A Bayesian Approach to Real-time Spatiotemporal Prediction Systems for Respiratory Syncytial Virus

Matthew J. Heaton

Department of Statistics, Brigham Young University, Provo, Utah, U.S.A

email: mheaton@stat.byu.edu

and

Celeste Ingersoll

Department of Statistics, Brigham Young University, Provo, Utah, U.S.A

email: celesteingersoll3@gmail.com

and

Candace Berrett

Department of Statistics, Brigham Young University, Provo, Utah, U.S.A

email: cberrett@stat.byu.edu

and

Brian Hartman

Department of Statistics, Brigham Young University, Provo, Utah, U.S.A

email: hartman@stat.byu.edu

and

Biometrics ?, 1–16

July 2020

Chantel Sloan

Department of Public Health, Brigham Young University, Provo, Utah, U.S.A

email: chantel.sloan@byu.edu

Summary: Respiratory Syncytial Virus (RSV) is a common cause of infant hospitalization and mortality. Unfortunately, there is no known cure for RSV but several vaccines are in various stages of clinical trials. Currently, immunoprophylaxis is a preventative measure consisting of a series of monthly shots that should be administered at the start, and throughout, peak RSV season. Thus, the successful implementation of immunoprophylaxis is contingent upon understanding when outbreak seasons will begin, peak, and end. In this research we estimate the seasonal epidemic curves of RSV using a spatially varying change point model. Further, using the fitted change point model, we develop a historical matching algorithm to generate real time predictions of seasonal curves for future years.

Key words: Spatiotemporal predictions; Markov chain Monte Carlo; Change point model

1. Introduction

Globally, respiratory syncytial virus (RSV) is a common cause of childhood acute lower respiratory infection (ALRI) and a major cause of hospital admissions in young children(Nair et al., 2013). About 45%, over 3.2 million,of hospital admissions and in-hospital deaths in children less than 6 months old were due to RSV-ALRI in 2015 (Shi et al., 2017). Further, children who are infected and survive have a much higher risk of developing longer term respiratory health issues. Such severe respiratory outcomes from RSV infection occur most often in children born preterm (Helfrich et al., 2015; Stagliano et al., 2015; Granbom et al., 2016).

Most RSV-attributed deaths are in children with a pre-existing condition and/or were born prematurely (Scheltema et al., 2017; Mazur et al., 2018). As RSV vaccine trials continue, we remain reliant on immunoprophylaxis shots being given to infants who are premature or predisposed to a severe RSV outcome (Hampp et al., 2011). Thus, to maximize the success rate for immunoprophylaxis (and any other preventative measure), it is important to not only understand when the RSV epidemic season will begin, peak, and end but also be able to predict the seasonal curves in real time (i.e. as soon as data becomes available). However, to date, very little is understood about the spatially varying dynamics of RSV year-to-year. Noyola and Mandeville (2008), Walton et al. (2010) and Sloan et al. (2017) show that RSV seasonality varies by climate and meteorogical conditions. In the contiguous US, studies by Stensballe et al. (2003) and Pugh et al. (2019) suggest that RSV seasonality generally starts in the south-east and moves north-west.

1.1 *Data*

The data motivating this research comes from the military data repository (MDR) of the military health system which serves all active and retired duty military personnel along with their dependents. The data contain birth dates, birth location and infection dates (for those infected with bronchiolitis) for all 50 states and US territories. However, for purposes of this research we only consider data from the 48 contiguous states and the District of Columbia. Inclusion criteria were children less than one year of age and born between January 1, 2003 and June 30, 2013. Subsequently, “controls” were defined as those children in the cohort who were never diagnosed with bronchiolitis and “cases” were those children who were diagnosed (using ICD-9 codes: 466.11, 466.19, 480.1, 0.79.6). Note that these ICD codes refer to bronchiolitis and not RSV directly. It has been shown that most bronchiolitis cases are RSV induced. Thus, in order to model and understand RSV, we will refer to bronchiolitis as RSV and assume that these cases were a result of RSV, as most are.

While the MDR contains individual-level information on birth location and time as well as infection information (for cases), such data is protected as personal health information. Hence, for this research, we analyze aggregated data from the MDR to preserve patient confidentiality. That is, for each state, the data analyzed here include the number of controls (hereafter referred to as “susceptible”) and the number of infant RSV cases in any given half month time period from 2004 to 2013 (time periods were defined as before the 15th and after or on the 15th of any month).

Figure 1 shows the trends of the ratio of the number of cases to the number of controls over time for a few selected states. While this figure displays data from 4 of 49 states analyzed in this project, they stand as an example of the variability in seasonal trends across both time and space. For example, in California, the peak proportion of cases in a season is around 0.02, but in Vermont the peak reaches close to 0.06. The data for California is consistent, but the data for Vermont is very sporadic with many zero case counts. This is mostly likely due to under-reporting, which occurs when the number of cases reported is less than the actual amount. This can happen when patients self-diagnose and seek over-the-counter medication, insurance claims are not filed, hospitals fail to report diagnosis, etc.

[Figure 1 about here.]

In addition to variability across states, each season within a state also varies from season to season. For example, Vermont has variable peaks, with some years peaking at significantly higher proportions of cases than others. Some states and years have long, wide seasons with low peak percentages, while others have short seasons with high peak percentages.

The variability of cases between states becomes more apparent in Figure 2. This map shows the total number of cases in each state, ranging from 60 to 15879, from 2004 to 2013. Given this raw data, it appears that Texas, California, Virginia, and North Carolina had many more reported cases of RSV than other states (again, due to the amount of military activity in these states). However, past research on RSV (Pugh et al., 2019) suggest that RSV is present in all states across the country.

[Figure 2 about here.]

1.2 *Statistical Challenges*

The task of estimating and predicting seasonal trends of RSV in infants using the aggregated MDR data faces several challenges including (but not limited to) space-time variation, big data, and prediction in real time. Consider each challenge in turn.

The aggregated MDR data is correlated and varies across space and time. That is, we need a model that will account for both the variability between states and variability across time. By modeling correlation in the data we will be able to borrow information across states so that states with little data can still have accurate seasonal RSV trend estimates. Further, accounting for the correlation in the data will ensure that we obtain accurate measures of uncertainty associated with our parameter estimates.

The MDR data covers 49 states over nine seasons, each season having 24 time periods, resulting in 10800 correlated data points. Traditional approaches to modeling spatial and temporal correlation results in a a 10800×10800 covariance matrix that is unreasonable to use computationally. While many methods have been proposed to circumvent this computational challenge (see Heaton et al., 2019, for a review), many of the proposed methods are not suitable for non-Gaussian data. As such, we plan to adopt the methods proposed in Hughes and Haran (2013) by using basis functions suitable to capturing state-to-state variation in bronchiolitis trends.

