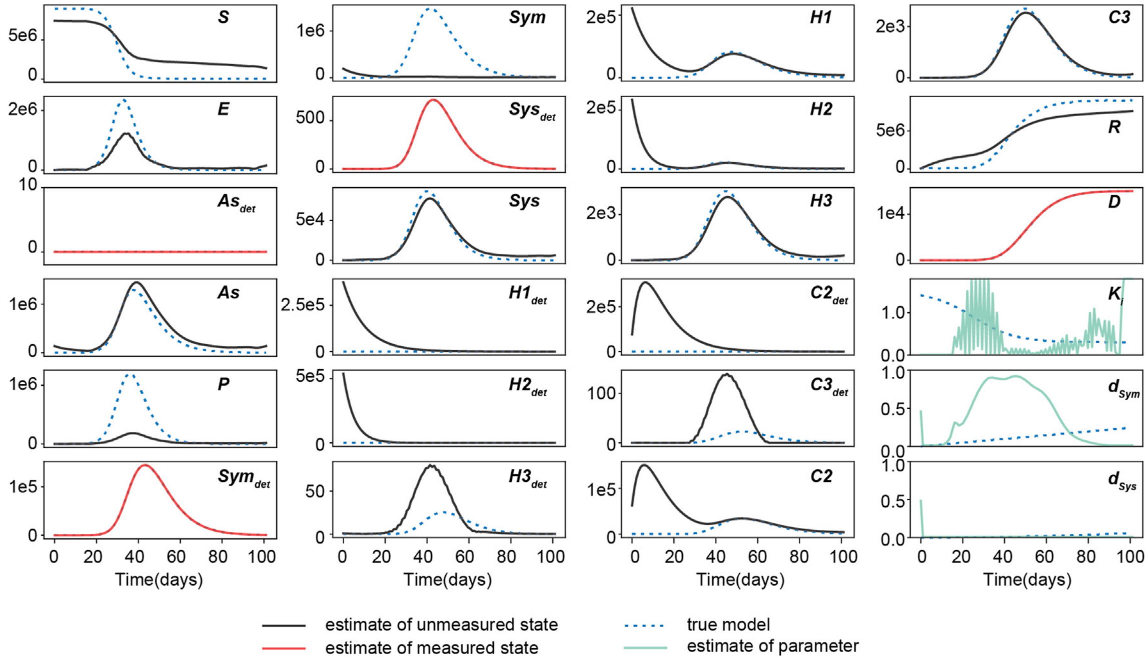


Fig. 6. Estimates for Experiment *ii*: without a measurement of Population R. This result is taken at β = 2, prior to the exponential runaway in the cost. Estimates of unmeasured states and time-varying parameters are poor.

