Accurate Calibration of Agent-based Epidemiological Models with Neural Network Surrogates

Rushil Anirudh [1] Jayaraman J. Thiagarajan [1] Peer-Timo Bremer [1] Timothy Germann [2] Sara Del Valle [2]
Frederick Streitz [1]

Abstract

Calibrating complex epidemiological models to observed data is a crucial step to provide both insights into the current disease dynamics, i.e. by estimating a reproductive number, as well as to provide reliable forecasts and scenario explorations. Here we present a new approach to calibrate an agent-based model – EpiCast – using a large set of simulation ensembles for different major metropolitan areas of the United States. In particular, we propose: a new neural network based surrogate model able to simultaneously emulate all different locations; and a novel posterior estimation that provides not only more accurate posterior estimates of all parameters but enables the joint fitting of global parameters across regions.

1. Introduction

Epidemiological models are playing a key role in the national response to COVID-19: Model parameters can provide intuitive metrics on disease progression, i.e. by computing an effective reproductive number, model forecasts represent a crucial planning tool, and exploring what-if scenarios can provide insights into the relative trade-off, for example, between social distancing measures and economic impacts. However, as with all computational models, all such benefits depend on the model parameters being well calibrated to the observed data and the model being generalizable to unknown situations. Generically, this describes a standard inverse problem of fitting model parameters, under the key assumption that the epidemiological model can effectively describe the observed data. In practice, epidemiological models face a number of additional challenges that significantly complicate the problem: First, even the most complex simulations cannot hope to account for all factors impacting disease progression as it may depend on everything from the weather to shopping habits and from socioeconomic status to average household size. Consequently, there exist a number of hidden variables and effects that are not explicitly modeled. Second, certain model parameters are expected to be localized with different communities having different starting conditions, i.e. household sizes, economic status, job classifications, etc., as well as different states or even different counties having varying levels of compliance to social distancing or other non-pharmaceutical (NPI) interventions. Finally, data collection is challenging with often noisy results due to errors, reporting delays, etc., and especially the more sophisticated models are naturally under-constrained as different effects can trade-off against each other, i.e. community transmission may be similarly affected by more mask wearing vs. fewer restaurant visits.

One common approach to compensate for both these challenges and uncertainties is to use a simpler model easier to constrain and fit. In the case of epidemiological models, this typically means a population based approach like the S(E)IR model (He et al., 2020) which expresses the disease progression in terms of **S**usceptible, **E**xposed, **I**infected, and **R**ecovered compartments through a set of coupled ordinary differential equations. However, these approaches are predominantly phenomological and provide limited insights into the underlying causes. Furthermore, it is challenging to explore different NPI strategies as none of the important factors, i.e. schools, businesses, travel, etc., can be directly modeled. Hence, we focus on more detailed agent-based models (Germann et al., 2006; Halloran et al., 2008) which explicitly model individual people, their community characteristics, and interactions.

In particular, we are using the EpiCast model (Germann et al., 2006) which creates populations of agents based on census track level information and is able to directly model school and workplace closures, household sizes, etc. However, depending on the population size that is modeled and the number of compute cores available, a single simulation may take minutes to hours and thus a brute-force search to match a model to observations is infeasible. Furthermore, some parameters of the model are local, i.e. dependent

[1] Lawrence Livermore National Laboratory [2] Los Alamos National Laboratory. Correspondence to: Rushil Anirudh <anirudh1@llnl.gov>.

Proceedings of the 1^{st} Workshop on Healthcare AI and COVID-19, ICML 2022.

on the geographical region like the number of currently infected, while others are global in that they describe biological properties of the virus, such as the fraction of asymptomatic infections. Fitting the latter on a per city, county, or even state level may lead to inconsistent results as the fundamental disease parameters are unlikely to differ significantly across regions. Conceptually, the ideal solution to this problem is to construct a nation-wide simulation with consistent global but flexible local parameters. Unfortunately, not only is even a single national simulation computationally expensive but the space of all local parameters could be prohibitively large and this is not yet considering the immense complexity in setting up such a simulation.

In this paper, we consider a more practical formulation of the problem: We develop a single neural network to act as an emulator for EpiCast across all geographical regions, using which we jointly estimate local and global parameters simultaneously. We train this surrogate using simulation runs at the level of Metropolitan Statistical Areas (MSAs), which are densely populated urban centers. By jointly training across MSAs in a population-normalized space (i.e., estimating the fraction of infections with respect to the MSA population), we show that the surrogate generalizes well to previously unseen MSAs.

This is particularly useful for modeling MSAs that are infected relatively later in the pandemic and more importantly, it eliminates the need to generate exploratory simulation runs for every MSA. Since inference with a neural network is orders of magnitude cheaper than EpiCast , we are able optimize for the optimal set of input parameters required for calibration. In order to make this highly ill-posed inverse problem more tractable, we propose new regularization objectives: (a) explicitly tying the global parameters across geographical regions while optimizing for the optimal set of global and local parameters, and (b) constraining the parameter values to be close to those observed during training. We find that they are effective at constraining the solution space, ultimately yielding accurate calibration of the epidemiological model. We validate our approaches with EpiCast and show that the proposed regularizers, when coupled with the pre-trained surrogate, can accurately recover the EpiCast parameters required to match the observed data well.

