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 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:

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

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. 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

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

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:

$$\begin{aligned}
&\text{Score}\{\widehat{f}(x_1|x_2,\ldots,x_D)\} \\
&\quad = \int\cdots\int \left[\text{Hellinger}\{f(x_1|x_2,\ldots,x_D), \widehat{f}(x_1|x_2,\ldots,x_D)\}\right] f(x_2,\ldots,x_D)dx_2\cdots dx_D \\
&\quad = \int\cdots\int \left[1 - \int \left\{f(x_1|x_2,\ldots,x_D)\widehat{f}(x_1|x_2,\ldots,x_D)\right\}^{\frac{1}{2}} dx_1\right]^{\frac{1}{2}} f(x_2,\ldots,x_D)dx_2\cdots dx_D \\
&\quad = \int\cdots\int \left[1 - \int \left\{\frac{\widehat{f}(x_1|x_2,\ldots,x_D)}{f(x_1|x_2,\ldots,x_D)}\right\}^{\frac{1}{2}} f(x_1|x_2,\ldots,x_D)\, dx_1\right]^{\frac{1}{2}} f(x_2,\ldots,x_D)dx_2\cdots dx_D
\end{aligned} \tag{1}$$

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 $h \geq 2$:

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

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

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 6 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 12 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 14 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.

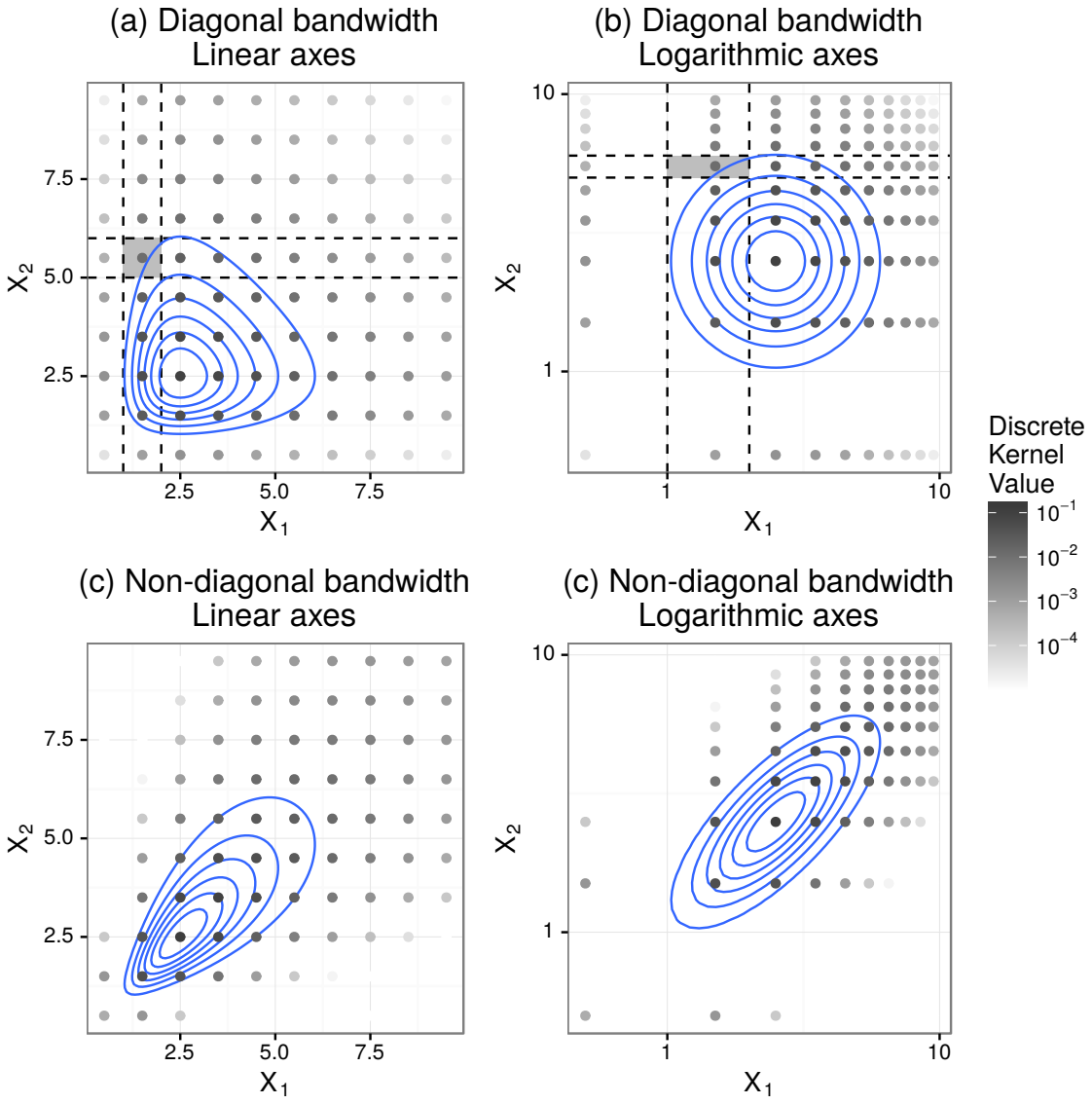
References

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

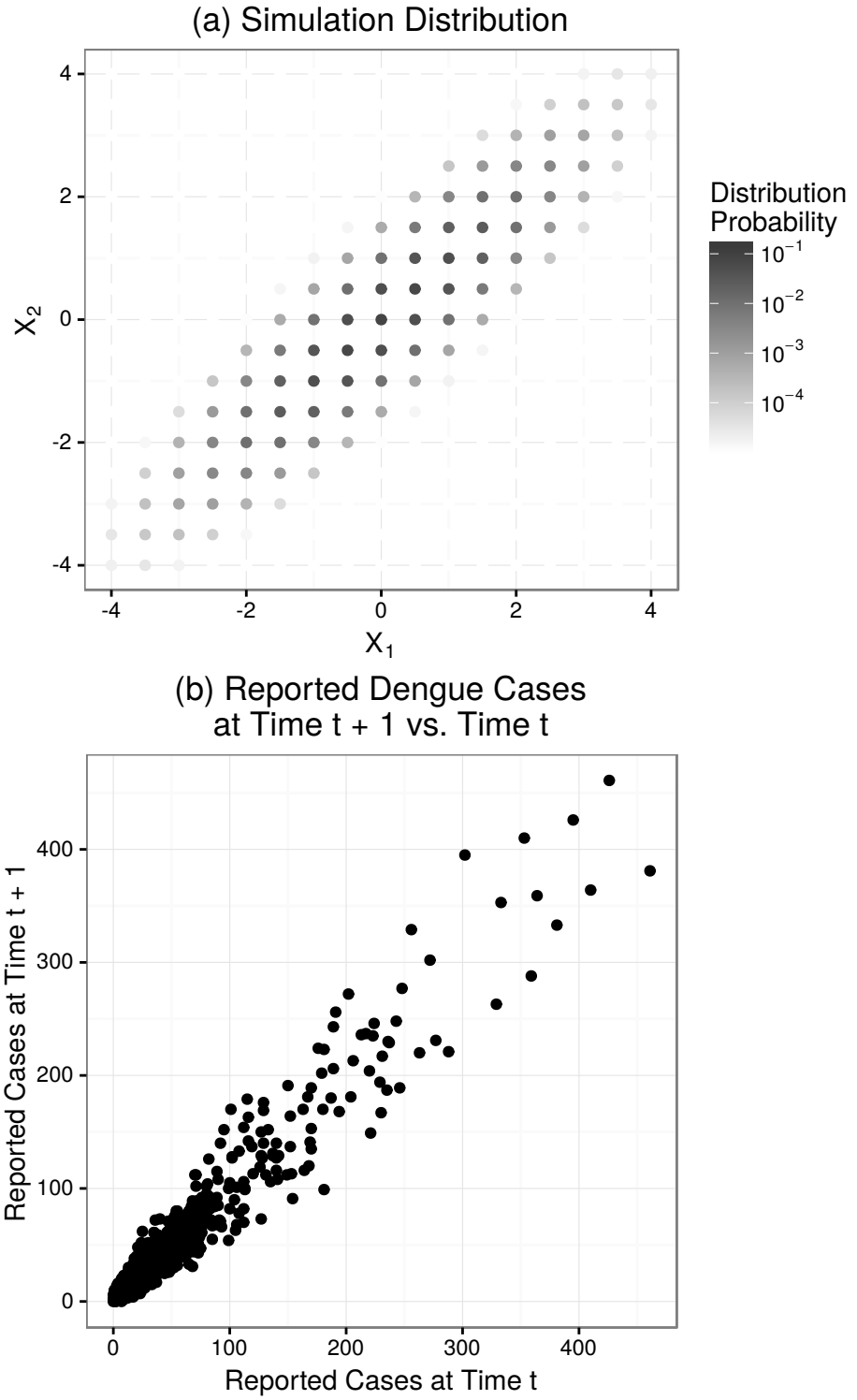
[2] Duong T, Hazelton ML. Cross-validation bandwidth matrices for multivariate kernel density estimation. *Scandinavian Journal of Statistics* 2005; **32**(3):485–506.