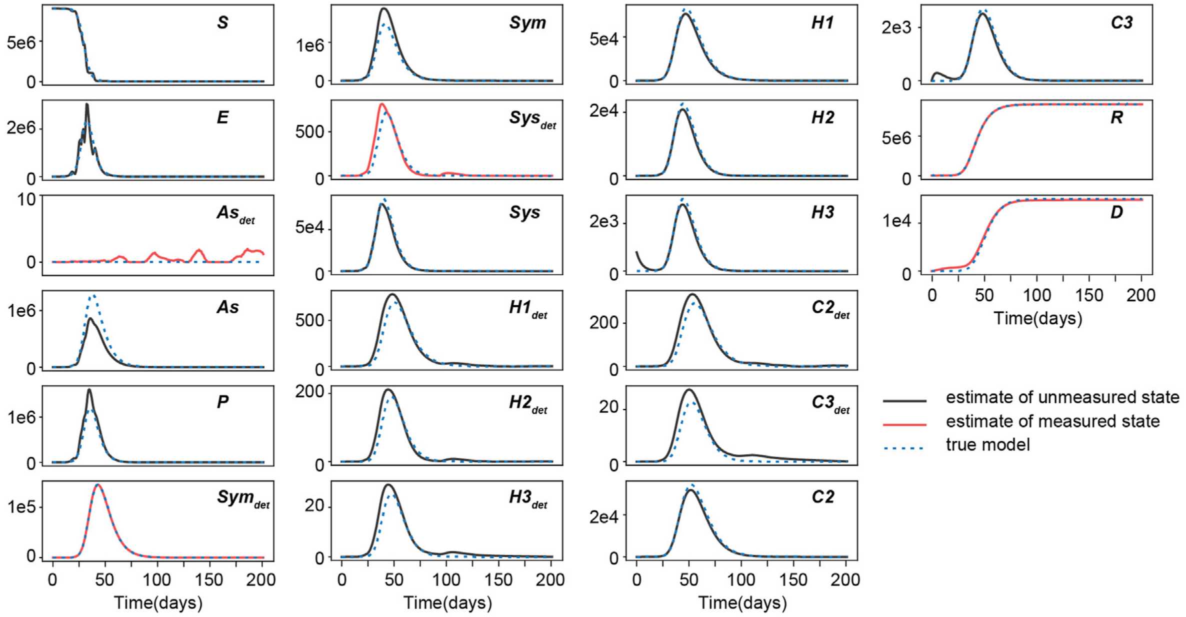


Fig. 7. Estimates for Experiment *iv*: low noise added to Population R and with a doubled temporal baseline of 201 days. The noise added to R propagates to some unmeasured States (S, E, As, and As_{det}), but the overall evolution is captured well. The noise precludes an estimate of the time-varying parameters (not shown). Results are reported using a value for β of 40.

some low state estimates (e.g. As) are not perfectly offset by high estimates (e.g. Sym). The addition of noise in these numbers - by definition - breaks the conservation of the population. Finally, the parameter estimates for Experiment *iv* do not survive the added contamination (not shown).

6. Conclusion

We have endeavoured to illustrate the potential of SDA to systematically identify the specific measurements, temporal baseline of measurements, and degree of measurement accuracy, required to estimate unknown model parameters in a high-dimensional model designed to examine the complex problems that COVID-19 presents to hospitals. In light of our assumed knowledge of some model parameters, we restrict our conclusions to general comments. We emphasize that estimation of the full model state requires measurements of the detected cases but not the undetected, provided that the recovered and dead are also measured. The state evolution is tolerant to low noise in these measurements, while the parameter estimates are not.

The ultimate aim of SDA is to test the validity of model estimation using real data, via prediction. In advance of that step, we are performing a detailed study of the model's sensitivity to contamination in the measurable populations As_{det}, Sym_{det}, Sys_{det}, R, and D. Concurrently we are examining means to render the parameter estimation less sensitive to noise, via various additional equality constraints in the cost function, and loosening the assumption of Gaussian-distributed noise. In particular, we shall require that the time-varying parameters be smoothly-varying. It will be important to examine the stability of the SDA procedure over a range of choices for parameter values and initial numbers for the infected populations.

This procedure can be expanded in many directions. Currently we are working to divide the model subpopulations by age, and to include age-specific parameters such as susceptibility and the likelihood of requiring hospitalization and intensive care. Specifically, SDA might inform the question of whether the contact matrices among age groups are non-stationary - a question of high interest for predicting age-dependent susceptibility during a second wave (ABC News, 2020) of the virus.

Other avenues for expansion are as follows: 1) define additional model parameters as unknowns to be estimated, including the fraction of patients hospitalized, the fraction who enter critical care, and the various timescales governing the reaction equations; 2) impose various constraints regarding the unknown time-varying quantities, particularly transmission rate $K_i(t)$, and identifying which forms permit a solution consistent with measurements; 3) examine model sensitivity to the initial numbers within each population; 4) examine model sensitivity to the temporal frequency of data sampling. Moreover, it is our hope that the procedure described in this paper can guide the application of SDA to a host of complicated questions surrounding COVID-19.

Acknowledgements

Thank you to Patrick Clay from the University of Michigan for discussions on inferring exposure rates given social distancing protocols.

Appendix A. Details of the model

Table 3
State variables of the COVID-19 transmission model. The "detected" qualifier signifies that the population has been tested and is positive for COVID-19.

Variable	Description
S	Susceptible
E	Exposed
As_{det}	Asymptomatic, detected
As	Asymptomatic, undetected
Sym_{det}	Symptomatic mild, detected
Sym	Symptomatic mild, undetected
Sys_{det}	Symptomatic severe, detected
Sys	Symptomatic severe, undetected
$H_{1,det}$	Hospitalized and will recover, detected
$H_{2,det}$	Hospitalized and will go to critical care and recover, detected
$H_{3,det}$	Hospitalized and will go to critical care and die, detected
H_1	Hospitalized and will recover, undetected
H_2	Hospitalized and will go to critical care and recover, undetected
H_3	Hospitalized and will go to critical care and die, undetected
$C_{2,det}$	In critical care and will recover, detected
$C_{3,det}$	In critical care and will die, detected
C_2	In critical care and will recover, undetected
C_3	In critical care and will die, undetected
R	Recovered
D	Dead

Model equations of motion

The blue notation specified by overbrackets denotes the correspondence of specific terms to the reactions between the populations depicted in Fig. 1.

