Supplement to Prediction of Infectious Disease Epidemics via Feature-Weighted Density Ensembles

Evan L Ray, Nicholas G Reich

August 2017

In this supplement, we include additional figures and results.

Component Model Log Scores and Weighting Features

Supplemental Figs 4, 5, and 6 illustrate the relationship between log scores and weighting features for predictions from the three component models made during the training phase in weeks before the season onset (for predictions of onset timing) or the season peak (for predictions of peak timing or peak incidence).

Permutation Test Procedure

In the manuscript, we conducted permutation tests to compare mean performance and worst-case performance for different methods across all combinations of prediction target, region, and season. Here, we outline the procedure used for those permutation tests.

The first step in our analysis was to compute the mean log score for predictions made before the onset week (for predictions of onset timing) or the peak week (for predictions of peak timing or peak incidence). As discussed in the manuscript, this ensures that results for seasons with early onset times count with equal weight as seasons with late onset times in model comparisons. It also means that the permutation test procedure may lose power that would be available from comparing results in individual weeks; however, due to the presence of serial auto-correlation in model performance in consecutive weeks, we surmised that the loss in power would not be very dramatic. After this step, we have 165 measures of mean performance for each of our 9 models: one for each combination of the 3 prediction targets, 11 spatial units, and 5 test phase seasons. These measures of mean performance were used directly in permutation tests for overall average performance. For tests of worst-case performance, we calculated the difference between the mean performance for a given model, target, region, and season and the median performance over all models we considered for that target, region, and season.

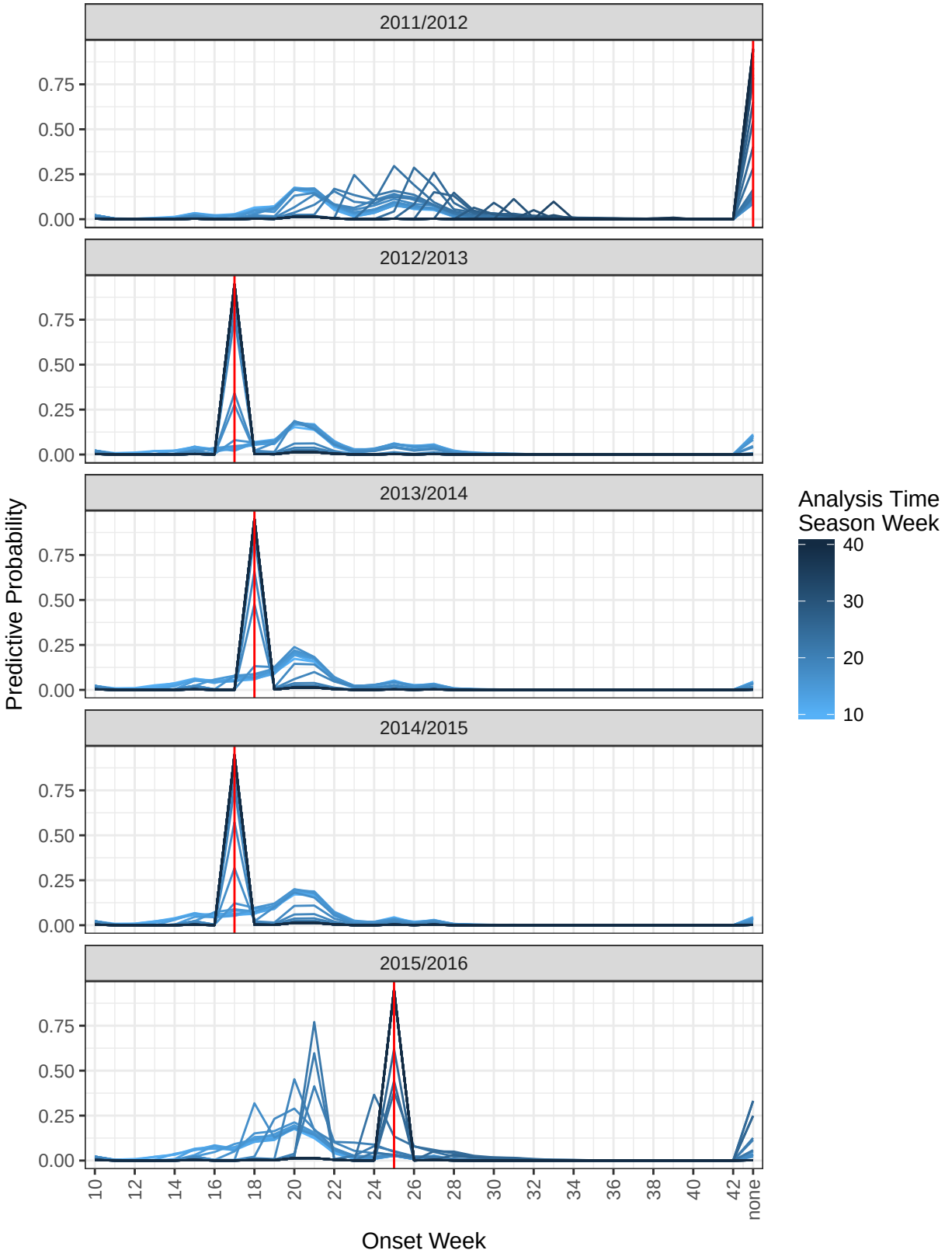
These computations give a score $\tau_{m,p,r,s}$, for each combination of model m, prediction target p, region r, and test phase season s (where which score is used depends on whether we are testing mean performance or worst-case performance). Denote the vector of these scores for a particular model by τ_m. To compare these values for a given pair of models m_1 and m_2, the observed test statistic is $|\text{mean}(\tau_{m_1}) - \text{mean}(\tau_{m_2})|$ for comparisons of mean performance and $|\text{min}(\tau_{m_1}) - \text{min}(\tau_{m_2})|$ for comparisons of worst-case performance.

