Supporting Information for "Comparison of Combination Methods to Create Calibrated Ensemble Forecasts for Seasonal Influenza in the U.S."

Nutcha Wattanachit[1] | Evan L. Ray[1] | Thomas C. McAndrew[2] | Nicholas G. Reich[1]

[1]School of Public Health and Health Sciences, University of Massachusetts Amherst, Massachusetts, USA
[2]College of Health, Lehigh University, Pennsylvania, USA

Correspondence
*Nutcha Wattanachit, University of Massachusetts Amherst. Email: nwattanachit@umass.edu

1 | INDIVIDUAL FORECASTING MODELS

Prospective forecasts of 1-4 week ahead wILI for the United States and the 10 Health and Human Services (HHS) regions from 27 individual models in Table S1 were used as component forecasts in the combination methods to create ensemble forecasts. The forecast submissions are available in the FluSight Network repository [1].

2 | MEAN OUT-OF-SAMPLE LOG SCORES AT DIFFERENT AGGREGATION LEVELS

In order to create a summary of forecast accuracy for all combination methods, observation-level out-of-sample log scores were aggregated for each target across all three test seasons and for each season across all four targets in the article. At fine-grained aggregation levels, mean out-of-sample log scores varied across targets in the three test seasons (Figure S1). For 8 out of 12 target-season pairs, the BMC_2 and BLP were the two top performing methods. Specifically, both the BLP and BMC_2 had the best mean out-of-sample log scores for the 2 week ahead horizon in the 2017/2018 and 2018/2019 seasons. The BLP outperformed other methods for 1 and 3 week ahead horizons in the 2016/2017 season and for 1-2 week ahead horizons in the 2016/2017 season. The BMC_2 outperformed the other four methods for the 3 week ahead horizon in the 2017/2018 season and for 1,3, and 4 week ahead horizons in the 2018/2019 season. The LP had the best score for the 3-4 week ahead horizons in the 2016/2017 season and the 4 week ahead horizon in the 2017/2018 season. The EW-LP was the worst performing method for most target and season pairs, while the EW-BLP and EW-BMC_2 yielded the worst log scores for the 3 and 4 week ahead targets in the 2017/2018 season.

For most locations, the EW-LP was also the worst performing method across all targets and seasons (see Figure S2, Figure S3, and Figure S4). However, the EW-LP's mean out-of-sample log scores for 3 and 4 week ahead forecasts ranked among the top three positions in 4 locations. Either the BLP or BMC_2 or both were the two top performing methods across all targets for about half of the locations in the 2016/2017 and 2017/2018 seasons and for most locations in 2018/2019 season. Nonetheless, the 3 and 4 week ahead forecasts from the BMC_2 method had worse mean out-of-sample log scores in multiple locations compared to those of other methods. The LP was one of the top two performing method for the 3 and 4 week ahead targets for most locations in the 2017/2018 season. These variations in out-of-sample log scores were also shown in Figure 4 in the article.

The ranks of mean out-of-sample log scores of 1 week ahead forecasts were relatively more consistent compared to the ranks of mean out-of-sample log scores of forecasts for farther forecast horizons. Additionally, Figure S2, Figure S3, and Figure S4

Table S1 List of individual forecasting models in the FluSight Network repository

