Advances in Bayesian methodology for collaborative probabilistic forecast hubs with applications in disease outbreak forecasting

by

Spencer Gordon Wadsworth

A dissertation submitted to the graduate faculty

in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

Major: Statistics

Program of Study Committee:
Jarad Niemi, Major Professor
Lynna Chu
Karin Dorman
Vivekananda Roy
Chong Wang

The student author, whose presentation of the scholarship herein was approved by the program of study committee, is solely responsible for the content of this dissertation. The Graduate College will ensure this dissertation is globally accessible and will not permit alterations after a degree is conferred.

Iowa State University

Ames, Iowa

2024

DEDICATION

This dissertation is dedicated to two of my older sisters. Heidi has always given the best advice and feedback on my education decisions. She was also the first person to recommend that I earn a PhD and maybe the only person I would have listened to. Before settling on what to study in college, Juline repeatedly insisted I take a statistics class, and despite my misgivings she prevailed upon me to enroll in an introductory course.

TABLE OF CONTENTS

Page

LIST OF TABLES

Page

LIST OF FIGURES

Page

ACKNOWLEDGMENTS

I have to first acknowledge Dr. Jarad Niemi for putting up with me for the last few years. He's able to get the best work out of me and has been totally supportive of my interests and goals. His questions, challenges, and advice on my work made our weekly meetings one of the highlights of my time. I'm grateful to Drs. Lynna Chu, Karin Dorman, Vivek Roy, Chong Wang, and Kori Khan for serving on my committee. Other professors who've had an outsized impact on me are Drs. Dan Nettleton, Kris De Brabanter, Farzad Sabzikar, and Kevin Kasper. I'm fortunate some of them were willing to indulge me in my childish ribbing.

Friends from Snedecor Hall made the time far more enjoyable. Studying with Federico Veneri and Eryn Blagg kept me sane leading up to the prelim exam in the middle of a pandemic. Regular pod visits from Paul Morris helped the time go faster and make the mood lighter. Sam Fox doubled as a free personal trainer and a chum on whom to try my stupidest jokes (his jokes weren't much better). Other meaningful friendships include Mattie, Grant, Ben, Zirou, Ricardo, and Josh. Friends from the Saylorville YSA branch who made me think I could live in Iowa forever were Anthony and Amber Borrino, Bennie Bresee, Molly Skouson, Kathryn Blair, and Avery Hill.

When working on a PhD a thousand miles from home, it really pays off to have the best Mom, Dad, and 11 siblings. There was always someone to call when I wanted to complain, vent, brag, or roast.

Finally, I have to thank D. Todd Christofferson. He doesn't know me, but the first time I felt I wanted to earn a PhD was 12 years while listening to him speak (Christofferson, 2012).

0.1 References

Christofferson, D. T. (November, 2012). Brethren, We Have Work to Do. *Ensign*, pages 47–50.

ABSTRACT

Probabilistic forecasting competitions and collaborative forecast hubs are becoming increasingly popular in many fields. One example is the recent surge in disease outbreak forecast hubs that have come about partially as a result of the COVID-19 pandemic. The increase in collaborative forecasting has led to many advancements in statistical and machine learning modeling and developments in how to best administer collaborative hubs, including in how multiple probabilistic forecasts should be represented, scored, and combined into an ensemble forecast. In this dissertation, we contribute to the ongoing development of collaborative forecast hubs by introducing new methodology for forecasting, representing forecasts, and combining forecasts into an ensemble forecast. Much of the work was motivated by the United States Centers for Disease Control and Prevention (CDC) flu forecasting competition.

In chapter 2, we introduce a two component Bayesian forecast modeling framework for flu hospitalizations. This modeling framework is able to incorporate forecasts of influenza-like illness (ILI) data allowing for the exploitation of several years of ILI data and successful nonlinear ILI forecast models for use in hospitalization forecasts. ILI is modeled as a Bayesian nonlinear hierarchical function, and flu hospitalizations are modeled as a linear model with ILI as either a linear or quadratic predictor variable. In a simulation study, we compare forecasts where the ILI model utilizes a compartmental disease outbreak mathematical model to an ILI model which uses smooth nonlinear model. Variations of the model are also compared where seasonal systematic discrepancy is or is not accounted for. Additional comparison is made for variations to the linear model of hospitalizations with model forecasts applied to the 2023-24 flu hospitalization data.

In chapter 3 we introduce a new model, the quantile Gaussian process (QGP), for quantile matching (QM) or estimating the continuous distribution of a probabilistic forecast given a set of estimated quantiles. The QGP is based on a well established central limit theorem for sample

quantiles, and unlike many QM methods, the QGP can provide accurate uncertainty quantification for making inference on estimated quantiles and perform accurate QM. For cases where the true distribution family from which estimated quantiles is unknown, we select a normal mixture distribution with four components for QM because of its ability to approximate many distributional shapes. The QM abilities of the QGP are assessed in several simulation studies and compared with other QM methods, and the QGP is applied to quantile forecasts from the 2023-24 CDC flu forecasting competition.

In chapter 4, we introduce the stacked Gibbs posterior (SGP) as a method for optimally selecting weights for a mixture distribution ensemble of several component forecasts. The SGP selects ensemble model weights by optimizing a proper scoring rule, the continuous ranked probability score, and as the SGP is a Gibbs posterior, it also provides a posterior probability distribution of the weights and allows for prior information to influence individual model importance. An asymptotic consistency result for independent and identically distributed data is given, and the SGP is analyzed in two simulation studies and for forecasts of the 2023-24 CDC flu forecast competition. In these studies, the SGP outperforms model averaging and equal average ensemble forecasts by producing superior forecasts.

CHAPTER 1. GENERAL INTRODUCTION

"Forecasts may tell you a great deal about the forecaster; they tell you nothing about the future."

-Warren Buffett (1981)

Predicting future events, or forecasting, with the goals of minimizing risk and maximizing reward is central to much of decision making for issues both public and private. The scientific literature on forecasting is immense, and there are many applications for it (Petropoulos et al., 2022; Gneiting and Katzfuss, 2014; De Gooijer and Hyndman, 2006). Forecasts will rarely be exact, and thus in many fields it is becoming increasingly important that forecasts be probabilistic in nature, or that values of uncertainty be attached to different possible outcomes (Gneiting and Katzfuss, 2014). It has also been shown that having forecasts which are probabilistic as opposed to point forecasts improves decision making (Ramos et al., 2013; Joslyn and LeClerc, 2012).

When producing a probabilistic forecast, one should keep in mind the needs for the problems they want to solve. Considerations such as what specific methodologies are appropriate and how the forecast uncertainty should be represented should be addressed. Some common representations for probabilistic forecasts include probability density/mass functions (PDF) or continuous distribution functions (CDF) (Hall and Mitchell, 2007; Gneiting et al., 2007), distribution samples such as Markov chain Monte Carlo samples (Krüger et al., 2021), discretized bin distributions (McGowan et al., 2019), and quantile or interval forecasts (Taylor, 2021; Bracher et al., 2021). Which of these representations to use may depend on the methodology for producing the forecast, the purpose of the forecast, and the how or by whom the forecast will be interpreted. For instance, some methodologies such as parametric statistical models naturally produce predictive distributions which include a PDF or CDF, whereas some machine learning methods may only produce point predictions including quantile predictions via quantile regression (Gneiting et al., 2023; Koenker, 2017; Gneiting, 2011; Koenker and Bassett Jr, 1978). When using

such a machine learning method, quantile forecasts may be the most reasonable representation for forecasts. Likewise, in collaborative forecasting projects where multiple participants submit their own forecast of the same event, it may be necessary to choose a representation which simplifies the computing, comparison, and combining of forecasts (Wadsworth et al., 2023).

Because for a given event it is unlikely that one can produce the "correct" forecast model, a common forecast approach is to produce multiple forecasts and either select the best forecast(s) or combine the forecasts into a single ensemble forecast. This approach has led to the creation of numerous forecasting competitions in fields such as finance/economics, global energy, and the influenza outbreak (Makridakis et al., 2020; Hyndman, 2020; Biggerstaff et al., 2016; Hong et al., 2016), and to the development of many methods for combining forecasts (Wang et al., 2023; Yao et al., 2018; Gneiting and Ranjan, 2013). In multi-forecast settings, there is also a need to compare forecast performance among the various methods. It has become the standard to use proper scoring rules for evaluating and comparing forecasts (Bracher et al., 2021; Gneiting and Raftery, 2007). Proper scoring rules are defined in such a way as to keep forecasters honest and only submit the forecasts which they believe best capture reality (Gneiting and Raftery, 2007). Besides use in forecast evaluation, proper scoring rules also make ideal functions for optimizing forecast combinations (Yao et al., 2018; Geweke and Amisano, 2011).

One area where collaborative forecasting has seen a major rise in the last decade is disease outbreak forecasting. The United States Centers for Disease Control and Prevention (CDC) has hosted a number of disease outbreak forecast competitions and hubs for several disease outbreaks with examples including influenza (McGowan et al., 2019; Mathis et al., 2024), the Dengue fever (Johansson et al., 2019), and the West Nile virus (Holcomb et al., 2023). The COVID-19 pandemic which began in 2020 spurred a tremendous forecasting effort which led to several collaborative forecast hubs being created in the US and Europe as well as the development of The Consortium of Infectious Disease Modeling Hubs and the creation of the hubverse, which provides tools for creating new forecast hubs (The Consortium of Infectious Disease Modeling Hubs, 2024).

One collaborative forecast project, from which much of the work in this dissertation is motivated, is the CDC flu forecasting competition, also known as FluSight.

FluSight began with 2013-14 flu season as a competition in which several teams around the US submitted weekly forecasts of the flu outbreak for certain targets. With the exception of the 2020-21 season, the CDC has hosted FluSight every flu season since 2013. FluSight has led to a number of improvements in collaborative disease outbreak modeling, and subsequent collaborative forecast projects were based on it (Mathis et al., 2024; Bracher et al., 2021; McGowan et al., 2019; Biggerstaff et al., 2016). Through the duration of a flu season, forecasts from the various teams are scored using appropriate proper scoring rules, and each week forecasts are combined into an ensemble forecast which is published as the official CDC forecast (Mathis et al., 2024). The interruption of FluSight by the COVID-19 pandemic and the creation of the COVID-19 Forecast Fub led to the cancellation of FluSight for a season and prompted a number of changes to FluSight when it was restarted in the 2021-22 season. For instance, the target data used for the first seven seasons of FluSight was replaced with other flu related data, and the forecast representation was changed.

The target data used for the first seven seasons of the competition was the influenza-like illness (ILI) data. ILI is a measure of the percentage of patients who visit a healthcare provider and display flu-like symptoms (CDC, 2024). ILI is only a proxy for the flu since it is not based on laboratory confirmed cases. Thus ILI may also be influenced by cases of COVID-19 and other respiratory illnesses. For the 2021-22 season, the change was made from targeting ILI to the number of hospitalizations of patients with a laboratory confirmed flu infection (Mathis et al., 2024; HealthData.gov, 2024). Another major change to FluSight which was adopted from the COVID-19 Forecast Hub was the use of a set of quantiles as the forecast representation. Originally the forecast representation was a set of discretized bins, but for practical reasons the change to quantiles was made (Mathis et al., 2024; Cramer et al., 2022a,b; McGowan et al., 2019; Biggerstaff et al., 2016).

In this dissertation, we contribute to the work already done in the disease outbreak and collaborative forecast hub setting by introducing statistical methodology for producing forecasts, estimating continuous distributions given quantile forecasts, and producing model combination methodology. This work addresses many of the aspects of forecasting in FluSight discussed above.

In chapter 2 we develop a forecast modeling framework for forecasting flu hospitalizations which incorporates both the ILI data and the hospitalization data. The modeling framework is two-fold and includes a nonlinear model of ILI data and a model of hospitalization data which is a linear or quadratic model with ILI as a predictive covariate. The modeling framework takes advantage of the years of weekly ILI data and successful forecast methodology development for FluSight during the seasons before the COVID-19 pandemic. Specific ILI models used are similar to two models used by Osthus et al. (2019) and Ulloa (2019), the former model being one of the most successful in FluSight from the 2015-16 and 2017-18 seasons (Osthus and Moran, 2021). Both of these models are based on nonlinear functions meant to capture the trajectory of the ILI. These models are hierarchical and are able to exploit the several seasons of ILI data. The hospitalization data has only been available since 2021, so a linear model with ILI as a covariate is able to take advantage of the ILI forecasts to produce hospitalization forecasts. The hospitalization model is an autoregressive model with exogenous variables (Raftery et al., 2010; Ljung, 1987). The evaluation of the modeling framework is done in a simulation study for data from the 2023-24 flu season.

In chapter 3 we consider the situation where distributions or forecasts are represented as a set of quantiles as is the case in FluSight. There are a number of reasons to represent uncertainty in terms of quantiles: data may be large or be private in nature (Simpson et al., 2023), prediction methods may only produce point predictions or quantile regression predictions (Gneiting, 2011; Koenker and Bassett Jr, 1978), quantiles provide simple interpretations, etc. In forecasting, however, when the forecast representation is quantiles, there are limitations on methodology that can be used for scoring forecasts and combining multiple forecasts into an ensemble forecast. There are a number of methods for estimating continuous distributions given a set of quantiles

(Gerding et al., 2023; Nirwan and Bertschinger, 2020; Gyamerah et al., 2020; Li et al., 2019), which then allow for scoring and ensemble building using methods which require a PDF or CDF. But these methods either fail to account for the inherent uncertainty of estimating quantiles or there are cases where the methods cannot be performed effectively. The contribution of the chapter is to introduce a Gaussian process model specifically designed to estimate with uncertainty a continuous distribution which is close to the distribution from which quantiles are estimated. The model is based on established asymptotic theory for sample quantiles (Walker, 1968). We include simulation studies which show the model's ability to provide accurate inference on parameters, estimate unknown distributions of various shapes, and we fit the model to the quantile forecasts submitted during the 2023-24 FluSight season.

In chapter 4 we introduce a method of combining continuous forecast distributions into a single mixture distribution ensemble forecast. An ensemble forecast often outperforms the individual forecasts from which it is constructed, and many methods already exist for combining forecasts (Wang et al., 2023). Many methods construct an ensemble by selecting combination parameters which optimize a loss function or some scoring rule, but these methods select point estimates for the optimization without allowing for the quantifying uncertainty of the optimization weights (Yao et al., 2018; Thorey et al., 2017; Geweke and Amisano, 2011). Other methods select model weights based on some variation of Bayesian model averaging, so while these methods are probabilistic they may not provide optimal weights for an ensemble forecast, and forecasts may not be well tailored to the specific problem (Lavine et al., 2021; Raftery, 1996). We introduce a method for combining models by optimizing combinations under a Gibbs posterior distribution. When a statistical likelihood is unavailable or when the interest is in minimizing a loss function rather than capture a data generating mechanism, a Gibbs posterior distribution may be used for estimating function minimizing parameters while also providing uncertainty quantification of those parameters. Using a Gibbs posterior for optimizing weights thus allows for uncertainty quantification of model combination weights which are estimated to minimize a loss function. Like a standard Bayesian posterior distribution, a Gibbs posterior uses Bayes theorem

to update a prior distribution given the exponential of a loss function evaluated on data (Martin and Syring, 2022; Bissiri et al., 2016). Thus the model optimization we present allows for uncertainty quantification in the combination scheme and also allows for using information in a prior distribution to influence the ensemble combination. We use this new model combination methodology to combine forecasts from the 2023-24 FluSight season and show that it outperforms several standard combination methods. The continuous distribution estimations from the quantile forecasts produced in chapter 3 were necessary to make the model combination possible.

As a whole, this dissertation provides novel methodologies which contribute to the literature on forecasting in collaborative forecast settings. It also provides new avenues for further research and development which are discussed in the conclusions of each chapter and in the concluding remarks in chapter 5.

1.1 References

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.

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.

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 FluView, U.S. influenza surveillance: Purpose and methods. https://www.cdc.gov/fluview/overview/?CDC_AAref_Val=https://www.cdc.gov/flu/weekly/overview.htm. Accessed: 2024-10-22.

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. (2022a). The United States COVID-19 forecast hub dataset. *Scientific Data*, 9(1):462.

Cramer, E. Y., Ray, E. L., Lopez, V. K., Bracher, J., Brennen, A., Castro Rivadeneira, A. J., Gerding, A., Gneiting, T., House, K. H., Huang, Y., et al. (2022b). Evaluation of individual and ensemble probabilistic forecasts of COVID-19 mortality in the United States. *Proceedings of the National Academy of Sciences*, 119(15):e2113561119.

De Gooijer, J. G. and Hyndman, R. J. (2006). 25 years of time series forecasting. *International Journal of Forecasting*, 22(3):443–473.

Gerding, A., Reich, N. G., Rogers, B., and Ray, E. L. (2023). Evaluating infectious disease forecasts with allocation scoring rules. *arXiv preprint arXiv:2312.16201*.

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

Gneiting, T. (2011). Quantiles as optimal point forecasts. *International Journal of Forecasting*, 27(2):197–207.

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

Gneiting, T., Wolffram, D., Resin, J., Kraus, K., Bracher, J., Dimitriadis, T., Hagenmeyer, V., Jordan, A. I., Lerch, S., Phipps, K., et al. (2023). Model diagnostics and forecast evaluation for quantiles. *Annual Review of Statistics and Its Application*, 10:597–621.

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.

HealthData.gov (2024). COVID-19 reported patient impact and hospital capacity by state (raw). https://healthdata.gov/dataset/COVID-19-Reported-Patient-Impact-and-Hospital-Capa/6xf2-c3ie/about_data. Accessed: 2024-10-22.

Holcomb, K. M., Mathis, S., Staples, J. E., Fischer, M., Barker, C. M., Beard, C. B., Nett, R. J., Keyel, A. C., Marcantonio, M., Childs, M. L., et al. (2023). Evaluation of an open forecasting challenge to assess skill of West Nile virus neuroinvasive disease prediction. *Parasites & Vectors*, 16(1):11.

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.

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

Johansson, M. A., Apfeldorf, K. M., Dobson, S., Devita, J., Buczak, A. L., Baugher, B., Moniz, L. J., Bagley, T., Babin, S. M., Guven, E., et al. (2019). An open challenge to advance probabilistic forecasting for Dengue epidemics. *Proceedings of the National Academy of Sciences*, 116(48):24268–24274.

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.

Koenker, R. (2017). Quantile regression: 40 years on. *Annual Review of Economics*, 9(1):155–176.

Koenker, R. and Bassett Jr, G. (1978). Regression quantiles. *Econometrica*, 46(1):33–50.

Krüger, F., Lerch, S., Thorarinsdottir, T., and Gneiting, T. (2021). Predictive inference based on Markov chain Monte Carlo output. *International Statistical Review*, 89(2):274–301.

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

Ljung, L. (1987). *System Identification: Theory for the User*. Prentice-Hall, Englewood Cliffs, NJ.

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.

McGowan, C. J., Biggerstaff, M., Johansson, M., Apfeldorf, K. M., Ben-Nun, M., Brooks, L., Convertino, M., Erraguntla, M., Farrow, D. C., Freeze, J., et al. (2019). Collaborative efforts to forecast seasonal influenza in the United States, 2015–2016. *Scientific Reports*, 9(1):683.

Nirwan, R.-S. and Bertschinger, N. (2020). Bayesian quantile matching estimation. *arXiv preprint arXiv:2008.06423*.

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.

Petropoulos, F., Apiletti, D., Assimakopoulos, V., Babai, M. Z., Barrow, D. K., Taieb, S. B., Bergmeir, C., Bessa, R. J., Bijak, J., Boylan, J. E., et al. (2022). Forecasting: theory and practice. *International Journal of Forecasting*, 38(3):705–871.

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.

Simpson, M., Holan, S. H., Wikle, C. K., and Bradley, J. R. (2023). Interpolating population distributions using public-use data: An application to income segregation using American Community Survey data. *Journal of the American Statistical Association*, 118(541):84–96.

Taylor, J. W. (2021). Evaluating quantile-bounded and expectile-bounded interval forecasts. *International Journal of Forecasting*, 37(2):800–811.

The Consortium of Infectious Disease Modeling Hubs (2024). The hubverse: open tools for collaborative modeling. https://github.com/hubverse-org. GitHub release v3.0.1, 5 Aug 2024.

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.

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

Walker, A. (1968). A note on the asymptotic distribution of sample quantiles. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 30(3):570–575.

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.

Warren Buffett (1981). To the Shareholders of Berkshire Hathaway Inc. https://www.berkshirehathaway.com/letters/1980.html. Accessed: 2024-10-17.

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.

CHAPTER 2. FORECASTING INFLUENZA HOSPITALIZATIONS USING A BAYESIAN HIERARCHICAL NONLINEAR MODEL WITH DISCREPANCY

Spencer Wadsworth and Jarad Niemi

Department of Statistics, Iowa State University, Ames, IA 50011

Modified from a manuscript to be submitted to *Bayesian Analysis*

Abstract

The annual influenza outbreak leads to significant public health and economic burdens making it desirable to have prompt and accurate probabilistic forecasts of the disease spread. The United States Centers for Disease Control and Prevention (CDC) hosts annually a national flu forecasting competition which has led to the development of a variety of flu forecast modeling methods. For the first several years of the competition, the target to be forecast was weekly percentage of patients with an influenza-like illness (ILI), but in 2021 the target was changed to weekly hospitalization counts. Reliable state and national hospitalization data has only been available since 2021, but for ILI the data has been available since 2010 and has been successfully forecast for several seasons. In this manuscript, we introduce a two component modeling framework for forecasting weekly hospitalizations utilizing both hospitalization data and ILI data. The first component is for modeling ILI data using a dynamic nonlinear Bayesian hierarchical model. The second component is for modeling hospitalizations as a function of ILI. For hospitalization forecasts, ILI is first forecasted and then hospitalizations are forecast with ILI forecasts used as a linear or quadratic predictor. In a simulation study, two ILI forecast models, including one similar to the winning model for two seasons of the CDC forecast competition from Osthus et al. (2019) and a nonlinear Bayesian hierarchical model from Ulloa (2019) are compared. Also assessed is the usefulness of including a systematic model discrepancy term in the ILI model.

Forecasts of state and national hospitalizations for the 2023-24 flu season are made, and different modeling decisions are compared. We found that including a discrepancy component in the ILI model tends to improve forecasts during certain weeks of the year. We also found that other modeling decisions such as the exact nonlinear function to use in the ILI model or the error distribution for hospitalization models may or may not be better than other decisions, depending on the season, location, or week of the forecast.