As mentioned, the success of immunoprophyalxis in preventing RSV hinges on the appropriate timing of administration. As such, up-to-date predictions of the current seasonal RSV curve is imperative for patient health. Given the nine seasons of data available in the MDR, we want to predict RSV trends over the next cycle in each state as data become available (i.e. as cases are reported). Traditionally, this is done via modeling spatial and temporal correlation and calculating the conditional distribution of the future cycle given the past cycle (referred to as the predictive distribution in Bayesian statistics or kriging in the spatial literature, see Banerjee et al. 2014 or Cressie and Wikle 2015 for details on kriging and Gelman et al. 2013 for details on predictive distributions). However, these standard prediction methods require expensive simulations or computations. When predicting in real time, these computations would have to be repeated each time new data became available. Hence, for real-time predictions, we need a computationally feasible yet accurate approach.

1.3 *Goals and Contributions*

The goal of this research is to estimate the start, peak and end of RSV seasonality in every state in the contiguous US over nine seasons from 2004 to 2013. Specifically, we use the aggregated MDR data described above to estimate parameters of a spatially varying change point seasonality curve with continuity constraints during the off-season. By borrowing information from neighboring states, we are able to better estimate parameters in states with few cases. Finally, we develop a real-time spatio-temporal RSV season prediction algorithm based on probabilistic historical matching.

The remainder of this paper is outlined as follows. In Section 2, we posit our spatially-varying change point model as well as detail our historical matching algorithm. In Section 3, we use the data to fit our change point model, display the results and use the resulting model fit to generate predictions for the 2013 RSV season. Finally, in Section 4 we draw conclusions and highlight future areas of research.

2. Model

2.1 *A Spatial Change Point Model*

The goal of this analysis is to estimate probability curves that tell us the likelihood of susceptible infants contracting RSV at different times within a cycle for every state. These curves will be seasonal, meaning that throughout one cycle the probability of contracting RSV rises, peaks, and falls. Let $s = 1, 2, \ldots, 49$ denote the state and $c = 1, 2, \ldots, 9$ denote the cycle of RSV where a cycle is from July of a given year to June of the following year. Finally, let $t = 0, \ldots, 23$ denote the half month time interval (i.e. before the 15th or on and after the 15th of each month as described in Section 1).

As a response variable, let Y_{sct} denote the number of cases in state s, cycle c and time t and N_{sct} denote the corresponding number of susceptibles (controls). As RSV infection is binary, it follows that the number of RSV cases given the number of susceptible infants, follows a binomial distribution such that

$$Y_{sct}|N_{sct} \sim Binomial(N_{sct}, \rho_{sct}) \tag{1}$$

where $\rho_{sct} \in (0, 1)$ is the probability of an RSV case occurring in state s, during cycle c, at time interval t. In essence, ρ_{sct} for all s, c and t are the primary parameters of interest in this study as this dictates the observed number of cases. As mentioned as a challenge in Section 1, this parameter varies but is correlated over time. Further, the seasonal cycle (i.e. the variation in ρ_{sct} for $t = 1, \ldots, 23$) should be well-defined with a start, peak and end.

In order to estimate an RSV probability curve with the appropriate seasonal shape, we propose the following change point model given by

$$\log\left(\frac{\rho_{sct}}{1-\rho_{sct}}\right) = \begin{cases} \beta_s & \text{if } t < \delta_{sc1} \\ \beta_s + \theta_{sc}(\frac{t-\delta_{sc1}}{\delta_{sc2}-\delta_{sc1}}) & \text{if } \delta_{sc1} < t < \delta_{sc2} \\ \beta_s + \theta_{sc} & \text{if } \delta_{sc2} < t < \delta_{sc3} \\ \beta_s + \theta_{sc} - \theta_{sc}(\frac{t-\delta_{sc3}}{\delta_{sc4}-\delta_{sc3}}) & \text{if } \delta_{sc3} < t < \delta_{sc4} \\ \beta_s & \text{if } t > \delta_{sc4} \end{cases} \tag{2}$$

where β_s is a baseline probability of RSV for each state, $\theta_{sc} > 0$ is the maximum increase in probability at the peak for each state and cycle and $\boldsymbol{\delta}_{sc} = \{\delta_{sc1}, \delta_{sc2}, \delta_{sc3}, \delta_{sc4}\}$ are the change point times for each curve. Notably, the logit of ρ_{sct} in (2) ensures $\rho_{sct} \in (0,1)$ as required. For further clarity on what each parameter in the above change point model represent, Figure 3 displays the resulting seasonal curve where $\beta_s = 0$, $\theta_{sc} = 0.9$, $\delta_{sc1} = 7$, $\delta_{sc2} = 11$, $\delta_{sc3} = 16$, and $\delta_{sc4} = 20$. Notably, each of these parameters are of scientific interest as they represent different facets of a seasonal RSV curve. For example, θ_{sc} corresponds to the increase in RSV probability at the peak of a season. Statistical inference for θ_{sc} can provide insight into which states and cycles were particularly burdensome.

[Figure 3 about here.]

The seasonal parameters θ_{sc} and changepoints $\boldsymbol{\delta}_{sc} = (\delta_{sc1}, \ldots, \delta_{sc4})'$ are allowed to vary by season and state capturing the variability seen in Figure 1. Further, as discussed below, we wish to correlate these parameters across space and time. In contrast, we allow the baseline rate β_s to be constant across cycles because off-season rates should be similar from cycle to cycle. Further, this allows for the resulting curves to be continuous across cycle.

By definition, the change points are constrained so that $0 < \delta_{sc1} < \cdots < \delta_{sc4} < 23$. To enforce this constraint we use the stick-breaking method similar to Pugh et al. (2019) and specify

$$\delta_{sc1} = \omega_{sc1}(23)$$

$$\delta_{sc2} = \omega_{sc2}[(1-\omega_{sc1})(23)] + \delta_{sc1}$$

$$\delta_{sc3} = \omega_{sc3}[(1-\omega_{sc2})(1-\omega_{sc1})(23)] + \delta_{sc2}$$

$$\delta_{sc4} = \omega_{sc4}[(1-\omega_{sc3})(1-\omega_{sc2})(1-\omega_{sc1})(23)] + \delta_{sc3}$$

where $\boldsymbol{\omega}_{sc} = \{\omega_{sc1}, \omega_{sc2}, \omega_{sc3}, \omega_{sc4}\} \in (0,1)$ correspond to the percentage of an entire seasonal cycle (23 time points) spent in that particular regime. Further, this parameterization also ensures that an entire cycle occurs within one season.