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

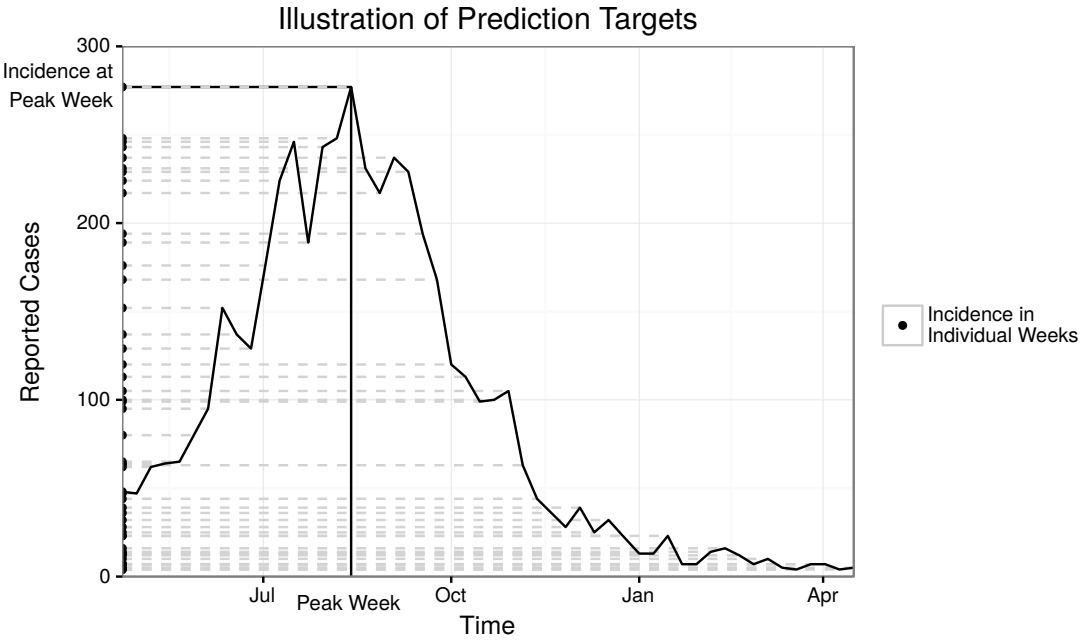
[4] Meyer S, Held L, Höhle M. Spatio-temporal analysis of epidemic phenomena using the R package surveillance. *Journal of Statistical Software* 2016; :(to appear).



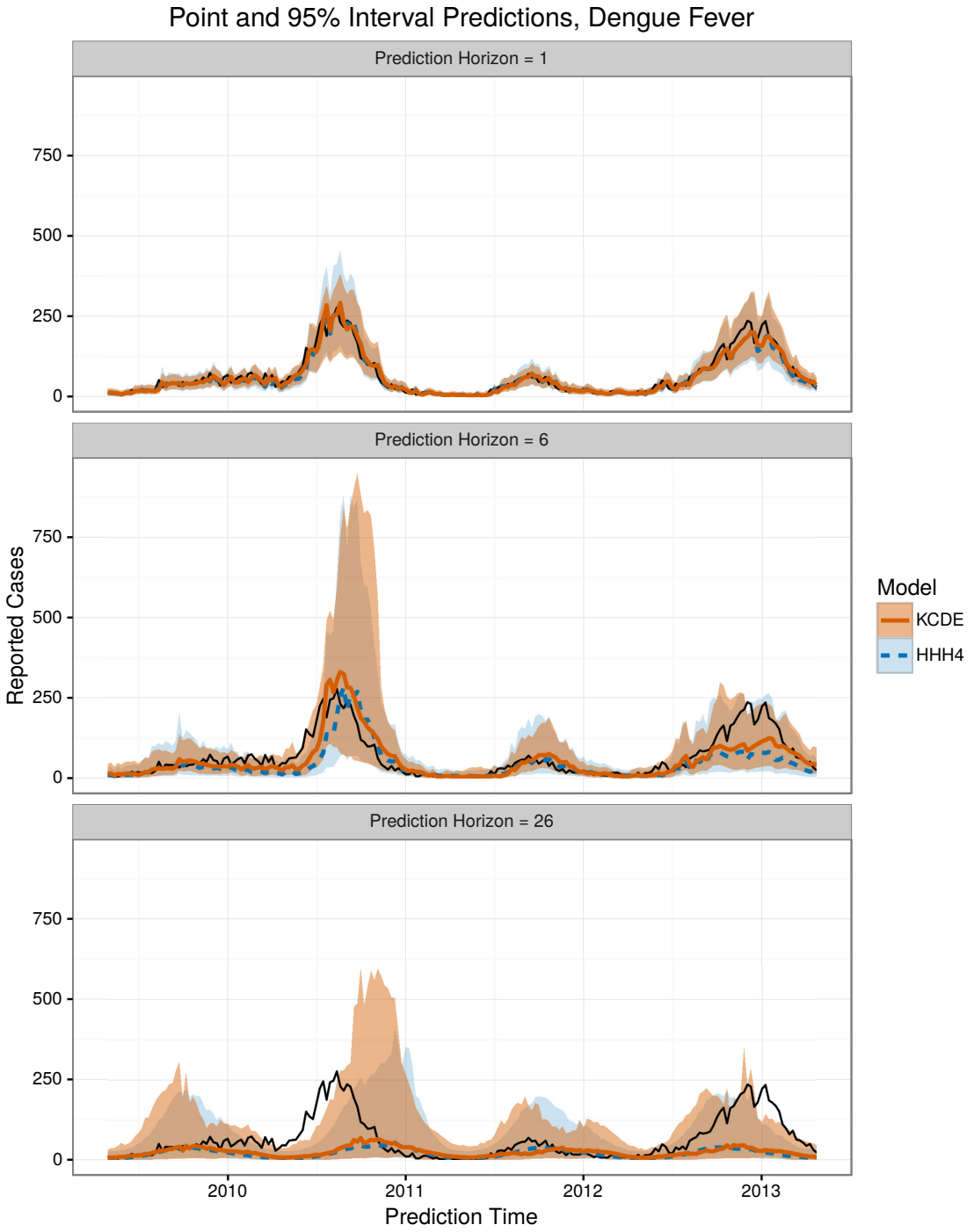
Supplemental Figure 1: Illustrations of $K^{\text{inc}}_{\text{cont}}$ and $K^{\text{inc}}_{\text{disc}}$ in the bivariate case. Solid lines show contours of the continuous kernel function. Grey dots indicate the value of the discrete kernel function. The value of the discrete kernel is obtained by integrating 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 $\begin{bmatrix} 0.2 & 0.15 \\ 0.15 & 0.2 \end{bmatrix}$.