2. Background

2.1. EpiCast

As discussed above, we are using the EpiCast framework as our epidemiological model. Originally, developed to understand influenza outbreaks (Germann et al., 2006) it has recently been adapted and modified to model the COVID-19 spread. EpiCast is an agent-based simulation, which explicitly models individual agents in a population, their main occupations and locations, and interactions. Based on a 12 hour time steps an agent can be at home, at work or school, or in the community representing tasks such as shopping. Randomized populations are created from census data based on census tract granularity with about 2,000 agents per tract. EpiCast considers a wide range of properties, such as age distributions, house hold sizes, occupation, commuting patterns, etc. In particular, the system allows epidemiologists to explore detailed scenarios such as different school schedules, the impact of travel restrictions, etc. However, calibrating the model to noisy data especially in the context of shifting public health orders, local variations between populations, and various other confounding factors is difficult.

EpiCast takes six parameters to produce the average number of expected infections in the regions of interest. These are: (a) INFECTED: The factor used for the starting number of infected population, (b) REMOVED: percent of the population assumed already removed (quarantined or hospitalized), (c) COMPLICANCE Fraction of people expected to comply with local guidelines, (d) TRANSPROP: transmission probability of the disease, (e) PROPASYM: Fraction of asymptomatic people in the population, and finally (f) RELINF: Relative infectiousness. The first three of these are considered local and can reasonably vary between different regions, while the last three are global as they describe biological properties of the virus.

We selected a set of self-consistent scenarios to demonstrate how modern surrogate modeling and advanced calibration provide us with highly accurate and globally consistent model parameters across different regions. In particular, we have created a large ensemble of 15 major statistical areas (MSAs), roughly corresponding to major metropolitan areas in the US, aimed to approximate their daily COVID-19 case loads as provided by the Johns Hopkins University (JHU) repository (Dong et al., 2020). Each ensemble explores the six parameters discussed above and is initialized according to the local characteristics of each MSA. For each MSA, the starting numbers are seeded in an tight interval $\pm 2\%$ of the JHU data assuming a under-reporting of cases by a factor of 3. The latter compensates for the fact that the simulation assumes perfect knowledge of all infections and is inline with best estimates for unreported cases. In separate experiments we have seen limited impact of the exact under-reporting factor and would expect identical results for other choices. All ensembles start at June 22nd and simulate four weeks of disease progression to provide a total of $7,500$ EpiCast training curves.

2.2. Calibration

Calibration (also called history matching) is the problem of matching the output of a simulation to real observed data.

This is often used to better understand the phenomenon of interest. In epidemiology, calibration allows us to determine the state of a pandemic and characterize properties of the infecting agent. Let us define the set of output curves to be $y \in \mathcal{Y} \subset \mathbb{R}^T$, where T is the number of days for which data is available. Next, the inputs are defined as $x \in \mathcal{X} \subset \mathbb{R}^d$, where d is the number of parameters of the epidemiological model; for EpiCast $d = 6$. The simulation outputs an expected number of daily infected cases over the period of interest. Now, given an observed curve $y_{obs} \subset \mathbb{R}^t, t \leq T$, the calibration problem is to find the posterior distribution $p(x|y_{obs})$ such that a sample x from this posterior matches the simulation output according to a goodness-of-fit measure (GOF) such as an ℓ_p norm. In other words, identifying a set of samples $\mathbf{x} = [x_1, x_2, \ldots, x_n], \forall i, x_i \in \mathcal{X}$ such that, $||\mathcal{E}(x_i) - y_{obs}||_p \leq \epsilon$, where $\mathcal{E}$ represents the EpiCast model and ϵ determines the desired error bound.

3. Proposed Methodology

In this section we outline the proposed surrogate that acts as an emulator for EpiCast in a wide range of MSAs. Next, we describe how this surrogate can be used for calibrating EpiCast in different scenarios.

3.1. Surrogate training

An accurate neural network based surrogate for the EpiCast simulation can speed up compute by orders of magnitude while producing outputs consistent with EpiCast . As a result, even an approximate surrogate can be very useful for calibration, by enabling several thousand evaluations in order to infer accurate posteriors. In contrast to general surrogate modeling problems, EpiCast has the critical challenge of making consistent predictions across all MSAs – potentially even ones which may not be accessible during training. This is a practical necessity that can help MSAs that are infected relatively later in the pandemic to leverage data and information from MSAs that are infected earlier.

In order to train such a “universal” surrogate to make predictions across MSAs, we need a few simple, reversible data transformations – (1) first, we consider the number of cumulative cases over time instead of the daily infected cases, since the former is easier to predict (smooth trajectory), (2) we transform the data into a *population normalized space* such that we operate with the number of infections relative to the total population; and (3) finally we reduce the dimensionality of the cumulative case curves using any standard dimensionality reduction method (10-component PCA in our experiments). As a result, surrogate fitting is now reposed as $\hat{\mathcal{E}} : x \mapsto z$, where $z \in \mathbb{R}^{10}$ correspond to the PCA components of the curve; the final predicted curve is obtained by $\hat{y} = \Pi^T z$, where Π is the PCA basis estimated using the training data.

Surrogate network architecture The surrogate is modeled as a fully connected 3-layer neural network that maps a 6-D input parameter to a 10-D space using LeakyRelu (Xu et al., 2015) activations, and batch-normalization (Ioffe & Szegedy, 2015). In general, we find that training in the reduced dimensionality space affords better surrogates with fewer training examples. For more complex curves and larger datasets, Π can be replaced with a more sophisticated pre-trained representations, e.g., auto-encoder or a generative model; similar observations have been reported for other surrogate modeling problems in (Anirudh et al., 2020).

