Contents lists available at ScienceDirect

Epidemics

journal homepage: www.elsevier.com/locate/epidemics

A semi-parametric mixed model for short-term projection of daily COVID-19 incidence in Canada

Muhammad Abu Shadeque Mullah [*], Ping Yan

Public Health Agency of Canada, Canada

ARTICLE INFO

Keywords:
Poisson–Gamma mixture
Penalized splines
Generalized linear mixed effects model
Semi-parametric mixed model
Bayesian Markov chain Monte Carlo

ABSTRACT

During a pandemic, data are very “noisy” with enormous amounts of local variation in daily counts, compared with any rapid changes in trend. Accurately characterizing the trends and reliable predictions on future trajectories are important for planning and public situation awareness. We describe a semi-parametric statistical model that is used for short-term predictions of daily counts of cases and deaths due to COVID-19 in Canada, which are routinely disseminated to the public by Public Health Agency of Canada. The main focus of the paper is the presentation of the model. Performance indicators of our model are defined and then evaluated through extensive sensitivity analyses. We also compare our model with other commonly used models such as generalizations of logistic models for similar purposes. The proposed model is shown to describe the historical trend very well with excellent ability to predict the short-term trajectory.

1. Introduction

Since the beginning of COVID-19 pandemic, predictive models have been applied at all levels of governments in Canada and worldwide. Public health officials from various jurisdictions regularly disseminate model based predictions to the public through the media in order to raise public situation awareness. The Public Health Agency of Canada (PHAC) releases approximately every month of long-term and short-term forecasts.

The models and methods for the long-term forecasts (Ogden et al., 2020) are dynamic, driven by assumptions at the levels of individuals including the adherence of public health measures, assumptions at the level of the virus as well as assumptions at the population level regarding the dynamic states of the system. The two sub-models in Ogden et al. (2020) are a deterministic compartmental model (Baily, 1975; Ludwige et al., 2020) and an agent-based model (Borshchev, 2016). Both sub-models are strategic and conceptual for planning purposes, to produce “what-if” scenarios, with respect to what must happen and what might happen, with respect to the trajectory of the epidemic according to the time of transmission.

Unlike in the long-term forecast models, parameters in the short-term forecasting models do not carry specific biological, environmental, social and behavioural meanings. They simply describe data features based on reported cases, such as the initial growth rate, inflexion point, dampening factor, turning points (e.g. knots), and so on. The short-term forecasting models are empirical. The focus is on their abilities to describe both the historical trajectory and the temporal variations based on established goodness-of-fit criteria. Provided that data generating mechanisms are well understood (e.g. case-definition, how data are organized and reported, reporting delays and under-reporting, retrospective ascertainment of time of events, etc.) and that data generating mechanisms remain relatively stable, fitting these empirical models tend to provide reliable predictions, as long as the time horizon is not too far into the future, because historical data carry no information on the future of societal behaviour, virus evolution and policy changes. A review of the modelling numbers the PHAC releases approximately every month shows its predictions have been within range of the actual results most of the time (Anon, 2020a,b).

The empirical models can be fully parametric or semi-parametric. The former is typically modelled using sigmoid growth-curve functions fitted to the cumulative case counts and extrapolated into the near future. This approach has been extensively described in the textbook (Yan and Chowell, 2019), published on the Eve of the COVID-19 global pandemic, with step-by-step examples based case counts on Zika disease. As soon as the COVID-19 outbreaks started, short-term forecasts based on these models were immediately applied to early data emerging in China (Roosa et al., 2020a,b). In Canada, short-term forecasts based on the same type of growth-curve models was included as part of the PHAC public releases since April 2020, with supporting document (Smith et al., 2021).

This paper focuses on semi-parametric methods developed at PHAC during the course of the COVID-19 pandemic to augment the performance of the existing short-term forecasting for cumulative cases

[*] Correspondence to: Public Helath Agency of Canada (PHAC), 130 Colonnade Road, Ottawa, Ontario, K1A 0K9, Canada.
E-mail address: muhammad.mullah@phac-aspc.gc.ca (M.A.S. Mullah).

https://doi.org/10.1016/j.epidem.2022.100537
Received 27 May 2021; Received in revised form 19 November 2021; Accepted 5 January 2022
Available online 19 January 2022

and deaths, and to forecast the daily incidence cases rather than ever growing cumulative numbers.

Data are very "noisy" with enormous amounts of "local" variation in daily counts, compared with any rapid changes in trend. A popular way of describing the trend, without the models, is measured with 7-day running averages as plotted in many dashboards. The 7-day running averages smooth out variations due to weekly effects such as reporting patterns during weekdays versus on weekends. They still have substantial local variability due to other abnormalities such as computer glitches, catching up previously under-reported data and effects of long holiday seasons. More importantly, the 7-day running averages are defined before the end of the data series and cannot be extended for prediction purposes.

These empirical models are first used to fit case count data. The semi-parametric model described in this paper, namely penalized splines (Ruppert, 2002; Ruppert et al., 2003), was initially designed to fit the model to describe the trend without over-fitting, which can cause near interpolation of the data. In the figures presented later in the paper, we shall see that the resulting fitted values are smooth trend functions and the representation of the historical trend is the usefulness of the model in its own right. Meanwhile, the statistically estimated credible intervals are wide enough to accommodate local variability in the 7-day running averages, but also narrow enough to exclude apparent outliers in daily counts. This is a good balance between the robustness and precision of the modelled trend.

Once a best fitted model is established with estimated coefficients, short-term predictions are produced by extrapolation using these estimated parameters. They should be interpreted as the prediction of the running averages into the near future along with the prediction intervals of the running averages. Since the future trajectory not only depends on the historical trajectory in the recent past, but also, and more importantly, depends on the transmission environment, people's behaviour, public health measures, as well as surveillance and reporting practices in the future. Therefore, we limit to 10-day as short-term forecasting.

2. Material and methods

2.1. The proposed model

2.1.1. Poisson-Gamma mixture model with penalized splines

Let Y_t denote the number of incident cases on day t. To describe the trajectories of the outbreak, we consider a Poisson–Gamma mixture model (Lawless, 1987) that can accommodates the latent heterogeneity in the count data. We use the log-linear form, with mean and variance of Y_t given t as

$$\mathbb{E}[Y_t \mid t] = \mu(t, \underline{\theta}),$$
$$\mathbb{V}[Y_t \mid t] = \mu(t, \underline{\theta}) + \kappa\{\mu(t, \underline{\theta})\}^2,$$

where $\underline{\theta}$ is the vector of parameters associated with the mean function and $\kappa > 0$ is the over-dispersion parameter. The above marginal specification of the mean and variance addresses the random component of the data generation mechanism, which tends to be over-dispersed. The mean function $\mu(t, \underline{\theta})$ models the systematic trend, which can be either fully parametric or semi-parametric.