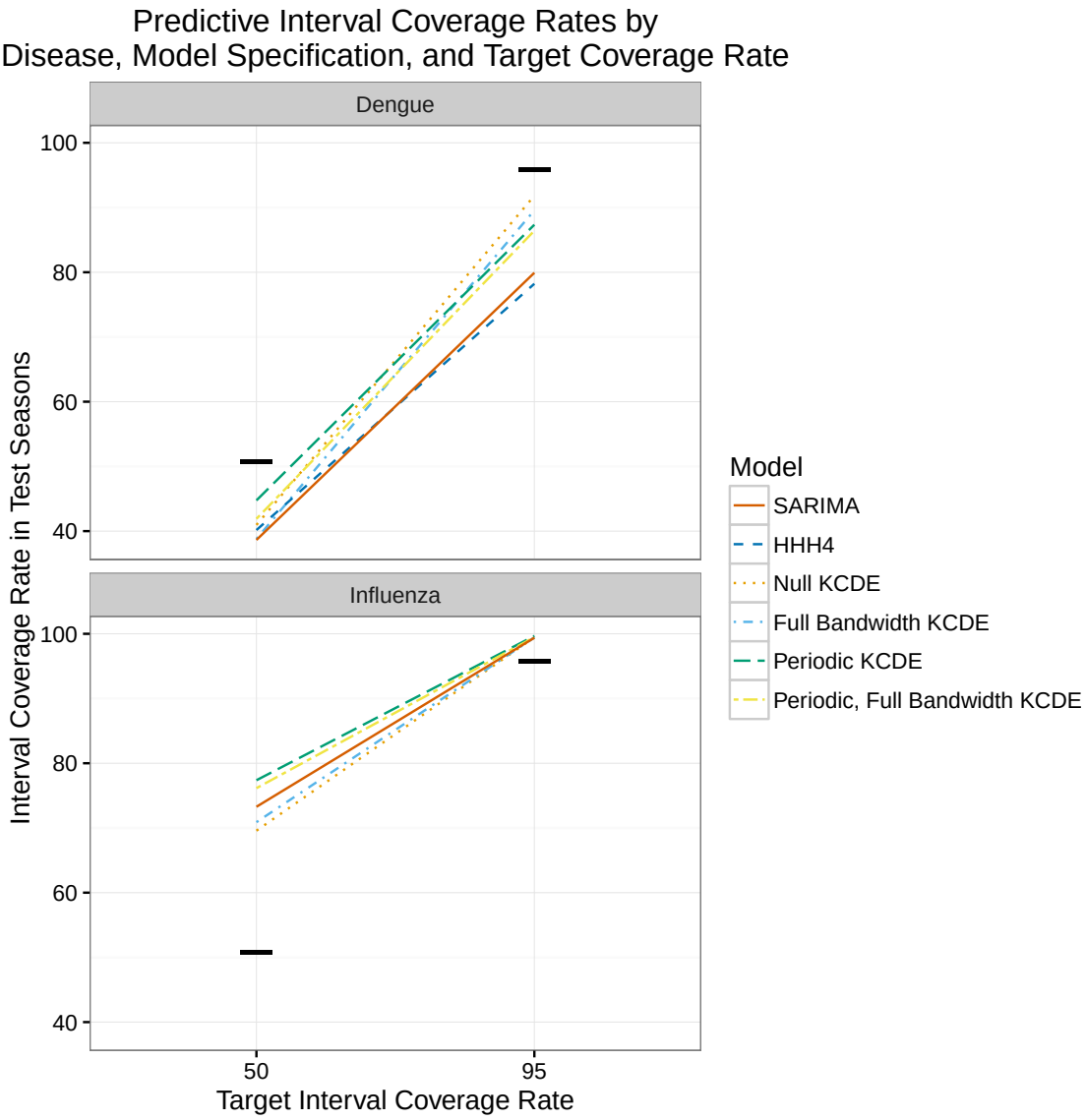
Supplemental Figure 2: Panel (a) shows the distribution that we simulate data from in the simulation study. Panel (b) shows an example motivating the choice of distribution for the simulation study: reported dengue cases at time $t+1$ vs. at time t.



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.



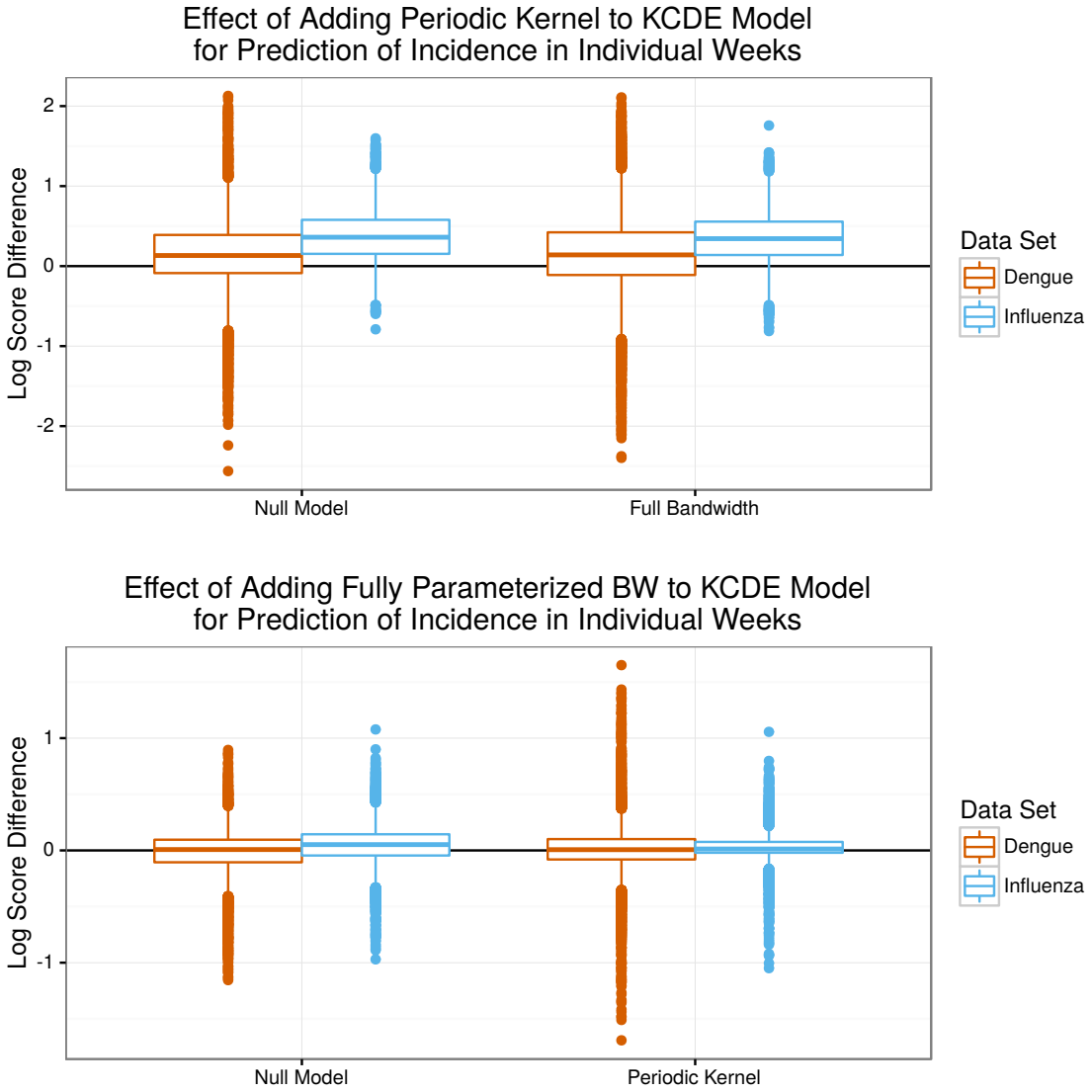
Supplemental Figure 4: Plots of point and interval predictions from HHH4 and the Periodic, Full Bandwidth KCDE model.