***K*eywords** Disease outbreak forecasting · Bayesian hierarchical modeling · Probabilistic forecasting · Model discrepancy

2.1 Introduction

Every year the seasonal influenza outbreak burdens the public health system by infecting millions, causing an influx of primary care visits and hospitalizations and leading to between 290,000 and 650,000 deaths worldwide (WHO, 2024). Molinari et al. (2007) estimated the United States' annual economic burden from medical costs, loss of income, and deaths to be over $87 billion. Accurate forecasting of infectious diseases can inform public decision making and ease the burden of an outbreak (Turtle et al., 2021; Lutz et al., 2019). There is a growing consensus that disease forecasts should be probabilistic in nature (Gneiting and Katzfuss, 2014; Bracher et al., 2021), and it has been shown that reporting forecast uncertainty along with predictions may lead to better decision making (Ramos et al., 2013; Joslyn and LeClerc, 2012; Winkler, 1971).

To better inform public decision making regarding the flu epidemic, in 2013 the United States Centers for Disease Control and Prevention (CDC) organized a national flu forecasting competition, also known as FluSight (Biggerstaff et al., 2016; Mathis et al., 2024; CDC, 2024a). Originally, over a dozen teams of researchers from academic and industry backgrounds participated in FluSight by contributing their own forecast models. Besides the 2020 season –or the flu season spanning the fall of 2020 and the winter of 2021– FluSight has been operated annually and researchers outside the CDC have been invited to participate. Initially the target data for forecasts was influenza-like illness (ILI) data. ILI is the proportion of patients who meet

a healthcare provider and who display flu like symptoms, and ILI data has been available at the state and national level since the 2010 flu season (CDC, 2024c,b). The collaborative ILI forecasting effort has led to a number of modeling developments in flu forecasting (McAndrew and Reich, 2021; Osthus and Moran, 2021; Osthus et al., 2019; Ulloa, 2019, see references therein for more examples), and in their paper's introduction, Osthus et al. (2019) categorized the most commonly used flu forecasting models into four classes including mechanistic models based on differential equation compartmental models, agent based models based on population simulation, machine learning/regression models including data driven machine learning and statistical models, and data assimilation models which are constructed by assimilating mechanistic models into a probabilistic framework. An additional forecast model used in FluSight involves the combination of several forecasts into a single ensemble forecast, which has been shown to perform well relative to individual models (McAndrew and Reich, 2021; Ray et al., 2020; Yamana et al., 2016).

The administration of FluSight saw few changes during the first seven seasons, but the onset of the COVID-19 pandemic and subsequent developments for COVID-19 forecasting led to major modifications. As a result of the COVID-19 pandemic which began during the 2019 flu season, the typical flu outbreak behavior was altered between the 2019 and 2022 seasons (Mathis et al., 2024). The COVID-19 pandemic led to the creation of the Health and Human Services (HHS) Patient Impact and Hospital Capacity Data System (HealthData.gov, 2024) which contains COVID-19 and flu hospitalization data, and the COVID-19 Forecast Hub was founded. The COVID-19 Forecast Hub was based on FluSight but with certain major adjustments including how the forecast uncertainty is represented and the addition of the weekly publication of a multi-model ensemble forecast as the official forecast of the CDC (Bracher et al., 2021; Cramer et al., 2022a). Using estimated quantiles for representing forecast uncertainty and creating a multi-model ensemble are both aspects of the COVID-19 Forecast Hub which were adopted by the flu forecast competition. Additionally the target of the flu forecasts changed from being ILI data to being HHS hospitalization data, which reports the number of hospitalizations due to a laboratory confirmed flu infection (Mathis et al., 2024; HealthData.gov, 2024). This is as a result

of having COVID-19 cases in the population making ILI data, already only a proxy for flu behavior, more difficult to interpret.

The contribution of this manuscript is to introduce a two component framework for modeling HHS hospitalization forecasts where hospitalization data and years of ILI data are used to inform forecast models. The first modeling component is a model of ILI data and the second is a model of hospitalization data with ILI as a predictive covariate. Herein we use ILI models similar to those in Osthus et al. (2019) and Ulloa (2019) for ILI forecasting. The model of Osthus et al. (2019) is a combined data assimilation and statistical regression model which involves a compartmental model in a probabilistic framework. Their model also includes an additional component for capturing a systematic discrepancy between the deterministic part of the model and the actual data, an idea which was first introduced by Kennedy and O'Hagan (2001). The model in Ulloa (2019) is a Bayesian hierarchical regression model with an underlying function intended to capture the trajectory of the seasonal ILI data. Herein, we provide a framework under which discrepancy modeling may be used along with a general function modeling ILI data, and we show the effectiveness of including discrepancy modeling during certain periods of the flu season.

In line with the newer FluSight standard of forecasting hospitalizations, we model hospitalizations as a linear function of ILI. Thus forecasts produced herein target flu hospitalizations and are a mapping of ILI forecasts to hospitalizations. This allows for ILI data from many seasons to be exploited and for ILI forecasts to assist in forecasting hospitalizations, which has fewer seasons of data than ILI. Several modeling schemes and their forecasts for the 2023 flu hospitalization season are compared, and it is shown that the modeling decisions produce good forecast results for different states or times during the flu season.

In section 2.2 we review the ILI and hospitalization data provided by the CDC and targeted by FluSight. In section 2.3 the modeling framework contributed by this manuscript is given. In the same section, functions similar to those used by Osthus et al. (2019) and Ulloa (2019) are defined. These functions are the susceptible-infectious-recovered (SIR) compartmental model and the asymmetric Gaussian (ASG) function respectively. Model fitting and implementation are

described at the end of the section. Section 2.4 is a simulation study where four ILI forecast models and their use in forecasting hospitalizations are compared. Commonly used proper scoring rules (Gneiting and Raftery, 2007), which are also introduced and defined in section 2.4, are used for comparing the forecasts. Forecasting of the 2023 flu outbreak along with assessment and comparison under several selected models is performed in section 2.5, with the analysis being done using the conventions of FluSight. Finally, the manuscript is concluded in section 2.6 with general observations and some discussion.

2.2 Flu outbreak data

In this section we introduce and define ILI and hospitalization data and evaluate the data visually. ILI and hospitalization data have been the object of forecasting for FluSight with ILI being the target for the first seven seasons and hospitalizations being the target since the 2022 season. Both of these data were collected at the state, territorial, and national level and were reported at least weekly. Overall the data is reported for 53 locations including the 50 US states, the District of Columbia (DC), Puerto Rico (PR), and at the US national level. We will refer to forecast targets throughout this manuscript. A target is the specific horizon, 1, 2, 3, or 4-weeks ahead, for a specific location and week during the season.

2.2.1 Influenza-like illness data

The US Outpatient Influenza-like Illness Surveillance Network (ILINet) collects information on respiratory illness from outpatient visits to health care providers. Over 3,400 outpatient health care providers in all 50 US states, PR, DC, and the US Virgin Islands report each week the total number of outpatient visits along with the number of ILI cases. An ILI case is defined as a "fever (temperature of 100°F[37.8°C] or greater) and a cough and/or a sore throat." Prior to the 2021 season, the definition included "without a known cause other than influenza" (CDC, 2024c). Because other illnesses such as COVID-19, RSV, and the common cold may induce similar

respiratory symptoms, ILI may include patients infected with some disease other than influenza. To know whether or not a sick patient is infected with influenza would require a laboratory test.

In 2013, when FluSight began, the ILI data was the object of the forecasts. The data was released publicly at HHS region levels, and forecast teams were asked to provide forecasts of several ILI targets on the regional levels including season onset, 1-4 week ahead ILI levels, and the week of peak ILI activity (Biggerstaff et al., 2016; McGowan et al., 2019). Currently, the ILI data is collected by the CDC and published on an online portal for viewing at the national, HHS region, census, and state levels (CDC, 2024b). To obtain ILI data, we used the `R` package `cdcfluview` which provides functions for downloading the data (Rudis, 2021). Weekly ILI data from the national, HHS region, and census levels are available from the 1997 flu season until the current season. At the state level, data is available from the 2010 flu season to the current season.

Figure 2.1 shows the ILI data at the national level for flu seasons 2010 to 2023. For most seasons there are 52 weeks listed, but for the 2010, 2015, and 2021 seasons there are 53 weeks because there were 53 Sundays during those seasons. To better align with the flu behavior, week 1 is set as the first week of August and week 52 or 53 is the last week in July of the following year. For example week 1 of the 2013 season corresponds to the first week of August 2013, and week 52 of the same season corresponds to the last week of July 2014. This convention is used for the remainder of this manuscript.

Notable from the plots in figure 2.1 is the regular trajectory of the ILI. With the exception of season 2020, the ILI begins low at week 1 and increases as the fall and winter progress until the ILI reaches a peak. As spring progresses to summer, the ILI decreases to low values. As Osthus et al. (2019) point out, there is nearly always either a global or local peak at week 22 which typically corresponds to the week between Christmas and New Year's day. Whether local or global, the ILI holiday peak is generally expected and thought to be due to widespread holiday travel, school closure, or other unique social behavior (Ewing et al., 2017; Garza et al., 2013). The only seasons when there was not a peak at week 22 were season 2022 where the season peak

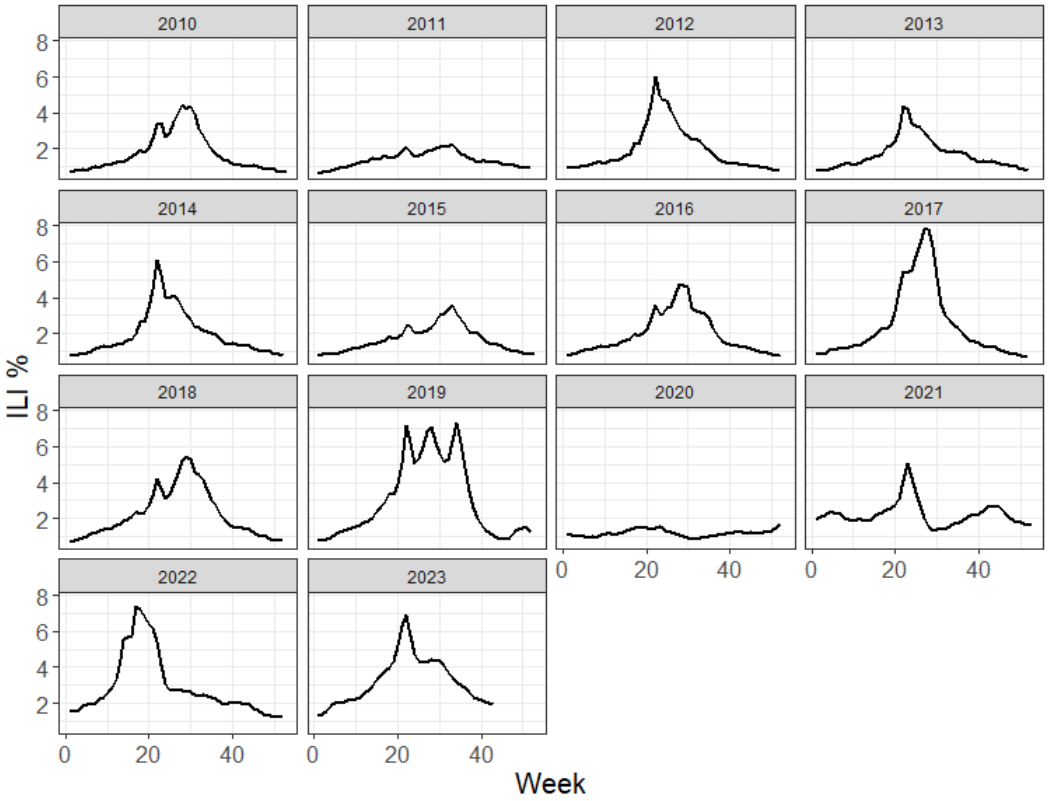

Figure 2.1: Percentage of outpatient visits with an influenza-like illness (ILI) in the US for seasons 2010 to 2023. Week 1 is the first week of August of the year the flu season begins and the last week of the season is the last week of July of the following year.

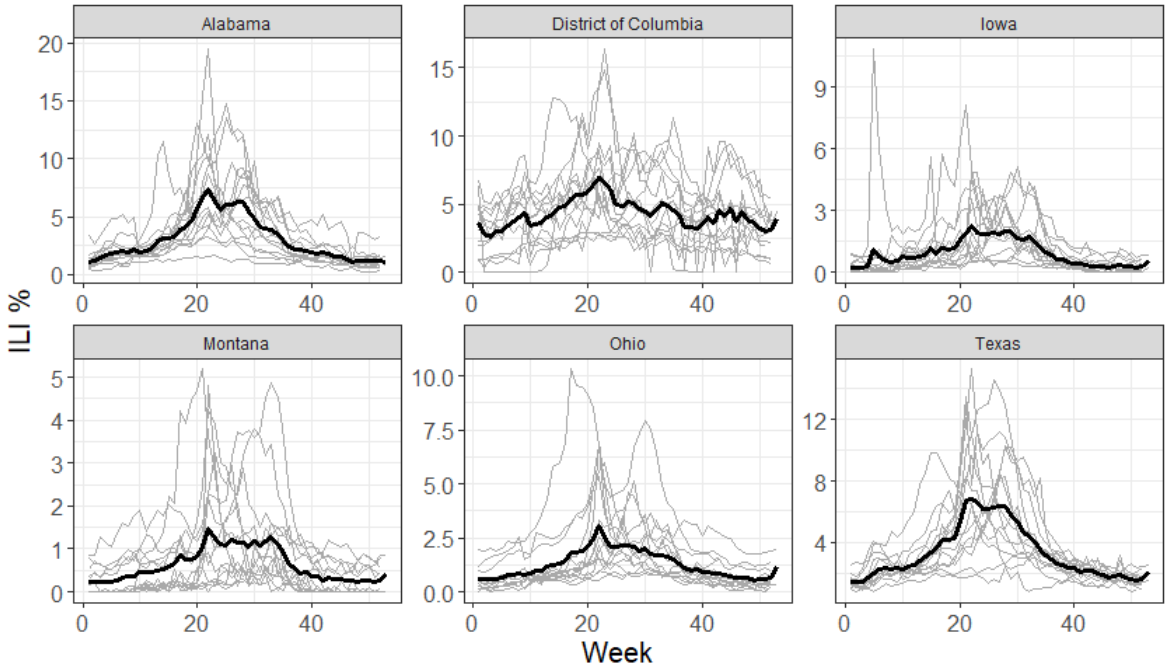

Figure 2.2: Percentage of outpatient visits with an influenza-like illness (ILI) in five different states and the District of Columbia for seasons 2010 to 2023. Week 1 is the first week of August of the year the flu season begins and the last week of the season is the last week of July of the following year. Plots include lines for ILI% from the 2010 flu season to 2023 (grey) and for the weekly ILI averaged over all seasons (black).

occurred particularly early and season 2020 which was greatly influenced by the COVID-19 pandemic.

Figure 2.2 shows ILI data from five states and the District of Columbia, locations which received particular attention in Osthus and Moran (2021). The plots include the ILI data for all seasons from 2010 to 2023 in grey, and the black line is the per week ILI average over seasons. The patterns in the individual states are similar to the national level plots in figure 2.1 in that the ILI rises in the fall and winter until it peaks and descends as the spring and summer progress. For these locations ILI regularly peaks, either locally or globally, at or near week 22.

2.2.2 Hospitalization Data

Hospital admission data, used as the object of FluSight forecasting for the 2022 and 2023 seasons, is based on the CDC's National Healthcare Safety Network (NHSN) dataset entitled

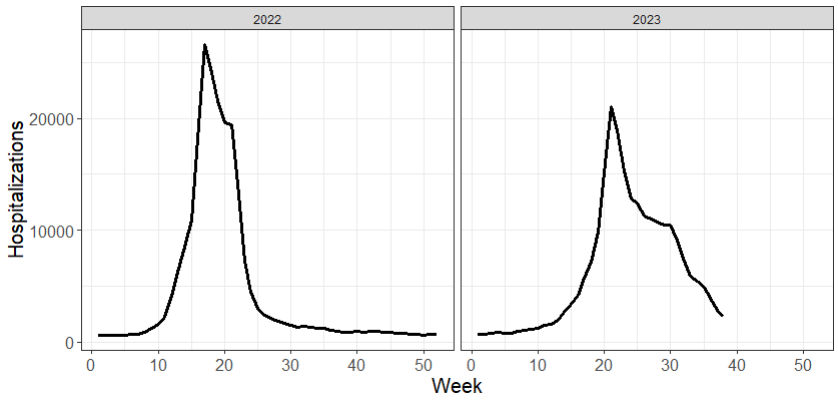

Figure 2.3: Weekly flu confirmed hospitalization counts at the national level for 2022 (left) and 2023 (right) flu seasons

HealthData.gov COVID-19 Reported Patient Impact and Hospital Capacity by State Timeseries. Several targets of respiratory illnesses including COVID-19, RSV, and influenza are reported weekly by most hospitals in the US. In February 2022 it became mandatory for all hospitals to report the number of COVID-19 and influenza hospitalizations, and since then reporting of hospitalizations has become widespread. These data were updated every Wednesday and Friday according to NHSN guidelines (HealthData.gov, 2024).

Figure 2.3 shows the weekly national hospitalizations for the 2022 and 2023 flu seasons. These plots show similarities to the ILI plots in 2.1 in that at the early weeks of the season hospitalizations are low, but they increase in the fall to a peak after which they decrease until the flu outbreak ends. For both the 2022 and 2023 seasons, the hospitalizations peaked during the same week as ILI, and in 2023 that peak occurred during the holiday week 22. Figure 2.4 shows the 2022 and 2023 weekly hospitalizations for the same states from 2.2. Similar to the national data, the peak in 2022 came early compared to the peak of 2023.

Comparing figures 2.2 and 2.4 shows that ILI and hospitalizations share the similar pattern of increasing to a peak in the winter and decreasing thereafter. Figure 2.5 shows scatter plots with ILI% on the x-axes and hospitalizations on the y-axes, revealing a positive somewhat linear relationship between the two variables. This relationship motivates the forecast models outlined in the next section.

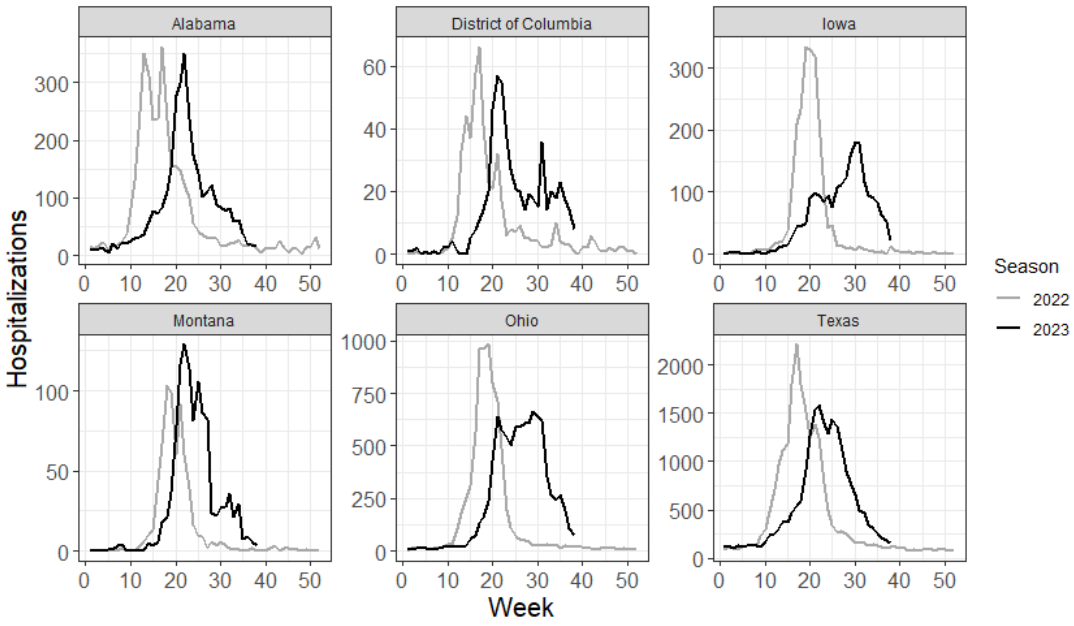

Figure 2.4: Weekly hospitalization counts for five states and DC for the 2022 (grey) and 2023 (black) flu seasons

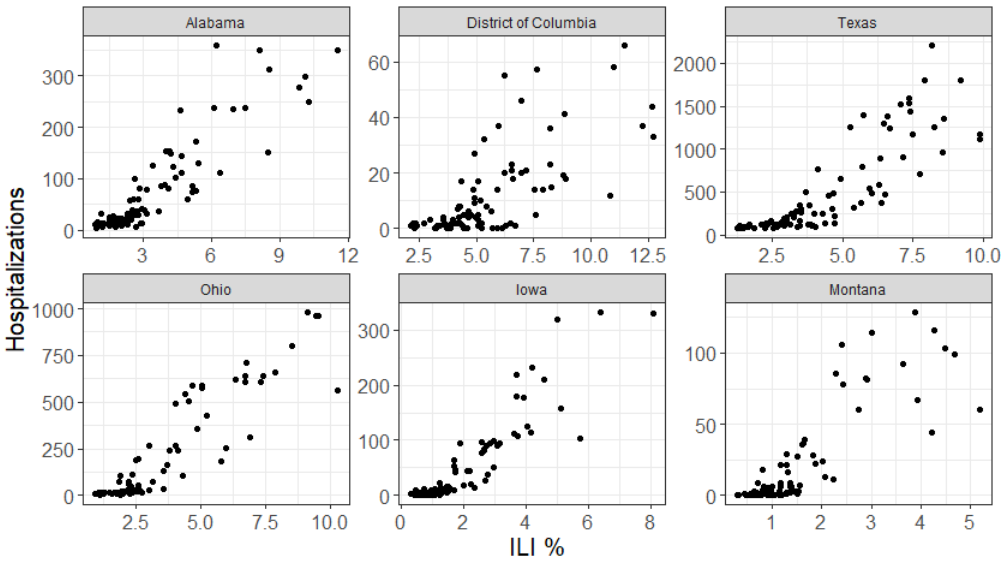

Figure 2.5: Weekly flu confirmed hospitalization counts (y-axis) for five states and the District of Columbia for the 2022 and 2023 flu seasons plotted against ILI% (x-axis)

2.3 ILI and hospitalization forecast modeling

The typical behavior of the ILI data which starts at lower values in the late summer and fall but which increases to a peak, usually in December or January, followed by a decline motivates the use of a nonlinear function which follows of a similar trajectory for modeling ILI. Compartmental models are standard mathematical models used for disease outbreaks. One important compartmental model is the susceptible-infectious-recovered (SIR) model, which is used by some to model ILI data (Osthus et al., 2019; Allen, 2017). Ulloa (2019) chose to use the asymmetric Gaussian (ASG) function to model ILI data. The SIR and ASG models may both be appropriate to describe ILI behavior over the course of a flu season, however there may also be systematic behavior not captured by either, necessitating an additional model component to capture the discrepancy. In the first part of this section, we present an ILI model similar to the model in Osthus et al. (2019). With some generalization, the model may incorporate any appropriate nonlinear function, though the focus here is on the SIR and ASG functions which are defined.

