BAYESIAN STACKING VIA PROPER SCORING RULE OPTIMIZATION USING A GIBBS POSTERIOR

Spencer Wadsworth
Iowa State University
Ames, Iowa
sgw96@iastate.edu

Jarad Niemi
Iowa State University
Ames, Iowa
niemi@iastate.edu

1 Abstract

Combining probabilistic forecasts into an ensemble forecast is standard practice in collaborative forecast projects where linear pooling is one of the most common methods. A common idea is that weight selection methods should be tailored to the specific research question and often selecting weights via optimization of a proper scoring rule. Bayesian predictive synthesis has also emerged as a model probability updating scheme which provides a Bayesian solution to selecting model weights more flexible than Bayesian model averaging. Existing methods may or may not improve ensembles for given datasets, and there is room for additional methodology. In this manuscript, we introduce a Gibbs posterior on stacked model weights based on minimizing a proper scoring rule. The Gibbs posterior extends stacking into a Bayesian framework by allowing for optimal weight solutions to be influenced by a prior distribution and include uncertainty quantification. We compare ensemble forecast performance with model averaging methods and equal weighted models in simulation studies and in a real data example from the 2023-24 US Centers for Disease Control FluSight competition. In both the simulation studies and the FluSight analysis, the stacked Gibbs posterior produces ensemble forecasts which perform better than the ensembles from other methods considered.

***K*eywords** Optimal linear pooling · Gibbs posterior · Probabilistic forecasting · Proper scoring rules

2 Introduction

Forecasting future events is the object of much scientific and social activity and informs many public and private decisions. Decision making tends to be better if uncertainties are attached to the forecasts (Ramos et al., 2013; Joslyn and LeClerc, 2012), and in many fields the focus on making forecasts probabilistic is growing (Wang et al., 2023; Hong et al., 2020; Kapetanios et al., 2015; Gneiting and Katzfuss, 2014; Collins, 2007; Palmer, 2002). It's rare that nature's data generating processes can be known, so a common forecasting approach is to produce multiple statistical models or predictive procedures each targeting the same event. One may then search for and select the best forecasts or in some way combine forecasts to improve overall forecasting (Yao et al., 2018; Clyde and Iversen, 2013; Biggerstaff et al., 2016; Bernardo and Smith, 1994). Large forecast competitions and official forecast hubs exist to exploit the skill of multiple forecasts and forecasters (Mathis et al., 2024; Cramer et al., 2022; Hyndman, 2020; Makridakis et al., 2020; Reich et al., 2019a; Biggerstaff et al., 2016; Hong et al., 2016). In these initiatives, multiple forecasters submit separate forecasts targeting the same event, and the performances of the forecasts, typically assessed via a *proper scoring rule*, are compared. Combining multiple candidate forecasts into an ensemble forecast is common practice in forecast hubs, and this often leads to forecasts which are superior to individual individual forecasts (Wang et al., 2023; Li et al., 2023; Gyamerah et al., 2020; Li et al., 2019; Reich et al., 2019b). For example in the 2023-24 United States Centers for Disease Control and Prevention (CDC) collaborative flu forecasting competition, also known as FluSight, ensemble forecasts largely outperformed individual forecasts in forecasting flu hospitalizations, and in fact the forecast published weekly by the CDC as the official forecast is an ensemble of all competing forecasts (Mathis et al., 2024). Figure 1 shows an example of FluSight hospitalization forecasts from many competing teams on the left where on the right is shown the ensemble forecast constructed by combining all competing forecasts.

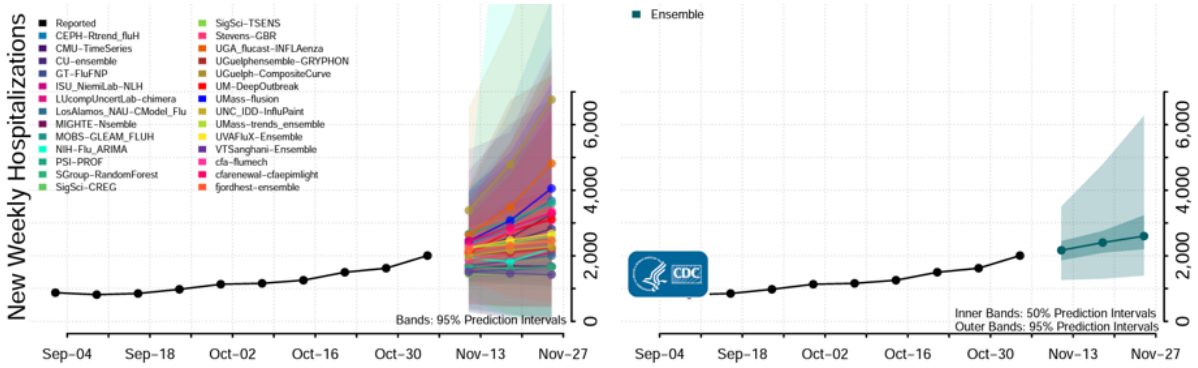

Figure 1: United States flu hospitalization forecasts from multiple competing forecast teams (left) for 1-4 week ahead horizons from the week of November 6, 2023 and an ensemble forecast (right) made by combining all competing forecasts into one. Image downloaded from https://www.cdc.gov/flu/weekly/flusight/flu-forecasts.htm

There are many methods for constructing an ensemble forecast including but not limited to linear and nonlinear methods (Yao et al., 2018; Geweke and Amisano, 2011; Hall and Mitchell, 2007; Bassetti et al., 2018; Gneiting and Ranjan, 2013; Ranjan and Gneiting, 2010) which combine probability density/mass functions (PDF), cumulative distribution functions (CDF), or quantile functions (Lichtendahl Jr et al., 2013). The focus in this manuscript is on combining forecasts via *linear pooling*, also known as model stacking (Yao et al., 2018; Geweke and Amisano, 2011; Hall and Mitchell, 2007; Stone, 1961). Most linear pooling combination methods involve selecting a set of combination weights, a point on the simplex, which optimize the ensemble according to some criteria but which leave out room for providing uncertainty in the optimization. Weights in an ensemble may also be selected by maximizing a likelihood or via a Bayesian posterior distribution, but when so constructed ensemble forecasts may not be suited for the problem specific criteria under which forecasts are assessed. Bayesian model probabilities may be used to select weights, but selecting weights this way may also not be well suited to problem specific needs. An important case of linear pooling is combining via a simple average over the forecasts. This equal weighting (EQW) scheme frequently outperforms sophisticated combination methods leading to what is known as the "forecast combination puzzle" (Frazier et al., 2023; Claeskens et al., 2016; Smith and Wallis, 2009).

In this manuscript, we introduce a new method which we call the stacked Gibbs posterior (SGP), for combining probabilistic forecasts by optimizing a linear pool. The SGP provides uncertainty quantification of the weights in the ensemble and allows the use of a prior distribution to inform or influence the optimized weights. The SGP is attained by constructing a Gibbs posterior which is a type of posterior probability distribution constructed not within the standard Bayesian framework but by specifying a risk function which one desires to optimize, assigning a prior distribution to the optimization parameters, and updating parameter distributions via Bayes theorem (Martin and Syring, 2022; Bissiri et al., 2016). The SGP is a Gibbs posterior designed specifically to optimize model weights in a forecast with the object being to minimize a proper scoring rule (Gneiting and Raftery, 2007).

The motivation for this work was ensemble modeling of the hospitalization forecasts for the FluSight project, with the hope of constructing a well performing forecast. Section 3 includes a brief review of probabilistic forecasts and linear pool combinations of forecasts. In section 4 the SGP is developed, and important details of the Gibbs posterior and proper scoring rules are covered. Section 5 includes two simulation studies to analyze the SGP and compare it with other forecast combination methods. In these studies the SGP outperforms model averaging and EQW methods for combining forecasts. In 6 the SGP is used to construct ensemble forecasts for the FluSight project and again outperforming model averaging and EQW approaches. The manuscript is concluded in section 7.

3 Probabilistic forecasts and linear pooling

We consider the situation where a set of data $y_1, ..., y_T$ is given, and one is tasked with producing a probabilistic forecast for the unobserved $\tilde{y}_{T+h}$ for h time steps ahead of the latest time T. The forecaster produces a forecast P which is probabilistic in nature and which may also be indexed by $t \in \{1, ..., T\}$. P may be a PDF, CDF, quantile function, set of quantiles or intervals, or any other representation used in probabilistic forecasting (Gneiting and Katzfuss, 2014). For the remainder of this paper, unless otherwise stated, we assume P is the CDF forecast of a continuous random variable and p is the corresponding PDF. Gneiting et al. (2007) argue that a probabilistic forecast should be evaluated based on sharpness subject to calibration. Sharpness refers to how concentrated the forecast uncertainties are, and calibration

refers to the uniformity of the probability integral transform (PIT). Sharpness can be evaluated by the width of predictive intervals but is often measured by using proper scoring rules. The PIT is made by evaluating the probabilistic forecast at the true value after it is observed. For the CDF P, the PIT of the observed y_{T+h}, $P(y_{T+h})$ should be roughly uniformly distributed if the forecast is well calibrated. The PIT is assessed by making and visualizing its histogram (Gneiting et al., 2007; Hamill, 2001).

When in the situation where there are multiple models used to create probabilistic forecasts, we define the set of models as $\mathcal{M} = (M_1, ..., M_C)$ where M_c is one model and $c \in (1, 2, ..., C)$. We say $M_\mathcal{T}$ is the true model which generates $\{y_t\}$. The setting where the true model $M_\mathcal{T} \notin \mathcal{M}$ is referred to as the $\mathcal{M}-$open setting (Yao et al., 2018; Bernardo and Smith, 1994). In real world problems, it is often the case that forecasts belong to the $\mathcal{M}-$open setting. Combining the candidate forecasts in the $\mathcal{M}-$open setting is a powerful way to improve forecasts and has been done now for over half a century (Wang et al., 2023).

3.1 Linear pooling

The most common method for combining several probabilistic forecasts is known as *linear opinion pooling* and is a linear combination of multiple probabilistic forecasts. For C candidate forecasts of the event $\tilde{y}_{t+h}$, h time steps into the future from an observed time t with CDFs $P_c(\tilde{y}_{t+h})$ for $c \in (1, ..., C)$, and forecast weights w_c, a linear pool is defined as the mixture distribution in (1). To be a proper mixture distribution, the weights must be constrained such that they are each nonnegative and together sum to 1, or $w_c \geq 0$ and $\sum_{c=1}^{C} w_c = 1$.

$$\bar{P}_\omega(\tilde{y}_{t+h}) = \sum_{c=1}^{C} w_c P_c(\tilde{y}_{t+h}) \tag{1}$$

Given several individual forecasts, the goal of a forecaster in combining them via linear pooling is to select the vector of weights $\omega = (w_1, ..., w_C)$ in such a way that the ensemble forecast is optimized (Wang et al., 2023; Stone, 1961).