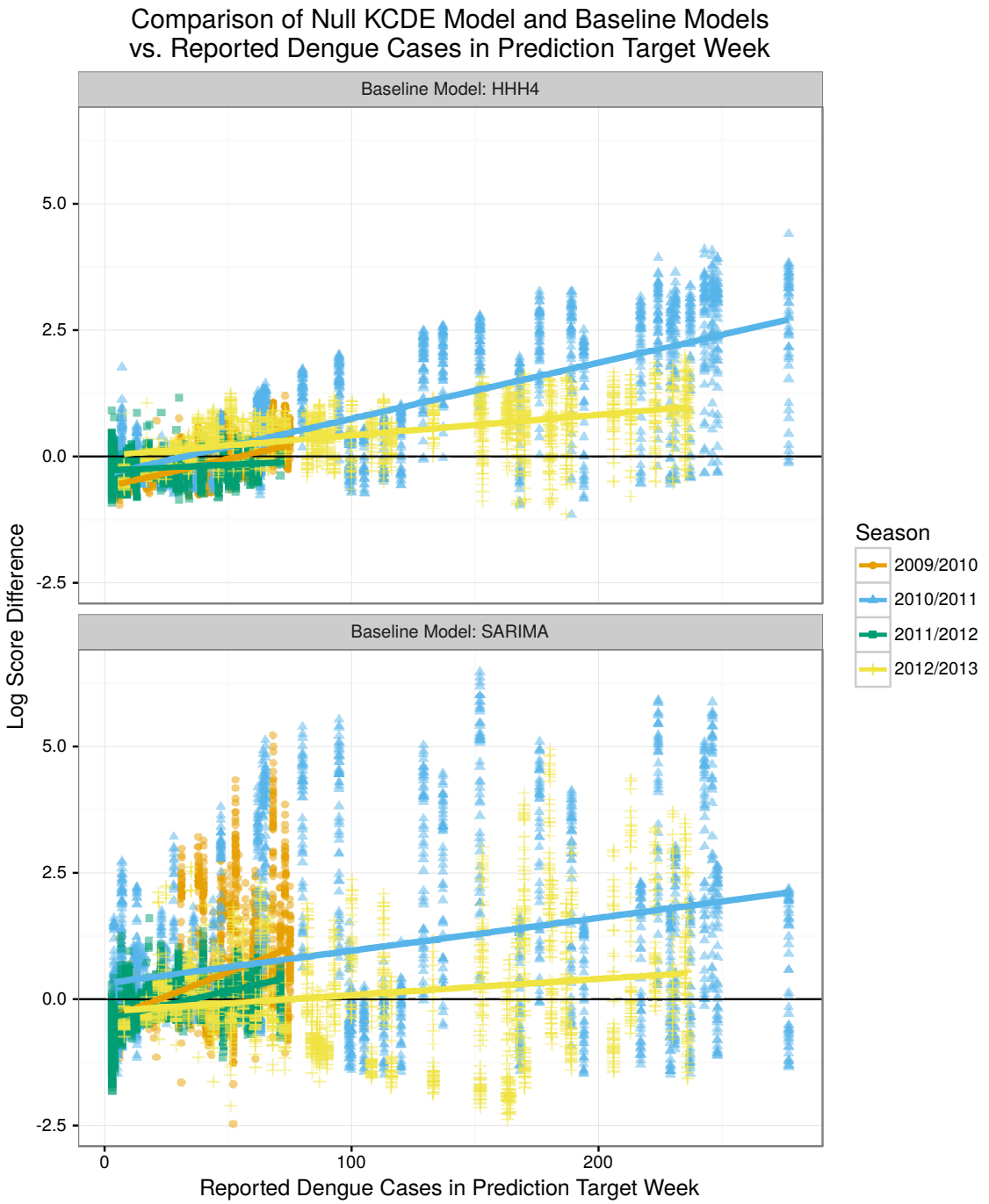
Supplemental Figure 5: Coverage rates for predictions of disease incidence in individual weeks during the test time frame. For each model specification, we have obtained the overall proportion of predictive intervals that contained the realized outcome, combining across all prediction horizons and all times in the test period at which the prediction was made.

Disease	Model	Target Coverage	Actual Coverage
Dengue	Null KCDE	50	40.96
Dengue	Full Bandwidth KCDE	50	38.79
Dengue	Periodic KCDE	50	**44.75**
Dengue	Periodic, Full Bandwidth KCDE	50	41.9
Dengue	HHH4	50	40.16
Dengue	SARIMA	50	38.64
Dengue	Null KCDE	95	**91.83**
Dengue	Full Bandwidth KCDE	95	89.57
Dengue	Periodic KCDE	95	87.34
Dengue	Periodic, Full Bandwidth KCDE	95	86.42
Dengue	HHH4	95	78.22
Dengue	SARIMA	95	79.92
Influenza	Null KCDE	50	**69.58**
Influenza	Full Bandwidth KCDE	50	70.9
Influenza	Periodic KCDE	50	77.37
Influenza	Periodic, Full Bandwidth KCDE	50	76.15
Influenza	SARIMA	50	73.27
Influenza	Null KCDE	95	99.46
Influenza	Full Bandwidth KCDE	95	99.42
Influenza	Periodic KCDE	95	99.68
Influenza	Periodic, Full Bandwidth KCDE	95	99.48
Influenza	SARIMA	95	**99.38**