An inherent modeling challenge associated with this data is correlating the parameters in (2) across state and cycle. While various approaches exist, we extend the approach developed in Hughes and Haran (2013) to develop spatio-temporal basis functions for these parameters that will also reduce the dimension of the parameter space and facilitate computation. In our approach, let $\mathbf{A}$ denote a 441×441 matrix that indicates space and time neighbors for all states and cycles, where $49 \times 9 = 441$ is the total number of spatio-temporal points in consideration. Under this spatio-temporal extension of Hughes and Haran (2013), we define $\mathbf{M}$ to be the first Q eigenvectors of the matrix $(\mathbf{I}-\mathbf{J}/441)\mathbf{A}(\mathbf{I}-\mathbf{J}/441)$ where $\mathbf{I}$ is the identity matrix and $\mathbf{J}$ is a matrix of ones. In this way, we reduce the dimension of space and time in $\mathbf{A}$ into $Q \ll 441$ components where Q is chosen to capture the majority of space-time variation as quantified by the associated eigenvalues.

The matrix $\mathbf{M}$ above corresponds to a basis that captures space-time variation across states and cycles. As such, we model:

$$\log\left(\frac{\boldsymbol{\omega}_i}{1-\boldsymbol{\omega}_i}\right) = \gamma_{\omega_i 0}\mathbf{1} + \mathbf{M}\boldsymbol{\gamma}_{\omega_i} \tag{3}$$

$$\log(\boldsymbol{\theta}) = \gamma_{\theta 0}\mathbf{1} + \mathbf{M}\boldsymbol{\gamma}_{\theta} \tag{4}$$

where $\boldsymbol{\gamma}_{\omega_i}$ represents a vector of coefficients for $\boldsymbol{\omega}_i = (\omega_{11i}, \ldots, \omega_{SCi})'$ and $\boldsymbol{\gamma}_\theta$ represents a vector of coefficients for $\boldsymbol{\theta} = (\theta_{11}, \ldots, \theta_{SC})'$. Note that the logit transform for $\boldsymbol{\omega}_i$ ensures $\omega_{SCi} \in (0, 1)$ for all s and c while the log transform ensures $\theta_{sc} > 0$.

Finally, recall that the baseline rates β_s are constant across cycle. Hence, as a modeling strategy we define $\boldsymbol{A}_s$ that indicates only spatial neighbors for all states. Using the eigen-decomposition as described above, we let $\mathbf{M}_s$ be the first Q_s eigenvectors of $(\mathbf{I} - \mathbf{J}/49)\mathbf{A}(\mathbf{I} - \mathbf{J}/49)$ where $Q_s < 49$ and set

$$\boldsymbol{\beta} = \gamma_{\beta 0}\mathbf{1} + \mathbf{M}_s\boldsymbol{\gamma}_\beta \tag{5}$$

where $\boldsymbol{\gamma}_\beta$ represents a vector of basis coefficients. In this way, we are modeling on a lower dimensional space which facilitates smoothing across regions.

The unknown parameters in the spatially varying change point model above are $\gamma_{\omega_i 0}$ and $\boldsymbol{\gamma}_{\omega_i}$ for $i = 1, 2, 3, 4$ in Equation (3) as well as $\gamma_{\theta 0}$ and $\boldsymbol{\gamma}_\theta$ in (4) and $\gamma_{\beta 0}$ and $\boldsymbol{\gamma}_\beta$ in (5). In general, our strategy for assigning prior distributions is to be vague, as we have little prior information at our disposal. However, if more information were available, such information could be incorporated in to a more informative prior. Specifically we assume all parameters are independent and identically distributed $\mathcal{N}(0, 1)$ random variables. These priors may appear informative due to the small variances. However, recall that each parameter is defined on the logit scale such that a standard deviation of 1 is quite vague. Further, because the basis functions in $\mathbf{M}_s$ and $\mathbf{M}$ are orthonormal, assuming prior independence among the associated coefficients is justified.

2.2 *Predicting Via Historical Matching*

Beyond estimating seasonal curves of RSV, a main goal in this analysis is to predict the seasonal curve of an upcoming season. As noted in the introduction, this can be done via refitting the above model each time data becomes available. However, this is computationally expensive and impractical for online predictions of the seasonal curve. Hence, here we propose a practical historical matching algorithm that is computationally feasible while still producing accurate results. That is, to make these predictions, we propose to match future data to past seasonal curves. Details are as follows.

Let $\boldsymbol{\rho}_{s(c+1)}$ represent all parameters associated with the seasonal curve in cycle $c+1$ and let $\mathbf{D}_{s(c+1)}$ denote all data available in cycle $c+1$ for state s (which may be the empty set if no data is observed or partial data from a season if only part of the season has been observed). We assume, *a priori*, that $\text{Prob}(\boldsymbol{\rho}_{s(c+1)} = \boldsymbol{\rho}_{sc}) = \pi_{sc}$ for $s = 1, \dots, S$ and $c = 1, \dots, C$ such that $\sum_{s,c} \pi_{sc} = 1$ and $\pi_{sc} \in [0, 1]$. In this case, the posterior distribution for $\boldsymbol{\rho}_{s(c+1)}$ is given by

$$\text{Prob}(\boldsymbol{\rho}_{s(c+1)} = \boldsymbol{\rho}_{sc}) \propto \mathcal{L}(\mathbf{D}_{s(c+1)} \mid \boldsymbol{\rho}_{sc})\pi_{sc} \tag{6}$$

where $\mathcal{L}(\cdot)$ denotes the binomial likelihood of data $\mathbf{D}_{c+1}$ if $\mathcal{D}_{c+1} \neq \emptyset$ and $\mathcal{L}(\cdot) = 1$ otherwise.

The above historical matching approach allows us to probabilistically match a future curve $\boldsymbol{\rho}_{s(c+1)}$ to past observed curves $\boldsymbol{\rho}_{sc}$ as data $\mathbf{D}_{c+1}$ becomes available. Further, the computation for this historical matching is substantially easier to implement than fully refitting the above model. That is, if $\boldsymbol{\Psi}$ denotes all parameters from the model in Section 2.1, given MCMC draws $\boldsymbol{\Psi}^{(1)}, \dots, \boldsymbol{\Psi}^{(N)}$ from the posterior distribution, the algorithm to obtain posterior draws of $\boldsymbol{\rho}_{s(c+1)}$ proceeds as follows for each $i = 1, \dots, N$:

(1) Using the available data, calculate the likelihood of $\mathbf{D}_{s(c+1)}$ under each state-cycle combination such that,

$$\mathcal{L}(\mathbf{D}_{s(c+1)} | \boldsymbol{\rho}_{sc}) = \prod_t \binom{N_{s,c+1,t}}{Y_{s,c+1,t}} (\rho_{sct}^{(i)})^{Y_{s,c+1,t}} (1 - \rho_{sct}^{(i)})^{N_{s,c+1,t} - Y_{s,c+1,t}}. \tag{7}$$

(2) Calculate posterior probabilities using the likelihood and prior as in (6).