With the aim of forecasting hospitalizations, we also introduce a linear model of hospitalizations data with ILI as a predictive covariate. To forecast hospitalizations, ILI data is first forecasted and the forecast is then plugged in as a covariate in the hospitalization model, thus producing hospitalization forecasts. The hospitalization model is also defined in this section, and the section is concluded with descriptions of selected prior distributions, model implementation, and posterior sampling.

2.3.1 ILI Model

The proposed model for ILI for any location is given in (2.1). Here $ILI_{s,w}$ is the ILI for flu season s and week $w = 1, 2, ..., W$, where $W = 52$ or $W = 53$, depending on how many Sundays there are in a given season. The ILI is a proportion, so the Beta random variable is a natural selection for modeling. Under the parameterization in of the Beta distribution used in (2.1) the expected value is $\pi_{s,w}$ and the variance is $\pi_{s,w}(1 - \pi_{s,w})/(1 + \kappa_s)$, making κ_s a scale parameter.

The nonlinear function $f_{\theta_s}(w)$ captures the trajectory of the ILI, and γ_w is a discrepancy term for capturing the systematic patterns which $f_{\theta_s}(w)$ does not capture.

$$\begin{aligned} ILI_{s,w} &\overset{ind}{\sim} \text{Beta}(\pi_{s,w}\kappa_s,\ \kappa_s(1-\pi_{s,w})) \\ \text{logit}(\pi_{s,w}) &= f_{\theta_s}(w) + \gamma_w \end{aligned} \tag{2.1}$$

In Osthus et al. (2019) $f_{\theta_s}(w) = \text{logit}(I_{s,w})$ where $I_{s,w}$ is the infectious compartment of the SIR model from (2.2) in section 2.3.2. In Ulloa (2019) $f_{\theta_s}(w) = ASG_\theta(w)$ from (2.3) in section 2.3.3. In Ulloa (2019) modeling is done hierarchically over seasons.

2.3.2 Susceptible-Infectious-Recovered (SIR) compartmental model

The SIR compartmental model is a mathematical model used for modeling disease outbreaks and was introduced by Kermack and McKendrick (1927). Since then, compartmental models have become standard for modeling infectious diseases (Allen et al., 2008), and many extensions have been made and studied (Simon, 2020; Allen, 2017; Van den Driessche, 2008, for example). The SIR mathematical model includes three compartments and assumes that at any time $t > 0$ every individual in a closed population belongs to exactly one compartment. The three compartments are susceptible (S), infectious (I), and recovered (R), and their interaction over the course of an outbreak is described by the differential equations in (2.2).

$$\frac{dS}{dt} = -\beta SI, \quad \frac{dI}{dt} = \beta SI - \delta I, \quad \frac{dR}{dt} = \delta I \tag{2.2}$$

Here S, I, and R represent the proportion of the population in each compartment such that $S + I + R = 1$ for all t. The trajectory is determined by the disease transmission rate $\beta > 0$ and the recovery rate $\delta > 0$. Respectively, these may be thought of as the expected proportion of susceptible individuals who will be infected by an infectious individual, and the expected rate of recovery to an immune state for a newly infected person. Whether or not a disease outbreak is classified as an epidemic is determined by the initial susceptible population S_0, or the susceptible population at time 0, and the parameter $\rho = \delta/\beta$. If $S_0/\rho > 1$, the outbreak is considered an

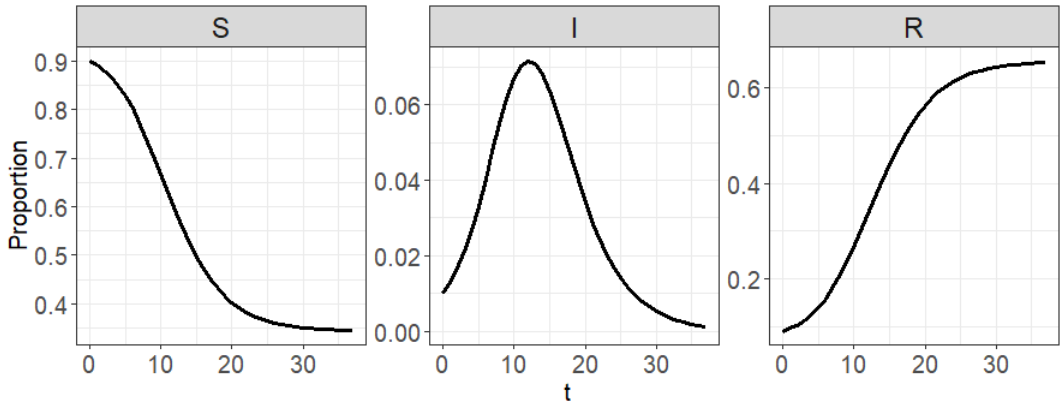

Figure 2.6: Susceptible-infectious-recovered (SIR) model separated by compartments. The three compartments are the susceptible compartment (left), infectious (center), and recovered (right). In this example, $S_0/\rho > 1$.

epidemic. It is non-epidemic if $S_0/\rho \leq 1$ (Osthus et al., 2019). Figure 2.6 shows the trajectory of the three compartments of an SIR model where the S_0 and ρ were selected to match an outbreak that would classify as epidemic. In the case where $S_0 \leq \rho$, the trajectory for the I compartment would never be increasing. The increase to a peak and subsequent decrease in the I compartment of figure 2.6 suggest it is reasonable to model ILI by this compartment. Thus in modeling the ILI data, we consider the data to be analogous to the I proportion of the population.

2.3.3 Asymmetric Gaussian (ASG) function

The ASG function is another example of a nonlinear function which can approximate the trajectory of the flu outbreak. The ASG was previously used by Ulloa to model and forecast ILI (Ulloa, 2019), and it has been used to model vegetation growth and satellite sensor data (Lewis-Beck et al., 2020; Jonsson and Eklundh, 2002; Hird and McDermid, 2009; Beck et al., 2006; Atkinson et al., 2012). The ASG is a modification of the asymmetric Gaussian distribution (Wallis, 2014) and is characterized by its rise to a peak and fall from that peak which may not occur at the same rate, as shown in figure 2.7. The ASG function is denoted as $ASG_\theta(w)$ where $\theta = (\lambda, \nu, \mu, \sigma_1^2, \sigma_2^2)$, $\nu > 0$, $\lambda > 0$, $\mu \in (-\infty, \infty)$, $\sigma_1, \sigma_2 > 0$ and $w \in (1, ..., W)$ is week. The function is defined in (2.3).

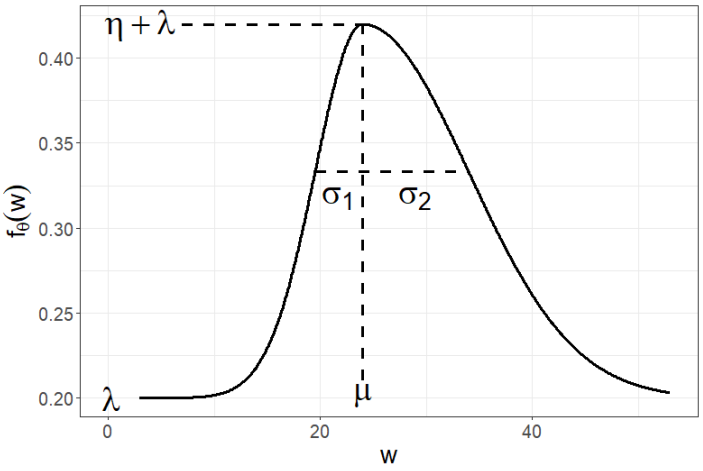

Figure 2.7: Example plot of asymmetric Gaussian (ASG) function showing the shape of the function in relation to the parameters λ, η, μ, σ_1, and σ_2

$$ASG_\theta(w) = \begin{cases} \lambda + (\nu - \lambda)\exp[-(w-\mu)^2/2\sigma_1^2], & w < \mu \\ \lambda + (\nu - \lambda)\exp[-(w-\mu)^2/2\sigma_2^2], & w \geq \mu \end{cases} \tag{2.3}$$

For modeling in this manuscript, we use a slightly reparameterized version of the function in (2.4), where $\eta = \nu - \lambda > 0$. This constraint guarantees that the function has a peak greater than λ.

$$ASG_\theta(w) = \begin{cases} \lambda + \eta\exp[-(w-\mu)^2/2\sigma_1^2], & w < \mu \\ \lambda + \eta\exp[-(w-\mu)^2/2\sigma_2^2], & w \geq \mu \end{cases} \tag{2.4}$$

2.3.4 Model discrepancy

The SIR and ASG functions are useful for capturing the main trend of the ILI data, but as Osthus et al. (2019) points out there may be systematic behavior that these or other possible functions may not capture. As noted in section 2.2, figures 2.1 and 2.2 show a regular peak at week 22 of the flu season. Figures 2.8 and 2.9 are used together to illustrate the systematic discrepancy from a fitted function. Figure 2.8 shows the US ILI percentage for all flu seasons from 2010 to 2022 excluding 2020 with a best fit ASG function plotted over the ILI. The fits for each season were made by obtaining the maximum likelihood estimate (MLE) of a model given the ILI

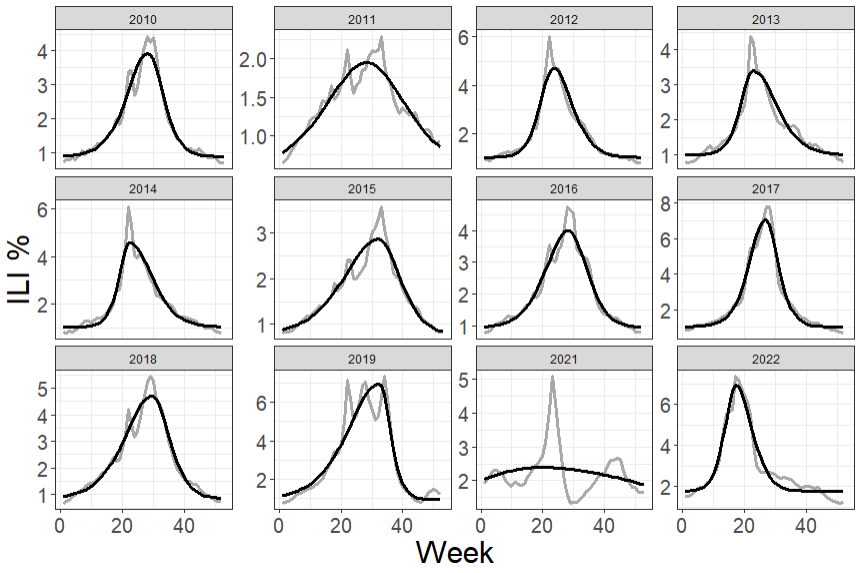

Figure 2.8: Observed US national influenza-like illness (ILI) percentage for seasons 2010 to 2022 excluding 2020 (grey) overlaid with MLE of an asymmetric Gaussian (ASG) model for the ILI data (black)

data where we assume the ASG function is the mean parameter of a Beta distributed random variable. Figure 2.9 shows the discrepancy between the model fit and the data for the same seasons. The grey lines show the difference between the data and the functions from figure 2.8 for each season, and the black line is the average by week over all seasons. The lines show that the ASG function typically underpredicts week 22 and overpredicts week 23. Perhaps for other weeks, like week 30 for example, there also tends to be systematic behavior not captured by the ASG function.

The term γ_w, where w is the season week, is included in (2.1) to capture the per week discrepancy between ILI and the function. Modeling discrepancy has been used in uncertainty analysis of simulators to capture systematic differences between mathematical models and reality (Ma et al., 2022; Brynjarsdóttir and O'Hagan, 2014; Arendt et al., 2012; Kennedy and O'Hagan, 2001). Modeling discrepancy can lead to overfitting, particularly in forecasting scenarios, and may also lead to identifiability issues. Thus, care must be taken in setting parameter constraints as well as in the selection of prior distributions (Osthus et al., 2019; Brynjarsdóttir and O'Hagan, 2014). Modeling of the discrepancy for ILI was done by Osthus et al. (2019) during the 2015 and 2016 flu

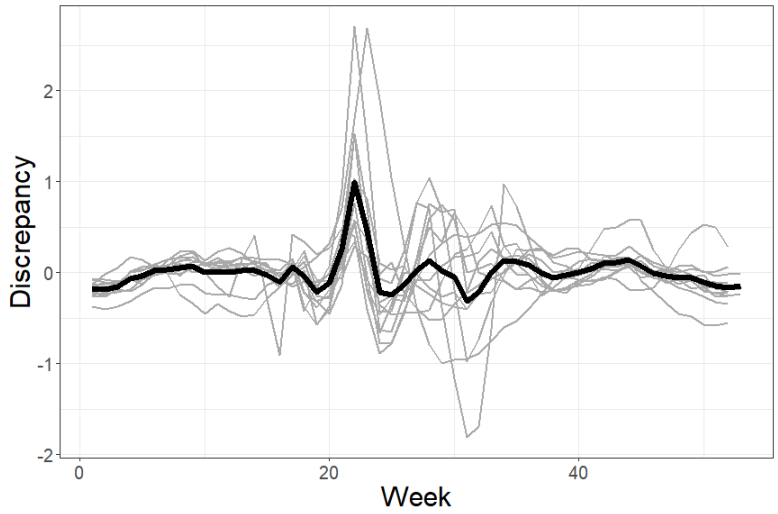

Figure 2.9: Difference between observed US national influenza-like illness (ILI) and MLE fits for an asymmetric Gaussian (ASG) model for each season 2010 to 2022 excluding 2020 (grey) and the average difference of all seasons (black)

seasons where their model outperformed all others in the CDC flu forecasting challenge (Osthus et al., 2019). As in Osthus's model, γ_w is modeled as a reverse random walk, as shown in (2.5).

$$\gamma_w|\gamma_{w+1} \overset{ind}{\sim} N(\gamma_{t+1}, \sigma^2_\gamma), \quad \gamma_W \sim N(0, \sigma^2_{\gamma_W}) \tag{2.5}$$

The idea for using the reverse random walk is that there are several previous seasons of ILI data, and assuming the random walk captures systematic behavior, fitting it hierarchically over seasons can assist in predicting future behavior in the current season. Reverse random walks have also been used with success in election forecasting and other flu forecasting models (Osthus and Moran, 2021; Osthus et al., 2019; Linzer, 2013). The sum to zero constraint $-\gamma_1 = \sum_{w=2}^{W} \gamma_w$ is imposed on (2.5) to improve identifiability.

2.3.5 ILI forecasts

Model (2.1) is fit via Bayesian posterior updating. Future ILI forecasts are obtained via the posterior predictive distribution where for week w, the predictive distribution is obtained by integrating over the parameters π_s, κ_s, σ^2_γ, and $\sigma^2_{\gamma_W}$ as in (2.6) where $p(\boldsymbol{\pi}_s, \kappa_s, \sigma^2_\gamma, \sigma^2_{\gamma_W}|\mathbf{ILI})$ is the

density function of the posterior distribution for the model parameters. If the current week is w^* then the desired forecasts are for weeks $w^* + i$ where i is a positive integer.

$$p(\widetilde{ILI}_{s,w^*}|\mathbf{ILI}) = \int\int\int\int p(\widetilde{ILI}_{s,w^*}|\boldsymbol{\pi}_s, \kappa_s, \sigma^2_\gamma, \sigma^2_{\gamma_W}) p(\boldsymbol{\pi}_s, \kappa_s, \sigma^2_\gamma, \sigma^2_{\gamma_W}|\mathbf{ILI}) d\boldsymbol{\pi}_s d\kappa_s d\sigma^2_\gamma d\sigma^2_{\gamma_W} \tag{2.6}$$

2.3.6 Hospitalization model

The second component for forecast modeling is the hospitalization model defined in (2.7). This is an example of an autoregressive model with exogenous variables where the autoregressive lag is one (ARX(1)) (Raftery et al., 2010; Ljung, 1987). Here $H_{s,w}$ is the number of hospitalizations for week w in season s, $\epsilon_{s,w}$ is an error term distributed according to some distribution D_s with mean parameter 0, scale parameter σ_{ϵ_s}, and the additional parameter ω_s is the degrees of freedom parameter when D_s belongs to the location-scale t (LST) family.

$$\begin{aligned} H_{s,w} &= \alpha_{0s} + \alpha_{1s}(ILI_{s,w} \times P) + \alpha_{2s}(ILI_{s,w} \times P)^2 + \phi H_{s,w-1} + \epsilon_{s,w} \\ \epsilon_{s,w} &\overset{iid}{\sim} D_s(0, \sigma^2_{\epsilon_s} \times P, \omega_s) \end{aligned} \tag{2.7}$$

This model is for any location, and for fitting purposes $ILI_{s,w}$ is always multiplied by P which is proportional to the population of the state or territory, in this case the total population divided by 50,000. This is done as a means of scaling so that the prior distribution assigned to $\boldsymbol{\alpha}_s = (\alpha_{0s}, \alpha_{1s}, \alpha_{2s})$, σ_{ϵ_s}, and ω_s might reasonably be the same for all states.

Like the ILI model, the hospitalization model in (2.7) is also fit via Bayesian posterior updating. To obtain forecasts for H_{s,w^*+i}, the ILI posterior predictive distribution is used along with the posterior distribution for the parameters in (2.7). We considered three scenarios for model (2.7). A model where D_s belongs to the normal family (NORM), D_s belongs to the LST family, and one where $H_{s,w}$ is replaced with $\log(H_{s,w} + 1)$ and D_s is from a normal family, or $H_{s,w} + 1$ is lognormally distributed (LNORM). The population value P is excluded from the LNORM model, and we set $\boldsymbol{\alpha}_{22} = \boldsymbol{\alpha}_{23}$ to help with fitting. In the LNORM model, if the linear parameters are not set the same for seasons 22 and 23, the final variance was more prone to be

extreme. Besides varying the distribution family of hospitalizations, we also considered $\alpha_{2s} = 0$ or there is no quadratic ILI term.

2.3.7 Prior selection

The priors selected for the ILI data model under both the SIR and ASG models largely follow the prior selections in Osthus et al. (2019) and Ulloa (2019) with a few exceptions where changes improved numerical stability and/or we felt the adjusted prior made more sense for the problem. For model (2.1), parameters which are common even when using different functions of $f_{\theta_s}(w)$ are κ_s, σ^2_γ, and $\sigma^2_{\gamma_W}$. For the SIR function $\theta_s = (S_{0s}, I_{0s}, R_{0s}, \alpha_s, \rho_s)$, and for the ASG function $\theta_s = (\alpha_s, \eta_s, \mu_s, \sigma^2_{1s}, \sigma^2_{2s})$. For the hospitalization model in (2.7) the parameter to be estimated is $\Psi = (\alpha_{0s}, \alpha_{1s}, \alpha_{2s}, \phi, \sigma_{\epsilon_s}, \omega_s)$.

The priors assigned were mostly noninformative, though in certain cases the prior distributions were selected for numerical stability as was the case for σ^2_γ and $\sigma^2_{\gamma_W}$. For these two scale parameters only, rather than assigning a half-normal prior to the standard deviation parameters, as recommended by Gelman (2006), the priors were assigned to the variance parameters. Univariate parameters were assigned either a normal distribution prior if the support is on $\mathbb{R}$, a half-normal prior if the support is nonnegative, or a truncated-normal prior to match a more specific support. Under the ASG model, θ_s is modeled hierarchically over seasons so that for each season the transformed parameter $T(\theta_s) \sim N(\theta, \Sigma)$ and priors distributions are assigned to θ and Σ.

Additional prior constraints were made to improve parameter identifiability. In Osthus et al. (2019) the initial value of the susceptible population compartment of the SIR model was set to $S_0 = 0.9$. The parameters I_{0s}, β_s, and ρ_s were assigned informative priors. To improve identifiability when $f_{\theta_s}(w) = ASG_{\theta_s}(w)$ in (2.1) we followed a modular Bayesian approach for fitting the parameters. A modular Bayesian approach involves multiple steps of parameter fitting where some parameters may be estimated without priors via maximum likelihood estimation or other means. Fitting the rest of the model parameters involves assigning priors and conditioning

on the previously fit parameters. This has been done in computer modeling to improve identifiability and other issues, though Liu et al. (2009) warn this approach is not probabilistically sound if parameter inference is a priority (Jiang et al., 2015; Arendt et al., 2012; Liu et al., 2009). We carried out the modular fit by first estimating the maximum likelihood estimate (MLE) for the parameter λ_s for each season in (2.4) and plugging in the MLE as a fixed value.

2.3.8 Parameter estimation and posterior predictive sampling

The models were fit via Markov chain Monte Carlo (MCMC) sampling using the `cmdstanr` package which was developed and is maintained by the Stan Development Team (2024) (Gabry et al., 2022). Stan implements Hamiltonian Monte Carlo (HMC) sampling with the No-U-turn sampler (Hoffman et al., 2014). The `cmdstanr` package provides several diagnostic statistics for assessing the sampler. As mentioned, most model parameter prior distributions were intended to be uninformative. Appendix 2.B shows plots of posterior distributions for select parameters from ILI and hospitalization models.

We assessed the model fit for four ILI models. These included the SIR and ASG models and models with and without discrepancy modeling. When discrepancy is included, the models are denoted as SIRD and ASGD. These models were fit using US national data from 2010 to 2023 flu seasons, where data from the 2020 season was excluded because of the unique behavior during that season. Assessment for hospitalization modeling was done for six different models. These include NORM, LNORM, and LST models, and models where a quadratic ILI term is included or excluded. To assess posterior sampling convergence, models were fit to data where ILI and hospitalization data up to week 14 of the 2023 season was included. Sampling was done with four chains where from each chain 60,000 posterior draws were sampled, and the first 10,000 draws were discarded as a burn-in. The $\hat{R}$ statistic (Vehtari et al., 2021) and the effective sample size (ESS) (Gelman et al., 2013) were calculated for each parameter. Tables 2.1 and 2.2 summarize the maximum $\hat{R}$ and the minimum ESS over all parameters for ILI and hospitalization models respectively. Forecast models for all other weeks of the season were fit using one chain of 60,000

Table 2.1: Maximum $\hat{R}$ and minimum ESS over all parameters for four ILI models fitted on US data for week 14 of the 2023 season

	ASG	ASGD	SIR	SIRD
$\hat{R}$	< 1.001	< 1.001	< 1.001	1.001
ESS	72,057	9,980	17,745	6,259

Table 2.2: Maximum $\hat{R}$ and minimum ESS over all parameters for six hospitalization models fitted on US data for week 14 of the 2023 season

	NORM	NORM^2	LNORM	LNORM^2	LST	LST^2
$\hat{R}$	< 1.001	< 1.001	< 1.001	< 1.001	< 1.001	< 1.001
ESS	71,266	73,554	46,747	54,975	7,660	62,663

draws where the first 10,000 draws were discarded as a burn-in. For parameters of the ASG models that were prone to cause trouble in posterior sampling, the starting values were set to be the MLEs.

