Published in final edited form as:

Stat Med. 2023 November 20; 42(26): 4696–4712. doi:10.1002/sim.9884.

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

Abstract

The characteristics of influenza seasons vary substantially from year to year, posing challenges for public health preparation and response. Influenza forecasting is used to inform seasonal outbreak response, which can in turn potentially reduce the impact of an epidemic. The United States Centers for Disease Control and Prevention, in collaboration with external researchers, has run an annual prospective influenza forecasting exercise, known as the FluSight challenge. Uniting theoretical results from the forecasting literature with domain-specific forecasts from influenza outbreaks, we applied parametric forecast combination methods that simultaneously optimize model weights and calibrate the ensemble via a beta transformation and made adjustments to the methods to reduce their complexity. We used the beta-transformed linear pool, the finite beta mixture model, and their equal weight adaptations to produce ensemble forecasts retrospectively for the 2016/2017, 2017/2018, 2018/2019 influenza seasons in the U.S. We compared their performance to methods that were used in the FluSight challenge to produce the FluSight Network ensemble, namely the equally weighted linear pool and the linear pool. Ensemble forecasts produced from methods with a beta transformation were shown to outperform those from the equally weighted linear pool and the linear pool for all week-ahead targets across in the test seasons based on average log scores. We observed improvements in overall accuracy despite the beta-transformed linear pool or beta mixture methods' modest under-prediction across all targets and seasons. Combination techniques that explicitly adjust for known calibration issues in linear pooling should be considered to improve probabilistic scores in outbreak settings.

Keywords

infectious disease forecasting; epidemiology; seasonal influenza; ensemble; combination method

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

SUPPORTING INFORMATION

The following supporting information is available as part of the online article:

Individual Forecasting Model Information and Additional Results. This supporting information contains sections 1–3 including individual forecasting model information and additional results of the application.

[0]The FluSight Challenge uses a slightly different binned probability format where the i^{th} bin is defined as $[\ell_i, u_i)$[23]; this detail does not have a practical impact on the set up because the influenza-like-illness measure is continuous.

[0]Since the target variables are discretized in this application, we adapt Definition 2.5 in Gneiting and Raftery[28] by sampling a uniformly distributed PIT value between $F(l_i)$ and $F(u_i)$ where $[l_i, u_i)$ is the bin containing the observed value.

1 | INTRODUCTION

Seasonal influenza outbreaks pose public health challenges and cause a large morbidity and mortality burden worldwide. The United States Centers for Disease Control and Prevention (CDC) estimates there were 35.5 million cases of influenza, 490,600 influenza-related hospitalizations, and 34,200 deaths from influenza during the 2018–2019 influenza season in the U.S.[1]. Influenza forecasting has become integral to public health decision making[2]. A forecasting model uses data to make projections of the future trajectory of an infectious disease target, such as cases, hospitalizations and deaths, and can provide uncertainty measures of its predictions. Thus, forecasting models are a powerful tool for public health officials to improve seasonal outbreak preparedness and response, which can in turn potentially reduce the burden of seasonal influenza. The CDC's establishment of the Center for Forecasting and Outbreak Analytics in August of 2021[3] highlights a critical need to advance the use of infectious disease forecasting and modeling.

To provide public health officials real-time, prospective information about the future trajectory of seasonal influenza, the CDC, in collaboration with external researchers, started an annual prospective influenza forecasting exercise in the U.S., known as the FluSight challenge, in 2013. This exercise has been conducted with the goal of improving forecast accuracy and the integration of forecasts with real-time public health decision making. Multiple academic and non-academic groups submit weekly forecasts to the FluSight challenge. A submission typically contains probabilistic and point forecasts for seven targets in each of the 10 Health and Human Services (HHS) regions in the U.S. as well as at the national level. In this manuscript, we focus on the probabilistic forecasts, in which a predictive distribution is specified for the outcome of interest. All forecast targets are based on the weighted percentage of outpatient visits for influenza-like illness (wILI) collected through the U.S. Outpatient Influenza-like Illness Surveillance Network (ILINet), weighted by state populations.

Constructing a single ensemble forecast that combines the forecasts from multiple individual models has advantages. An ensemble forecast unifies signals from many models into a single forecast, making it easier for stakeholders to understand. In addition, ensemble forecasts have been shown to consistently achieve a high degree of accuracy and often outperform individual forecasts of infectious disease targets[4,5,6,7,8,9,10,11,12]. A subset of teams participating in the FluSight challenge has produced a collaborative multi-model ensemble, the FluSight Network ensemble, using stacked generalization —in particular, the FluSight Network ensemble is calculated as a linear combination of the individual forecasts.

Despite the success of linear combination methods such as the one used to produce the FluSight Network ensemble, their forecasts lack calibration[13,14]. Gneiting and Ranjan[14] proved that the linear aggregation increases the dispersion of the combined predictive distribution and therefore may result in overdispersed ensemble forecasts even when the individual forecasts are well-calibrated. More generally, a simple linear combination of individual forecasts may produce miscalibrated ensemble forecasts unless their calibration is adjusted for.

Previous work has presented parametric and nonparametric approaches to combining and calibrating ensemble forecasts. The beta-transformed linear pool is a combination formula that calibrates the combined predictive distribution by overlaying the linear pool with a beta transformation[15,14]. An extension to the beta-transformed linear pool using a Bayesian nonparametric approach to estimate infinite beta mixture models was proposed[16]. This method achieves a theoretically stronger result of probabilistic calibration compared to the beta-transformed linear pool by extending the flexibility of the combination function. Kuleshov and Deshpande[17] introduced calibrated risk minimization as a principle that maximizes sharpness subject to calibration by adding calibration loss as a constraint in the loss function. Rumack, Tibshirani and Rosenfeld[18] presented a post-processing method called the recalibration ensemble that combines and calibrates forecasts in separate steps and applied this method to recalibrating epidemic forecasts.

In practice, there is merit in selecting parsimonious models and combination methods with computationally efficient estimation[19,20,21]. The optimal degree of flexibility and computational complexity of combination methods often vary for different applications. Baran and Lerch[20] compare the performance of multiple forecast combination methods and assesses the degree of flexibility combination methods needed to yield the best practical results for post-processing applications in forecasting wind speed and precipitation. In influenza probabilistic forecasting, Ray and Reich[22] study a range of individual model weighing schemes with different levels of complexity in generating ensemble forecasts via the feature-weighted ensemble approach that combines aspects of linear pooling or stacking and gradient boosting. In both of these studies, the methods with an intermediate level of flexibility yielded better predictive performances in their respective applications.

This work aims to add to the growing field of infectious disease probabilistic forecasting by investigating the accuracy and probabilistic calibration of ensemble forecasts produced from combination methods that combine and calibrate simultaneously while not having any knowledge of the underlying model structure of the individual models or the ability to reproduce their forecasts in the U.S. seasonal influenza setting. Using 27 individual models from the FluSight network, we apply the linear pool, beta-transformed linear pool, and the finite beta mixture approach to combine predictive distributions. We also adapt the beta-transformed linear pool and the finite beta mixture approach by fixing individual model weights to be equal to investigate whether these more parsimonious approaches are sufficiently flexible for producing accurate and well-calibrated ensemble forecasts. We modify estimation approaches of the methods with beta transformation to accommodate the binned probability distribution representation used in the FluSight challenge[23].

Section 2 reviews the CDC influenza data, forecast targets, and forecast combination methods. Section 3 describes the application of the combination methods in seasonal influenza forecasting and presents results. Section 4 contains discussions of results in the context of related work, real-time forecasting operations, and data-driven public health decision making.

2 | METHODS

2.1 | Influenza Data

The U.S. Outpatient Influenza-like Illness Surveillance Network (ILINet) publishes the weekly percentage of outpatient doctor's office visits due to influenza-like illness weighted by state populations (wILI). ILINet is a syndromic surveillance system that includes more than 3,000 providers[24]. The CDC Influenza Division reports weekly estimates of wILI for the United States and for the 10 Health and Human Services (HHS) regions (Figure 1).

2.2 | Forecast targets

Forecasts submitted to the CDC FluSight challenge typically consists of three seasonal targets and four short-term targets. We produce ensemble forecasts of short-term 1–4 week ahead wILI for all locations from the 2016/2017 to 2018/2019 influenza season in this study. We do not include forecasts of seasonal targets, such as the peak week, peak incidence, and seasonal onset, due to the lack of importance of probabilistic forecasts after those events have been observed in a particular season.

2.3 | Forecast combination methods

Let $f_1, \ldots, f_M$ and $F_1, \ldots, F_M$ be predictive probability density functions (PDFs) and cumulative distribution functions (CDFs), respectively, for a real-valued forecast target, y, from M individual models. The combination methods described in this section include the linear pool as a baseline method and the beta-transformed linear pool and finite beta mixture combination as the methods that combine and calibrate forecasts.

2.3.1 | Linear pool (LP and EW-LP)—The linear pool is a mixture model with a predictive density

$$f_{\mathrm{LP}}(y) = \sum_{m=1}^{M} \omega_m f_m(y), \tag{1}$$

where ω_m is a nonnegative weight for the m^{th} individual model and $\sum_{m=1}^{M} \omega_m = 1$. The equally weighted linear pool (EW-LP) is a special case of the LP with the weights fixed to $\omega_m = \frac{1}{M}$.

