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.

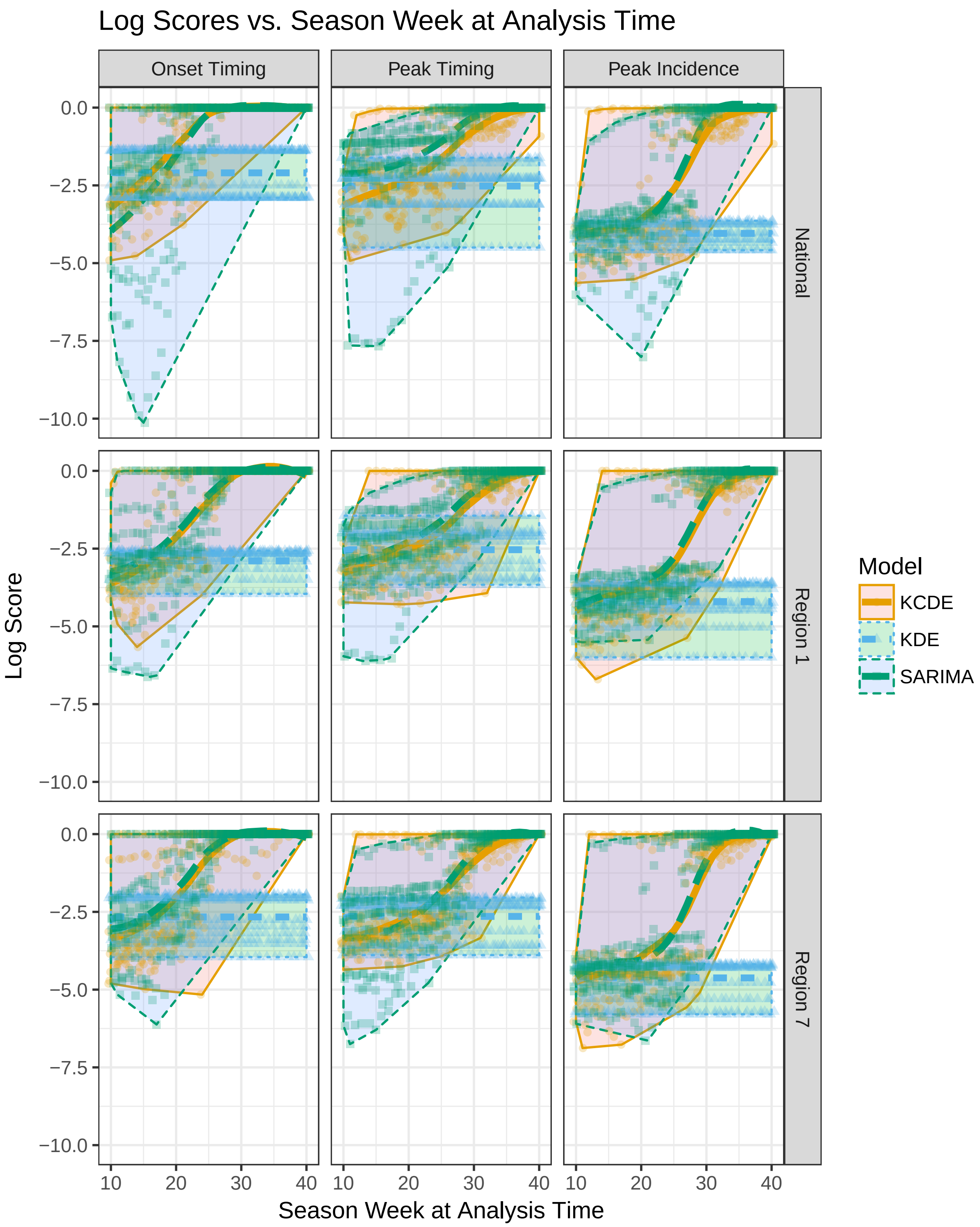
Component Model Log Scores and Weighting Features

Supplemental Figs 1, 2, and 3 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).

