Infectious disease prediction with kernel conditional density estimation Supplementary Material

Evan L. Ray[a], Krzysztof Sakrejda[a], Stephen A. Lauer[a], Michael A. Johansson[b], Nicholas G. Reich[a]

[a] *Department of Biostatistics and Epidemiology, School of Public Health and Health Sciences, University of Massachusetts, Amherst*
415 Arnold House, 715 N. Pleasant Street, Amherst, MA 01003, USA

[b] *Dengue Branch, Division of Vector-Borne Infectious Diseases, Centers for Disease Control and Prevention, San Juan, Puerto Rico, USA*

1 Methodological Details

1.1 Discretizing the Kernel Function

We obtain the discrete kernel function by discretizing an underlying continuous kernel function. For each component of the vector $\tilde{\mathbf{z}}_{t^*} = (z_{t^*-l_1}, \ldots, z_{t^*-l_M}, z_{t^*+h})'$, we associate lower and upper bounds of integration a_{z_j} and b_{z_j} with each value in the domain of that random variable. The value of the kernel function is obtained by integrating over the hyper-rectangle specified by these bounds:

$$K^{\text{inc}}_{\text{disc}}(\tilde{\mathbf{z}}_{t^*}, \tilde{\mathbf{z}}_t; \mathbf{B}^h) = \int_{a_{z_{t^*-l_1}}}^{b_{z_{t^*-l_1}}} \cdots \int_{a_{z_{t^*+h}}}^{b_{z_{t^*+h}}} K^{\text{inc}}_{\text{cont}}(\tilde{\mathbf{z}}_{t^*}, \tilde{\mathbf{z}}_t; \mathbf{B}^h)\, dz_{t^*-l_1} \cdots dz_{t^*+h}.$$

In our application, the possible values of the random variables are non-negative integer case counts. In order to facilitate use of the log-normal kernel, we add 0.5 to the observed case counts; the corresponding integration bounds are the non-negative integers as illustrated in Supplemental Figure 1.

1.2 Alternative Copula Specifications

In Section 2.2 of the main manuscript, we introduced the use of copulas to model dependence in incidence values across different weeks in the season. In our model, we have used a separate copula function for each trajectory length (where the length of the trajectory is equal to the number of weeks remaining in the season). This specification allows the amount of dependence among future incidence values to vary with the time in the season at which we are making the predictions. Alternative formulations are also possible: for instance, we could fit a single copula function for the longest trajectory required, and use the relevant subset of those dependence parameters to make predictions at shorter trajectory lengths. We have not tried this alternative formulation, but we have observed that in our formulation the estimated dependence parameters vary strongly with the trajectory length. For instance, in the application to influenza using the KCDE specification with a fully parameterized bandwidth and a periodic kernel component the estimate of ξ_1^H is 0.909 for a trajectory of length $H = 17$, but it is 0.769 for a trajectory of length $H = 31$. Thus, the amount

of dependence between incidence values $d = 1$ week apart is much lower for $H = 31$ than it is for $H = 17$. This level of variation is fairly consistent across all values of H and d, and the standard errors of these estimates are in the range of about 0.02 to 0.06, suggesting that these differences are statistically significant (though we have not conducted a formal hypothesis test of this). This indicates that a simpler model with the same dependence parameters used at all points in the season would not adequately capture variation in dependence levels over the course of the season.

1.3 Parameter Estimation

We follow a two-stage strategy for parameter estimation [1]:

1. Estimate the parameters for marginal predictive distributions using the cross-validation procedure described in Section 2.1 of the main text.

2. Estimate the copula parameters, holding the parameters for the marginal predictive distributions fixed:

   (a) Form vectors of "pseudo-observations" by passing observed incidence trajectories from previous seasons through the marginal predictive c.d.f.s obtained in step 1:

   $$(u_{k,1}, \ldots, u_{k,H}) = \{F^1(z_{t_k^*+1} \mid t_k^*, z_{t_k^*-l_1}, \ldots, z_{t_k^*-l_M}; \boldsymbol{\theta}^1), \ldots, F^H(z_{t_k^*+H} \mid t_k^*, z_{t_k^*-l_1}, \ldots, z_{t_k^*-l_M}; \boldsymbol{\theta}^H)\}$$

   We form one such vector of pseudo-observations for each season in the training data; in the notation here, these seasons are indexed by k. The relevant time points t_k^* are the times in those previous seasons falling H time steps before the end of the season.

   (b) Estimate the copula parameters $\boldsymbol{\xi}^H$ by maximizing the likelihood of the pseudo-observations.

2 Simulation Study Details

2.1 Simulation Distributions

In the simulation study, we simulate data from a discretized multivariate normal distribution. The method for discretizing the underlying multivariate normal is the same as we described above for descritizing the kernel function. The normal distribution has mean $\underline{0}$ and covariance matrix with 1 on the diagonal and 0.9 off of the diagonal. Alternatively, this distribution can be characterized as the sum of the column vectors $(U, U)'$ and $\mathbf{V} = (V_1, V_2)'$, where $U \sim N(0, 0.9)$ is a random offset generated independently from $\mathbf{V} \sim N(\underline{0}, 0.1\mathbb{I})$ (where $\mathbb{I}$ denotes the 2×2 identity matrix). This multivariate normal distribution was used in one of the simulation studies conducted by Duong and Hazelton [2] demonstrating that a fully parameterized bandwidth matrix could yield improved density estimates for joint density estimation with continuous distributions. We discretize this distribution at the half-integers as illustrated in panel (a) of Supplemental Figure 2.

Panel (b) of Supplemental Figure 2 gives a motivating example for using this particular distribution in the simulation study: prediction of incidence at a prediction horizon of one week. In the simplest formulation this task omitting the periodic kernel component and predicting using only the most recent observation, our goal is to estimate the conditional distribution of incidence at time $t + 1$ given incidence at time t. A key feature of the observed disease incidence in our data sets is the high autocorrelation of the time series, which appears as a linear trend in the scatter plot of incidence at adjacent time points. The simulation study examines how the bandwidth matrix parameterization relates to performance of KCDE in estimating conditional densities in the presence of such correlation between the variables being conditioned on and the variables whose density is being estimated.