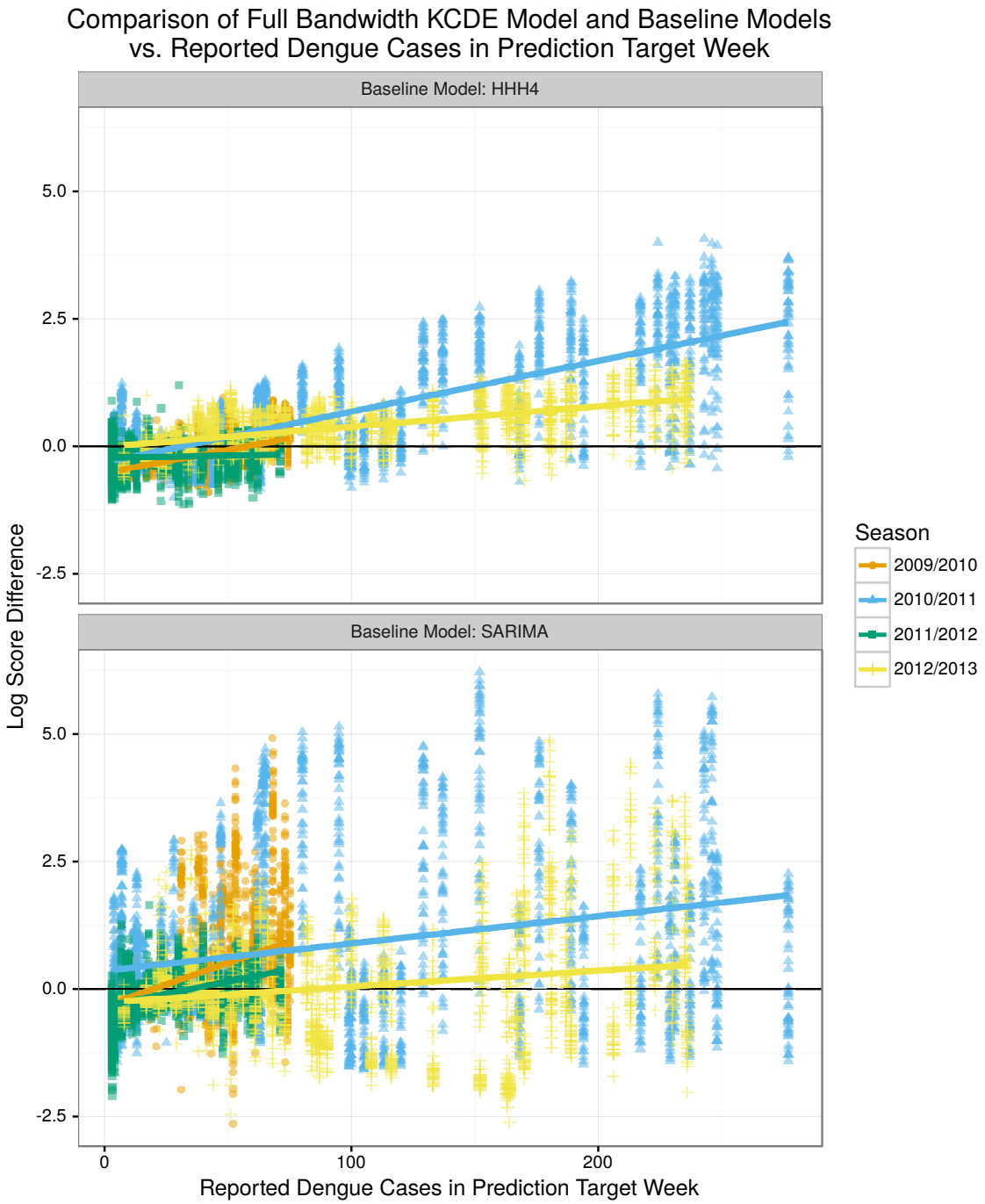
Supplemental Table 1: Coverage rates for predictions of disease incidence in individual weeks during the test time frame. For each model specification, we have obtained the overall proportion of predictive intervals that contained the realized outcome, combining across all prediction horizons and all times in the test period at which the prediction was made. For each combination of disease and target coverage rate, the result for the model with actual coverage rate closest to the target coverate rate is highlighted.



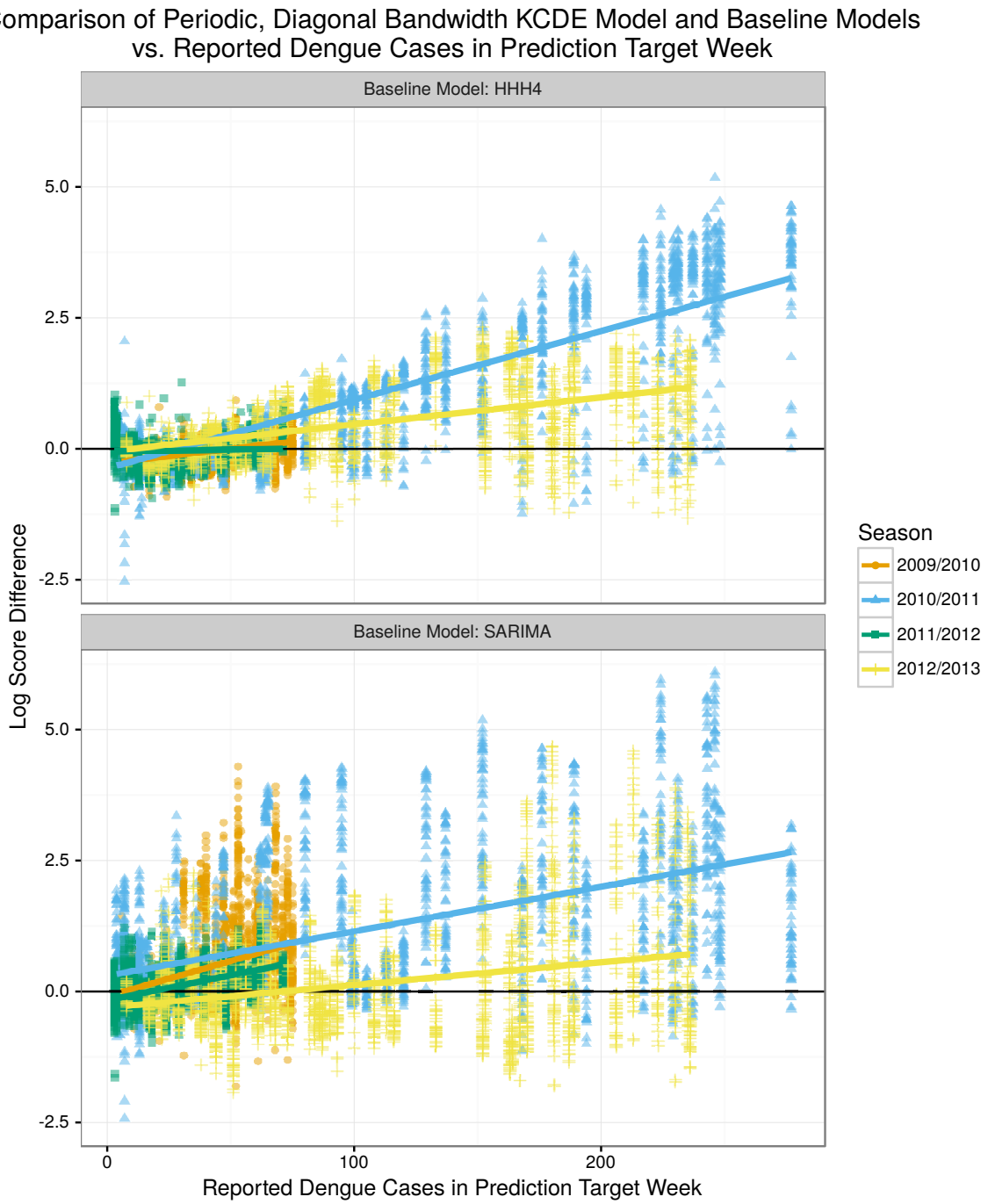
Supplemental Figure 6: Differences in log scores for the weekly predictive distributions among pairs of models across all combinations of prediction horizon and prediction time in the test period. In the upper panel positive values indicate cases when the specification of KCDE with the periodic kernel outperformed the corresponding specification without the periodic kernel. In the lower panel positive values indicate cases when the specification of KCDE with a fully parameterized bandwidth outperformed the corresponding KCDE specification with a diagonal bandwidth matrix.