3.2. Surrogate-based calibration of epidemiological models

Calibration is an inverse problem with a non-linear forward operator, i.e., the epidemiological simulation. Like typical inverse problems, calibration is highly ill-posed, i.e., there are several solutions (input parameters) that may yield the same simulation output. For example, the cumulative number of new infections over the next T days maybe very similar either when the initial number of infections is high and the transmission probability is low or if the number of infections is low but transmission rate is high. As a result, we need to resolve these potential solutions further using prior knowledge of the pandemic, where for instance, an unusually low (or high) transmission rate may not be possible and similarly using information about the current number of infected cases. We formally describe the posterior estimation problem and present the novel regularization strategies next.

Formally, calibration is the problem of determining $p(x|y_{obs})$, commonly referred to as the posterior distribution, given an observation y_{obs}. Here, the observations specify the trajectory of cumulative new infections over the last T days of observation. In addition, $p(y)$ may not be computationally tractable, and as a result determining the posterior in closed form is non-trivial.

Initialization: The quality of samples obtained is dependent on the initialization for x. For example, a relatively safe choice is to draw the initial seed set from an uniform distribution, which is a weak prior since it assumes all values of the parameters are equally likely which matches the desired sampling of the training data. However, given the relatively low number of samples per ensemble (500 in 6-dimensional space) the actual samples produced need not be uniform. Hence, we found it useful to draw $\mathbf{x}_{init}$ from a multivariate normal (MVN) distribution, whose means and covariance are estimated from the input parameter settings in the training set.

3.2.1. Naïve posteriors

The easiest way to estimate the posterior is to choose an appropriate initial set of samples that are expected to cover the range of potential solutions and filter the top-K based on their similarity according to a metric of choice (such as an ℓ_p-norm). The tightness of the posterior estimate improves with a larger set of initial samples since it increases the odds of finding closer matches. As a result, often requiring tens of thousands of evaluations, which become infeasible with a complex, computationally intensive simulation like EpiCast . Instead, here we propose to use the surrogate to evaluate a large number of initial samples ($\sim$ 200K) drawn from a multivariate distribution, $\mathbf{x}_{init}$. We then find the top-K matches based only on the data from the observed time-steps, and use that as the approximate posterior. In other words, we find the top-K samples that minimize: $\sum_{i=1}^{n} ||y_{obs} - \Pi^T(\hat{\mathcal{E}}(x_i))||_p$, where $\|.\|_p$ denotes the ℓ_p norm ($p = 2$ in our experiments).

3.2.2. Optimized posteriors using gradient descent

The naïve approach relies on a large initial set in order to find good matches, which may not always feasible, and requires an exponentially large number of such seed samples to fit a specified error. Alternatively, we explore a more direct optimization strategy using a much smaller initial seed set (1K samples) and find a better posterior. Here the problem is reformulated as an optimization problem in x, and solved using gradient descent. In particular, we can use an optimizer like Adam (Kingma & Ba, 2014) to search the parameter $\mathcal{X}$ for the optimal x^*. This is reminiscent of projected gradient descent (PGD) style optimization strategies that find an appropriate sample in the range space of a pretrained generator to solve inverse imaging problems (Yeh et al., 2017). For this optimization, we compute the loss directly on the curves, instead of their PCA coefficients, as we find that provides better gradients for our optimization. Mathematically,

$$\min_{\{x_i \in \mathcal{X}\}} \left[\sum_{i=1}^{n} \left\| (y_{obs} - \Pi^T(\hat{\mathcal{E}}(x_i)) \right\|_2 \right] + \lambda \mathcal{R}(\{x_i\}), \quad (1)$$

where $\mathcal{R}(.)$ is an appropriate regularization term that reduces the potential solution space. Like in many inverse problems, the choice of the regularizer plays a key role in finding high quality solutions. However, commonly used regularizers such as sparsity or total variation do not apply in the case of epidemiological models since the parameters have a physical meaning (and specific set of ranges) that must be respected.

Regularization objectives for epidemiological model calibration We consider two different regularization objectives in this work, inspired by the physical quantities represented by the parameters. First, we use a Kullback-Liebler divergence objective between $\mathbf{x}$ and $\mathbf{x}_{init}$ to ensure that the gradient descent does not ignore the meaningful relationships across input parameters. We approximate $\mathbf{x}$ at each step to be drawn from a multi-variate normal (MVN) distribution, and as a result we can compute the KL divergence in closed form between two MVNs (See (Duchi, 2007) for derivation) as

$$\mathrm{KL}(p(x)||p(x_{init})) = \frac{1}{2}\Bigg(\log \frac{\det \Sigma_2}{\det \Sigma_1} - d + \mathrm{Tr}(\Sigma_2^{-1}\Sigma_1) + (\mu_2 - \mu_1)^T \Sigma_2^{-1}(\mu_2 - \mu_1) \Bigg) \quad (2)$$

where $\mu_1, \mu_2, \Sigma_1, \Sigma_2$ correspond to the mean and covariances for $\mathbf{x}$ and the initialization $\mathbf{x}_{init}$ respectively.