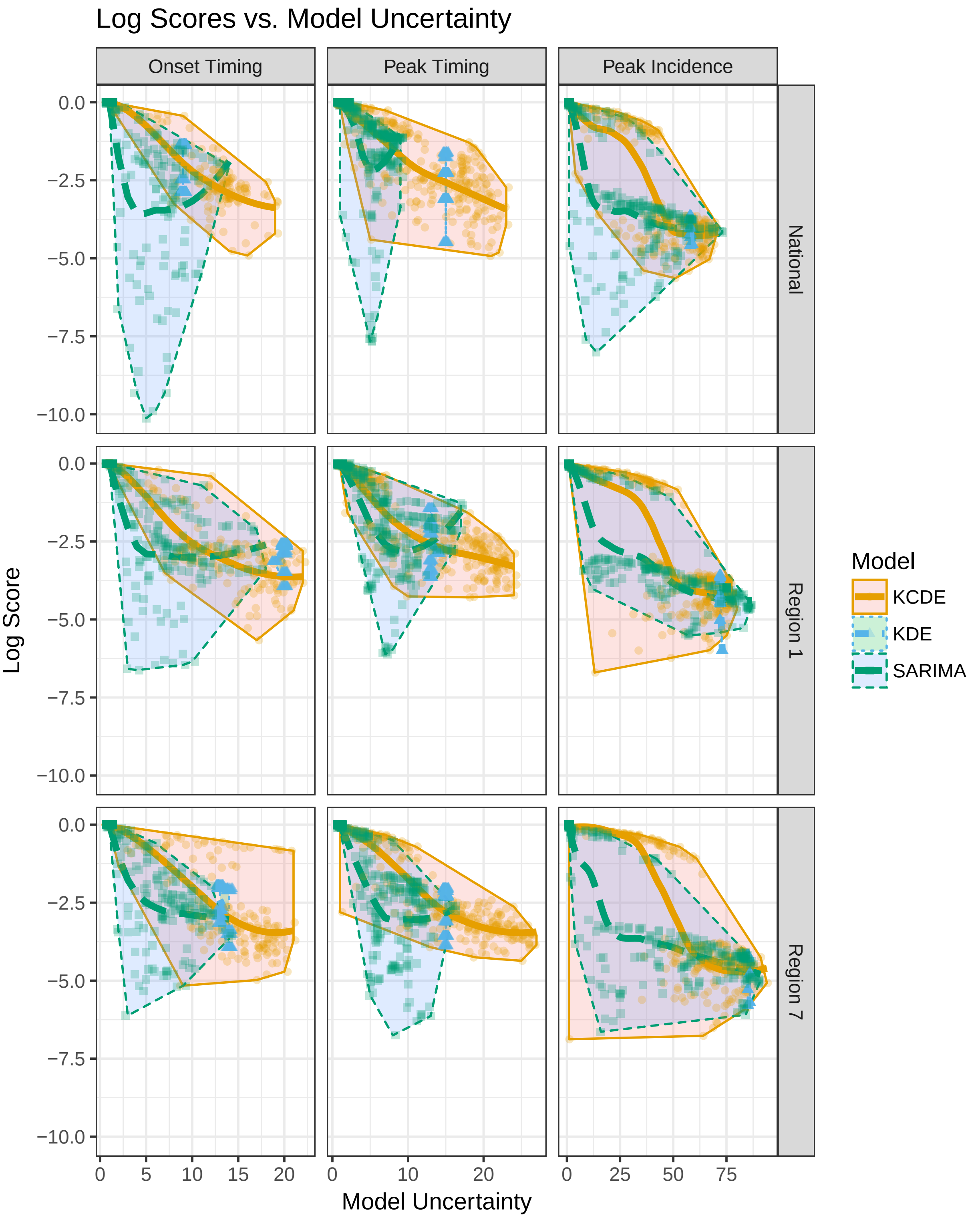
Permutation Test Procedure

In the manuscript, we conducted permutation tests to compare mean performance and worst-case performance across

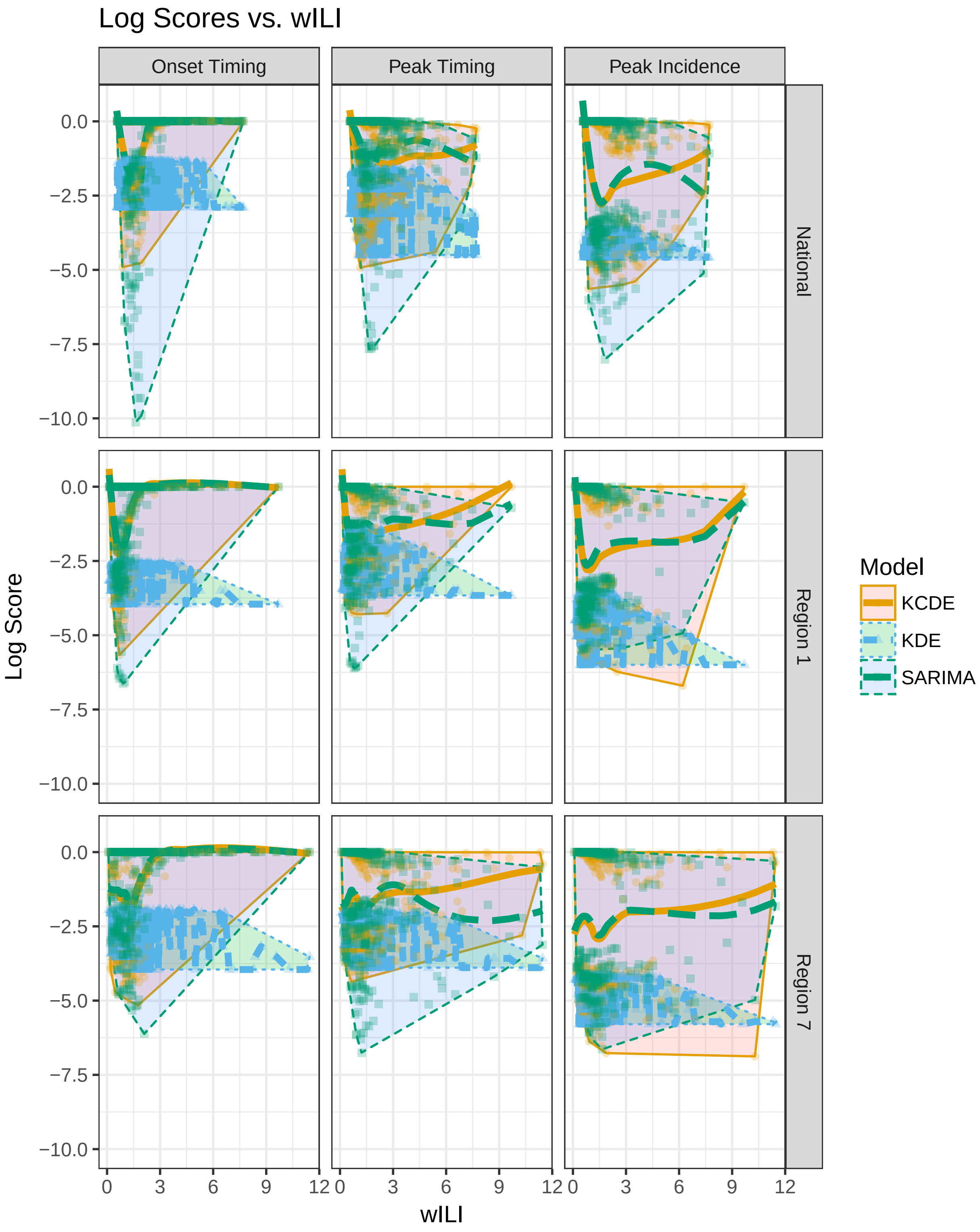
```
## Joining, by = c("model", "prediction_target")
```