If the weights in (1) are optimized according to some score function, it is known as an *optimal prediction pool* (Geweke and Amisano, 2011; Hall and Mitchell, 2007). Yao et al. (2018) introduced a semi-Bayesian approach where a leave-one-out (LOO) estimator for combination weights consists of LOO posterior predictive distributions of candidate forecast models optimized over a proper scoring rule (Gneiting and Raftery, 2007). It is common to select weights which minimize the logarithmic score (LogS) (Li et al., 2023; Yao et al., 2018; Geweke and Amisano, 2011; Hall and Mitchell, 2007) or weights that minimize the continuous ranked probability score (CRPS) (Berrisch and Ziel, 2023; Li et al., 2019; Thorey et al., 2017), both of which are negatively oriented proper scoring rules. More discussion on proper scoring rules is given in section 4. Weight selection methods are often tailored to be problem specific, for example they may be formulated to be dynamic (Li et al., 2023; Billio et al., 2013) or to vary at different points on the PDF (Berrisch and Ziel, 2023).

Another way to select weights in (1) is through model averaging which Wang et al. (2023) emphasize is distinct from forecast combination. Forecast combination is combining forecasts with the aim of creating an optimal forecast, whereas model averaging determines model probabilities and provides a measure of model uncertainty. The most common model averaging approach is called Bayesian model averaging (BMA). In BMA a prior probability, $p(\mathcal{M}_c)$, is assigned to each model, and the posterior model probability is updated as data is observed according to Bayes theorem as

$$p_{BMA}(\mathcal{M}_c|\boldsymbol{y}_t) = \frac{p(\boldsymbol{y}_t|\mathcal{M}_c)p(\mathcal{M}_c)}{\sum_{j=1}^{C} p(\boldsymbol{y}_t|\mathcal{M}_j)p(\mathcal{M}_j)} \tag{2}$$

where $\boldsymbol{y}_t$ represents all data observed up to time t. To forecast a future event $\tilde{y}_{t+h}$, one may formulate a predictive mixture distribution of the form of (1) where the c^{th} forecast model is $P_c(y_{t+h}) = P(y_t|\mathcal{M}_c)$ and the weight given to the model is $w_c = p_{BMA}(\mathcal{M}_c|\boldsymbol{y}_t)$ (Raftery, 1996). In the model combination setting, a major drawback of BMA is that when the data is large enough, BMA assigns probability 1 to a single candidate model and probability 0 to all other models. This is demonstrated in Yao et al. (2018) and in a simulation study in section 5 herein. Adaptive variable selection (AVS) is an alternative model averaging method introduced by Lavine et al. (2021), and it falls within the Bayesian predictive synthesis framework (Tallman and West, 2024; McAlinn and West, 2019). In AVS, a prior distribution is assigned to each candidate model, and a Gibbs posterior model probability is computed for each candidate model where posterior updating is based on the exponential of some problem specific function, $S(\cdot)$, rather than on a model likelihood as in BMA. The AVS model probability is then

$$p_{AVS}(\mathcal{M}_c|\boldsymbol{y}_t) \propto \exp\{-\eta S(\mathcal{M}_c)\}p(\mathcal{M}_c) \tag{3}$$

where η is a model tuning parameter to be selected based on the needs of the problem. AVS allows for much more flexibility in updating posterior model averages, but as we show in a simulation in section 5 AVS may, like BMA, tend to prefer one candidate model over the other candidate models.

Among all weight selection methods for linear pooling, one of the most consistently well performing combinations is to give equal weight to all models in (1) so that $w_1 = w_2 = ... = w_C = 1/C$. This phenomenon has received much attention (see for example Frazier et al. (2023); Claeskens et al. (2016); Smith and Wallis (2009)), and fitting a simple average is obviously easier than using most weight optimization methods. Thus for another weight selection method to be worth implementing, it should outperform the EQW scheme. It has thus be recommended to use an EQW average as a baseline by which other methods should be compared (Li et al., 2023; Clemen, 1989).

4 Stacked Gibbs posterior

For weight selection of a linear pooled ensemble forecast, the stacked Gibbs posterior (SGP), introduced in this section, is used to estimate forecast combination weights ω, the vector of weights in the linear pool of (1). The SGP is defined in (4) and consists of three components, including a prior distribution on ω, $\pi(\omega)$, an empirical risk function $S_n(\cdot)$ based on a proper scoring rule, and a tuning parameter η. Each of these is discussed in further detail in sections 4.1 and 4.2.

$$\pi_n^{(\eta)}(\omega) \propto \exp\{-\eta n S_n(\omega)\}\pi(\omega) \tag{4}$$

The SGP is a probability distribution over weights which allows for inferential analysis of the weights. This is in contrast to the stacking done by Yao et al. (2018) where the weights are optimized as a point on the simplex. The SGP also differs from the posterior model averaging in (2) and (3). In the case of posterior model averaging, the posterior probability is assigned directly to candidate models rather than model weights. In the remainder of this section a general Gibbs posterior is defined in more detail following Martin and Syring (2022), proper scoring rules are defined following Gneiting and Raftery (2007) with specific emphasis on the CRPS, and a posterior consistency result is presented.

4.1 Gibbs posterior

The standard Bayesian posterior distribution for an unknown parameter θ given some data y is defined by Bayes theorem as $\pi(\theta|y) = cL(y|\theta)\pi(\theta)$ so that $\pi(\theta|y)$ is equal to the joint distribution of a statistical model $L(y|\theta)$ and a prior distribution $\pi(\theta)$ times a normalizing constant c. When the model $L(y|\theta)$ does not exist or the goal is inference on parameters which minimize some loss function rather than maximize a likelihood, the Gibbs posterior is a alternative to the standard posterior distribution which allows for inference on the minimizer of a selected function. The Gibbs posterior has been used in many applications, and has shown promising results often outperforming the Bayesian posterior in question specific parameter estimation, prediction, model selection, and model averaging (Martin and Syring, 2022; Loaiza-Maya et al., 2021; Lavine et al., 2021; Syring and Martin, 2017; Jiang and Tanner, 2008).

The setup for constructing a Gibbs posterior is as follows. We are given data, $y_1, ..., y_n \in \mathcal{Y}$ typically assumed to be independent and identically distributed (i.i.d.) from some inaccessible distribution $\mathcal{L}$. We do not assume a likelihood model for the data, either because one is not available or because inference via a likelihood is not useful for our problem. Instead we define a loss function parameterized by $\theta \in \Theta$, $l_\theta(y) : \Theta \times \mathcal{Y} \to \mathbb{R}$. Two common loss functions are the squared-error and the absolute-error losses. The goal for prediction is to minimize the loss function.

A risk function $R(\theta)$ is the expectation of the loss function $l_\theta(y)$ over the distribution $\mathcal{L}$, or $R(\theta) = El_\theta(Y)$. The primary interest is in the risk minimizer θ^* defined in (5) where $\in$ indicates that θ^* may not be unique.

$$\theta^* \in \underset{\theta\in\Theta}{\operatorname{argmin}} R(\theta) \tag{5}$$

Since $R(\theta)$ is inaccessible, an empirical risk function $R_n(\theta)$ based on the given data is formulated as in (6). Then $\hat{\theta}_n$, an estimate of θ^*, is estimated by the estimator in (7).

$$R_n(\theta) = \frac{1}{n}\sum_{i=1}^{n} l_\theta(y_i) \tag{6}$$

$$\hat{\theta}_n = \underset{\theta\in\Theta}{\operatorname{argmin}} R_n(\theta) \tag{7}$$

The estimator can be viewed as an M-estimator, the statistical properties of which have been extensively studied (Boos and Stefanski, 2013; Van der Vaart, 1998).

The purpose of the Gibbs posterior is to produce probabilistic uncertainty quantification for the unknown risk minimizer. In this setting, the Gibbs posterior is defined in (8) where R_n is the empirical risk from (6). The Gibbs posterior is proportional to the product of a prior distribution assigned to θ and the exponential of the empirical risk times a parameter $\eta > 0$ which Martin and Syring (2022) refer to as the *learning rate*.

$$\pi_n^{(\eta)}(\theta) \propto \exp\{-\eta n R_n(\theta)\}\pi(\theta) \tag{8}$$

The learning rate η is a tuning parameter used to control the balance between the prior distribution and the data. It should not be considered a model parameter to which one may assign a prior, but rather should be selected by a data driven tuning or some other problem specific selection. Often the learning rate is tuned so that posterior credible intervals have frequentist coverage probability, or it is selected via cross validation to optimize some criteria (Martin and Syring, 2022; Syring and Martin, 2017; Bissiri et al., 2016; Zhang, 2006). Importantly the asymptotic consistency of the Gibbs posterior is not dependent on η, so selecting η by most well reasoned methods should be appropriate (Martin and Syring, 2022; Loaiza-Maya et al., 2021).

For the SGP in (4), the more general $\theta \in \Theta$ is replaced with $\omega \in \{w_1, ..., w_C : 0 \leq w_c \leq 1, \sum_c w_c = 1\}$, and we consider empirical risk functions based two loss functions. The first of these functions is in (9), and is used for the simple case of i.i.d. data $y_1, ..., y_n$ and candidate models which are provided once and remain constant as more data becomes available.

$$S_n(\omega) = G_n(\omega) = \frac{1}{n}\sum_{i=1}^{n} G(\bar{P}_\omega, y_i). \tag{9}$$

Here $\bar{P}_\omega$ is the CDF of an ensemble forecast with C competing forecast models from (1). Each component P_c is fully specified, so that the only parameters to be estimated are model weights. The function $G(\cdot, \cdot)$ is a proper scoring rule (see section 4.2). A second empirical risk, used in a time dynamic setting is given in (10). Here we have an observed time series $y_1, ..., y_T$, and we want to construct an ensemble model for forecasting the future observation $\tilde{y}_{T+h}$.

$$S_T(\omega) = G_T(\omega) = \frac{1}{T}\sum_{t=1}^{T} \alpha^{T-t} G(\bar{P}_{t,\omega}, y_t) \tag{10}$$

In this dynamic setting, α^{T-t} is a discount factor (Koop and Korobilis, 2013; Raftery et al., 2010) used to lessen the influence of previous forecasts and making the performance of the most recent forecasts the most influential. In Lavine et al. (2021) the discount factor is fixed at $\alpha = 0.98$, and we use the same value herein. The linear pool $\bar{P}_{t,\omega}$ is made up of competing models which may or may not be time dependent. In the second simulation in section 5, the risk in (10) is used where competing models are fixed in time whereas in the analysis of flu forecasts in section 6, the forecasts vary in time.