$$\frac{\mathrm{d}S}{\mathrm{d}t} = -\frac{\overbrace{K_i \bullet S \bullet [\mathit{infectious} + (\mathit{infectious}_{det} \times \mathit{reduced})]}^{\to E}}{N}$$

$$\bullet \mathit{infectious} = As + P + Sym + Sys + H_1 + H_2 + H_3 + C_2 + C_3$$

$$\bullet \mathit{infectious}_{det} = As_{det} + Sym_{det} + Sys_{det}$$

$$\frac{\mathrm{d}E}{\mathrm{d}t} = \overbrace{K_i \bullet S \bullet [\mathit{infectious} + (\mathit{infectious}_{det} \times \mathit{reduced})]/N}^{S\to} - \overbrace{\frac{1-f_{sympt}}{t_{infection}} \bullet E \bullet d_{As}}^{\to AS_{det}} - \overbrace{\frac{1-f_{sympt}}{t_{infection}} \bullet E \bullet (1.0 - d_{As})}^{\to As} - \overbrace{\frac{f_{sympt}}{t_{infection}} \bullet E}^{\to P}$$

$$\frac{\mathrm{d}As_{det}}{\mathrm{d}t} = \overbrace{\frac{1-f_{sympt}}{t_{infection}} \bullet E \bullet d_{As}}^{E\to} - \overbrace{\frac{1}{t_{R,a}} \bullet As_{det}}^{\to R}$$

$$\frac{\mathrm{d}As}{\mathrm{d}t} = \overbrace{\frac{1-f_{sympt}}{t_{infection}} \bullet E \bullet (1.0 - d_{As})}^{E\to} - \overbrace{\frac{1}{t_{R,a}} \bullet As}^{\to R}$$

$$\frac{\mathrm{d}P}{\mathrm{d}t} = \overbrace{\frac{f_{sympt}}{t_{infection}} \bullet E}^{E\to} - \overbrace{\frac{1-f_{severe}}{t_{sympt}} \bullet P \bullet d_{Sym}}^{\to Sym_{det}} - \overbrace{\frac{1-f_{severe}}{t_{sympt}} \bullet P \bullet (1.0 - d_{Sym})}^{\to Sym} - \overbrace{\frac{f_{severe}}{t_{sympt}} \bullet P \bullet d_{Sys}}^{\to Sys_{det}} - \overbrace{\frac{f_{severe}}{t_{sympt}} \bullet P \bullet (1.0 - d_{Sys})}^{\to Sys}$$

$$\frac{\mathrm{d}Sym_{det}}{\mathrm{d}t} = \overbrace{\frac{1-f_{severe}}{t_{sympt}} \bullet P \bullet d_{Sym}}^{P\to} - \overbrace{\frac{1}{t_{R,m}} \bullet Sym_{det}}^{\to R}$$

$$\frac{\mathrm{d}Sym}{\mathrm{d}t} = \overbrace{\frac{1-f_{severe}}{t_{sympt}} \bullet P \bullet (1.0 - d_{Sym})}^{P\to} - \overbrace{\frac{1}{t_{R,m}} \bullet Sym}^{\to R}$$

$$\frac{\mathrm{d}Sys_{det}}{\mathrm{d}t} = \overbrace{\frac{f_{severe}}{t_{sympt}} \bullet P \bullet d_{Sys}}^{P\to} - \overbrace{\frac{f_H}{t_H} \bullet Sys_{det}}^{\to H_{1,det}} - \overbrace{\frac{f_C}{t_H} \bullet Sys_{det}}^{\to H_{2,det}} - \overbrace{\frac{f_D}{t_H} \bullet Sys_{det}}^{\to H_{3,det}}$$

$$\frac{\mathrm{d}Sys}{\mathrm{d}t} = \overbrace{\frac{f_{severe}}{t_{sympt}} \bullet P \bullet (1.0 - d_{Sys})}^{P\to} - \overbrace{\frac{f_H}{t_H} \bullet Sys}^{\to H_1} - \overbrace{\frac{f_C}{t_H} \bullet Sys}^{\to H_2} - \overbrace{\frac{f_D}{t_H} \bullet Sys}^{\to H_3}$$

$$\frac{\mathrm{d}H_{1,det}}{\mathrm{d}t} = \overbrace{\frac{f_H}{t_H} \bullet Sys_{det}}^{Sys_{det}\to} - \overbrace{\frac{1}{t_{R,h}} \bullet H_{1,det}}^{\to R}$$