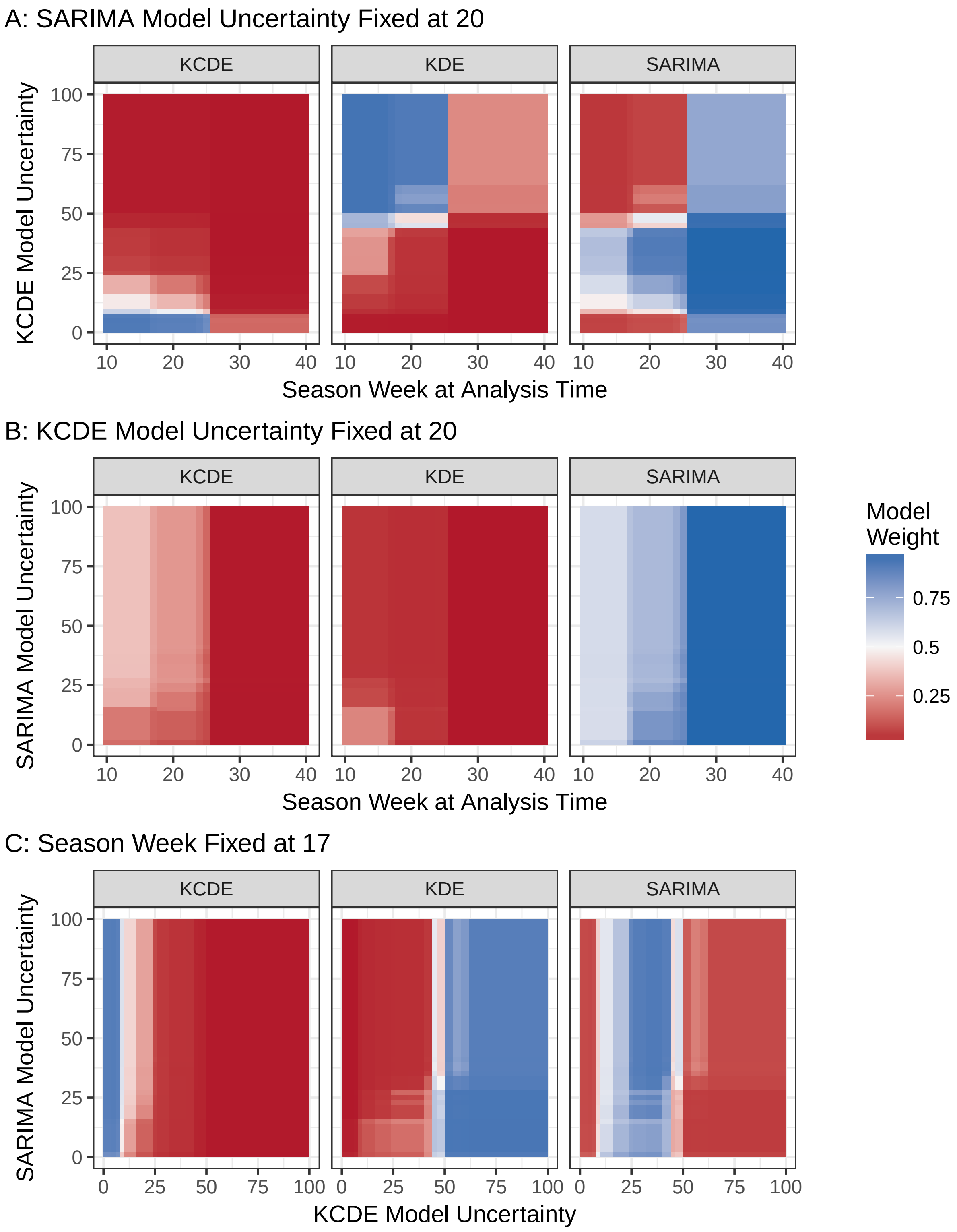
Supplemental Figure 1: Log scores achieved by each component model in each week of the season, summarizing across all seasons in both the training phase when all three component models produced predictions. The thick line is a smoothed estimate of mean log score at each week in the season; the shaded 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.



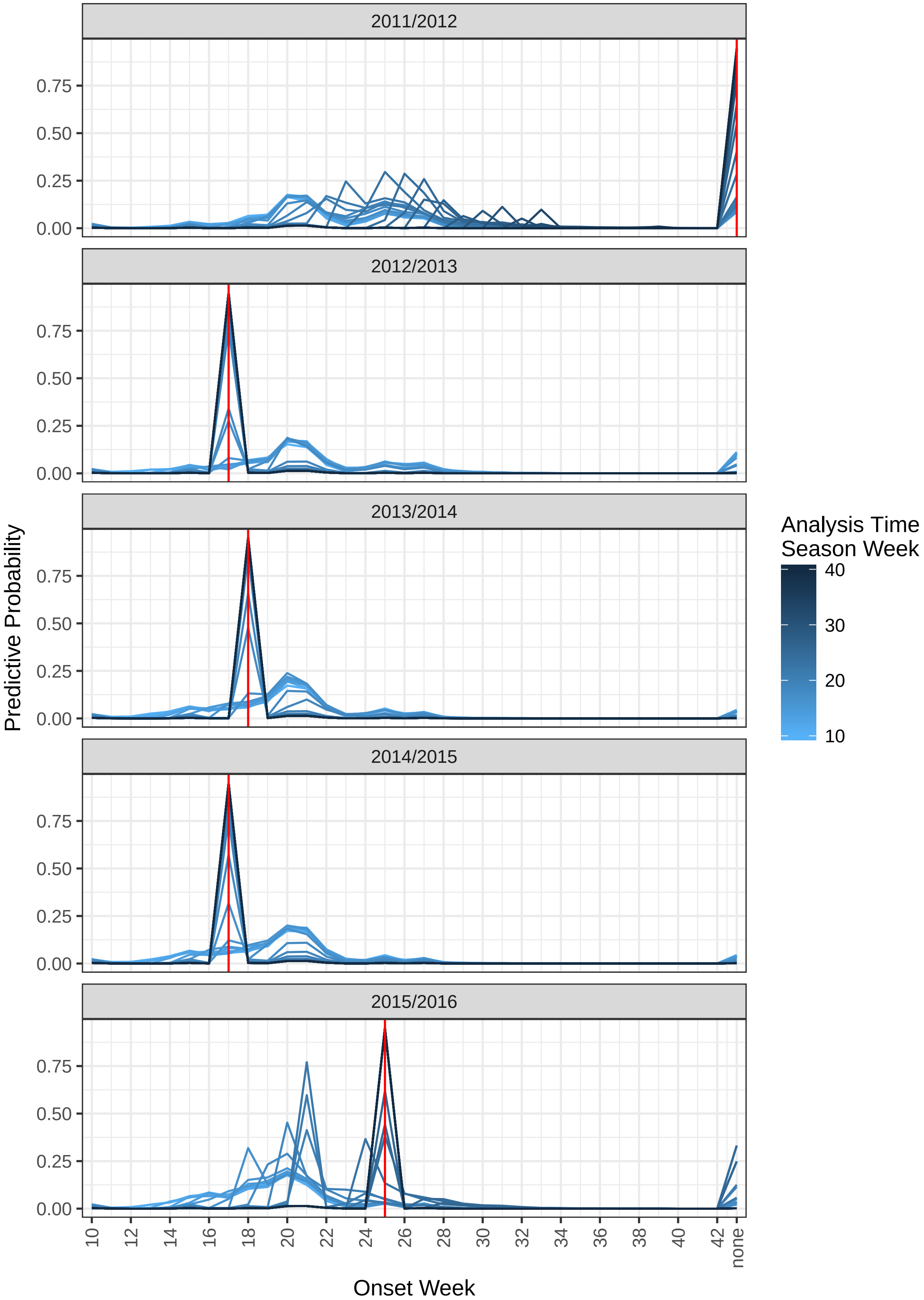
Supplemental Figure 2: Log scores achieved by each component model vs. model uncertainty as measured by the number of bins required to cover 90% of the predictive distribution. The plot summarizes results across all seasons in the training phase when all three component models produced predictions. The 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.