The permutation test evaluated whether the scores $\tau_{m_1,p,r,s}$ and $\tau_{m_2,p,r,s}$ were drawn from the same distribution within each combination of p, r, and s. Specifically for each each combination of values p, r and s, we permuted

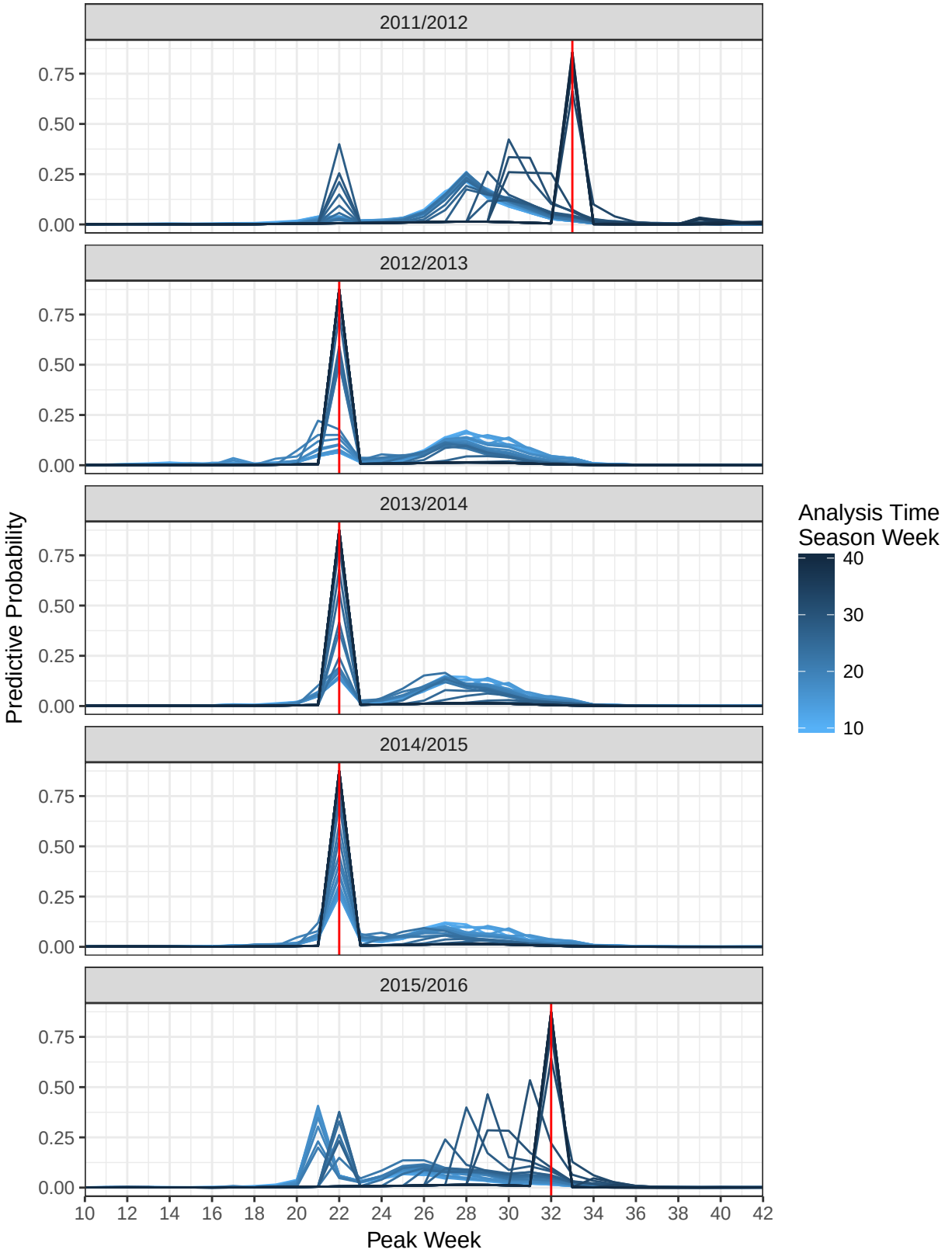
the values of $\tau_{m_1,p,r,s}$ and $\tau_{m_2,p,r,s}$. This yields a new pair of permuted vectors $\tilde{\tau}_{m_1}$ and $\tilde{\tau}_{m_2}$ and a corresponding test statistic value. Repeating the permutation process 100,000 times yielded an approximate sampling distribution for the test statistic under the null hypothesis, from which we calculated an approximate p-value.

We note that the "paired permutation" strategy presented here accounts for the fact that some prediction targets, regions, and seasons are more difficult to predict than others, so scores are not exchangeable across different combinations of those factors. However, this procedure does not capture possible correlation in the performance of a single model across different regions within a given season or different seasons within a given region.