$$\frac{\mathrm{d}H_{2,det}}{\mathrm{d}t} = \overbrace{\frac{f_C}{t_H} \bullet Sys_{det}}^{Sys_{det}\to} - \overbrace{\frac{1}{t_C} \bullet H_{2,det}}^{\to C_{2,det}}$$

$$\frac{\mathrm{d}H_{3,det}}{\mathrm{d}t} = \overbrace{\frac{f_D}{t_H} \bullet Sys_{det}}^{Sys_{det}\to} - \overbrace{\frac{1}{t_C} \bullet H_{3,det}}^{\to C_{3,det}}$$

$$\frac{dH_1}{dt}=\overbrace{\frac{f_H}{t_H}\cdot Sys}^{Sys\rightarrow}-\overbrace{\frac{1}{t_{R,h}}\cdot H_1}^{\rightarrow R}$$

$$\frac{dH_2}{dt}=\overbrace{\frac{f_C}{t_H}\cdot Sys}^{Sys\rightarrow}-\overbrace{\frac{1}{t_C}\cdot H_2}^{\rightarrow C_2}$$

$$\frac{dH_3}{dt}=\overbrace{\frac{f_D}{t_H}\cdot Sys}^{Sys\rightarrow}-\overbrace{\frac{1}{t_C}\cdot H_3}^{\rightarrow C_3}$$

$$\frac{dC_{2,det}}{dt}=\overbrace{\frac{1}{t_C}\cdot H_{2,det}}^{H_{2,det}\rightarrow}-\overbrace{\frac{1}{t_{R,c}}\cdot C_{2det}}^{\rightarrow R}$$

$$\frac{dC_{3,det}}{dt}=\overbrace{\frac{1}{t_C}\cdot H_{3,det}}^{H_{3,det}\rightarrow}-\overbrace{\frac{1}{t_D}\cdot C_{3,det}}^{\rightarrow D}$$

$$\frac{dC_2}{dt}=\overbrace{\frac{1}{t_C}\cdot H_2}^{H_2\rightarrow}-\overbrace{\frac{1}{t_{R,c}}\cdot C_2}^{\rightarrow R}$$

$$\frac{dC_3}{dt}=\overbrace{\frac{1}{t_C}\cdot H_3}^{H_3\rightarrow}-\overbrace{\frac{1}{t_D}\cdot C_3}^{\rightarrow D}$$

$$\frac{dR}{dt}=\overbrace{\frac{1}{t_{R,a}}\cdot As_{det}}^{AS_{det}\rightarrow}+\overbrace{\frac{1}{t_{R,a}}\cdot As}^{As\rightarrow}+\overbrace{\frac{1}{t_{R,m}}\cdot Sym_{det}}^{Sym_{det}\rightarrow}+\overbrace{\frac{1}{t_{R,m}}\cdot Sym}^{Sym\rightarrow}+\overbrace{\frac{1}{t_{R,h}}\cdot H_{1,det}}^{H_{1,det}\rightarrow}+\overbrace{\frac{1}{t_{R,h}}\cdot H_1}^{H_1\rightarrow}+\overbrace{\frac{1}{t_{R,c}}\cdot C_{2,det}}^{C_{2,det}\rightarrow}+\overbrace{\frac{1}{t_{R,c}}\cdot C_2}^{C_2\rightarrow}$$

$$\frac{dD}{dt}=\overbrace{\frac{1}{t_D}\cdot C_{3,det}}^{C_{3,det}\rightarrow}+\overbrace{\frac{1}{t_D}\cdot C_3}^{C_3\rightarrow}$$

Table 4

The model parameters, with the unknown parameters to be estimated denoted in boldface. The unknown parameters K_i, Sym, and d_{Sys} are taken to be time-varying. The unknown parameters fsympt and fsevere are taken to be intrinsic properties of the disease and therefore constant numbers. The detection probability of asymptomatic cases is taken to be known and zero. Units of time are days.

