Modeling Epidemic Spread: A Gaussian Process Regression Approach

Baike She[1], Lei Xin[2,†], Philip E. Paré[3], and Matthew Hale[1,*]

***Abstract*— Modeling epidemic spread is critical for informing policy decisions aimed at mitigation. Accordingly, in this work we present a new data-driven method based on Gaussian process regression (GPR) to model epidemic spread through the difference on the logarithmic scale of the infected cases. We bound the variance of the predictions made by GPR, which quantifies the impact of epidemic data on the proposed model. Next, we derive a high-probability error bound on the prediction error in terms of the distance between the training points and a testing point, the posterior variance, and the level of change in the spreading process, and we assess how the characteristics of the epidemic spread and infection data influence this error bound. We present examples that use GPR to model and predict epidemic spread by using real-world infection data gathered in the UK during the COVID-19 epidemic. These examples illustrate that, under typical conditions, the prediction for the next twenty days has 94.29% of the noisy data located within the 95% confidence interval, validating these predictions. We further compare the modeling and prediction results with other methods, such as polynomial regression, k-nearest neighbors (KNN) regression, and neural networks, to demonstrate the benefits of leveraging GPR in disease spread modeling.**

***Index Terms*— Epidemic Modeling, Epidemic Prediction, Error Bound, Gaussian process regression**

I. INTRODUCTION

MODELING and predicting the spread of diseases is critical for understanding spreading patterns and decision-making for epidemic mitigation [1], [2]. Existing epidemic modeling and prediction techniques typically construct compartmental models by selecting model structures and parameters to fit spreading data [3], [4], e.g., in the susceptible-infected-recovered (SIR) model. Distinct from existing works, we leverage Gaussian process regression to model spreading trends by studying the number of infected cases directly, without using any particular compartmental model.

Gaussian process regression excels at capturing complex, nonlinear relationships without relying on predefined functional forms and can effectively handle small datasets, which are common in measuring disease spreading [5], [6], [7]. For instance, [5] employs a spatio-temporal covariance function and data from various states and all weeks of the year to model influenza-like illness forecasting. Meanwhile, [7] trains individual Gaussian process (GP) models for each forecast based on a relatively small set of features from previous weeks, resulting in small, but reliable prediction intervals. Moreover, although existing works leverage Gaussian processes to model daily infected cases, they often overlook the fact that both the daily infected cases and the associated noise cannot follow a normal or log-normal distribution [8]. Therefore, applying Gaussian process regression directly to daily infected cases may not produce accurate results. In addition, these studies typically focus on empirical results of applying GPR to model and predict disease dynamics in time-series data, without delving into theoretical aspects.

To address these challenges, we introduce an approach to modeling the change on the logarithmic scale of the number of infected cases using Gaussian process regression, while also providing insights into model uncertainty. We propose an upper bound on posterior variance to assess the impact of epidemic data and develop a high-probability error bound to examine how epidemic spread and infection data influence the accuracy of predictions and confidence in them. These results help bridge the gap between theoretical analyses and practical applications in epidemic modeling, paving the way for predictive control methods in future efforts.

To illustrate this framework, we apply it to real COVID-19 infection data from the United Kingdom. These results show that by selecting appropriate parameters in the modeling process, predictions for twenty days into the future capture 94.29% of the actual data from those days within a 95% confidence interval, which validates the prediction accuracy in practice. As we show, most prediction errors arise from either drastic changes in the spreading trend or a limited number of available data samples, both of which increase uncertainty. We further use the same dataset to compare the modeling and prediction results with other methods, such as polynomial regression, k-nearest neighbors (KNN) regression, and neural networks, to demonstrate the benefits of leveraging GPR for disease spread modeling.

The rest of the paper is organized as follows: Section II provides background and problem statements; Section III proposes the model, analyzes prediction uncertainty, and illustrates the results with examples; Section IV concludes.

*This work is supported by the grant NSF-ECCS #2238388 and by DARPA under grant HR00112220038.

† Corresponding Author.

[1]Baike She and Matthew Hale are with the School of Electrical and Computer Engineering, Georgia Institute of Technology, Atlanta, GA, 30318, USA. bshe6@gatech.edu; matthale@gatech.edu

[2]Lei Xin is with the Department of Computer Science and Engineering, The Chinese University of Hong Kong, Hong Kong, 999077, China. lxinshenqing@gmail.com

[3]Philip E. Paré is with the Elmore Family School of Electrical and Computer Engineering, Purdue University, West Lafayette, IN 47907, USA. philpare@purdue.edu

Notation: We use $\mathbb{R}$ and $\mathbb{N}$ to denote the sets of reals and naturals, respectively. We define $\underline{n} = \{1, 2, \ldots, n\}$ for $n \in \mathbb{N}_{>0}$. We use $|S|$ to denote the cardinality of a finite set S. A compact interval $\mathbb{T}$ is of the form $[a, b]$, where $a, b \in \mathbb{R}_{\geq 0}$ and $a < b$. Its length is given by $\hat{\mathbb{T}} = b - a$. For $v \in \mathbb{R}^n$, we use $\text{diag}\{v\} \in \mathbb{R}^{n\times n}$ for the diagonal matrix whose i^{th} diagonal entry is v_i, for $i \in \underline{n}$. For a real symmetric matrix $A \in \mathbb{R}^{n\times n}$, let $[A]_{ij}$ denote its $i^{th}j^{th}$ entry and $\lambda_{\max}(A)$ denote its largest eigenvalue. We write $I_n \in \mathbb{R}^{n\times n}$ for the identity matrix. Let $\mathcal{N}(\mu, \sigma^2)$ denote the one-dimensional normal distribution with mean μ and variance σ^2. We use $\exp(\cdot)$ and $\log(\cdot)$ to denote the exponential function and the natural logarithmic function, respectively. We use $p(\cdot)$ to represent the probability distribution of a random variable.