2.3.2 | Beta-transformed linear pool (BLP and EW-BLP)—Gneiting and Ranjan[14] demonstrate that the LP produces forecasts that lack calibration when the individual forecasts are well-calibrated and propose a flexible alternative approach, the beta-transformed linear pool (BLP), which has a predictive CDF defined by

$$F_{\mathrm{BLP}}(y) = B_{\alpha,\beta}\left(\sum_{m=1}^{M} \omega_m F_m(y)\right), \tag{2}$$

where $B_{\alpha,\beta}$ denotes the CDF of the beta distribution with the parameters $\alpha, \beta > 0$, ω_m is a nonnegative weight for the m^{th} individual model weight and $\sum_{m=1}^{M} \omega_m = 1$. To find the predictive PDF of the BLP we can differentiate the above CDF, finding

$$f_{\mathrm{BLP}}(y) = \left(\sum_{m=1}^{M} \omega_m f_m(y)\right) b_{\alpha,\beta}\left(\sum_{m=1}^{M} \omega_m F_m(y)\right) \tag{3}$$

where $b_{\alpha,\beta}$ is the PDF of the beta distribution. The LP is a special case of the BLP when $\alpha = \beta = 1$. The equal weight variation of this method is the equally weighted beta-transformed linear pool (EW-BLP), which is a special case of the BLP with fixed weights $\omega_m = \frac{1}{M}$. Figure 2 demonstrates how the BLP's beta transformation operates on the LP's predictive CDF.

2.3.3 | Finite beta mixture combination (BMC_K and $\mathrm{EW}-\mathrm{BMC}_K$)—

A Bayesian approach is used to extend the BLP to finite and infinite beta mixtures for combining and calibrating predictive distributions[16]. Baran and Lerch[20] note the high computational costs of the estimating this approach. Due to the computational burden of this Bayesian approach, we choose to employ a frequentist approach to estimate a finite beta mixture model

$$F_{\mathrm{BMC}_K}(y) = \sum_{k=1}^{K} \theta_k B_{\alpha_k,\beta_k}\left(\sum_{m=1}^{M} \omega_{km} F_m(y)\right), \tag{4}$$

where K is the number of beta components, θ_k is a beta mixture weight for the k^{th} beta component, B_{α_k,β_k} denotes the CDF of the beta distribution with the parameters $\alpha_k, \beta_k > 0$, and $\omega_k = (\omega_{k1}, \ldots, \omega_{kM})$ comprises the individual model weights specific to each beta component. Differentiating the CDF, the predictive density of the BMC_K is

$$f_{\mathrm{BMC}_K}(y) = \sum_{k=1}^{K} \theta_k \left(\sum_{m=1}^{M} \omega_{km} f_m(y)\right) b_{\alpha_k,\beta_k}\left(\sum_{m=1}^{M} \omega_{km} F_m(y)\right). \tag{5}$$

The equally weighted variation of the finite beta mixture combination approach ($\mathrm{EW}-\mathrm{BMC}_K$) is a special case of the BMC_K with $\boldsymbol{\omega}_k = \left(\frac{1}{M}, \ldots, \frac{1}{M}\right)$ With $K = 1$, the BMC_K and the $\mathrm{EW}-\mathrm{BMC}_K$ become the BLP and the EW-BLP, respectively.

2.4 | Modification of the BLP and BMC_K methods for combining discrete distributions

The predictive density functions of the BLP and the BMC_K are given by the equation (3) and (5), respectively. However, the forecasts of 1–4 week ahead wILI, which is a continuous measure of disease incidence, are represented using a binned probability format in submissions to the FluSight challenge. Here we describe a modification to BLP and BMC_K models to handle this discretized representation of the target variable.

Let F denote a predictive CDF of a forecasting model, Y be the outcome variable, and $\{(\ell_i, u_i] : i = 1, \ldots, I\}$ be a collection of disjoint bins covering the set of possible outcomes for Y, with $u_i = \ell_{i+1}$ for $i < I$. An individual forecast of Y consists of an assignment of probabilities to each of the I bins:

$$P_i = Pr(\ell_i < Y \leq u_i) \tag{6}$$

$$= F(u_i) - F(\ell_i). \tag{7}$$

In order to estimate the parameters of the BMC_K, we modify the log-likelihood function, which is the log of equation (5), for a single observation y that falls in bin j to be

$$\begin{aligned}\log\left[f_{\mathrm{BMC}_K}(y)\right] &= \log\left[P_{\mathrm{BMC}_K, j}\right] \\ &= \log\left[F_{\mathrm{BMC}_K}(u_j) - F_{\mathrm{BMC}_K}(\ell_j)\right] \\ &= \log\left[\sum_{k=1}^{K}\theta_k \boldsymbol{B}_{\alpha_k, \beta_k}\left(\sum_{m=1}^{M}\omega_{km}F_m(u_j)\right) - \sum_{k=1}^{K}\theta_k \boldsymbol{B}_{\alpha_k, \beta_k}\left(\sum_{m=1}^{M}\omega_{km}F_m(\ell_j)\right)\right] \\ &= \log\left[\sum_{k=1}^{K}\theta_k \boldsymbol{B}_{\alpha_k, \beta_k}\left(\sum_{m=1}^{M}\omega_{km}\sum_{i \leq j}P_{m,i}\right) - \sum_{k=1}^{K}\theta_k \boldsymbol{B}_{\alpha_k, \beta_k}\left(\sum_{m=1}^{M}\omega_{km}\sum_{i < j}P_{m,i}\right)\right]\end{aligned}$$

where $P_{\mathrm{BMC}_K, j}$ is the probability assigned to bin j by the BMC_K's discretized predictive distribution, $F_{\mathrm{BMC}_K}(y)$ is the continuous predictive CDF of the BMC_K, $F_m(u_i)$ and $F_m(\ell_i)$ are the predictive CDFs of a individual model m, and $P_{m,i}$ is the probability assigned to bin i by individual model m's discretized predictive distribution.

Since the BLP is a special case of the BMC_K where $K = 1$, the modified log-likelihood function of the BLP is the same as above with a single beta component term in the outer summation.

3 | APPLICATION IN SEASONAL INFLUENZA FORECASTING IN THE U.S.

We apply the combination methods introduced in Section 2 to prospective forecasts from 27 individual forecasting models (Table S1, Supporting Information[25]) available in the FluSight Network repository[26] to generate weekly ensemble forecasts of 1–4 week ahead wILI for the United States and the 10 Health and Human Services (HHS) regions from the 2016/2017 to 2018/2019 influenza seasons.

3.1 | Forecast evaluation

We follow the FluSight Challenge guidelines[27] by using the logarithmic score or log score which is defined as the logarithm of the predictive density or mass function evaluated at the observed data point. The log score is a proper scoring rule that assesses the sharpness and calibration of probabilistic forecasts simultaneously[28]. In the FluSight challenge where forecasts are represented in a binned probability format, the log score is defined as

$$\begin{aligned}\mathrm{LogS}(f, y^*) &= \log\int_{\ell_i}^{u_i} f(y)dy \\ &= \log P_i\end{aligned}$$

where y^* is the observed value of the forecast target y and ℓ_i and u_i are the pre-specified lower and upper bounds of bin i such that $y^* \in [\ \ell_i\ , u_i)$.

We generate forecasts from each combination method for all combinations of week, region, target, and season, and calculate their log scores. Following the CDC scoring convention[23], we truncate log scores to be no lower than −10. The benefit of this approach is that it enables us to average log scores for a method even when that method receives a log score of $-\infty$ (assigning zero probability to an observed value) for any forecasts. However, this modified log score is formally no longer a proper score. Log scores are averaged across all forecast regions and weeks for each target and test season to get summary measures of accuracy for each method to compare their performance.

The calibration of a probabilistic forecast addresses the statistical consistency between the predictive distributions of forecasts and the observations. The concept of calibration allows us to assess whether a model produces reliable forecasts, i.e. whether an event that the model assigns a particular predicted probability really occurs that at that frequency in the long run. The forecast f is probabilistically calibrated if its probability integral transform (PIT) values are uniformly distributed on the unit interval[29]. The probability integral transform, $z_i = F(y^*)$, is the probability obtained from evaluating the predictive CDF of a model at the observed value, y^*.

To assess the probabilistic calibration of the ensemble forecasts (i.e., the uniformity of the PIT values), we use the graphical tool called the probability plot, which plots the empirical CDF of the PIT values. Specifically, we compute the PIT values of all observations in the test seasons and plot their empirical CDFs by target and season. The empirical CDF curve should follow a 45-degree line bisecting the plot if the forecasts are probabilistically calibrated. In the case where deviations from uniformity are observed, the shape of empirical CDF curve of the PIT values suggests the causes behind the lack of probabilistic calibration[30]. For example, PIT values concentrating near 0 and 1 indicates that the observed values fall on the tails of the predictive distribution of the forecasts more frequently than expected, i.e., the probability plot shows the slope steeper than 1 near the PIT values of 0 and 1, so that the predictive distributions were too narrow. To quantitatively measure the deviation of a PIT CDF curve from a standard uniform CDF, we compute the Cramer distance[28,31], $\int_{-\infty}^{\infty} (F(x) - G(x))^2 dx$, where $F(x)$ is an empirical CDF of PIT values and $G(x)$ is a standard uniform CDF. The Cramer distance can be viewed as a summary measure of calibration, however, it lacks the diagnostic property of the probability plot. The lower a Cramer distance is, the less an empirical CDF of PIT values deviates from a standard uniform CDF overall.

3.2 | Parameter estimation

The test data set includes the 2016/2017, 2017/2018, and 2018/2019 influenza seasons. When generating ensemble forecasts for a test season, the training data set consists of all the influenza seasons preceding that test season, starting with the 2010/2011 season. Under this framework, a different number of training seasons is used for each test season.