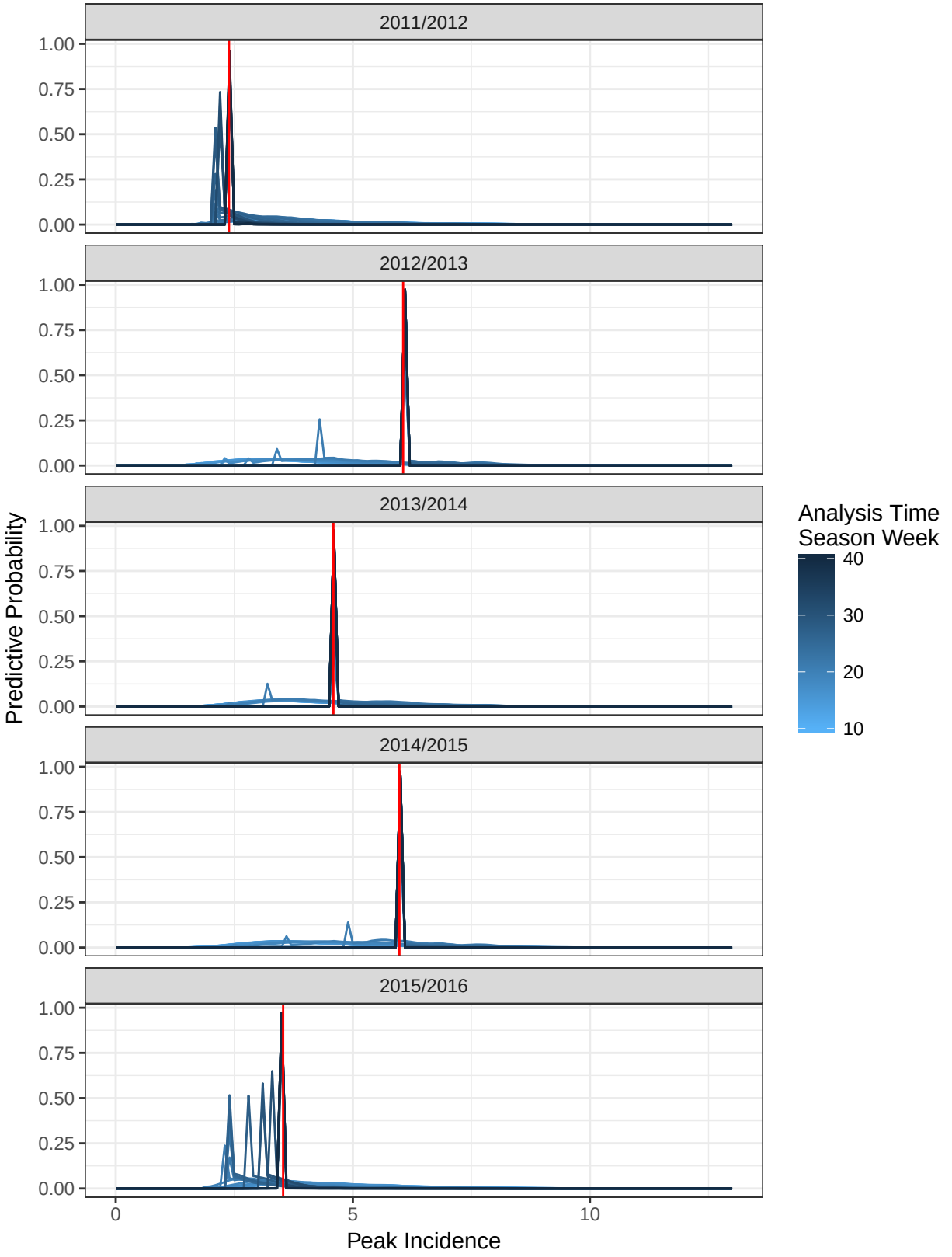
```
## Joining, by = c("model", "prediction_target")
```



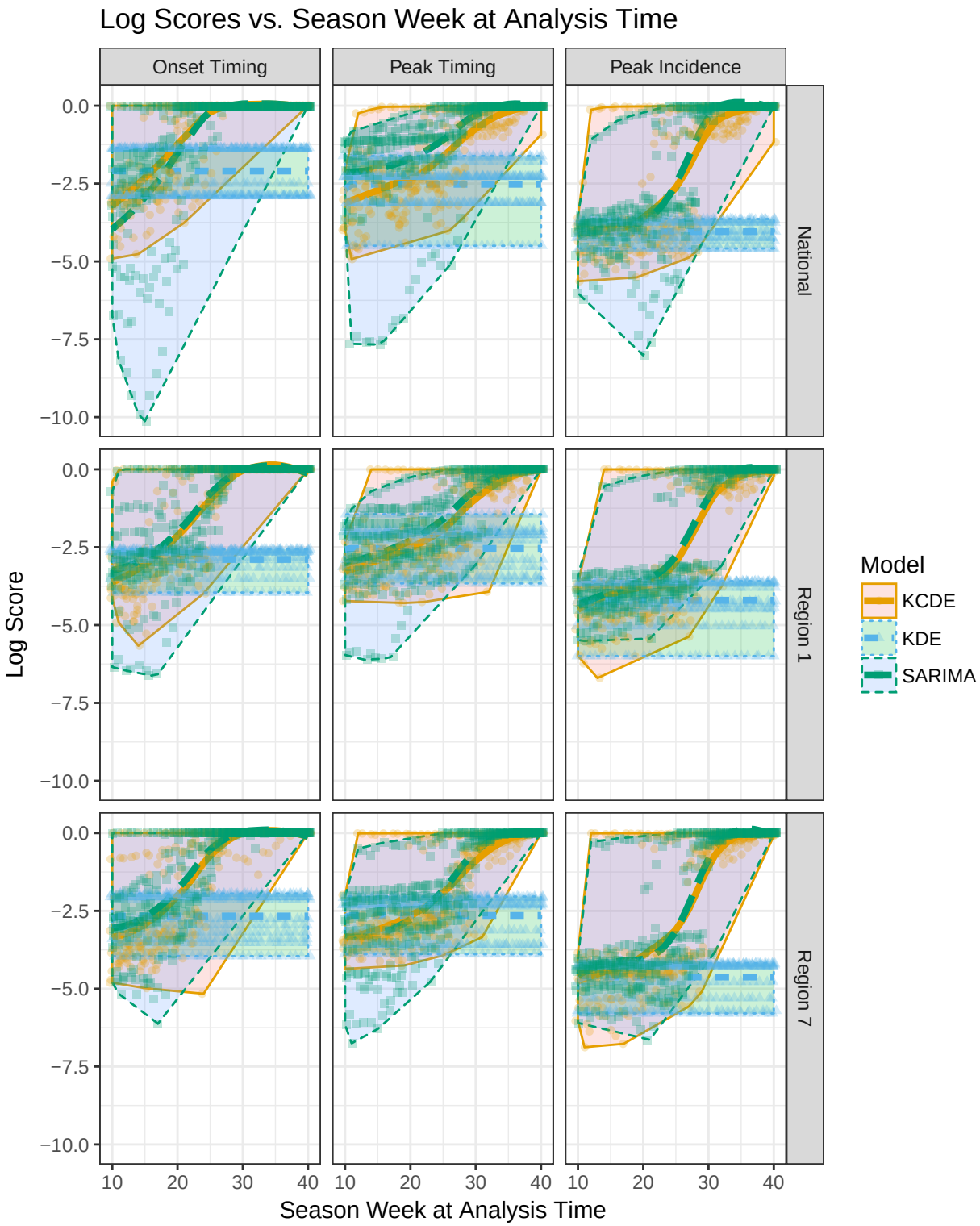
Supplemental Figure 1: Predictive distributions for onset timing at the national level from just the FW-reg-w method, facetted by season.