Team	Model Abbreviation	Model Description
CU	EAKFC_SEIRS	Ensemble Adjustment Kalman Filter SEIRS [2]
	EAKFC_SIRS	Ensemble Adjustment Kalman Filter SIRS [2]
	EKF_SEIRS	Ensemble Kalman Filter SEIRS [3]
	EKF_SIRS	Ensemble Kalman Filter SIRS [3]
	RHF_SEIRS	Rank Histogram Filter SEIRS [3]
	RHF_SIRS	Rank Histogram Filter SIRS [3]
	BMA	Bayesian Model Averaging [4]
Delphi	BasisRegression	Basis Regression (`epiforecast` defaults) [5]
	DeltaDensity1	Delta Density (`epiforecast` defaults) [6]
	DeltaDensity2	Markovian Delta Density (`epiforecast` defaults) [6]
	EmpiricalFuture	Empirical Futures (`epiforecast` defaults) [5]
	EmpiricalTraj	Empirical Trajectories (`epiforecast` defaults) [5]
	Uniform	Uniform Distribution [1]
LANL	DBMplus	Dynamic Bayesian SIR Model with discrepancy [7]
ReichLab	KCDE	Kernel Conditional Density Estimation [8]
	KCDE backfill	Kernel Conditional Density Estimation with backfill [8]
	KDE	Kernel Density Estimation and penalized splines [9]
	SARIMA1	SARIMA model without seasonal differencing [9]
	SARIMA2	SARIMA model with seasonal differencing [9]
FluOutlook	Mech	Mechanistic GLEAM Ensemble [1]
	MechAug	Augmented Mechanistic GLEAM Ensemble [1]
Protea	Cheetah	Ensemble of dynamic harmonic model and historical averages [1]
	Kudu	Subtype weighted historical average model [1]
	Springbok	Dynamic Harmonic Model with ARIMA errors [1]
FluX	ARLR	Auto Regressive model with Likelihood Ratio based Model Selection [1]
	LSTM	Recurrent Neural Network (Long Short-Term Memory) [1]
UA	EpiCos	Epidemic Cosine with Variational Data Assimilation [1]

showed that accuracy of forecasts for some locations were worse than others across all combination methods. In the 2017/2018 season, mean out-of-sample log scores of forecasts for HHS region 2 and HHS region 6 were notably poorer compared to those for other locations. In the 2018/2019 season, mean out-of-sample log scores of forecasts for HHS region 6 and HHS region 8 were worse than scores of forecasts for other locations.

3 | PROBABILITY PLOTS OF ENSEMBLE FORECASTS BY TARGET-SEASON PAIR

Figure S5 and Figure S6 highlight season-to-season variations in the probabilistic calibration of the ensemble forecasts. For all targets, the out-of-sample forecasts from the EW-LP and LP methods were slightly too wide in the 2016/2017 season as the distributions of PIT values were concentrated around intermediate PIT values. However, the probability plots indicated under-prediction for 2 to 4 week ahead horizons in the 2017/2018 seasons as too few forecasts had PIT values below approximately 0.6. In the 2018/2019 season, they produced too wide forecasts for 1 to 2 week ahead horizons, while under-prediction drove their miscalibration for 3 to 4 week ahead horizons.

Out-of-sample forecasts produced from the beta-transformed combination methods exhibited modest under-prediction for all targets in the 2016/2017 seasons as the empirical CDF curves are below the reference line across all PIT values. The Cramer distances between the empirical CDF curves and the uniform distribution indicated the PIT values of 2 and 3 week ahead forecasts from the BLP and BMC_2 deviated farther from the uniform distribution compared to the LP's forecasts. In the 2017/2018 season,