There exist a number of parametric models most of which are nested within the generalized-Richards model (see, Yan and Chowell, 2019 for details). Considering $\mu(t, \underline{\theta})$ is the difference of the cumulative function $C(t)$, the generalized-Richards model is given by

$$\frac{dt}{d}C(t) = rC(t)^p\left[1 - \left(\frac{C(t)}{N}\right)^{\gamma}\right] \tag{1}$$

with four parameters (r, N, p, γ) : $r > 0$ is the growth rate, $N = \lim_{t\to\infty} C(t) > 0$ is the final epidemic size and carrying capacity, $p \in (0, 1)$ is the growth scaling parameter and $\gamma > 0$ is the shape parameter that measures the deviation from symmetry of the S-shaped standard logistic curve. This model (1) includes special cases: (i) exponential growth ($p = 1, N \to \infty$); (ii) sub-exponential growth ($0 < p < 1, N \to \infty$); (iii) logistic ($p = \gamma = 1$); (iv) Richards ($p = 1, 0 < \gamma < 1$). During different phases of the pandemic, special cases of the generalized-Richards model were assuming to describe "local data features" of the cumulative counts.

The parametric model (1) has limited ability to capture complex shapes of the disease trajectory because of the restricted sigmoid shape of $C(t)$. It may be applied locally to a specific section of the case count data during which the daily case counts $C(t) - C(t-1)$ can be approximated by a "single wave". This brings two difficulties. First, the choices of number of data points used from the recent past determine the numbers of parameters that can be estimated and affect the future predicted trajectories. Furthermore, when daily cases during the recent period of time represent a "trough" so that the cumulative numbers $C(t)$ first increase in a concave fashion followed by a turning point to become convex, the parametric model (1) will not only fail to provide good fit to data, but also, in terms of using algorithms to search for the best fit, the algorithms may fail to converge. For these reasons, we consider the following semi-parametric approach.

For modelling the daily COVID-19 incidence in Canada, we estimate the systematic trend semi-parametrically. In particular, we consider

$$\log\{\mu(t, \underline{\theta})\} = m(t, \underline{\theta}), \tag{2}$$

where $m(t, \underline{\theta})$ is a smooth function that is centred and twice-differentiable. We model the time dependent intensity function $m(t, \underline{\theta})$ by using the thin plate regression splines of Wood (2003). One key advantage of this splines is that it avoids the knot placement problems of conventional regression spline modelling. Thin plate regression splines are constructed by placing knots at every unique value of the covariate t to form the basis for a full thin plate spline and then truncating this basis in an optimal manner, to obtain a low rank smoother. This is done by eigen-decomposition of the full basis and keeping only the K eigenvectors associated with the K largest eigen values to form a new optimal basis. Considering a large dimension of the basis (K, hereinafter referred to as knots), we represent $m(t, \underline{\theta})$ as

$$m(t, \underline{\theta}) = \sum_{j=1}^{K} \beta_j b_j(t) = B(t)\underline{\theta}, \tag{3}$$

where $\underline{\theta} = (\beta_1, \ldots, \beta_K)$ is the vector of regression coefficients that represent changes in slope from one segment to the next and $B(t) = (b_1(t), \ldots, b_K(t))$ is the design matrix that contains thin plate regression spline basis functions. The un-constraint estimation of β_j would lead to a "overly fluctuating" fit due to the large number of knots. To prevent over-fitting, we impose the restriction

$$\underline{\theta}^T S \underline{\theta} \leq c$$

for some nonnegative constant c, in which S is a positive semi-definite penalty matrix. This restricted smoothing approach is known as penalized splines (P-splines). To ensure identifiability, we construct the smooth by absorbing the centring constraint into the basis function so that $\sum_{i=1}^{n} m(t_i, \underline{\theta}) = 0$, where n is the number of data points of the time-series at the time of the analysis. More specifically, we choose $\underline{\theta}$ and $B(t)$ such that $\mathbb{1}^T B(t)\underline{\theta} = 0$, where $\mathbb{1}$ is a vector of all 1s.

2.1.2. The penalized splines as mixed model

Generalized Linear Mixed Models (GLMM) which is typically used for correlated data analysis can also be used for curve fitting. P-splines can be viewed as a particular case of the GLMMs. To achieve a smooth function, we can use the GLMM to shrink the regression coefficients of knot points towards zero, by including them as random effects. Following Wood (2004), we re-parametrize $m(t, \underline{\theta})$ (3) in terms of a fixed effects parameter vector $\underline{\theta}_F$ and a random effects $\underline{\theta}_R$ as

$$m(t, \underline{\theta}) = B_F(t)\underline{\theta}_F + B_R(t)\underline{\theta}_R, \tag{4}$$

where $\underline{\theta}_R \sim N(0, \sigma^2 I)$ with smoothing parameter $\lambda = 1/\sigma^2$, $B_F(t)$ are the columns of $B(t)$ for which the penalty matrix S has zero eigenvalues, and $B_R(t) = B(t)U\sqrt{D_+^{-1}}$ in which U is the matrix containing eigenvectors of S corresponding to E, the strictly positive eigenvalues arranged in descending order of magnitude; and D_+ is the diagonal matrix containing E on the leading diagonal. Substituting (4) in (2), we have a Poisson–Gamma mixture model with mixed effects which we call Gamma–Poisson semi-parametric mixed model (SPMM) as the smooth function $m(t, \underline{\theta})$ is represented by a semi-parametric function.

The Poisson–Gamma SPMM is capable of fitting flexible shapes. It is capable to fit well to the entire data series with finite number of parameters. It will be shown that, once the number of knots (i.e., basis dimension) reaches $K = 25$, adding more knots gain little in terms of the goodness-of-fit to data although the future trajectory of the daily incidence counts is sensitive to the choice of K; and that, fitting the entire data series from the beginning of the pandemic yields more precision in terms of the short-term predictions.

2.1.3. Prediction

Given the estimated coefficients $\hat{\underline{\theta}}_F$ and $\hat{\underline{\theta}}_R$, we evaluate $B_F(t)$ and $B_R(t)$ for the prediction data t_{new} by identifying the knot locations that define the basis functions for the original data t. We then obtain the predicted values as

$$\hat{m}(t, \underline{\theta}) = B_F(t_{new})\hat{\underline{\theta}}_F + B_R(t_{new})\hat{\underline{\theta}}_R. \tag{5}$$

2.2. Estimation

We considered the Bayesian estimation for model fitting due to at least two potential problems of using the likelihood-based estimation as noted in Crainiceanu et al. (2005). First, the likelihood of the Poisson–Gamma SPMM is a high dimensional integral over the random effects that does not have a closed form and need to be approximated. This approximation can sometimes have a considerable effect on the parameter estimation. Second, the confidence intervals are obtained by replacing the estimated parameters instead of the true parameters and ignoring the inherent additional variability. This results in narrower (than they should be) confidence intervals. We, therefore, adopted a Bayesian approach to fit the models.

We used the following non-informative priors for the parameters used in our model:

$$\begin{aligned} \beta_j &\sim \text{Normal}(0, 10^6) \quad \text{for all fixed effects} \\ \sigma &\sim \text{Uniform}(0, 10^3) \\ 1/\kappa &\sim \text{Uniform}(0, 500). \end{aligned}$$

The number of knots K was chosen following a simple rule (adapted from Wand, 2003)

$$K = \max(5, \min(n/5, 25)). \tag{6}$$

While several software platforms (such as WinBUGS, OpenBUGS, JAGS, INLA, STAN) are now available for model fitting via MCMC sampling, we used JAGS (Just Another Gibbs Sampler) (Plummer, 2009) to fit Bayesian models. JAGS is a mature and declarative language for Bayesian model fitting with reasonable computation time and a nice link to `R`. We call JAGS from inside of `R`using the `R` package `R2jags` (Su and Yajima, 2012) and export results to `R`.