II. Background and Problem Formulation

This section introduces the Gaussian process regression and states the problems that we solve in the rest of the paper.

A. Gaussian process regression

We briefly introduce the one-dimensional Gaussian process regression [9]. Consider an unknown function $f : \mathbb{R} \to \mathbb{R}$ and n inputs captured by $X = [x_1, \ldots, x_n]^\top \in \mathbb{R}^n$, where $x_i \in \mathbb{R}$, $i \in \underline{n}$, and $n \in \mathbb{N}_{>0}$. The corresponding outputs are given by the vector $F = [f(x_1), \ldots, f(x_n)]^\top \in \mathbb{R}^n$, where the n outputs in F follow a joint Gaussian distribution. The mean of the distribution is given by $m(X) = [m(x_1), \cdots, m(x_n)]^\top \in \mathbb{R}^n$. Suppose that the observation of each output $f(x_i)$ is corrupted with zero-mean independent Gaussian noise, i.e., $y(x_i) = f(x_i) + \varepsilon_i$, where $\varepsilon_i \sim \mathcal{N}(0, \sigma_i^2)$, and $\sigma_i^2 > 0$ denotes the variance of ε_i. Then, the covariance matrix of the noise is $\Sigma = \text{diag}\{\sigma_1^2, \ldots, \sigma_n^2\} \in \mathbb{R}^{n\times n}$. Using the noisy training dataset $\{(x_i, y(x_i))\}_{i=1}^n$, we can employ GPR to model the input-output relation $f : \mathbb{R} \to \mathbb{R}$ at a training location x_i, for $i \in \underline{n}$, as well as to predict the output $f(x^*) \in \mathbb{R}$ at some testing location $x^* \in \mathbb{R}$, where x^* is not necessarily one of the x_i's.

Gaussian process regression is a kernel-based approach. We use $k(\cdot, \cdot) : \mathbb{R} \times \mathbb{R} \to \mathbb{R}_{\geq 0}$ to represent the potential kernel function [10]. Let $K(X, X) \in \mathbb{R}_{\geq 0}^{n\times n}$ denote the kernel matrix of the training points, where $[K(X, X)]_{ij} = k(x_i, x_j)$ denotes the covariance between two training points x_i and x_j, for $i, j \in \underline{n}$. For a testing point x^*, we define $K(x^*, X) \in \mathbb{R}_{\geq 0}^{1\times n}$ as the kernel vector such that $[K(x^*, X)]_j = k(x^*, x_j)$, for $j \in \underline{n}$. Therefore, $K(x^*, X)$ captures the covariances between the testing point x^* and all training points. As Gaussian process regression operates as a Bayesian inference approach, we consider a zero-mean prior for generality [9], though the results we develop can be generalized for any other prior.

Consider the posterior distribution for the predicted random variable $f(x^*)$ at the testing location x^* conditioned on the noisy training data $\{(x_i, y(x_i))\}_{i=1}^n$. The posterior mean $m(x^*)$ and posterior variance $\sigma^2(x^*)$ at the testing location x^* are given by the following result [9, Equation (2.22)].

Proposition 1. *Let $Y = [y(x_1), \cdots, y(x_n)]^\top$. Then we have $p(f(x^*) \mid \{y(x_i)\}_{i=1}^n) = \mathcal{N}(m_Y(x^*), \sigma_Y^2(x^*))$, where*

$$m_Y(x^*) = K(x^*, X)(K(X, X) + \Sigma)^{-1}Y,$$
$$\sigma_Y^2(x^*) = k(x^*, x^*) - K(x^*, X)(K(X, X) + \Sigma)^{-1}K(x^*, X)^\top.$$

B. Problem Formulation

As discussed in the introduction, critical metrics such as the number of infected cases are essential for assessing epidemic spread. However, both these cases and their associated noise do not follow a normal or log-normal distribution [8]. Furthermore, epidemic severity is typically monitored through population testing and data reporting, where observation noise is unavoidable. However, existing works often overlook these limitations and use Gaussian processes to model the spread by directly focusing on the trend of time-series infection cases. Additionally, missing cases frequently arise due to insufficient testing or underreporting. Therefore, in this work, we solve the following problems:

Problem 1. *Develop a new model that uses Gaussian process regression to model epidemic spread, and show the effectiveness of this approach analytically and numerically.*

Problem 2. *Quantify how noise and data sample size affect the prediction results when using the model from Problem 1.*

Problem 3. *Develop a high-probability error bound on the prediction error to analyze the impact of data, and validate this result on real-world data.*

Problem 4. *Compare the GPR modeling and prediction results with other methods to demonstrate its benefits using real-world spread data.*

III. Gaussian process regression for Modeling and Predicting Epidemic Spread

In Section III-A, we model epidemic spread using Gaussian process regression, and then quantify its accuracy in a lemma and an example. This solves Problem 1. In Section III-B, we establish an upper bound on the prediction variance to assess the impact of infection data, which solves Problem 2. Section III-C presents a high-probability error bound, explaining the relationship between the spreading dynamics, data, and prediction error, solving Problem 3. We compare GPR with other modeling methods, through Examples 1 and 2, addressing Problem 4.