2.2 Hellinger Distance

The Hellinger distance of the estimated density $\widehat{f}(x)$ from the true density $f(x)$ is given by

$$\text{Hellinger}(f, \widehat{f}) = \left[1 - \int \left\{ f(x)\widehat{f}(x) \right\}^{\frac{1}{2}} \, \mathrm{d}x \right]^{\frac{1}{2}}$$

In the simulation study, we measure the quality of a conditional density estimate by integrating the Hellinger distance over the range of the conditioning variables, weighting according to the density of those conditioning variables:

$$\begin{aligned}
&\text{Score}\{\widehat{f}(x_1 \mid x_2, \ldots, x_D)\} \\
&\quad = \int \cdots \int \left[\text{Hellinger}\{f(x_1 \mid x_2, \ldots, x_D), \widehat{f}(x_1 \mid x_2, \ldots, x_D)\} \right] f(x_2, \ldots, x_D) \mathrm{d}x_2 \cdots \mathrm{d}x_D \\
&\quad = \int \cdots \int \left[1 - \int \left\{ f(x_1 \mid x_2, \ldots, x_D) \widehat{f}(x_1 \mid x_2, \ldots, x_D) \right\}^{\frac{1}{2}} \mathrm{d}x_1 \right]^{\frac{1}{2}} f(x_2, \ldots, x_D) \mathrm{d}x_2 \cdots \mathrm{d}x_D \\
&\quad = \int \cdots \int \left[1 - \int \left\{ \frac{\widehat{f}(x_1 \mid x_2, \ldots, x_D)}{f(x_1 \mid x_2, \ldots, x_D)} \right\}^{\frac{1}{2}} f(x_1 \mid x_2, \ldots, x_D) \, \mathrm{d}x_1 \right]^{\frac{1}{2}} f(x_2, \ldots, x_D) \mathrm{d}x_2 \cdots \mathrm{d}x_D
\end{aligned} \tag{1}$$

We perform Monte Carlo integration to evaluate the integrals in Equation (1) by sampling observations $(x_{i,1}, \ldots, x_{i,D})$ from the joint distribution of $\mathbf{X}$.

3 Application Details

3.1 Prediction Targets

As we discussed in the main article, there are three prediction targets for each data set:

1. For each week in the test data, we obtain a predictive distribution for the incidence measure in that week at each prediction horizon from 1 to 52 weeks ahead.
2. In each week of the test data set, we make predictions for the timing of the peak week of the corresponding season.
3. In each week of the test data set we predict incidence in the peak week for the corresponding season. Following the precedent set in the competitions, we make predictions for *binned* incidence in the peak week.

These prediction targets are illustrated in Supplemental Figure 3.

3.2 HHH4 Model

The HHH4 model for a single infectious disease incidence time series specifies that observed incidence Z_t follows either a Poisson or a Negative Binomial distribution with mean parameterized as

$$\begin{aligned}
E[Z_t] &= \lambda_t Z_{t-l} + \nu_t, \text{ where} \\
\log(\lambda_t) &= \alpha^{(\lambda)} + \sum_{s-1}^{S^{(\lambda)}} \left\{ \gamma_s^{(\lambda)} \sin(\omega_w t) + \delta_s^{(\lambda)} \cos(\omega_s t) \right\} \\
\log(\nu_t) &= \alpha^{(\nu)} + \sum_{s-1}^{S^{(\nu)}} \left\{ \gamma_s^{(\nu)} \sin(\omega_w t) + \delta_s^{(\nu)} \cos(\omega_s t) \right\}
\end{aligned}$$

In these equations, l is a lag to use in the autoregressive term and $S^{(\lambda)}$ and $S^{(\nu)}$ specify the number of sinusoidal terms used to capture seasonality. We used Aikake's Information Criterion (CITE) to perform model selection. We considered all possible model specifications that could be obtained by varying the following four factors:

1. Parametric family: $\{\text{Poisson}, \text{Negative Binomial}\}$
2. $l \in \{1, 2, 3\}$
3. $S^{(\lambda)} \in \{0, 1, 2, 3\}$
4. $S^{(\nu)} \in \{0, 1, 2, 3\}$

This is similar to the approach taken by Held and Paul [3]. The selected model (with lowest AIC among the candidate specifications considered) had a Negative Binomial family, $l = 1$, $S^{(\lambda)} = 2$, and $S^{(\nu)} = 1$.

The surveillance package provides functionality to compute one-step-ahead predictive distributions and to iteratively sample trajectories over multiple time steps [4], but it does not provide functionality to compute the predictive distributions at horizons more than one step ahead. For this article, we used an importance sampling estimate of the predictive density at horizons $h \geq 2$:

$$\begin{aligned} P(Z_{t+h} = z_{t+h} \mid z_t) &= \iint P(Z_{t+h} = z_{t+h}, \ldots, Z_{t+1} = z_{t+1} \mid z_t) \, \mathrm{d}z_{t+1} \cdots \mathrm{d}z_{t+h-1} \\ &\approx \frac{1}{J} \sum_{j=1}^{J} P(Z_{t+h} = z_{t+h} \mid z_{t+h-1}^{(j)}, \ldots, z_{t+1}^{(j)}, z_t), \text{ where} \end{aligned}$$

$(z_{t+h-1}^{(j)}, \ldots, z_{t+1}^{(j)})$, $j = 1, \ldots, J$ are sampled from the joint distribution of $(Z_{t+h-1}, \ldots, Z_{t+1}) \mid z_t$.