The parameters of each combination method are estimated simultaneously by maximizing the average log score, which is positively oriented (i.e., higher scores are better), via maximum likelihood estimation over a training data set. The parameters are chosen to be target-specific to allow for variations among targets. We update the parameter estimates for each test season, so there are 12 sets of parameters to be estimated for four targets and three test seasons. We modify the estimation approach for the BLP, EW-BLP, BMC_K, and $\mathrm{EW}-\mathrm{BMC}_K$ as outlined in the Methods section in order to apply the combination methods to individual forecasts in a binned probability representation.

3.2.1 | Choice of K for finite beta mixture combination approaches—

We use a leave-one-season-out cross validation approach to select the number of beta components, K, in the BMC_K and $\mathrm{EW}-\mathrm{BMC}_K$ for each target-test season pair. Specifically, we train the BMC_K and $\mathrm{EW}-\mathrm{BMC}_K$ using $K = 2$ through 5 on each subset of data in the training data with one season left out and use those ensemble fits to generate forecasts for the left out influenza season. Log scores for all combination methods are calculated for all unique forecasts, then averaged across all weeks, regions, and validation seasons to obtain a single mean validation log score for each target and method. In order to take model complexity into account, we calculate mean validation log scores across all locations for each validation season in training seasons, compute a standard error for each target-test season pair, and select the smallest K for BMC_K and $\mathrm{EW}-\mathrm{BMC}_K$ with mean validation log scores within 1 standard error of the best log score in a particular target-test season pair. As a result, models with best mean validation log scores are not necessarily selected if a more parsimonious model could achieve similar mean validation log scores.

Based on the mean validation log scores in Table 1, $K = 2$ is selected for the BMC_K and $\mathrm{EW}-\mathrm{BMC}_K$ for all four targets and three seasons. The finite beta mixture combination methods with $K = 2$ had the best mean validation log scores in every instance other than the 2 week ahead target in the 2018/2019 season. Overall, using a higher number of beta components in the finite beta mixture approaches does not substantially improve mean out-of-sample log scores in our application. Thus, the finite beta mixture methods with the most parsimonious number of parameters are selected.

3.3 | Results

3.3.1 | Overall Summary—

Based on mean out-of-sample log scores across all targets and seasons (Figure 3 panel (b)), the BMC_2 is the most accurate method, followed by the BLP and LP. Across all three test seasons, the BMC_2 outperformed the other five methods for 3 and 4 week ahead horizons, and performed as equally well as the BLP for the 2 week ahead horizon (Figure 3 panel (a)). The BLP is the best performing method for the 1 week ahead horizon. The BMC_2 is also the best performing method for the 2017/2018 and 2018/2019 season based on mean out-of-sample log scores across all four horizons, while the BLP is the best performing method for the 2016/2017 season. These results indicate that the BLP and BMC_2 can consistently improve the accuracy of ensemble forecasts compared to the other commonly used methods included in this study despite season-to-season and target variations.

Across all test seasons, the 1 week ahead forecasts from the BLP and BMC$_2$ methods are more probabilistically calibrated than other methods based on the probability plots and their Cramer distances from the standard uniform CDF (Figure 5(a) and Figure 6(a)). However, the forecasts produced from all beta-transformed methods became less calibrated as forecast horizons increased. Across all targets, the LP and BLP methods produced most calibrated forecasts in the 2016/2017 and 2018/2019 seasons, followed by the BMC$_2$ method. The LP is the most calibrated across all targets and and test seasons, which is due to its stable performance across all test seasons and in particular its substantially better calibration relative to the other methods in the 2017/2018 season.

3.3.2 | Comparison of combination methods' accuracy—Across all targets and seasons, the BMC$_2$ has the best mean out-of-sample log score of −3.02, though it only marginally outperformed the BLP and LP, which have mean out-of-sample log scores of −3.03 and −3.06, respectively (Figure 3(b)). Across all three test seasons, the BLP has the best mean out-of-sample log scores for the 1 week ahead horizon (Figure 3(a)). The BMC$_2$, which is the most flexible method in this study, has the best mean out-of-sample log scores of −3.19 and −3.34 for 3 and 4 week ahead horizons, respectively. It also performed as well as the BLP, which also has the best mean out-of-sample log score of −2.95 for the 2 week ahead horizon. Across all four target horizons, the BLP is the most accurate method for the 2016/2017 season, while the BMC$_2$ is the most accurate for the 2017/2018 and 2018/2019 season.

The observation-level log scores of the ensemble forecasts from the beta-transformed combination methods exhibit higher variation compared to those from the EW-LP and the LP for all targets and test seasons (Figure 4). Mean out-of-sample log scores of 1–4 week ahead national-level forecasts in test seasons by epiweek in Figure 7(b) indicate that accuracy deteriorates near the time at which the peak incidence was observed, especially for the BLP and BMC$_2$. These results are in alignment with the theoretical finding that the LP tends to produce overdispersed or wider forecasts, as evident by their wider prediction intervals compared to those of the BLP and BMC$_2$ forecasts in Figure 7(a), resulting in less extreme log scores.

The performance of the BMC$_2$ method, which is the method with the highest number of estimated parameters, was slightly less consistent in the test seasons compared to its superior performance across all targets and seasons in the training periods. Nonetheless, it was always among the top two methods in terms of out-of-sample log score across all targets and seasons (Figure 3). We also see notably higher variation in log scores across all methods for all targets in the 2017/2018 season, which was one of the most severe and longest flu seasons in the recent years[32]. Section 2 in the Supporting Information[25] provides detailed results of mean out-of-sample log scores by location, target, test season.

3.3.3 | Comparison of combination methods' calibration—The empirical CDF curves of PIT values from probabilistically calibrated forecasts should follow the CDF of a standard uniform distribution, that is, a diagonal line between 0 and 1. The more an empirical CDF curve of the PIT values deviates from the reference line, the more miscalibrated the forecasts are. To quantify the deviation from the CDF of a standard uniform distribution, we computed the Cramer distances between the empirical CDF of PIT values and the CDF of a standard uniform distribution.

Overall, all combination methods produced forecasts that lack probabilistic calibration in the test period. The BLP and BMC_2 methods are more probabilistically calibrated than other methods for the 1 week ahead horizon, as their empirical CDF curves are less deviated from the reference line (Figure 5(a)) and their Cramer distances are the lowest (Figure 6(a)). However, the forecasts produced from all beta-transformed methods became less calibrated as forecast horizons increased. Across 2 to 4 week ahead forecast horizons in the test period, the empirical CDF curves of the PIT values from the forecasts from the EW-BMC_2 were the most miscalibrated among all beta-transformed methods as indicated by its Cramer distances being the highest.

Probability plots by season (Figure 5(b)) show that ensemble forecasts from all methods are relatively well-calibrated in the 2016/2017 season, while they are most miscalibrated in the 2017/2018 season. In the 2018/2019 season, the LP, BLP and BMC_2 methods produced noticeably better calibrated forecasts compared to forecasts from the EW-BLP and EW-BMC_2. In the 2017/2018 season, the forecasts from all methods tended to too low, but the observed wILI were generally still captured in the upper tail of the predictions from the LP and EW-LP methods, whereas the beta-transformed combination methods under-predicted more systematically. This is illustrated for the national-level forecasts in Figure 7.

Recall that according to theory, the LP will produce ensemble forecasts with too wide predictive distributions when individual models are well-calibrated[14]. The results (Figure 5) in our application during the training period are consistent with this theory. Specifically, the forecasts from the LP and EW-LP tended to be too wide, i.e., more observed values concentrated near the center of predictive distributions than expected for a well-calibrated model as indicated by the slopes of PIT CDF curves being higher than 1 for intermediate PIT values and lower than 1 near 0 and 1 (Figure 5).

Despite the under-prediction across horizons observed in both training and test periods, the beta-transformed combination methods' probabilistic calibration was notably better in the training period, especially for 3 and 4 week ahead horizons (Figure 5(a) and Figure 6(a)) and for the 2017/2018 season (Figure 5(b) and Figure 6(b)). The added flexibility afforded by these methods enables them to adjust for observed dynamics in the training seasons. Coupled with substantially larger observed disease incidence in the 2017/2018 season compared to the training seasons (Figure 1), beta-transformed combination methods' poor out-of-sample calibration was misaligned with its performance on the training period as a result, which also adversely affected the calibration performance aggregated by target across all test seasons.

The probability plots by target-season pairs (Figure S6, Supporting Information[25]) show similar calibration results as in Figure 5 —the empirical CDF curves of the PIT values of forecasts produced from the LP methods in the test seasons also appear miscalibrated in the lower tail. The calibration of all methods by target-season pairs are discussed in more details in Section 3 in Supporting Information[25].

3.3.4 | Comparison of accuracy and calibration between combination methods and their more regularized counterparts—

Comparing performance of particular pairs of methods provide insights into how using more or less regularized methods affect forecasts' performance. The LP method can be considered a regularized approach of the BLP method, as the calibration parameters, α and β, are regularized to one. The BLP can be deemed a regularized approach of BMC_2, as the number of beta components (K) is fixed to one, resulting in an incrementally lower level of complexity. Likewise, the EW-LP and EW-BLP methods are also regularized versions of the EW-BLP and the EW-BMC_2 methods, respectively.