The Bayesian estimates were medians from 350,000 iterations of the MCMC algorithm after discarding the first 150,000 iterations as burn-in. We ran two chains and thinned them by keeping every 200th iteration. The 95% posterior credible interval for each parameter of interest was obtained as 2.5th and 97.5th percentiles of the posterior sample. We evaluated convergence of the chains by visually examining the trace plot, density plot, sample autocorrelation function for each parameter, and also following Gelman and Rubin (1992) to quantify the between-chain and the within-chain variability of a quantity of interest.

Note that the Gamma mixture onto the mean of Poisson was not intended to frame as a gamma prior. Rather, the Poisson–Gamma was used as a model to describe the random component of the data (i.e., over-dispersion) whereas the Bayesian priors were applied to the parameters associated with the systematic component of the trend and hyper-parameters. With this framework of the Poisson–Gamma mixture to represent the random component and the Bayesian approach to fit the model, we can actually achieve two objectives: (i) increase the accuracy; and (ii) accommodate over-dispersion.

2.3. Evaluation of model performance

We assessed the projection accuracy of the proposed Poisson-Gamma SPMM model using a retrospective analysis. To evaluate both the goodness of the model fit and its ability to short-term forecasts, we used the following criteria: (i) Bayesian information criteria (BIC); (ii) deviance information criteria (DIC); (iii) mean absolute error (MAE); (iv) root mean squared error (rMSE); (v) average coverage probabilities (ACP) of the 95% prediction intervals (PI); and (vi) mean interval score (MIS).

The BIC and DIC are defined as

$$\begin{aligned} \text{BIC} &= D + k\log(n), \\ \text{DIC} &= \bar{D} + p_D, \end{aligned}$$

where $D = -2\log(\hat{L})$ is the deviance in which $\hat{L}$ is the maximized value of the likelihood function, k is the number of parameters estimated by the model, $\bar{D} = E(D)$ is the ‘posterior mean deviance’ and p_D is the ‘effective number of parameters’ defined as the posterior mean deviance minus deviance evaluated at the posterior mean of the parameters (see, Schwarz, 1978; Speigelhalter et al., 2002 for details).

The MAE and rMSE were computed as

$$\text{MAE} = \frac{1}{n}\sum_{t=1}^{n} |\hat{y}_t - y_t|,$$

$$\text{rMSE} = \sqrt{\frac{1}{n}\sum_{t=1}^{n}\left(\hat{y}_t - y_t\right)^2},$$

where y_t is the observed incident cases on day t and $\hat{y}_t$ is the corresponding predicted cases.

The average coverage probability of the 95% PI (i.e., the proportion of the observed incidences that fall within the 95% PI) was computed

$$\text{ACP} = \frac{1}{h}\sum_{t=1}^{h} \mathbb{1}\left(\hat{L}_t < y_t < \hat{U}_t\right),$$

where $\hat{L}_t$ and $\hat{U}_t$ are the lower and upper limits of the 95% point-wise PI, $\mathbb{1}(.)$ is an indicator function, and h is the number of additional days for which forecasting has been made. The mean interval score (MIS) which is a measure of the width of the 95% PI as well as the coverage was defined as

$$\begin{aligned} \text{MIS} = \frac{1}{h}\sum_{t=1}^{h} \Big[&\left(\hat{U}_t - \hat{L}_t\right) + \frac{2}{\alpha}\left(\hat{L}_t - y_t\right)\mathbb{1}\left(y_t < \hat{L}_t\right) \\ &+ \frac{2}{\alpha}\left(y_t - \hat{U}_t\right)\mathbb{1}\left(y_t > \hat{U}_t\right)\Big] \end{aligned}$$

with the type I error rate, $\alpha = 0.05$. The MIS penalizes both for wider prediction interval and also for the observations that fall outside the 95% PI (Gneiting and Raftery, 2007).

The BIC and DIC measure both the ‘goodness of fit’ and ‘complexity’ of the model. We used BIC for comparing the Bayesian SPMM to other commonly used frequentist methods, and DIC for Bayesian model comparison. The MAE and rMSE measure the closeness of the estimated trajectories to the observed cases whereas the MIS and ACP assess the uncertainty of the predictions.

3. Analysis of COVID-19 data for Canada

The proposed model was fitted to reported daily incidence data from March 10, 2020 to March 05, 2021. The day March 10, 2020 was selected as starting point because on that day the reported total cases in Canada was nearly 100, and we assumed that as a proxy signal for community transmission. The forecasting was performed every week (every Saturday, starting from March 21) to predict for the next 10 days using all the past data that ended on Fridays. We chose each Friday as the end point because some provinces do not report numbers on Saturdays and Sundays. More specifically, the first 10-day projection was made on March 21 (Projection Week-1, PW-1) using data from March 10 to March 20 predicting March 21 to March 30, a subsequent projection was made on March 28 (PW-2) using data from March 10 to March 27 predicting March 28 through April 06, and next projection was made on April 04 using data from March 10–April 03 forecasting April 04–April 13, and so on using all the past days data. The last projection week is referred as PW-51 and considered data reported from March 10, 2020 to March 05, 2021. Details on the calibration period and forecasting horizons are shown in the Supplementary Table (Table $S1$). The daily number of new COVID-19 cases data were collected from http://www.covid-19canada.com.

3.1. Results

Figs. 1 and 2 show the plots of projected incidence produced by the Poisson–Gamma SPMM at different phases of the COVID-19 pandemic in Canada with 95% credible intervals that are extended into the predictions (prediction intervals). Results from model fits generated in every week during the pandemic are summarized in Supplementary Table $S1$. The SPMM model nicely captured the trajectory of the outbreak in all phases. The fitted curve indicated a sharp decline in COVID-19 cases in Canada since January 2021 and predicted (expected) 3413 incident cases (95% prediction interval: (2662, 4342)) to be reported on March 15, 2021.

The 7-day running averages were also plotted in Fig. 2 along with the model fitted curve. Both the fitted function and the 7-day running average described the trend very well, and the extension of the fitting function into the near future could be viewed as the prediction of the trajectory of the moving average. The moving average still carries substantial fluctuation, and is especially affected by irregularities in reporting of daily data, such as during the holiday seasons (e.g. Fig. 2).

The shaded region (in Figs. 1 and 2) represents the 95% credible interval that make an "envelope" to incorporate the variations of the data, as well as the prediction intervals into the future. If the reported incidence cases since the day of forecasting stay within the envelope, it implies the model is performing as expected and data generated by the epidemic and reporting mechanisms are as expected. In that context, our proposed model captured the majority of future incident cases in all phases. The percentage coverage of the 95% prediction interval was more than 80% in most cases and varied between 60% and 100% during the pandemic (see, Supplementary Table $S1$). Note that the prediction intervals entirely contained the variability in the future 7-day running averages (e.g., Fig. 2).