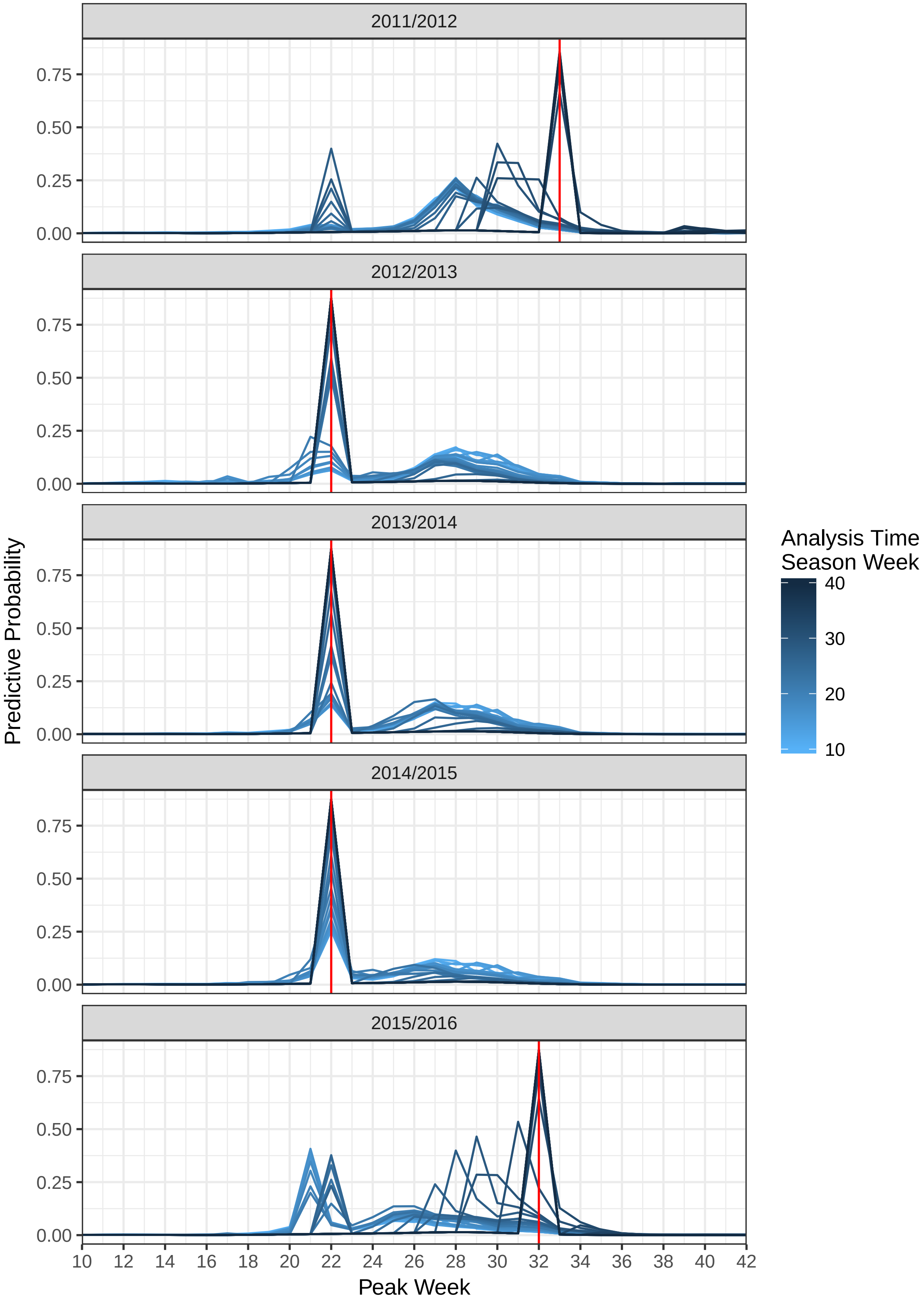
Supplemental Figure 3: Log scores achieved by each component model vs. wILI in the week of the season when predictions were made. The plot summarizes results across all seasons in the training phase when all three component models produced predictions. The thick line is a smoothed estimate of mean log 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.