3.3 Predictive Distributions for Individual Weeks: Additional Results

Here we present some additional results for predicting incidence in individual weeks in the applications. Supplemental Figure **??** shows that including the periodic kernel in the KCDE specification yielded consistent performance gains in the application to influenza. The performance gains in the application to dengue fever were smaller, but average performance was still higher when the periodic kernel was included. The figure also shows that the gains from using a fully parameterized bandwidth instead of a diagonal bandwidth are negligible, though there is a small gain on average in the application to influenza.

3.4 Predictive Distributions for Peak Week and Peak Incidence: Additional Results

Figure 6 in the main text shows log scores for prediction of incidence in the peak week. Supplemental Figure **??** in this supplement shows the corresponding results for prediction of peak week timing. As with predictions of peak incidence, there is no clear evidence that KCDE either outperforms or underperforms relative to the SARIMA model. The log scores give us information about the values that the predictive distributions take at a single value: the eventual realized outcome. Supplemental Figures **??** through **??** give more information, about the predictive distributions for peak week height and timing obtained from SARIMA and the Periodic, Full Bandwidth KCDE specification.

As we discussed in the main text, the predictive distributions for peak week timing and incidence are obtained by performing an appropriate Monte Carlo integral of the joint distribution for incidence in all remaining weeks in the season. In more plain language, we sample incidence trajectories from the joint predictive distribution of incidence in all remaining weeks and calculate the proportion of those sampled trajectories where the peak fell in each incidence bin or at each week of the season.

Supplemental Figure 7 illustrates this with the Periodic, Full Bandwidth KCDE specification and the SARIMA model. For reference, we have also included all observed trajectories for the seasons in the training and test data sets and trajectories sampled from the predictive distribution that would be obtained by combining the KCDE predictions at different horizons using an independence assumption instead of a copula. We can see that the effect of the copula is to induce correlation in the incidence across different weeks. The trajectories obtained with the copula are much smoother than the trajectories obtained with an independence assumption.

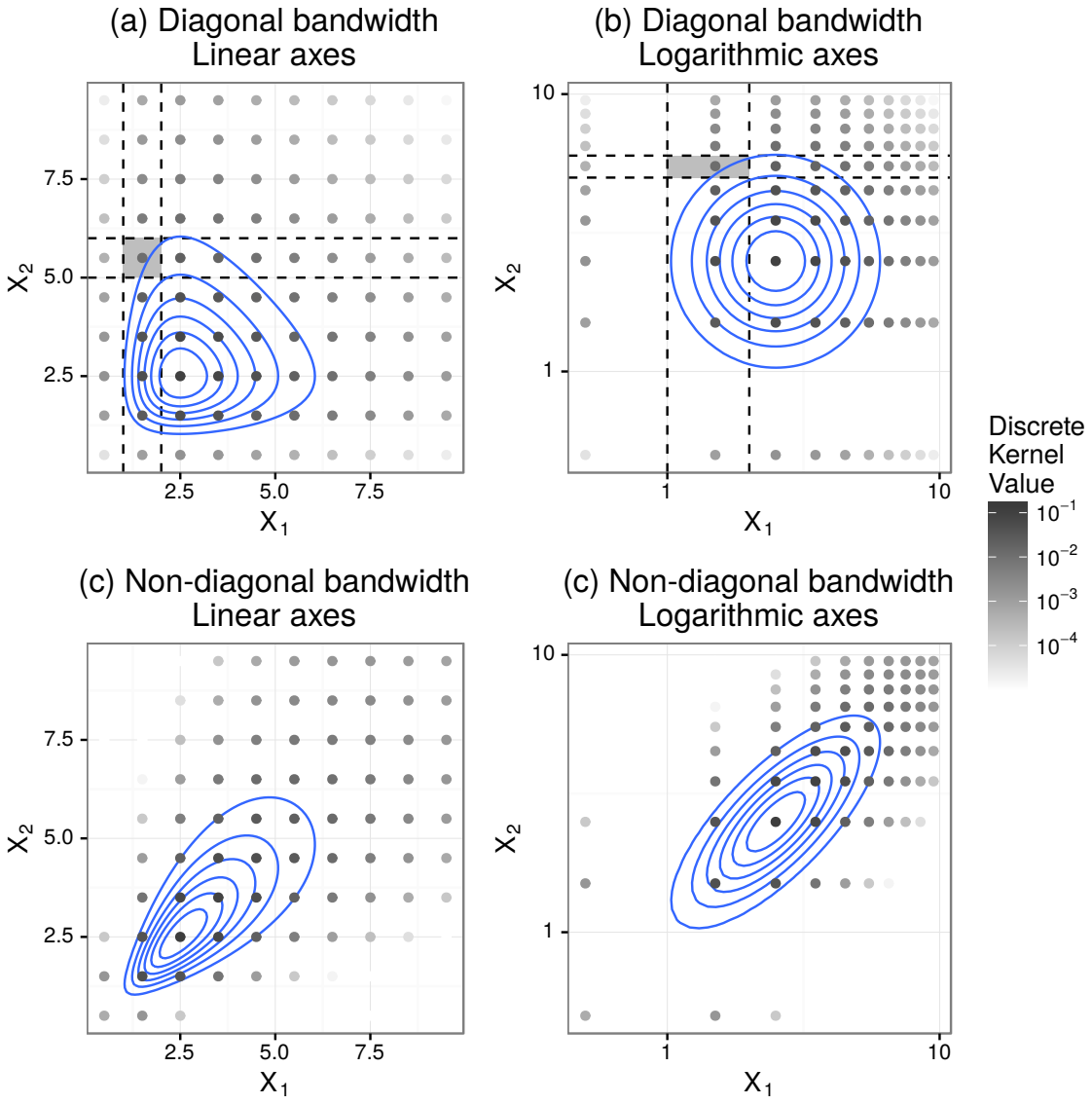
References

[1] Joe H. Asymptotic efficiency of the two-stage estimation method for copula-based models. *Journal of Multivariate Analysis* 2005; **94**(2):401–419.