Parameter	Description	Value
N	Total population	9,000,000
Reduced	The property that a detected case is likely to transmit less, via successful quarantine)	0.2
$K_i(t)$	Transmission rate	See Appendix B
$d_{As}(t)$	Detection probability of asymptomatic cases	0.0
f_{sympt}	Fraction of positive cases that produce symptoms	0.6 (Oran and Topol, 2020)
$t_{infection}$	Time from exposure to infection	4.0 (Li et al., 2020c)
$t_{R,a}$	Time to recovery for asymptomatics	8.0 Assumed to be same as tR,m
$d_{Sym}(t)$	Detection probability of mild symptomatics	See Appendix B
$d_{Sys}(t)$	Detection probability of severe symptomatics	See Appendix B
f_{severe}	Fraction of symptomatics that are severe	0.07 (Salje et al., 2020)
t_{sympt}	Time to symptoms, for symptomatics	4.0 (Roman et al., 2020; Jing et al., 2020)
$t_{R,m}$	Time from symptoms to recovery, for mild symptomatics	8.0 (Roman et al., 2020)a
f_H	Fraction of severe cases that are hospitalized and then recover: $f_H = 1.0 - f_C - f_D$	0.66
f_C	Fraction of severe cases that require critical care and then recover	0.3 (Lewnard et al., 2020)
f_D	Fraction of severe cases that die	0.04 (Wang et al., 2019)
t_H	Time from symptoms to hospital, for severe symptomatics	5.0 (Huang et al., 2020)
$t_{R,h}$	Time from entering hospital to recovery, for severe symptomatics that do not require critical care	10.0 (Lewnard et al., 2020; Wang et al., 2019)

Table 4 (*continued*)

Parameter	Description	Value
t_C	Time from entering hospital to critical care, for severe symptomatics	5.0 (Huang et al., 2020)
$t_{R,c}$	Time from entering critical care to recovery for severe symptomatics	10.0 (Bi et al., 2020)
t_D	Time from entering critical care to death, for severe symptomatics	5.0 (Yang et al., 2020)

[a] As described in (Roman et al., 2020), viral load can be high and detectable for up to 20 days. We choose a shorter duration of infectiousness to capture the time during which transmissibility is highest.

Appendix B. Unknown time-varying parameters to be estimated

The unknown parameters assumed to be time-varying are the transmission rate K_i, and the detection probabilities d_{Sym} and d_{Sys} for mild and severe symptomatic cases, respectively.

The transmission rate in a given population for a given infectious disease is measured in effective contacts per unit time. This may be expressed as the total contact rate (the total number of contacts, effective or not, per unit time), multiplied by the risk of infection, given contact between an infectious and a susceptible individual. The total contact rate can be impacted by social behavior.

In this first employment of SDA upon a pandemic model of such high dimensionality, we chose to represent K_i as a relatively constant value that undergoes one rapid transition corresponding to a single social distancing mandate. As noted in Experiments, social distancing rules were imposed in New York City roughly 25 days following the first reported case. We thus chose K_i to transition between two relatively constant levels, roughly 25 days following time t0. Specifically, we wrote $K_i(t)$ (Tang et al., 2020) as:

$$K_i(t) = -f \cdot \frac{1}{e^{(T-t)/s} + 1} + \xi.$$

The parameter T was set to 25, beginning four days prior to the first report of a detection in NYC (NY Times, 2020) to the imposition of a stay-home order in NYC on March 22 (NY Governor's Office, 2020). The parameter s governs the steepness of the transformation, and was set to 10. Parameters f and ξ were then adjusted to 1.2 and 1.5, to achieve a transition from about 1.4 to 0.3. For detection probabilities d_{Sym} and d_{Sys}, a linear and quadratic form, respectively, were chosen to preclude symmetries, and both were optimistically taken to increase with time:

$$\begin{aligned} d_{Sym}(t) &= 0.2 \cdot t \\ d_{Sys}(t) &= 0.1 \cdot t^2 \end{aligned}$$

Finally, each time series was normalized to the range: [0:1], via division by their respective maximum values.

Appendix C. Technical details of the inference experiments

The simulated data were generated by integrating the reaction equations (Appendix A) via a fourth-order adaptive Runge-Kutta method encoded in the Python package odeINT. A step size of one (day) was used to record the output. Except for the one instance noted in *Results* regarding Experiment i, we did not examine the sensitivity of estimations to the temporal sparsity of measurements. The initial conditions on the populations were: $S_0 = N - 1$ (where N is the total population), $As_0 = 1$, and zero for all others.

For the noise experiments, the noise added to the simulated Symdet, Sysdet, and R data were generated by Python's numpy. random.normal package, which defines a normal distribution of noise. For the "low-noise" experiments, we set the standard deviation to be the respective mean of each distribution, divided by 100. For the experiments using higher noise, we multiplied that original level by a factor of ten. For each noisy data set, the absolute value of the minimum was then added to each data point, so that the population did not drop below zero

The optimization was performed via the open-source Interior-point Optimizer (Ipopt) (Wächter, 2009). Ipopt uses a Simpson's rule method of finite differences to discretize the state space, a Newton's method to search, and a barrier method to impose user-defined bounds that are placed upon the searches. We note that Ipopt's search algorithm treats state variables as independent quantities, which is not the case for a model involving a closed population. This feature did not affect the results of this paper. Those interested in expanding the use of this tool, however, might keep in mind this feature. One might negate undesired effects by, for example, imposing equality constraints into the cost function that enforce the conservation of N.