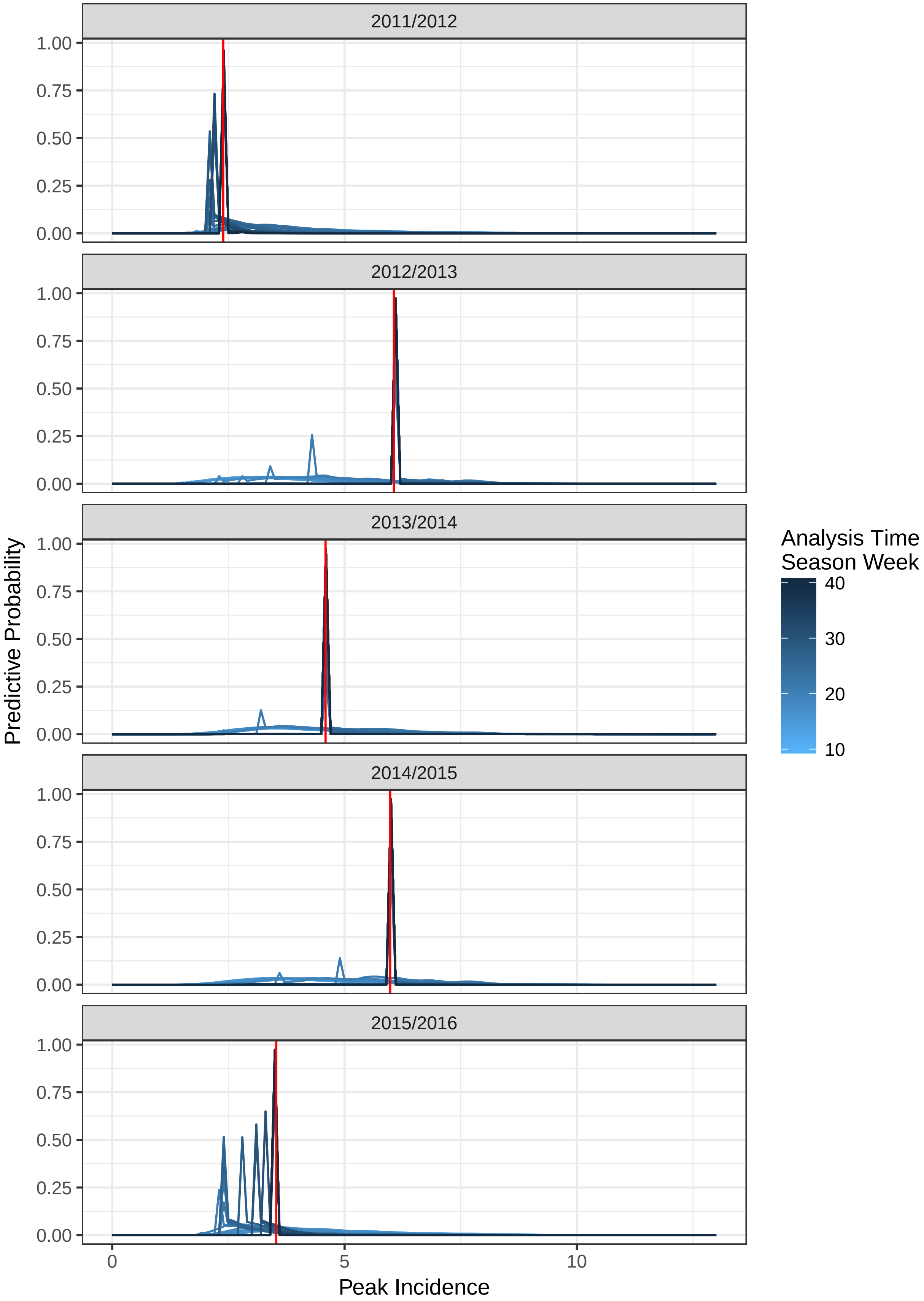
Supplemental Figure 4: Weights assigned to each component model by the FW-reg-wu model for the prediction of season peak incidence at the national level. There are three weighting functions (one for each component model) represented in each row of the figure. The value of the weight is depicted by the 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.