There were wide ranges of the 95% prediction limits and the average ranges (average coverage length) varied between 244.67 and 3744.77 (see, Supplementary Table $S1$). With regards to the forecasting performance indicator rMSE, the model appears to perform better during mid-May and early-June in 2020, i.e., when Canada's incidence numbers were declining in the course of first wave.

The over-dispersion parameter κ was estimated implicitly from the model along with the other parameters and was positive for all projection weeks that varied between 0.01 and 0.4. The κ has sizeable influence on the curve fitting. We illustrated this in Fig. 4 of the following sensitivity analysis.

3.2. Sensitivity analysis

We assess the sensitivity of the proposed Poisson–Gamma SPMM model to the (i) number of past data points used for model fitting; (ii) magnitude of the over-dispersion parameter κ; and (iii) number of knots K. To that end, we chose a projection week in December 2020 (specifically, PW-38: December 12, 2020) and fit the model by (i) using all previous data (from March 10 through December 11) vs. using only recent past 20 days data (from November 21 to December 11); (ii) varying κ: 0, 0.1 and 0.5; and (iii) varying K: 10 knots vs. 20 knots vs. 25 knots vs. 30 knots vs. 35 knots. For PW-38, the model estimated κ was 0.1. The other two choices of κ were to explore how the results differ when $\kappa = 0$ or when a higher value of κ is considered (e.g., $\kappa = 0.5$). For $\kappa = 0$ and 0.5, the curve fitting was performed assuming and treating κ as fixed parameter.

Results from the sensitivity analyses are shown in Figs. 3–5 and Table 1. Clearly, the model is sensitive to the number of previous data points used for model fitting, dispersion parameter κ as well as to the choice of knots. The forecasting trends varied based on the number of previous data points used for model fitting and using the recent past 20 days data yielded a wider prediction interval compared to using all previous data points (Fig. 3). Also, as evident from Fig. 4, different values of κ led to different trajectories (Fig. 4-a) and for larger κ the prediction intervals were wider (Fig. 4-b). Further, it is apparent in Fig. 5-a that different choices of knots led to different forecast trajectories. However, the best predict ability of the model was observed when $K = 25$ with minimum rMSE and maximum coverage probability (Table 1). In general, larger number of knots (K) was associated with higher coverage and narrower prediction interval (Table 1 and Fig. 5-b).

3.3. Comparison to other methods

We compare the performance of our Poisson–Gamma SPMM to the well known phenomenological models (logistic growth model and generalized-Richards model) as well as to the cubic polynomial model fitted to the data within three different phases of the pandemic: (i) March 10–September 25 that covers a period when daily data series showed recurrence waves; (ii) March 10–June 12 covering a bell shaped single web period; and (iii) April 26–September 25 covering a small segment when the daily incidence looks like a U-shape. In all of these scenarios, the SPMM consistently outperformed the other simpler models both in curve fitting and short term forecasting based on the all performance metrics as shown in Table 2. For scenario (i), the SPMM successfully captured the multiple-wave pattern of the outbreak (Fig. 6-a). In contrast, the logistic and the generalized Richards models were unable to reproduce the recurrence waves, yielding poor performance (Fig. 6-a and Table 2). For scenario (ii), all the models performed well in capturing the single wave pattern of the outbreak (Fig. 6-b) but nonetheless the SPMM performed best in terms of the all performance indicators. For scenario (iii), the logistic and generalized-Richards models completely fail to capture the U-shape pattern of the data series whereas the SPMM reproduced it nicely (Fig. 6-c). Because the logistic and the generalized Richards models conform to the specific trend shape (bell shape for daily incidence and S-shape for the cumulative), they are not flexible enough to deal with recurrence waves in a long period of time series or to capture a U-shape pattern of the trend.

It is also worthwhile to discuss and compare our SPMM with the sub-epidemic models (Chowell et al., 2019). The sub-epidemic models are sequential of sectional (generalized) logistic functions with smoothed connectors. They are confined with finite number of parameters which do not increase with data points. The sub-epidemic models are able to fit to long period of time series with recurrent waves. Our SPMM model share the same features of the sub-epidemic model with almost the same number of parameters; they can be viewed as sequential polynomial models with smoothed connectors.