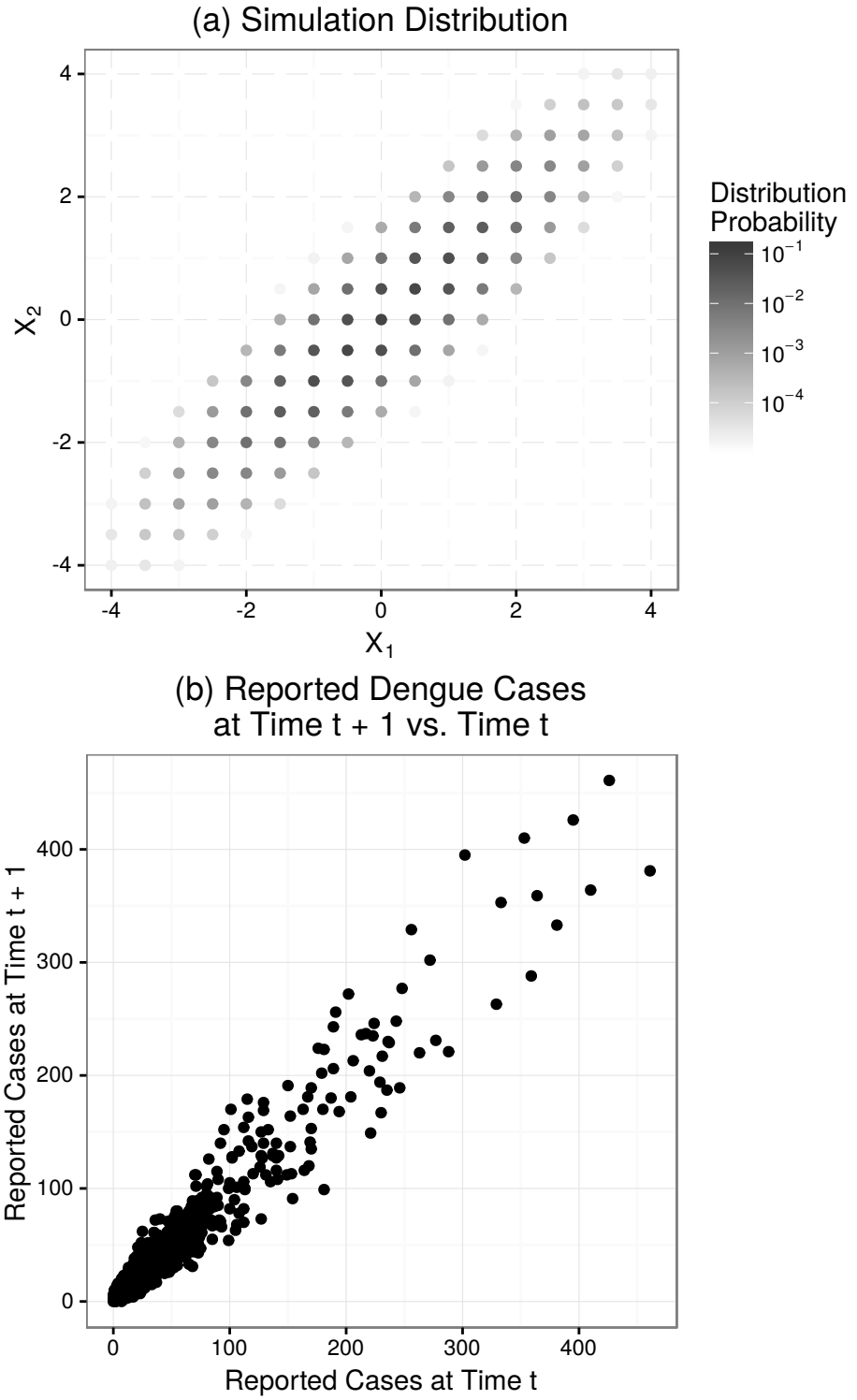
[2] Duong T, Hazelton ML. Cross-validation bandwidth matrices for multivariate kernel density estimation. *Scandinavian Journal of Statistics* 2005; **32**(3):485–506.

[3] Held L, Paul M. Modeling seasonality in space-time infectious disease surveillance data. *Biometrical Journal* 2012; **54**(6):824–843.

[4] Meyer S, Held L, Höhle M. Spatio-temporal analysis of epidemic phenomena using the R package surveillance. *Journal of Statistical Software* 2016; :(to appear).