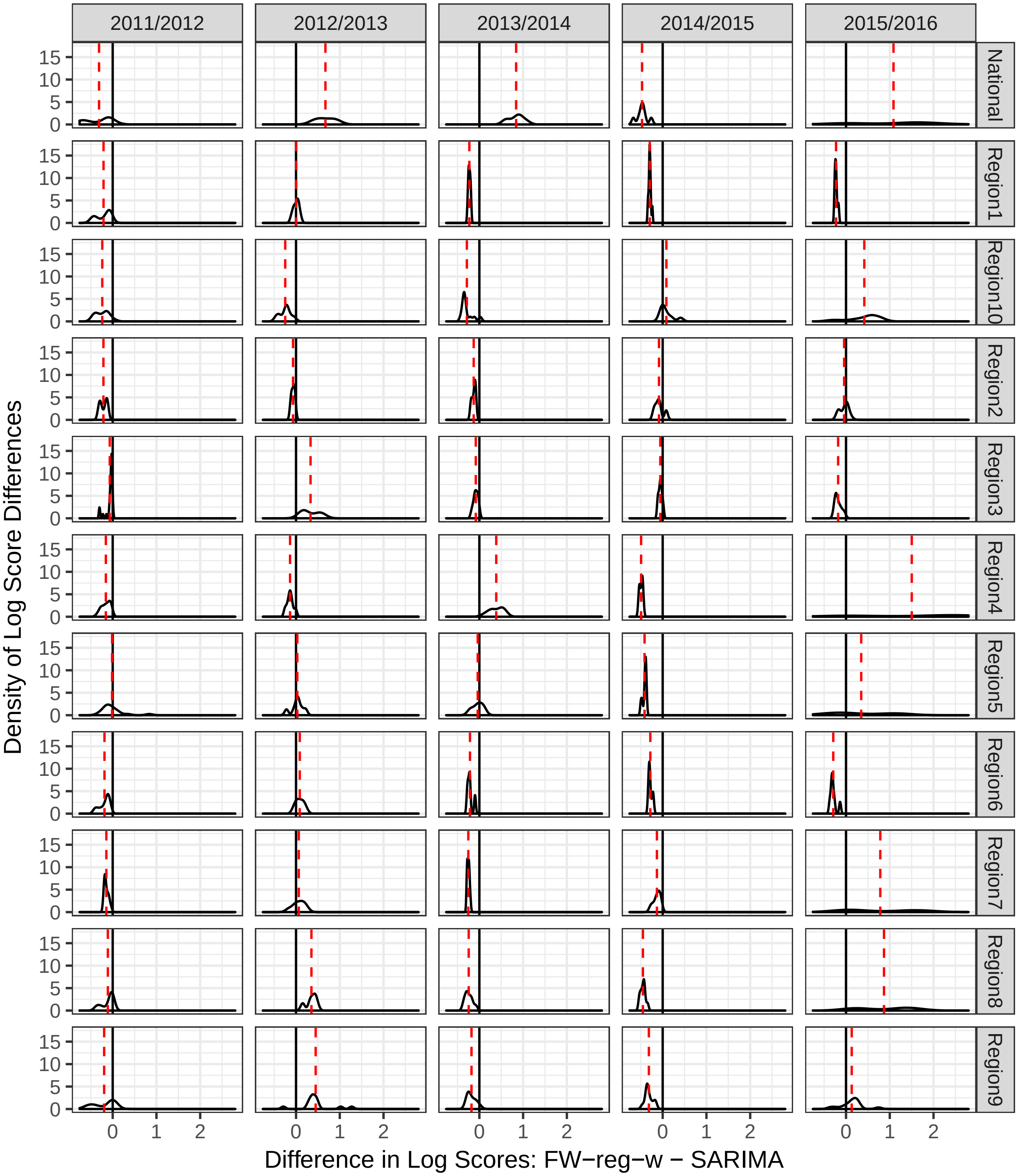
Supplemental Figure 5: Predictive distributions for onset timing at the national level from just the FW-reg-w method, facetted by season.



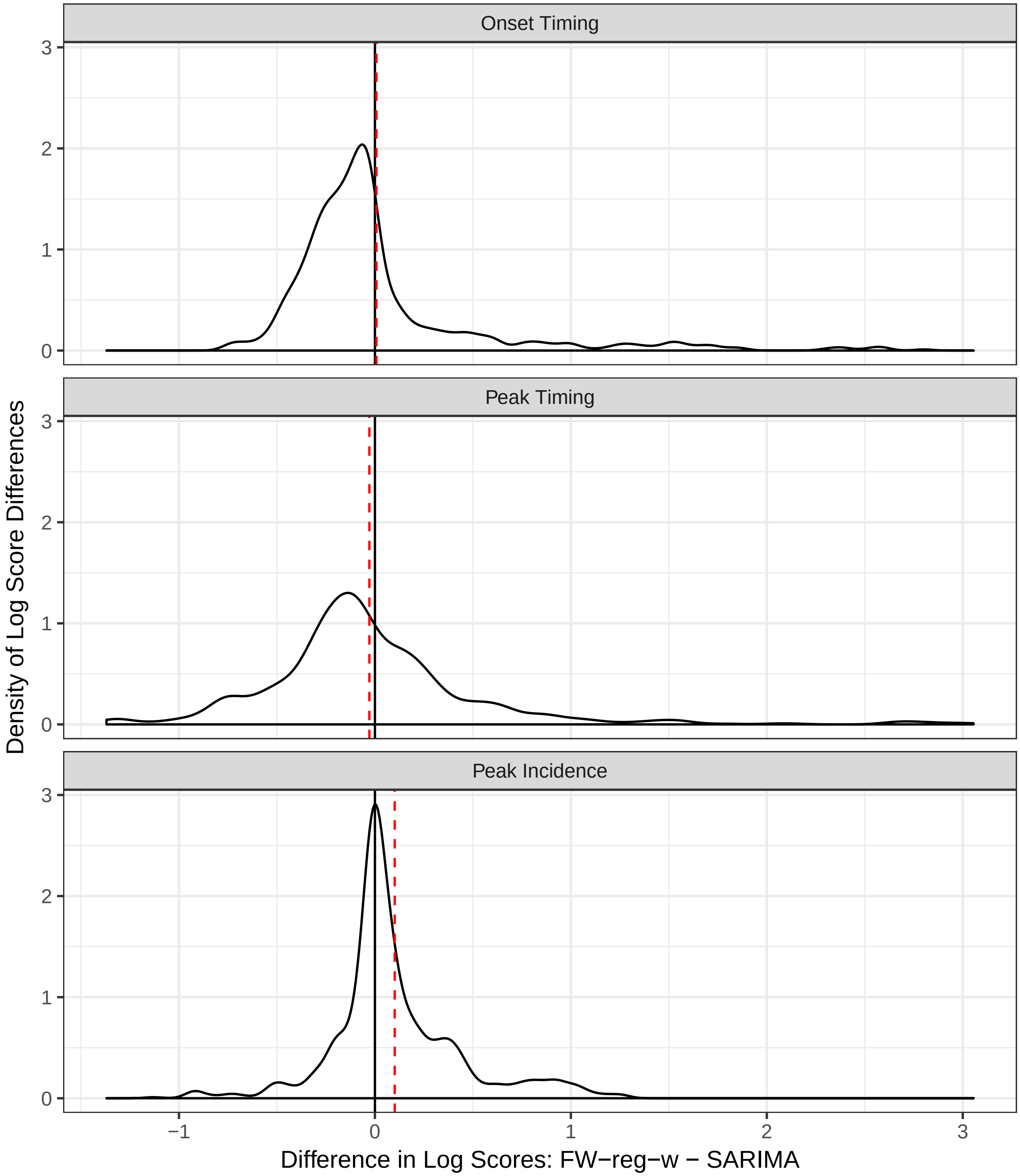
Supplemental Figure 6: Predictive distributions for peak timing at the national level from just the FW-reg-w method, facetted by season.