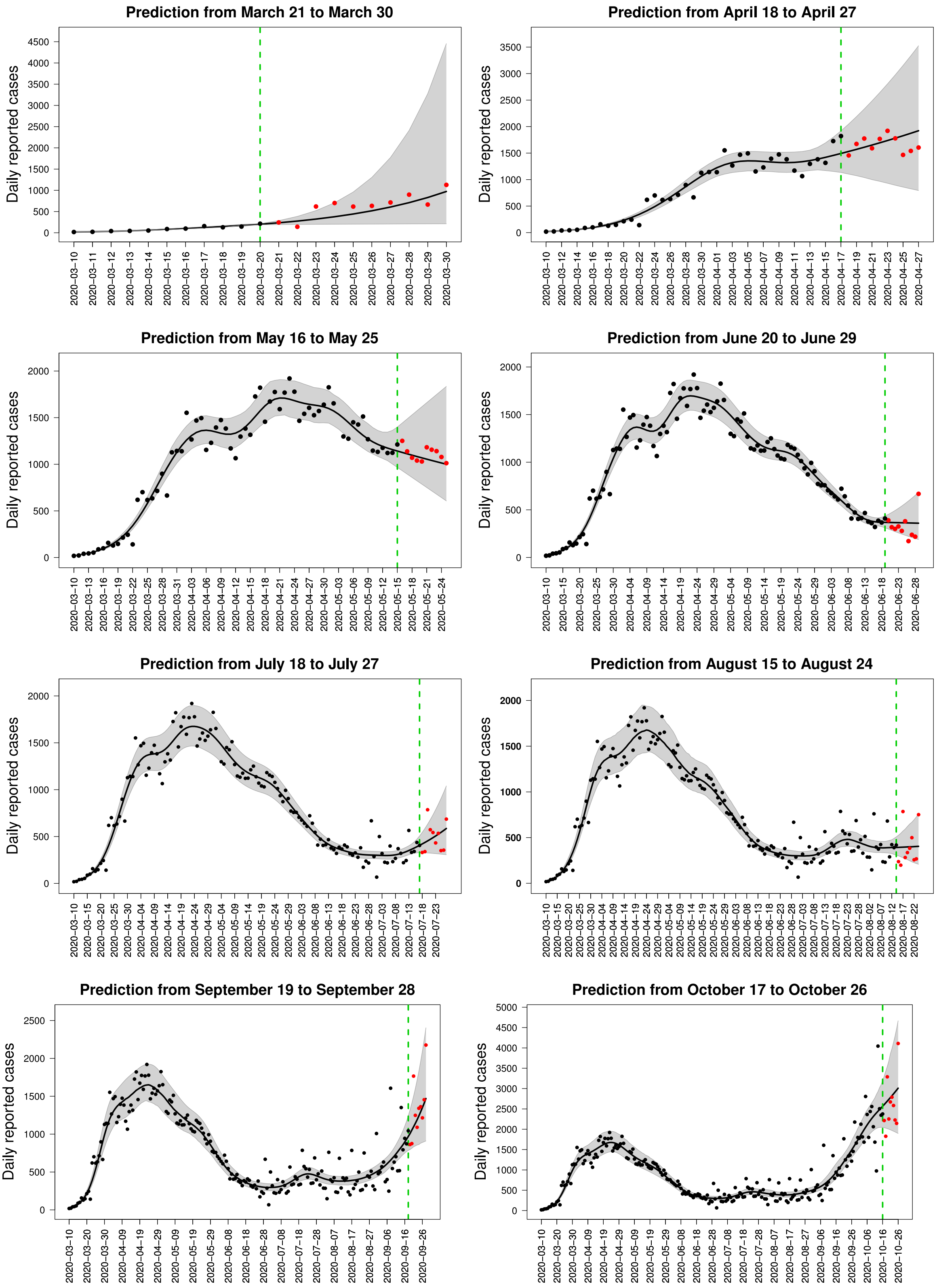

Fig. 1. Ten-day ahead forecasts of daily incident cases by the Poisson–Gamma SPMM model at various phases of the COVID-19 outbreak in Canada. Projections were made using all data reported since March 10, 2020. The black dots at the left of the green vertical line are actual daily incidence data that were used for model fitting. The black line is the model fitted (expected) values going through the dots; this line extends into the future (for next 10 days) as prediction. The red dots at the right of the green vertical line correspond to the daily incident cases reported since the projections. The shaded light-grey region is the 95% credible interval that are extended into the prediction; and the green vertical dashed line indicates the start time of the forecast and separates the calibration and forecasting periods. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

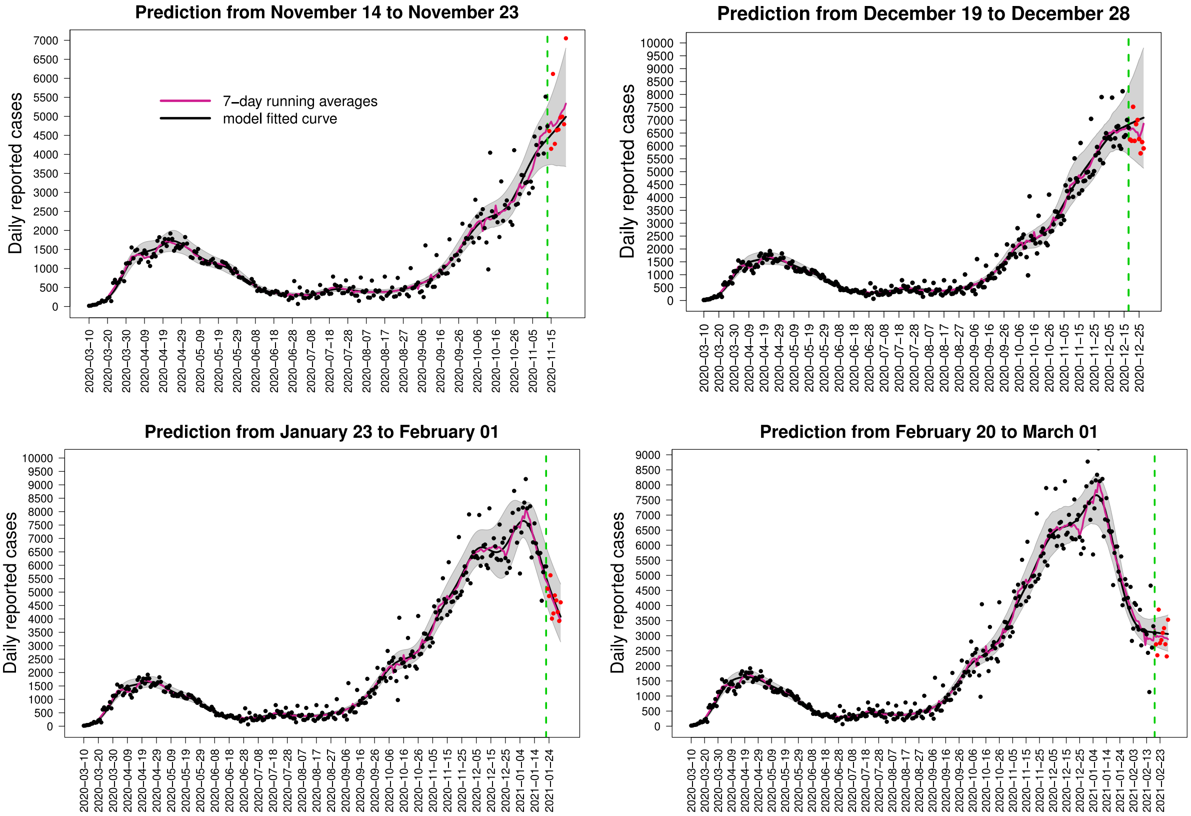

Fig. 2. Ten-day ahead forecasts of daily incident cases by the Poisson–Gamma SPMM model at various phases of the COVID-19 outbreak in Canada. Projections were made using all data reported since March 10, 2020. The black dots at the left of the green vertical line are actual daily incidence data that were used for model fitting. The black line is the model fitted (expected) values going through the dots; this line extends into the future (for next 10 days) as prediction. The violet line is the 7-day running averages. The red dots at the right of the green vertical line correspond to the daily incident cases reported since the projections. The shaded light-grey region is the 95% credible interval that are extended into the prediction; and the green vertical dashed line indicates the start time of the forecast and separates the calibration and forecasting periods. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

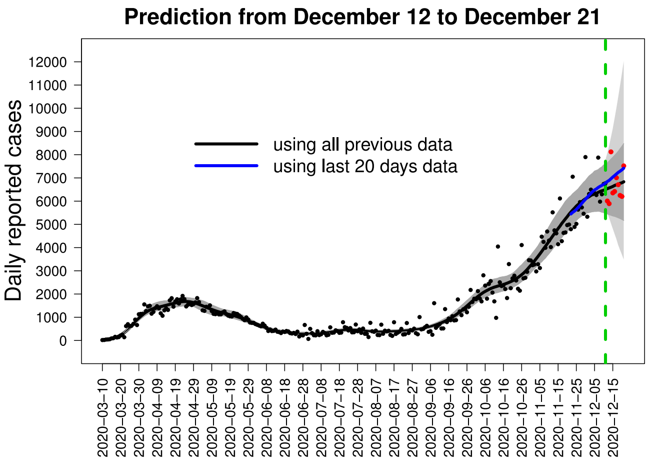

Fig. 3. Projection trajectories and the 95% credible intervals from the sensitivity analysis of the Poisson–Gamma SPMM with respect to the number of previous data used for model fitting. The shaded grey areas are 95% prediction intervals when past 20 days data were used for model fitting whereas the light-grey areas are the 95% prediction intervals when all previous data points were used for model fitting. The green vertical dashed line separates the calibration and forecasting periods.

4. Conclusion remarks and discussion

We have demonstrated the usefulness of the Poisson–Gamma SPMM for short-term predictions of the trajectory of future daily reported cases based on previously observed data. The statistical literatures related to theories and methods we referred (Ruppert, 2002; Wood, 2003, 2004) are focused on fitting models to data. In this paper, we put equal emphasis on how the fitted model adequately describes the underlying trajectory through noisy daily counts and on extrapolation of the model into the near future.