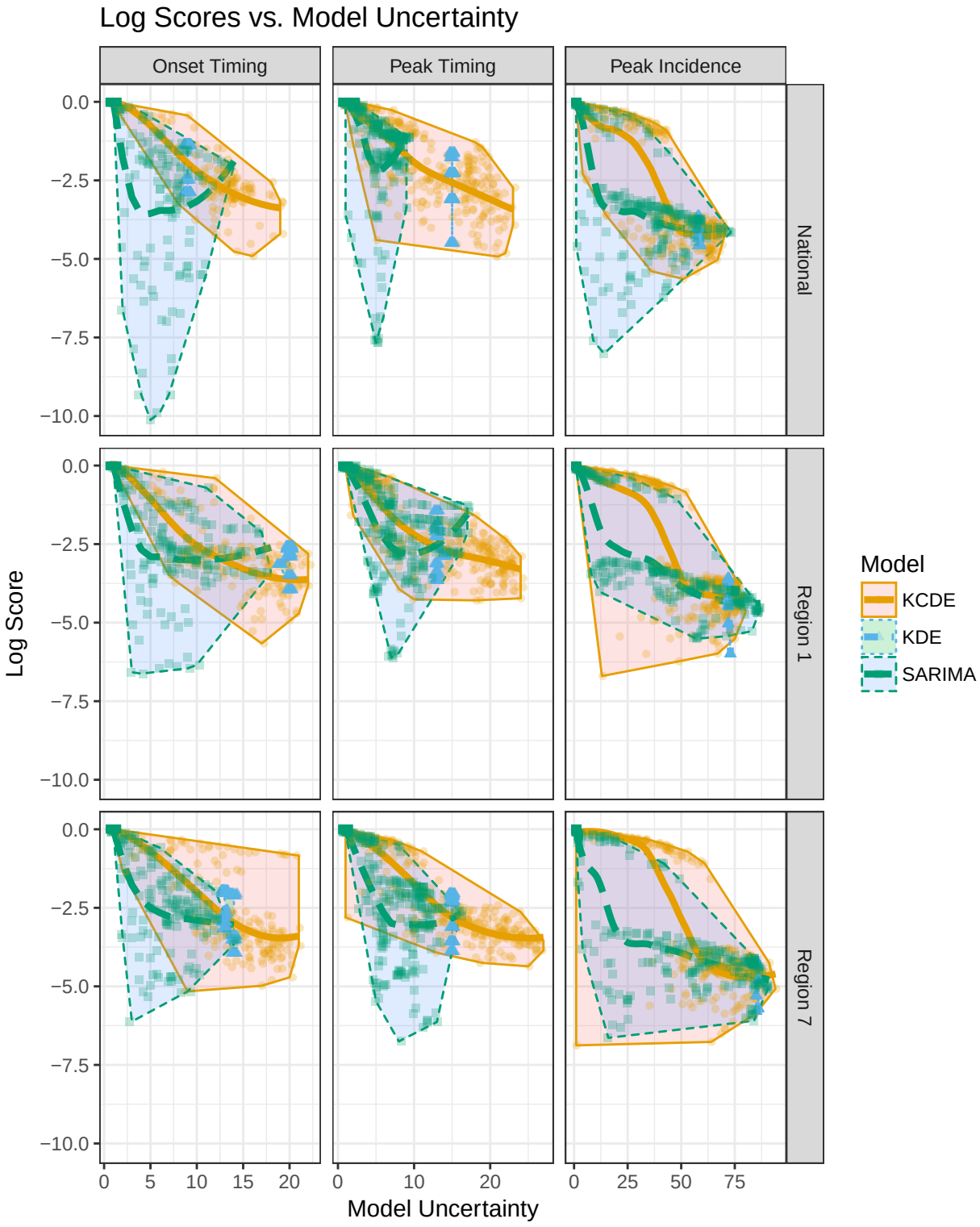
Supplemental Figure 2: Predictive distributions for peak timing at the national level from just the FW-reg-w method, facetted by season.