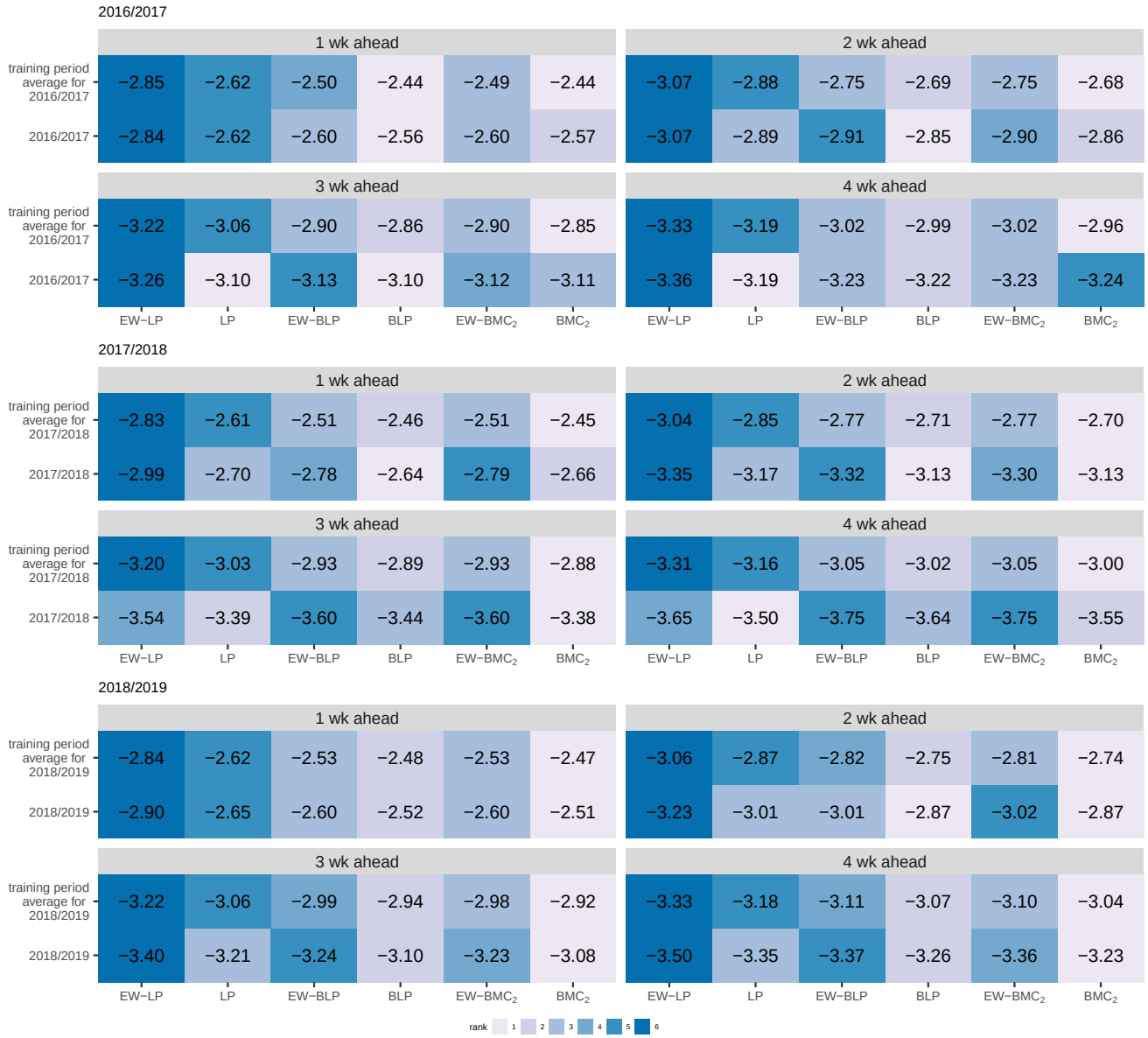

Figure S1 Mean training and out-of-sample log scores for the 1-4 week ahead targets in the 2016/2017, 2017/2018, and 2018/2019 season

their forecasts were also more miscalibrated than the LP's for all targets, while the EW-BLP and EW-BMC_2 were the most miscalibrated methods for three out of four targets, as shown in both Figure S5 and Figure S6.

The beta-transformed combination methods yielded better calibrated out-of-sample forecasts compared to the EW-LP and LP for the 1 week ahead target, but showed substantial under-prediction for the rest of the targets in the 2018/2019 season. This is in alignment with the results of summary measure of probabilistic calibration by Cramer distances, as the Cramer distances between the empirical CDF curves of the PIT values of beta-transformed combination methods and the uniform distribution are lower than Cramer distances between the empirical CDF curves of the PIT values of the EW-LP and LP and the uniform distribution for the 1 week ahead horizon, but higher for all other horizons. The EW-BLP and EW-BMC_2 were more miscalibrated compared to the BLP and BMC_2, despite having similar degrees of calibration in the previous two test seasons.

The lack of calibration of the ensemble forecasts generated from the beta-transformed combination methods for most targets in the test seasons, evident in Figure S5 and Figure S6, was in contrast with modest under-prediction and lower Cramer distances