To obtain forecast distributions of hospitalizations, draws from the posterior predictive distribution from the ILI model were used in conjunction with the posterior distribution of the hospitalizations model. When fitting model (2.6), MCMC samples of $\widetilde{ILI}_{s,w:(w+4)}$ were saved. Model (2.7) was fit and MCMC samples for the marginal distributions for the model parameters were saved. To obtain forecast distributions for $H_{s,w+i}$ where $i \in \{1, 2, 3, 4\}$, the following steps are repeated K times where K is an integer for the number of desired samples. We set $K = 50,000$.

Step 1: Sample $\widetilde{ILI}^*_{s,w:(w+4)}$

Step 2: Sample α^*_{0s}, α^*_{1s}, α^*_{2s}, ϕ^*, $\sigma^*_{\epsilon_s}$, ω^*_s from respective marginal posterior distributions

Step 3: Sample $H^*_{s,w+i}$ from $D(\omega^*_s, \mu^*_{s,w+i}, \sigma^2_{\epsilon_s})$, where

$$\mu^*_{s,w+i} = \alpha^*_{0s} + \alpha^*_{1s}(ILI^*_{s,w+i} \times P) + \alpha^*_{2s}(ILI^*_{s,w+i} \times P)^2 + \phi^* H^*_{s,w+i-1}$$

Step 4: Repeat step 3 for $i \in \{1, 2, 3, 4\}$ to obtain $H^*_{s,(w+1):(w+4)}$

Step 5: Repeat steps 1-4 K times

The sample $\{H^*_{s,w+i}\}^K$ was then used as the probabilistic forecast for hospitalizations at week $w+i$. For the forecast competition analysis in section 2.5, all negative values of $\{H^*_{s,w+i}\}^K$ were set to 0 to reflect realistic values of hospitalizations and comply with the FluSight forecasting rules.

2.4 Simulation Study

In this section, we present a simulation study conducted for comparing ILI models and further assessing the hospitalization forecast model. US ILI data is used, and hospitalization data is simulated. A leave-one-season-out (LOSO) approach was combined with a Monte Carlo simulation approach. For each replication, we simulated log-hospitalizations for all weeks during seasons 2010, ..., 2022, excluding 2020, using the existing ILI data as a predictive covariate. Each season was in turn "left-out" and treated as if it was the most recent season which we desired to forecast. Fitting and forecasting was then done for weeks 14, 20, 26, 32, and 38 of the left out season, giving two weeks that tend to occur as flu cases increase, two as cases decrease, and one that occurs when the cases may be increasing or decreasing. Week 20 is a week leading up to the holiday week 22 where ILI typically has a local peak. We were particularly interested in how important modeling discrepancy is for forecasting at week 20.

For the simulation of hospitalizations, the parameters $\boldsymbol{\alpha}_s = (\alpha_{0s}, \alpha_{1s}, \alpha_{2s})$ and $\sigma^2_{\epsilon_s}$ from the hospitalization model in (2.7) were considered the same across all seasons so that all $\boldsymbol{\alpha}_s = \boldsymbol{\alpha}$. The values for $\boldsymbol{\alpha}$, σ^2_ϵ, and ϕ were estimated by fitting model (2.7) using ILI and hospitalization data from the 2022 season. For fitting, the hospitalization data was first log-transformed. We took $\boldsymbol{\alpha}_\phi = (\boldsymbol{\alpha}, \phi)$ and assigned the noninformative prior $p(\boldsymbol{\alpha}_\phi, \sigma^2_\epsilon) \propto 1/\sigma^2_\epsilon$. The marginal posterior distribution $\boldsymbol{\alpha}_\phi | \sigma^2_\epsilon, \boldsymbol{H_{22}}$ was then the established posterior multivariate normal distribution and $\sigma^2_\epsilon | \boldsymbol{H_{22}}$ the inverse-χ^2 posterior distribution (Gelman et al., 2013). The posterior means of those parameters were used as the values from which log-hospitalizations were simulated. The number of replicates in the simulation was 500.

Model comparison was done by calculating the continuous ranked probability score (CRPS) and the logarithmic score (LogS) for each forecast. These scores are both proper scoring rules which evaluate the forecast distribution and density functions respectively. Proper scoring rules are the current standard for comparing performance between probabilistic forecasts and selecting the best forecasts according to the notion of maximizing sharpness subject to (auto-)calibration (Gneiting et al., 2007; Tsyplakov, 2013). Proper scoring rules are commonly used in forecast comparison and have the property that a forecaster is incentivized to be honest in the reporting of their forecasts (Gneiting and Raftery, 2007; Gneiting and Katzfuss, 2014). The CRPS is defined in (2.8) and the LogS in (2.9). Here $F(\cdot)$ is the forecast distribution function, $f(\cdot)$ is the forecast density function, and y^* is the observed targeted value of the forecast. The orientation of both scores is negative, meaning the smaller the score the better.

$$\text{CRPS}(F, y^*) = \int_{-\infty}^{\infty} (F(x) - \mathbb{1}(y^* \leq x))^2 dx \tag{2.8}$$

$$\text{LogS}(f, y^*) = -\log(f(y^*)) \tag{2.9}$$

Both the CRPS and LogS are calculated using the `scoringRules` package in `R` (Jordan et al., 2019). To calculate the LogS when given MCMC samples from a posterior predictive distribution, a continuous density function was first estimated via kernel density estimation. The CRPS is calculated using a quantile decomposition from (Laio and Tamea, 2007).

Figures 2.10 - 2.14 show boxplots of the CRPS and LogS for the four models. Figure 2.10 shows that the variation of overall CRPS is smallest for the ASG models and larger for the SIR models. The median scores for the SIR models also appears slightly higher than for the ASG models. When faceted by season in figure 2.11, the boxplots of the CRPS often show the same pattern for ASG and SIR models but not always. For example, the bulk of CRPS values for the SIRD model in 2019 appears to have smaller variation than the other models. Figure 2.12 shows CRPS boxplots faceted by week and horizon. Notable plots here are for week 20 where the two models accounting for the ILI discrepancy, ASGD and SIRD, have CRPS values with lower

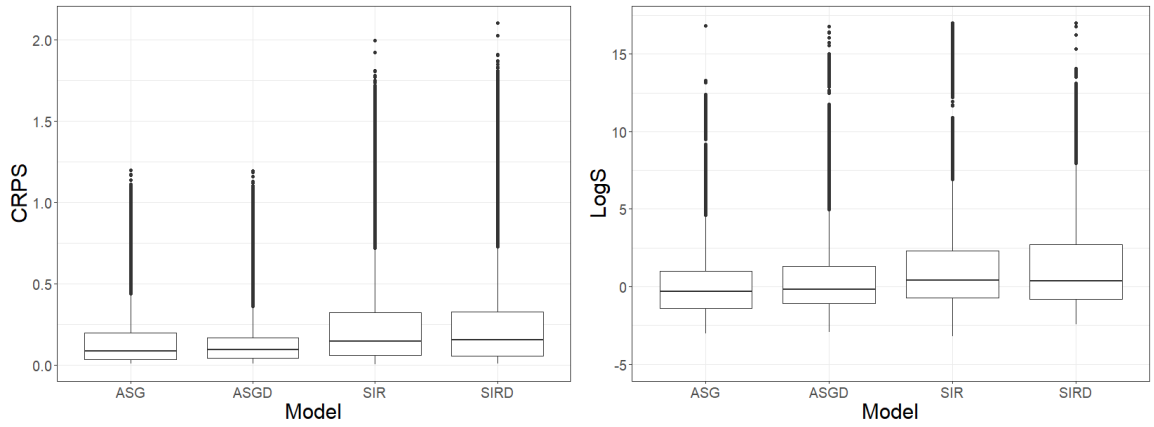

Figure 2.10: Boxplots of the continuous ranked probability score (CRPS) (left) and logarithmic score (LogS) (right) for the four ILI models over all seasons, weeks, and horizons in the simulation study

medians and lower variation than the two models not accounting for ILI discrepancy. This makes sense given the forecasts at week 20 are forecasting weeks, 22 and 23 where figure 2.9 shows a seasonal peak and trough.

Figures 2.13 and 2.14 show boxplots of the LogS for the simulated forecasts. Note that in these plots only values of less than 17 are included to improve visualization. The LogS plots show similar results to the CRPS, though there are some differences. For example, SIRD in figure 2.13 tends to show smaller LogS variation relative to the other three models than is seen by the CRPS of SIRD in figure 2.11. We also note the smaller relative LogS variation than CRPS variation for SIRD when comparing figure 2.14 with figure 2.12, particularly at week 14. Among the four models, ASG tends to have the lowest values in CRPS and LogS though this is not always the case. When forecasting weeks 21-23, the two models which model discrepancy in the ILI, ASGD and SIRD, tend to outperform the models not modeling the discrepancy. This confirms that there is value in modeling discrepancy at least during around the holiday weeks near week 22.

2.5 Analysis of forecasts for 2023 flu season

In this section we apply the forecast models to make forecasts of the 2023 flu season weekly hospitalizations. The scoring of the forecasts is in the context of the FluSight competition where each competing forecast was submitted as a set of 23 quantiles corresponding to the given

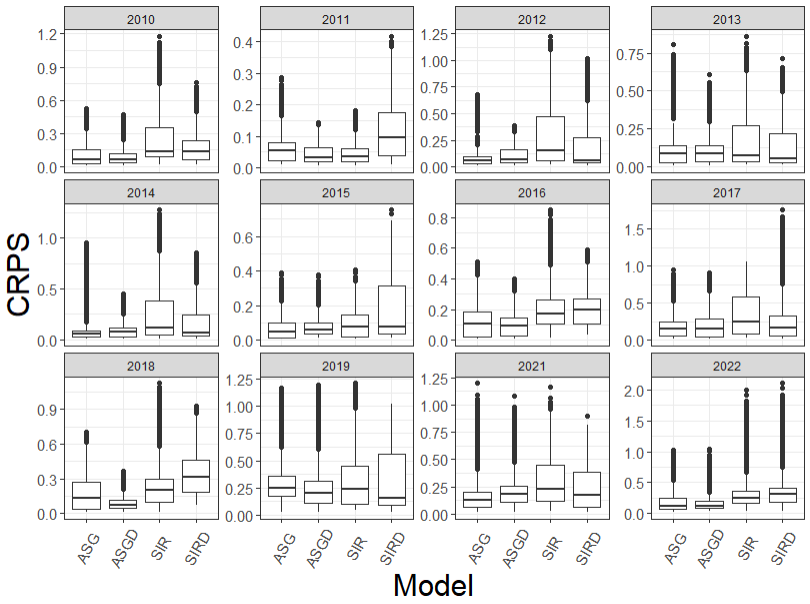

Figure 2.11: Boxplots of continuous ranked probability score (CRPS) for the four ILI models over all weeks and horizons in the simulation study faceted by season and including seasons 2010-2022, excluding 2020

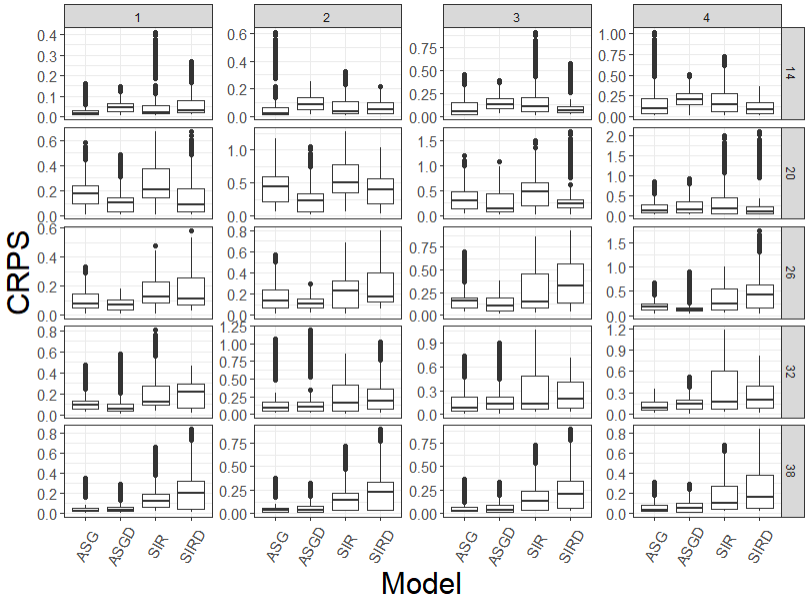

Figure 2.12: Boxplots of continuous ranked probability score (CRPS) for the four ILI models over all seasons in the simulation study faceted by horizon (x-axis) and week (y-axis). Horizons include 1-4 week ahead forecasts and weeks include weeks 14, 20, 26, 32, and 38 of the flu season

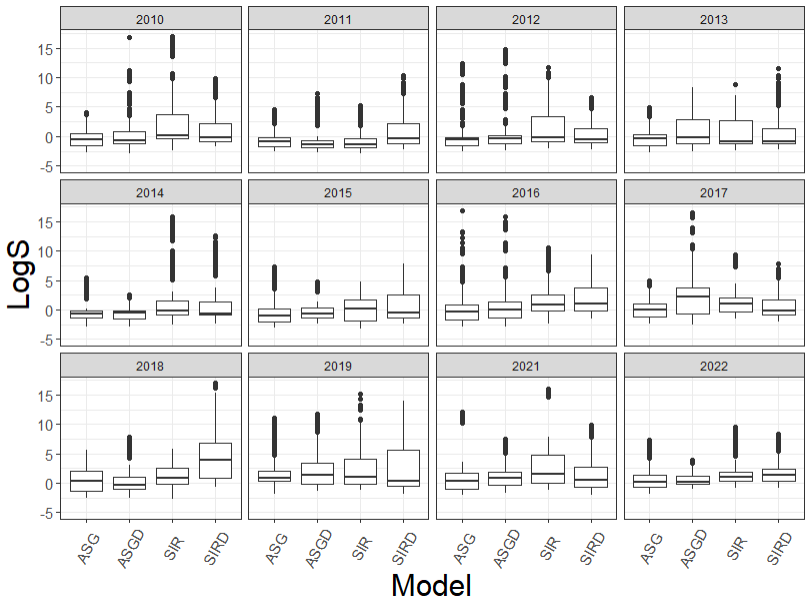

Figure 2.13: Boxplots of logarithmic score (LogS) for the four ILI models over all weeks and horizons in the simulation study faceted by season and including seasons 2010-2022, excluding 2020.

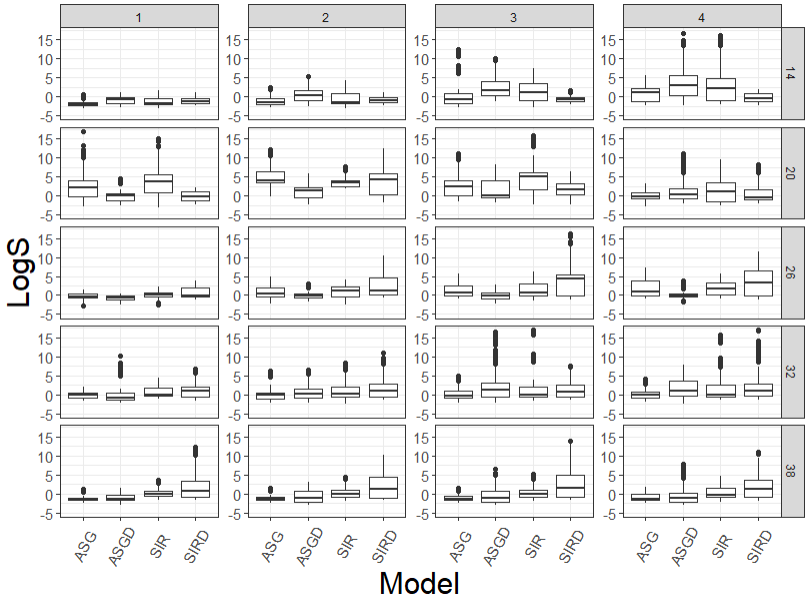

Figure 2.14: Boxplots of logarithmic (LogS) for the four ILI models over all seasons in the simulation study faceted by horizon (x-axis) and week (y-axis). Horizons include 1-4 week ahead forecasts and weeks include weeks 14, 20, 26, 32, and 38 of the flu season.

probability levels $(0.010, 0.025, 0.050, 0.100, 0.150, \ldots, 0.950, 0.975, 0.990)$. A single forecast is thus comprised of 11 predictive intervals and a median. Forecasts of 1, 2, 3, and 4-week ahead hospitalization counts were requested, and forecasts were made at the state and national level. The first week of forecasting took place during the week of October 7, 2023, and the final week was the week of April 27, 2024 making 29 total weeks of forecasts. The same format was used during the 2021 and 2022 seasons and for the COVID-19 Forecast Hub (Mathis et al., 2024; Bracher et al., 2021). Primary scores for evaluating each forecast were the weighted interval score (WIS), the log-weighted interval score (LWIS), and the relative weighted interval score (RWIS). The focus in this section will be on the LWIS (see appendix 2.A for RWIS based results).

The WIS is a proper scoring rule used for scoring quantile or interval forecasts (Gneiting and Raftery, 2007; Gneiting and Katzfuss, 2014; Bracher et al., 2021) and is defined in (2.10) where Q is a forecast represented by all included quantiles, B is the number of intervals, y^* is the observed value targeted by the forecast, $w_0 = 1/2$ and $w_b = \alpha_b/2$ are weights for each interval, and α_b is the nominal level of the b^{th} interval. IS_α is the interval score (IS), a proper scoring rule for a single interval as defined in (4.18). The goal of the forecaster is to minimize the WIS. The LWIS is the same as the WIS except that it is evaluated over the log of quantiles and the log of the observed value.

$$WIS_{0,B}(Q, y^*) = \frac{1}{B + 1/2} \times (w_0 \times |y^* - median| + \sum_{b=1}^{B} \{w_k \times IS_{\alpha_b}(Q, y^*)\}) \tag{2.10}$$

$$IS_\alpha(l, r; y^*) = (r - l) + \frac{2}{\alpha}(l - y^*)\mathbb{1}\{y^* < l\} + \frac{2}{\alpha}(y^* - r)\mathbb{1}\{y^* > r\} \tag{2.11}$$

We fit 24 forecast models for each location for all 29 weeks, and for each week forecast 1-4 week ahead hospitalizations. The 24 models included all combinations of ASG, ASGD, SIR, and SIRD ILI models, the NORM, LNORM, and LST hospitalization models, and both quadratic and linear hospitalization models. The prior distributions under the SIR model are in (2.12). Because we set $S_{0s} = 0.9$, prior distributions were assigned only to I_{0s}, β_s and ρ_s, recalling the parameter

for the recovery rate $\delta_s = \rho_s \beta_s$. Here $\mathbb{1}_A$ represents the indicator function for values within the set A.

$$\begin{aligned} I_{0s} &\sim N(0.005, 0.03)\mathbb{1}_{(0,.1)} \\ \beta_s &\sim N^+(0.8, 0.3) \\ \rho_s &\sim N^+(0.68, 0.08) \end{aligned} \tag{2.12}$$

For the ASG model, the MLE $\hat{\theta}_s = (\hat{\lambda}_s, \hat{\eta}_s, \hat{\mu}_s, \hat{\sigma}^2_{1s}, \hat{\sigma}^2_{2s})$ was calculated and $\hat{\lambda}_s$ was accepted as a fixed value. The remaining estimates were used as starting values for posterior sampling. The transformation $T(\theta_s) = (\log(\eta_s), \mu_s, \log(\sigma^2_{1s}), \log(\sigma^2_{2s}))$ was made, and a prior distribution was assigned to $T(\theta_s)$. The prior distributions for the ILI model under the ASG function are shown in (2.13). These are slightly informative priors because for most parameters we have an idea what reasonable values may be. Here $m = (0.3, 23, 3.69, 4.7)$ and $C = \text{diag}(0.2, 5, 2, 2)$ where $\text{diag}(\cdot)$ is the diagonal matrix for the given entries.

$$\begin{aligned} T(\theta_s) &\overset{ind}{\sim} MVN(\theta, \Sigma) \\ T(\theta) &\overset{ind}{\sim} MVN(m, C) \\ \Sigma &= \text{diag}(\zeta_1^2, ..., \zeta_4^2) \\ \zeta_i &\overset{ind}{\sim} N^+(0, 4^2) \end{aligned} \tag{2.13}$$

Parameters shared by both the SIR and ASG models are the scale parameter κ_s and the discrepancy parameters σ^2_γ and $\sigma^2_{\gamma_W}$. These priors are in (2.14).

$$\begin{aligned} \kappa_s &\overset{ind}{\sim} N^+(0, 10,000^2) \\ \sigma^2_\gamma &\sim N^+(0, .02^2) \\ \sigma^2_{\gamma_W} &\sim N^+(\hat{\sigma}^2_W, 1^2) \end{aligned} \tag{2.14}$$

Because of the limited information for estimating $\sigma^2_{\gamma_W}$, we selected an informative prior distribution by first estimating $\hat{\sigma}^2_{\gamma_W}$. This was estimated for a given state by first calculating the MLE for θ_s. $\widehat{ILI}_{s,w}$ was then predicted such that $\text{logit}(\widehat{ILI}_{s,W}) = f_{\hat{\theta}_s}(W)$ for each season. Then

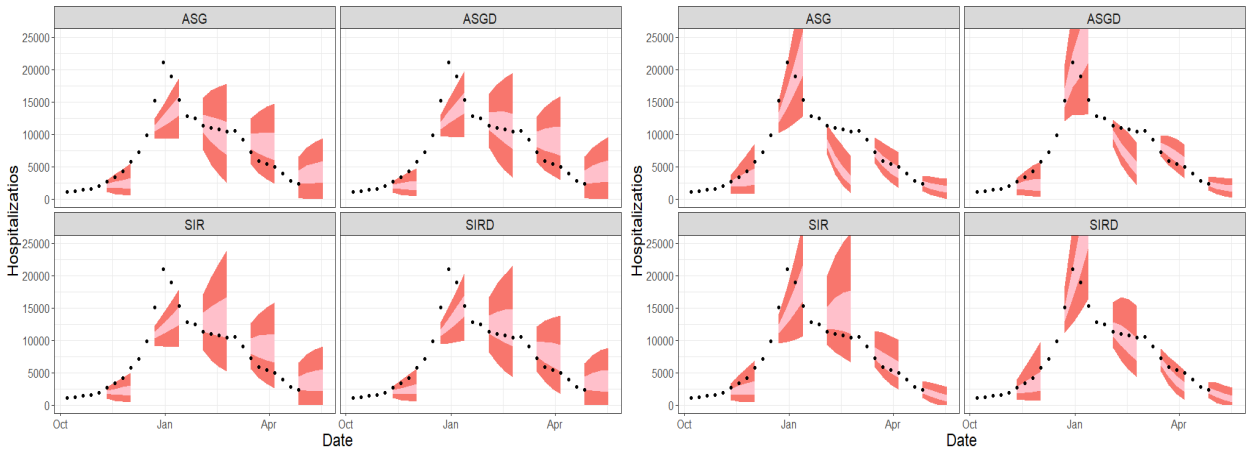

Figure 2.15: Forecasts 1-4 weeks ahead for US hospitalizations during the 2023 season for weeks 14, 20, 26, and 32. Forecasts are separated by ILI model, and the hospitalization models are all normally distribution. The figure includes hospitalization forecasts where ILI is a linear predictor (left) and where ILI is a quadratic predictor (right). 50% predictive intervals are pink and 95% predictive intervals are red.