(3) Randomly sample a historical match based on the posterior probabilities.

The algorithm results in N draws of $\boldsymbol{\rho}_{s(c+1)}$ that can be used for posterior analysis.

We make predictions using two different priors: (1) uniform prior and (2) spatial prior. The uniform prior gives each state and cycle an equal probability of being matched to the data. That is, under the uniform prior $\pi_{sc} = 1/(S \times C)$. Whereas, the spatial prior allows only states neighboring the predicted state to be matched. That is, $\pi_{sc} \neq 0$ only for those states neighboring state s and π_{sc} is uniform across the spatial neighbors.

3. Application to RSV Data

3.1 *Model Inference*

Figure 4 shows posterior mean estimates of the probability of contracting RSV (see Equation (2)) during the 2007-2008 cycle. While our estimates reveal that every season is different for each state, there are a few commonalities between seasons that are demonstrated by the 2007-2008 cycle which we wish to highlight. Primarily, at the beginning of the cycle, probabilities first rise in the south-east states (e.g. between Louisiana and Florida). Over the next two months, RSV moves through the Midwest and up the east coast. Finally, rates in the northwest finally hit their respective maximum in early February before starting to decline.

[Figure 4 about here.]

While the above south-east to north-west movement is fairly consistent from year to year, some states have highly variable seasons year over year. Figure 5 shows the posterior mean probability of contracting RSV in California, Nebraska, Texas, and Utah along with 95% credible interval bands. In California, RSV seasonality is consistent where the same relatively mild RSV season repeats every year, with peak probabilities extending no higher than 2%. In Nebraska (top right panel), we see a different RSV season almost every year with some being more severe than others. In various states, RSV seasonality is short but strong (e.g. Utah) whereas other states have longer seasons with wider peaks (e.g. Texas). In all four plots, we see very little uncertainty in the estimates.

[Figure 5 about here.]

Each season, RSV affects different areas in different ways. Figure 6 map peak probabilities from four different seasonal cycles. In the 2005-2006 cycle, RSV had the highest impact in the Northern Midwest with an unusually bad season in Wisconsin. Two seasons later (2007-2008), the peak of RSV infection seems to have been observed further south. In the 2011-2012 season, RSV again impacted the south the most, but this season was still relatively light compared to the peak rate in other seasons.

[Figure 6 about here.]

3.2 *Prediction Via Historical Matching*

Each of Figures 4-6 emphasize the importance of modeling season-to-season and state-to-state variablity in RSV seasonality. Due to this variability, health care professionals may not be able to base timing of treatments of immunoprophylaxis (or other preventative/counter measures) simply on past seasons of RSV in their area. Thus, predicting the timing of future RSV seasons is incredibly valuable for immunoprophylaxis administration.

To demonstrate the historical matching prediction algorithm, we refit the model from Section 2.1 with the last cycle of data held out for prediction validation. For this cross-validation study, we make predictions of upcoming RSV seasons by matching future data to past seasonal curves using two different priors for $\{\pi_{sc} : s = 1, \ldots, S; c = 1, \ldots, C\}$: (1) a uniform prior and (2) a spatial prior. The uniform prior gives each cycle and state equal probability of being matched to any state and any cycle (mathematically $\pi_{\rho sc} = (SC)^{-1}$) whereas the spatial prior only allows states neighboring the predicted state to match (mathematically $\pi_{\rho sc} = 0$ for all $s' \notin \mathcal{N}_s$ where $\mathcal{N}_s$ is the set of neighbors of state s).

The results of predicting the last season are displayed in Figure 7 which compare predicted curves using a uniform prior and a spatial prior for three states side by side with the true data points overlaid in black. Notably, these online predictions only use data up to and including the time of prediction. For example, at time $t = 5$, the associated predictions only use data from time point $1, \ldots, 5$.

[Figure 7 about here.]

In both prior cases, the predicted RSV seasonal curves change and improve as more data is observed. In the uniform case, the predicted season for Washington remains low until time period $t = 10$ after which the predicted curves start to change significantly as more data is observed and the predicted curves begin to accurately represent the overall season. When using a spatial prior, the predicted curves shift toward to the true season much sooner. As seen with Georgia, the predictions using the uniform prior take much more data to estimate the true seasonal curve. The spatial prior helps the predicted curves move closer to the true season with fewer known data points.

[Figure 8 about here.]

The uncertainty we have regarding these estimates is shown in Figure 8. We calculate 95% uncertainty bands for the predicted RSV curves using the MCMC draws. In many cases, the uncertainty bands largely overlap with curves from neighboring time periods because the estimated curves are so similar. As a result, we only show the estimated curves with uncertainty bands for three time periods that have relatively different predictions within a season. In most cases, the the uncertainty bands are thinner when a spatial prior is used.

4. Conclusions

In this research we estimated the seasonality of RSV over nine seasons using a spatially varying change point model with continuity constraints during the off-season. Using this model, we estimated the start, peak, and end of RSV seasonality for every state in the contiguous United States from 2004 to 2013. By borrowing information from neighboring states using spatial correlation, we were better able to predict RSV season in states with few and/or under-reported cases. In addition, correlating the model parameters across space and time allowed us to capture the variability of RSV seasons between and within states while appropriately accounting for uncertainty.

Using the fitted change point model, we developed a computationally practical historical matching algorithm for online predictions of RSV season. This algorithm matches future data to past seasonal curves to generate real time predictions of RSV seasonal curves for future years with out having to refit the model each time new data becomes available. We made these predictions under both a uniform prior and a spatial prior, and concluded that a spatial prior produces more accurate estimates of RSV seasonal curves with less available data. We can successfully identify peak times and rates of RSV in every state of the contiguous US.

We found that this model successfully captured the variability of RSV seasons within and between states. Although every year differs, a common trend found year over year is RSV season starting strong in the South and moving through the Midwest with lighter seasons on the coasts. Some states, like California, have consistent RSV seasonal patterns, while other states, like Louisiana, have vastly different seasons each year. Using this model and historical matching, we also predict which states will likely have the worst RSV season by estimating peak probabilities of getting RSV. All this information about RSV seasons can support health care professionals in their efforts to administer treatments of immunoprophylaxis or other preventative treatments effectively.

In this current model, we force each season to have a unimodal curve, but in reality some seasonal curves may be bimodal. While the unimodal assumption is likely true for RSV due to the regular seasonality, if these methods were applied to some other disease then possible multimodality may need to be accounted for. However, we leave the multimodal extension of these models to future research.

One shortcoming of the proposed historical matching algorithm is if historical data is scarce. Certainly, for this application, we had 8 years of previous data in 49 states (392 total

RSV cycles) which is sufficient to observe a wide variety of RSV seasonal curves. However, if less data were available then less historical curves would be available to match to. Hence, alternative prediction strategies beyond the historical matching algorithm used here may be a useful avenue for future research.