In the public media, a popular and useful description the underlying trajectory is the 7-day running average. As shown in Fig. 2, both our model fitted function and the 7-day running average describe the trend very well. Furthermore, the running averages still carry substantial fluctuation which is smoothed in our fitted function. Since the running averages cannot be naturally extended beyond the last data point, the

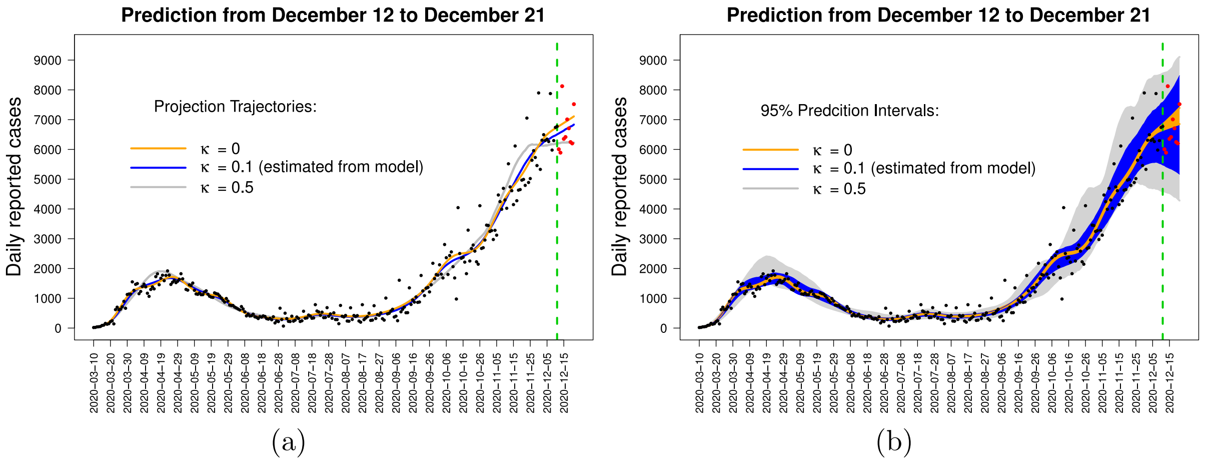

Fig. 4. Projection trajectories (left panel, a) and 95% credible and prediction intervals (shaded regions in the right panel, b) for different values of the over-dispersion parameter κ to assess the role of κ in curve fitting. Daily incidence cases reported since the projections were plotted (red dots) to the right of the vertical green line. The black circles correspond to the daily cases reported up until projection day; and the green vertical dashed line separates the calibration and forecasting periods. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

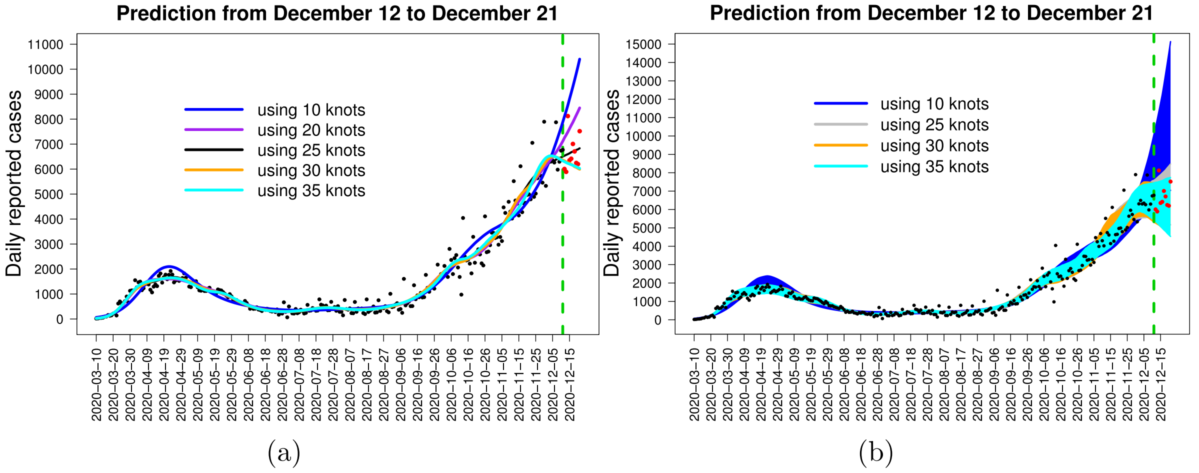

Fig. 5. Projection trajectories (left panel, a) and 95% credible intervals (shaded regions in the right panel, b) from the sensitivity analysis of the Poisson–Gamma SPMM model with respect to the choice of knots. Daily incidence cases reported since the projections were plotted (red dots) to the right of the vertical green line. The black circles correspond to the daily cases reported up until projection day; and the green vertical dashed line separates the calibration and forecasting periods. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

Table 1
Short-term forecasting performance from the sensitivity analyses to assess the robustness of the Poisson–Gamma SPMM model with respect to the choice of knots, and to the number of previous data points used for model fitting. We report deviance information criteria (DIC), mean absolute error (MAE), root mean square error (rMSE), mean interval score (MIS), and percentage coverage of the 95% prediction interval or average coverage probability (ACP).

Projection Week: December 12, 2020					
Sensitivity to	DIC	MAE	rMSE	MIS	ACP
Number of past data points used, n (when K = 25; κ estimated from data)					
all past data	2948.0	214.2	360.8	3963.8	0.9
past 20 days data	NA[a]	584.2	754.1	5421.4	1.0
Dispersion Parameter, κ (when K = 25; n = all past data)					
$\kappa = 0$	18344.9	215.3	360.9	19293.4	0.1
$\kappa = 0.1$	2948.0	214.2	360.8	3963.8	0.9
$\kappa = 0.5$	2957.1	260.6	420.7	4321.5	1.0
Number of Knots, K (when n = all past data; κ estimated from data)					
K = 10	2974.1	354.9	657.3	15756.4	0.3
K = 20	2980.5	243.6	425.4	4324.0	0.8
K = 25	2948.0	214.2	360.8	3963.8	0.9
K = 30	2970.7	217.4	372.4	5146.0	0.9
K = 35	2939.9	221.8	381.4	5094.1	0.9

[a] DIC of a model estimated using past 20 days data is not comparable to that estimated using all past data.

Table 2

Comparison of the short-term forecasting performance of the SPMM to other models fitted to the data within three different time segments covering a multiple-wave pattern, a single-wave pattern and a U-shape pattern. We report Bayesian information criteria (BIC), mean absolute error (MAE), root mean square error (rMSE), mean interval score (MIS), and percentage coverage of the 95% prediction interval or average coverage probability (ACP).