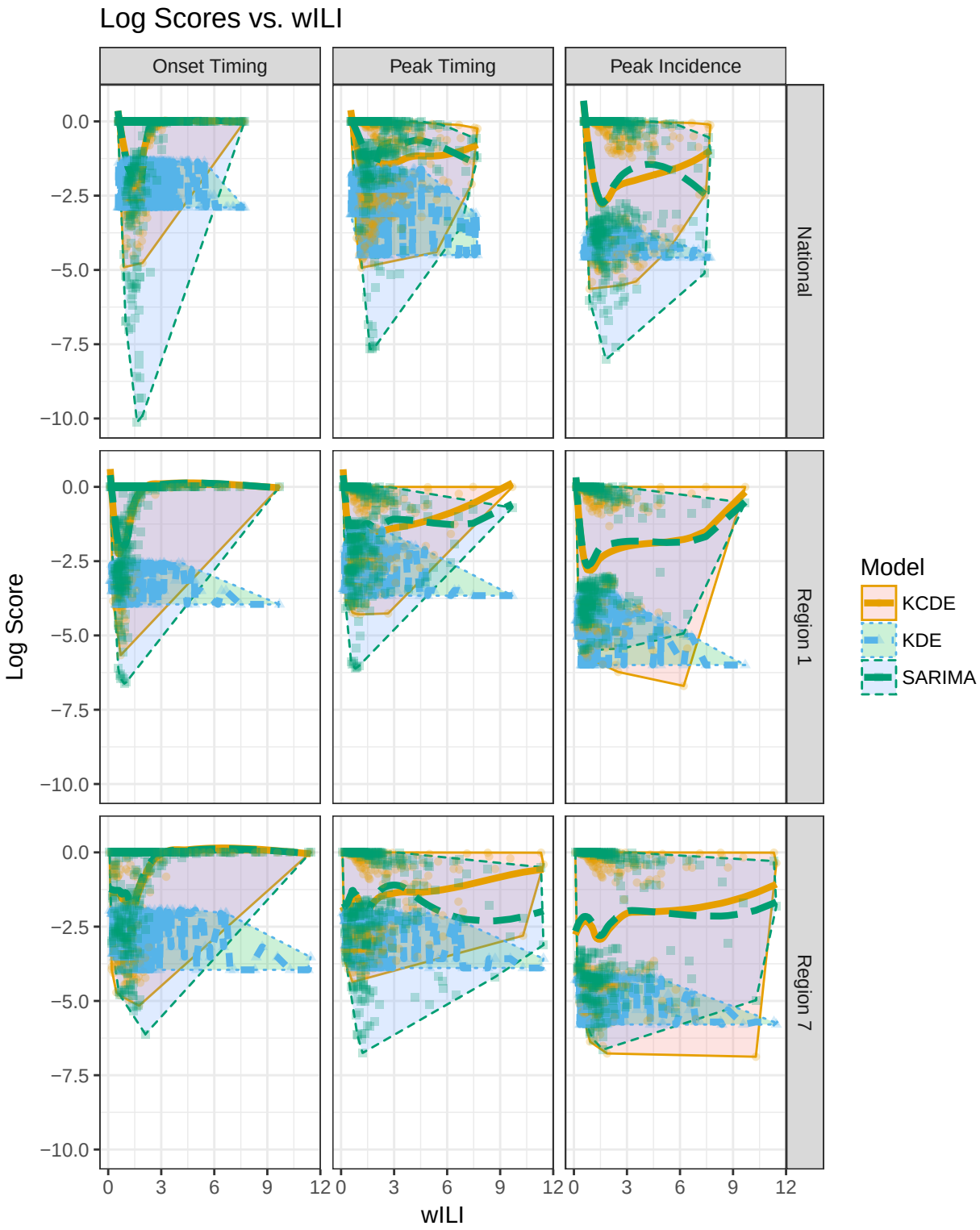
Supplemental Figure 3: Predictive distributions for peak incidence at the national level from just the FW-reg-w method, facetted by season.