4.2 Proper scoring rules and the continuous ranked probability score

In this subsection, we review the definition of a proper scoring rule and give the definitions for the LogS and the CRPS. Also included are additional properties of the CRPS. Following Gneiting and Raftery (2007), a scoring rule is a function $G : \mathcal{P} \times \mathcal{Y} \to [-\infty, \infty]$ where $\mathcal{P}$ is a convex class of probability measures on $(\mathcal{Y}, \mathcal{A})$, $\mathcal{A}$ being a σ-algebra on the set $\mathcal{Y}$. $G(P, y)$ is then a score on how well a predictive distribution, typically a probabilistic forecast, $P \in \mathcal{P}$ predicts a realized value y on the sample space. A negatively oriented scoring rule is a proper scoring rule if for two forecasts $P, Q \in \mathcal{P}$, the function $G(P, Q) = \int G(P, y)dQ(y)$ is such that $G(Q, Q) \leq G(P, Q)$. If the inequality is strict, then $G(\cdot, \cdot)$ is a *strictly proper scoring rule*.

Proper scoring rules are used not only to score probabilistic forecasts but as functions to optimize over when selecting weights for an optimal linear pool (Lavine et al., 2021; Li et al., 2019; Yao et al., 2018; Thorey et al., 2017; Geweke and Amisano, 2011). In practice, a number of scoring rules for continuous distributions are used to estimate parameters and evaluate forecasts (Gneiting and Katzfuss, 2014; Gneiting et al., 2007, for some examples), the LogS being perhaps the most widely used. The LogS is defined in (11) and evaluates a distribution by its PDF p.

$$\text{LogS}(p, y) = -\log(p(y)) \tag{11}$$

The CRPS is an increasingly popular scoring rule which, depending on the application, has certain advantages over the LogS. The CRPS is a global scoring rule in that it evaluates the whole distribution of a forecast whereas the LogS is a local scoring rule in that it evaluates the distribution only at a point (Gneiting and Katzfuss, 2014). When used for estimation, the CRPS is often a more robust estimator than the LogS and can lead to sharper and better calibrated forecasts under model misspecification (Gebetsberger et al., 2018; Gneiting et al., 2005). The CRPS is defined in (12) and evaluates a distribution through its CDF P.

$$\text{CRPS}(P, y) = \int_{-\infty}^{\infty} (P(x) - 1(y \leq x))^2 dx \tag{12}$$

For P belonging to certain distribution families, there may be a closed form solution to (12) (see table 1 of Zamo and Naveau (2018)). Particularly relevant for the work herein is the closed form solution of the CRPS when P is a mixture distribution with normal components. Li et al. (2019) provide the derivation for the normal mixture CRPS in (13).

$$CRPS(P, y) = \sum_{c=1}^{C} \sum_{c'=1}^{C} \alpha_{c,c'} w_c w_{c'} + \sum_{c=1}^{C} \beta_c w_c; \quad \text{where}$$

$$\begin{aligned} \alpha_{c,c'} =& -\frac{1}{\sqrt{2\pi}} \sqrt{\sigma_c^2 + \sigma_{c'}^2} \exp\left(\frac{(\mu_c - \mu_{c'})^2}{2(\sigma_c^2 + \sigma_{c'}^2)}\right) \\ & - \frac{\mu_c - \mu_{c'}}{2} \left[2\Phi\left(\frac{(\mu_c - \mu_{c'})}{\sqrt{\sigma_c^2 + \sigma_{c'}^2}}\right) - 1\right] \\ \beta_c =& \sqrt{\frac{2}{\pi}} \sigma_c \exp\left(-\frac{(\mu_c - y)^2)}{2\sigma_c^2}\right) \\ & + (\mu_c - y)\left[2\Phi\left(\frac{\mu_c - y}{\sigma_c}\right) - 1\right] \end{aligned} \tag{13}$$

The CRPS has some general forms other than (12) including (14) which was shown by Gneiting and Raftery (2007).

$$\text{CRPS}(P, y) = E|X - y| - \frac{1}{2} E|X - X'| \tag{14}$$

Here X and X' are independent copies of the same random variable with CDF P. When there is no closed form solution to the CRPS but where one can sample from the forecast distribution, the CRPS may still be estimated via Monte Carlo approximation by estimating the expectations in (14). We also note that if P takes the form of an ensemble forecast where all components are continuous distributions, then (14) can be further reduced to (15). A simple proof for this is in appendix 9.1.

$$CRPS(\bar{P}, y) = \sum_{c=1}^{C} w_c E|X_c - y| - \frac{1}{2} \sum_{c=1}^{C} \sum_{c'=1}^{C} w_c w_{c'} E|X_c - X_{c'}| \tag{15}$$

Li et al. (2019) showed this result for all mixture components being Gaussian, but it's simple to show that it holds when all components are continuous. This result allows for straightforward CRPS estimation for a continuous mixture distribution forecast when samples of the component distributions are available. To complete the definition of the SGP in (4), we let the function $G(\cdot, \cdot) = CRPS(\cdot, \cdot)$ and G_n take the form of either (9) or (10).

4.3 SGP consistency

In Martin and Syring (2022), parameter asymptotic consistency results of the Gibbs posterior similar to consistency results pertaining to the standard Bayesian posterior are established, and minimal conditions for consistency are given for i.i.d. data. The difference for the Gibbs posterior is that the posterior mass converges to the risk minimizer. The conditions for consistency follow the conditions for consistency in M-estimation (Van der Vaart, 1998). We define posterior consistency in 4.3 which is the same definition used by Martin and Syring (2022).

For a given distance $d : \Theta \times \Theta \to \mathbb{R}^+$, the Gibbs posterior distribution $\pi_n^{(\eta)}$ is consistent at θ^* if

$$\pi_n^{(\eta)}(\{\theta : d(\theta, \theta^*) > \epsilon\}) \to 0$$

in probability as $n \to \infty$

For the SGP in (4) with the empirical risk function in (9) the posterior consistency is shown for the distance function being the Euclidean distance, $d(\boldsymbol{a}, \boldsymbol{b}) = ||\boldsymbol{a} - \boldsymbol{b}||$, and the prior distribution on the weight vector ω to be the Dirichlet distribution $\pi(\omega) \sim Dirichlet(\lambda)$. Other distance functions and prior distributions may work, but importantly one of the conditions for consistency is that the prior distribution gives sufficient mass to the risk minimizer or that $\pi(\{\theta : R(\theta) - R(\theta^*) < \delta\}) > 0$ for all $\delta > 0$. The support of the Dirichlet distribution is over the simplex and thus meets this condition. For i.i.d. data $y_1, ..., y_n$ with probability law $\mathcal{L}$ where $E|Y| < \infty$ and where the empirical risk function is the function in (9), posterior consistency of the SGP proposed in (4) is affirmed by theorem 4.3.

If $\boldsymbol{\omega}^* \in \Omega = \{w1, ..., w_C : \sum_c w_c = 1, 0 \leq w_c\}$ uniquely minimizes $G(\omega) = E[CRPS(\bar{P}_\omega, Y)]$ and for the prior distribution π, $\pi(\{\boldsymbol{\omega} : G(\boldsymbol{\omega}) - G(\boldsymbol{\omega}^*) < \delta\}) > 0$ for all $\delta > 0$, then for any $\epsilon > 0$

$$\pi_n^{(\eta)}(\{\boldsymbol{\omega} : d(\boldsymbol{\omega}, \boldsymbol{\omega}^*) > \epsilon\}) \to 0$$

in $\mathcal{L}$-probability as $n \to \infty$

This manuscript does not include consistency results beyond the case of data being i.i.d. and having competing models which are fixed, or unaltered by additional data. Such is the setting of the first simulation study in the next section. One reason it may be difficult to develop theory in a dynamic setting is that the models which have the best predictive abilities may be changing in time making ω^* a moving target. However, this does not necessarily discount the SGP's ability to select model weights which produce well performing ensemble forecasts in dynamic settings. This is explored in second simulation study and in a real data example below.

5 Simulation studies

This section contains two simulation studies to demonstrate the SGP's ability to estimate optimal weights for model combination. The first study is done in an i.i.d. data setting with fixed competing predictive models. The second study includes the same predictive models, but the data is generated dynamically and weights are optimized for one step ahead forecasts.

5.1 I.I.D. data

The first study in this section shows how the SGP may be used to optimize weights among competing models in an $\mathcal{M}$-open scenario and is similar to the first simulation study in Yao et al. (2018). In this study there are six candidate models for predicting the data. Data is generated from the mixture of normals distribution $\nu \times N(3, 1) + (1 - \nu) \times N(6.5, 1)$ where $\nu = 0.65$. Because all simulated data is i.i.d., the empirical risk function from (5) is used, and the consistency results from theorem 4.3 hold. The six competing models are each normally distributed with variance 1, but each model has a unique mean parameter, the unique means being $\{0, 2, 4, 6, 8, 10\}$. Figure 2 shows in red the mixture distribution from which data is simulated as well as the candidate normal predictive models in grey.

The predictive performance of SGP is compared with that of AVS and BMA. For both AVS and SGP, the learning rate η needed to be selected, and we opted to select η with the aim of minimizing the CRPS for prediction. Figure 3 shows an example of how the CRPS varies for different values of η for the AVS and SGP. In the case of AVS, LOO cross validation over a grid of values for η was performed and the η which minimized the predictive CRPS was selected. For the SGP, the value of the CRPS continues to decrease as η increases. The improvement became less stark for higher values, so η was fixed at 15 so as to retain influence from the prior distribution.

Figure 4 shows examples from estimating model weights of constructing a linear pool from the candidate models for samples sizes of 10, 20, 50, 100, and 200 for BMA, AVS, and SGP weight selection methods. The top row shows how with a sample size as small as 10, BMA concentrates nearly all model probability into one model. AVS does better at spreading the model probability around, but as the sample size increases, it too will often favor one model over all others. SGP on the other hand tends to better capture the spread of the true distribution as the sample size increases.