A. Modeling the Spread Using Gaussian Process Regression

We first solve Problem 1 by introducing a model to capture disease spreading trends. For an epidemic spreading process, we use $I(t)$ to denote a noisy observation of the number of the infected cases at time step $t \geq 0$, i.e., $I(t)$ equals the number of infected cases at time t plus some noise. We use $\bar{I}(t)$ to denote the true infected cases without observation noise. Consider the datasets $\{I(t_1), I(t_2), \ldots, I(t_n)\}$ and $\{I(t_1 - \eta), I(t_2 - \eta), \ldots, I(t_n - \eta)\}$ of an epidemic spreading process, measured at times $\{t_1, t_2, \ldots, t_n\}$ and $\{t_1 - \eta, t_2 - \eta, \ldots, t_n - \eta\}$, respectively, where $\eta > 0$, $t - \eta > 0$, and $n \in \mathbb{N}_{>0}$. We assume that the change in the logarithmic scale of the number of infected cases between consecutive time steps approximates a Gaussian process. Thus, for $i \in \underline{n}$, we define

$$\Delta(t_i) = \underbrace{\log(\bar{I}(t_i)) - \log(\bar{I}(t_i - \eta))}_{\bar{\Delta}(t_i)} + \varepsilon(t_i). \quad (1)$$

Remark 1. *We mentioned in Section II that infected cases typically do not follow a Gaussian or even a log-normal distribution. As discussed in [11], both the cases and the noise are often better modeled using Poisson or negative binomial distributions. The phenomenon where differences of log-transformed variables can approximate a Gaussian distribution is rooted in statistical principles such as variance stabilization [12], the Central Limit Theorem, and the Delta Method [13]. While the log-transformation of Poisson-distributed data may introduce skewness, differencing tends to normalize highly-skewed distributions. This behavior is often leveraged in fields like epidemiology and economics to model relative changes in data empirically, including changes in disease counts [14], [15], [16]. For example, studies such as [14] and [15] confirm that modeling differences on the logarithmic scale of case counts improves predictability and aligns with Gaussian-based inference methods. Widely-used tools (e.g., [16] for COVID-19 modeling) rely on relative changes in case counts, often log-differenced, to ensure statistical validity under Gaussian assumptions. Therefore, we provide an analysis based on the first-order differences on the logarithmic scale of the infected cases in* (1)*.*

Proposition 2. *Consider $\bar{\Delta}(t_i) = \log(\bar{I}(t_i)) - \log(\bar{I}(t_i - \eta))$, and $\bar{I}(t) > 0$, for all $(t_i - \eta) \in \mathbb{R}_{>0}$. If the number of infected cases increases during the time interval $[t_1, t_2]$, then for any time t_i such that $t_i, t_i - \eta \in [t_1, t_2]$, we have $\bar{\Delta}(t_i) > 0$, and vice versa. If the number of infected cases remains unchanged, then $\bar{\Delta}(t_i) = 0$.*

Remark 2. *The proof of Proposition 2 is in the Appendix [17]. Proposition 2 is a direct result of the definition of $\bar{\Delta}(t_i)$ and the properties of the logarithmic function. Note that $\bar{\Delta}(t_i) = \log(\bar{I}(t_i)/\bar{I}(t_i - \eta))$ captures the ratio between the number of infected cases at time step t_i and time step $t_i - \eta$, for $\eta > 0$. We use the function $\bar{\Delta} : \mathbb{R}_{\geq 0} \to \mathbb{R}$ to model and analyze spreading trends. Similar to the reproduction number [18], which uses the threshold of one to determine whether infected cases are increasing or decreasing, we can use the threshold of zero for $\bar{\Delta}$ to assess epidemic spread. If $\bar{\Delta}$ is greater than zero, the spread is increasing; if it is less than zero, the spread is decreasing.*

To simplify the analysis, we consider $I(t) > 0$ for all $t \geq 0$. In addition, we use i.i.d. noise $\varepsilon(t_i) \sim \mathcal{N}(0, \sigma^2)$ to capture the noise term in (1). Thus, the covariance matrix of the noise terms is $\Sigma = \sigma^2 I_n$. For an epidemic spreading process, we consider a set of n time steps $T = \{t_1, t_2, \ldots, t_n\}$ as the input batch, for $n \in \mathbb{N}_{>0}$. The corresponding output batch of n entries is given by $\{\Delta(t_1), \Delta(t_2), \ldots, \Delta(t_n)\}$. We define $\mathbf{\Delta} = [\Delta(t_1), \Delta(t_2), \ldots, \Delta(t_n)]^\top$, where $\Delta(t_i)$ is defined in (1), for $t_i \in T$ and $\eta > 0$. Our goal is to model and predict the first-order difference on the logarithmic scale of the infected cases. Hence, let the testing location be the time step $t^* \in \mathbb{T}$, $\mathbb{T} = [a, b]$. We can apply Proposition 1 to this model and the results are in Proposition 3.