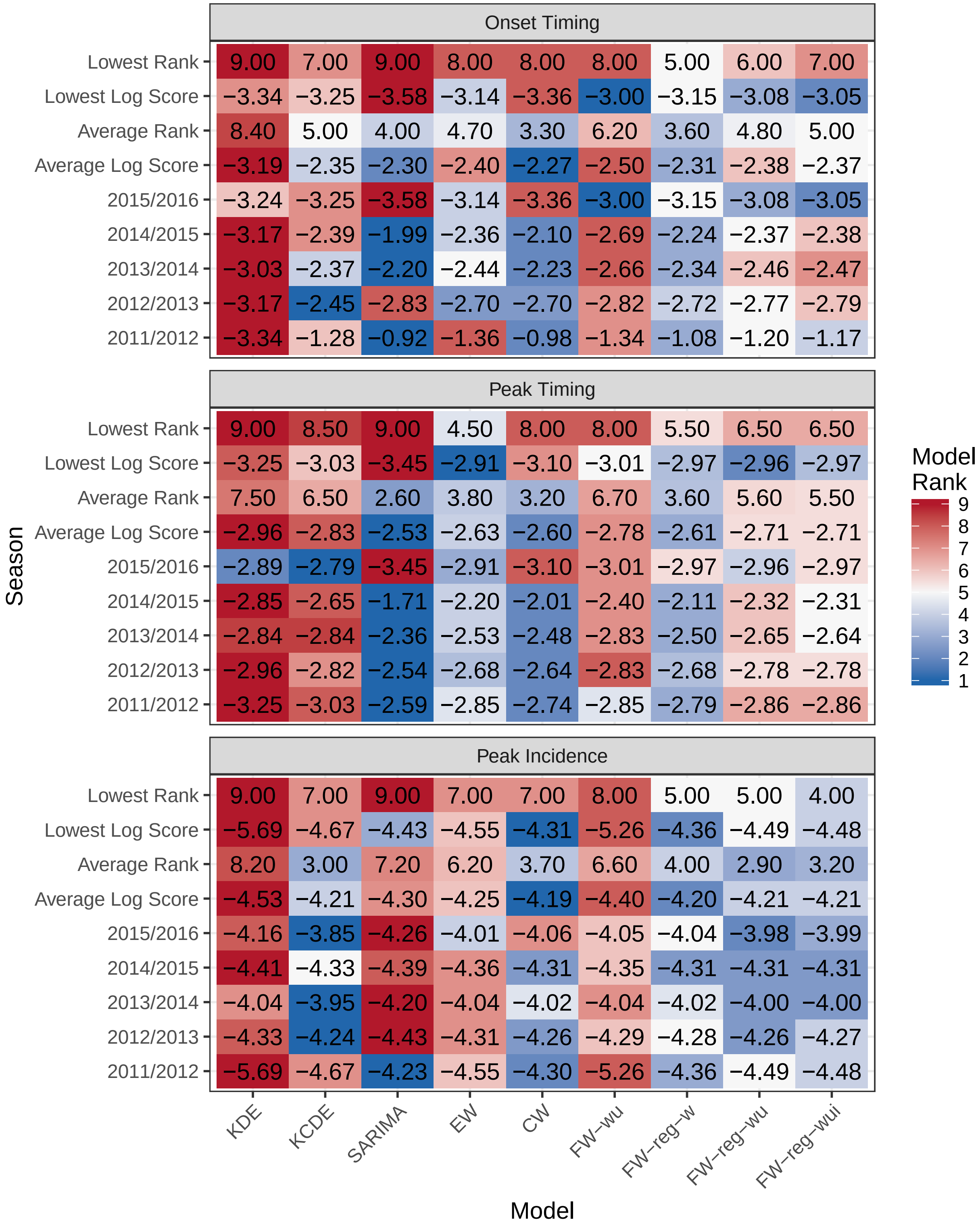
Supplemental Figure 7: Predictive distributions for peak incidence at the national level from just the FW-reg-w method, facetted by season.



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 onset timing 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 occurred in the given region and season.



Supplemental Figure 9: 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.



Onset Timing

Season	KDE	KCDE	SARIMA	EW	CW	FW−wu	FW−reg−w	FW−reg−wu	FW−reg−wui
Lowest Rank	9.00	7.00	9.00	8.00	8.00	8.00	5.00	6.00	7.00
Lowest Log Score	−3.34	−3.25	−3.58	−3.14	−3.36	−3.00	−3.15	−3.08	−3.05
Average Rank	8.40	5.00	4.00	4.70	3.30	6.20	3.60	4.80	5.00
Average Log Score	−3.19	−2.35	−2.30	−2.40	−2.27	−2.50	−2.31	−2.38	−2.37
2015/2016	−3.24	−3.25	−3.58	−3.14	−3.36	−3.00	−3.15	−3.08	−3.05
2014/2015	−3.17	−2.39	−1.99	−2.36	−2.10	−2.69	−2.24	−2.37	−2.38
2013/2014	−3.03	−2.37	−2.20	−2.44	−2.23	−2.66	−2.34	−2.46	−2.47
2012/2013	−3.17	−2.45	−2.83	−2.70	−2.70	−2.82	−2.72	−2.77	−2.79
2011/2012	−3.34	−1.28	−0.92	−1.36	−0.98	−1.34	−1.08	−1.20	−1.17