As measured by log scores, methods that used beta calibration were generally better than methods that did not, but adding more flexibility for calibration generally did not lead to additional gains in performance (Figure 3). The EW-BLP method outperformed the EW-LP method for all four targets and all three test seasons, and the BLP method outperformed the LP method for three out of four targets, especially the one and two week ahead targets, and two out of three test seasons. However, the EW-BLP and EW-BMC_2 methods had similar mean log scores for all targets and test seasons, as did the BLP and BMC_2 methods. Comparing the calibration performance of the BLP and BMC_2 methods to their regularized counterparts, it was noted that more regularized methods had a less severe degradation of calibration performance moving from the training period to the test period, especially when calibration results were aggregated across all seasons and in the 2017/2018 season when calibration results were aggregated across all targets (Figure 5 and Figure 6). Comparing EW-LP to EW-BLP and EW-BLP to EW-BMC_2 shows a similar pattern of calibration performance.

Similarly, the methods with equal individual weights (EW-LP, EW-BLP, EW-BMC_2) can also be viewed regularized versions of the corresponding methods with optimally weighted individual models, as the individual models' weights are regularized to $\frac{1}{M}$. The equally weighted variations of the combination methods (EW-LP, EW-BLP, and EW-BMC_2), though more parsimonious, had sub-optimal forecast accuracy compared to their counterparts that assigned weights to the individual models in this application. The EW-LP was the worst overall method across all targets and seasons, and all equally weighted variations had worse mean out-of-sample log scores compared to their more complex counterparts (Figure 3). Additionally, the equally weighted ensembles generally had poorer calibration than the corresponding weighted variations in both the training period and the test period, as their Cramer distances are notably higher than those of the BLP and BMC_2 in the test period (Figure 5 and Figure 6).

4 | DISCUSSION

As demonstrated in the forecasting literature, in many settings ensemble forecasts have consistent superior performance and give decision makers the ability to unify the strengths and diversity of individual models into one forecast. These particular advantages are of great importance in practice for infectious disease forecasting[6,33,9,34,8,12]. This work aims to offer insight into forecast accuracy and calibration of parametric combination methods in which calibration and individual model weight estimation happen simultaneously in the application of seasonal influenza forecasting in the U.S.

We applied the linear pool, beta-transformed linear pool, and the finite beta mixture combination method to available forecasts in the FluSight challenge to produce ensemble forecasts for seasonal influenza in the U.S. retrospectively for three test seasons and compared their performance. Our results showed that two of the combination methods included in this study, the BLP and BMC_2, offered consistently superior forecast accuracy relative to the LP and EW-LP. Either the BLP or the BMC_2 or both delivered better mean log scores across the test seasons compared to the other methods for all targets and across all targets for all seasons. Despite using different methods to create ensemble forecasts, our findings are in agreement with the findings in Rumack, Tibshirani and Rosenfeld[18] that combination methods that take into account forecast calibration improve accuracy of seasonal influenza ensemble forecasts in the U.S.

The BMC_2 uses twice as many parameters as the BLP, but only marginally outperformed it in two out of four targets and five out of twelve target-season pairs. Considering the large number of individual models in the FluSight challenge, the BLP may be more easily applicable in practice compared to the BMC_2 as it has half as many individual model weights and beta parameters to estimate. Although the LP under-performed relative to the BLP and BMC_2 for most targets and seasons, the differences in mean log scores were typically small. It was also observed that the BLP and BMC_2 methods' poor observation-level mean log scores are less frequent, but more extreme than the LP's, leading to better aggregated mean log scores. More parsimonious combination methods with fixed, equally weighted individual model weights, namely the EW-LP, EW-BLP and EW-BMC_2, appear to not be flexible enough to deliver superior performance compared to the other methods in this study. While this is the case for this application, combination methods using equal weights with or without the beta transformation could be useful in other applications where it might be difficult to estimate individual model weights when available models change over time or training data are limited[12].

The results on the probabilistic calibration of the ensemble forecasts measured by the uniformity of the PIT values are less straightforward. The results for the LP and EW-LP forecasts indicate that they had too wide predictive distributions across all targets. Despite the beta-transformed combination methods' success at correcting the overdispersion of the ensemble forecasts from the LP and EW-LP, their forecasts exhibited a pattern of systematic under-prediction. This under-prediction is relatively more pronounced at longer forecast horizons, especially in the 2017/2018 influenza season. In the 2017/2018 season, which was a large influenza season in the U.S.[32], the ensemble forecasts from all combination methods under-predicted to some extent. However, the BLP and BMC_2 methods had particularly poor calibration that season, indicating that they may have adapted to the dynamics of the training seasons, which were smaller in scale. Note that an overdispersed forecaster, such as the LP and EW-LP, has the advantage of being more likely to capture an extreme season, though it may not be optimal overall based on proper scores. These more conservative methods may be desirable in applications in which stakeholders want to avoid missing a large season at the expense of having too wide forecasts.

Additionally, the parameters of the the finite beta mixture combination methods were estimated by maximum likelihood, which is not equivalent to a measure of probabilistic calibration. If probabilistic calibration is of critical importance for the application at hand, other approaches such as post-hoc calibration techniques that directly target a calibrated forecast distribution may be appropriate.

Exploring different approaches to select training periods can be useful. For instance, Baran and Lerch[20] use rolling training periods in the application of a similar set of combination methods to wind speed and precipitation forecasting and Rumack, Tibshirani and Rosenfeld[18] constructs a training period that takes into account seasonality of epidemic forecasting. In addition, combination methods that require the joint estimation of all parameters, including the parameters in the individual models, may be considered in a setting where the underlying model structure of individual models are known. An example of one of these methods is the mixture EMOS model[35]. Since the FluSight challenge provides forecasts only, in this work we selected combination methods that do not require reproduction of forecasts from individual models.

For combining forecasts in outbreak settings, the beta-transformed linear pool (BLP) is a promising alternative to standard linear pooling (LP) methods. Compared with LP, the BLP has only two additional parameters, α and β, and the simple modification of the log likelihood function makes the BLP applicable to combining forecasts in a binned probability format. The BMC_2 may add value in instances where the BLP is not flexible enough, though we only see marginal improvement in the mean out-of-sample log scores in our application.

As infectious disease forecasting has come to the forefront of the public health effort in formulating well-informed policies in response to outbreaks, it is critical to gain insight on model combination approaches in order to combine individual models' strengths and to produce accurate ensemble forecasts. This study demonstrates an effort to improve our understanding of how forecast combination methods compare in a setting of an infectious disease with well-established surveillance data pipeline like seasonal influenza.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

ACKNOWLEDGMENTS

This work has been supported by the National Institutes of General Medical Sciences (R35GM119582) and the U.S. CDC (1U01IP001122). The content is solely the responsibility of the authors and does not necessarily represent the official views of NIGMS, the National Institutes of Health, or CDC.

DATA AVAILABILITY STATEMENT

The data that support the findings of this study are openly available in Zenodo at https://doi.org/10.5281/zenodo.3454212. The code that implements the methods in this study is available on Github at https://github.com/NutchaW/forecast_calibration.

References

1. Centers for Disease Control and Prevention. Estimated Influenza Illnesses, Medical visits, Hospitalizations, and Deaths in the United States — 2018–2019 influenza season | CDC. URL: https://www.cdc.gov/flu/about/burden/2018-2019.html; 2021.
2. Lutz CS, Huynh MP, Schroeder M, et al. Applying infectious disease forecasting to public health: a path forward using influenza forecasting examples. BMC Public Health 2019; 19(1): 1659. doi: 10.1186/s12889-019-7966-8 [PubMed: 31823751]
3. Centers for Disease Control and Prevention. CDC Stands Up New Disease Forecasting Center. URL: https://www.cdc.gov/media/releases/2021/p0818-disease-forecasting-center.html; 2021.
4. Yamana TK, Kandula S, Shaman J. Superensemble forecasts of dengue outbreaks. Journal of The Royal Society Interface 2016; 13(123): 20160410. Publisher: Royal Societydoi: 10.1098/rsif.2016.0410 [PubMed: 27733698]
5. 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 [PubMed: 29107987]
6. DeFelice NB, Little E, Campbell SR, Shaman J. Ensemble forecast of human West Nile virus cases and mosquito infection rates. Nature Communications 2017; 8(1): 14592. Publisher: Nature Publishing Groupdoi: 10.1038/ncomms14592
7. Viboud C, Sun K, Gaffey R, et al. The RAPIDD ebola forecasting challenge: Synthesis and lessons learnt. Epidemics 2018; 22: 13–21. doi: 10.1016/j.epidem.2017.08.002 [PubMed: 28958414]
8. Ray EL, Wattanachit N, Niemi J, et al. Ensemble Forecasts of Coronavirus Disease 2019 (COVID-19) in the U.S.. medRxiv 2020: 2020.08.19.. doi: 10.1101/2020.08.19.20177493 [PubMed: 20177493]
9. Chowell G, Luo R, Sun K, Roosa K, Tariq A, Viboud C. Real-time forecasting of epidemic trajectories using computational dynamic ensembles. Epidemics 2020; 30: 100379. doi: 10.1016/j.epidem.2019.100379
10. Cramer EY, Ray EL, Lopez VK, et al. Evaluation of individual and ensemble probabilistic forecasts of COVID-19 mortality in the United States. Proceedings of the National Academy of Sciences 2022; 119(15): e2113561119. Publisher: Proceedings of the National Academy of Sciencesdoi: 10.1073/pnas.2113561119
11. Oidtman RJ, Omodei E, Kraemer MUG, et al. Trade-offs between individual and ensemble forecasts of an emerging infectious disease. Nature Communications 2021; 12(1): 5379. Publisher: Nature Publishing Groupdoi: 10.1038/s41467-021-25695-0
12. Ray EL, Brooks LC, Bien J, et al. Comparing trained and untrained probabilistic ensemble forecasts of COVID-19 cases and deaths in the United States. arXiv:2201.12387 [stat] 2022. arXiv: 2201.12387.
13. Hora SC. Probability Judgments for Continuous Quantities: Linear Combinations and Calibration. Management Science 2004; 50(5): 597–604. Publisher: INFORMS.
14. Gneiting T, Ranjan R. Combining predictive distributions. Electronic Journal of Statistics 2013; 7(0): 1747–1782. doi: 10.1214/13-EJS823
15. Ranjan R, Gneiting T. Combining probability forecasts. Journal of the Royal Statistical Society: Series B (Statistical Methodology) 2010; 72(1): 71–91. doi: 10.1111/j.1467-9868.2009.00726.x
16. Bassetti F, Casarin R, Ravazzolo F. Bayesian Nonparametric Calibration and Combination of Predictive Distributions. Journal of the American Statistical Association 2018; 113(522): 675–685. doi: 10.1080/01621459.2016.1273117
17. Kuleshov V, Deshpande S. Calibrated and Sharp Uncertainties in Deep Learning via Simple Density Estimation. arXiv:2112.07184 [cs] 2021. arXiv: 2112.07184.
18. Rumack A, Tibshirani RJ, Rosenfeld R. Recalibrating probabilistic forecasts of epidemics. PLOS Computational Biology 2022; 18(12): e1010771. Publisher: Public Library of Sciencedoi: 10.1371/journal.pcbi.1010771 [PubMed: 36520949]
19. Claeskens G, Magnus JR, Vasnev AL, Wang W. The forecast combination puzzle: A simple theoretical explanation. International Journal of Forecasting 2016; 32(3): 754–762. doi: 10.1016/j.ijforecast.2015.12.005