Proposition 3. *Consider an unknown function $\bar{\Delta} : \mathbb{R}_{\geq 0} \to \mathbb{R}$. The posterior mean at time t^* is given by $m_\Delta(t^*) = K(t^*, T)(K(T, T) + \sigma^2 I_n)^{-1}\mathbf{\Delta}$, and the posterior variance is $\sigma^2_\Delta(t^*) = k(t^*, t^*) - K(t^*, T)(K(T, T) + \sigma^2 I_n)^{-1} K(t^*, T)^\top$, where $K(T, T)$ is the kernel matrix between the training points, $k(t^*, T)$ is the vector of covariances between the testing point t^* and all training points, $k(t^*, t^*)$ is the variance at the testing point, and $\sigma^2 I_n$ is the covariance matrix of the i.i.d noise.*

For analysis, we specify the kernel function as the squared exponential kernel to capture the covariance between any pair of points $a, b \in \mathbb{R}_{\geq 0}$. The kernel itself is

$$k(a, b) = \alpha^2 \exp\left(-\frac{(a - b)^2}{2\beta^2}\right), \quad (2)$$

where $\beta > 0$ is the length scale of the kernel, and α^2 is the signal variance.

Remark 3. *While our results can extend to other stationary kernels that are Lipschitz continuous, we use the squared exponential kernel due to its effectiveness in modeling epidemics [19]. The squared exponential kernel, commonly used in Gaussian process regression, depends on the distance between variables rather than their absolute positions, promoting smoothness and Lipschitz continuity in the modeled functions. These properties are essential for deriving error bounds [9], and we will use them below.*

The Gaussian process model in Proposition 3 and the spread modeling approach in (1) together address Problem 1. We use the following example to illustrate the proposed Gaussian process model for modeling epidemic spread, which partially addresses Problem 4..

Example 1. *We use Gaussian process regression to model the real-world epidemic spread of COVID-19 in the United Kingdom, using infection data from March 1^{st}, 2022, to February 28^{th}, 2023 [20]. Figure 1 shows the daily number of infected cases per million people in the UK population during this period. This dataset is chosen due to its multiple waves of infection and the variability in data size over time. To better capture the changes in daily infected cases, we preprocess the data by applying a thirty-day rolling average. For example, the average number of cases on March 30^{th}, 2022, is calculated by averaging the cases from March 1^{st} to March 30^{th}.*

By indexing March 30^{th}, 2022 as day one, we represent the daily infected cases on day t as $I(t)$, with $t \in \underline{335}$. We set $\eta = 7$. Then the difference on the logarithmic scale of consecutive daily observed infected cases is defined as $\Delta(t) = \log(I(t)) - \log(I(t - 7))$, where $t \geq 8$ and $t \in \underline{335}$. For example, $\Delta(8)$ on April 6^{th}, 2022, is calculated by subtracting the logarithmic scale of the infected cases on April 6^{th} from the logarithmic scale of those of March 31^{st}. We plot $\Delta(t)$, for $t \geq 8$ and $t \in \underline{335}$, from April 6^{th}, 2022 to February 28^{th}, 2023 in Figure 2 (dotted yellow line). Comparing Figure 2 to Figure 1, we see that $\Delta(t)$ effectively captures the trend in the spread. For instance, when $\Delta(t)$ is less than zero from April 2022 to May 2022, the trend in the number of infected cases decreases, as shown in Figure 1. This example demonstrates that when Δ is positive, daily infected cases are increasing, and when it is negative, they are decreasing.

Next, we apply GPR to model the spread. The training time steps are $\{8, 9, \ldots, 365\}$, corresponding to the training data

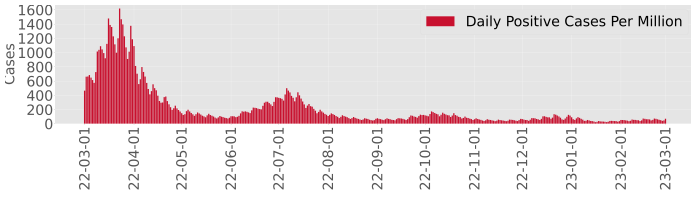

Fig. 1. The daily infected cases in the United Kingdom from $\mathbf{03-01-2022}$ to $\mathbf{02-28-2023}$ [20].

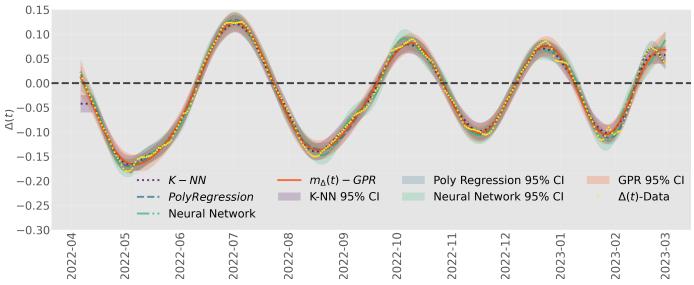

Fig. 2. Gaussian process regression model for $\boldsymbol{\Delta(t)}$.

$\{\Delta(8), \Delta(9), \ldots, \Delta(365)\}$, *as depicted by the dotted yellow line in Figure 2. Using the GPR algorithm from Proposition 3, we visualize the posterior means at the training time steps in Figure 2. We use* $\varepsilon \sim \mathcal{N}(0, 0.002)$ *to ensure that* 98.18% *of the noisy training data points fall within the* 95%*-confidence interval. The red solid line in Figure 2 represents the posterior mean, denoted as* $m^*_\Delta(t)$*, at the training locations* $t \in \underline{365}$ *and* $t \geq 8$*. We compute the mean square error (MSE) using the posterior mean and the training data, resulting in an MSE of* 0.000069 *for the Gaussian process regression model. The example shows how modeling the difference on the logarithmic scale of the infected cases effectively captures the spread dynamics and highlights the utility of Gaussian process regression in doing so*[1].