Consistency conditions for the global parameters As discussed above, EpiCast uses three local parameters that are expected to vary between regions and three global ones expected to be the same. Therefore, when calibrating the model for a particular region, we can use observations from other regions for which data is available to constrain the unknown global parameters to be consistent. We achieve this by simultaneously optimizing for the parameters across K different MSAs, $\mathbf{x}^{(j)} = \{x_i^{(j)}\} \forall j = 1, \cdots, K$, using the following cost:

$$\mathcal{L}_{multi} = \sum_{j=1}^{K} \sum_{i=1}^{n} \|y_{obs}^{(j)} - \Pi^T(\hat{\mathcal{E}}(x_i^{(j)})\|_2, \quad (3)$$

We then impose a KL-divergence based cost using only the global parameters, to define this global consistency constraint for each MSA j:

$$\mathcal{G}^{(j)} = \sum_{k \neq j} \mathrm{KL}\left(p(x^{(j)})||p(x^{(k)})\right), \quad (4)$$

where $x^{(j)}$ is the random variable corresponding to the observations in the sample set $\mathbf{x}^{(j)}$. Note, the consistency is imposed across the K different MSAs. The overall optimization objective to calibrate EpiCast for all the K different MSAs simultaneously is

$$\left\{\mathbf{x}^{(j)*}\right\}_{j=1}^{K} = \arg \min_{\{\mathbf{x}^{(j)}\}_{j=1}^{K}} \mathcal{L} = \mathcal{L}_{multi} + \lambda_1 \sum_{j=1}^{K} \mathrm{KL}(p(x^{(j)})||p(x_{init})) + \lambda_2 \sum_{j=1}^{K} \mathcal{G}^{(j)}, \quad (5)$$

where the KL cost is only measured on the k^{th}–MSA of interest, while the rest of the terms are used on all the test

MSAs available for calibration. λ_1, λ_2 are regularization weights whose values are fixed using cross validation. In all our experiments, we used $\lambda_1 = 1\mathrm{e}-6, \lambda_2 = 1\mathrm{e}-4$.

The complete calibration procedure for a single MSA is outlined in algorithm 1, it can be trivially extended to the multi-MSA setting. Following the suggested reporting guideline for calibration methods outlined in (Stout et al., 2009), we summarize the key aspects of our calibration strategy: (a) *Target:* Cumulative curve of new infections over a T-day period. (b) *GOF metric:* Mean squared error, (c) *Search algorithm:* Gradient descent using a NN-surrogate as outlined in alg. 1 with 1000 initial samples. (d) *Acceptance criteria:* Convergence of GOF measure (e) *Stopping rule:* $N_{max} = 25000$ steps.

Algorithm 1 Proposed Algorithm for Posterior Estimation

```
procedure CALIBRATION($y_{obs}^{(j)}, \hat{\mathcal{E}}, \Pi, \mathbf{x}_{init}$)
    Initialization $\mathbf{x}^{(j)} = \mathbf{x}_{init}, \forall j$
    for $n \leftarrow 1$ to $N_{max}$ do
        $\hat{y}^{(j)} = \Pi^T \hat{\mathcal{E}}(\mathbf{x}^{(j)}), \forall j$        ▷ $\Pi$: PCA Basis
        Compute objection in eqn. (5)
        $\mathbf{x}^{(j)} \leftarrow \mathbf{x}^{(j)} - \gamma \nabla \mathcal{L}, \forall j$        ▷ Update estimate
    return $\{\mathbf{x}^{(j)}\}$        ▷ samples from the posterior
```

4. Experiments and Results

In this section we outline the experimental details and describe their results. We first describe the training protocol for the surrogate, followed by its use in posterior estimation for calibrating the main epidemiological model.

4.1. EpiCast Simulation Dataset

As noted in section 2.1, we create a dataset that is representative of the functionality of EpiCast using 7500 runs that consist of 15 MSAs with 500 runs each. The names and descriptions of individual MSAs considered in this work are available in the supplement. Each simulation takes about three minutes on 64 cores of a Knights Landing (KNL) node on one of the top 20 supercomputer. Together with pre- and post-processing simulations, workflow overheads etc., the entire ensemble required $\sim 40,000$ core hours. We use 10 MSAs to train our surrogate the remaining for testing. We perform all our calibration experiments using 15 test curves from MSA015.

4.2. Surrogate Training and Evaluation

The surrogate network consists of a 3 dense layers, along with a LeakyReLu activation and batch normalization after the first and second layers. The surrogate predicts into the PCA latent space, which are decoded using the PCA components to obtain the final curve, as described earlier.

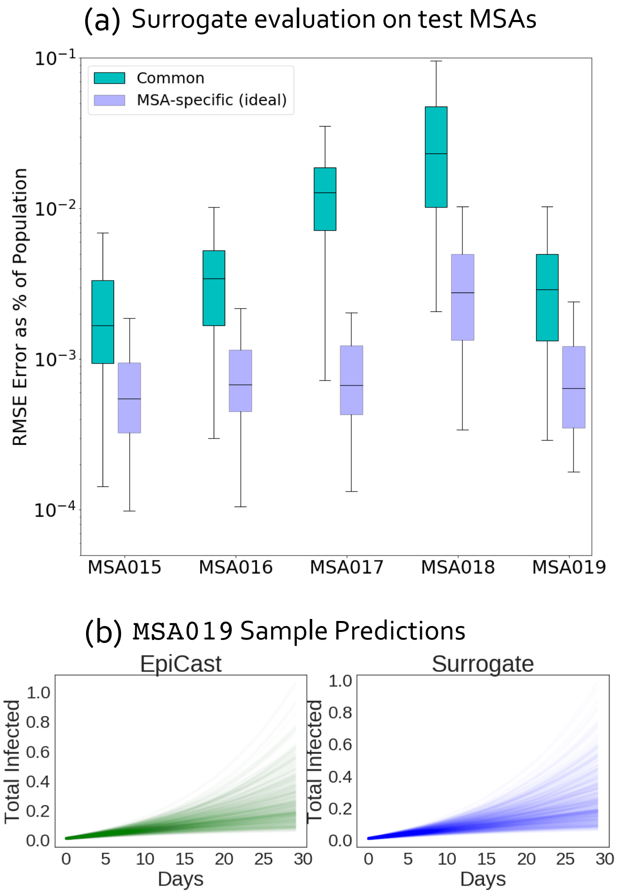

Figure 1. Our EpiCast surrogate can generalize to previously unseen MSAs in population-normalized space, often predicting within 0.1% error of the total population. Here, we show the 25% and 75% confidence intervals estimated using 500 curves across the 5 test-MSAs not accessed during training.