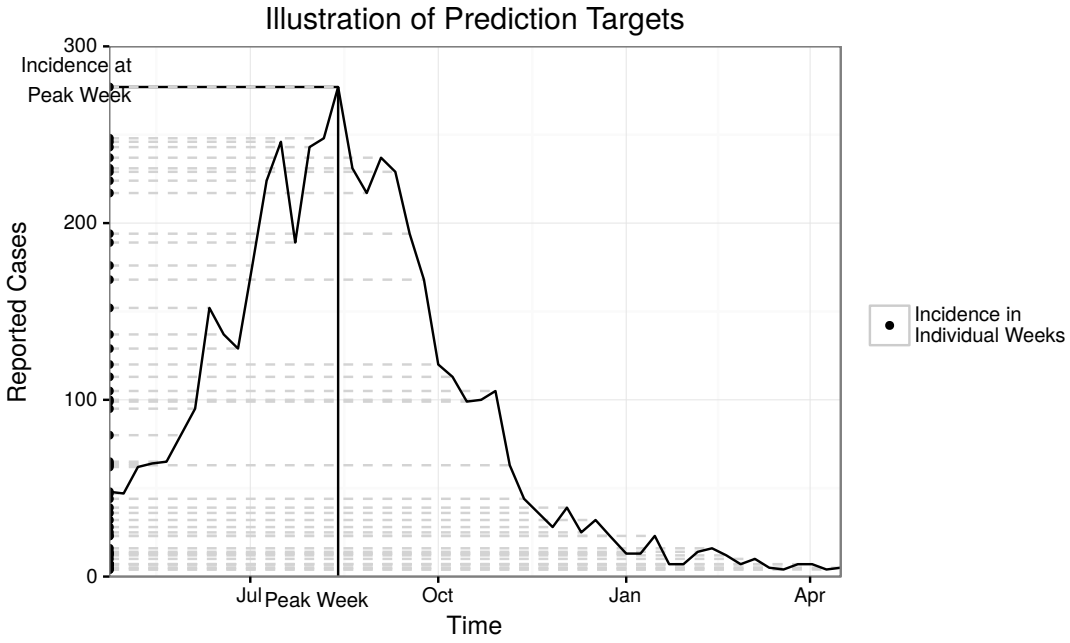
Supplemental Figure 1: Illustrations of $K^{\text{inc}}_{\text{cont}}$ and $K^{\text{inc}}_{\text{disc}}$ in the bivariate case. Solid lines show contours of the continuous kernel function. Grey dots indicate the value of the discrete kernel function. The value of the discrete kernel is obtained by integrating the continuous kernel over regions as illustrated by the dashed lines in panels (a) and (b). In all panels the kernel function is centered at $(2.5, 2.5)$. Panels (a) and (b) show the same kernel function with different axis scales; the bandwidth matrix is $\begin{bmatrix} 0.2 & 0 \\ 0 & 0.2 \end{bmatrix}$. Panels (c) and (d) show the same kernel function, with bandwidth matrix $\begin{bmatrix} 0.2 & 0.15 \\ 0.15 & 0.2 \end{bmatrix}$.



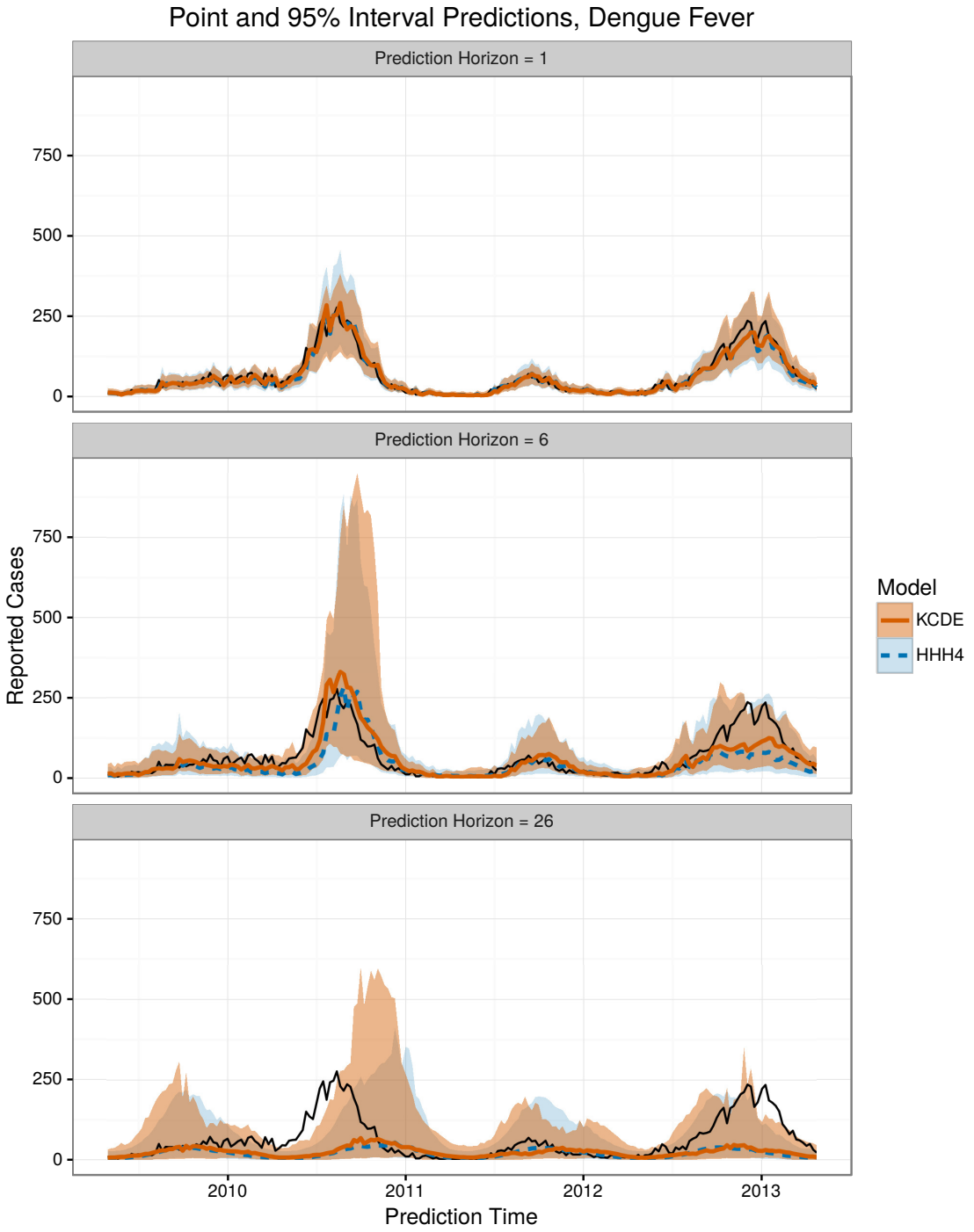
Supplemental Figure 2: Panel (a) shows the distribution that we simulate data from in the simulation study. Panel (b) shows an example motivating the choice of distribution for the simulation study: reported dengue cases at time $t+1$ vs. at time t.