20. Baran S, Lerch S. Combining predictive distributions for the statistical post-processing of ensemble forecasts. International Journal of Forecasting 2018; 34(3): 477–496. doi: 10.1016/j.ijforecast.2018.01.005

21. Stanescu A, Pandey G. Learning parsimonious ensembles for unbalanced computational genomics problems. In: World Scientific. 2016 (pp. 288–299)

22. 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 [PubMed: 29462167]

23. McGowan CJ, Biggerstaff M, Johansson M, et al. Collaborative efforts to forecast seasonal influenza in the United States, 2015–2016. Scientific Reports 2019; 9(1): 683. Publisher: Nature Publishing Groupdoi: 10.1038/s41598-018-36361-9 [PubMed: 30679458]

24. Centers for Disease Control and Prevention. U.S. Influenza Surveillance System: Purpose and Methods | CDC. URL: https://www.cdc.gov/flu/weekly/overview.htm; 2020.

25. Wattanachit N, Ray EL, McAndrew T, Reich NG. Supplement to “Comparison of Combination Methods to Create Calibrated Ensemble Forecasts for Seasonal Influenza in the U.S.”. 2022.

26. Tushar A, Reich NG, Yamana TK, et al. FluSightNetwork/cdc-flusight-ensemble: End of 2018/2019 US influenza season. URL: 10.5281/zenodo.3454212; 2019

27. Centers for Disease Control and Prevention. Epidemic Prediction Initiative. URL: https://predict.cdc.gov/post/5d827e75fba2091084d47b96; 2019.

28. Gneiting T, Raftery AE. Strictly Proper Scoring Rules, Prediction, and Estimation. Journal of the American Statistical Association 2007; 102(477): 359–378. doi: 10.1198/016214506000001437

29. Gneiting T, Balabdaoui F, Raftery AE. Probabilistic forecasts, calibration and sharpness. Journal of the Royal Statistical Society: Series B (Statistical Methodology) 2007; 69(2): 243–268. doi: 10.1111/j.1467-9868.2007.00587.x

30. Laio F, Tamea S. Verification tools for probabilistic forecasts of continuous hydrological variables. Hydrology and Earth System Sciences 2007; 11(4): 1267–1277. Publisher: Copernicus GmbHdoi: 10.5194/hess-11-1267-2007

31. Rizzo ML, Székely GJ. Energy distance. WIREs Computational Statistics 2016; 8(1): 27–38. doi: 10.1002/wics.1375

32. Centers for Disease Control and Prevention. What You Should Know for the 2017–2018 Influenza Season. URL: https://www.cdc.gov/flu/about/season/flu-season-2017-2018.htm; 2019.

33. Reich NG, McGowan CJ, Yamana TK, et al. Accuracy of real-time multi-model ensemble forecasts for seasonal influenza in the U.S.. PLOS Computational Biology 2019; 15(11): e1007486. Publisher: Public Library of Sciencedoi: 10.1371/journal.pcbi.1007486 [PubMed: 31756193]

34. McAndrew T, Reich NG. Adaptively stacking ensembles for influenza forecasting with incomplete data. arXiv:1908.01675 [cs, stat] 2020. arXiv: 1908.01675.

35. Baran S, Lerch S. Mixture EMOS model for calibrating ensemble forecasts of wind speed. Environmetrics 2016; 27(2): 116–130. doi: 10.1002/env.2380 [PubMed: 27812298]

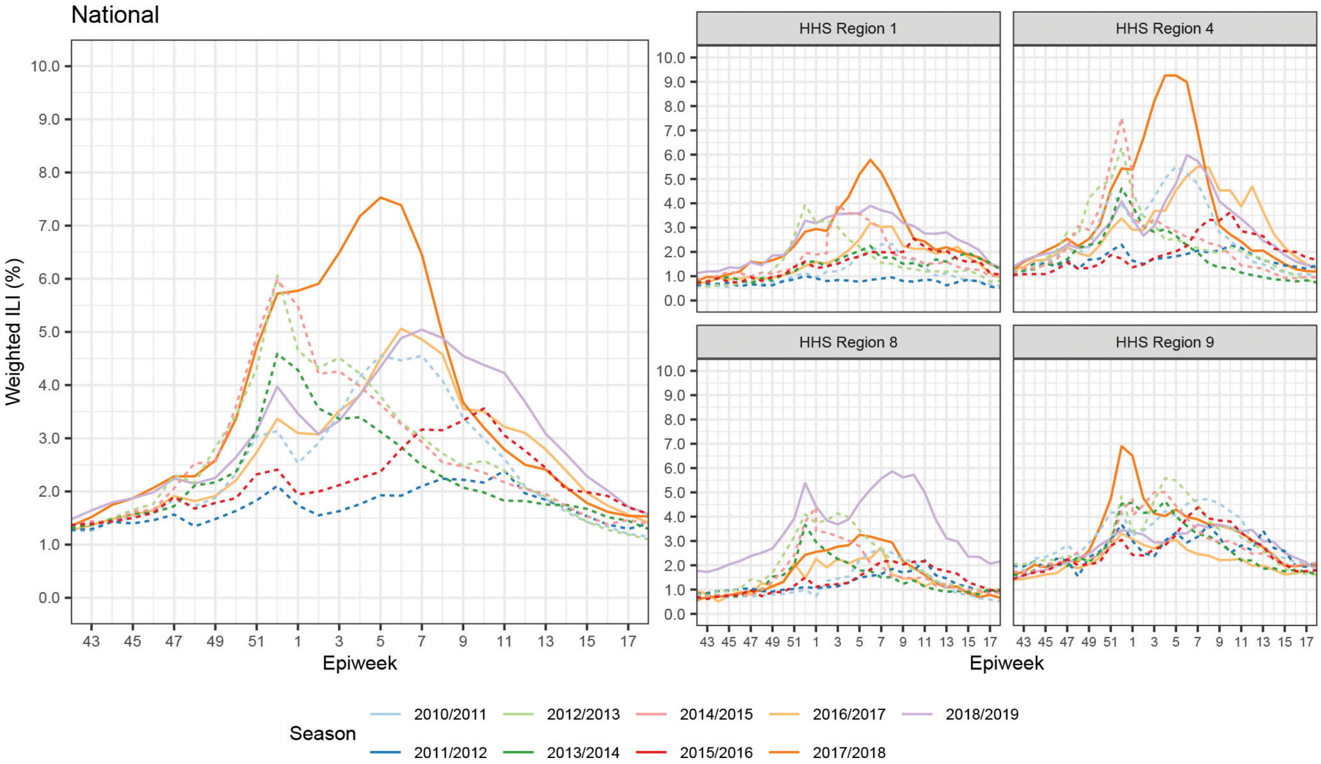

Figure 1.

Influenza-like illness weighted by state populations (wILI) data at the national level (left panel) and four HHS regions (right panel) from the 2010/2011 to 2018/2019 influenza season published by the U.S. Outpatient Influenza-like Illness Surveillance Network (ILINet). The 2016/2017, 2017/2018, and 2018/2019 seasons, which are the test seasons, are represented in solid lines. A high level of weighted ILI was observed at a national level and in three selected HHS regions in the 2017/2018 season and in HHS region 8 in the 2018/2019 season.