Overall, the benefits of this research on understanding and predicting RSV seasonality is in determining how and when to effectively administer preventative measures. Health care providers can use these methods and predictions to continually evaluate how RSV is behaving in their respective regions and best serve infants and families around them.

Acknowledgements

Is there anyone we need to acknowledge?

References

Banerjee, S., Carlin, B. P., and Gelfand, A. E. (2014). *Hierarchical modeling and analysis for spatial data.* Chapman and Hall/CRC.

Cressie, N. and Wikle, C. K. (2015). *Statistics for spatio-temporal data.* John Wiley & Sons.

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., and Rubin, D. B. (2013). *Bayesian data analysis.* Chapman and Hall/CRC.

Granbom, E., Fernlund, E., Sunnegårdh, J., Lundell, B., and Naumburg, E. (2016). Respiratory tract infection and risk of hospitalization in children with congenital heart defects during season and off-season: A swedish national study. *Pediatric cardiology* **37,** 1098–1105.

Hampp, C., Kauf, T. L., Saidi, A. S., and Winterstein, A. G. (2011). Cost-effectiveness of respiratory syncytial virus prophylaxis in various indications. *Archives of pediatrics & adolescent medicine* **165,** 498–505.

Heaton, M. J., Datta, A., Finley, A. O., Furrer, R., Guinness, J., Guhaniyogi, R., Gerber, F., Gramacy, R. B., Hammerling, D., Katzfuss, M., et al. (2019). A case study competition among methods for analyzing large spatial data. *Journal of Agricultural, Biological and Environmental Statistics* **24,** 398–425.

Helfrich, A. M., Nylund, C. M., Eberly, M. D., Eide, M. B., and Stagliano, D. R. (2015). Healthy late-preterm infants born 33–36+ 6 weeks gestational age have higher risk for respiratory syncytial virus hospitalization. *Early human development* **91,** 541–546.

Hughes, J. and Haran, M. (2013). Dimension reduction and alleviation of confounding for spatial generalized linear mixed models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **75,** 139–159.

Mazur, N. I., Higgins, D., Nunes, M. C., Melero, J. A., Langedijk, A. C., Horsley, N., Buchholz, U. J., Openshaw, P. J., McLellan, J. S., Englund, J. A., et al. (2018). The respiratory syncytial virus vaccine landscape: lessons from the graveyard and promising candidates. *The Lancet Infectious diseases* **18,** e295–e311.

Noyola, D. and Mandeville, P. (2008). Effect of climatological factors on respiratory syncytial virus epidemics. *Epidemiology & Infection* **136,** 1328–1332.

Pugh, S., Heaton, M. J., Hartman, B., Berrett, C., Sloan, C., Evans, A. M., Gebretsadik, T., Wu, P., Hartert, T. V., and Lee, R. L. (2019). Estimating seasonal onsets and peaks of bronchiolitis with spatially and temporally uncertain data. *Statistics in medicine* **38,** 1991–2001.

Scheltema, N. M., Gentile, A., Lucion, F., Nokes, D. J., Munywoki, P. K., Madhi, S. A., Groome, M. J., Cohen, C., Moyes, J., Thorburn, K., et al. (2017). Global respiratory syncytial virus-associated mortality in young children (rsv gold): a retrospective case series. *The Lancet Global Health* **5,** e984–e991.

Shi, T., McAllister, D. A., O'Brien, K. L., Simoes, E. A., Madhi, S. A., Gessner, B. D., Polack, F. P., Balsells, E., Acacio, S., Aguayo, C., et al. (2017). Global, regional, and national disease burden estimates of acute lower respiratory infections due to respiratory syncytial virus in young children in 2015: a systematic review and modelling study. *The Lancet* **390,** 946–958.

Sloan, C., Heaton, M., Kang, S., Berrett, C., Wu, P., Gebretsadik, T., Sicignano, N., Evans, A., Lee, R., and Hartert, T. (2017). The impact of temperature and relative humidity on spatiotemporal patterns of infant bronchiolitis epidemics in the contiguous united states. *Health & place* **45,** 46–54.

Stagliano, D. R., Nylund, C. M., Eide, M. B., and Eberly, M. D. (2015). Children with down syndrome are high-risk for severe respiratory syncytial virus disease. *The Journal of pediatrics* **166,** 703–709.

Stensballe, L., Devasundaram, J., and Simoes, E. (2003). Respiratory syncytial virus epidemics: the ups and downs of a seasonal virus. *The Pediatric infectious disease journal* **22,** S21–S32.

Walton, N. A., Poynton, M. R., Gesteland, P. H., Maloney, C., Staes, C., and Facelli, J. C. (2010). Predicting the start week of respiratory syncytial virus outbreaks using real time weather variables. *BMC medical informatics and decision making* **10,** 68.

Received 2 June 2020.

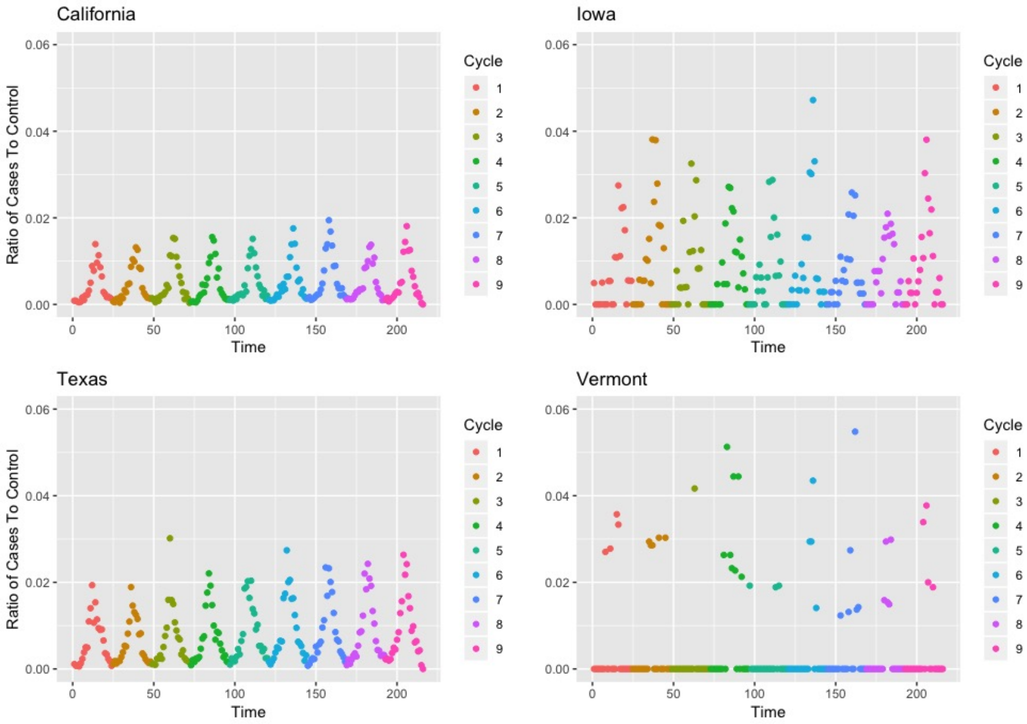


Figure 1: Ratio of the number of cases to the number of controls over time for California, Iowa, Texas, and Vermont