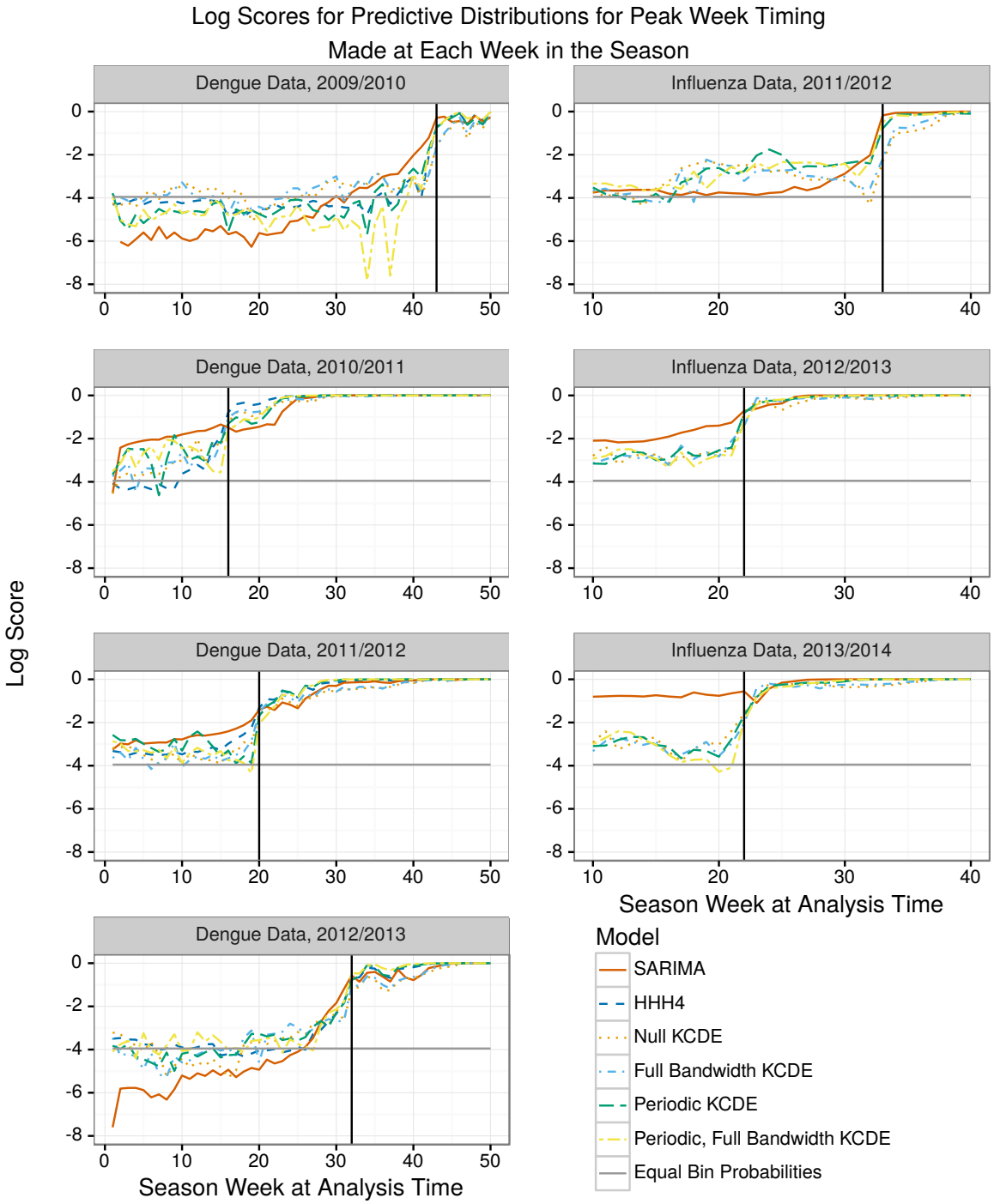
Supplemental Figure 7: Differences in log scores for the weekly predictive distributions obtained from the Null KCDE model and the baseline models, plotted against the observed incidence in the week being predicted. For reference, a log score difference of 2.3 (4.6) indicates that the predictive density from KCDE was about 10 (100) times as large as the predictive density from the baseline model at the realized outcome. Each point corresponds to a unique combination of prediction target week and prediction horizon.



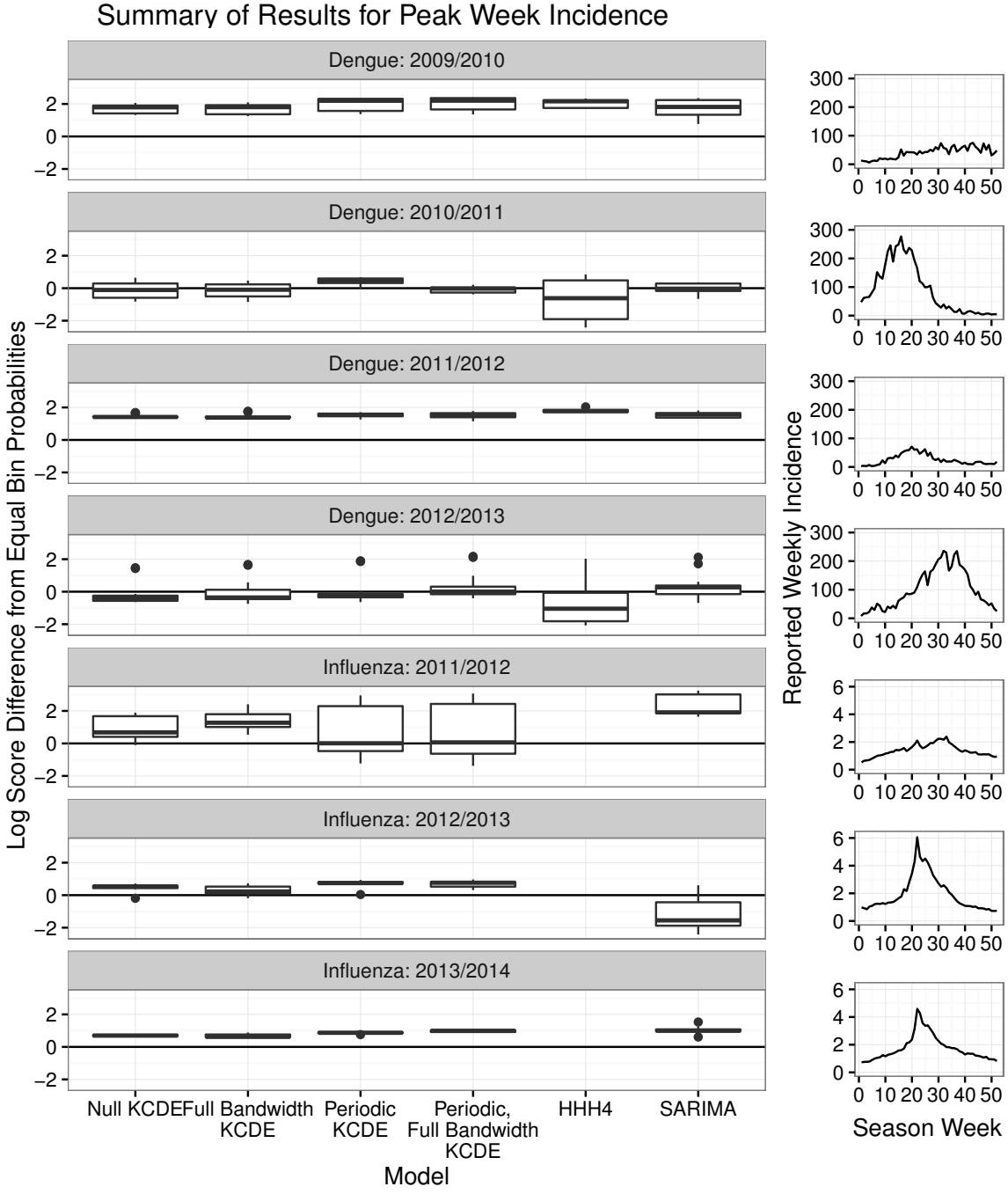
Supplemental Figure 8: Differences in log scores for the weekly predictive distributions obtained from the Full Bandwidth KCDE model and the baseline models, plotted against the observed incidence in the week being predicted. For reference, a log score difference of 2.3 (4.6) indicates that the predictive density from KCDE was about 10 (100) times as large as the predictive density from the baseline model at the realized outcome. Each point corresponds to a unique combination of prediction target week and prediction horizon.