Figure 5 shows boxplots of the CRPS and the LogS for the three methods of weight selection. The simulation was replicated 500 times for each sample size, and after selecting weights, the CRPS and LogS in these plots were calculated as the Monte Carlo mean scores for 1,000 additional draws from the true distribution after selecting weights. As the

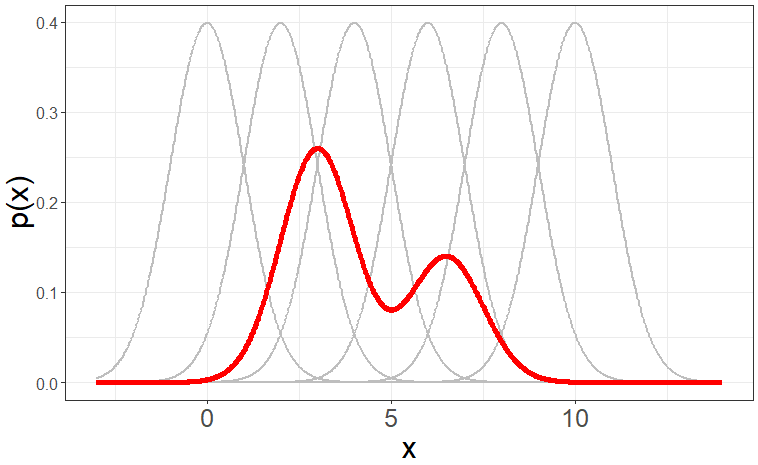

Figure 2: Mixture distribution density function from which data are simulated (red) and 6 candidate normal predictive densities (grey). The mixture distribution is $\nu \times N(3,1) + (1-\nu) \times N(6.5,1); w = 0.65)$. The candidate predictive distributions are each normal with variance 1 and means 0, 2, 4, 6, 8, and 10.

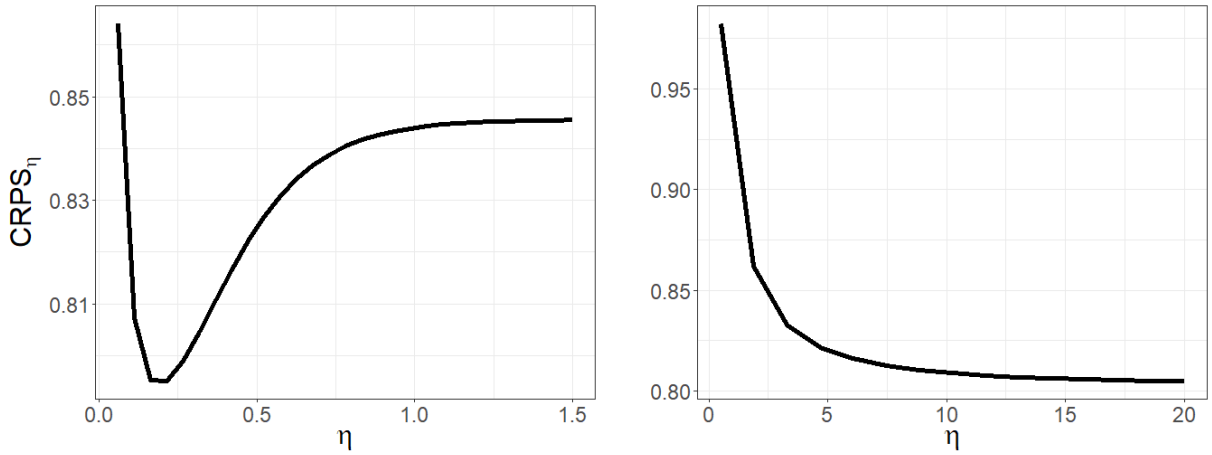

Figure 3: An example of CRPS values for different values of η after optimizing competing model weights for AVS (left) and for SGP (right).

sample size increases, the SGP clearly separates itself from the model averaging approaches. AVS appears to perform well for smaller sample sizes but worsens as the sample size increases, and it tends to favor one model of the others.

5.2 Dynamic data simulation forecast

The simulation study in this section has similarities to the previous study including that it utilizes the same fixed candidate predictive models and the data is simulated from a normal mixture distribution. Here, however, the mixture distribution from which the data is simulated is dynamic over time and the focus of making an ensemble model is to perform one step ahead forecasting. The data was simulated from the model in (16). The two component normals in the mixture distribution are the same as in the previous study, but here the weights for each model are updated at each time

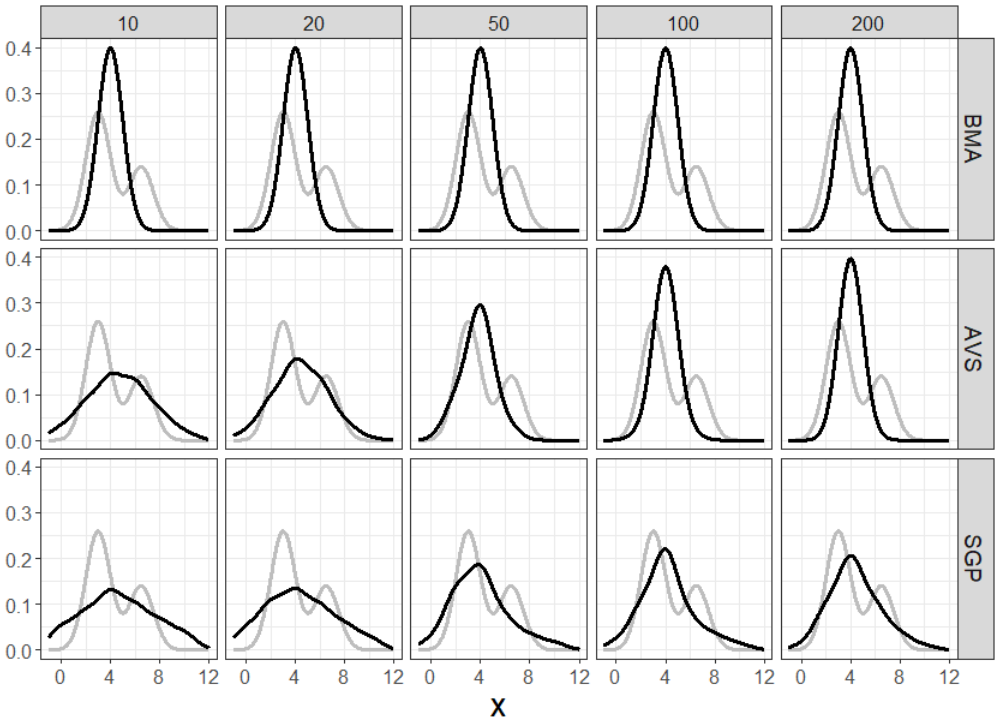

Figure 4: Examples of estimated densities after weighting of competing models faceted by weighting method and sample size. The estimated densities (black) overlay the true model density (grey).

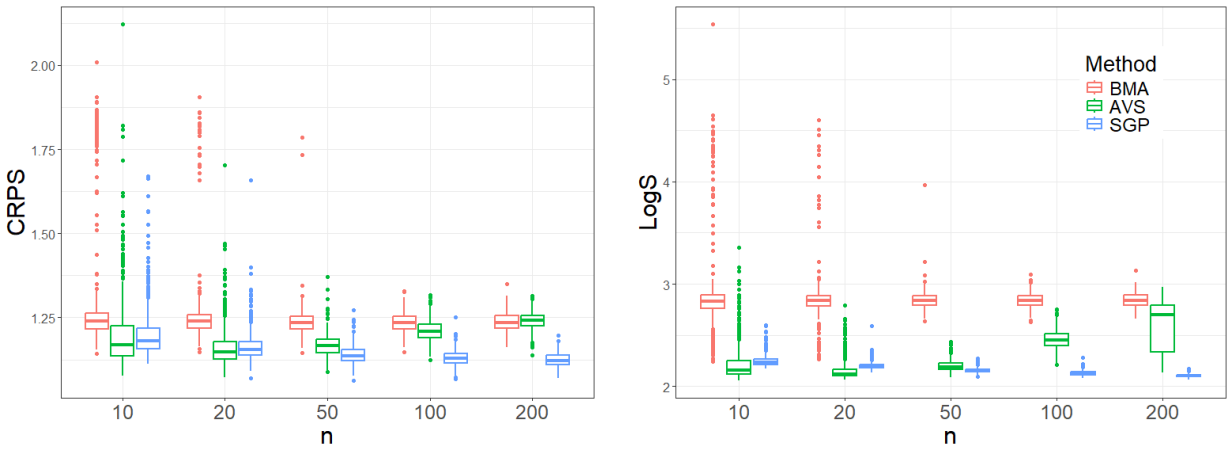

Figure 5: Boxplots of the mean of 1,000 Monte Carlo CRPSs (left) and LogSs (right) for 500 replicates. Plots are colored by weighting method.

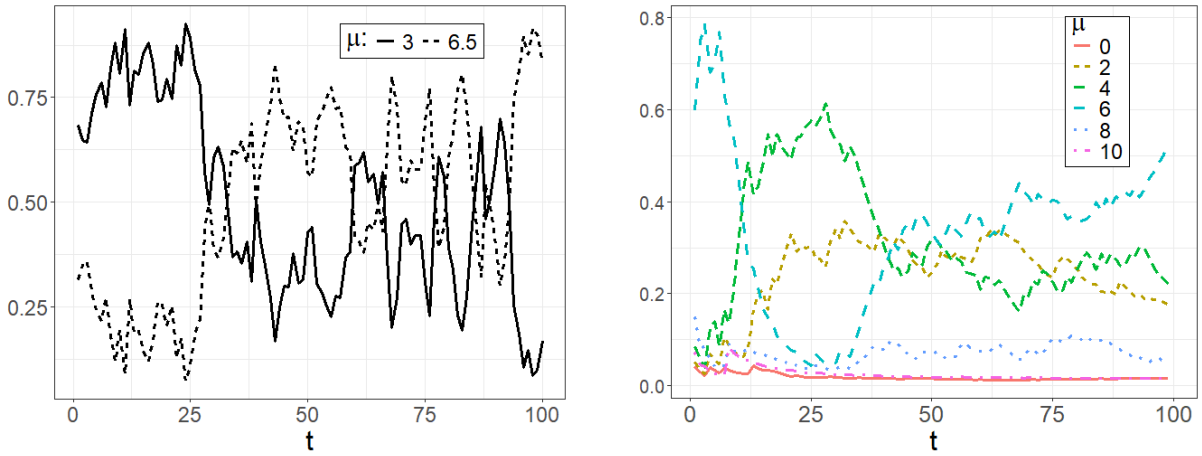

Figure 6: An example of simulated component weights for 100 time steps. The figure shows the simulated weights for the two components of the true distribution (left) and the weights optimized under SGP for the six competing component models (right).

point according to a random walk process.

$$\begin{aligned} Y_t &\sim \omega_{1,t} \times N(3,1) + \omega_{2,t} \times N(6.5,1) \\ \omega_{i,t} &= \frac{\exp(z_{i,t})}{\sum_a \exp(z_{i,t})}; \quad i \in \{1,2\} \\ \boldsymbol{z}_t &\sim N(\boldsymbol{z}_{t-1}, \sigma^2 I) \\ (\omega_{1,1}, \omega_{2,1}) &= (0.65, 0.35) \\ \sigma^2 &= 0.01 \end{aligned} \tag{16}$$

A single observation y_t is simulated at each time step $t \in \{1, ..., T = 100\}$, and at each step the ensemble weights are selected and an ensemble for forecasting $\tilde{y}_{t+1}$ is constructed. Figure 6 gives an example of the dynamic weights simulated from (16). The left plot shows the simulated weight values, and the right plot shows the SGP optimized weights of the competing forecasts for times 2 to 100.