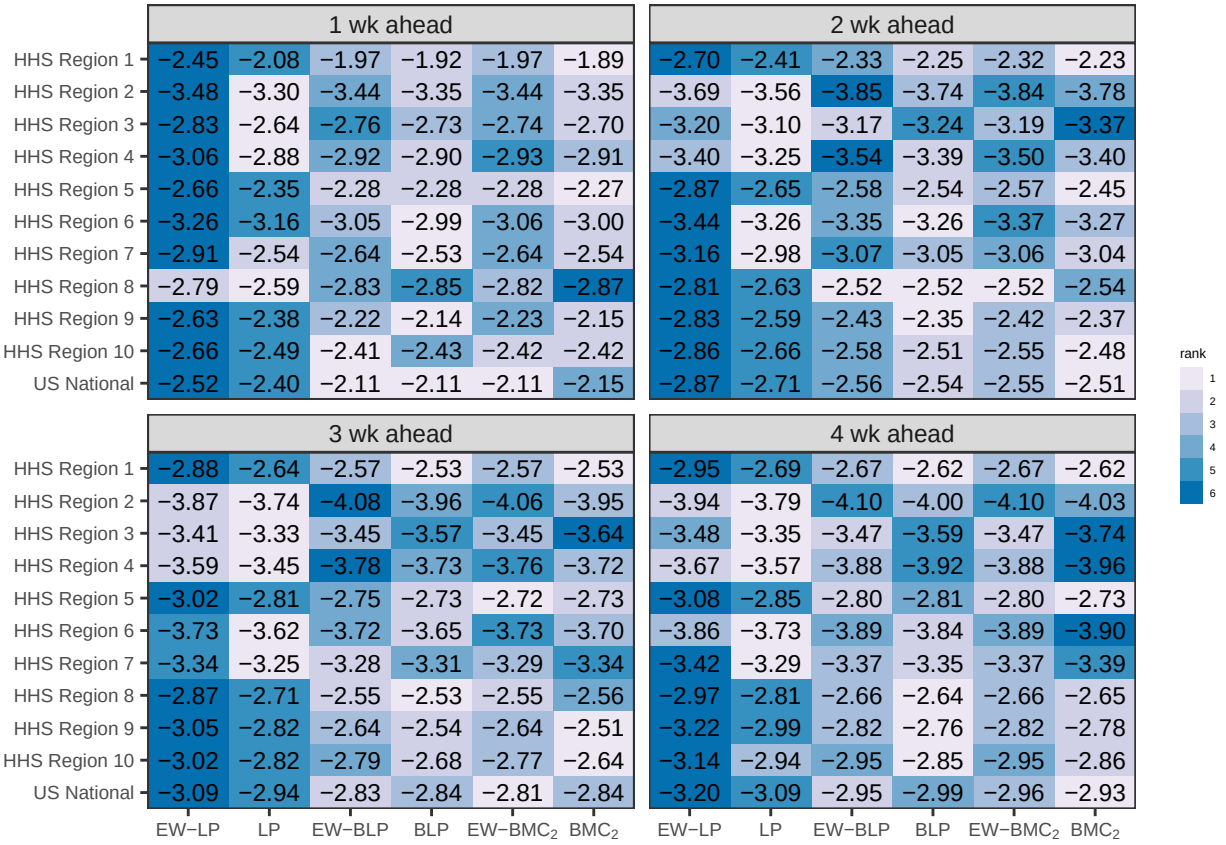

Figure S2 Mean out-of-sample log score across all weeks by HHS region in the 2016/2017 influenza season

from the uniform distribution shown in the training period. In the 2017/2018 season, which was a large influenza season in the U.S., the ensemble forecasts from all combination methods appeared to under-predict.