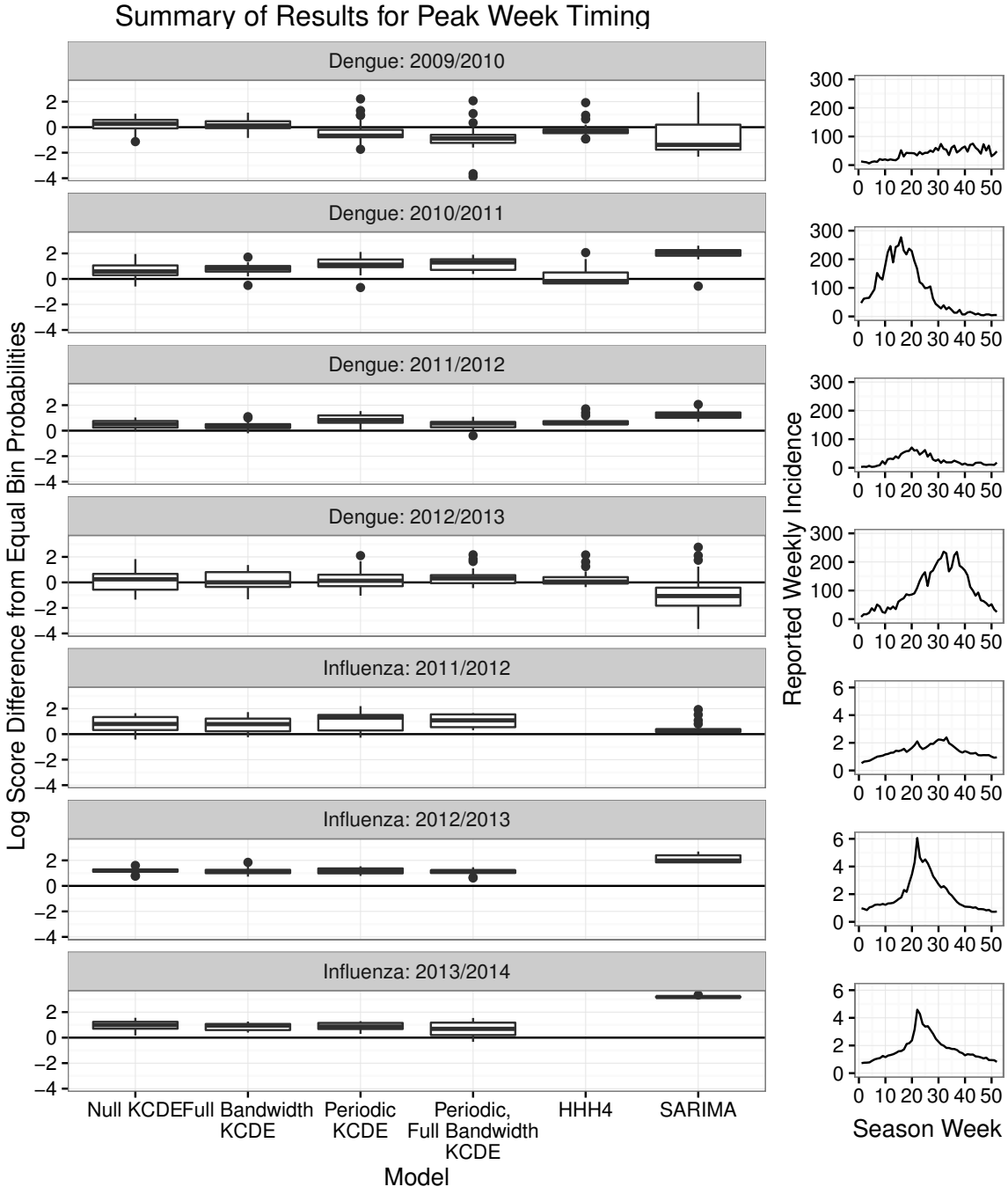
Supplemental Figure 9: Differences in log scores for the weekly predictive distributions obtained from the Periodic, Diagonal Bandwidth KCDE model and the baseline models, plotted against the observed incidence in the week being predicted. For reference, a log score difference of 2.3 (4.6) indicates that the predictive density from KCDE was about 10 (100) times as large as the predictive density from the baseline model at the realized outcome. Each point corresponds to a unique combination of prediction target week and prediction horizon.