The performance of the ensemble forecasts are assessed in terms of proper scoring rules as well as by the PIT, that is the uniformity of the histogram of observations evaluated by the probabilistic forecast. The PIT is evaluated both by looking at the PIT histogram and by measuring the distance between the PIT and the uniform distribution using the unit Wasserstein distance metric (UWD1) used in chapter 3 of Wadsworth and Niemi (2024). SGP is compared with AVS and BMA, and because this is a forecasting problem the performance of ensemble forecasts under EQW is also included. Figure 7 shows the PIT results for one step ahead forecasts for forecasts of times 2 to 100 over 500 replications of simulated data. The histogram plots clearly show the SGP PIT histogram is much nearer to uniformity than the PIT histograms for the other methods. The boxplots of the 500 UWD1 distances show smaller UWD1 values for the SGP PIT meaning the distances between the PIT and the standard uniform distribution are largely smaller for SGP, further showing superior calibration compared to the other methods.

Figure 8 shows how the CRPS and LogS for one step ahead forecasts behave over time. The lines represent the mean scores over the 500 simulation replicates. Again the SGP outperforms AVS, BMA, and EQW ensembles with the superior performance appearing more starkly for LogS than for CRPS. Unsurprisingly the EQW forecast scores remain relatively constant through time while the other methods improve as time goes on. Interestingly, the behavior of AVS shown in the i.i.d. case in the previous study to perform very well with small amounts of data but to worsen with larger data does not manifest in this study of dynamic forecasting.

The two studies in this section show the SGP's ability to construct accurate forecasts for cases of both i.i.d. and dynamic data. The SGP outperformed AVS, BMA, and EQW ensembles both in terms of proper scoring rules and in PIT uniformity. In the following section, the SGP ensemble construction is applied to forecasts from the CDC flu forecast competition with comparisons of performance against AVS, BMA, and EQW.

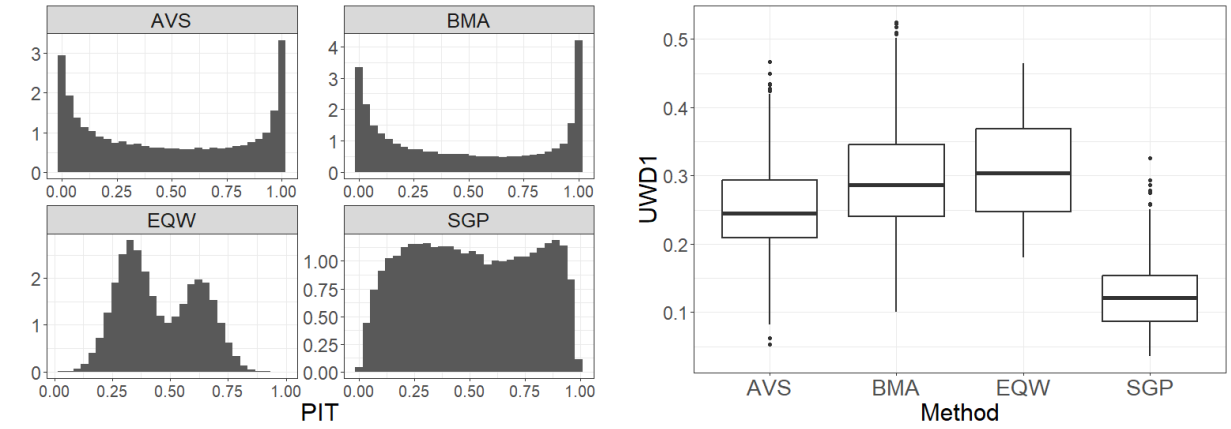

Figure 7: Plots for assessing calibration of one step ahead forecasts for the four weighting methods. The figure shows the PIT for all 99×500 simulated one-step-ahead forecasts for AVS, BMA, EQW, and SGP (left) and boxplots of the UWD1 distance between a standard uniform distribution and the empirical distribution of the PIT for the weighting schemes for 99 predictions (right).

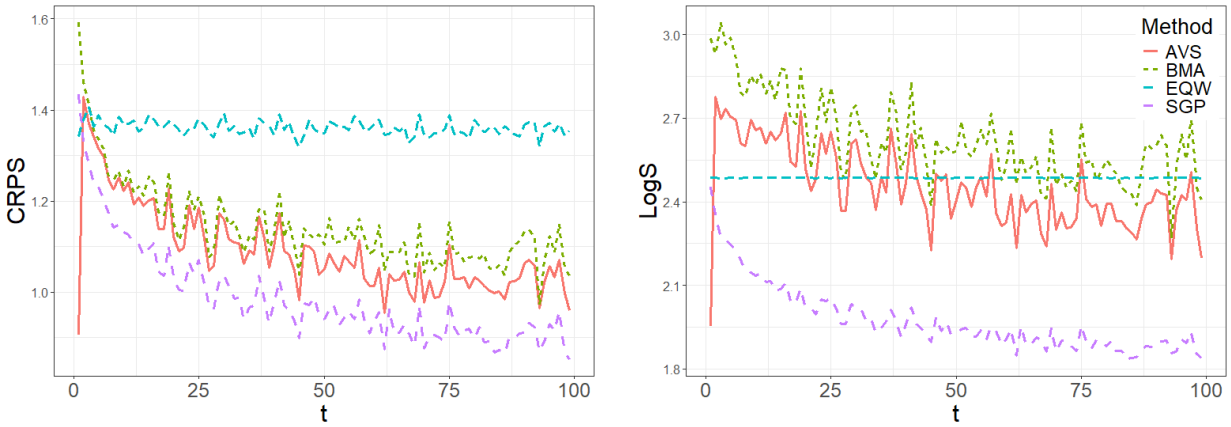

Figure 8: The mean CRPS (left) and LogS (right) over 500 replicates of one step ahead predictions at 99 time points colored by weighting methods.

6 Analysis on 2023-24 CDC flu forecast competition

Every flu season, beginning in 2013 with the exception of the 2020-21 flu season, the CDC has hosted a national flu forecasting competition, also known as FluSight. The competition begins in October and lasts around 30 weeks until May of the following year. There are generally a few dozen academic and industry research teams who each create separate probabilistic forecasts targeting certain flu events for future weeks. At the end of the season, forecast performance of all forecasts from each team are evaluated, and a winning team is announced (Mathis et al., 2024; Biggerstaff et al., 2016). During the 2022-23 and 2023-24 flu seasons, forecasts targeted 1, 2, 3, and 4-week ahead hospitalizations due to a laboratory confirmed flu infection for each of the 50 US states, District of Columbia, Puerto Rico, and the nation as a whole. Every week during the competition, a team could submit a forecast for each target, and after the hospitalization count for the target were observed the forecast was then scored and compared with all other forecasts. Besides reporting and scoring all forecasts individually, all submitted forecasts were combined into an ensemble forecast, which ensemble was published at CDC (2024) as the official CDC forecast and also scored along with the individual forecasts (Mathis et al., 2024; Wadsworth et al., 2023).

Teams were given complete freedom in how they developed their forecasts, and since the beginning of the competition there has been a large variety of methodologies used (Mathis et al., 2024; Github, 2024; McAndrew and Reich, 2021; Osthus and Moran, 2021; Osthus et al., 2019; Ulloa, 2019; Morgan et al., 2018; Farrow et al., 2017). To make scoring and ensemble modeling straightforward, each submitted forecast was required to follow a specified representation.

During the 2022-23 and 2023-24 seasons, the requested forecast representation was a set of estimated quantiles for each of 23 probabilities values provided by FluSight. For notation, we say each forecast team is given K probabilities $0 \leq p_1, p_2, ..., p_K \leq 1$. Then, the c^{th} competing team would submit quantiles -estimated however they please- $q^{(c)}_{r,w+h,1}, ..., q^{(c)}_{r,w+h,K}$ where

$$P(H_{r,w+h} < q^{(c)}_{r,w+h,k}) = p_k; \quad k \in (1, 2, ..., K)$$

for $H_{r,w+h}$ hospitalizations in location or region r, for time horizon $h \in \{1, 2, 3, 4\}$ weeks ahead, and where $H_{r,w}$ is the most recently observed hospitalization count at week w.

The forecasts were scored by the weighted interval score (WIS), defined in (17) (Bracher et al., 2021). The WIS is a proper scoring rule made for scoring quantile or interval representation forecasts. In the definition, P represents a probabilistic forecast with an interval representation including I predictive intervals with different nominal levels. The values $v_0 = 1/2$ and $v_i = \alpha_i/2$ are weights for each interval where α_i is the nominal level of the i^{th} interval, IS_α is the interval score (IS) a proper scoring rule for a single interval as defined in (18), and y^* is the true observation targeted by P.

$$WIS_{0,I}(P, y^*) = \frac{1}{I + 1/2} \times (v_0 \times |y^* - median| + \sum_{k=1}^{K} \{v_i \times IS_{\alpha_i}(P, y^*)\}) \tag{17}$$

$$IS_\alpha(l, r; y^*) = (r - l) + \frac{2}{\alpha}(l - y^*)1\{y^* < l\} + \frac{2}{\alpha}(y^* - r)1\{y^* > r\} \tag{18}$$

Like the CRPS the WIS is a global scoring rule, allowing for evaluation over the whole range of the forecast. In fact, as the number of forecast intervals I increases, the WIS becomes arbitrarily close to the CRPS (Bracher et al., 2021; Gneiting and Ranjan, 2011).