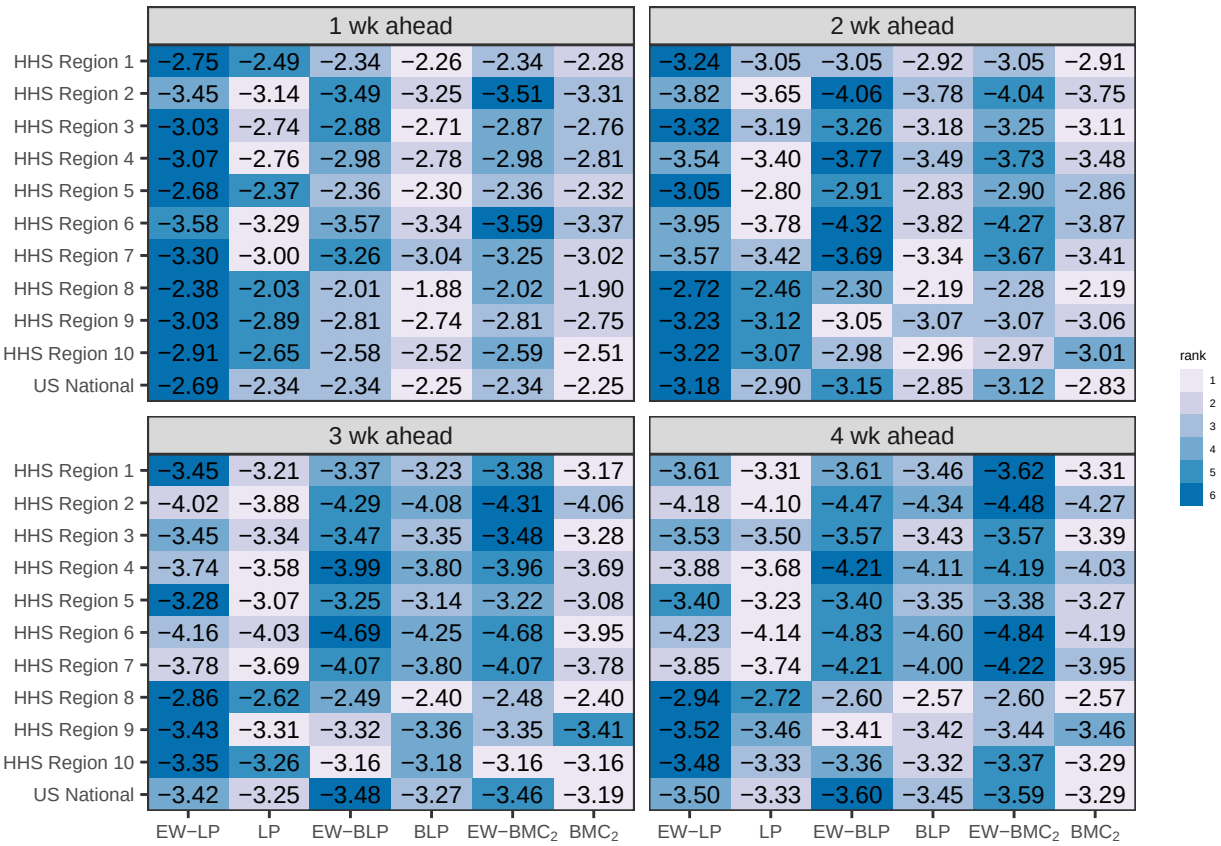

Figure S3 Mean out-of-sample log score across all weeks by HHS region in the 2017/2018 influenza season