Model	BIC	MAE	rMSE	MIS	ACP
Scenario 1: Recurrence waves period, March 10–September 25					
SPMM	1855.8	122.4	192.8	1369.6	1.0
Logistic	3020.5	379.7	592.1	73890.3	0.0
Generalized-Richards	2997.3	353.4	569.1	73258.4	0.0
Cubic Polynomial	2862.4	234.0	292.5	5087.1	0.4
Scenario 2: A bell shaped single wave period, March 10–June 12					
SPMM	921.6	94.0	162.2	246.7	1.0
Logistic	1326.9	176.3	228.2	2075.2	0.3
Generalized-Richards	1296.9	127.2	189.0	338.1	0.7
Cubic Polynomial	1307.2	152.8	216.6	1548.7	0.4
Scenario 3: A U-shape period, April 26–September 25					
SPMM	1423.5	129.0	205.5	1662.1	0.9
Logistic	2278.1	354.9	554.8	60005.0	0.0
Generalized-Richards	2283.2	355.0	554.9	59998.0	0.0
Cubic Polynomial	2107.7	178.6	269.6	16071.4	0.1

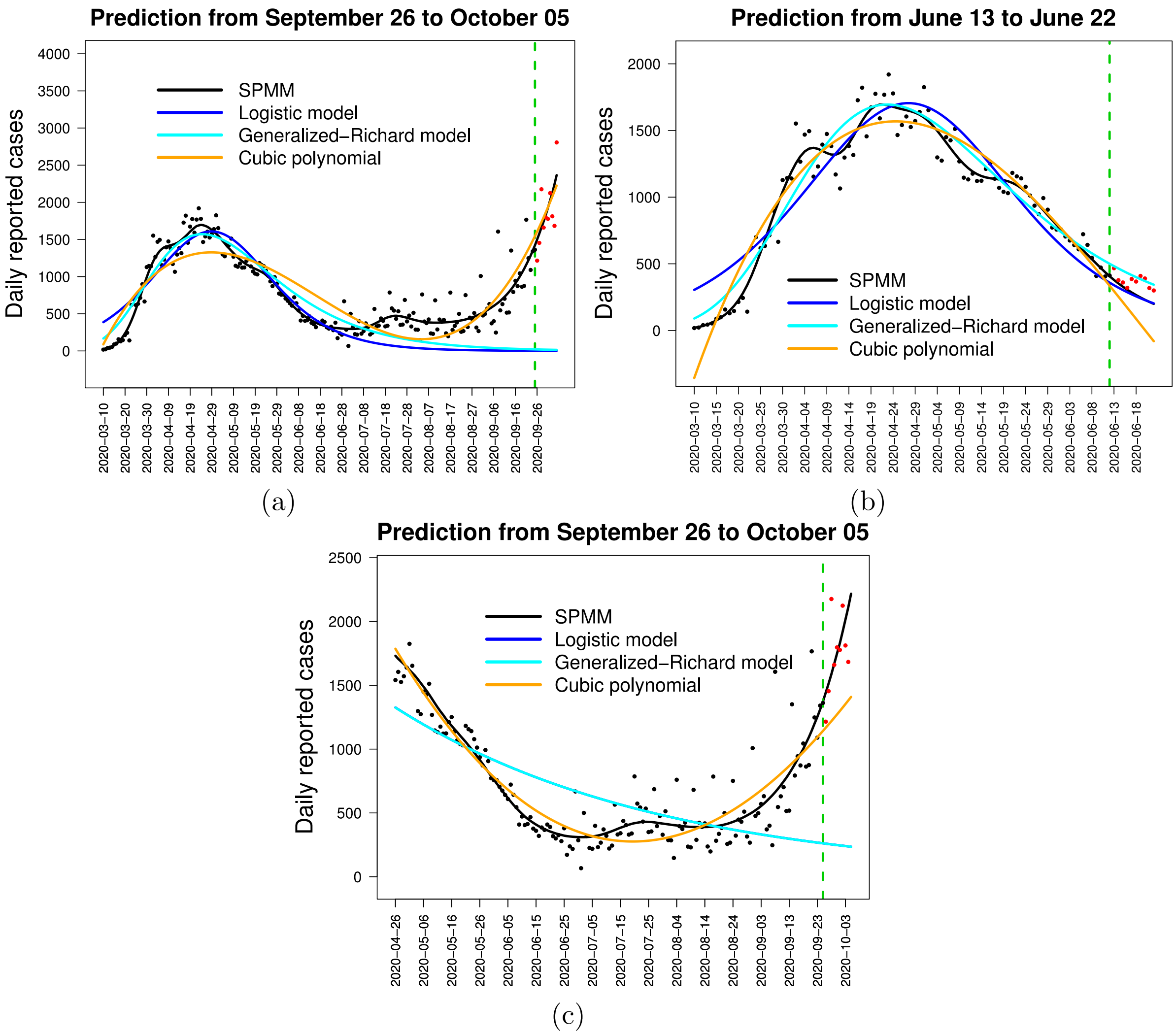

Fig. 6. A comparison of the curve fitting by the SPMM to the other simpler growth models in three different scenarios, covering a time period when the daily incidence shows a multiple-wave pattern (top left panel, a), a single-wave pattern (top right panel, b) and a U-shape pattern (bottom panel, c). For U-shape pattern in panel c, the logistic and generalized-Richards models yielded identical fit. The green vertical dashed line indicates the start time of the forecast and separates the calibration and forecasting periods. The black and red dots are the observed daily cases before and after the projection day, respectively.

short term predictions based on the SPMM may be interpreted as the prediction of the trajectory of the moving average.

We take a balanced approach among robustness, precision and accuracy, both in fitting the model to data and in the prediction.

We use the term credible intervals for characterizing the local variations in data and prediction intervals for the uncertainties in the prediction. The widths of these intervals depend on several factors. Our sensitivity analyses have shown that using all available previous data points produces narrower credible intervals as data far back tell us a lot about local variability (noise) due to reporting and testing processes. Figs. 1 and 2 show that the longer the time series used in the analysis, the narrower the estimated the credible intervals which give more precise description of the trend. By extension, they also yield narrower prediction intervals into the near future (Fig. 3).

The over-dispersion parameter κ and the number of knots K both affect the estimated range of uncertainty. The former is demonstrated in Fig. 4 whereas the later is shown in Fig. 5.

With the prior distribution $\kappa^{-1} \sim U(0, 500)$ along with K determined by (6), fitting the model to all available previous data points yields the estimated credible intervals wide enough to take into account the natural randomness in the reported data, while leaving large fluctuation in the data due to irregular reporting such as data catching up due to long weekends, data cleaning and computer system migration etc. (which happens from time to time in different jurisdiction) as “outliers”. This is a balance between precision and robustness that we take to best describe the historical trend. Had additional efforts been carried by diligently reviewing data jurisdiction by jurisdiction, it would have been possible to re-distribute the batch reported daily numbers on Mondays or the day immediately following status holidays to make most of these apparent outliers to disappear. We did not attempt to make these due diligences.

One of the reasons that we take a Bayesian approach is relevant to the prediction intervals. It is well recognized that the prediction intervals obtained by replacing the estimated parameters instead of the true parameters ignore the inherent additional variability, which results in narrower (than they should be) prediction intervals and under communicates the uncertainty of the predicted values. This is a core issue in the foundations of statistics (Geisser, 1993). The Bayesian prediction intervals are one of the ways to overcome this.

Figs. 4 and 5 also demonstrate how sensitive the directions of the predicted trajectories are with respect to κ and K. We limit our prediction horizon into the following 10 days, with carefully chosen K values according to the best fit criterion and the narrowest credible interval. The criteria on goodness of fit and the precision only summarize how well the fitted model explains historical trend and local variability.