To validate the GPR model, we compare it with polynomial regression, k-nearest neighbors (KNN), and a neural network, commonly used in disease spread modeling [5]. For polynomial regression, we use a degree of 20, achieving an MSE of 0.000048*, with* 93.64% *of training data within the* 95% *confidence intervals, represented by a dashed blue line and blue-shaded areas. For KNN, setting* $\kappa = 15$ *results in an MSE of* 0.000080 *and* 95.78% *of the data within the intervals, represented by a dotted purple line and purple-shaded areas. The neural network employs sinusoidal feature generation to capture periodic patterns, using a three-layer structure* $(200 - 150 - 100$ *neurons) and achieves an MSE of* 0.000090 *with* 93.33% *of data within the intervals, represented by a dash-dotted green line and green-shaded areas.*

All four methods produce similarly accurate models, with MSEs on the same order of magnitude and most data within the 95% *confidence intervals. GPR stands out for its flexibility in kernel selection, such as adapting periodic kernels when necessary [5], while the other methods require careful parameter tuning—polynomial degree,* κ*, or neural network architecture. Notably, without sinusoidal feature augmentation, the neural network would underfit, yielding an MSE three orders of magnitude higher. Although effective, the alternative models' complexity can lead to overfitting, as explored in Example 2.*

[1]We perform a sensitivity analysis in GPR modeling by varying the averaging window size ($\{1, 3, 5, 10, 20, 30, 50\}$) and observe that shorter windows capture short-term perturbations, while longer windows highlight periodic patterns. Regardless of the window size, GPR models retain at least 92.48% of the training data within the 95% confidence intervals. Additionally, fixing the window size at 30, we vary η from the set $\{3, 7, 14, 21, 28\}$. Higher values of η yield narrower confidence intervals by better capturing long-term spreading trends and reducing short-term noise. Details of the sensitivity analysis are available at `https://github.com/baikeshe/GPR_Epi_Modeling.git`.

B. The Impact of Spread Data on Prediction Variance

We next propose an upper bound for the prediction variance to study the impact of data on predictions.

Lemma 1. *Consider a set of time steps* $T = \{t_1, t_2, \ldots, t_n\}$*. Let* $\mathbb{T}_r(t^*)$ *represent all points in* T *that lie within a ball of radius* r*, centered at the testing location* t^**, i.e.,* $\mathbb{T}_r(t^*) = \{t \in T : |t - t^*| \leq r\}$*. Consider the squared exponential kernel from* (2) *for the Gaussian Process model in Proposition 3. The posterior variance* $\sigma^2_\Delta(t^*)$ *at the testing time step* t^* *obeys the bound* $\sigma^2_\Delta(t^*) \leq \alpha^2 - \frac{\alpha^2}{\alpha^2 + \frac{\sigma^2}{|\mathbb{T}_r(t^*)|}}$.

The proof of Lemma 1 is provided in the Appendix [17]. Lemma 1 provides a data-dependent posterior variance bound on $\sigma^2_\Delta(t^*)$ at the testing time step t^* in terms of the number of the training data points around the testing location. The bound on the prediction variance given by Lemma 1 shows that more training data in $\mathbb{T}_r(t^*)$, captured by $|\mathbb{T}_r(t^*)|$, leads to a lower variance bound. Additionally, higher variance noise in the data, captured by σ^2, increases the bound on the posterior variance. In the absence of observation noise or with infinitely many data points near t^*, the variance bound approaches zero. Lemma 1 illustrates how the available data, the spreading trend, and noise can affect the posterior variance bound, and it solves Problem 2.

C. High Probability Error Bound on the Prediction

Having discussed the impact of data on the upper bound of the posterior variance, we now analyze the error bound on the prediction error. We first introduce the Lipschitz constant of the squared exponential kernel [21]. We consider the space of sample functions corresponding to the space of continuous functions on the time interval $\mathbb{T} \subset [t_a, t_b]$, where $t_a, t_b \in \mathbb{R}_{>0}$, such that the input batch $T \subset \mathbb{T}$.

Lemma 2. *[21, Corollary 8] Consider* $t_i, t_j \in \mathbb{T}$*. Then for all* t*,* $k(\cdot, t)$ *is* L_k*-Lipschitz, where* $L_k = \frac{\alpha^2}{\beta e^{1/2}}$.

We see that L_k is determined by the kernel parameters α and β in (2). In addition, consider our continuous unknown function $\bar{\Delta}$ from Remark 2, where $\bar{\Delta} : \mathbb{T}_{\geq 0} \to \mathbb{R}$ with Lipschitz constant $L_{\bar{\Delta}}$, such that $|\bar{\Delta}(t_1) - \bar{\Delta}(t_2)| \leq L_{\bar{\Delta}} |t_1 - t_2|$ for all $t_1, t_2 \in \mathbb{T}$. The Lipschitz continuity of $\bar{\Delta}(t)$ indicates that the change in spread over some time interval is limited by the length of that interval.

Recall the training dataset $\{(t_i, \Delta(t_i))\}_{i=1}^n$ where $t_i \in T$. We consider the training dataset to be within a specific time interval of interest, i.e., $T \subset \mathbb{T}$, and the testing point is also in this interval, i.e., $t^* \in \mathbb{T}$. The time interval of interest may be extensive, causing the testing location t^* to be far from any of the training points in the dataset T.