$\hat{\sigma}^2_{\gamma_W}$ was calculated as the estimated variance over seasons of the differences $\text{logit}(\widehat{ILI}_{s,w}) - \text{logit}(ILI_{s,w})$. When $f_{\theta_s}(w)$ was the SIR function, $\widehat{\theta}_s$ was calculated using the `mle2` function in the `bblme` package (Bolker and R Development Core Team, 2023) and the `ode` function in the `deSolve` package (Karline Soetaert et al., 2010) function in `R`. Where the ASG function was used, $\hat{\theta}_s$ was calculated using the `optim` function. The prior for the hospitalization models are in (2.15).

$$\begin{aligned}
\alpha_{0s} &\overset{ind}{\sim} N(0, 5^2) \\
\alpha_{1s} &\overset{ind}{\sim} N(0, 5^2) \\
\alpha_{2s} &\overset{ind}{\sim} N(0, 5^2) \\
\phi &\sim N(0, .4)\mathbb{1}_{(-1,1)} \\
\sigma_{\epsilon_s} &\overset{ind}{\sim} N^+(0, 4^2) \\
\omega_s &\overset{ind}{\sim} N^+(0, 15^2)
\end{aligned} \tag{2.15}$$

Figure 2.15 shows 1-4-week ahead forecasts for US flu hospitalizations during the 2023 season under the NORM hospitalization model. The forecasts shown are for weeks 14, 20, 26, and 32. The predictive bands are the 50% and 95% predictive intervals. These plots show a tendency to

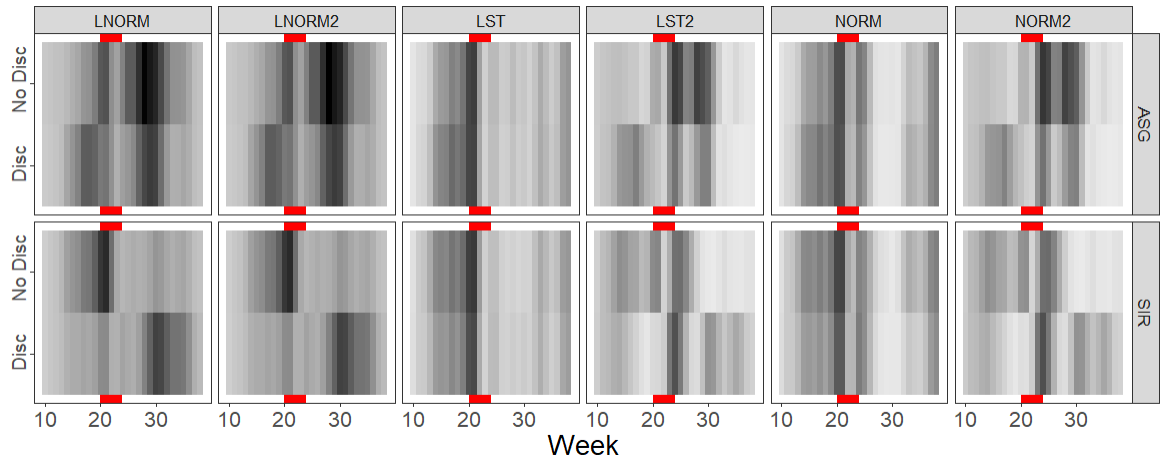

Figure 2.16: Each plot shows the log weighted interval score (LWIS) for every week of the 2023 flu season with scores for models including discrepancy in the ILI model (top) and excluding discrepancy (bottom). Scores are separated by hospitalization distribution family and by ILI as a linear or quadratic predictor. Scores for models with an ASG ILI model are above while those with an SIR model are below. The lighter the shade, the lower the LWIS with low LWIS being better.

often underpredict hospitalizations. The quadratic models appear to do a better job predicting hospitalizations at the season peak but a poorer job predicting after the peak, whereas the linear models seem to predict well or slightly overpredict after the peak.

Figure 2.16 shows model performance by LWIS for each week of the season for all 24 models of US hospitalizations. LWIS scores are grouped by ILI model, hospitalization model distribution, and by the linear or quadratic modeling. Here, a darker value represents a larger LWIS and worse forecast compared to lighter values. The weeks around the holiday week 22 are highlighted by a red band. In most cases, there appears to be a turning point in performance at or near week 22. The models which include discrepancy appear to forecast better around week 22, though they may not outperform the models without discrepancy through the whole season. Such is the case for the NORM and LST quadratic models and the linear LNORM models. It's also notable that for the SIR model, including discrepancy appears to make for poorer forecasts after week 22 whereas for ASG, including discrepancy appears to improve forecasts.

Figures 2.17 and 2.18 also show weekly forecast performance by LWIS, but all 53 locations are included. Overall, the weeks leading up to week 22 are the most difficult to forecast. In figure

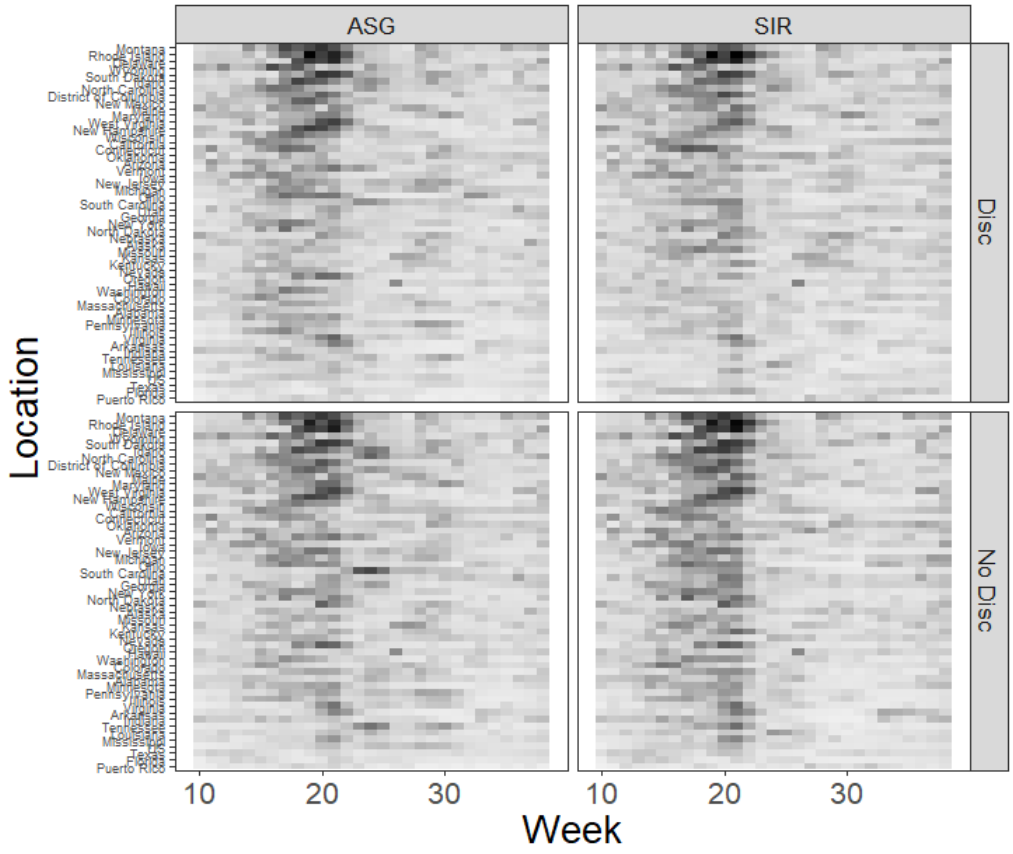

Figure 2.17: Each plot shows the log weighted interval scores (LWIS) for all 50 US states, PR, DC, and national level forecasts at each week during the 2023 flu season. Scores are averaged over all horizons 1-4 weeks ahead. Scores are faceted by ILI model function (columns) and by whether or not discrepancy modeling was included (rows). The lighter the shade, the lower the LWIS with low LWIS being better.

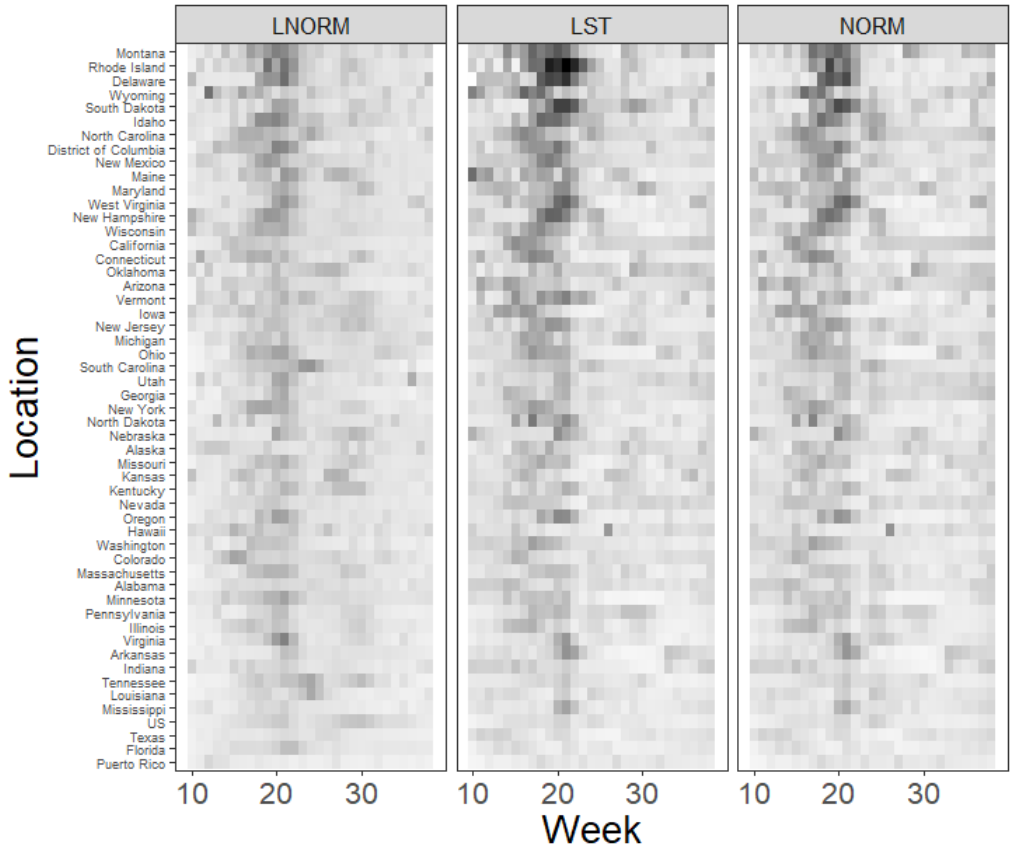

Figure 2.18: Each plot shows the log weighted interval scores (LWIS) for all 50 US states, PR, DC, and national level forecasts at each week during the 2023 flu season. Scores are averaged over all horizons 1-4 weeks ahead. Scores are faceted by hospitalization model distribution. The lighter the shade, the lower the LWIS with low LWIS being better.

Table 2.3: Overall scores for each of the 24 forecast models. The overall score is the log weighted interval score (LWIS) averaged over all locations, weeks, and horizons. The scores in the first two rows are for linear models, and the scores in the third and fourth rows are for quadratic models. The lowest WISs are bolded.

		ASG			SIR		
		LNORM	LST	NORM	LNORM	LST	NORM
Linear	No Disc	0.366	0.392	0.365	**0.355**	0.416	0.394
	Disc	0.358	0.391	0.365	**0.343**	0.382	0.354
Quadratic	No Disc	0.366	0.398	0.377	**0.355**	0.401	0.377
	Disc	0.358	0.390	0.369	**0.343**	0.378	0.356

2.17, it appears that both the SIR and ASG models which include discrepancy perform slightly better than the models which do not. Figure 2.18 shows the LWIS across the whole season faceted by hospitalization model distribution. The LNORM model appears to perform the best during the weeks leading up to week 22. The season overall score, calculated as the mean LWIS over all locations and weeks, for all 24 models is shown in table 2.3. The first main takeaway is that the top four performing models are SIR ILI models with the LNORM hospitalization models. The next main takeaway is that overall the models which include discrepancy outperformed the models without discrepancy.

It should not be assumed that these two takeaways will apply for future flu seasons. As suggested by the simulation study in the previous section, it may often be the case that the ASG ILI model is better for forecasting. Indeed, a close examination of figures 2.16, 2.17, and 2.18 shows that it is common for the ASG models to show better forecasting skill than the SIR models.

2.6 Conclusion

In this manuscript we introduce a statistical modeling framework which allows for the incorporation of several ILI forecast modeling methods. Specifically, we built upon Osthus et al. (2019) and introduced a framework for modeling ILI which includes the use of an arbitrary function for modeling the main trajectory of ILI along with modeling the discrepancy. We model

flu hospitalizations by incorporating the ILI forecast model into a model forecasting hospitalizations where hospitalization predictions are a linear or quadratic function of ILI.

The simulation study in section 2.4 suggests the ASG function in ILI modeling may slightly outperform the SIR model according to the LogS and CRPS scoring rules, but in the analysis of the 2023 flu season forecasts, the SIR model was overall superior. The results from both the simulation study and the real data analysis suggest that the addition of a discrepancy component in ILI modeling may improve forecasts especially near the holiday week between Christmas and New Years day. It should not be assumed that these conclusions may be generalized for all locations in the US, weeks of a season, or for future flu seasons.

Forecasting the seasonal influenza outbreak remains a challenging task for forecasters. The general modeling framework in this manuscript is successful under diverse modeling conditions for all locations in the US and may contribute to future forecasting efforts. All forecast models were presented as separate from one another, but in the various locations and times of the flu season, different models perform better than others. To build on the work done here, a natural step forward would be to combine all or a few selected forecasts into an ensemble forecast. Such an ensemble may work to cancel out certain model biases or highlight model strengths, leading to more robust forecasts.

2.7 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.

2.8 References

Allen, L. J. (2017). A primer on stochastic epidemic models: Formulation, numerical simulation, and analysis. *Infectious Disease Modelling*, 2(2):128–142.

Allen, L. J., Brauer, F., Van den Driessche, P., and Wu, J. (2008). *Mathematical Epidemiology*, volume 1945. Springer.

Arendt, P. D., Apley, D. W., Chen, W., Lamb, D., and Gorsich, D. (2012). Improving identifiability in model calibration using multiple responses. *Journal of Mechanical Design*, 134(10):100909.

Atkinson, P. M., Jeganathan, C., Dash, J., and Atzberger, C. (2012). Inter-comparison of four models for smoothing satellite sensor time-series data to estimate vegetation phenology. *Remote Sensing of Environment*, 123:400–417.

Beck, P. S., Atzberger, C., Høgda, K. A., Johansen, B., and Skidmore, A. K. (2006). Improved monitoring of vegetation dynamics at very high latitudes: A new method using MODIS NDVI. *Remote Sensing of Environment*, 100(3):321–334.

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.

Bolker, B. and R Development Core Team (2023). *bbmle: Tools for General Maximum Likelihood Estimation*. R package version 1.0.25.1.

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.

Brynjarsdóttir, J. and O'Hagan, A. (2014). Learning about physical parameters: The importance of model discrepancy. *Inverse Problems*, 30(11):114007.

CDC (2024a). Centers for Disease Control and Prevention FluSight: About flu forecasting. https://www.cdc.gov/flu-forecasting/about/index.html?CDC_AAref_Val=https://www.cdc.gov/flu/weekly/flusight/how-flu-forecasting.htm. Accessed: 2024-10-22.

CDC (2024b). Centers for Disease Control and Prevention FluView portal. https://gis.cdc.gov/grasp/fluview/fluportaldashboard.html. Accessed: 2024-10-22.

CDC (2024c). Centers for Disease Control and Prevention FluView, U.S. influenza surveillance: Purpose and methods. https://www.cdc.gov/fluview/overview/?CDC_AAref_Val=https://www.cdc.gov/flu/weekly/overview.htm. Accessed: 2024-10-22.

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. (2022a). The United States COVID-19 forecast hub dataset. *Scientific Data*, 9(1):462.

Cramer, E. Y., Ray, E. L., Lopez, V. K., Bracher, J., Brennen, A., Castro Rivadeneira, A. J., Gerding, A., Gneiting, T., House, K. H., Huang, Y., et al. (2022b). Evaluation of individual and ensemble probabilistic forecasts of COVID-19 mortality in the United States. *Proceedings of the National Academy of Sciences*, 119(15):e2113561119.

Ewing, A., Lee, E. C., Viboud, C., and Bansal, S. (2017). Contact, travel, and transmission: the impact of winter holidays on influenza dynamics in the United States. *The Journal of Infectious Diseases*, 215(5):732–739.

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

Garza, R. C., Basurto-Dávila, R., Ortega-Sanchez, I. R., Carlino, L. O., Meltzer, M. I., Albalak, R., Balbuena, K., Orellano, P., Widdowson, M.-A., and Averhoff, F. (2013). Effect of winter school breaks on influenza-like illness, Argentina, 2005–2008. *Emerging Infectious Diseases*, 19(6):938.

Gelman, A. (2006). Prior distributions for variance parameters in hierarchical models. *Bayesian Analysis*, 1(3):515–533.

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.

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.

HealthData.gov (2024). COVID-19 reported patient impact and hospital capacity by state (raw). https://healthdata.gov/dataset/COVID-19-Reported-Patient-Impact-and-Hospital-Capa/6xf2-c3ie/about_data. Accessed: 2024-10-22.

Hird, J. N. and McDermid, G. J. (2009). Noise reduction of NDVI time series: An empirical comparison of selected techniques. *Remote Sensing of Environment*, 113(1):248–258.

Hoffman, M. D., Gelman, A., et al. (2014). The No-U-Turn sampler: adaptively setting path lengths in Hamiltonian Monte Carlo. *Journal of Machine Learning Research*, 15(1):1593–1623.

Jiang, Z., Apley, D. W., and Chen, W. (2015). Surrogate preposterior analyses for predicting and enhancing identifiability in model calibration. *International Journal for Uncertainty Quantification*, 5(4).

Jonsson, P. and Eklundh, L. (2002). Seasonality extraction by function fitting to time-series of satellite sensor data. *IEEE ransactions on Geoscience and Remote Sensing*, 40(8):1824–1832.

Jordan, A., Krüger, F., and Lerch, S. (2019). Evaluating probabilistic forecasts with scoringRules. *Journal of Statistical Software*, 90(12):1–37.

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.

Karline Soetaert, Thomas Petzoldt, and R. Woodrow Setzer (2010). Solving differential equations in R: Package deSolve. *Journal of Statistical Software*, 33(9):1–25.

Kennedy, M. C. and O'Hagan, A. (2001). Bayesian calibration of computer models. *Journal of the Royal Statistical Society: Series B Statistical Methodology*, 63(3):425–464.

Kermack, W. O. and McKendrick, A. G. (1927). A contribution to the mathematical theory of epidemics. *Proceedings of the Royal Society of London. Series A, Containing papers of a mathematical and physical character*, 115(772):700–721.

Laio, F. and Tamea, S. (2007). Verification tools for probabilistic forecasts of continuous hydrological variables. *Hydrology and Earth System Sciences*, 11(4):1267–1277.

Lewis-Beck, C., Walker, V. A., Niemi, J., Caragea, P., and Hornbuckle, B. K. (2020). Extracting agronomic information from SMOS vegetation optical depth in the US Corn Belt using a nonlinear hierarchical model. *Remote Sensing*, 12(5):827.

Linzer, D. A. (2013). Dynamic Bayesian forecasting of presidential elections in the states. *Journal of the American Statistical Association*, 108(501):124–134.

Liu, F., Bayarri, M., Berger, J., et al. (2009). Modularization in Bayesian analysis, with emphasis on analysis of computer models. *Bayesian Analysis*, 4(1):119–150.

Ljung, L. (1987). *System Identification: Theory for the User*. Prentice-Hall, Englewood Cliffs, NJ.

Lutz, C. S., Huynh, M. P., Schroeder, M., Anyatonwu, S., Dahlgren, F. S., Danyluk, G., Fernandez, D., Greene, S. K., Kipshidze, N., Liu, L., et al. (2019). Applying infectious disease forecasting to public health: a path forward using influenza forecasting examples. *BMC Public Health*, 19(1):1–12.

Ma, P., Karagiannis, G., Konomi, B. A., Asher, T. G., Toro, G. R., and Cox, A. T. (2022). Multifidelity computer model emulation with high-dimensional output: An application to storm surge. *Journal of the Royal Statistical Society Series C: Applied Statistics*, 71(4):861–883.

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.

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

McGowan, C. J., Biggerstaff, M., Johansson, M., Apfeldorf, K. M., Ben-Nun, M., Brooks, L., Convertino, M., Erraguntla, M., Farrow, D. C., Freeze, J., et al. (2019). Collaborative efforts to forecast seasonal influenza in the United States, 2015–2016. *Scientific Reports*, 9(1):683.

Molinari, N.-A. M., Ortega-Sanchez, I. R., Messonnier, M. L., Thompson, W. W., Wortley, P. M., Weintraub, E., and Bridges, C. B. (2007). The annual impact of seasonal influenza in the US: measuring disease burden and costs. *Vaccine*, 25(27):5086–5096.

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.

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.

Ray, E. L., Wattanachit, N., Niemi, J., Kanji, A. H., House, K., Cramer, E. Y., Bracher, J., Zheng, A., Yamana, T. K., Xiong, X., et al. (2020). Ensemble forecasts of coronavirus disease 2019 (COVID-19) in the US. *medRxiv*, pages 2020–08.

Rudis, B. (2021). *cdcfluview: Retrieve Flu Season Data from the United States Centers for Disease Control and Prevention ('CDC') 'FluView' Portal*. R package version 0.9.4.

Simon, C. M. (2020). The SIR dynamic model of infectious disease transmission and its analogy with chemical kinetics. *PeerJ Physical Chemistry*, 2:e14.

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

Tsyplakov, A. (2013). Evaluation of probabilistic forecasts: proper scoring rules and moments. *Available at SSRN 2236605*.

Turtle, J., Riley, P., Ben-Nun, M., and Riley, S. (2021). Accurate influenza forecasts using type-specific incidence data for small geographic units. *PLOS Computational Biology*, 17(7):e1009230.

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

Van den Driessche, P. (2008). Deterministic compartmental models: extensions of basic models. In *Mathematical Epidemiology*, pages 147–157. Springer.

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.

Wallis, K. F. (2014). The two-piece normal, binormal, or double Gaussian distribution: its origin and rediscoveries. *Statistical Science*, pages 106–112.