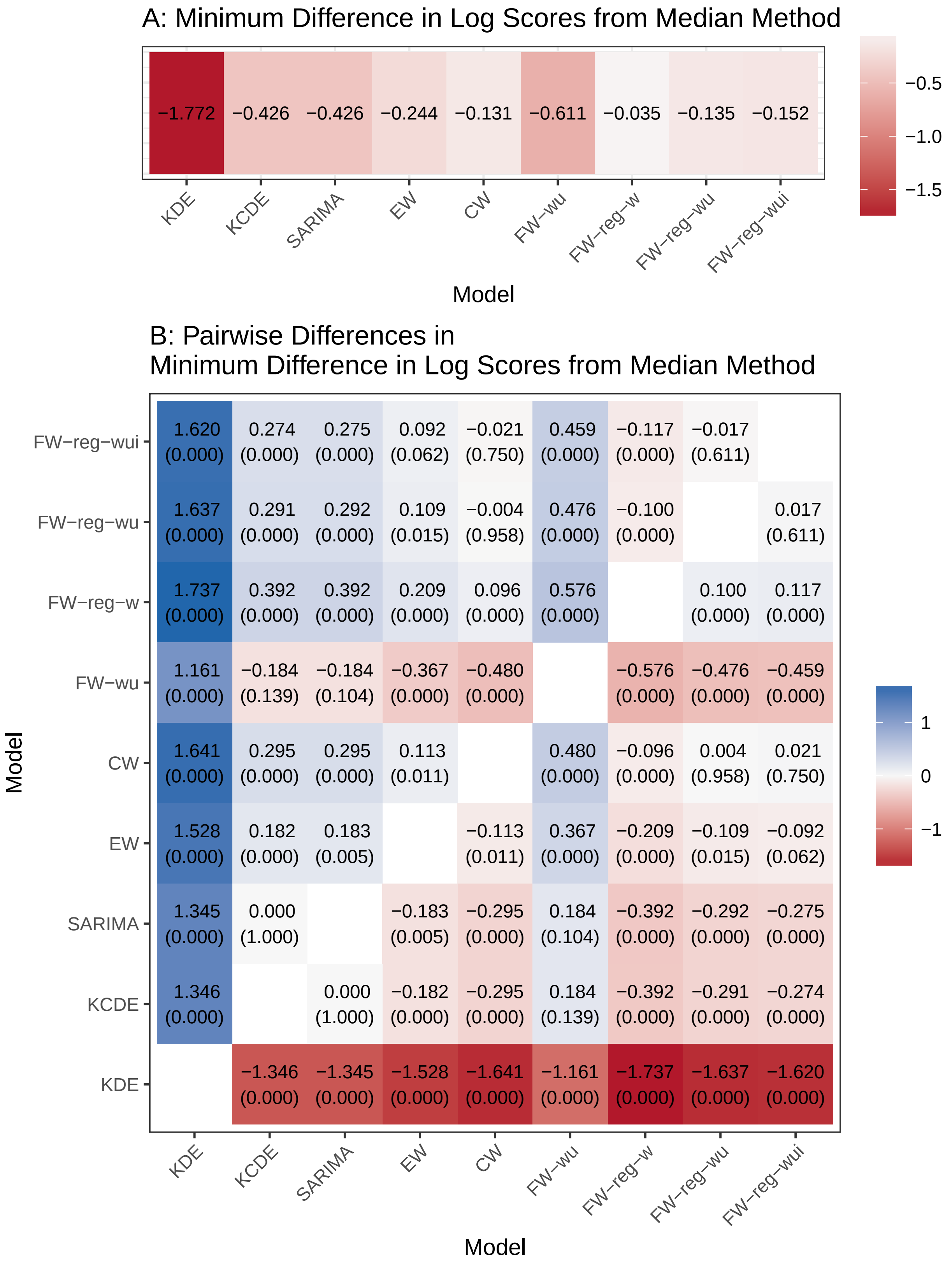
Peak Timing

Season	KDE	KCDE	SARIMA	EW	CW	FW−wu	FW−reg−w	FW−reg−wu	FW−reg−wui
Lowest Rank	9.00	8.50	9.00	4.50	8.00	8.00	5.50	6.50	6.50
Lowest Log Score	−3.25	−3.03	−3.45	−2.91	−3.10	−3.01	−2.97	−2.96	−2.97
Average Rank	7.50	6.50	2.60	3.80	3.20	6.70	3.60	5.60	5.50
Average Log Score	−2.96	−2.83	−2.53	−2.63	−2.60	−2.78	−2.61	−2.71	−2.71
2015/2016	−2.89	−2.79	−3.45	−2.91	−3.10	−3.01	−2.97	−2.96	−2.97
2014/2015	−2.85	−2.65	−1.71	−2.20	−2.01	−2.40	−2.11	−2.32	−2.31
2013/2014	−2.84	−2.84	−2.36	−2.53	−2.48	−2.83	−2.50	−2.65	−2.64
2012/2013	−2.96	−2.82	−2.54	−2.68	−2.64	−2.83	−2.68	−2.78	−2.78
2011/2012	−3.25	−3.03	−2.59	−2.85	−2.74	−2.85	−2.79	−2.86	−2.86

Peak Incidence

Season	KDE	KCDE	SARIMA	EW	CW	FW−wu	FW−reg−w	FW−reg−wu	FW−reg−wui
Lowest Rank	9.00	7.00	9.00	7.00	7.00	8.00	5.00	5.00	4.00
Lowest Log Score	−5.69	−4.67	−4.43	−4.55	−4.31	−5.26	−4.36	−4.49	−4.48
Average Rank	8.20	3.00	7.20	6.20	3.70	6.60	4.00	2.90	3.20
Average Log Score	−4.53	−4.21	−4.30	−4.25	−4.19	−4.40	−4.20	−4.21	−4.21
2015/2016	−4.16	−3.85	−4.26	−4.01	−4.06	−4.05	−4.04	−3.98	−3.99
2014/2015	−4.41	−4.33	−4.39	−4.36	−4.31	−4.35	−4.31	−4.31	−4.31
2013/2014	−4.04	−3.95	−4.20	−4.04	−4.02	−4.04	−4.02	−4.00	−4.00
2012/2013	−4.33	−4.24	−4.43	−4.31	−4.26	−4.29	−4.28	−4.26	−4.27
2011/2012	−5.69	−4.67	−4.23	−4.55	−4.30	−5.26	−4.36	−4.49	−4.48

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.


A: Minimum Difference in Log Scores from Median Method

B: Pairwise Differences in Minimum Difference in Log Scores from Median Method

Supplemental Figure 11: 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 minimum difference 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.