We train the model on 500 evaluations of EpiCast on 10 MSAs, and test it on 100 curves from 5 test MSAs not accessed during training. Although EpiCast produces curves specific to local regions based on demographic factors, we show that a surrogate in population normalized space is able to achieve reasonable performance accuracy. Figure 1(a) shows the prediction performance using the error as a percentage of the population for each of the 5 test MSAs for both a common surrogate across all MSAs and an MSA-specific surrogate. For the latter, we train on 400 curves from each of the MSAs and evaluate them on the remaining 100 (which are the same test curves as the common model). We observe that in all cases, a surrogate trained on a set of MSAs generalizes well to previously unseen MSAs often with a prediction error of under 0.1% of the total population, indicating similarities in infection spread across dense metropolitan regions. As expected, a ideal surrogate that has seen data from a particular MSA performs better, but not significantly so. In Figure 1(b) we also show sample predictions from both EpiCast epidemiological model and our NN-surrogate and see that for one of the test MSA019, our predictions match the model accurately, while being more efficient to compute.

4.3. Calibration and posterior estimation

To evaluate calibration performance we use 15 curves from MSA015, which is one of the given test MSAs. We use the inputs that were originally used to generate these curves using EpiCast as the "true values" in our evaluations. Since the problem is highly ill-posed we expect several combinations of the same parameters to produce the same curve, as a result, we evaluate the performance in two quantitative ways (a) a root-mean squared error (RMSE) goodness of fit compared to the observed curve and those produced by evaluating EpiCast on samples from the posterior for each method in consideration. (b) We also measure the RMSE of the estimated parameters on the two most sensitive parameters directly – `INFECTED` and `TRANSPROP` which correspond to the number of infected at the start of measurements, and the transmission probability of the infections respectively. In reality, the global consistency cost can be directly applied across MSAs, but since we are working with simulated data, we find samples across the test MSAs with similar global parameters and use the corresponding curves across MSA 015, 016, 017 in computing the cost according to (4).

4.3.1. Qualitative results

In Figure 2 we show how the quality of calibration changes with more available data, and better constraints. In each case we run EpiCast on the samples generated and show the observed curve, the true curve, and the curves produced by the simulation. Ideally, we want the curves from the simulation to be as close as possible to the true curve, given only the observed curve. As we observe in Fig 2, when very few observations are available the naïve model and the least constrained optimized posterior are most likely to contain the true solution, mainly because these methods have a very large variance and their posteriors are weak. Next, as we observe more data, we observe that including the proposed regularization objectives significantly improves the quality of the fits for the same number of EpiCast evaluations.

In Figure 3, we show the marginal distribution of samples from the posterior distribution on all six parameters – the first three are "local" parameters, i.e., specific to the geographical region and the last three are "global". We compare all four methods on a test curve from RMSA015, with respect to the true value which is shown as a black dashed line. As expected, we observe that the marginals for samples obtained using the naive and the unconstrained optimization methods yield very wide posteriors, even on highly sensitive parameters like `INFECTED` and `TRANSPROP`, that are expected to be tight. On the other hand, the samples obtained using the regularized optimization yield very confident estimates for all the parameters. We also show the corresponding EpiCast evaluations of these samples and observe that the proposed constraints significantly help improve the posterior fits, yielding accurate curves compared to the ground truth.

4.3.2. Quantitative evaluation

In table 2, we show the RMSE error for the top-100 curves (based on GOF), across 15 test curves from RMSA015. For each test curve, we estimate the samples from the posterior followed by evaluating the top 100 sample candidates using EpiCast . In all cases, the cost for optimization or sample selection/filtering is done based on curves for the number of days they can be observed. Next, in table 1 we evaluate the quality of estimating the two mose sensitive parameters to EpiCast . Since the problem of calibration is highly ill-posed, average metrics comparing samples from the posterior distribution to the ground truth may not be directly meaningful, however the sensitive parameters are expected to be easier to estimate (as seen in Fig 2). We observe that in all scenarios considered across both the parameters, the regularization objectives help in optimization and in the accurate recovery of sensitive parameters.

5. Related Work

Agent-based models (ABMs) (Perez & Dragicevic, 2009; Germann et al., 2006) are computational models that are commonly used to inform policy-making regarding public health. Such models must be 'calibrated' in order to explain observed data, which involves identifying the input parameters to the computational model that can exactly match the observed data according to a goodness of fit (GOF) metric. Since this process is under specified, there is often a large set of solutions that yield the same output to match the observation. Within epidemiology there are several studies that address the calibration problem for both agent-based and SEIR models (Dantas et al., 2018; Ward et al., 2016; Farah et al., 2014), including numerous recent works related to the COVID-19 pandemic(Bertozzi et al., 2020; Hazelbag et al., 2020). The focus of our work is specifically on ABMs, where previous works have mostly focused on calibration achieved via primarily parameter searches or optimization strategies based on GOF metrics such as R-squared, RMSE, absolute error etc. We refer the interested reader to Tables 1&2 of (Hazelbag et al., 2020) for a comprehensive list of these works.