WHO (2024). World Health Organization website influenza (seasonal) fact sheet. https://www.who.int/news-room/fact-sheets/detail/influenza-(seasonal). Accessed: 2024-10-22.

Winkler, R. L. (1971). Probabilistic prediction: Some experimental results. *Journal of the American Statistical Association*, 66(336):675–685.

Yamana, T. K., Kandula, S., and Shaman, J. (2016). Superensemble forecasts of Dengue outbreaks. *Journal of The Royal Society Interface*, 13(123):20160410.

2.A FluSight forecast competition scoring results

In the CDC flu forecast competition, a baseline forecast was made to which all other forecasts were compared. The baseline forecast for a given state and week had as a median the most recent observed hospitalization count. The uncertainty was based on differences between previous hospitalizations, and is similar to the baseline forecast used in previous flu forecasting seasons and in the COVID-19 hub (Mathis et al., 2024; Cramer et al., 2022b). The RWIS for one model was calculated by first taking the ratio of the average LWIS paired with every other model. This was then diveded by the same ratio for the baseline forecasts (see methods in Mathis et al. (2024) for

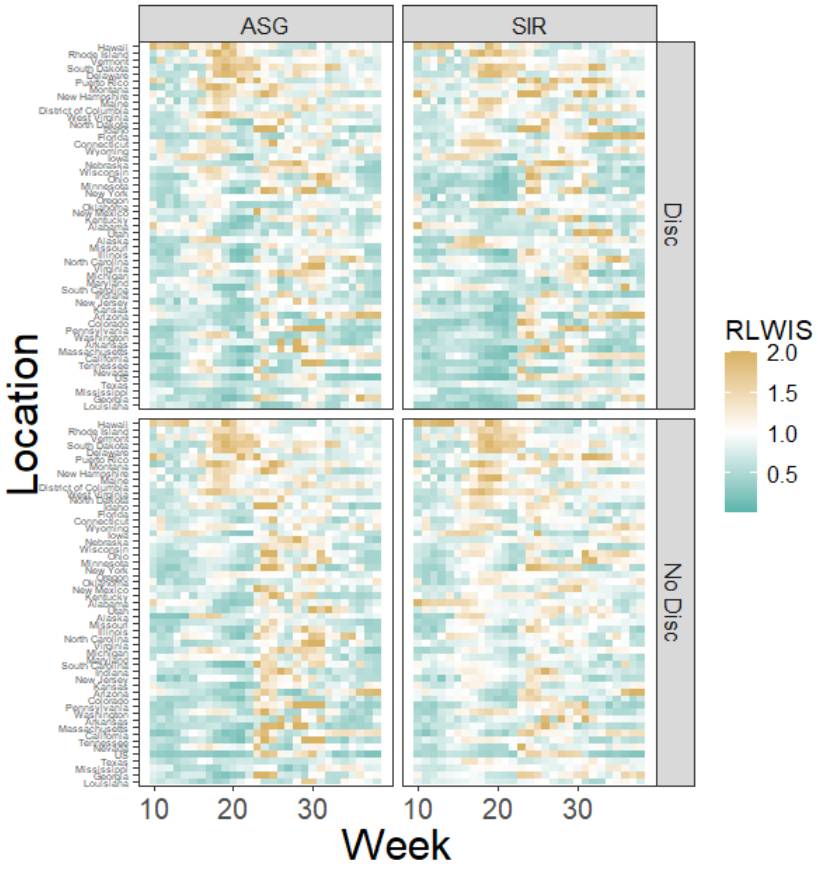

Figure 19: RWIS averaged over all 24 models and horizons for forecasts across all locations and weeks. The plot is separated by the four ILI models. Dark blue is lower RWIS, and dark tan is higher RWIS where white is RWIS = 1

details, and see also the supplementary material). An RWIS less than 1 indicates the forecast outperformed the baseline forecast. Figures 19 and 20 show the mean RWIS of hospitalization forecasts over all 24 models from section 2.5 for all locations and weeks of the 2023 flu season. Some similar patterns to those noted in section 2.5 emerge, including a slightly better performance by forecasts including discrepancy modeling over those which do not include discrepancy modeling and better overall performance by the SIR models than by the ASG forecast models. Interestingly, figure 20 shows the LNORM model showing especially poor RWIS performance about three quarters into the season. This was not seen in the LWIS in figure 2.18.

Figure 21 shows the mean over weeks RWIS for the SIR and ASG forecast models with an RWIS less than 1 for the most locations. The plot on the left is from a forecast by an SIRD ILI

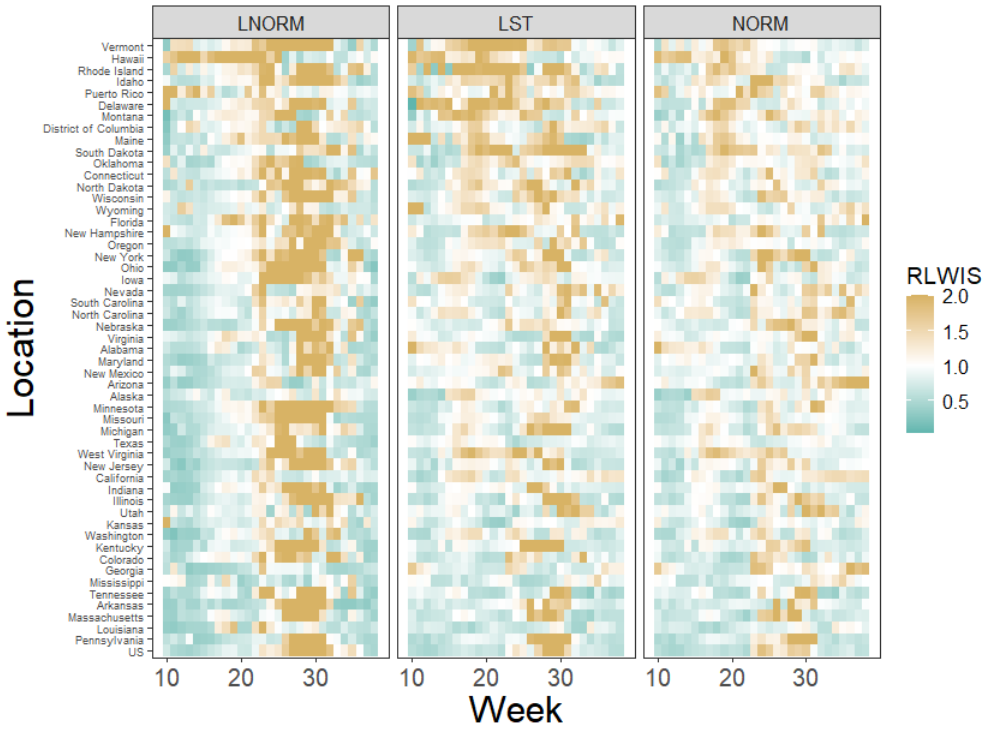

Figure 20: RWIS averaged over all 24 models and horizons for forecasts across all locations and weeks. The plot is separated by the three distribution choices for the hospitalization model. Dark blue is lower RWIS, and dark tan is higher RWIS where white is RWIS = 1

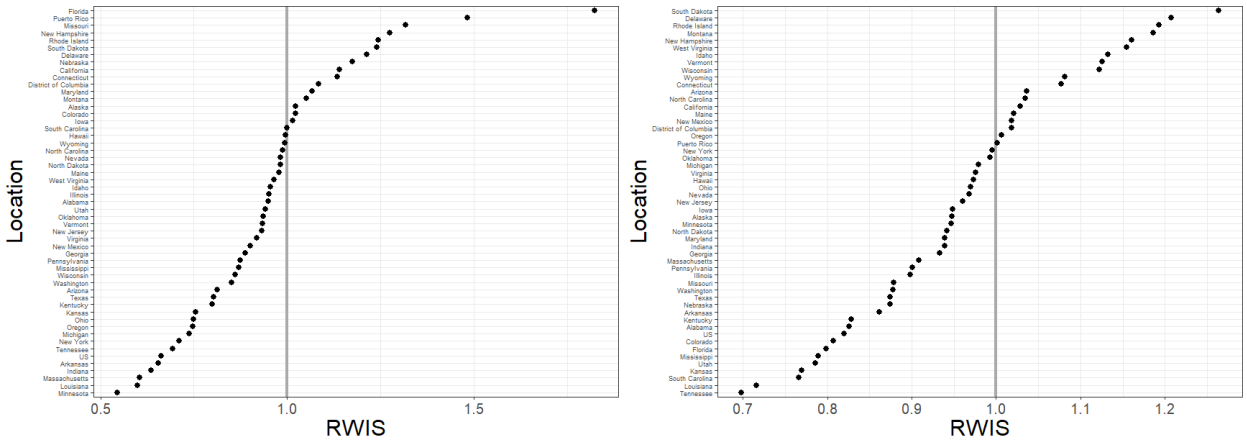

Figure 21: RWIS over the whole 2023 flu season for all 53 locations for hospitalization forecasts from an SIRD quadratic hospitalization model with normal errors (left) and forecasts from an ASGD linear hospitalization model with normal errors (right). An RWIS less than 1 (left of vertical grey line) indicates the model forecasts outperformed the baseline forecasts for that particular region.

model and a quadratic NORM hospitalization model. This forecast model had an RWIS less than 1 for 37 out of 53 locations. The plot on the right is from a forecast by an ASGD ILI model and linear NORM hospitalization model. This model had an RWIS less than 1 for 34 out of 53 locations.

2.B Posterior distribution plots for select parameters

The figures in this section show 95% credible intervals for model parameters under the several modeling schemes along with the prior distributions assigned to the parameters. Figure 22 is for parameters unique to the SIR ILI model. Figures 23 and 24 are for parameters unique to the ASG models. Figure 25 is for parameters shared by SIR and ASG models, including parameters used for modeling discrepancy. Figures 26 and 27 are for parameters used in hospitalization modeling.

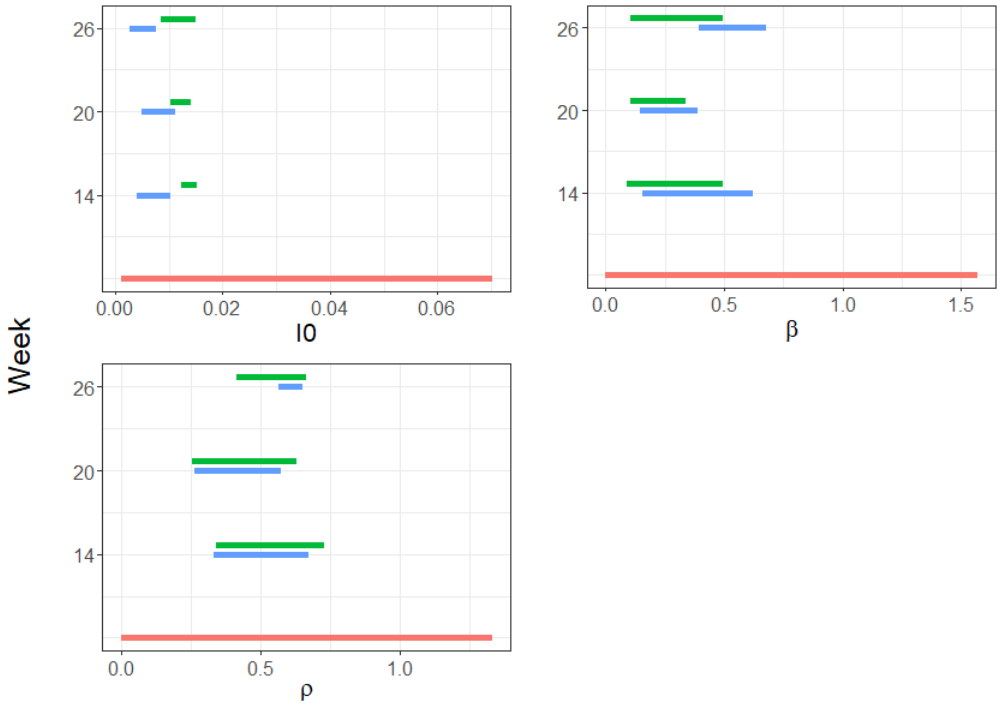

Figure 22: Posterior 95% credible intervals from ILI model for parameters of SIR differential equations. Shown are intervals from the US model of ILI for weeks 14, 20, and 26. The green is from the posterior where discrepancy is not modeled, and the blue is from the model where it is. The red interval is the 95% interval of the prior distribution.

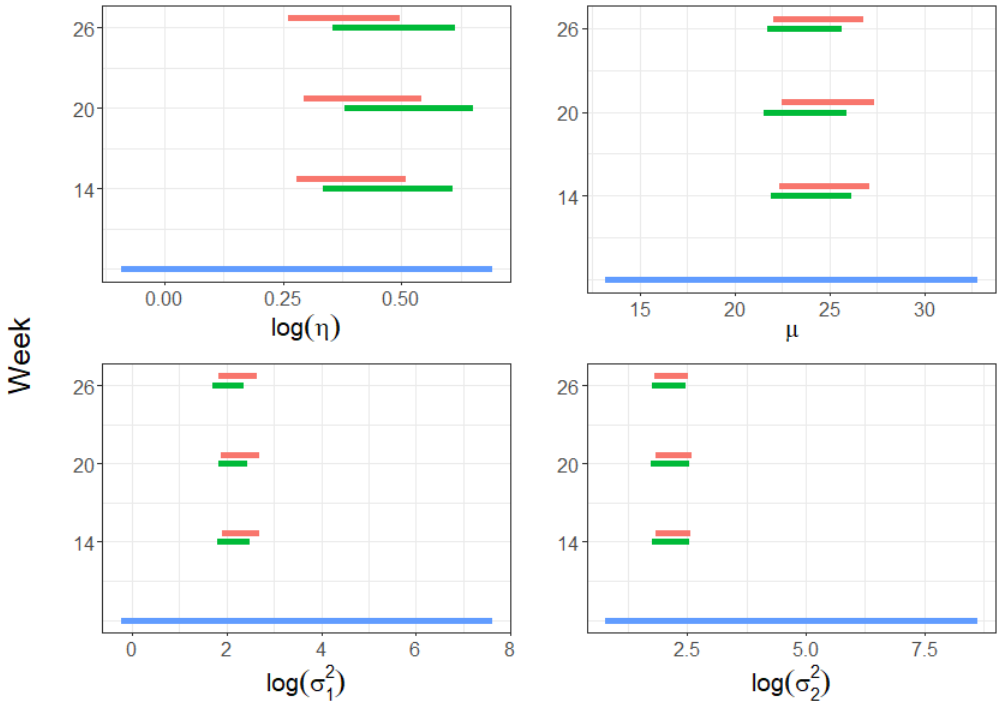

Figure 23: Posterior 95% credible intervals from ILI model for parameters of ASG function. Shown are intervals from the US model of ILI for weeks 14, 20, and 26. The red is from the posterior where discrepancy is not modeled, and the green is from the model where it is. The blue interval is the 95% interval of the prior distribution.

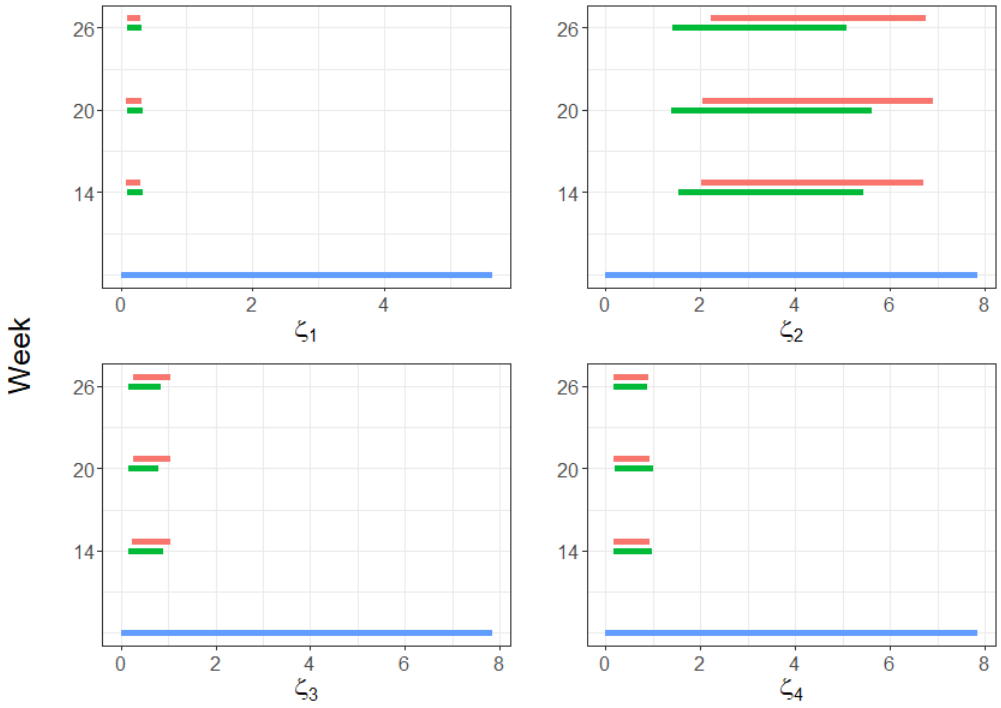

Figure 24: Posterior 95% credible intervals from ILI model for variance parameters of the ASG function. Shown are intervals from the US model of ILI for weeks 14, 20, and 26. The red is from the posterior where discrepancy is not modeled, and the green is from the model where it is. The blue interval is the 95% interval of the prior distribution.

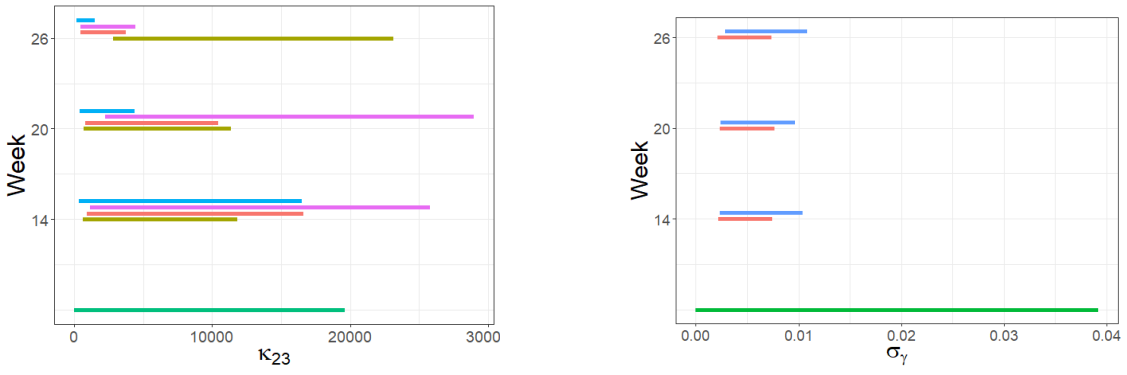

Figure 25: Posterior 95% credible intervals from ILI model for scale parameter κ_s. Blue is from the SIR model, purple from SIRD, red from ASG and yellow from ASGD (left). Posterior 95% credible intervals from ILI model for scale parameter of modeled discrepancy σ_γ (right). Shown are intervals from the US model of ILI for weeks 14, 20, and 26. Blue is from the SIRD model and red from ASGD. The green interval is the 95% interval of the prior distribution.

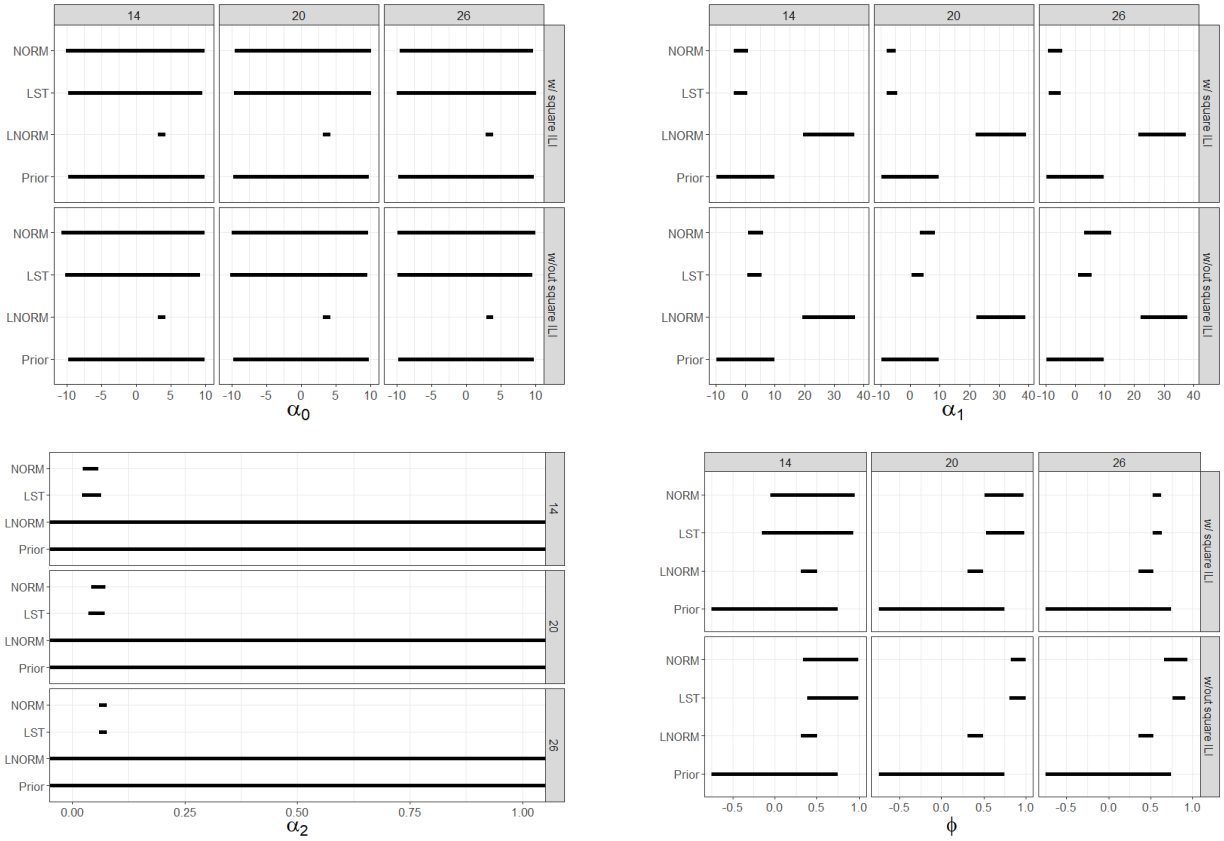

Figure 26: Posterior 95% credible intervals for α_0 for the three hospitalization models separated by whether or not the squared ILI term was included. Intervals are for the US hospitalizations and weeks 14, 20 and 26 are shown (top left). Prior distribution 95% interval is also included. 95% posterior credible intervals for α_1 for the three hospitalization models separated by whether or not the squared ILI term was included. Intervals are for the US hospitalizations and weeks 14, 20 and 26 are shown (top right). Prior distribution 95% interval is also included. 95% posterior credible intervals for α_2 for the three hospitalization models. Intervals are for the US hospitalizations and weeks 14, 20 and 26 are shown (bottom left). Prior distribution 95% interval is also included. 95% posterior credible intervals for ϕ for the three hospitalization models separated by whether or not the squared ILI term was included (bottom right). Prior distribution 95% interval is also included. In each plot intervals are for the US hospitalizations and weeks 14, 20 and 26 are shown.