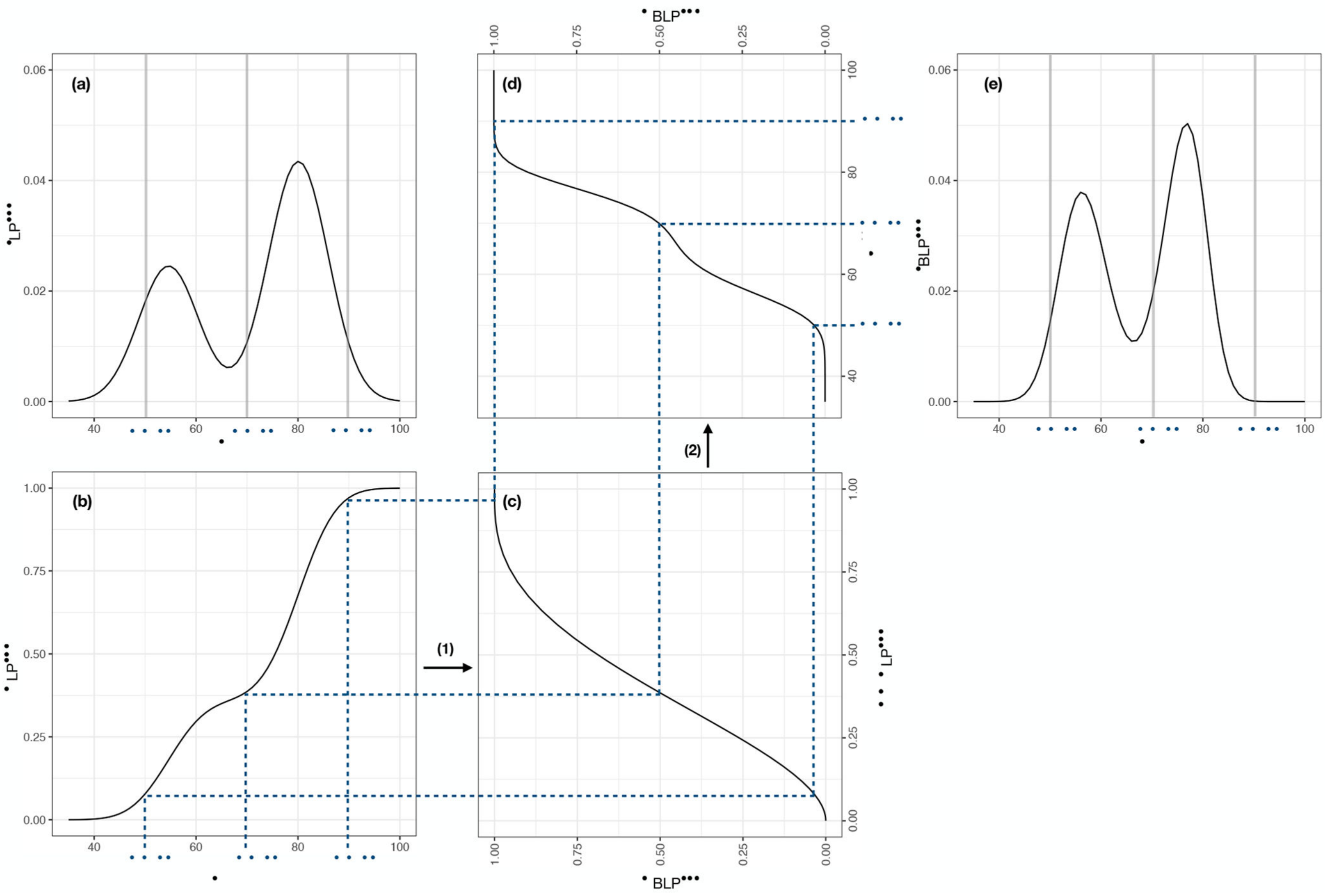

Figure 2.

Illustrative example of the BLP’s beta transformation of $F_{\mathrm{LP}}(x)$. For demonstration purposes, the individual model weights of the LP and BLP are fixed to be the same. We start with a predictive density from a linear pool, $f_{\mathrm{LP}}(x)$ in panel (a), and the corresponding CDF $F_{\mathrm{LP}}(x)$ in panel (b). Step (1) shows the beta transformation of $z = F_{\mathrm{LP}}(x)$ in panel (b) through $F_{\mathrm{BLP}}(z) = \mathrm{B}_{\alpha,\beta}(z)$ with $\alpha = 2$ and $\beta = 3$ in panel (c). Step (2) shows the BLP’s predictive CDF $F_{\mathrm{BLP}}(x)$, panel (d), as a result of the beta transformation in the first step. Panel (d) shows that the beta-transformed CDF concentrates probability closer to the median and its predictive density, panel (e), is narrower compared to that of the LP in panel (a). Other choices of the parameters for the beta transform could lead to a still narrower distribution after the transformation, a wider distribution, or an asymmetric adjustment that acts differently in the left and right tails.

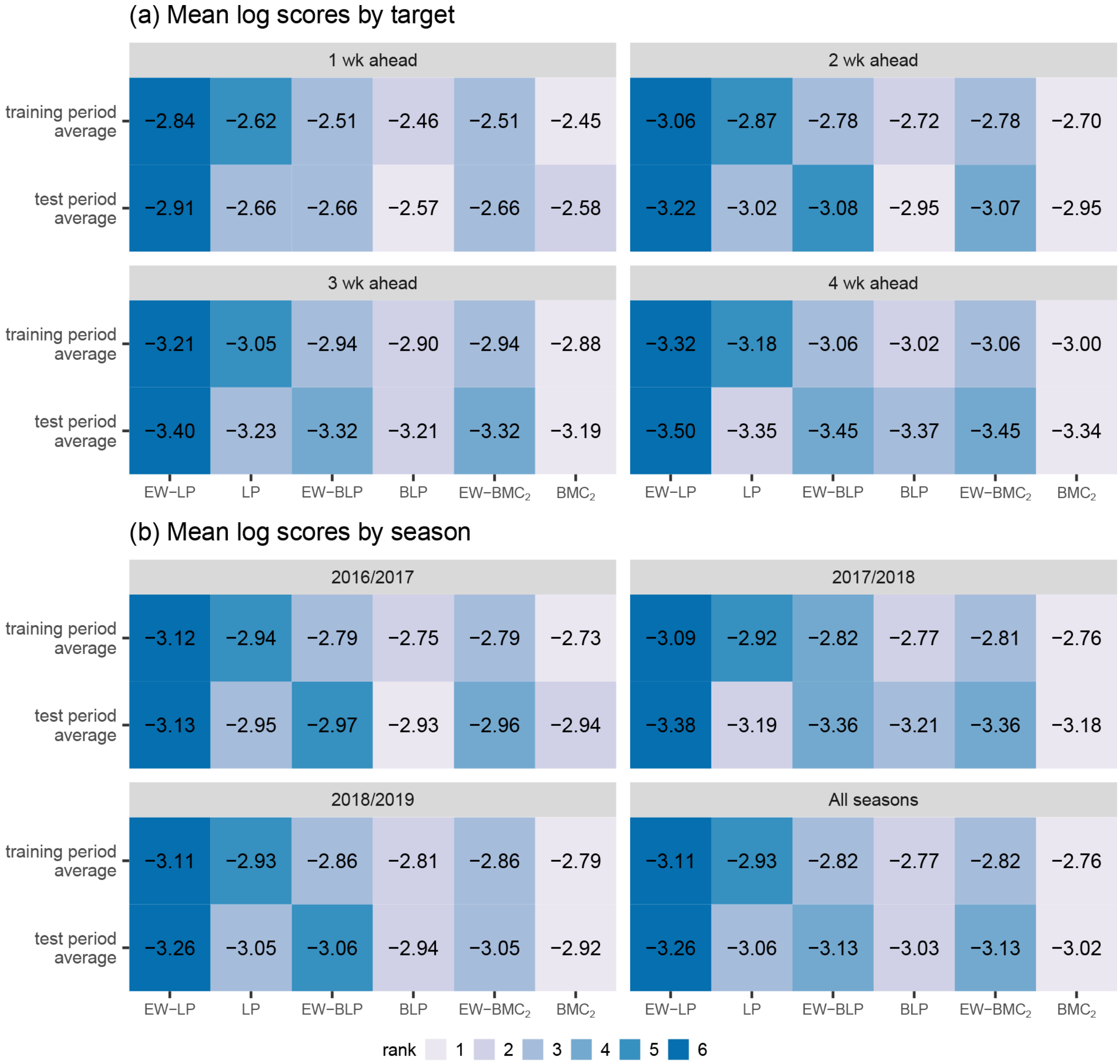

Figure 3.

Mean training and out-of-sample log scores of ensemble forecasts of wILI in the U.S. Higher log scores (lower ranks) indicate better accuracy. Panel (a) shows mean training and out-of-sample log scores by target. The BLP and BMC$_2$ are two best performing methods for three out of four forecast horizons based on out-of-sample log scores. Panel (b) shows mean training and out-of-sample log scores by season and across all seasons. The BLP and BMC$_2$ are two best performing methods for two out of three test seasons.