Most existing works have either focused on optimization strategies to obtain point estimates, or sampling algorithms

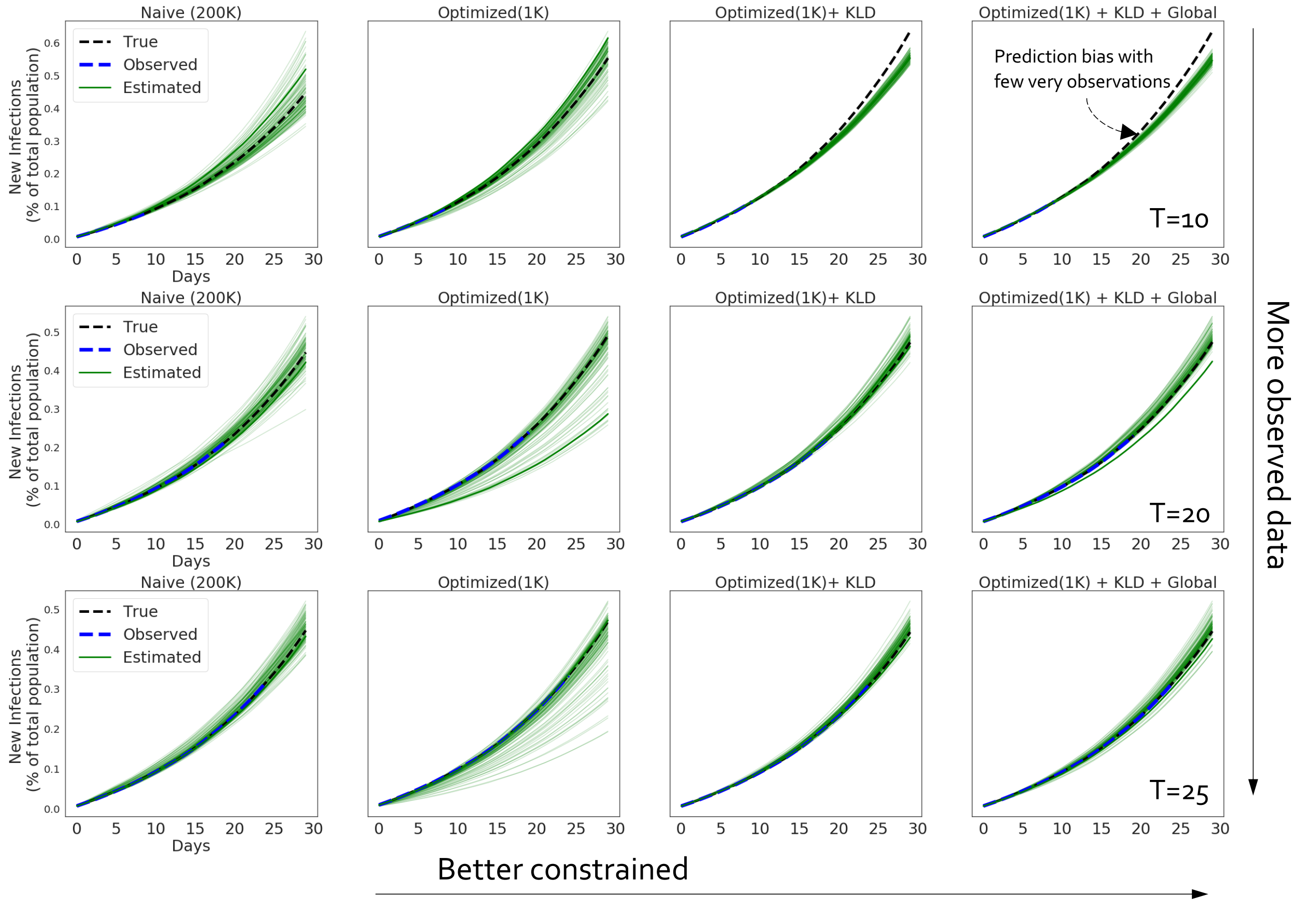

Figure 2. Comparing EpiCast evaluations from the samples obtained using different posterior estimation methods considered here. We show how the estimates change with more data observed.

Method	$\mathbf{T_{obs} = 10}$	$\mathbf{T_{obs} = 20}$	$\mathbf{T_{obs} = 25}$	$\mathbf{T_{obs} = 10}$	$\mathbf{T_{obs} = 20}$	$\mathbf{T_{obs} = 25}$
		INFECTED			**TRANSPROP**	
Naïve	0.029 ± 0.023	0.029 ± 0.025	0.029 ± 0.023	0.022 ± 0.015	0.021 ± 0.014	0.021 ± 0.014
Optimized	0.035 ± 0.032	0.063 ± 0.030	0.062 ± 0.033	0.029 ± 0.020	0.035 ± 0.019	0.029 ± 0.018
+KLD	0.015 ± 0.005	$\mathbf{0.025 \pm 0.024}$	0.024 ± 0.022	$\mathbf{0.021 \pm 0.008}$	$\mathbf{0.017 \pm 0.009}$	0.017 ± 0.009
+Global	$\mathbf{0.014 \pm 0.006}$	0.028 ± 0.022	$\mathbf{0.023 \pm 0.014}$	$\mathbf{0.021 \pm 0.009}$	$\mathbf{0.017 \pm 0.008}$	$\mathbf{0.014 \pm 0.006}$

Table 1. Evaluating RMSE on the two most sensitive parameters `INFECTED, TRANSPROP`. We observe that the additional constraints studied in this paper improve the quality of the estimate. The mean prediction and standard deviation are averaged across 20 different observations from a test set corresponding to MSA015.