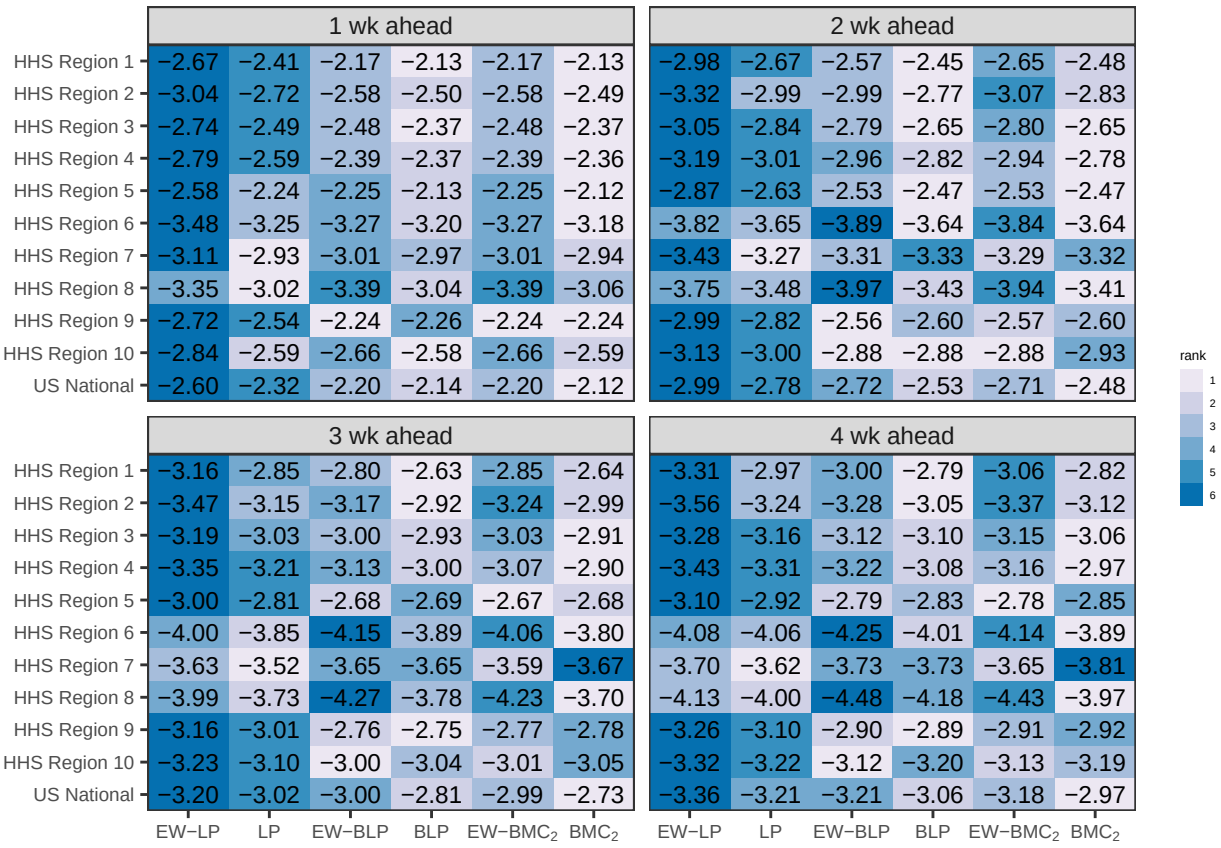

Figure S4 Mean out-of-sample log score across all weeks by HHS region in the 2018/2019 influenza season