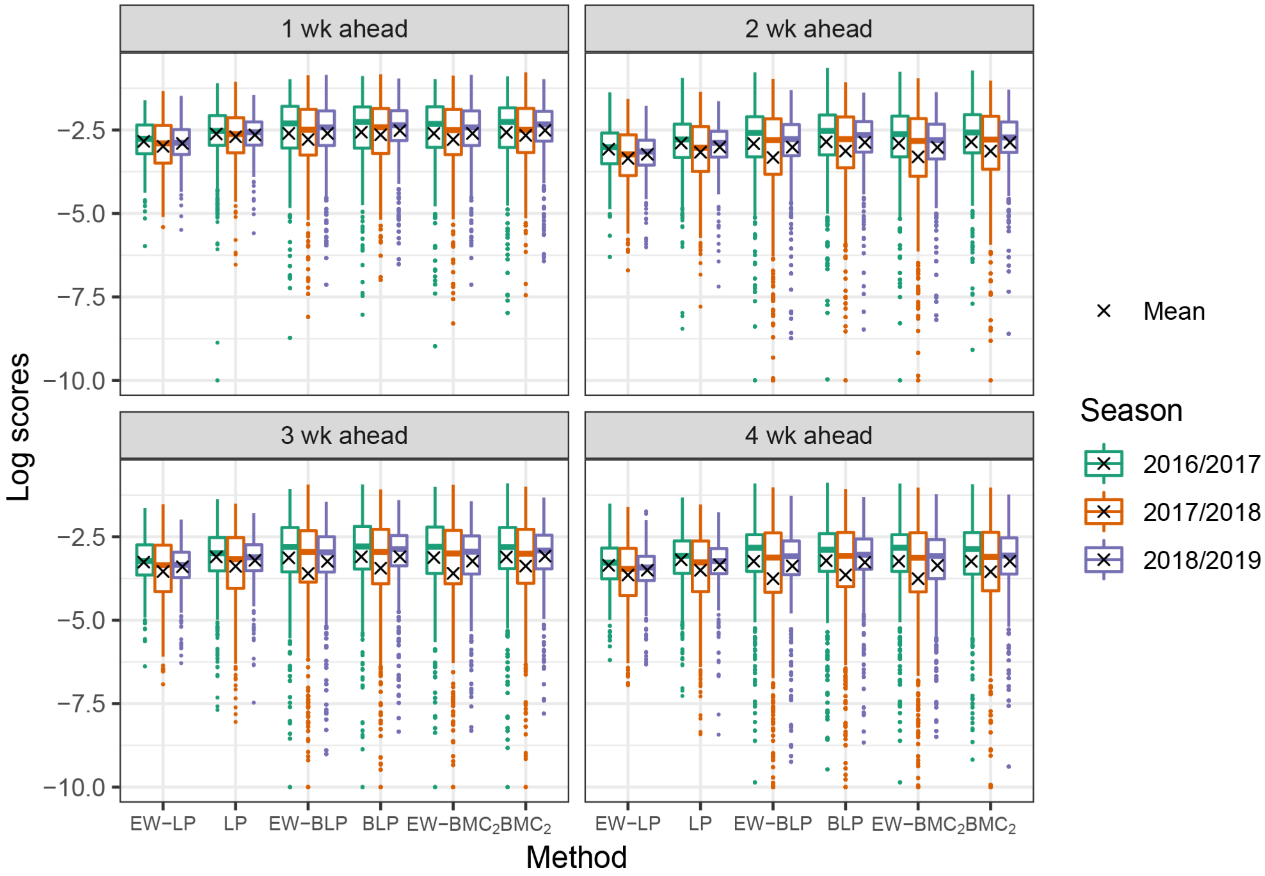

Figure 4.

Boxplots of out-of-sample, observation-level log scores of ensemble forecasts of wILI in the U.S. by target and season. Each black cross marker represents a mean out-of-sample log score for a particular target and season. The observation-level log scores of forecasts from the beta-transformed combination methods are highly variable compared to those from the EW-LP and the LP for all four targets and three test seasons.

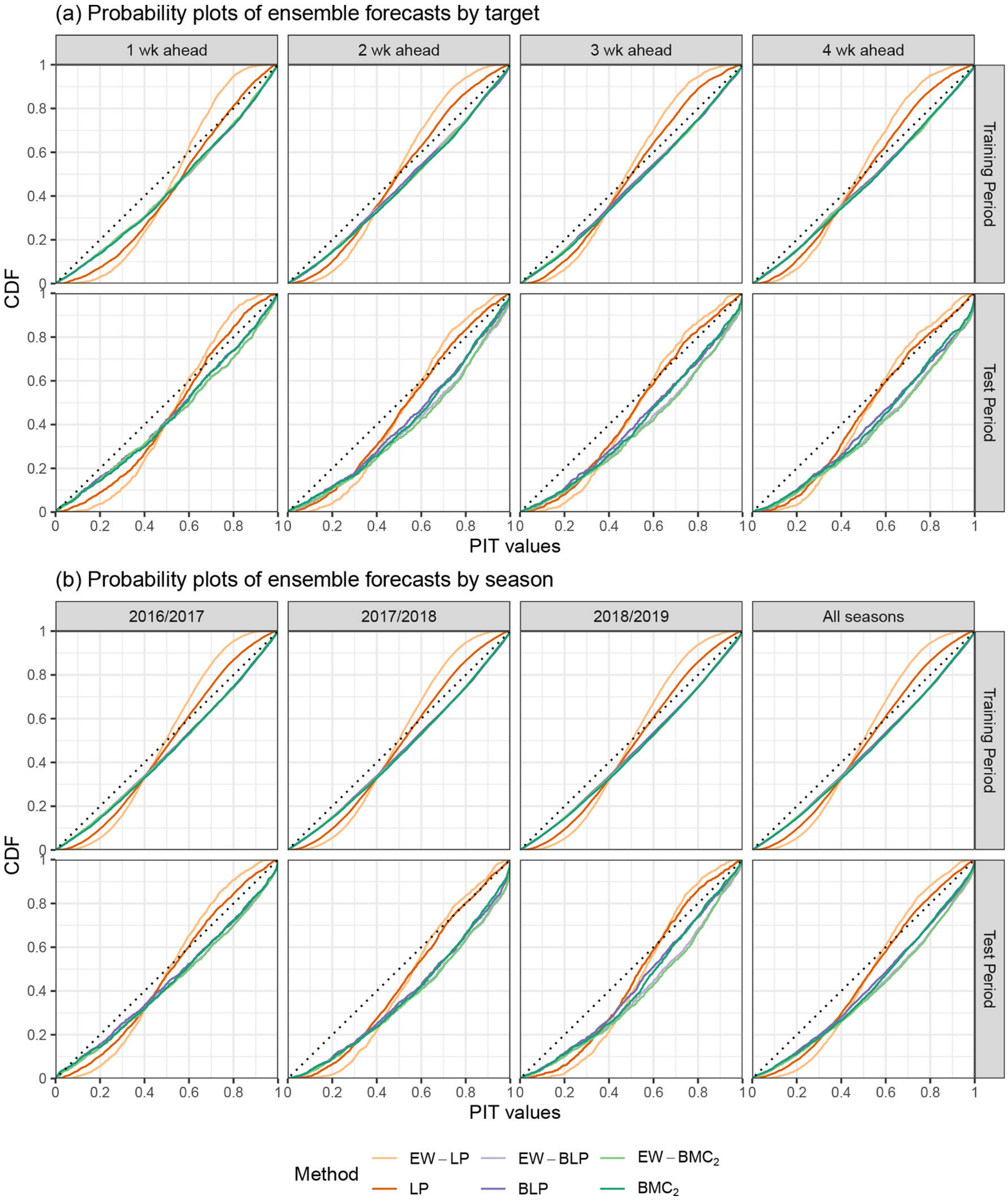

Figure 5.

Probability plots show the empirical CDF curves of PIT values of ensemble forecasts in the training and test periods by target and season. The black diagonal dashed line is the CDF of a standard uniform distribution used as the reference line for assessing probabilistic calibration. The more the empirical CDF curves of PIT values are deviated from uniformity, the less calibrated forecasts are. Panel (a) shows probability plots of ensemble forecasts by target. In the test period, the BLP and BMC$_2$ methods are more calibrated than other methods for the 1 week ahead horizon, while the LP method are more calibrated at farther forecast horizons. Panel (b) shows probability plots of ensemble forecasts by season. Forecasts from all methods are least calibrated in the 2017/2018 season. The LP, BLP and BMC_2 methods achieve a similar degree of calibration in the 2016/2017 and 2018/2019 season. In both panels, all beta-transformed combination methods display varying degrees of under-prediction and the EW-LP and LP methods are often miscalibrated in the lower tail of the predictions.