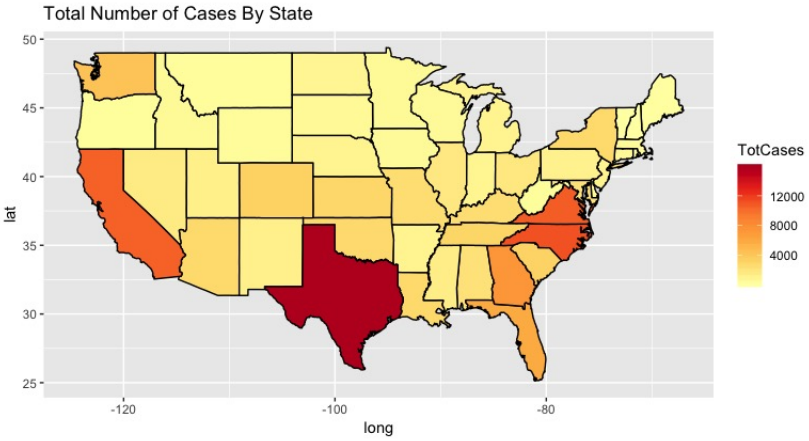


Figure 2: Total number of cases over all years in each state.

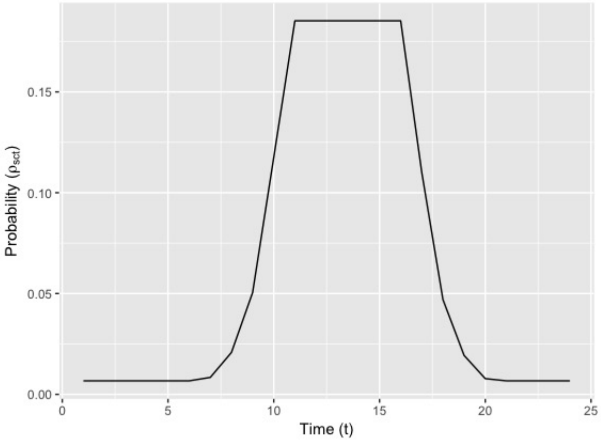


Figure 3: Example seasonality curve for the probability, ρ_{sct}, of contracting RSV in cycle c for state s.

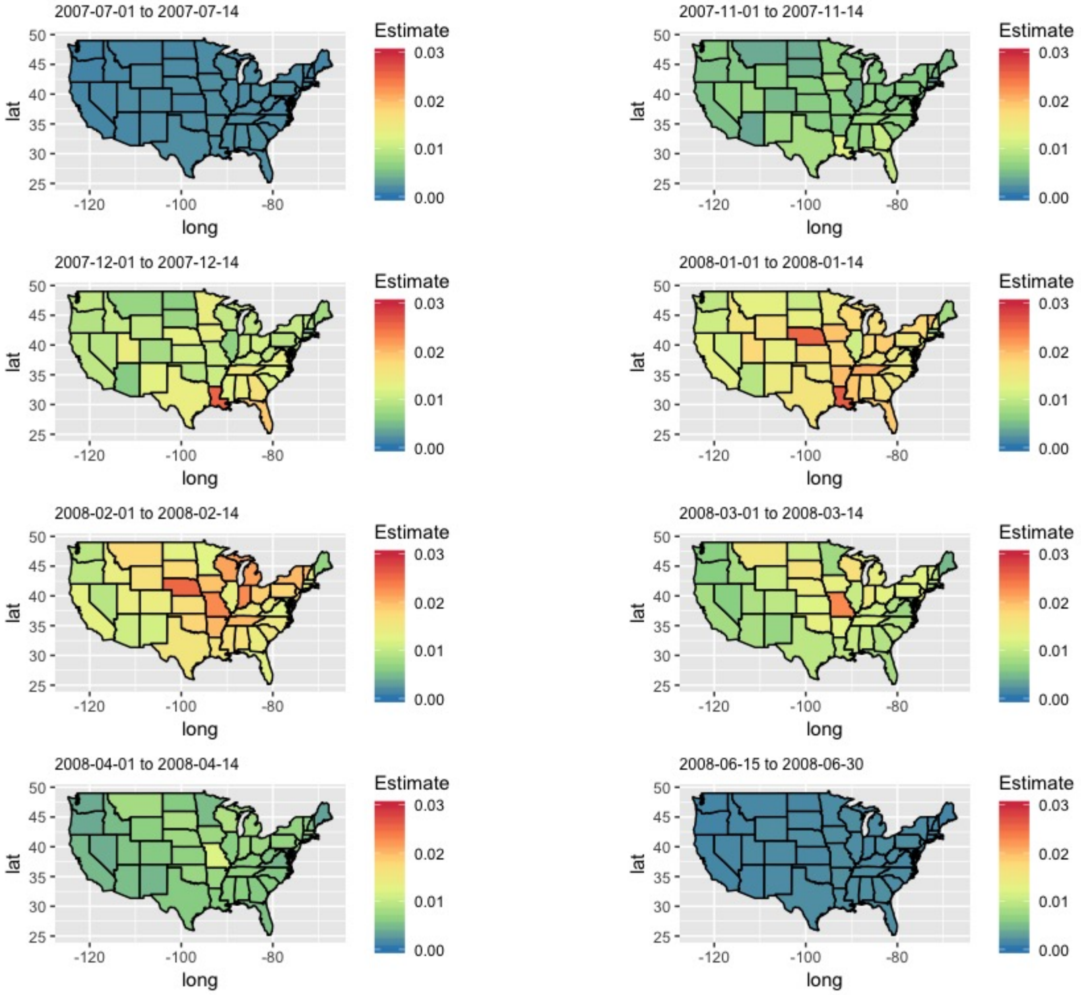


Figure 4: Progression of the RSV season across the contiguous United States from July 2007 to June 2008.