Method	$\mathbf{T_{obs} = 10}$	$\mathbf{T_{obs} = 20}$	$\mathbf{T_{obs} = 25}$
Naïve	9.93 ± 6.88	7.07 ± 5.61	6.70 ± 5.57
Optimized	$\mathbf{8.93 \pm 9.11}$	16.64 ± 17.17	13.06 ± 12.17
+ KLD	11.42 ± 2.38	$\mathbf{5.20 \pm 4.59}$	5.47 ± 4.81
+ Global	11.63 ± 2.55	6.23 ± 5.26	$\mathbf{5.42 \pm 4.37}$

Table 2. Evaluating RMSE quality (as percentage of population $\times 1\mathrm{e}-3$) for curve fits by running samples by evaluating samples from the posterior with the EpiCast simulation.

to find a distribution of parameter values to approximate the posterior distribution. In contrast, our approach relies on a non-linear neural network-based surrogate that acts as a proxy for the computationally intensive epidemiological model. Since inference with a neural network is orders of magnitude cheaper than the original simulation, we are able to employ complex optimization strategies for calibration. To the best of our knowledge we are the first to propose calibration using a neural network based surrogate for ABMs. Our work is also related to an increasing number of covid

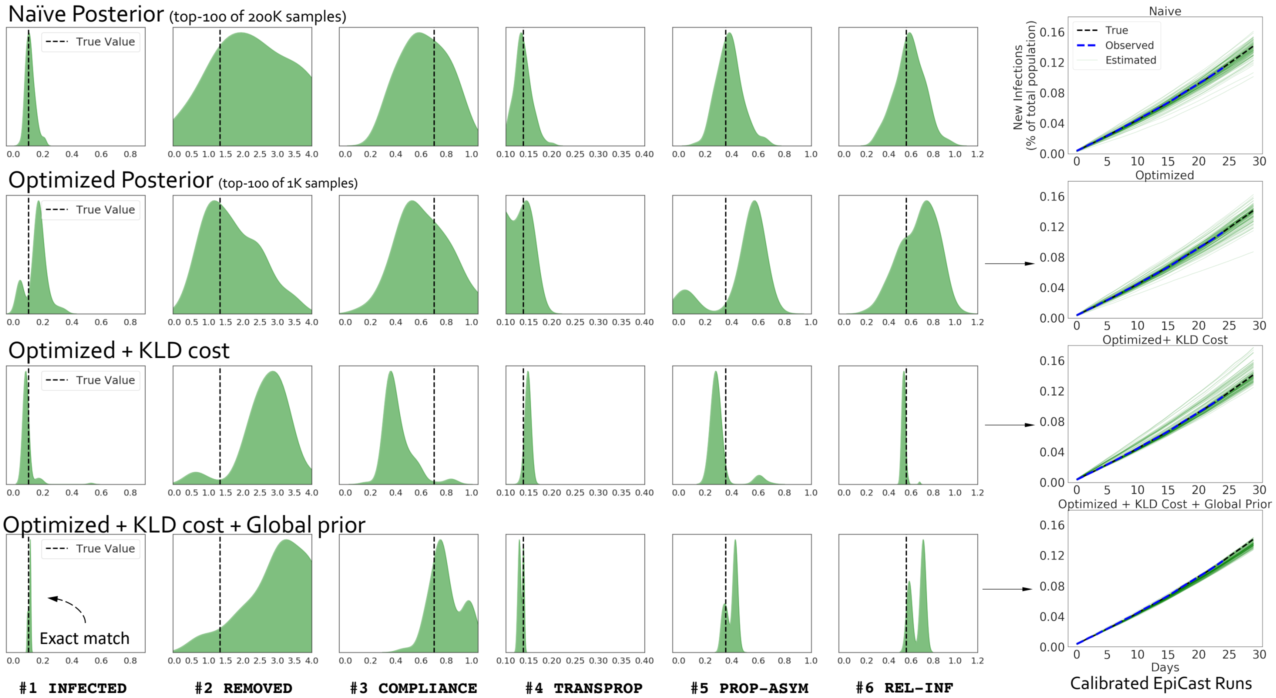

Figure 3. Comparing the marginals of posteriors obtained using different methods outlined in this work. At the bottom, we show the EpiCast evaluations on samples from these posteriors. We observe that the proposed priors yield tighter posteriors, and consequently more accurate matches with the epidemiological model.

modeling efforts using machine learning (Davis et al., 2020; Soures et al., 2020; Libin et al., 2020; Gu, 2020) that have been used to model and design mitigations with simpler epidemiological models like the SEIR models. In contrast, we are interested in the more complex, computationally intensive ABMs. Similar to our work, (Davis et al., 2020) focus on the specific problem of surrogate modeling for SIR models and other individual-based models. Our focus is on building an effective surrogate that can also be used in calibration across geographical regions not accessed during training.

6. Discussion

Epidemiological models play a crucial role in assessing the state of a public health crisis like the ongoing COVID-19 pandemic. These computational models enable decision makers to explore what-if scenarios and forecasts acting as an important planning tool for implementing health policies nationwide. Like all computational models in order for them to be effective, they must be calibrated to fit observed data, which is the process of finding the right combination of input parameters that can match the observed data to some desired degree of fidelity. We are specifically interested in calibration of complex agent-based models (ABMs), that are computationally intensive. We propose an alternative strategy for using an ABM called EpiCast (Germann et al., 2006) with the help of a neural network-based surrogate. We demonstrate that our surrogate can effecitvely emulate the ABM's behaviour – including across geographical regions that were not included in the training dataset. We additionally propose novel regularization objectivesto make the highly ill-posed inverse problem more tractable, and can successfully matches observed curves when samples from our estimated posteriors are evaluated usingEpiCast . We expect calibration strategies like those proposed in this work to help in more effective usage of ABMs in characterizing and fighting the spread of COVID-19.