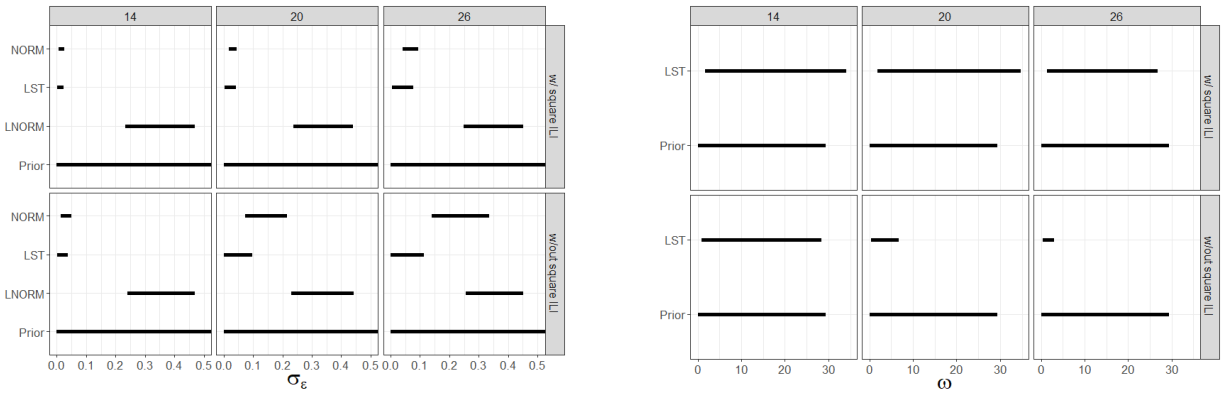

Figure 27: Posterior 95% credible intervals for σ_ϵ for hospitalization models with and without the ILI squared term. Intervals are for the US hospitalizations and weeks 14, 20 and 26 are shown (left). Prior distribution 95% interval is also included. 95% posterior credible intervals for ω for LST hospitalization models with and without the ILI squared term (right). Prior distribution 95% interval is also included. In all plots intervals are for the US hospitalizations and weeks 14, 20 and 26 are shown.

CHAPTER 3. QUANTILE FORECAST MATCHING WITH A BAYESIAN QUANTILE GAUSSIAN PROCESS MODEL

Spencer Wadsworth and Jarad Niemi

Department of Statistics, Iowa State University, Ames, IA 50011

3.1 Abstract

For various reasons a relatively small number of probabilities along with corresponding quantiles are often used to define predictive distributions or probabilistic forecasts. These quantile predictions offer easily interpreted uncertainty quantification of an event, and quantiles are generally straightforward to estimate using standard statistical and machine learning methods. However, compared to a distribution defined by a probability density or cumulative distribution function, a set of quantiles has far less distributional information. When given estimated quantiles it may be desirable to estimate a fully defined continuous distribution function, and many researchers do so to make evaluation or ensemble modeling simpler. Most existing methods for fitting a distribution to quantiles lack accurate representation of the inherent uncertainty from quantile estimation or are limited in their applications. In this manuscript we present a Gaussian process model, the quantile Gaussian process, which is based on established theory of quantile functions and sample quantiles, to construct a probability distribution given estimated quantiles. A Bayesian application of the quantile Gaussian process is evaluated for parameter inference and distribution approximation in simulation studies, and the quantile Gaussian process is used to approximate the distributions of quantile forecasts from the 2023-24 US Centers for Disease Control collaborative flu forecasting competition. The simulation studies and data analysis show that the quantile Gaussian process leads to accurate inference on model parameters, estimation of a continuous distribution, and uncertainty quantification of sample quantiles.

***K*eywords** Sample quantiles · Quantile regression · Probabilistic forecasting · Disease outbreaks

3.2 Introduction

The use of quantiles in statistical modeling and in reporting inferential or predictive uncertainty is widespread, and reporting several quantiles or predictive intervals is common for probabilistic forecasting (Gneiting et al., 2023). Quantiles can be easier to interpret than statistical model parameters and are often used to define confidence or prediction intervals. Thus with multiple quantiles for a particular outcome, one has a measure of uncertainty. Estimating quantiles in the presence of covariates via quantile regression is a common statistical and machine learning strategy, and where parametric models are complicated or nonexistent quantiles may be easier to estimate via machine learning methods (Martin and Syring, 2022; Chung et al., 2021; Koenker, 2017; Koenker and Bassett Jr, 1978). In estimating multiple quantiles, quantile regression often provides a tradeoff with parametric modeling in that for some problems quantile regression may be easier to perform but that predicted quantiles lack the detailed information of a fully defined predictive distribution (Pohle, 2020). Quantile regression may also lead to other issues such as quantile crossing where an estimated quantile may be smaller than another estimated quantile with for a lower probability level (Chernozhukov et al., 2010; He, 1997). Perhaps, for data privacy reasons, data quantiles are often reported as summaries of the data. Census or medical data, which can be very large or personal to the subjects, may be published as summary or aggregate data including percentiles (quantiles) and medians (Simpson et al., 2023; CDC, 2022; Nirwan and Bertschinger, 2020). In collaborative forecast initiatives and competitions, probabilistic forecasts are often submitted as a set of predictive quantiles for given probabilities (Gneiting et al., 2023; Hong et al., 2016). In recent years disease outbreak forecast hubs require that all forecasts submitted by outside participants be represented by several predictive intervals or quantiles. This standardized representation allows for straightforward forecast scoring and ensemble building (Mathis et al., 2024; Github, 2024; Cramer et al., 2022a,b; Sherratt et al., 2023; Bracher et al., 2021).

A set of quantiles may provide distributional information for an event, but it is not as informative as a distribution defined by a cumulative distribution function (CDF), the inverse-CDF or quantile function QF, or a probability density/mass function (PDF), and quantiles provide no distribution tail information –information for values below the smallest quantile or above the largest quantile. Another drawback for using quantiles to define a distribution is that many tools for evaluating or scoring continuous distributions require a CDF or PDF (Gneiting and Raftery, 2007; Gneiting and Katzfuss, 2014). Combining distributions into an ensemble distribution is commonly done by aggregating multiple QFs or CDFs/PDFs, the latter being possible only when a CDF/PDF is available (Gneiting et al., 2005; Wang et al., 2023).

Fitting quantiles to recover or estimate a continuous distribution is done in many fields, often for the purpose of evaluating forecasts using rules that require a CDF or PDF (Simpson et al., 2023; Gerding et al., 2023). Fitting may also be done to be able to combine forecasts using aggregation methods that require CDFs or PDFs (Gyamerah et al., 2020; Li et al., 2019; Baran and Lerch, 2018; Bogner et al., 2017; He et al., 2016; Gneiting et al., 2005). An example of fitting quantiles is given in figure 3.1. The 12 points in the figure are quantiles estimated from a random sample of size 100 from a standard normal distribution for given probabilities. The grey line is the fitted QF of a normal distribution that was estimated by selecting the mean and standard deviation parameters which minimize the least squares distance between the estimated quantiles and the QF.

Hereafter we refer to estimating a continuous distribution by fitting quantiles as quantile matching (QM). Sgouropoulos et al. (2015) performed QM by minimizing the mean square difference between quantiles of a response variable and a linear combination of quantiles of covariates. Selecting distribution parameters which minimize the mean square error between quantiles and a QF is a common QM method (Dilger et al., 2022; Li et al., 2019; Belgorodski et al., 2017), kernel density estimation and spline interpolation are common nonparametric QM methods (Gerding et al., 2023; Gyamerah et al., 2020; He et al., 2016), and Keelin (2016) introduced a semiparametric method based on defining a flexible QF to fit quantiles. An issue

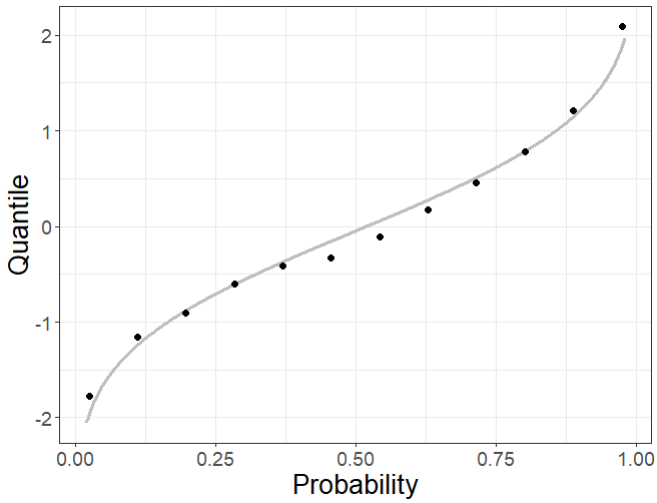

Figure 3.1: Points representing 12 quantiles (y-axis) for given probability values (x-axis) estimated from a random sample of size 100 from a standard normal distribution. The quantile function (QF) of normal distribution (grey) was fit by selecting the mean and standard deviation parameters which minimize the least squares distance between the quantiles and the QF.

shared by these methods is that they do not adequately account for the uncertainty inherent from the estimation of quantiles. Nirwan and Bertschinger (2020) introduced a QM model based on the definition of sample quantiles as order statistics and formulated the model likelihood as the joint PDF of multiple order statistics. This model provides exact inference for quantile uncertainty, but relies on the CDF and PDF of the distribution, making it less practical for fitting quantile defined distributions which have a QF but for which the CDF and PDF either do not exist in closed form or are difficult to evaluate (Perepolkin et al., 2023; Joiner and Rosenblatt, 1971; Tukey, 1960).

In this manuscript we introduce a novel model, the quantile Gaussian process (QGP), used specifically for QM. The context for using the QGP is that the available data is a set of quantiles estimated for a given set of probabilities. The QGP relies on established theory underlying the relationship between the QF of a continuous distribution and sample quantiles (Parzen, 2004; Gilchrist, 2000; Hyndman and Fan, 1996; Walker, 1968; Cramér, 1951). The QGP can provide accurate asymptotic uncertainty quantification on quantiles, and it works well for QM for quantile defined distributions. In section 3.3 the QF of a continuous random variable and sample quantiles are each defined, and important properties and asymptotic theory are reviewed. In section 3.4 the

QGP model for QM is introduced. Section 3.5 contains simulation studies illustrating the QGP's ability to make parameter inference and match the distribution from which the quantiles where estimated. Section 3.6 contains an application of QM using the QGP model for quantile forecasts of the 2023-24 United States Centers for Disease Control (CDC) flu forecasting competition, also known as FluSight. Section 3.7 concludes this manuscript with a short summary and ideas for further development in QM modeling.

3.3 Quantile function and sample quantiles

Along with the CDF and PDF, the QF is a defining function of a random variable. The QF, however, often receives less attention in standard statistical training than do the CDF and PDF (Parzen, 2004), and the CDF and PDF are more used often for modeling and inference. This section contains the definition of the QF, important properties, the definition of sample quantiles, and key central limit theorems (CLTs) of sample quantiles.

3.3.1 Quantile function definition and properties

The QF is defined in definition 3.1.

Definition 3.1. *For a cumulative distribution function $F : \mathbb{R} \to [0, 1]$, the quantile function $Q : [0, 1] \to \mathbb{R}$ is defined as*

$$Q(p) = F^{-1}(p) := \inf\{x \in \mathbb{R} : F(x) \geq p\}, \quad p \in [0, 1]$$

An alternative definition not in terms of the cumulative distribution function is a function $Q : [0, 1] \to \mathbb{R} \cup \{-\infty, \infty\}$, which is nondecreasing and left-continuous.

For most statistical applications, it is useful to define the QF as the inverse of the CDF. However, it is sometimes the case that defining a distribution by its QF first is advantageous, for instance when a distribution is defined by a QF but has no closed form CDF or when simple QFs may be aggregated to define more complex distributions (Perepolkin et al., 2023; Gasthaus et al.,

2019; Alvarez and Orestes, 2023). This section continues with important properties of the QF and functions including the location-scale property, the quantile density function, the probability integral transform, and examples of quantile defined distributions.

3.3.1.1 Location-scale property

For a random variable with PDF $f_0(x)$ and QF $Q_0(p)$ for $p \in [0,1]$, a location-scale family takes the PDF form $(1/\sigma)f_0((x-\mu)/\sigma)$ where the location parameter $\mu \in (-\infty, \infty)$ and the scale parameter $\sigma \in (0, \infty)$ (Casella and Berger, 2002). The QF of the location-scale family then takes the form $\mu + \sigma Q_0(p)$ (Parzen, 2004). By the addition and multiplication rules from section 3.2 of Gilchrist (2000), this is a valid QF. Gilchrist (2000) outlines the standardization rule which states that if the random variable Y with QF $Q_Y(p)$ has a standard distribution, that is the random variable Y is centered (by a mean or median) at 0 and has a scale of 1, then a random variable with quantile function $Q_{T(Y)}(p) = \mu + \sigma Q_Y(p)$ is centered at μ and has scale σ.

The linear form for a QF is already inherent in some distributions families such as the normal and logistic distribution families, with respective QFs $\mu + \sigma\Phi^{-1}(p)$ and $\mu + \sigma\log(p/(1-p))$. For the normal distribution, $\Phi^{-1}(p)$ is the QF of a standard normal distribution and μ and σ are the mean and standard deviation of the location-scale transformed normal distribution. For the logistic distribution $\log(p/(1-p))$ is the QF of a logistic distribution with mean 0 and scale 1, and μ and σ become the mean and scale parameters of the location-scale transformed logistic distribution. The linear form is convenient for quantile regression modeling and is key to the development of the metalog distribution family which was designed specifically for modeling quantiles (Keelin, 2016).

3.3.1.2 Quantile density function

Besides the CDF, PDF, and QF a fourth infrequently mentioned defining function of a continuous random variable is the quantile density function (QDF). As the PDF is the derivative

of the CDF, the QDF $q : [0, 1] \to \mathbb{R}$ is the derivative of the QF defined as

$$q(p) = \frac{dQ(p)}{dp}.$$

If $Q(p)$ is the inverse of a CDF $F(x)$, a calculus result shows that the reciprocal QDF $[q(p)]^{-1} = f(Q(p))$, or the PDF evaluated at the QF at p (Perepolkin et al., 2023; Gilchrist, 2000). The QDF is important for establishing the CLTs in section 3.3.2 and for the QGP models introduced in section 3.4.

3.3.1.3 Probability integral transform

An important application of the QF is using it to sample from continuous probability distributions via the probability integral transform (PIT). The PIT states that for any continuous random variable Y with CDF $F_Y(y)$, the transformed random variable $X = F_Y(Y)$ is uniformly distributed on $(0, 1)$, or $X \sim \text{Unif}(0, 1)$. Thus if one can sample from a standard uniform distribution, one may also sample from a continuous distribution provided the QF of that distribution can be evaluated or reasonably approximated (Wilkinson, 2018). The PIT is useful for assessing model fit via generalized residuals (Yang, 2024; Cox and Snell, 1968) and the calibrating of probabilistic forecasts (Gneiting et al., 2007). Herein the PIT is important for developing one version of the QGP model, and it is used as part of a distance measure for assessing QM.

3.3.1.4 Examples of quantile defined distributions

An example of a quantile defined distribution family that lacks a CDF in closed form is the generalized lambda distribution (GLD) family (Ramberg and Schmeiser, 1974; Perepolkin et al., 2023). A special case of the GLD family is the Tukey lambda distribution (TLD) family, a one parameter distribution family introduced by Tukey (1960), (Joiner and Rosenblatt, 1971). The QF for the TLD is in (3.1) and the QDF in (3.2).

$$Q_\lambda(p) = \begin{cases} \frac{1}{\lambda}\left[p^\lambda - (1-p)^\lambda\right], & \lambda \neq 0 \\ \log\left(\frac{p}{1-p}\right), & \lambda = 0 \end{cases} \tag{3.1}$$

$$q_\lambda(p) = p^{\lambda-1} + (1-p)^{\lambda-1} \tag{3.2}$$

Another distribution family defined by its quantile function but without a closed form CDF is the metalog distribution (MLD) family, a generalization of the logistic distribution family (Keelin, 2016). One version of the MLD is defined in (3.3).

$$Q_{\mathbf{a}}(p) = a_1 + a_3(p-0.5) + a_2 \log\left(\frac{p}{1-p}\right) \tag{3.3}$$

These two distributions will be used in section 3.5 for comparing different quantile matching methods.

3.3.2 Sample quantiles

Using samples of a random variable to estimate quantiles is prevalent in statistical modeling. Given a random sample from some distribution, the sample quantile function may be defined in terms of the order statistics. Hyndman and Fan (1996) review several definitions of sample quantiles used in statistical packages and state definition 3.2 as a general definition of the functions reviewed.

Definition 3.2. *For n independent observations $Y_1, ..., Y_n$ from a distribution with corresponding order statistics $Y_{(1)}, ..., Y_{(n)}$, the sample quantile function $\hat{Q}_n(p)$ may be defined as*

$$\hat{Q}_n(p) = (1-\gamma)Y_{(j)} + \gamma Y_{(j+1)} \tag{3.4}$$

where

$$\frac{j-m}{n} \leq p < \frac{j-m+1}{n}$$

for some $m \in \mathbb{R}$ and $0 \leq \gamma \leq 1$ where γ is some function of $j = \lfloor pn + m \rfloor$, $g = pn + m - j$, and $\lfloor \cdot \rfloor$ is the floor function.

Sample quantiles are used to give a probability summary of a dataset. Sample quantiles of Markov chain Monte Carlo draws from a Bayesian posterior distribution are used to analyze

posterior distributions, for example by using quantiles to form credible intervals. Well known asymptotic results of sample quantiles are presented below, and it may be noted that asymptotic results for quantile regression quantiles are similar (Kocherginsky et al., 2005; Koenker and Bassett Jr, 1978).

3.3.2.1 Sample quantile central limit theorem

The asymptotic distribution of a set of sample quantiles is given in (3.5) and is the basis for the QGP model for QM in section 3.4. For independent draws $Y_1, ..., Y_n$ from a continuous random variable with CDF F, PDF f, and QF Q, and with a set of sample quantiles calculated by (3.4), theorem 3.1 holds. Note that in (3.5), $N_K(\cdot, \cdot)$ refers to a multivariate normal distribution with K dimensions, and in (3.6) the operator $a \wedge b$ is the minimum of the values a and b.

Theorem 3.1. Sample quantile central limit theorem

Given a vector of length K of probabilities $\boldsymbol{p} = (p_1, ..., p_K)$ and the corresponding quantile vector $\hat{\boldsymbol{Q}}_n(\boldsymbol{p}) = (\hat{Q}_n(p_1), ..., \hat{Q}_n(p_K))$, if F is absolutely continuous for all $y \in \mathcal{Y}$ and is strictly increasing, then

$$\sqrt{n}(\hat{\boldsymbol{Q}}_n(\boldsymbol{p}) - \boldsymbol{Q}(\boldsymbol{p})) \xrightarrow{D} N_K(0, \boldsymbol{\mathcal{K}}) \tag{3.5}$$

where $\boldsymbol{\mathcal{K}}$ is a covariance matrix with the ij^{th} entry κ_{ij} where

$$\kappa_{ij} = \frac{p_i \wedge p_j - p_i p_j}{f(Q(p_i)) f(Q(p_j))} \tag{3.6}$$

Theorem 3.1 has been known since the mid 20th century (Cramér, 1951), and a thorough though not unique proof is found in Walker (1968). The entries in the covariance matrix in (3.6) may equivalently be written as

$$\rho_{ij} = [p_i \wedge p_j - p_i p_j] q(p_i) q(p_j) \tag{3.7}$$

where q is the QDF. Another CLT in (3.8) for a set of quantiles transformed by the underlying CDF $F(\hat{\boldsymbol{Q}}_n(\boldsymbol{p})) = (F(\hat{Q}_n(p_1)), ..., F(\hat{Q}_n(p_k)))$ is in (3.5) and can be derived from (3.5) by using the PIT and the Delta method (Parzen, 2004).

$$\sqrt{n}(F(\hat{\boldsymbol{Q}}_n(\boldsymbol{p})) - \boldsymbol{p}) \xrightarrow{D} N_K(0, \boldsymbol{\Gamma}) \tag{3.8}$$

Here the covariance matrix Γ has entries $\Gamma_{ij} = p_i \wedge p_j - p_i p_j$, making the asymptotic distribution in (3.8) a Brownian bridge (Chow, 2009). This second result makes QM using the QGP model possible where $Q(p)$ is difficult to solve as is shown in the following section.

3.4 Quantile Gaussian process model

We consider the situation where one is given a set of data including a vector of K probabilities $\boldsymbol{p} = (p_1, ..., p_K)$ and a vector of estimated quantiles at the given probability levels $\hat{\boldsymbol{Q}}_n(\boldsymbol{p}) = (\hat{Q}_n(p_1), ..., \hat{Q}_n(p_K))$. The set of quantiles may provide useful information about a distribution, but the information is limited and one may desire a more complete distribution. The QDP model for QM and estimating a continuous distribution based on the CLT in (3.5) is in (3.9).

$$\hat{\boldsymbol{Q}}_n(\boldsymbol{p}) \sim N_K(\boldsymbol{Q}_\theta(\boldsymbol{p}), n^{-1}\boldsymbol{\mathcal{K}}_\theta) \tag{3.9}$$

Here θ is an unknown parameter to be estimated from the data. The covariance matrix $\boldsymbol{\mathcal{K}}_\theta$ has entries from the covariance function $\kappa_\theta(\cdot, \cdot)$ where

$$\kappa_\theta(p, p') = \frac{p \wedge p' - pp'}{f_\theta(Q_\theta(p)) f_\theta(Q_\theta(p'))}, \quad p, p' \in (0, 1)$$

The covariance function $\kappa_\theta(\cdot, \cdot)$ initially appears as if it will be difficult to evaluate and indeed can be, depending on the functional forms of f_θ and Q_θ. But where these functions belong to a location-scale family, the covariance function can be greatly simplified. Noting that the covariance is a function of $f_\theta(Q_\theta(p)) = q_\theta(p)$, the QDF, if $\theta = (\mu, \sigma)$ where μ is a location parameter and σ is a scale parameter, then $q_\theta(p) = \sigma q(p)$ where $q(p)$ is the QDF of a standard distribution having location 0 and shape 1. Staudte (2017) calls this the location invariant and scale equivariant

property. For location-scale distribution families, this greatly simplifies the form of $\kappa_\theta(\cdot,\cdot)$ so that it becomes

$$\kappa_\theta(p,p') = \sigma^2[p \wedge p' - pp']q(p)q(p')$$

Thus for estimating θ one only needs to estimate σ^2. A special case of the QGP for QM of a location-scale family is the normal QGP which we define and explore below.