A drawback of the quantile representation and the WIS is that no information about the forecast distribution exists below the 1^{st} quantile or above the K^{th} quantile. Another drawback is that even with a large number of quantiles the information is less than would available from a PDF or CDF, making a more detailed forecast representation preferable (). In order to apply the SGP from (4) on the forecasts, the quantile Gaussian process quantile matching model from chapter 3 of Wadsworth and Niemi (2024) was fit to the quantiles to approximate the distribution from which the quantiles were estimated. This was done via Bayesian posterior MCMC sampling from a normal mixture distribution. The result was 50,000 posterior predictive draws for forecasts of each location, week, and horizon. Monte Carlo approximations of the expected values in (15) were calculated from the posterior predictive samples allowing for estimation of the CRPS for individual forecasts and ensemble forecasts. This allowed for using the SGP to be used for selecting weights and constructing an ensemble forecast.

The analysis of ensemble forecasts of flu hospitalizations below was limited to 1-week ahead forecasts. For each location, only the forecasts from teams who submitted forecasts for every week during the competition were included so as to avoid having to deal with missing forecasts. We expected that competing forecasts would vary in skill throughout the flu season and thus used the empirical risk function from (10) used in the dynamic setting. Unlike in the second simulation study in the previous section, the candidate forecasts were dynamic, changing every week as the individual forecast teams updated their forecasts and the flu season progressed. To complete the SGP, an uninformative Dirichlet prior with parameter $\mathbf{1}_C$ was assigned to ω. The SGP was fit via Hamiltonian Monte Carlo sampling using the `cmdstanr` R package (Stan Development Team, 2024; Gabry et al., 2022). To assess convergence of the fit, the SGP for US forecasts at the national level for the week of January 20, 2024 (week 15 of 29 of the flu season) was fit with 4 chains. Each chain consisted of 60,000 posterior draws, the first 10,000 draws being discarded as a burn-in. From this fit, among all posterior parameters, the largest $\hat{R}$ statistic (Vehtari et al., 2021) was 1.000055, giving no reason to be concerned about about MCMC convergence. The smallest effective sample size (Gelman et al., 2013) over all chains was 74,167. The SGPs for all other regions and weeks were fit using one chain of 60,000 draws, the first 10,000 being discarded as a burn-in. The result was 50,000 Monte Carlo vector draws for the weights $\{\omega^{(m)}\}_{m=1}^{M}$. For each c, the Monte Carlo mean of each weight $\bar{w}_{c,t} = M^{-1} w^{(m)}_{c,t}$ was calculated, and the mean weights were used in the ensemble forecast for hospitalizations at time $t + 1$. The learning rate η was selected by fitting the SGP for all forecasts up to time $t - 1$ and estimating model combination weights at all values of η on a pre-selected grid. The value of η which produced combination weights leading to the smallest CRPS of an ensemble model at time t was used in the SGP for time $t + 1$. Using this procedure, 1-week ahead ensemble forecasts were constructed for flu season weeks 2-29 for each location, noting that for the week 1 forecast ensemble, model weights were all selected to be equal. The CRPS was

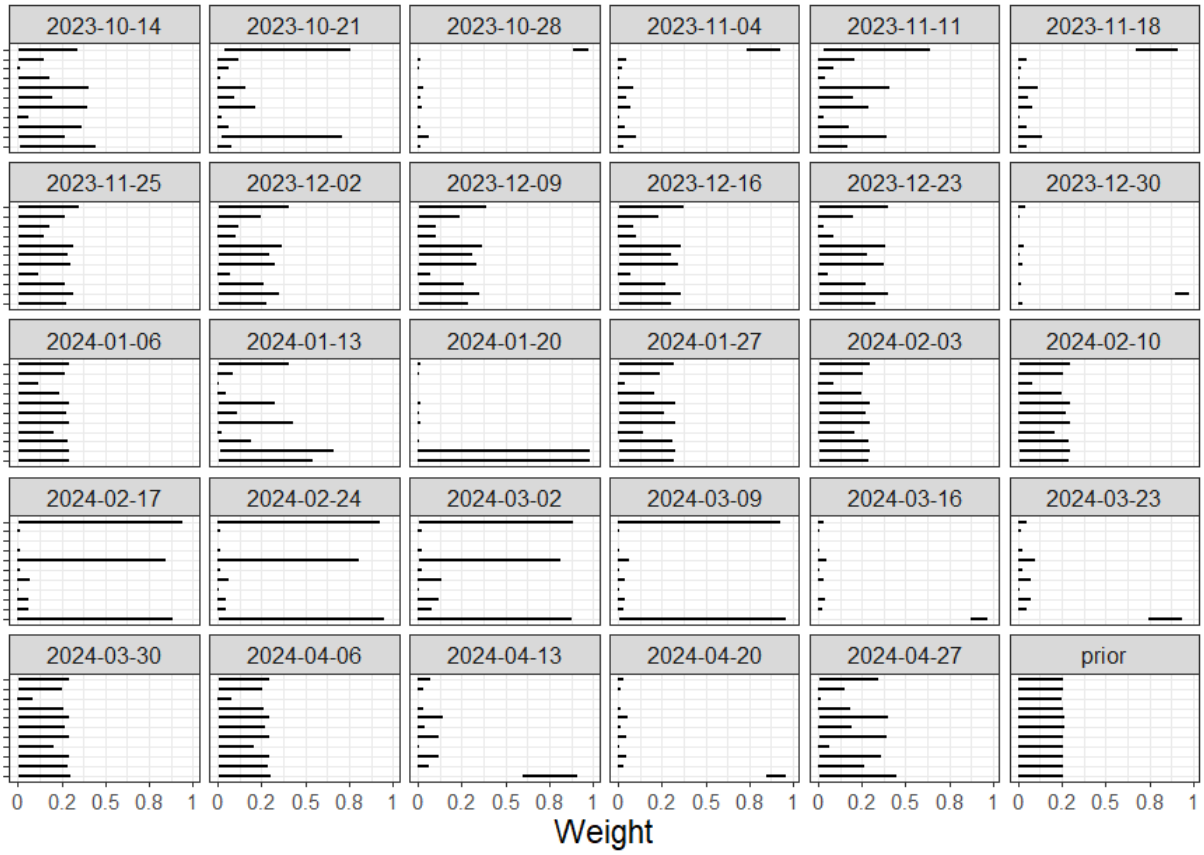

Figure 9: 90% credible intervals for forecast weights estimated via SGP for the 2023-24 CDC FluSite targetting flu hospitalizations at the national level of the United States. The intervals in the bottom right corner are those from the uninformative Dirichlet prior.

evaluated at the log forecast and log hospitalization. The mean CRPS for region r across the season was calculated as

$$\overline{CRPS}(\bar{P}, \boldsymbol{H}_r) = \frac{1}{W} \sum_{w=1}^{W-1} CRPS(\bar{P}_w, H_{r,w+1})$$

for each region. A similar mean for each week where the average was over all regions was also calculated.

Figure 9 shows 90% SGP credible intervals for the 11 included forecast models for flu hospitalizations at the US national level for all weeks during the 2023-24 flu season. The plots show how the SGP model probabilities change through the season where in some weeks one or two models are strongly favored over the others, whereas in other weeks models appear relatively equal in importance or display other patterns. Figure 10 shows the mean scores for SGP, AVS, BMA and EQW over all regions in one plot and over all weeks in the other. In all 53 locations, the overall CRPS for forecasts weighted by SGP is the lowest of the four methods compared, clearly showing superior skill in weight selection and forecasting for the hospitalization data and forecasts. Where the ensemble forecasts are evaluated by week, there is only one week where the SGP does not outperform the other methods.

Table 1 summarises the forecast performance for the four methods by showing in how many instances forecasts for each method were ranked 1, 2, 3, or 4 over regions or weeks, a smaller ranking meaning better performance. After SGP, AVS has the best performance followed by EQW and finally BMA.

This analysis demonstrates the effectiveness of selecting weights for a linear pooled ensemble forecast by optimizing the dynamic risk function in (10) via a Gibbs posterior distribution. Given that the forecasts used were posterior predictive samples from a model fit on quantile forecasts, it was key to be able to make a Monte Carlo approximation of the CRPS from (15). Of the methods compared, the SGP is arguably the most complicated, but in forecast performance it greatly outperformed the other methods.

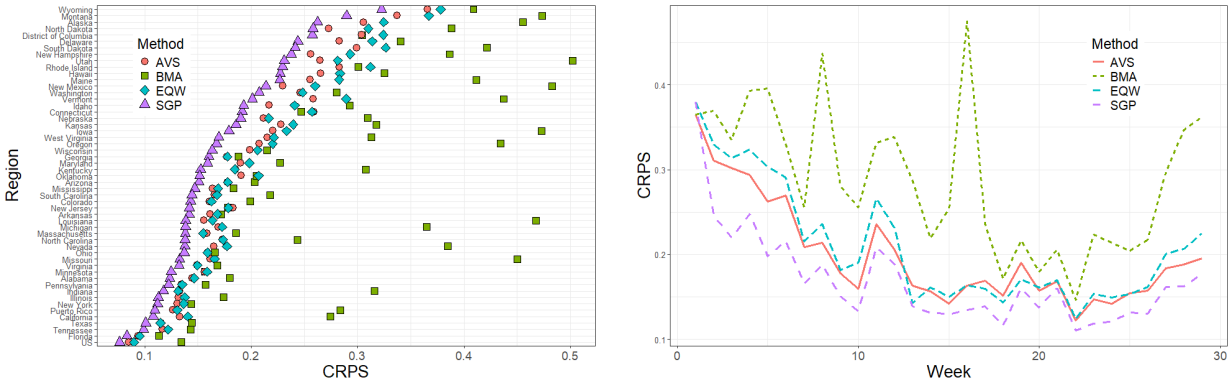

Figure 10: Mean over regions CRPS (left) for the four ensemble methods separated by shape and color, and mean over weeks CRPS (right) for the four methods

Table 1: Table showing the number of times ensemble forecasts for each method, SGP, AVS, BMA, and EQW, were ranked 1st, 2nd, 3rd, and 4th. There were 53 regions evaluated and 29 weeks, so each column under region should sum to 53 and each column under week to 29.