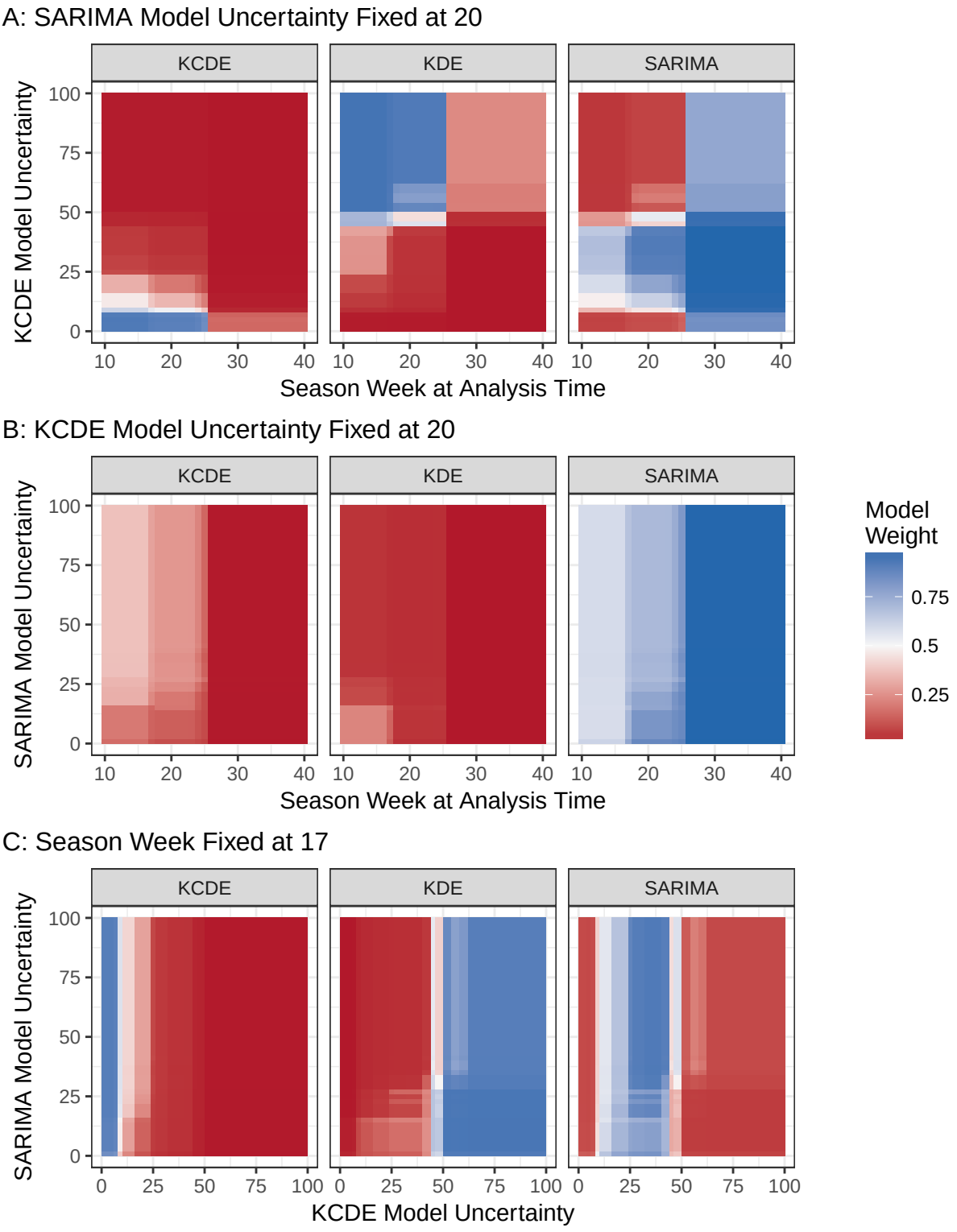
Supplemental Figure 4: Log scores achieved by each component model in each week of the season, summarizing across all seasons in both the training phase when all three component models produced predictions. The thick line is a smoothed estimate of mean log score at each week in the season; the shaded region indicates the convex hull of log scores achieved by each model; and the actual log scores achieved in each week are indicated with points.



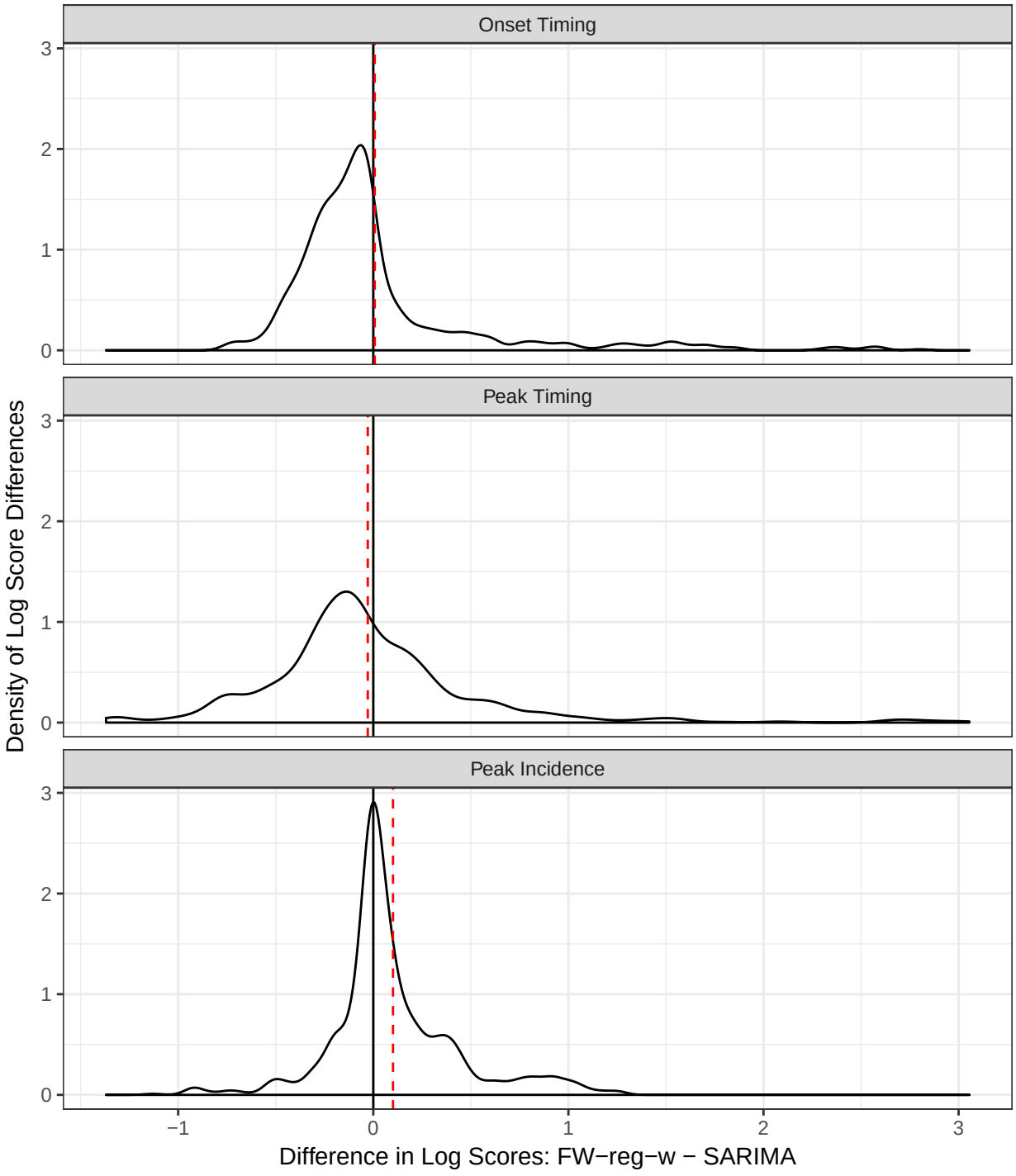
Supplemental Figure 5: Log scores achieved by each component model vs. model uncertainty as measured by the number of bins required to cover 90% of the predictive distribution. The plot summarizes results across all seasons in the training phase when all three component models produced predictions. The thick line is a smoothed estimate of mean log score at each value of model uncertainty; the shaded region indicates the convex hull of log scores achieved by each model; and the actual log scores achieved in each week are indicated with points. The KCDE and SARIMA models condition on all previously observed data within the current season, and generally have high certainly when the target event (season onset or season peak) has almost occurred or has already occurred.