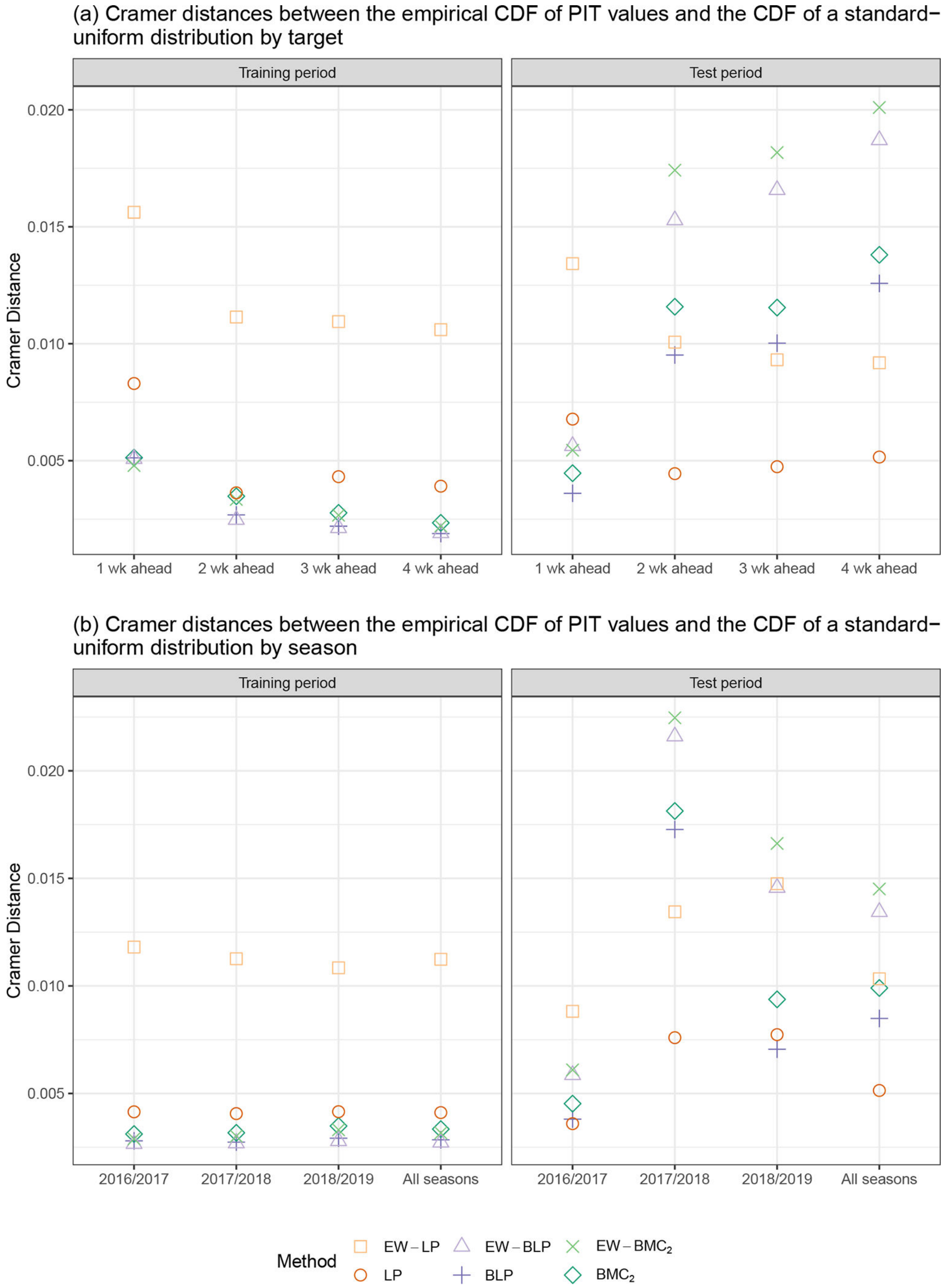

Figure 6.

Cramer distances between empirical CDF curves of PIT values and the CDF of a standard uniform distribution. The higher the Cramer distance is, the more miscalibrated the forecasts produced from the corresponding combination method are. Panel (a) shows Cramer distances by target. The LP method has the lowest Cramer distances for three out of four targets in the test period. Panel (b) shows Cramer distances by season and across all targets and seasons. The LP method has the lowest Cramer distances in two out of three seasons, while the BLP has the lowest Cramer distance in the 2018/2019 season.

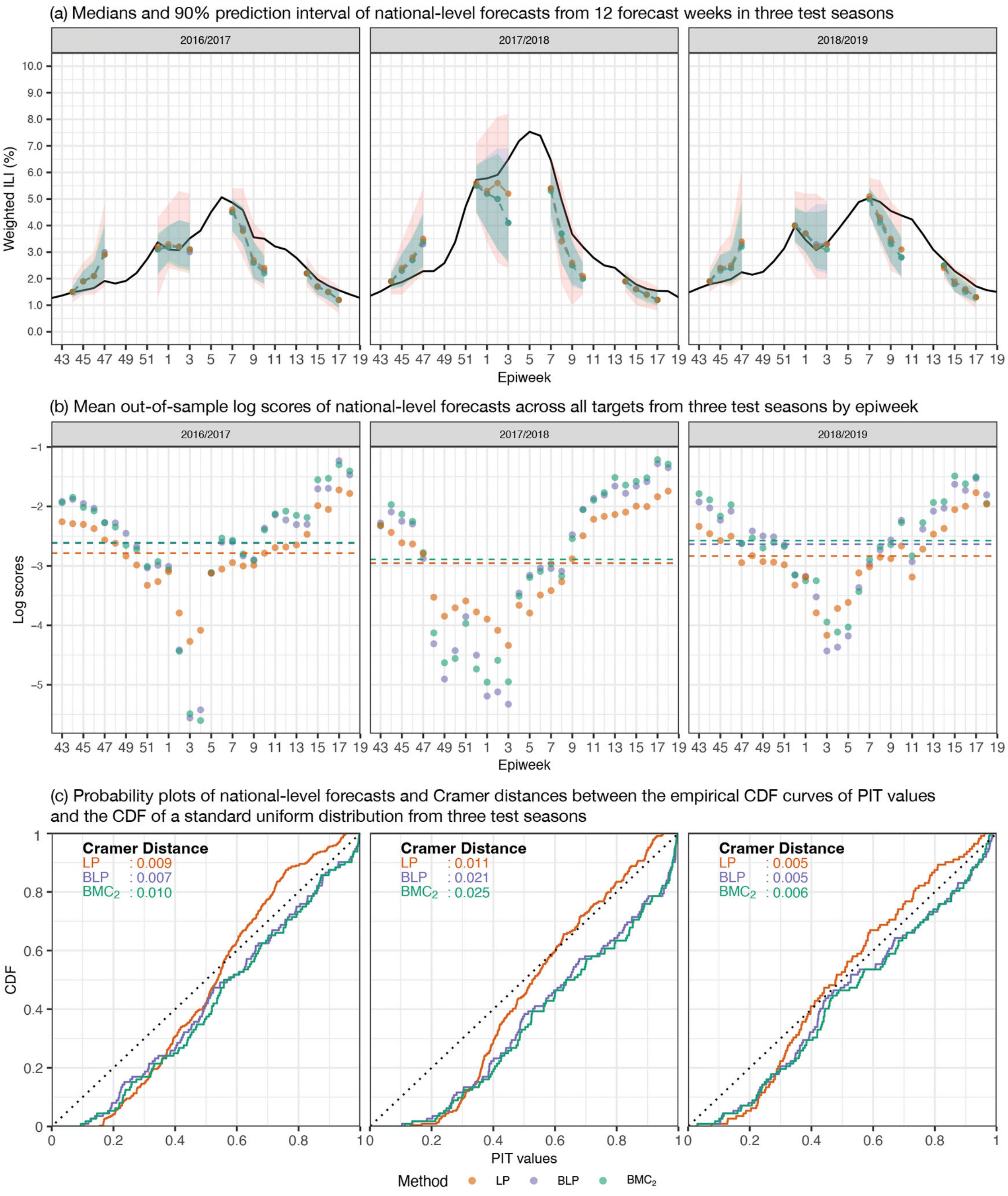

Figure 7.

Panel (a) shows point estimates (medians) with 90% prediction intervals from 1–4 week ahead probabilistic forecasts of national-level wILI generated from the LP, BLP, and BMC$_2$ with observed wILI in black solid lines. The LP has wider prediction intervals than the other two methods. Panel (b) shows mean out-of-sample log scores for national-level forecasts, averaging across all targets by test season with method-specific seasonal means in dashed lines. Mean log scores worsen near the time when the peak incidence occurred relative to the means. The BLP's and BMC$_2$'s seasonal means are similar in the first two test seasons.

Panel (c) shows the probability plots of national-level forecasts and Cramer distances by test season (from left to right, 2016/2017, 2017/2018, and 2018/2019). All methods' Cramer distances are similar in the first and last test seasons. In the 2017/2018 season, the LP is better calibrated in the upper tail where the BLP and BMC_2 substantially under-predicted.

Table 1

Mean validation log scores of BMC_K and $\mathrm{EW} - \mathrm{BMC}_K$ for all target-season pairs. The selection of the number of beta components (K) outlined in 3.2.1 takes both model complexity and mean log scores into account. Scores shown in bold and in italics are the selected methods and methods with best mean validation log scores, respectively.

2016/2017

Week ahead	**BMC$_2$**	**BMC$_3$**	**BMC$_4$**	**BMC$_5$**	**EW-BMc$_2$**	**EW-BMC$_3$**	**EW-BMC$_4$**	**EW-BMC$_5$**
1	***−2.49***	−2.50	−2.50	−2.50	***−2.50***	*−2.50*	*−2.50*	*−2.50*
2	***−2.74***	−2.75	−2.75	−2.76	**−2.76**	*−2.76*	*−2.76*	*−2.76*
3	***−2.95***	*−2.95*	*−2.95*	−2.97	***−2.92***	*−2.92*	*−2.92*	*−2.92*
4	***−3.08***	−3.09	−3.10	−3.11	***−3.03***	*−3.03*	−3.04	−3.04
2017/2018								
1	***−2.49***	−2.50	−2.50	−2.51	***−2.51***	*−2.51*	*−2.51*	−2.52
2	***−2.75***	*−2.75*	−2.76	−2.77	***−2.78***	*−2.78*	*−2.78*	*−2.78*
3	***−2.96***	*−2.96*	−2.97	−2.98	***−2.94***	−2.95	−2.95	−2.95
4	***−3.09***	*−3.09*	−3.10	−3.10	***−3.06***	*−3.06*	*−3.06*	*−3.06*
2018/2019								
1	***−2.51***	−2.52	−2.52	−2.53	***−2.55***	*−2.55*	*−2.55*	*−2.55*
2	**−2.80**	*−2.79*	−2.81	−2.80	**−2.85**	*−2.84*	−2.85	−2.85
3	***−2.99***	−3.00	−3.00	−3.01	***−3.03***	*−3.03*	*−3.03*	*−3.03*
4	***−3.13***	−3.14	*−3.13*	−3.14	***−3.15***	*−3.15*	*−3.15*	*−3.15*