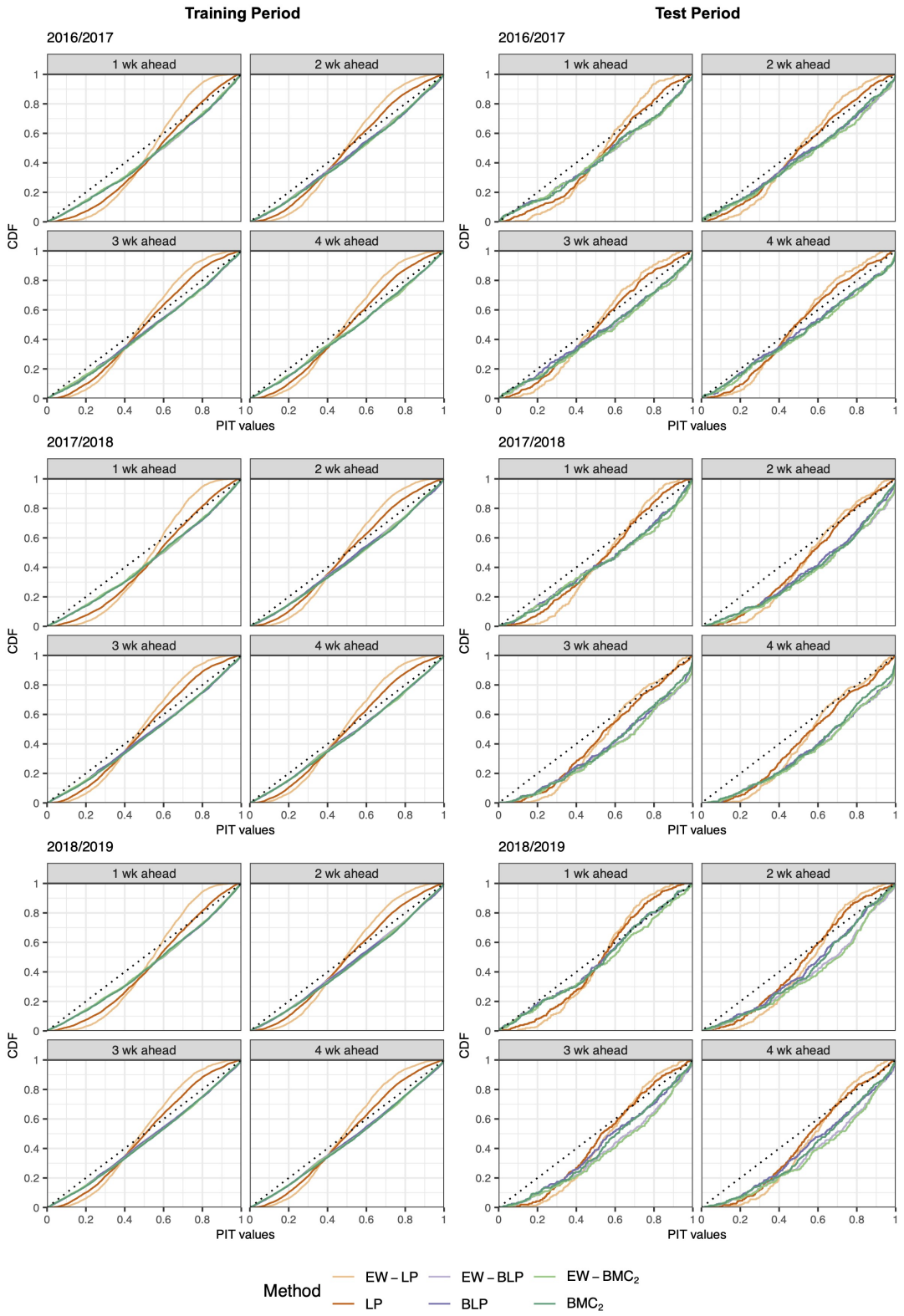

Figure S5 Probability plots by target and season. The black diagonal line is the reference line for assessing probabilistic calibration. The more an empirical CDF curve deviates from the reference line, the more miscalibrated the forecasts produced from the corresponding combination method is.