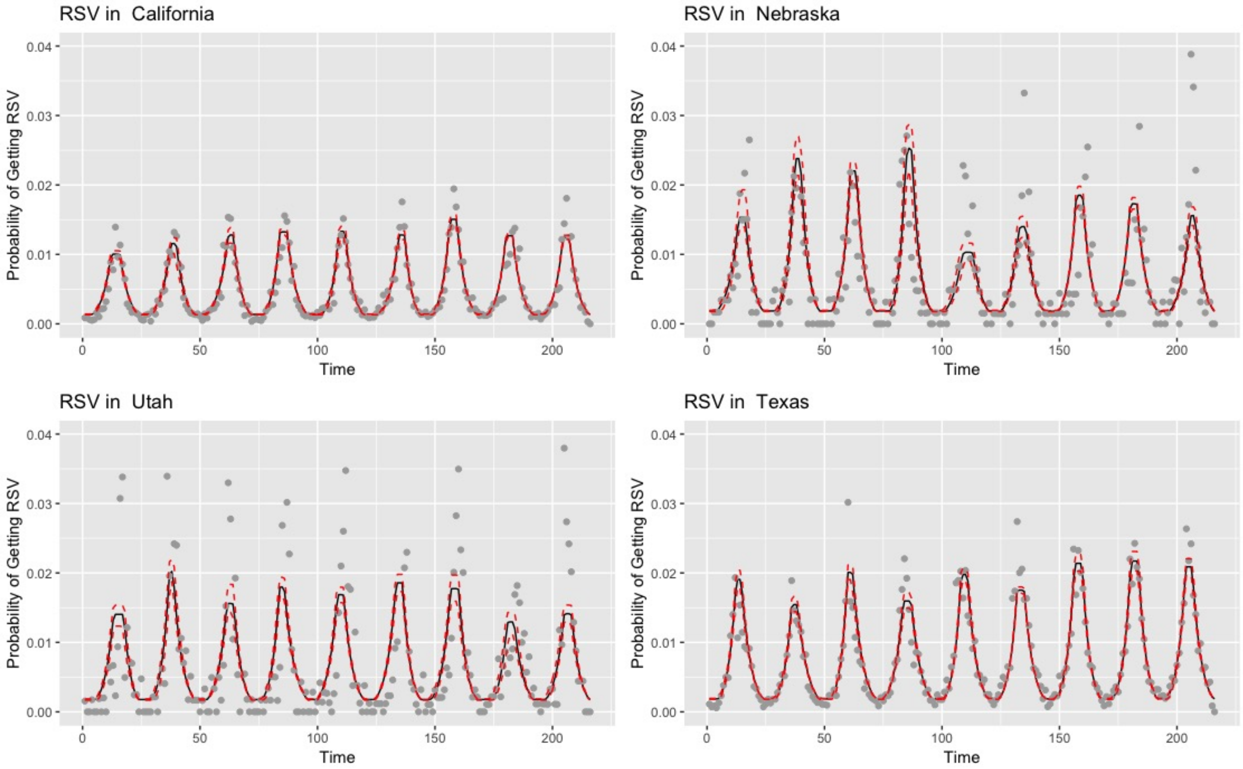
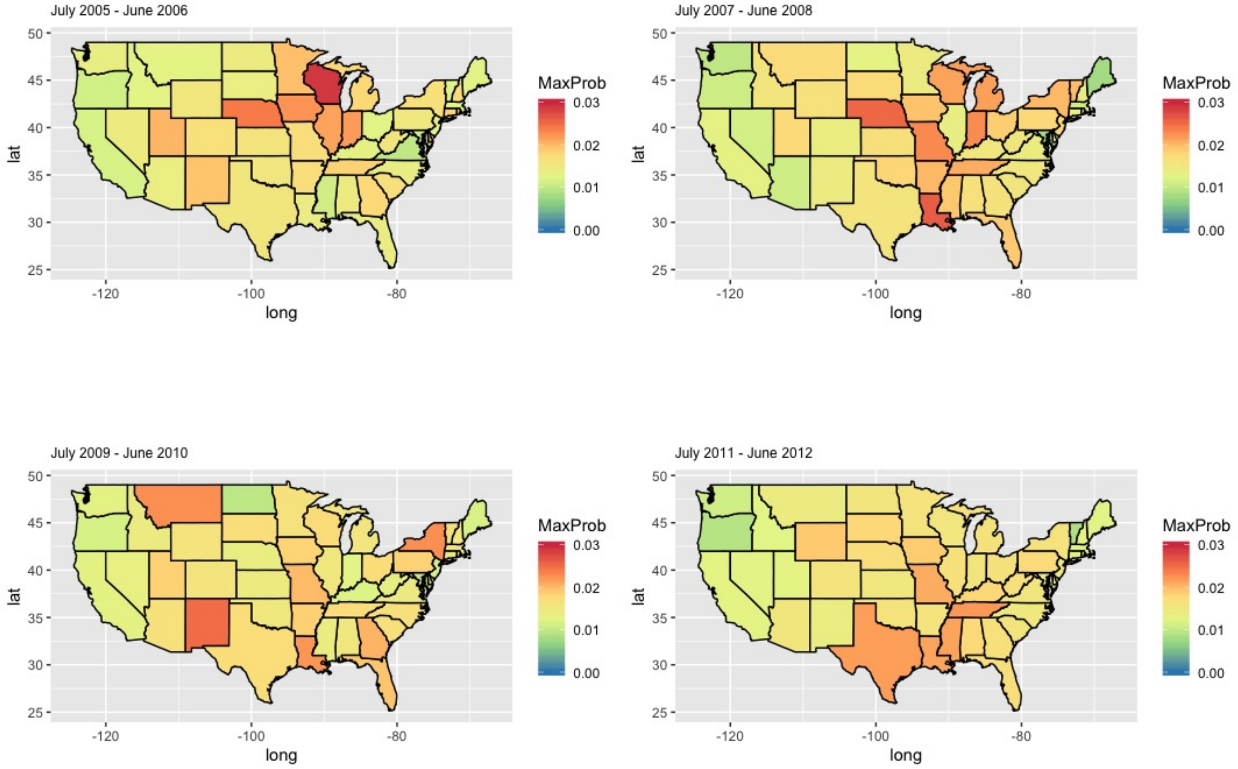


Figure 5: Fitted probability curves with 95% credible bands for RSV in California, Nebraska, Utah, and Texas.



July 2005 - June 2006

July 2007 - June 2008

July 2009 - June 2010

July 2011 - June 2012

Figure 6: Peak probabilities of RSV infection in each state from four selected cycles.