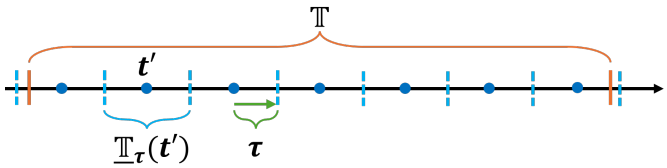

Fig. 3. A time interval $\mathbb{T}$ with $M(\tau, \mathbb{T}) = 7$.

To address this issue when studying the posterior mean, we consider a set of grid points $\mathbb{M}$ that are evenly distributed on $\mathbb{R}_{\geq 0}$. We define the radius associated with the point $t' \in \mathbb{M}$ as τ, such that $\mathbb{T}_\tau(t') := [t' - \tau, t' + \tau]$ for all $t' \in \mathbb{M}$. Then we consider $T \subset \mathbb{T} \subseteq \{t \mid t \in \bigcup_{t' \in \mathbb{M}} [t' - \tau, t' + \tau]\}$. To ensure that the union of the intervals of length 2τ can cover the continuous-time interval $\mathbb{T}$, the length of the interval between any two neighboring points in $\mathbb{M}$ must be smaller than 2τ. We define the covering number of $\mathbb{T}$, denoted by $M(\tau, \mathbb{T})$, as the cardinality of the set $\mathbb{M}$ with the minimum number of grid points to satisfy $\mathbb{T} \subseteq \{t \mid t \in \bigcup_{t' \in \mathbb{M}} [t' - \tau, t' + \tau]\}$. Figure 3 illustrates a time interval $\mathbb{T}$ with $M(\tau, \mathbb{T}) = 7$.

We derive the following theorem in part by using results from [22]. This result gives a high-probability bound on prediction accuracy, and it solves Problem 3.

Theorem 1. *Consider a zero-mean Gaussian process defined through the kernel $k(\cdot, \cdot)$ in (2) with Lipschitz constant L_k on time interval $\mathbb{T}$. Consider a continuous unknown function $\bar{\Delta} : \mathbb{T}_{\geq 0} \to \mathbb{R}$ that that captures the spreading process through the difference on the logarithmic scale of the infected cases, with Lipschitz constant $L_{\bar{\Delta}}$. The posterior mean of a Gaussian process conditioned on the training dataset $\{(t_i, \Delta(t_i))\}_{i=1}^n$, for all $t_i \in T$, at the testing time step t^* is given by $m_\Delta(t^*)$. Consider a set of grid points $\mathbb{M}$. If, for all $t^* \in \mathbb{T}$, there exists $t_i \in T$ and $t' \in \mathbb{M}$ such that $t_i \in \mathbb{T}_\tau(t')$ and $|t^* - t_i| \leq \tau$, then*

$$P\left(\left|\bar{\Delta}(t^*) - m_\Delta(t^*)\right| \leq \sqrt{\gamma_\delta(\tau)}\sigma_\Delta(t^*) + \xi_\delta(\tau)\right) \geq 1 - \delta,$$

for all $\delta \in (0, 1)$. Note that $\bar{\Delta}(t^*)$ is the noise-free variable at the testing time step t^*, and $m_\Delta(t^*)$ and $\sigma_\Delta(t^*)$ are the posterior mean and posterior standard deviation from Proposition 3, respectively. Further, we have

$$\gamma_\delta(\tau) = 2\log\left(\frac{\hat{\mathbb{T}}}{2\tau\delta} + \frac{1}{\delta}\right) \tag{3}$$

$$\xi_\delta(\tau) = (L_{\bar{\Delta}} + L_m)\tau + \sqrt{\gamma_\delta(\tau)L_{\sigma^2}\tau}, \tag{4}$$

where $L_{\bar{\Delta}}$ is the Lipschitz constant of the function $\bar{\Delta}$, $L_m = \frac{\alpha^2}{\beta e^{1/2}}\sqrt{n}\|(K(T,T) + \sigma^2 I_n)^{-1}\mathbf{\Delta}\|$ is the Lipschitz constant of the posterior mean function m_Δ, and $L_{\sigma^2} = \frac{2n\alpha^4}{\beta e^{1/2}}\left\|(K(T,T) + \sigma^2 I_n)^{-1}\right\|$ is the Lipschitz constant of the posterior variance function σ^2_Δ.

The proof of Theorem 1 is provided in the Appendix [17]. This theorem gives a high-probability error bound on the prediction error, which depends on the distance between training and testing points, captured by the length of the time interval $\hat{\mathbb{T}}$. A larger $\hat{\mathbb{T}}$ results in a larger error bound, and vice versa. The bound also shows that higher posterior variance at the testing time step $\sigma_\Delta(t^*)$ increases the error bound. Additionally, a lower Lipschitz constant of the spreading function $L_{\bar{\Delta}}$ reduces the error bound, as smaller changes in the spreading process correspond to a lower Lipschitz constant. While Proposition 3 provides an analytical solution for the posterior mean and variance of the GPR model, the bounds from Lemma 1 and Theorem 1 offer valuable insights for selecting datasets and data collection methods, helping to improve modeling outcomes.