Within the annealing procedure described in Methods, the parameter α was set to 2.0, and β ran from 0 to 38 in increments of 1. The inverse covariance matrix for measurement error (R_m) was set to 1.0, and the initial value of the inverse covariance matrix for model error ($R_{f,0}$) was set to 10^{-7}.

For each of the four simulated experiments, twenty paths were searched, beginning at randomly-generated initial conditions for parameters and state variables. All simulations were run on a 720-core, 1440-GB, 64-bit CPU cluster.

References

Abarbanel, H. (2013). *Predicting the future: Completing models of observed complex systems*. Springer.

Abarbanel, H. D. I., Bryant, P., Gill, P. E., Kostuk, M., Rofeh, J., Singer, Z., Toth, B., Wong, E., & Ding, M. (2011). *Dynamical parameter and state estimation in neuron models, the dynamic brain: An exploration of neuronal variability and its functional significance*.

An, Z., Rey, D., Ye, J., & Abarbanel, H. D. I. (2017). Estimating the state of a geophysical system with sparse observations: Time delay methods to achieve accurate initial states for prediction. *Nonlinear Processes in Geophysics, 24*(1).

Armstrong, E. (2020). Statistical data assimilation for estimating electrophysiology simultaneously with connectivity within a biological neuronal network. *Physical Review E, 101*(1), Article 012415.

Armstrong, E., Amol, V. P., Johns, L., Kishimoto, C. T., Abarbanel, H. D. I., & Fuller, G. M. (2017). An optimization-based approach to calculating neutrino flavor evolution. *Physical Review D, 96*(8), Article 083008.

As Americans Brace for Second Wave of COVID-19 … https://abcnews.go.com/Health/americans-brace-2nd-wave-covid-19-experts-predict/story?id=72817318 Accessed: 2020-10-26.

Bettencourt, L. M. A., & Ribeiro, R. M. (2008). Real time bayesian estimation of the epidemic potential of emerging infectious diseases. *PloS One, 3*(5).

Bettencourt, L. M. A., Ribeiro, R. M., Chowell, G., Lant, T., & Castillo-Chavez, C. (2007). Towards real time epidemiology: Data assimilation, modeling and anomaly detection of health surveillance data streams. *NSF workshop on intelligence and security informatics* (pp. 79–90). Springer.

Betts, J. T. (2010). *Practical methods for optimal control and estimation using nonlinear programming, ume 19*. Siam.

Bi, Q., Wu, Y., Mei, S., Ye, C., Zou, X., Zhang, Z., Liu, X., Lan, W., Truelove, S. A., Zhang, T., et al. (2020). *Epidemiology and transmission of covid-19 in shenzhen China: Analysis of 391 cases and 1,286 of their close contacts*. MedRxiv.

Cobb, L., Krishnamurthy, A., Mandel, J., & Beezley, J. D. (2014). Bayesian tracking of emerging epidemics using ensemble optimal statistical interpolation. *Spatial and spatio-temporal epidemiology, 10*, 39–48.

Cori, A., Ferguson, N. M., Fraser, C., & Simon, C. (2013). A new framework and software to estimate time-varying reproduction numbers during epidemics. *American Journal of Epidemiology, 178*(9), 1505–1512.

Daniel P. Oran and eric J. Topol. Getting a handle on asymptomatic SARS-CoV-2 infection. https://www.scripps.edu/science-and-medicine/translational-institute/about/news/sarc-cov-2-infection/. Accessed: 2020-05-24.

Evensen, G. (2009). *Data assimilation: The ensemble kalman filter*. Springer Science & Business Media.

First reported confirmation of coronavirus in New York City.(2020). https://www.nytimes.com/2020/03/01/nyregion/new-york-coronvirus-confirmed.html Accessed: 2020-05-19.

Hamilton, F., Berry, T., Peixoto, N., & Sauer, T. (2013). Real-time tracking of neuronal network structure using data assimilation. *Physical Review E, 88*(5), Article 052715.

M. Heggeness. The need for data innovation in the time of covid-19. https://www.minneapolisfed.org/article/2020/the-need-for-data-innovation-in-the-time-of-covid-19. Accessed: 2020-05-17.

Huang, C., Wang, Y., Li, X., Ren, L., Zhao, J., Hu, Y., Zhang, L., Fan, G., Xu, J., Gu, X., et al. (2020). Clinical features of patients infected with 2019 novel coronavirus in wuhan, China. *The lancet, 395*(10223), 497–506.