Disease	KCDE Specification	Baseline	Subset	Summary of Log Score Differences					
				Min	Q1	Q2	Mean	Q3	Max
Dengue	Null	HHH4	All Weeks	-1.60	-0.51	-0.26	0.22	0.39	8.89
			Low Incidence	-1.32	-0.51	-0.29	0.05	0.08	5.05
			High Incidence	-1.60	-0.23	0.49	0.81	1.45	6.29
Dengue	Null	SARIMA	All Weeks	-2.46	-0.37	-0.03	0.31	0.57	6.47
			Low Incidence	-2.46	-0.35	-0.07	0.17	0.37	5.39
			High Incidence	-1.57	-0.35	0.94	1.15	2.23	5.92
Dengue	Full Bandwidth	HHH4	All Weeks	-1.57	-0.50	-0.24	0.20	0.39	8.29
			Low Incidence	-1.50	-0.50	-0.28	0.05	0.10	4.91
			High Incidence	-1.57	-0.30	0.42	0.71	1.32	6.11
Dengue	Full Bandwidth	SARIMA	All Weeks	-2.64	-0.34	-0.01	0.29	0.56	6.22
			Low Incidence	-2.64	-0.31	-0.04	0.18	0.37	5.04
			High Incidence	-2.02	-0.33	0.66	1.04	2.13	5.80
Dengue	Periodic	HHH4	All Weeks	-2.47	-0.24	-0.02	0.35	0.49	7.21
			Low Incidence	-2.47	-0.27	-0.09	0.14	0.18	4.19
			High Incidence	-1.40	0.27	0.90	1.21	1.98	6.05
Dengue	Periodic	SARIMA	All Weeks	-2.42	-0.15	0.13	0.43	0.66	6.10
			Low Incidence	-2.42	-0.14	0.09	0.26	0.47	4.30
			High Incidence	-1.72	0.07	1.48	1.54	2.70	6.10
Dengue	Periodic, Full Bandwidth	HHH4	All Weeks	-2.09	-0.25	0.00	0.35	0.57	7.23
			Low Incidence	-2.09	-0.30	-0.09	0.13	0.22	3.98
			High Incidence	-1.28	0.41	0.94	1.26	1.95	5.68
Dengue	Periodic, Full Bandwidth	SARIMA	All Weeks	-2.16	-0.13	0.16	0.44	0.66	6.12
			Low Incidence	-2.16	-0.14	0.11	0.25	0.47	4.15
			High Incidence	-1.60	0.25	1.48	1.59	2.72	6.12
Influenza	Null	SARIMA	All Weeks	-2.26	-0.62	-0.39	-0.37	-0.16	3.56
			Low Incidence	-1.60	-0.59	-0.35	-0.34	-0.14	3.56
			High Incidence	-2.26	-1.01	-0.66	-0.54	-0.28	3.09
Influenza	Full Bandwidth	SARIMA	All Weeks	-2.83	-0.56	-0.33	-0.33	-0.14	3.79
			Low Incidence	-1.43	-0.46	-0.27	-0.26	-0.10	3.79
			High Incidence	-2.83	-1.11	-0.79	-0.68	-0.36	2.89
Influenza	Periodic	SARIMA	All Weeks	-1.72	-0.21	-0.04	0.00	0.15	3.90
			Low Incidence	-1.68	-0.17	-0.01	0.03	0.17	3.90
			High Incidence	-1.72	-0.57	-0.10	-0.06	0.26	3.18
Influenza	Periodic, Full Bandwidth	SARIMA	All Weeks	-2.18	-0.20	0.00	0.02	0.19	4.07
			Low Incidence	-1.67	-0.13	0.04	0.07	0.21	4.07
			High Incidence	-2.18	-0.58	-0.16	-0.10	0.24	3.11

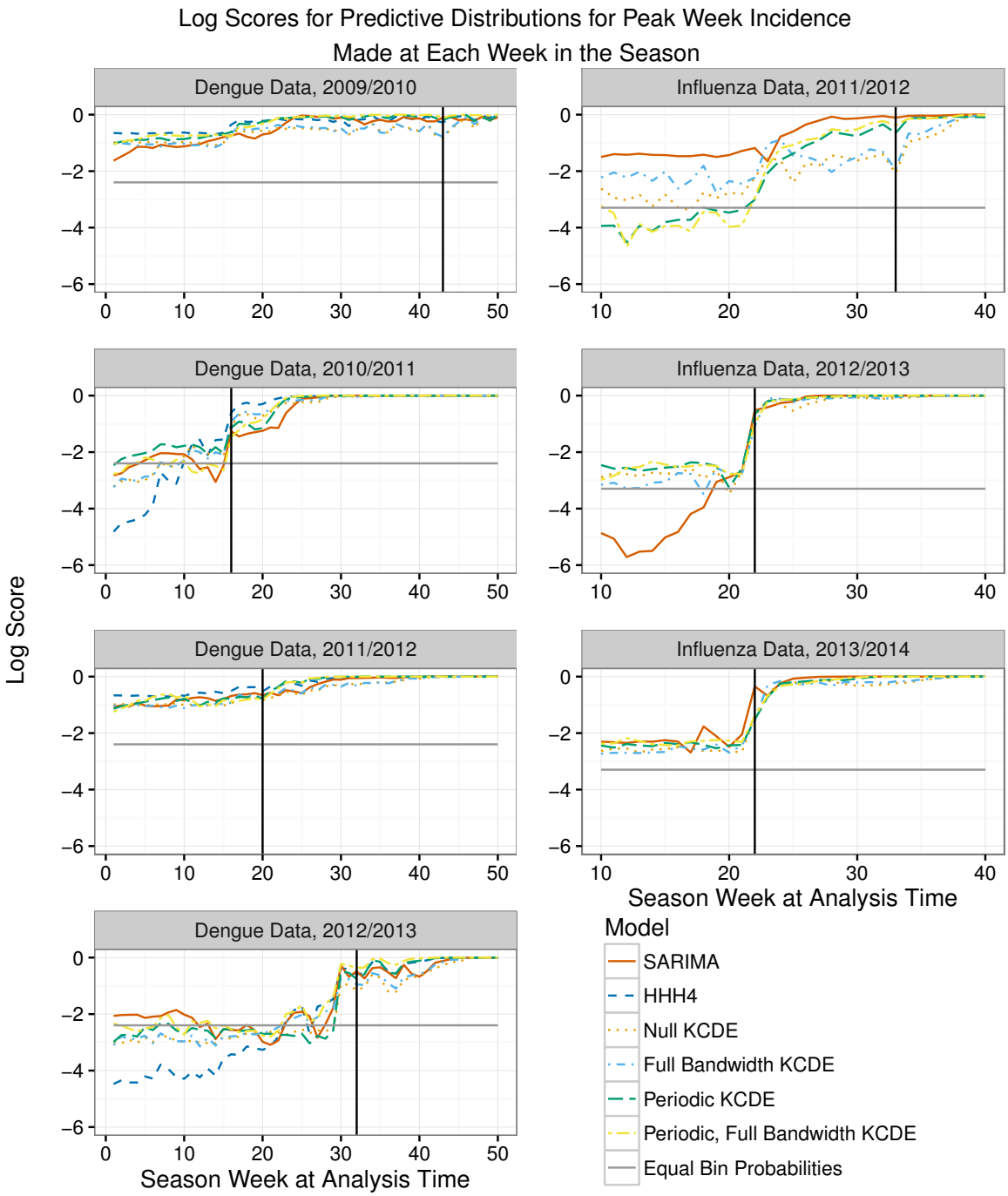
Supplemental Table 1: Summaries of model performance relative to the baseline models for predictions of incidence in individual weeks. Each row summarizes the log score differences for predictions of incidence made by one of the KCDE specifications and one of the baseline models. The first row for each model pair summarizes results for all combinations of target week in the test period and prediction horizon; the "Low Incidence" rows summarize results for predictions in weeks where the observed incidence in the target week was less than one third of the maximum weekly incidence in the test period; the "High Incidence" rows summarize results for weeks where the observed incidence was at least two thirds of the maximum weekly incidence in the test period.