Example 2. *Consider the dataset representing the difference on the logarithmic scale of the infected cases in the UK, shown by the solid red line in Figure 2. As in Example 1, the dataset is preprocessed using a thirty-day rolling average, with η set to 7. Then, we use a moving window of 30 days of data to train the GPR model and predict the posterior mean for the subsequent 20 days. For example, data from April 6^{th} to May 6^{th}, 2022, is used to predict the spread from May 7^{th} to May 27^{th}, 2022. In Figure 4, the solid red line represents the posterior mean of the predictions, with the shaded red area showing the 95% confidence interval. The dotted yellow line represents the noisy data. Around 94.29% of noisy data are located within the 95% confidence interval. Most predictions that fail to include the data within the 95% confidence interval arise from either abrupt changes in the spreading trend and/or the limited number of available data samples. This figure illustrates that the prediction captures the trend of the spread when compared to the noisy observations. Note that the confidence interval represents the uncertainty in predicting the true value, which is not directly available* [2].

We compare the GPR prediction results with polynomial regression, KNN, and neural networks using adjusted parameters to prevent overfitting, given the prediction task's smaller dataset (30 training points predicting 20 future points). For polynomial regression with degree 3, the MSE is 0.0183, with 72.00% of the testing data within the 95% confidence intervals, represented by a dashed blue line with blue-shaded areas. The KNN model ($\kappa = 3$) achieves an MSE of 0.0046, with 51.00% of data within the intervals, shown as a dotted purple line with purple-shaded areas. A neural network ($50-25-10$ neurons) yields an MSE of 0.0070, with 43.67% of data within the intervals, represented by a dash-dotted green line with green-shaded areas.

To address overfitting, we reduce model complexity compared to Example 1. Polynomial regression and KNN models previously optimized for higher orders (20 and 15, respectively) led to overfitting in prediction tasks, necessitating order reduction. Similarly, the neural network's architecture was simplified, and sinusoidal feature augmentation was excluded to suit the smaller dataset lacking periodic patterns. Despite these adjustments, the GPR model remains superior, capturing the spread without parameter tuning and achieving higher proportions of data within the 95% confidence intervals. In contrast, the other methods, while achieving MSEs of similar

[2] A 30-day moving average of the data provides an optimal balance for data preprocessing, capturing periodic trends while filtering short-term noise. Shorter windows (1, 3, 5 days) retain detail but risk overfitting, while moderate windows (10, 20 days) struggle to capture both short- and long-term dynamics. Overly long windows (50 days) oversmooth the data, reducing model adaptability. Detailed sensitivity analysis and insights on smoothing window effects on GPR prediction accuracy are available at `https://github.com/baikeshe/GPR_Epi_Modeling.git`.

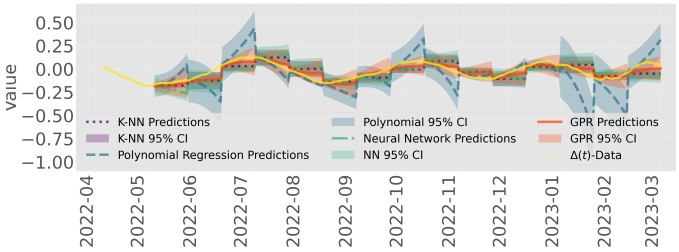

Fig. 4. Posterior mean of the prediction on spreading trend. Since we lack additional information when modeling the spread, such as periodic spreading patterns, we observe lower prediction accuracy when $\boldsymbol{\Delta}$ shifts from increasing to decreasing, or vice versa.

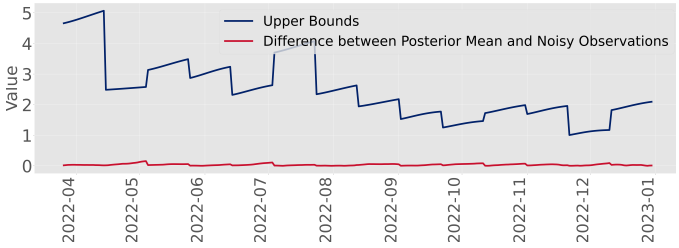

Fig. 5. Probability upper bound with $\boldsymbol{\delta = 0.05}$. The upper bound changes depending on how far into the future the prediction length extends. For the same twenty-day period, the further into the future, the larger the prediction error bound will be.

magnitude, require significant parameter adjustments, making them less robust at capturing the complex dynamics of disease spread with limited data.

In addition to comparing to other prediction models, Figure 5 illustrates the difference between the posterior mean and the noisy observations at each prediction, shown by the solid red line. The upper bound, computed based on Theorem 1, is plotted in blue. We assume that $M(\mathbb{T}, \tau) = 5$*, where* $\tau = 5$ *and* $\hat{\mathbb{T}} = 50$*. We assume that the Lipschitz constant of the function* Δ *is* $L_{\bar{\Delta}} = 0.01$ *and* $\delta = 0.05$*. We observe that, during each twenty-day prediction period, the bound increases monotonically as the prediction time steps move further away from the training data. By comparing the bound from April 2022 to September 2022 with the bound from September 2022 to March 2023, we find that the bound is looser during periods of significant changes in the spread (April 2022 to September 2022), leading to a more conservative bound. Conversely, when the spread is more consistent (September 2022 to March 2023), the bound tightens. The bound is generally larger when considering a longer time interval into the future. The bound can guide us in selecting appropriate datasets and data collection methods to improve prediction results.*

IV. Conclusion

In this work, we propose an approach to modeling epidemic spreading processes by using Gaussian process regression. We model and predict the spread through the difference on the logarithmic scale of the infected cases. We provide an upper bound on the posterior variance and mean error, highlighting the impact of the spreading trend and available data. Additionally, we discuss the impact of data preprocessing on modeling and prediction performance. We further highlight the benefits of using GPR methods by empirically comparing them with other modeling and prediction approaches, including polynomial regression, KNN, and neural networks. In future work, we plan to utilize this model and prediction mechanism to design a new data-driven predictive control strategy for epidemic mitigation.