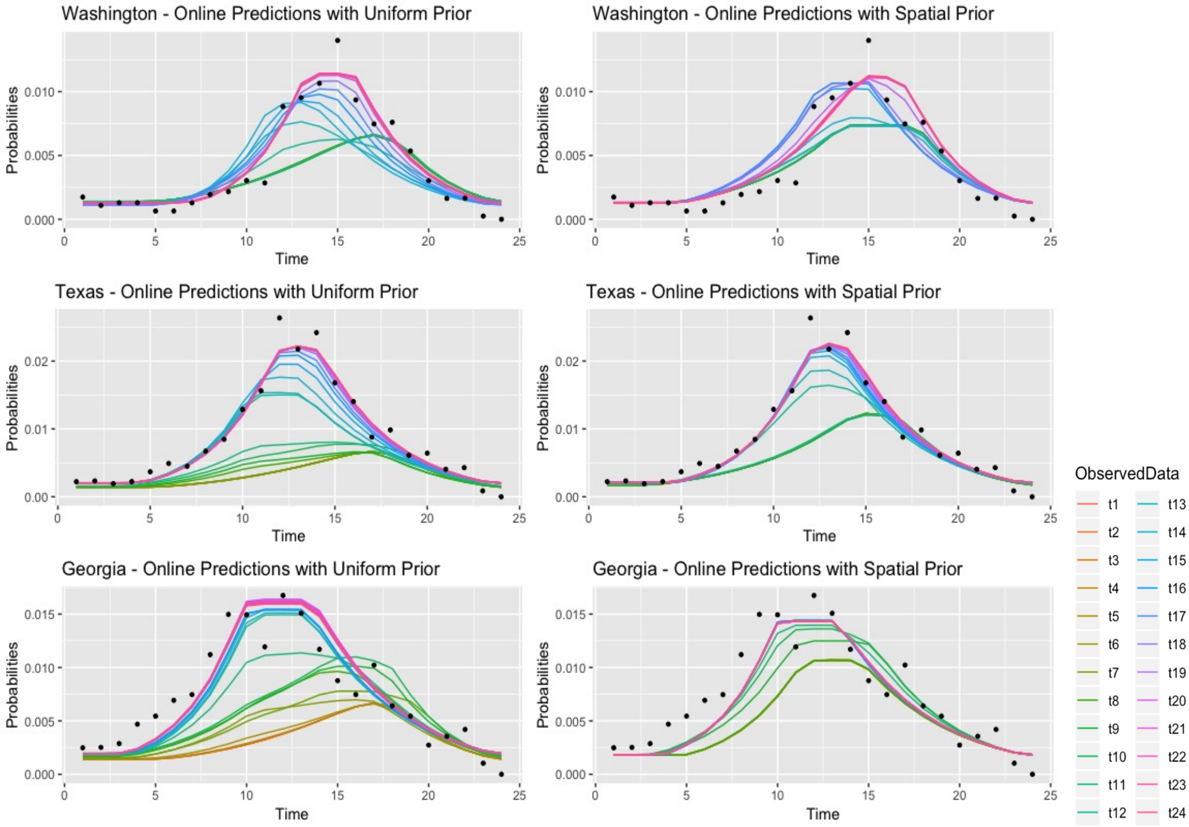


Figure 7: Online predictions via historical matching for Washington, Texas, and Georgia using both uniform and spatial priors.

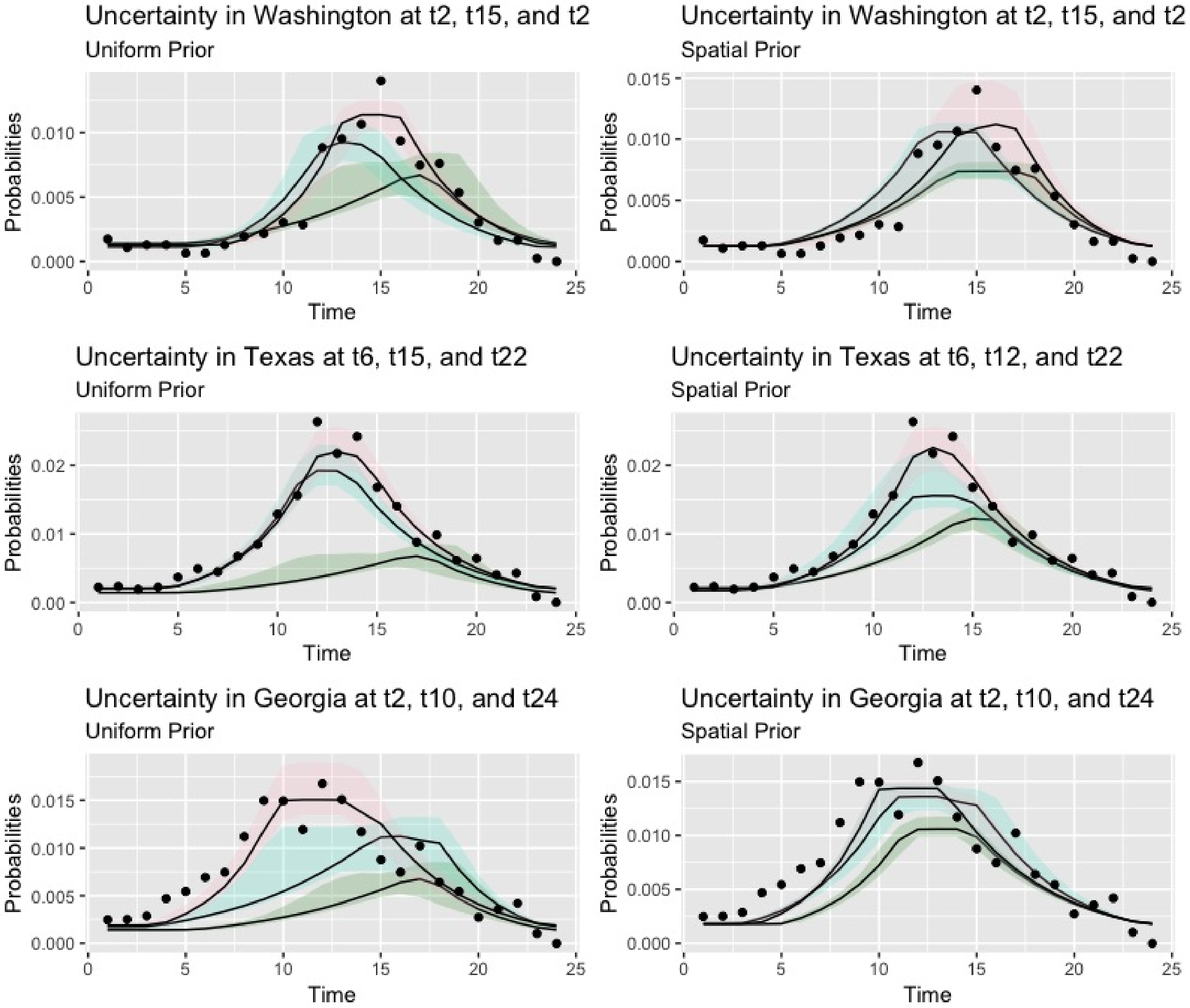


Figure 8: Uncertainty bands on predicted RSV curves for Washington, Texas, and Georgia at three time periods.