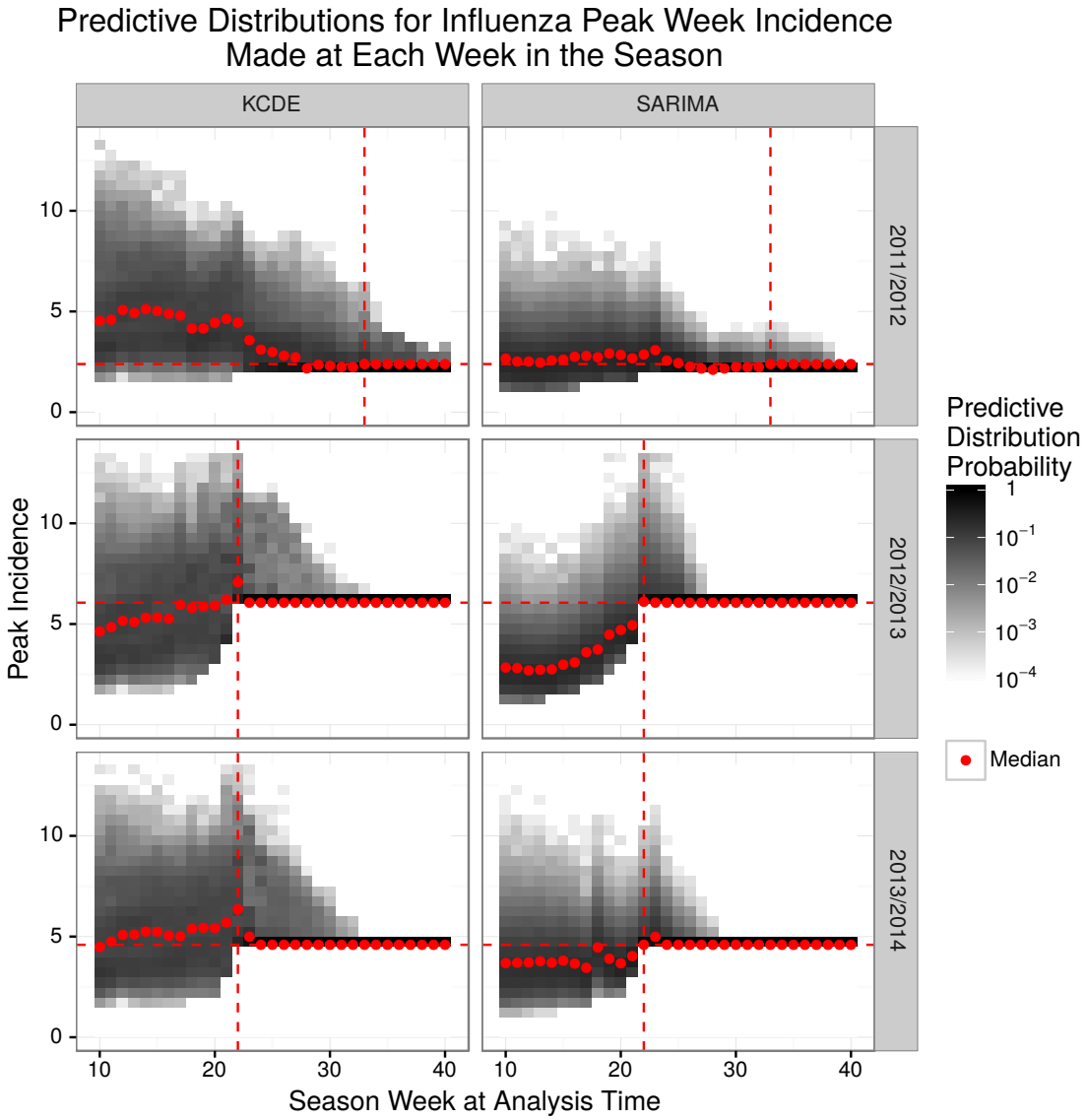
Supplemental Figure 3: Illustration of the prediction targets using one season of the dengue data. The solid vertical line indicates the timing of the peak week. The solid horizontal line indicates the incidence at the peak week. The points along the vertical axis indicate the incidence at every week for the 52 weeks after the time at which predictions are made.