References

[1] G. Giordano, F. Blanchini, R. Bruno, P. Colaneri, A. Di Filippo, A. Di Matteo, and M. Colaneri, "Modelling the COVID-19 epidemic and implementation of population-wide interventions in Italy," *Nature Medicine*, vol. 26, no. 6, pp. 855–860, 2020.

[2] M. Woolhouse, "How to make predictions about future infectious disease risks," *Philosophical Transactions of the Royal Society B: Biological Sciences*, vol. 366, no. 1573, pp. 2045–2054, 2011.

[3] B. She, S. Sundaram, and P. E. Paré, "A learning-based model predictive control framework for real-time SIR epidemic mitigation," in *Proc. of the American Contr. Conf.*, 2022, pp. 2565 –2570.

[4] B. She, S. Sundaram, and P. E. Paré, "Optimal mitigation of SIR epidemics under model uncertainty," in *Proc. of the 2022 IEEE Conf. on Deci. and Con. (CDC)*. IEEE, 2022, pp. 4333–4338.

[5] R. Senanayake, S. O'Callaghan, and F. Ramos, "Predicting spatio-temporal propagation of seasonal influenza using variational Gaussian process regression," in *Proc. of the AAAI Conf. on Arti. Intel.*, vol. 30, 2016.

[6] R. M. A. Velásquez and J. V. M. Lara, "Forecast and evaluation of COVID-19 spreading in USA with reduced-space Gaussian process regression," *Chaos, Solitons & Fractals*, vol. 136, p. 109924, 2020.

[7] C. Zimmer and R. Yaesoubi, "Influenza forecasting framework based on Gaussian Processes," in *Int. Conf. on Mach. Lea.* PMLR, 2020, pp. 11 671–11 679.

[8] L. Gonçalves, M. A. A. Turkman, C. Geraldes, T. A. Marques, and L. Sousa, "COVID-19: Nothing is normal in this pandemic," *Journal of Epidemiology and Global Health*, vol. 11, no. 2, pp. 146–149, 2021.

[9] C. K. Williams and C. E. Rasmussen, *Gaussian Processes For Machine Learning*. MIT press Cambridge, MA, 2006, vol. 2, no. 3.

[10] M. G. Genton, "Classes of kernels for machine learning: A statistics perspective," *J. of Mach. Lear. Res.*, vol. 2, no. Dec, pp. 299–312, 2001.

[11] S. Bhatt, N. Ferguson, S. Flaxman, A. Gandy, S. Mishra, and J. A. Scott, "Semi-mechanistic Bayesian modelling of COVID-19 with renewal processes," *J. of the Roy. Stat. Soci. Ser. A: Stat. in Soci.*, vol. 186, no. 4, pp. 601–615, 2023.

[12] J. Shao, *Mathematical Statistics*. Springer Science & Business Media, 2008.

[13] G. W. Oehlert, "A note on the delta method," *The American Statistician*, vol. 46, no. 1, pp. 27–29, 1992.

[14] J. M. Bland and D. G. Altman, "Statistics notes: The use of transformation when comparing two means," *BMJ*, vol. 312, no. 7039, p. 1153, 1996.

[15] N. I. Bosse, S. Abbott, A. Cori, E. van Leeuwen, J. Bracher, and S. Funk, "Scoring epidemiological forecasts on transformed scales," *PLoS Comp. Bio.*, vol. 19, no. 8, p. e1011393, 2023.

[16] Sam Abbott, Joel Hellewell, Katharine Sherratt, Katelyn Gostic, Joe Hickson, Hamada S. Badr, Michael DeWitt, Robin Thompson, EpiForecasts, and Sebastian Funk, *EpiNow2: Estimate Real-Time Case Counts and Time-Varying Epidemiological Parameters*, 2020.

[17] B. She, L. Xin, P. E. Paré, and M. Hale, "Modeling and Predicting Epidemic Spread: A Gaussian Process Regression Approach," *arXiv preprint arXiv:2312.09384*, 2023.

[18] P. van den Driessche, "Reproduction numbers of infectious disease models," *Inf. Dis. Mod.*, vol. 2, no. 3, pp. 288–303, 2017.

[19] S. Abbott, J. Hellewell, R. N. Thompson, K. Sherratt, H. P. Gibbs, N. I. Bosse, J. D. Munday, S. Meakin, E. L. Doughty, J. Y. Chun, *et al.*, "Estimating the time-varying reproduction number of SARS-CoV-2 using national and subnational case counts," *Wellcome Open Research*, vol. 5, no. 112, p. 112, 2020.

[20] E. Mathieu, H. Ritchie, L. Rodés-Guirao, C. Appel, C. Giattino, J. Hasell, B. Macdonald, S. Dattani, D. Beltekian, E. Ortiz-Ospina, and M. Roser, "Coronavirus pandemic (COVID-19)," *Our World in Data*, 2020, https://ourworldindata.org/coronavirus.

[21] A. Lederer, J. Umlauft, and S. Hirche, "Uniform error and posterior variance bounds for Gaussian process regression with application to safe control," *arXiv preprint arXiv:2101.05328*, 2021.

[22] ——, "Uniform error bounds for Gaussian process regression with application to safe control," *Proc. NeurIPS*, vol. 32, 2019.