Ranking	By region CRPS				By week CRPS			
	AVS	BMA	EQW	SGP	AVS	BMA	EQW	SGP
1^{st}	0	0	0	53	1	0	0	28
2^{nd}	44	3	6	0	23	1	5	0
3^{rd}	7	3	43	0	5	0	24	0
4^{th}	2	47	4	0	0	28	0	1

7 Discussion

In this manuscript, we introduced a new method for optimizing model weights in a linear pooled forecast ensemble. The method, which is based on optimization of the CRPS using a Gibbs posterior distribution, provides uncertainty quantification for optimal linear pool weights and allows for a prior distribution to inform and influence weight selection. Included with the development of the SGP is a theorem for the asymptotic consistency of the Gibbs posterior. The SGP was used in two simulation studies and a data analysis on the 2023-24 FluSight forecasts to evaluate performance for model weighting, and it was compared with two posterior model averaging methods and an equally weighted linear pool ensemble. Performance of the methods was determined by forecast performance in terms of minimizing the CRPS and LogS, and in the second simulation study by analyzing the uniformity of the PIT. In both the simulations and the flu forecast analysis, the SGP showed superior performance when compared to the AVS, BMA, and EQW.

The promising results show that the SGP is a valuable contribution to the field of probabilistic forecast combination. The content of the SGP explored in this paper was primarily directed at forecasting a future event, but in a scenario where model uncertainty is of more concern, further exploration of the Gibbs posterior should be carried out. This may include further exploration into the selection or design of objective prior distributions (Giummolè et al., 2019). Other aspects of the SGP to be explored could be its use on scoring rules other than the CRPS or in nonlinear model combination methods. While the CRPS is a popular and useful measure of probabilistic forecast performance and one which we show is often simple to estimate in the continuous mixture distribution case, other scoring rules exist and different rules may be preferable for use in different applications. The WIS, for example, remains the preferred scoring method of the quantile forecasts for FluSight. An adaption of the SGP where the empirical risk is based on the WIS instead of the CRPS may be more fit for FluSight and other forecast hubs. Likewise, optimizing the linear pool forecast model in (1) may not be the ideal method for ensemble building. In fact the linear pool, while relatively simple to construct, has its drawbacks such as the difficulty of producing a completely uniform PIT (Gneiting and Ranjan, 2013). Applications of the SGP in the context of the so called generalized linear pool would be another direction of further development.

A limitation of this work is the asymptotic consistency theory of the SGP. Theorem 4.3 is limited to the case where the data is i.i.d. and where the models of the linear pool are fixed. An issue with posterior consistency in the dynamic setting is the tendency for the skill of probabilistic forecast models to vary in time making a single risk minimizer

unlikely to exist without strict assumptions. Another limitation, present in the flu forecast application is that forecasts are often missing. In the analysis in section 6, only models with 100% submission over the course of the flu season for a given state were included. Over the season, many forecast teams failed to submit forecasts during some weeks, and even if their forecasts were highly skilled these could not be included in the SGP as it is developed here. Modifying the SGP to account for missing forecasts so that they may factor into an ensemble when included, or by considering any of the above recommendations could improve the overall performance and/or lead to additional developments.

8 Acknowledgments

This work is partially supported by the National Science Foundation under Grant No. 2152117. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

References

Bassetti, F., Casarin, R., and Ravazzolo, F. (2018). Bayesian nonparametric calibration and combination of predictive distributions. *Journal of the American Statistical Association*, 113(522):675–685.

Bernardo, J. M. and Smith, A. F. (1994). *Bayesian Theory*. John Wiley & Sons.

Berrisch, J. and Ziel, F. (2023). CRPS learning. *Journal of Econometrics*, 237(2):105221.

Biggerstaff, M., Alper, D., Dredze, M., Fox, S., Fung, I. C.-H., Hickmann, K. S., Lewis, B., Rosenfeld, R., Shaman, J., Tsou, M.-H., et al. (2016). Results from the Centers for Disease Control and Prevention's predict the 2013–2014 Influenza Season Challenge. *BMC Infectious Diseases*, 16(1):1–10.

Billio, M., Casarin, R., Ravazzolo, F., and Van Dijk, H. K. (2013). Time-varying combinations of predictive densities using nonlinear filtering. *Journal of Econometrics*, 177(2):213–232.

Bissiri, P. G., Holmes, C. C., and Walker, S. G. (2016). A general framework for updating belief distributions. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 78(5):1103–1130.

Boos, D. D. and Stefanski, L. A. (2013). *Essential Statistical Inference: Theory and Methods*. Springer, New York.

Bracher, J., Ray, E. L., Gneiting, T., and Reich, N. G. (2021). Evaluating epidemic forecasts in an interval format. *PLOS Computational Biology*, 17(2):e1010592.

CDC (2024). Centers for Disease Control and Prevention FluSight: Flu Forecasting. https://www.cdc.gov/flu-forecasting/data-vis/index.html. Accessed: 2024-09-24.

Claeskens, G., Magnus, J. R., Vasnev, A. L., and Wang, W. (2016). The forecast combination puzzle: A simple theoretical explanation. *International Journal of Forecasting*, 32(3):754–762.

Clemen, R. T. (1989). Combining forecasts: A review and annotated bibliography. *International Journal of Forecasting*, 5(4):559–583.

Clyde, M. and Iversen, E. S. (2013). Bayesian model averaging in the M-open framework. In Damien, P., Dellaportas, P., Polson, N. G., and Stephens, D. A., editors, *Bayesian Theory and Applications*, page 483–498. Oxford University Press.

Collins, M. (2007). Ensembles and probabilities: a new era in the prediction of climate change. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 365(1857):1957–1970.

Cramer, E. Y., Huang, Y., Wang, Y., Ray, E. L., Cornell, M., Bracher, J., Brennen, A., Castro Rivadeneira, A. J., Gerding, A., House, K., Jayawardena, D., Kanji, A. H., Khandelwal, A., Le, K., Niemi, J., Stark, A., Shah, A., Wattanachit, N., Zorn, M. W., Reich, N. G., and Consortium, U. C.-. F. H. (2022). The United States COVID-19 forecast hub dataset. *Scientific Data*, 9(1):462.

Farrow, D. C., Brooks, L. C., Hyun, S., Tibshirani, R. J., Burke, D. S., and Rosenfeld, R. (2017). A human judgment approach to epidemiological forecasting. *PLOS Computational Biology*, 13(3):e1005248.

Frazier, D. T., Covey, R., Martin, G. M., and Poskitt, D. (2023). Solving the forecast combination puzzle. *arXiv preprint arXiv:2308.05263*.

Gabry, J., Češnovar, R., and Johnson, A. (2022). *cmdstanr: R Interface to 'CmdStan'*. https://mc-stan.org/cmdstanr/, https://discourse.mc-stan.org.

Gebetsberger, M., Messner, J. W., Mayr, G. J., and Zeileis, A. (2018). Estimation methods for nonhomogeneous regression models: Minimum continuous ranked probability score versus maximum likelihood. *Monthly Weather Review*, 146(12):4323–4338.

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., and Rubin, D. B. (2013). *Bayesian Data Analysis, Third Edition*. Chapman and Hall/CRC.

Geweke, J. and Amisano, G. (2011). Optimal prediction pools. *Journal of Econometrics*, 164(1):130–141.

Github (2024). FluSight-forecast-hub. https://github.com/cdcepi/FluSight-forecast-hub. Accessed: 2024-10-22.

Giummolè, F., Mameli, V., Ruli, E., and Ventura, L. (2019). Objective bayesian inference with proper scoring rules. *Test*, 28(3):728–755.

Gneiting, T., Balabdaoui, F., and Raftery, A. E. (2007). Probabilistic forecasts, calibration and sharpness. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 69(2):243–268.

Gneiting, T. and Katzfuss, M. (2014). Probabilistic forecasting. *Annual Review of Statistics and Its Application*, 1:125–151.

Gneiting, T. and Raftery, A. E. (2007). Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, 102(477):359–378.

Gneiting, T., Raftery, A. E., Westveld, A. H., and Goldman, T. (2005). Calibrated probabilistic forecasting using ensemble model output statistics and minimum CRPS estimation. *Monthly Weather Review*, 133(5):1098–1118.

Gneiting, T. and Ranjan, R. (2011). Comparing density forecasts using threshold-and quantile-weighted scoring rules. *Journal of Business & Economic Statistics*, 29(3):411–422.

Gneiting, T. and Ranjan, R. (2013). Combining predictive distributions. *Electronic Journal of Statistics*, 7:1747.

Gyamerah, S. A., Ngare, P., and Ikpe, D. (2020). Probabilistic forecasting of crop yields via quantile random forest and Epanechnikov kernel function. *Agricultural and Forest Meteorology*, 280:107808.

Hall, S. G. and Mitchell, J. (2007). Combining density forecasts. *International Journal of Forecasting*, 23(1):1–13.

Hamill, T. M. (2001). Interpretation of rank histograms for verifying ensemble forecasts. *Monthly Weather Review*, 129(3):550–560.

Hong, T., Pinson, P., Fan, S., Zareipour, H., Troccoli, A., and Hyndman, R. J. (2016). Probabilistic energy forecasting: Global energy forecasting competition 2014 and beyond. *International Journal of Forecasting*, 32(3):896–913.

Hong, T., Pinson, P., Wang, Y., Weron, R., Yang, D., and Zareipour, H. (2020). Energy forecasting: A review and outlook. *IEEE Open Access Journal of Power and Energy*, 7:376–388.

Hyndman, R. J. (2020). A brief history of forecasting competitions. *International Journal of Forecasting*, 36(1):7–14.

Jiang, W. and Tanner, M. A. (2008). Gibbs posterior for variable selection in high-dimensional classification and data mining. *The Annals of Statistics*, 36(5):2207–2231.

Joslyn, S. L. and LeClerc, J. E. (2012). Uncertainty forecasts improve weather-related decisions and attenuate the effects of forecast error. *Journal of Experimental Psychology: Applied*, 18(1):126.

Kapetanios, G., Mitchell, J., Price, S., and Fawcett, N. (2015). Generalised density forecast combinations. *Journal of Econometrics*, 188(1):150–165.

Koop, G. and Korobilis, D. (2013). Large time-varying parameter VARs. *Journal of Econometrics*, 177(2):185–198.

Lavine, I., Lindon, M., and West, M. (2021). Adaptive variable selection for sequential prediction in multivariate dynamic models. *Bayesian Analysis*, 16(4):1059–1083.

Li, L., Kang, Y., and Li, F. (2023). Bayesian forecast combination using time-varying features. *International Journal of Forecasting*, 39(3):1287–1302.

Li, T., Wang, Y., and Zhang, N. (2019). Combining probability density forecasts for power electrical loads. *IEEE Transactions on Smart Grid*, 11(2):1679–1690.