We make the following comments:

1 The SPMM, the generalized Richards model and the cubic polynomial model have the same number of independent parameters. The generalized Richards model has 3 parameters as given by (1) plus an additional parameter $i_0 = C(0)$ for the initial condition. The cubic polynomial function, with an intercept parameter, also has 4 independent parameters. Although the SPMM seems to have very large number of parameters, but unlike the usual spline function, it only has 4 independent parameters to estimate: (i) a fixed effect associated with non-penalized part; (ii) the variance parameter sigma; (iii) the over dispersion parameter κ; and (iv) the intercept parameter. The parameters that define the spline function are implicit functions of these independent parameters as a thin-plate regression spline which rooted from a cubic spline but after some transformations and combinations.

2 Both the generalized Richards model and the cubic polynomial model can be only used to fit a section of the data. For prediction purposes, this section is referred to “recent data”, but there is no criterion to decide how “recent” is recent and different choices of recency determine the predicted trajectories. The SPMM, with the same number of independent parameters, can be used to fit to data as far as possible.

3 The fully parametric model is more restrictive in such a way that the underlying trend must conform to a specific shape as defined by the parametric model. In the case of the generalized Richards model, the expected cumulative counts must confirm to a sigmoidal curve and the expected daily counts must conform to a bell-shape. In reality, the underline trend of daily counts can be U-shaped, such as a resurgence of cases after declining into a trough. These data give very poor fit to such models.

A common limitation in all these models is that the trend of daily cases can change abruptly driven by factors such as the emergence of new variants, people’s behaviour, public health measures, as well as surveillance and reporting practices. It is beyond the scope of any of these empirical models to predict what might happen regarding these external transmission environment. However, we provide short term forecasts so we have to live with the noise in the data while still trying to capture the “signal” and forecast it.

CRediT authorship contribution statement

Muhammad Abu Shadeque Mullah: Conceptualization, Methodology, Analytical strategy, Software, Formal analysis, Interpretation of results, Writing – original draft, Writing – review & editing. **Ping Yan:** Methodology, Analytical strategy, Interpretation of results, Writing – original draft, Writing – review & editing.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgement

We are thankful to Professor J.F. Lawless from Department of Statistics and Actuarial Science, University of Waterloo for his valuable comments and feedback that greatly helped to improve this manuscript.

Funding

This work was supported by the Public Health Agency of Canada.

Appendix A. Supplementary data

The Supplementary Table, referenced in Section 3.1, is available online at https://doi.org/10.1016/j.epidem.2022.100537.

References

Anon, 2020a. CTV News: https://www.ctvnews.ca/health/coronavirus/how-canada-s-covid-19-pandemic-modelling-forecasts-compare-to-reality-1.5222193.
Anon, 2020b. National Post: https://nationalpost.com/news/canada/canadas-covid-models-have-been-largely-accurate-but-worst-cases-have-not-materialized.
Baily, N.T.J., 1975. The Mathematical Theory of Infectious Diseases and Its Applications, second ed. The Griffin & Company, London.
Borshchev, A., 2016. The Big Book of Simulation Modelling. AnyLogic North America.
Chowell, G., Tariq, A., Hyman, J.M., 2019. A novel sub-epidemic modeling framework for short-term forecasting epidemic waves. BMC Med. 17 (164), http://dx.doi.org/10.1186/s12916-019-1406-6.
Crainiceanu, C.M., Ruppert, D., Wand, M.P., 2005. Bayesian analysis for penalized spline regression using WinBUGS. J. Stat. Softw. 14 (14), 1–24.
Geisser, S., 1993. Predictive Inference, an Introduction. Chapman and Hall, New York.
Gelman, A., Rubin, D.B., 1992. Inference from iterative simulation using multiple sequences (with discussion). Statist. Sci. 7, 457–472.

Gneiting, T., Raftery, A.E., 2007. Strictly proper scoring rules, prediction, and estimation. J. Amer. Statist. Assoc. 102 (477), 122–127.

Lawless, J.F., 1987. Negative binomial and mixed Poisson regression. Canad. J. Statist. 9 (2), 209–225.

Ludwige, A., Berthiaume, P., Orpana, H., et al., 2020. Assessing the impact of varying levels of case detection and contact tracing on COVID-19 transmission in Canada during lifting of restrictive closures using a dynamic compartmental model. Can. Commun. Dis. Rep. 46 (11–12), 409–421.

Ogden, N.H., Fazil, A., Arino, J., Berthiaume, P., Fishman, D.N., Greer, A.L., Ludwig, A., Ng, V., Tuite, A.R., Turgeon, P., Waddell, L.A., 2020. Modelling scenarios of the epidemic of COVID-19 in Canada. Can. Commun. Dis. Rep. 46 (6), 198–204.

Plummer, M., 2009. Jags Version 1.0.3 Manual. Technical Report.

Roosa, K., Lee, Y., Luo, R., Kirpich, A., Rothenberg, R., Hyman, J.M., Yan, P., Chowell, G., 2020a. Real-time forecasts of the COVID-19 epidemic in China from February 5th to February 24th. Infect. Dis. Model. (ISSN: 2468-0427) 5, 256–263. http://dx.doi.org/10.1016/j.idm.2020.02.002.

Roosa, K., Lee, Y., Luo, R., Kirpich, A., Rothenberg, R., Hyman, J.M., Yan, P., Chowell, G., 2020b. Short-term forecasts of the COVID-19 epidemic in Guangdong and Zhejiang, China: February (2020) 13-23. J. Clin. Med. 9 (2), 596.

Ruppert, D., 2002. Selecting the number of knots for penalized splines. J. Comput. Graph. Statist. 11, 735–757.

Ruppert, D., Wand, M.P., Carroll, R.J., 2003. Semiparametric Regression. Cambridge University Press, Cambridge.

Schwarz, G.E., 1978. Estimating the dimension of a model. Ann. Statist. 6 (2), 461–464.

Smith, B., Bancej, C., Fazil, A., Mullah, M., Yan, P., Zhang, S., 2021. The performance of phenomenological models in providing near-term Canadian case projections in the midst of the COVID-19 pandemic. Epidemics http://dx.doi.org/10.1016/j.epidem.2021.100457.

Speigelhalter, D.J., Best, N.G., Carlin, B.O.P., van dar Linde, A., 2002. BayesIan measures of model complexity and fit (with discussion). J. R. Stat. Soc. B 64, 583–639.

Su, Y.S., Yajima, M., 2012. R2jags: A package for running jags from R. R package version 0.03-08. Available at http://CRAN.R-project.org//package=R2jags.

Wand, M.P., 2003. Smoothing and mixed models. Comput. Statist. 18, 223–249.

Wood, S., 2003. Thin plate regression splines. J. R. Stat. Soc. Ser. B 65, 95–114.

Wood, S., 2004. Stable and efficient multiple smoothing parameter estimation for generalized additive models. J. Amer. Statist. Assoc. 99, 673–686.

Yan, P., Chowell, G., 2019. Quantitative Methods for Investigating Infectious Disease Outbreaks. Springer Nature, Customer Service Center LLC, Switzerland.