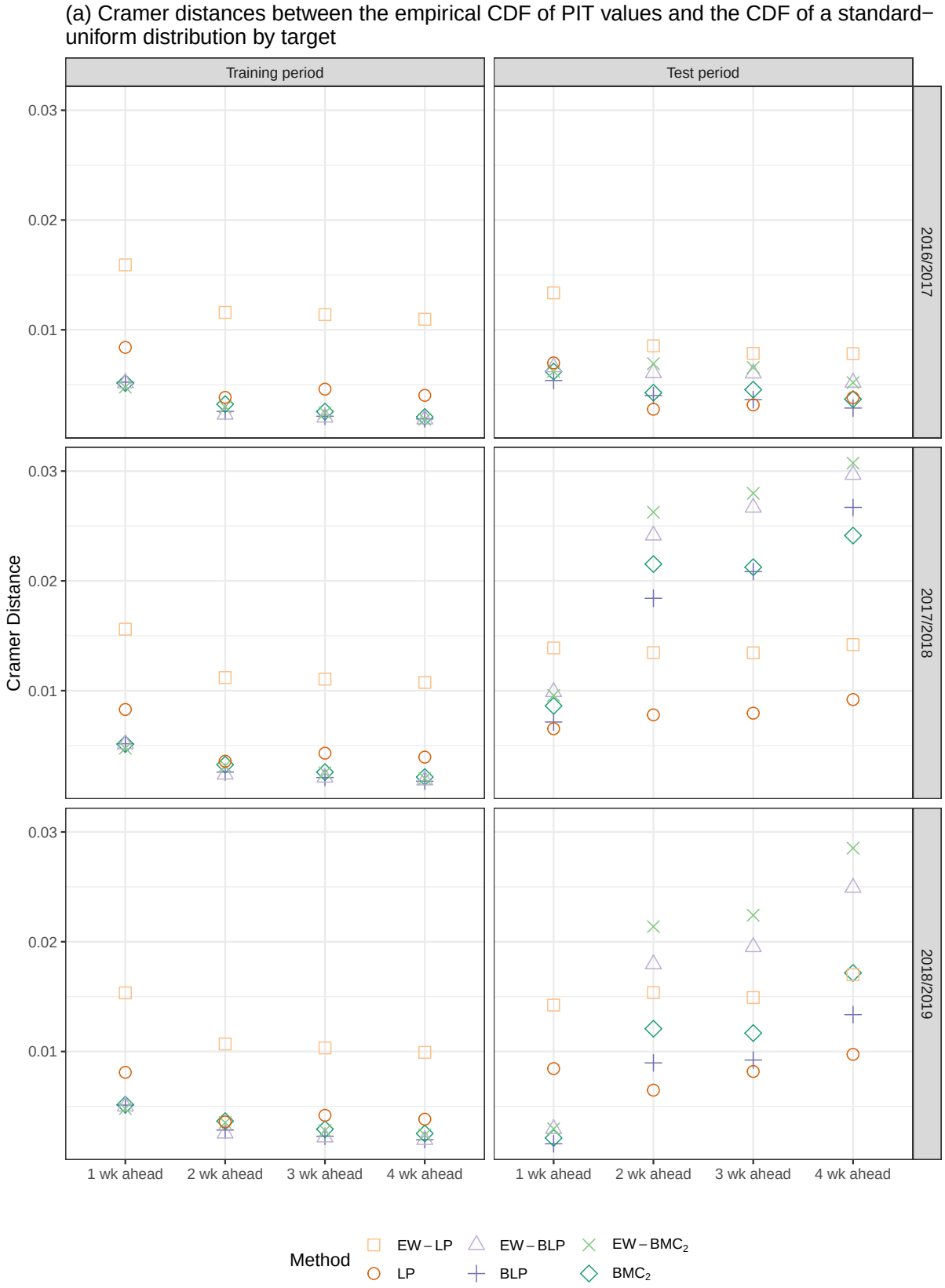

Figure S6 Cramer distances between empirical CDF curves of PIT values and the reference line in Figure S5. The higher the Cramer distance between an empirical CDF curve and the uniform distribution is, the more miscalibrated the forecasts produced from the corresponding combination method is.

References

1. FluSight Network . GitHub - FluSightNetwork/cdc-flusight-ensemble: Guidelines and forecasts for a collaborative U.S. influenza forecasting project.. URL: https://github.com/FluSightNetwork/cdc-flusight-ensemble; 2020.

2. Pei S, Shaman J. Counteracting structural errors in ensemble forecast of influenza outbreaks. *Nature Communications* 2017; 8(1): 925. doi: 10.1038/s41467-017-01033-1

3. Yang W, Karspeck A, Shaman J. Comparison of Filtering Methods for the Modeling and Retrospective Forecasting of Influenza Epidemics. *PLOS Computational Biology* 2014; 10(4): e1003583. Publisher: Public Library of Sciencedoi: 10.1371/journal.pcbi.1003583

4. Yamana TK, Kandula S, Shaman J. Individual versus superensemble forecasts of seasonal influenza outbreaks in the United States. *PLOS Computational Biology* 2017; 13(11): e1005801. Publisher: Public Library of Sciencedoi: 10.1371/journal.pcbi.1005801

5. Brooks LC, Farrow DC, Hyun S, Tibshirani RJ, Rosenfeld R. Flexible Modeling of Epidemics with an Empirical Bayes Framework. *PLOS Computational Biology* 2015; 11(8): e1004382. Publisher: Public Library of Sciencedoi: 10.1371/journal.pcbi.1004382

6. Brooks LC, Farrow DC, Hyun S, Tibshirani RJ, Rosenfeld R. Nonmechanistic forecasts of seasonal influenza with iterative one-week-ahead distributions. *PLOS Computational Biology* 2018; 14(6): e1006134. Publisher: Public Library of Sciencedoi: 10.1371/journal.pcbi.1006134

7. Dave Osthus , James Gattiker , Reid Priedhorsky , Sara Y. Del Valle . Dynamic Bayesian Influenza Forecasting in the United States with Hierarchical Discrepancy (with Discussion). *Bayesian Analysis* 2019; 14(1): 261–312. doi: 10.1214/18-BA1117

8. Ray EL, Sakrejda K, Lauer SA, Johansson MA, Reich NG. Infectious disease prediction with kernel conditional density estimation. *Statistics in Medicine* 2017; 36(30): 4908–4929. doi: 10.1002/sim.7488

9. Ray EL, Reich NG. Prediction of infectious disease epidemics via weighted density ensembles. *PLOS Computational Biology* 2018; 14(2): e1005910. arXiv: 1703.10936doi: 10.1371/journal.pcbi.1005910