Acknowledgements

This work was performed under the auspices of the U.S. Department of Energy by the Lawrence Livermore National Laboratory under Contract No. DE-AC52-07NA27344, and was supported by the DOE Office of Science through the National Virtual Biotechnology Laboratory, a consortium of DOE national laboratories focused on response to COVID-19, with funding provided by the Coronavirus CARES Act.

References

Anirudh, R., Thiagarajan, J. J., Bremer, P.-T., and Spears, B. K. Improved surrogates in inertial confinement fusion with manifold and cycle consistencies. *Proceed-*

ings of the National Academy of Sciences, 117(18): 9741–9746, 2020. ISSN 0027-8424. doi: 10.1073/pnas.1916634117. URL https://www.pnas.org/content/117/18/9741.

Bertozzi, A. L., Franco, E., Mohler, G., Short, M. B., and Sledge, D. The challenges of modeling and forecasting the spread of covid-19. *arXiv preprint arXiv:2004.04741*, 2020.

Dantas, E., Tosin, M., and Cunha Jr, A. Calibration of a seir–sei epidemic model to describe the zika virus outbreak in brazil. *Applied Mathematics and Computation*, 338: 249–259, 2018.

Davis, C. N., Hollingsworth, T. D., Caudron, Q., and Irvine, M. A. The use of mixture density networks in the emulation of complex epidemiological individual-based models. *PLoS computational biology*, 16(3):e1006869, 2020.

Dong, E., Du, H., and Gardner, L. An interactive web-based dashboard to track covid-19 in real time. *The Lancet infectious diseases*, 20(5):533–534, 2020.

Duchi, J. Derivations for linear algebra and optimization. *Berkeley, California*, 3:2325–5870, 2007.

Farah, M., Birrell, P., Conti, S., and Angelis, D. D. Bayesian emulation and calibration of a dynamic epidemic model for a/h1n1 influenza. *Journal of the American Statistical Association*, 109(508):1398–1411, 2014.

Germann, T. C., Kadau, K., Longini, I. M., and Macken, C. A. Mitigation strategies for pandemic influenza in the united states. *Proceedings of the National Academy of Sciences*, 103(15):5935–5940, 2006. doi: 10.1073/pnas.0601266103. URL https://www.pnas.org/content/103/15/5935.

Gu, Y. Covid-19 projections using machine learning. https://covid19-projections.com/, 2020. Accessed: 2020-Sept-09.

Halloran, M. E., Ferguson, N. M., Eubank, S., Longini, I. M., Cummings, D. A. T., Lewis, B., Xu, S., Fraser, C., Vullikanti, A., Germann, T. C., Wagener, D., Beckman, R., Kadau, K., Barrett, C., Macken, C. A., Burke, D. S., and Cooley, P. Modeling targeted layered containment of an influenza pandemic in the united states. *Proceedings of the National Academy of Sciences*, 105 (12):4639–4644, 2008. ISSN 0027-8424. doi: 10.1073/pnas.0706849105. URL https://www.pnas.org/content/105/12/4639.

Hazelbag, C. M., Dushoff, J., Dominic, E. M., Mthombothi, Z. E., and Delva, W. Calibration of individual-based models to epidemiological data: A systematic review. *PLoS computational biology*, 16(5):e1007893, 2020.

He, S., Peng, Y., and Sun, K. Seir modeling of the covid-19 and its dynamics. *Nonlinear Dynamics*, pp. 1–14, 2020.

Ioffe, S. and Szegedy, C. Batch normalization: Accelerating deep network training by reducing internal covariate shift. *arXiv preprint arXiv:1502.03167*, 2015.

Kingma, D. P. and Ba, J. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

Libin, P., Moonens, A., Verstraeten, T., Perez-Sanjines, F., Hens, N., Lemey, P., and Nowé, A. Deep reinforcement learning for large-scale epidemic control. *arXiv preprint arXiv:2003.13676*, 2020.

Perez, L. and Dragicevic, S. An agent-based approach for modeling dynamics of contagious disease spread. *International journal of health geographics*, 8(1):50, 2009.

Soures, N., Chambers, D., Carmichael, Z., Daram, A., Shah, D. P., Clark, K., Potter, L., and Kudithipudi, D. Sirnet: Understanding social distancing measures with hybrid neural network model for covid-19 infectious spread. *arXiv preprint arXiv:2004.10376*, 2020.

Stout, N. K., Knudsen, A. B., Kong, C. Y., McMahon, P. M., and Gazelle, G. S. Calibration methods used in cancer simulation models and suggested reporting guidelines. *Pharmacoeconomics*, 27(7):533–545, 2009.

Ward, J. A., Evans, A. J., and Malleson, N. S. Dynamic calibration of agent-based models using data assimilation. *Royal Society open science*, 3(4):150703, 2016.

Xu, B., Wang, N., Chen, T., and Li, M. Empirical evaluation of rectified activations in convolutional network. *arXiv preprint arXiv:1505.00853*, 2015.

Yeh, R. A., Chen, C., Yian Lim, T., Schwing, A. G., Hasegawa-Johnson, M., and Do, M. N. Semantic image inpainting with deep generative models. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 5485–5493, 2017.