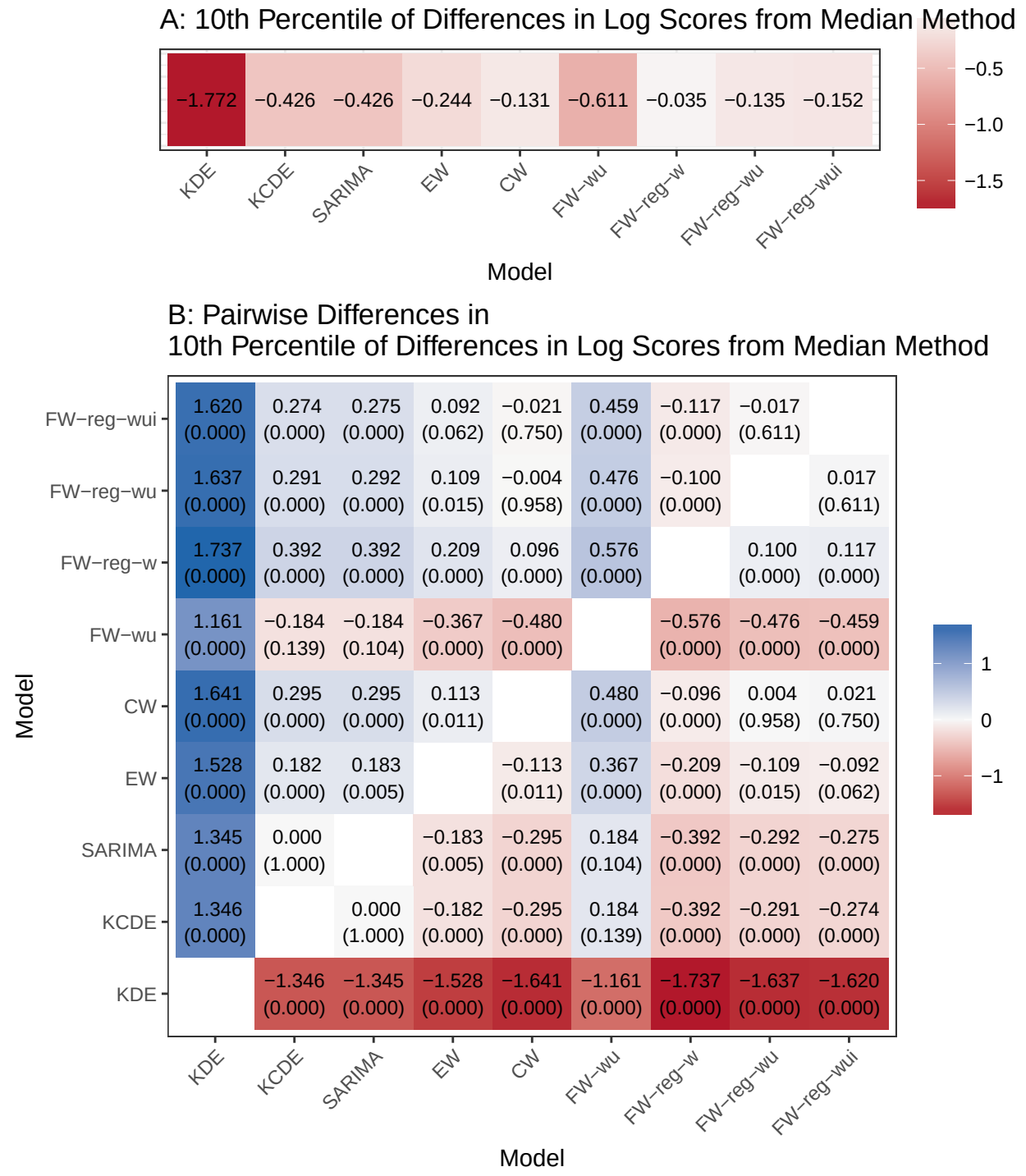
Supplemental Figure 6: Log scores achieved by each component model vs. wILI in the week of the season when predictions were made. The plot summarizes results across all seasons in the training phase when all three component models produced predictions. The thick line is a smoothed estimate of mean log score at each week in the season; the shaded region indicates the convex hull of log scores achieved by each model; and the actual log scores achieved in each week are indicated with points.