IHME COVID-19 health service utilization forecasting Team and Christopher JL Murray. (2020). *Forecasting COVID-19 impact on hospital bed-days, ICU-days, ventilator-days and deaths by US state in the next 4 months*. medRxiv: Cold Spring Harbor Laboratory Press, 03.27.20043752, March 2020.

Jörn Lothar Sesterhenn. (2020). *Adjoint-based data assimilation of an epidemiology model for the covid-19 pandemic in 2020*. arXiv preprint arXiv:2003.13071.

Jing, Q., You, C., Lin, Q., Hu, T., Yu, S., & Zhou, X.-H. (2020). *Estimation of incubation period distribution of COVID-19 using disease onset forward time: A novel cross-sectional and forward follow-up study. medRxiv*. Publisher: Cold Spring Harbor Laboratory Press, 03.06.20032417, March 2020.

Kadakia, N., Armstrong, E., Breen, D., Morone, U., Daou, A., Margoliash, D., & Abarbanel, H. D. I. (2016). Nonlinear statistical data assimilation for HVCRA neurons in the avian song system. *Biological Cybernetics, 110*(6), 417–434.

Kalnay, E. (2003). *Atmospheric modeling, data assimilation and predictability*. Cambridge university press.

Kimura, R. (2002). Numerical weather prediction. *Journal of Wind Engineering and Industrial Aerodynamics, 90*(12–15), 1403–1414.

Kostuk, M., Toth, B. A., Daniel Meliza, C., Margoliash, D., & Abarbanel, H. D. I. (2012). Dynamical estimation of neuron and network properties ii: Path integral Monte Carlo methods. *Biological Cybernetics, 106*(3), 155–167.

Lewnard, J. A., Liu, V. X., Jackson, M. L., Schmidt, M. A., Jewell, B. L., Flores, J. P., Jentz, C., Northrup, G. R., Mahmud, A., Reingold, A. L., et al. (2020). *Incidence, clinical outcomes, and transmission dynamics of hospitalized 2019 coronavirus disease among 9,596,321 individuals residing in California and Washington, United States: A prospective cohort study*. medRxiv.

Li, R., Pei, S., Chen, B., Song, Y., Zhang, T., Yang, W., & Shaman, J. (2020a). Substantial undocumented infection facilitates the rapid dissemination of novel coronavirus (sars-cov-2). *Science, 368*(6490), 489–493.

Li, R., Pei, S., Chen, B., Song, Y., Zhang, T., Yang, W., & Shaman, J. (May 2020). Substantial undocumented infection facilitates the rapid dissemination of novel coronavirus (SARS-CoV-2). *Science, 368*(6490), 489–493.

Li, Y. I., Turk, G., Rohrbach, P. B., Pietzonka, P., Kappler, J., Singh, R., Dolezal, J., Ekeh, T., Kikuchi, L., Peterson, J. D., et al. (2020b). *Efficient bayesian inference of fully stochastic epidemiological models with applications to covid-19*. arXiv preprint arXiv:2010.11783.

Meliza, C. D., Kostuk, M., Huang, H., Nogaret, A., Margoliash, D., & Abarbanel, Henry DI. (2014). Estimating parameters and predicting membrane voltages with conductance-based neuron models. *Biological Cybernetics, 108*(4), 495–516.

Nadler, P., Wang, S., Arcucci, R., Yang, X., & Guo, Y. (2020). *An epidemiological modelling approach for covid 19 via data assimilation*. arXiv preprint arXiv:2004.12130.

Nogaret, A., Daniel Meliza, C., Margoliash, D., & Abarbanel, H. D. I. (2016). Automatic construction of predictive neuron models through large scale assimilation of electrophysiological data. *Scientific Reports, 6*(1), 1–14.

PAUSE order in New York City takes effect 2020 March 22. https://www.governor.ny.gov/news/governor-cuomo-signs-new-york-state-pause-executive-order Accessed: 2020-05-19.

Rey, D., Eldridge, M., Kostuk, M., Abarbanel, H. D. I., Schumann-Bischoff, J., & Ulrich, P. (2014). Accurate state and parameter estimation in nonlinear systems with sparse observations. *Physics Letters A, 378*(11–12), 869–873.

Rhodes, C. J., & Hollingsworth, T. D. (2009). Variational data assimilation with epidemic models. *Journal of Theoretical Biology, 258*(4), 591–602.