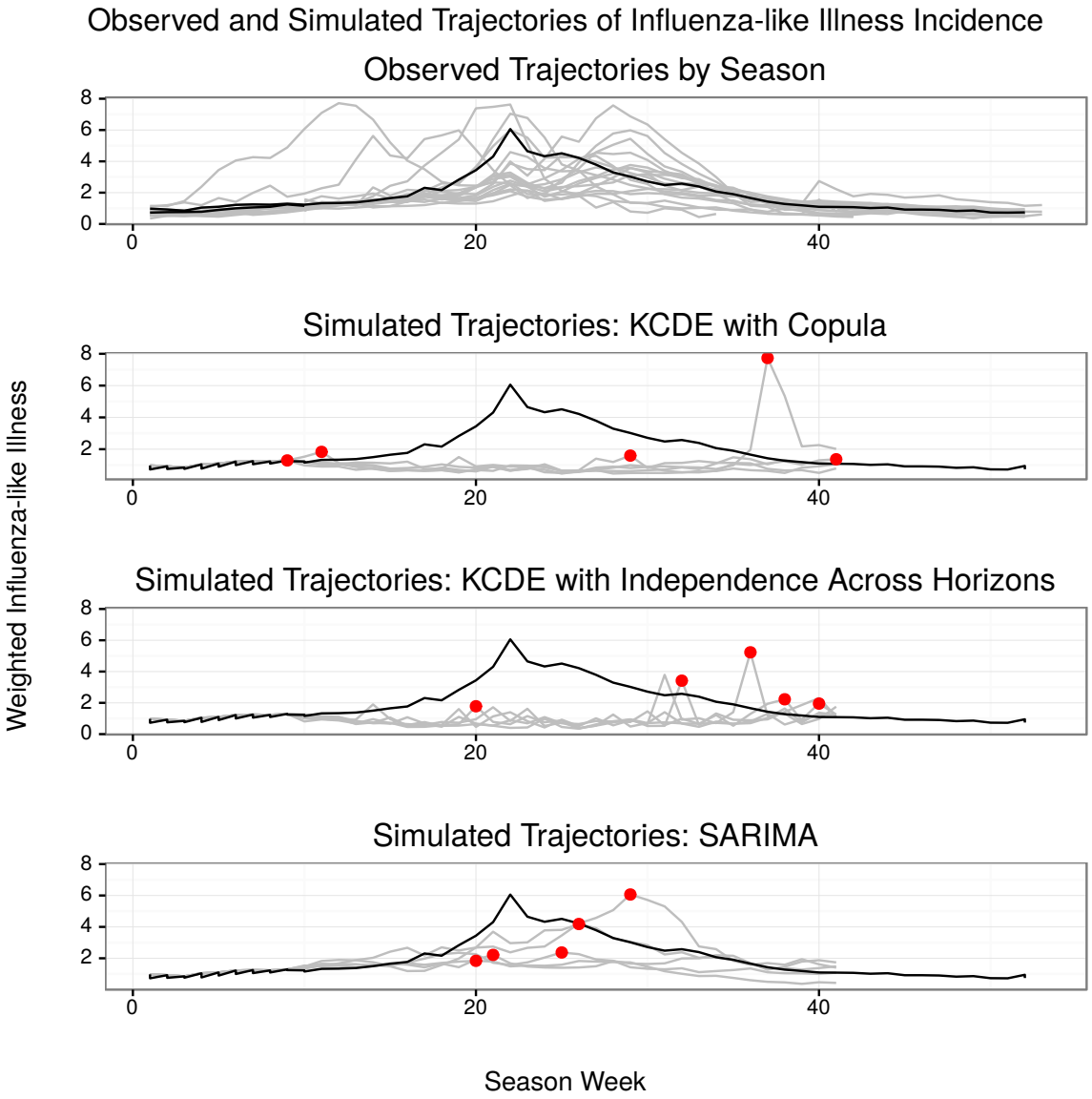
Supplemental Figure 4: Plots of point and interval predictions from HHH4 and the Periodic, Full Bandwidth KCDE model.



Supplemental Figure 5: Log scores for predictions of peak week incidence by predictive model and analysis time. The vertical line is placed at the peak week for each season. The log score for "Equal Bin Probabilities" is obtained by assigning equal probability that the peak incidence will be in each of the specified incidence bins. There are 11 incidence bins for dengue and 27 bins for influenza.



Supplemental Figure 6: Predictive distributions for predictions of peak week incidence for influenza. The horizontal axis represents the week in the season at which the prediction is made. The vertical axis represents binned incidence in the peak week, as described in the main text. Each "column" represents one predictive distribution. The horizontal dashed line is at the observed peak incidence for the season. The vertical dashed line is at the observed peak week for the season. The medians are calculated based on the unbinned predictions. The upper bin extends to infinity; we have cut it off at 13.5 for purposes of the display.



Supplemental Figure 7: Incidence trajectories for the influenza data set. The top panel displays the observed trajectories for all seasons in the data set, with the 2012/2013 season in darker color. The lower three panels display the observed trajectory from the 2012/2013 season and five simulated incidence trajectories from each of three models: the KCDE model with copula as implemented in our applications; a KCDE model using an independence assumption across prediction horizons; and the SARIMA model. The simulated trajectories are generated from the predictive distribution obtained 10 weeks into the 2012/2013 season. The red points indicate the peak week in each simulated trajectory.