Lichtendahl Jr, K. C., Grushka-Cockayne, Y., and Winkler, R. L. (2013). Is it better to average probabilities or quantiles? *Management Science*, 59(7):1594–1611.

Loaiza-Maya, R., Martin, G. M., and Frazier, D. T. (2021). Focused Bayesian prediction. *Journal of Applied Econometrics*, 36(5):517–543.

Makridakis, S., Spiliotis, E., and Assimakopoulos, V. (2020). The M4 competition: 100,000 time series and 61 forecasting methods. *International Journal of Forecasting*, 36(1):54–74.

Martin, R. and Syring, N. (2022). Direct Gibbs posterior inference on risk minimizers: Construction, concentration, and calibration. In *Handbook of Statistics*, volume 47, pages 1–41. Elsevier.

Mathis, S. M., Webber, A. E., León, T. M., Murray, E. L., Sun, M., White, L. A., Brooks, L. C., Green, A., Hu, A. J., Rosenfeld, R., et al. (2024). Evaluation of FluSight influenza forecasting in the 2021–22 and 2022–23 seasons with a new target laboratory-confirmed influenza hospitalizations. *Nature Communications*, 15(1):6289.

McAlinn, K. and West, M. (2019). Dynamic Bayesian predictive synthesis in time series forecasting. *Journal of Econometrics*, 210(1):155–169.

McAndrew, T. and Reich, N. G. (2021). Adaptively stacking ensembles for influenza forecasting. *Statistics in Medicine*, 40(30):6931–6952.

Morgan, J. J., Wilson, O. C., and Menon, P. G. (2018). The wisdom of crowds approach to influenza-rate forecasting. In *ASME international mechanical engineering congress and exposition*, volume 52026, page V003T04A048. American Society of Mechanical Engineers.

Newey, W. K. and McFadden, D. (1994). Large sample estimation and hypothesis testing. *Handbook of Econometrics*, 4:2111–2245.

Osthus, D., Gattiker, J., Priedhorsky, R., and Del Valle, S. Y. (2019). Dynamic Bayesian influenza forecasting in the United States with hierarchical discrepancy (with discussion). *Bayesian Analysis*, 14(1):261–312.

Osthus, D. and Moran, K. R. (2021). Multiscale influenza forecasting. *Nature Communications*, 12(1):2991.

Palmer, T. N. (2002). The economic value of ensemble forecasts as a tool for risk assessment: From days to decades. *Quarterly Journal of the Royal Meteorological Society: A journal of the atmospheric sciences, applied meteorology and physical oceanography*, 128(581):747–774.

Raftery, A. E. (1996). Hypothesis testing and model selection. In Gilks, W., Richardson, S., and Spiegelhalter, D., editors, *Markov Chain Monte Carlo in Practice: Interdisciplinary Statistics*, pages 163–187. Chapman and Hall.

Raftery, A. E., Kárnỳ, M., and Ettler, P. (2010). Online prediction under model uncertainty via dynamic model averaging: Application to a cold rolling mill. *Technometrics*, 52(1):52–66.

Ramos, M. H., Van Andel, S. J., and Pappenberger, F. (2013). Do probabilistic forecasts lead to better decisions? *Hydrology and Earth System Sciences*, 17(6):2219–2232.

Ranjan, R. and Gneiting, T. (2010). Combining probability forecasts. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 72(1):71–91.

Reich, N. G., Brooks, L. C., Fox, S. J., Kandula, S., McGowan, C. J., Moore, E., Osthus, D., Ray, E. L., Tushar, A., Yamana, T. K., et al. (2019a). A collaborative multiyear, multimodel assessment of seasonal influenza forecasting in the united states. *Proceedings of the National Academy of Sciences*, 116(8):3146–3154.

Reich, N. G., McGowan, C. J., Yamana, T. K., Tushar, A., Ray, E. L., Osthus, D., Kandula, S., Brooks, L. C., Crawford-Crudell, W., Gibson, G. C., et al. (2019b). A collaborative multi-model ensemble for real-time influenza season forecasting in the us. *bioRxiv*, page 566604.

Smith, J. and Wallis, K. F. (2009). A simple explanation of the forecast combination puzzle. *Oxford Bulletin of Economics and Statistics*, 71(3):331–355.

Stan Development Team (2024). Stan modeling language users guide and reference manual, 2.34. https://mc-stan.org. Accessed: 2024-10-22.

Stone, M. (1961). The opinion pool. *The Annals of Mathematical Statistics*, 32:1339–1342.

Syring, N. and Martin, R. (2017). Gibbs posterior inference on the minimum clinically important difference. *Journal of Statistical Planning and Inference*, 187:67–77.

Tallman, E. and West, M. (2024). Bayesian predictive decision synthesis. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 86(2):340–363.

Thorey, J., Mallet, V., and Baudin, P. (2017). Online learning with the continuous ranked probability score for ensemble forecasting. *Quarterly Journal of the Royal Meteorological Society*, 143(702):521–529.

Ulloa, N. (2019). *Bayesian hierarchical modeling for disease outbreaks*. PhD thesis, Iowa State University Department of Statistics.

Van der Vaart, A. W. (1998). *Asymptotic Statistics*. Cambridge University Press, Cambridge, United Kingdom.

Vehtari, A., Gelman, A., Simpson, D., Carpenter, B., and Bürkner, P.-C. (2021). Rank-normalization, folding, and localization: An improved $\hat{R}$ for assessing convergence of MCMC (with discussion). *Bayesian Analysis*, 16(2):667–718.

Wadsworth, S. and Niemi, J. (2024). *Advances in Bayesian methodology for collaborative probabilistic forecast hubs with applications in disease outbreak forecasting*. PhD thesis, Iowa State University Department of Statistics.

Wadsworth, S., Niemi, J., and Reich, N. (2023). Mixture distributions for probabilistic forecasts of disease outbreaks. *arXiv preprint arXiv:2310.11939*.

Wang, X., Hyndman, R. J., Li, F., and Kang, Y. (2023). Forecast combinations: An over 50-year review. *International Journal of Forecasting*, 39(4):1518–1547.

Yao, Y., Vehtari, A., Simpson, D., and Gelman, A. (2018). Using stacking to average Bayesian predictive distributions (with discussion). *Bayesian Analysis*, 13(3):917–1003.

Zamo, M. and Naveau, P. (2018). Estimation of the continuous ranked probability score with limited information and applications to ensemble weather forecasts. *Mathematical Geosciences*, 50(2):209–234.

Zhang, T. (2006). Information-theoretic upper and lower bounds for statistical estimation. *IEEE Transactions on Information Theory*, 52(4):1307–1321.

9 Appendix: Proofs of equation (15) and theorem 4.3

9.1 Proof of equation (15)

Proof. The random variables X and X' in (14) are independent copies of the same random variable with mixture distribution PDF

$$\bar{p}(x) = \sum_{c=1}^{C} w_c p_c(x)$$

where each $p_c(x)$ is a PDF of a continuous random variable. The random variables X_c and X_c' are independent copies of a random variable with PDF $p_c(x)$. Then the second expectation in (14) is

$$\begin{aligned} E|X - X'| &= \int_{\mathcal{X}} \int_{\mathcal{X}} |x - x'| p(x) p(x') dx dx' \\ &= \int_{\mathcal{X}} \int_{\mathcal{X}} |x - x'| \sum_{i=1}^{C} w_c p_c(x) \sum_{d=1}^{C} w_d p_d(x') dx dx' \\ &= \sum_{c=1}^{C} \sum_{d=1}^{C} w_c w_d \int_{\mathcal{X}} \int_{\mathcal{X}} |x - x'| p_c(x) p_d(x') dx dx' \\ &= \sum_{c=1}^{C} \sum_{d=1}^{C} w_c w_d E|X_c - X_d'| \end{aligned}$$

Similary for the first expectation

$$E|X - y| = \sum_{c=1}^{C} w_c E|X_c - y|$$

and the result in (15) follows. □

9.2 Proof of theorem 4.3

Proof. Martin and Syring (2022) identify three sufficient conditions needed to prove Gibbs posterior consistency. The first is that the prior distribution give sufficient mass to the risk minimizer ω^*. This is met with the selected Dirichlet prior. The other two conditions are a uniform law of large numbers and a separation or identifiability condition. Respectively these two conditions are (19) and (20).

$$\sup_{\omega \in \Omega} |G_n(\omega) - G(\omega)| \to 0 \text{ in } \mathcal{L} - \text{probability} \tag{19}$$

$$\inf_{\omega : d(\omega, \omega^*) > \delta} \{G(\omega) - G(\omega^*)\} > 0 \text{ for any } \delta > 0 \tag{20}$$

We first show (19). By lemma 2.4 in Newey and McFadden (1994), if

> [i)]the data are iid Ω is compact $CRPS(\bar{P}_\omega, y)$ is continuous at each $\omega \in \Omega$ with probability 1 there is a function $m(y)$ such that $|CRPS(\bar{P}_\omega, y)| \leq m(y)$ and $E[m(Y)] < \infty$

then $G(\omega)$ is continuous and (19) follows. Conditions i) and ii) are met by assumption and iii) is met by the continuity of the CRPS. To show iv) we let

$$\begin{aligned}\text{CRPS}(\bar{P}_\omega, y) &= E|X - y| - \frac{1}{2}E|X - X'| \\ &\leq E|X - y| \\ &\leq E\left[|X| + |y|\right] \\ &= E|X| + |y| := m(y)\end{aligned}$$

Then by the assumption that $E|Y| < \infty$, $E[m(Y)] < \infty$ meeting condition iv) and showing (19).

Now to show (20). Consider the open set $B_{\delta'} = \{\omega : d(\omega, \omega^*) < \delta'\}$ and the closed set $\overline{B}_\delta = \{\omega : d(\omega, \omega^*) \leq \delta\}$ where $0 < \delta' < \delta$. Since $B_{\delta'}$ is open, the set $\Omega \setminus B_{\delta'} \subset \Omega$ is a closed subset of a compact set and is thus compact. Because $G(\cdot)$ is continuous, the set $G(\Omega \setminus B_{\delta'})$ is also compact and thus contains its infimum. Now $\Omega \setminus \overline{B}_\delta \subset \Omega \setminus B_{\delta'}$, so $G(\Omega \setminus \overline{B}_{\delta'}) \subset G(\Omega \setminus B_{\delta'})$. Thus

$$G(\omega^*) < \inf_{\Omega \setminus B_{\delta'}} G(\omega) \leq \inf_{\Omega \setminus \overline{B}_\delta} G(\omega)$$

and (20) holds. □