Roman, W., Corman, V. M., Guggemos, W., Michael Seilmaier, Zange, S., Müller, M. A., Niemeyer, D., Jones, T. C., Vollmar, P., Rothe, C., Hoelscher, M., & Tobias, B. (May 2020). Sebastian brünink, julia schneider, rosina ehmann, katrin zwirglmaier, christian drosten, and clemens wendtner. Virological assessment of hospitalized patients with COVID-2019. *Nature, 581*(7809), 465–469.

Salje, H., Tran Kiem, C., Lefrancq, N., Courtejoie, N., Bosetti, P., Paireau, J., , … Claire-Lise Dubost, et al. (2020). Estimating the burden of sars-cov-2 in France. *Science, 369*(6500), 208–211. https://doi.org/10.1126/science.abc3517

Schiff, S. J. (2009). Kalman meets neuron: The emerging intersection of control theory with neuroscience. *2009 annual international conference of the IEEE engineering in medicine and biology society* (pp. 3318–3321). IEEE.

Tang, B., Nicola Luigi Bragazzi, Qian, L., Tang, S., Xiao, Y., & Wu, J. (2020). An updated estimation of the risk of transmission of the novel coronavirus (2019-ncov). *Infectious disease modelling, 5*, 248–255.

Tarantola, A. (2005). *Inverse problem theory and methods for model parameter estimation*. SIAM.

Thompson, R. N., Stockwin, J. E., van Gaalen, R. D., Polonsky, J. A., Kamvar, Z. N., Demarsh, P. A., Dahlqwist, E., Li, S., Miguel, E., Jombart, T., et al. (2019). Improved inference of time-varying reproduction numbers during infectious disease outbreaks. *Epidemics, 29*, 100356.

Toth, B. A., Kostuk, M., Daniel Meliza, C., Margoliash, D., & Abarbanel, H. D. I. (2011). Dynamical estimation of neuron and network properties i: Variational methods. *Biological Cybernetics, 105*(3–4), 217–237.

Wächter, A. (2009). Short tutorial: Getting started with ipopt in 90 minutes. In *Dagstuhl Seminar Proceedings. Schloss Dagstuhl-Leibniz-Zentrum für Informatik*.

Wallinga, J., & Teunis, P. (2004). Different epidemic curves for severe acute respiratory syndrome reveal similar impacts of control measures. *American Journal of Epidemiology, 160*(6), 509–516.

D. Wang, B. Hu, C. Hu, F. Zhu, X. Liu, J. Zhang, B. Wang, H. Xiang, Z. Cheng, Y. Xiong, Y. Zhao, Y. Li, X. Wang, and Z. Peng. Clinical characteristics of 138 hospitalized patients With 2019 novel coronavirus–infected pneumonia in Wuhan, China. JAMA, 323(11):1061–1069, March 2020. Publisher: American Medical Association.

Wang, J., Breen, D., Abraham, A., Abarbanel, H. D. I., & Cauwenberghs, G. (2016). Data assimilation of membrane dynamics and channel kinetics with a neuromorphic integrated circuit. *Biomedical circuits and systems conference (BioCAS), 2016 IEEE* (pp. 584–587). IEEE.

Weinberger, D., Cohen, T., Crawford, F., Mostashari, F., Olson, D., Pitzer, V. E., Reich, N. G., Russi, M., Simonsen, L., Watkins, A., et al. (2020). *Estimating the early death toll of covid-19 in the United States*. Medrxiv.

Whartenby, W. G., Quinn, J. C., & Abarbanel, H. D. I. (2013). The number of required observations in data assimilation for a shallow-water flow. *Monthly Weather Review, 141*(7), 2502–2518.

Yang, X., Yuan, Y., Xu, J., Shu, H., Liu, H., Wu, Y., Zhang, L., Yu, Z., Fang, M., Ting, Y., et al. (2020). *Clinical course and outcomes of critically ill patients with sars-cov-2 pneumonia in wuhan, China: A single-centered, retrospective, observational study*. The Lancet Respiratory Medicine.

Ye, J., Rey, D., Kadakia, N., Eldridge, M., Morone, U. I., Paul, R., & Quinn, J. C. (2015). Systematic variational method for statistical nonlinear state and parameter estimation. *Physical Review E, 92*(5), Article 052901.

Ye, J., Kadakia, N., Rozdeba, P. J., Abarbanel, H. D. I., & Quinn, J. C. (2015). Improved variational methods in statistical data assimilation. *Nonlinear Processes in Geophysics, 22*(2), 205–213.

Ye, J., Rozdeba, P. J., Morone, U. I., Daou, A., & Abarbanel, H. D. I. (2014). Estimating the biophysical properties of neurons with intracellular calcium dynamics. *Physical Review E, 89*(6), Article 062714.