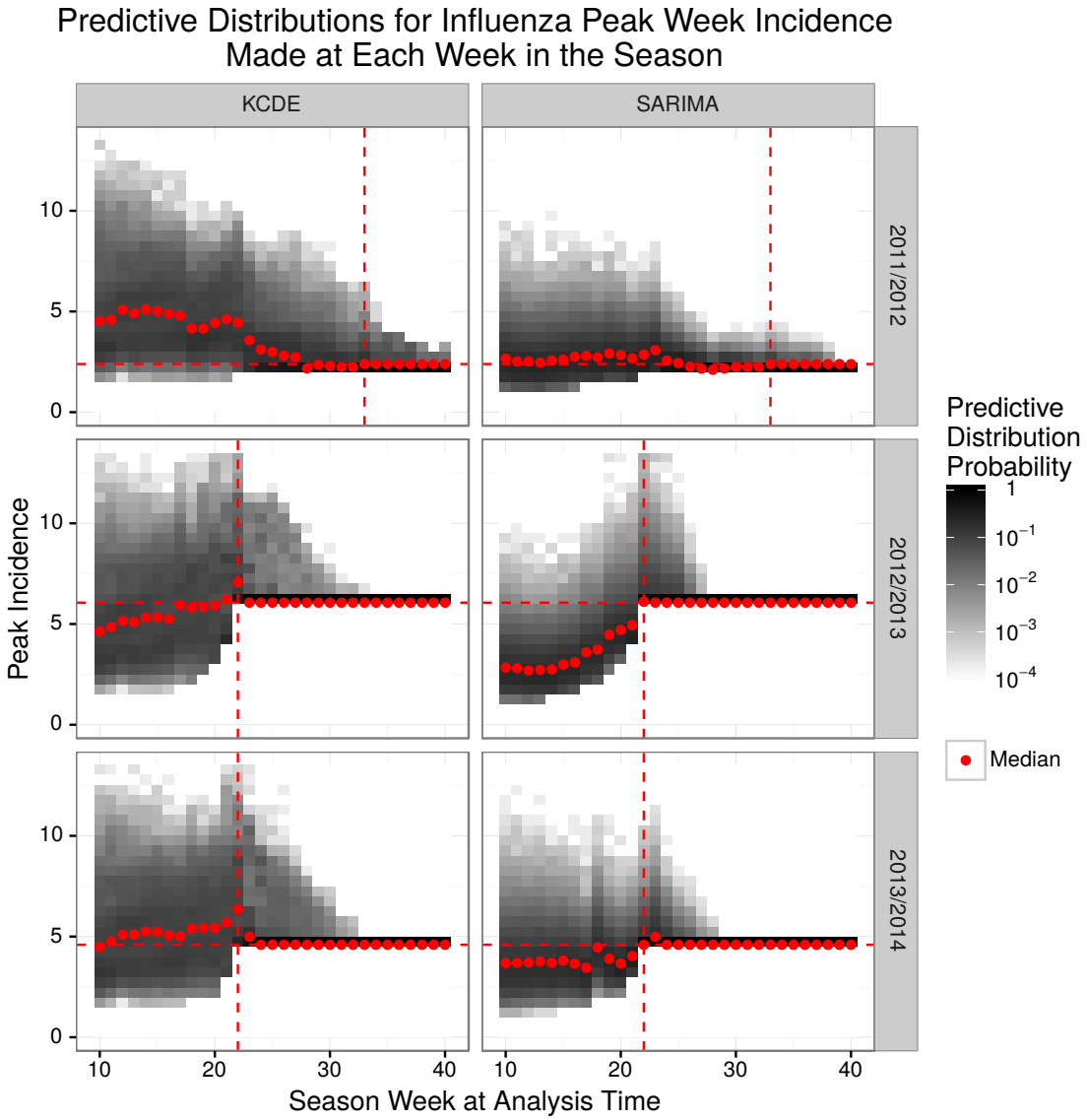
Supplemental Figure 10: Log scores for predictions of peak week timing by predictive model and analysis time. The vertical line is placed at the peak week for each season. The log score for "Equal Bin Probabilities" is obtained by assigning equal probability that the peak will occur in each week of the year.



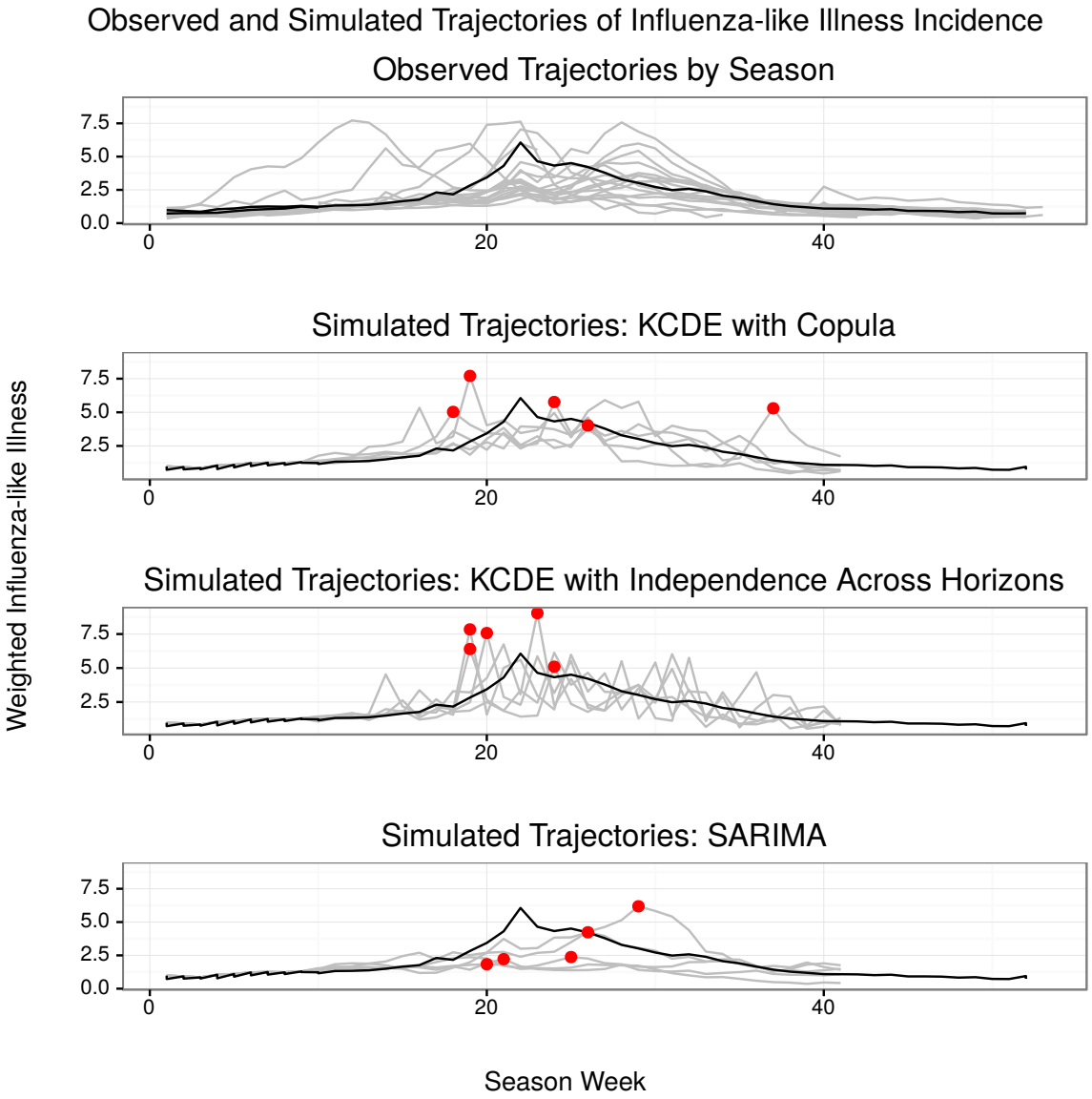
Supplemental Figure 11: A summary of performance of each method for predicting incidence in the peak week. Each boxplot summarizes all predictions made by a method in a given season in weeks before the actual peak week for that season. The vertical axis is the difference in log scores between the given method and a naive approach assigning equal probability to each incidence bin. Positive values indicate cases when the method did better than using equal bin probabilities. There are 11 incidence bins for dengue and 27 bins for influenza. The plots on the right display the trajectory of incidence over each season.



Supplemental Figure 12: A summary of performance of each method for predicting peak week timing. Each boxplot summarizes all predictions made by a method in a given season in weeks before the actual peak week for that season. The vertical axis is the difference in log scores between the given method and a naive approach assigning equal probability to each week of the year. Positive values indicate cases when the method did better than using equal bin probabilities. The plots on the right display the trajectory of incidence over each season.



Supplemental Figure 13: 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 14: 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.