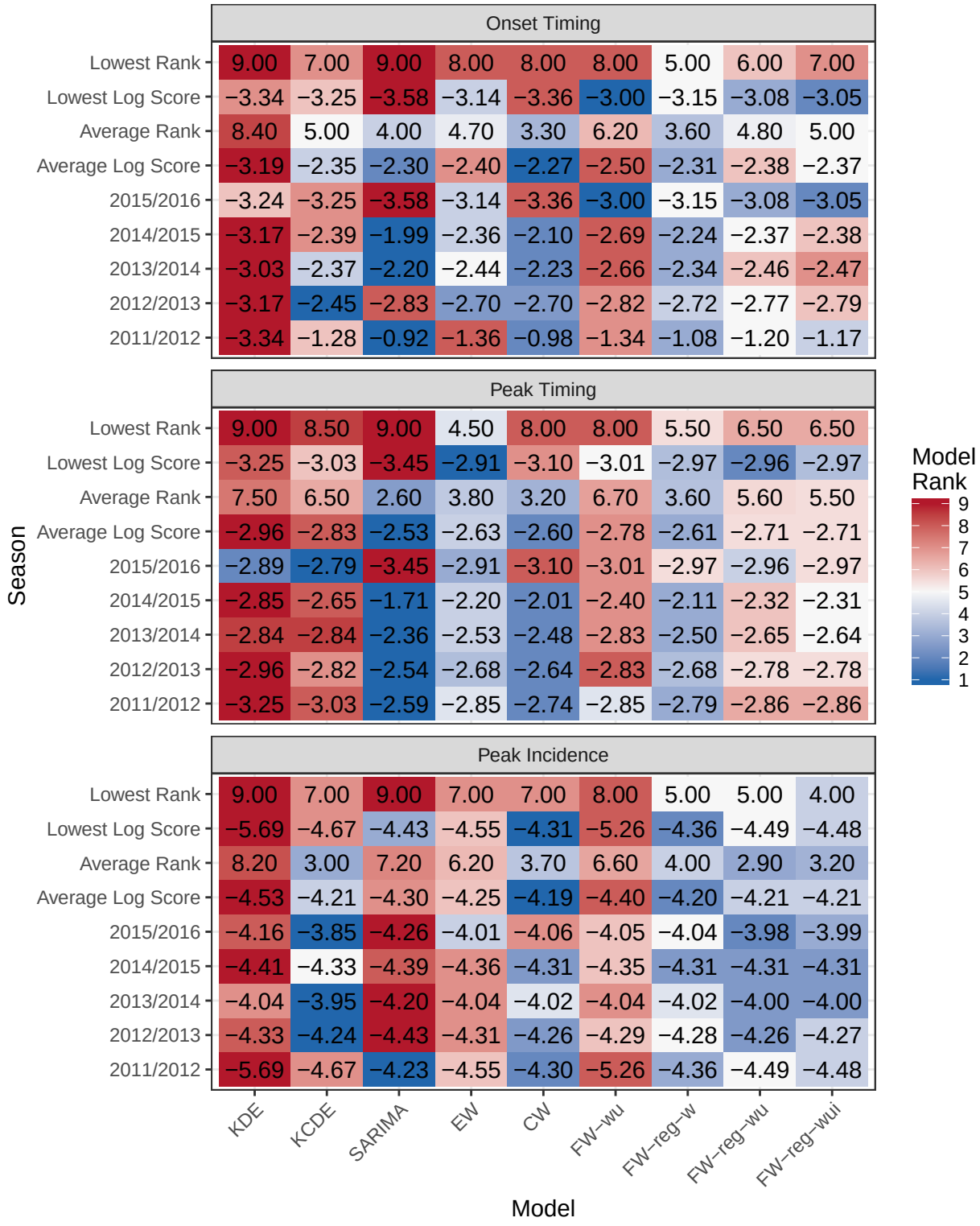
Supplemental Figure 7: Weights assigned to each component model by the FW-reg-wu model for the prediction of season peak incidence at the national level. There are three weighting functions (one for each component model) represented in each row of the figure. The value of the weight is depicted by the color. Each function depends on three features: the week of the season at the time when the predictions are made, KCDE model uncertainty, and SARIMA model uncertainty. Model uncertainty represents the minimum number of predictive distribution bins required to cover 90% probability of the predictive distribution, so the higher this number is the more uncertain the model is.



Supplemental Figure 8: Density plots representing the distribution of log score differences from predictions made by the FW-reg-w and SARIMA models for predictions of each prediction target, aggregated across all regions and test phase seasons. The horizontal axis represents the difference in log scores achieved by the FW-reg-w and SARIMA models for predictions made in a particular week; positive values indicate that FW-reg-w outperformed SARIMA for that prediction. The vertical line indicates the mean log score difference for all predictions made before the onset or season peak occurred.



Supplemental Figure 9: For each combination of 3 prediction targets, 11 regions, and 5 test phase seasons, we calculated the difference in mean log scores between each method and the method with median performance for that target, region, and season. Panel A presents the 10th percentile of these differences from the median model for each method across all combinations of target, region, and season. Larger values of this quantity indicate that the given model has better worst-case performance. Panel B displays the difference in this measure of worst-case performance for each pair of models. Positive values indicate that the model on the vertical axis had better worst-case performance than the model on the horizontal axis. A permutation test was used to obtain approximate p-values for these differences (see supplement for details). For reference, a Bonferroni correction at a familywise significance level of 0.05 for all pairwise comparisons leads to a significance cutoff of approximately 0.0014.



Supplemental Figure 10: Model performance ranked by mean log score within each of the five test seasons for predictions made before the target (season onset or peak) occurred. Averages are taken across all regions.