3.4.1 Normal QGP

If the quantiles in a dataset $(\hat{\boldsymbol{Q}}_n(\boldsymbol{p}), \boldsymbol{p})$ are assumed to be calculated from a normal distribution $N(\mu, \sigma^2)$, and the desire is to estimate the parameters μ and σ by modeling the quantiles according to the QGP in (3.9), then one may use the model (3.10).

$$\hat{\boldsymbol{Q}}_n(\boldsymbol{p}) \sim N_K\left(\mu + \sigma\boldsymbol{\Phi}^{-1}(\boldsymbol{p}), \frac{\sigma^2}{n}\boldsymbol{\Psi}\right) \tag{3.10}$$

Here $\boldsymbol{\Phi}^{-1}(\boldsymbol{p}) = (\Phi^{-1}(p_1), ..., \Phi^{-1}(p_K))$ is a known vector since $\boldsymbol{p}$ is given, and $\boldsymbol{\Psi}$ is a known $K \times K$ matrix with ij^{th} entry

$$\Psi_{ij} = \frac{2\pi(p \wedge p' - pp')}{\exp\{-\frac{1}{2}[\Phi^{-1}(p)^2 + \Phi^{-1}(p')^2]\}}.$$

Model (3.10) can then be viewed as an atypical normal linear regression model. Two aspects that make this model atypical relative to the standard linear regression model are that the slope parameter σ must be greater than 0 and that σ is both a regression coefficient and part of the variance. Note also that the sample size n is included as part of the variance. If n is known, then it may be multiplied through Ψ and forgotten. If unknown, it can then be accounted for as an unknown part of the variance.

Take for instance the data $\boldsymbol{X} = \left(\boldsymbol{1}, \Phi^{-1}(\boldsymbol{p})\right)$ and the parameter vector $\beta = (\mu, \sigma)$. Assuming the sample size n is unknown and setting $\sigma^2/n = \gamma^2$, model (3.10) then becomes

$$\hat{\boldsymbol{Q}}_n(\boldsymbol{p}) \sim N_K\left(X\beta, \gamma^2\boldsymbol{\Psi}\right)$$

In this model, one simplifying assumption may be to treat σ and γ^2 separately and proceed to fit a standard linear regression model. All frequentist and Bayesian results of the linear regression model apply, including the existence of conditionally conjugate prior distributions. The positive constraint on σ can also be dealt with without adding much complication (see Gelman et al. (2013) pgs. 377-378). If for a certain problem it is important to make inference on the unknown n, or σ and γ^2 cannot treated as separate, then one may estimate the two parameters by assigning appropriate prior distributions to σ and n and reverting back to (3.10). In section 3.5, we analyze a normal QGP in a simulation study where we assign independent prior distributions to σ and n, and σ is treated as both a regression coefficient and as part of the variance.

For any QGP model where the QF belongs to a location-scale family, the discussed relation to the normal linear regression model applies. The only difference in modeling becomes formulating the matrix Ψ using the proper QDF, and the data $\boldsymbol{X}$ with the proper quantile function. Many situations will require a more complicated QF which doesn't allow for modeling the data as a linear regression. Below we consider for instance a finite mixture distribution.

3.4.2 Finite normal mixture QGP

For a continuous random variable distributed according to a finite mixture distribution, the CDF takes the form of (3.11). Here there are C component distributions where $c \in \{1, ..., C\}$, F_{θ_c} is a continuous CDF with parameter θ_c, and $w_c > 0$ is a weight such that $\sum_{c=1}^{C} w_c = 1$.

$$F_\theta(x) = \sum_{c=1}^{C} w_c F_{\theta_c}(x) \tag{3.11}$$

Often F_θ will not be easily invertible, thus Q_θ will be a complicated function which can be evaluated only via numerical optimization which is often computationally expensive and/or inaccurate. A simple solution is to model the PIT transformed quantiles from the data rather than modeling the quantiles directly. Model (3.9) is then reformulated to be model (3.12), where $\Gamma_{ij} = p_i \wedge p_j - p_i p_j$ as in (3.8).

$$F_\theta(\hat{\boldsymbol{Q}}_n(\boldsymbol{p})) \sim N_K(\boldsymbol{p}, n^{-1}\boldsymbol{\Gamma}) \tag{3.12}$$

By modeling the PIT quantiles, the only function which requires evaluating is the CDF F_θ. Here both $\boldsymbol{p}$ and $\boldsymbol{\Gamma}$ are given, and solving for the transformation $F_\theta(\hat{\boldsymbol{Q}}_n(\boldsymbol{p}))$ is unnecessary for our purposes. A Bayesian fit of (3.12) is straightforward, requiring only that $F_\theta(\hat{\boldsymbol{Q}}_n(\boldsymbol{p}))$ be evaluated as part of the acceptance ratio of a Markov chain Monte Carlo (MCMC) iteration.

In section 3.5, we analyze model (3.12) where F_θ is set as a normal mixture distribution with $C = 4$ component normal distributions. Fitting this model proved faster and produced better results than trying to fit model (3.9) and evaluating Q_θ the QF of a normal mixture. In fact we found that in cases where F_θ was a simpler function, such as a normal or exponential distribution, fitting model (3.12) was at least as fast and the results were at least as good as fitting (3.9). Thus when possible, we elected to fit (3.12).

3.4.3 Competing quantile matching methods

A number of methods already exist for QM. Here we briefly review four of those methods, and in section 3.5 we compare performance between these and the QGP. The four methods include spline interpolation (SPL), kernel density estimation (KDE), an order statistics based model (ORD), and an independent quantile model (IND). The first two of these methods are non-parametric methods and neither of which include modeling the uncertainty of the quantiles.

The SPL method was used by Gerding et al. (2023) and Shandross et al. (2024) in order to estimate a CDF function given quantile forecasts from disease outbreak forecast hubs. SPL is an interpolation where monotonic cubic splines are fit to pass through each given quantile and predict a function for all values between given quantiles and beyond the extreme values. An `R` package used for fitting SPL and which we use in section 3.5 is `distfromq` (Ray and Gerding, 2024).

The KDE method treats given quantiles as if they constitute a random sample from a distribution and applies kernel smoothing to estimate a density function. KDE requires selecting a kernel function, often a Gaussian kernel is chosen, and applying that function to each draw in a

sample. Gyamerah et al. (2020) applied KDE smoothing with an Epanechnikov kernel to quantiles estimated via three different machine learning methods used for predicting crop yield. They then combined the estimated densities into an ensemble prediction. He et al. (2016) also use KDE to estimate density forecasts from quantiles to produce energy forecasts. We use the `evmix` package for performing QM using KDE (Hu and Scarrott, 2018).

One of the more common methods for QM is to select a CDF or QF function and then select model parameters which give a least squares fit to the estimated quantiles as done by Li et al. (2019). The `R` package `rriskDistributions` performs this least squares fit for some standard distributions (Belgorodski et al., 2017). Independent random error may be included to the best fit QF allowing for estimation of the variability of the quantiles (Nirwan and Bertschinger, 2020). In section 3.5, we analyze model (3.13) as a proxy for the IND model to compare with other methods.

$$F_\theta(\hat{Q}(p_k)) \overset{ind}{\sim} N(p_k, \sigma_\rho^2) \tag{3.13}$$

Here $\hat{Q}_k$ is the estimated sample quantile at probability p_k. The major difference between model (3.13) and model (3.12) is that the error for each quantile in a set of quantiles is considered independent, and the effects of sample size and correlation suggested by the CLT in (3.8) are ignored. Note that the parameter σ_ρ is not a part of θ, the parameter of the "true" distribution estimated by QM but instead is meant to capture independent error among the sample quantiles.

The final method we include is the ORD model introduced by Nirwan and Bertschinger (2020). This model relies on the definition of sample quantiles being order statistics. In the ORD model, the joint distribution of a set of order statistics is the model likelihood. The likelihood of a set of order statistics from a continuous distribution is a function of both the CDF and PDF of a continuous distribution. ORD is the method most similar to QGP and it provides exact inference of quantile uncertainty whereas the QGP provides asymptotic uncertainty.

For assessing QM, there are two aspects we analyze. The first is a model's ability to estimate parameters and the second is how far away a fit QM distribution is from a true distribution. Of the methods previously listed, parameter inference is only possible for the QGP, IND, and ORD models. These are also the only methods which provide uncertainty quantification for quantile

values. To analyze how far an estimated distribution is from a true distribution, a distance measure must be utilized. We outline three of these distances below.

3.4.4 Distance measures

A number of metrics for measuring the distance between two continuous univariate distributions exist. Those we consider in this paper are the Wasserstein distance (WD) and a slight modification of it, the total variation (TV) or the statistical difference, and the Kullback-Leibler divergence (KLD). A brief overview of these and other metrics, some of their statistical uses and properties, and relationships between metrics is found in Gibbs and Su (2002). These metrics were used in section 3.5 to assess the how well different QM methods approximate a true distribution.

The WD is used to measure the distance between two univariate random variables by their CDFs or QFs and has many applications in mathematics, optimization, and statistics, see Panaretos and Zemel (2019) for a review of the WD and its use in statistics. The p-WD is defined for two continuous random variables X and Y with respective CDFs F_X and F_Y. In terms of the two CDFs the p-WD is defined in (3.14), and the p-WD in terms of the corresponding QFs is defined in (3.15).

$$WD_p(F_X, F_Y) = \left(\int_{\mathbb{R}} |F_X(t) - F_Y(t)|^p dt \right)^{1/p} \tag{3.14}$$

$$WD_p(F, G) = \left(\int_0^1 |F_X^{-1}(t) - G_Y^{-1}(t)|^p dt \right)^{1/p} \tag{3.15}$$

We define a modified version of the WD, the uniform 1-WD (UWD1) in (3.16). Here we take F_X to be the CDF of a continuous random variable X. For a different continuous random variable ξ with CDF F_ξ, $F_X(\xi)$ is a random variable with support on $[0, 1]$. We let $F_{X,\xi}$ be the CDF of the random variable $F_X(\xi)$. Then the measure in (3.16) is a measure of how close $F_X(\xi)$ is the the CDF of a standard uniform distribution, noting that by the PIT, if $\xi \overset{D}{=} X$, $F_X(\xi)$ is the random variable of a standard uniform distribution. Multiplying the integral in (3.16) by two ensures the

measure takes values between 0 and 1 ((https://stats.stackexchange.com/users/20519/zhanxiong), 2024) where a value near 0 means ξ and X are "near" each other.

$$UWD_1(F, \mathcal{L}) = 2 \int_0^1 |F_{X,\xi}(x) - x| dx \tag{3.16}$$

The TV, defined in (3.17), is a distance between distributions measured in terms of the PDFs f and g. The TV takes values between 0 and 1, and it is closely related to the distance used in Sgouropoulos et al. (2015) to measure the goodness of QM.

$$TV(f, g) = \frac{1}{2} \int_{-\infty}^{\infty} |f(x) - g(x)| dx \tag{3.17}$$

The final metric between distributions we consider is the KLD defined in (3.18). The KLD is not a true metric but has many useful properties and applications. It may be interpreted as the expected divergence if one is using f to approximate g.

$$KLD(g||f) = \int_{\mathcal{X}} g(x) \log \left(\frac{g(x)}{f(x)} \right) dx \tag{3.18}$$

3.5 Quantile matching methods comparison on simulated data

In this section we present several simulation studies assessing the QM of the QGP model and compare results with those of the SPL, KDE, IND, and ORD matching methods. The QGP, IND, and ORD models were each fit via Hamiltonian Monte Carlo (HMC) sampling using the `Stan` software and the `cmdstanr` package developed and maintained by the Stan Development Team (2024) (Gabry et al., 2022). Simulation studies are done to assess parameter estimation and inference as well as QM estimation of a distribution.

3.5.1 Parameter estimation and quantile matching for known distribution families

Where the distribution family is known, one goal of QM may be to estimate and make inference on model parameters. In the first two studies, we analyzed parameter estimation and inference for normal and exponential family distributions with known parameters. We analyzed

the QGP model from (3.12), the IND model, and the ORD model. We note that model (3.12) allows one to do inference for both the model parameter θ and for an unknown sample size n. The ORD model also allows for estimation of an unknown n. For the QGP and ORD models, we include modeling cases where n is known and where n is unknown and estimated. We denote the models where n is known as QGPN and ORDN. We also compared the distance between the true model and the predictive distributions estimated via QGP, IND, ORD, SPL, and KDE. We also include a comparison of the QGP and ORD models where the true distribution is a quantile defined distribution.

3.5.1.1 Normal parameter estimation

In this simulation study we fit the ORD and IND models to quantiles estimated from independent samples from a normal distribution and compare parameter estimation with fits from the QGP model. Quantiles were simulated by first drawing a sample of size n from a known distribution then estimating K quantiles given probabilities $\{p_1, ..., p_K\}$. The ORD, IND, and QGP models were then fit to the quantiles where F was set to be the CDF of a normal distribution. For each combination of $n \in \{50, 150, 500, 1{,}000, 5{,}000\}$ and $K \in \{3, 7, 15, 50\}$, there were 500 simulation replicates. The data were simulated from a normal distribution with mean $\mu = 4$ and standard deviation $\sigma = 3.5$. For each of QGP, ORD, and IND models, we assigned the same prior distributions to μ and σ. When n is unknown, we also assigned it a prior distribution. The prior distributions were

$$\begin{aligned}
\mu &\sim N(5, 7^2) \\
\sigma &\sim N(0, 6^2)\mathbb{1}\{\sigma > 0\} \\
n &\sim N(0, 3000^2)\mathbb{1}\{n > 0\}.
\end{aligned}$$

For the IND model the prior on the independent parameter $\sigma_\rho \mathbb{1}\{n > 0\}$ from model (3.13) is $1/\sigma_\rho \sim N(0, 3000^2)$, which is similar to the prior on n for the QGP and ORD models. For each fit, the HMC chain was run for 60,000 draws with the first 10,000 discarded as a burn-in.

Figure 3.2 shows examples of marginal posterior densities for μ, σ, and unknown n where the true $n \in \{50, 150, 500, 1,000\}$ and $K \in \{7, 15, 23\}$. The parameter uncertainty of the QGP and ORD models are unsurprisingly similar in most cases, and where they differ it is hard to say that one is estimating the true parameter better than the other. The IND model on the other hand has tighter densities, and often the bulk of the posterior is far from the true parameter.

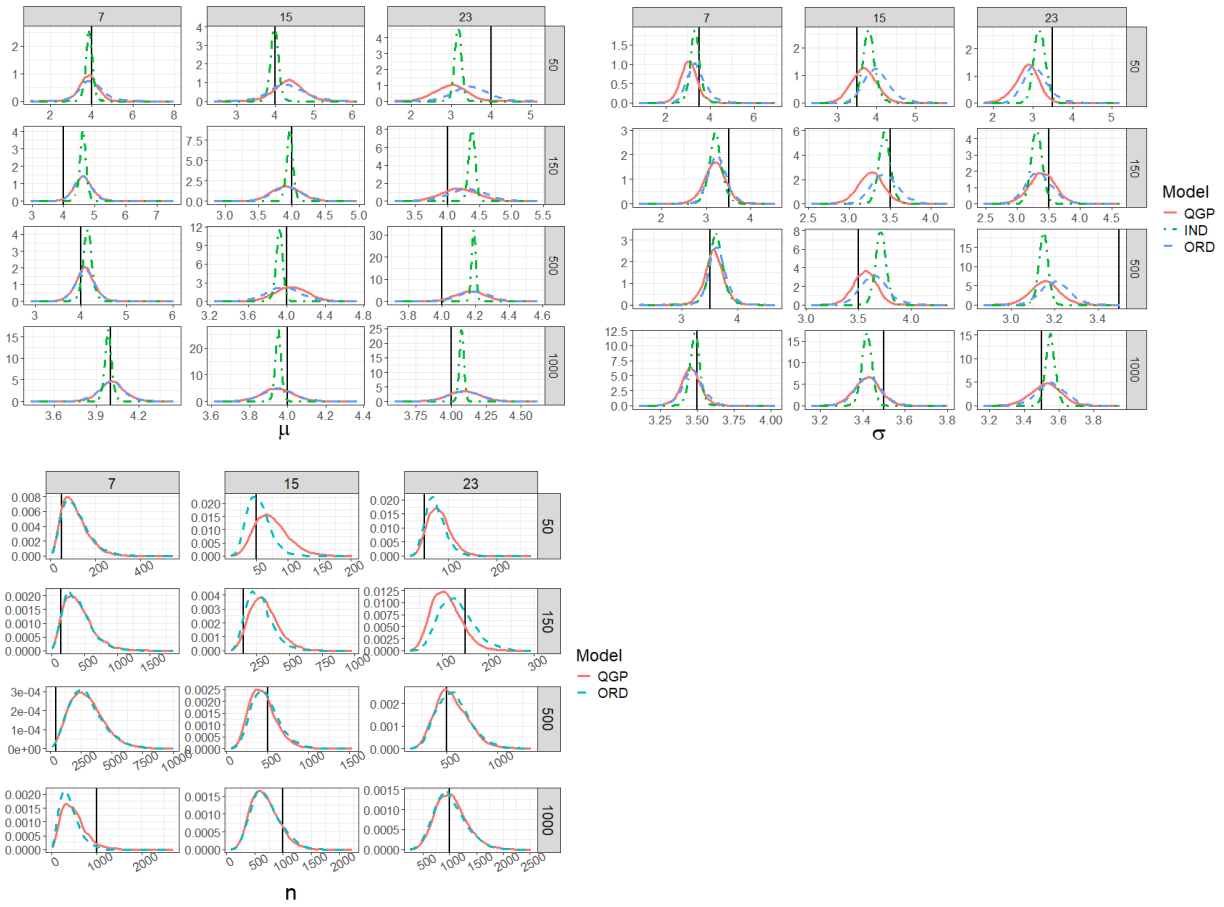

Figure 3.2: Density plots of posterior distribution samples for normal parameters by QM for QGP, ORD, and IND models. QM was done on estimated quantiles from a normal distribution with mean 4 and standard deviation 3.5. The posterior densities are for μ (top left), σ (top right), and sample size n (bottom). Plots are faceted by the sample size $n \in \{50, 150, 500, 1{,}000\}$ (y-axis) and number of quantiles $K \in \{7, 15, 23\}$ (x-axis). Vertical lines (black) show the value of the true parameter.

The left side of figure 3.3 shows the 90% coverage of the posterior distribution credible intervals as n increases for μ, σ and n for five models. QGPN and ORDN models where n is known are included along with QGP, ORD, and IND. 90% credible intervals of the parameters

were calculated by computing the 0.05^{th} and 0.95^{th} quantiles from the posterior distribution samples. The coverage percentage is calculated as the percentage of the 500 replicates which the true parameter value was within the 90% credible interval. As expected, the nominal coverage for the QGP and ORD models is better than that of the IND model, and for the two models where n is known the coverage is better than where it is unknown, particularly for lower values of n and K. The coverage for the ORD models appears to be slightly better than for the QGP models though not by much and not for every combination of n and K. This is likely because the ORD model provides exact inference whereas the QGP model provides asymptotic inference.

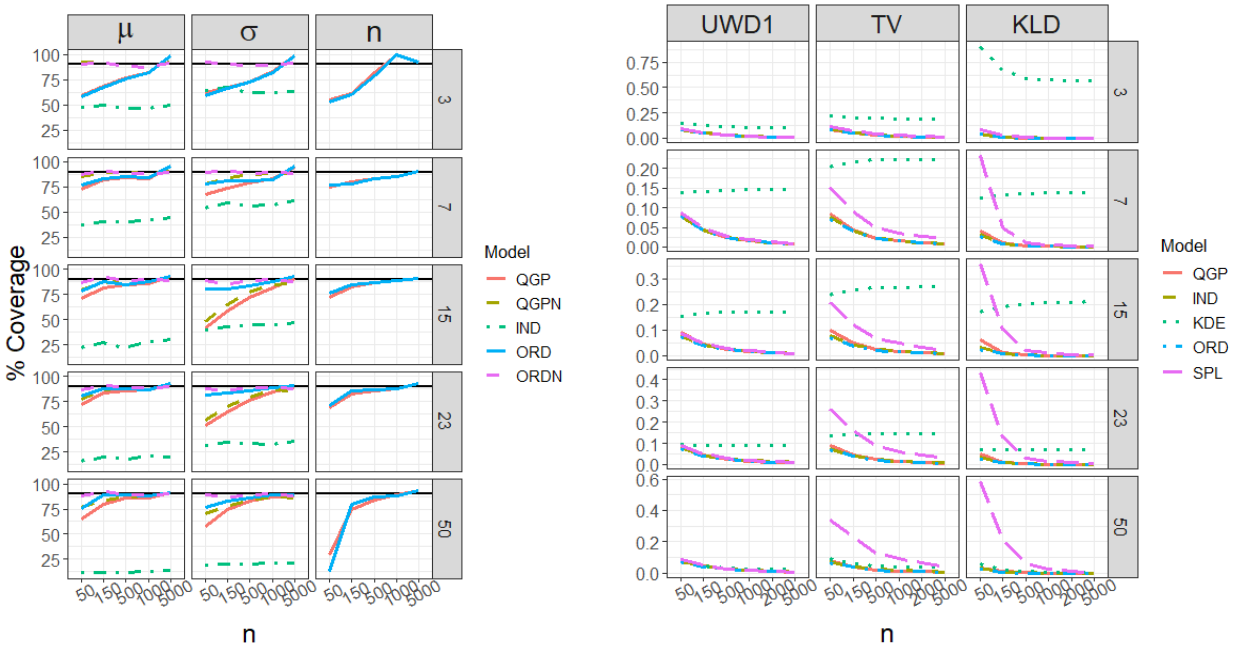

Figure 3.3: Posterior coverage (left) calculated as the percentage of times the true parameter fell within the modeled 90% credible interval over the 500 replications. Coverage is faceted by the normal parameters μ, σ, and n with $K \in \{3, 7, 15, 23, 50\}$, and by increasing sample size (x-axis). The five models QGP, ORD, QGPN, ORDN, and IND are colored as shown the legend. The horizontal line (black) is at the nominal 90% level. Only QGP and ORD appear for the parameter n as they are the only two which estimate an unknown n. Distance between the true distribution and the estimated QM predictive distribution (right) averaged over the 500 replications. Distances include the UWD1, TV, and KLD for $K \in \{3, 7, 15, 23, 50\}$, and by increasing sample size (x-axis).

To analyze how well QM methods produced predictive distributions which approximate the true distribution $N(4, 3.5^2)$, we measured the distance between the estimated predictive distributions of QGP, ORD, and IND fits to the true distribution. We also measured the

distances of the SPL and KDE QM fits to the true distribution. The distances were measured via the UWD1, TV, and KLD metrics. The UWD1 for each of QGP, ORD, QGPN, ORDN, and IND were calculated as follows. For $\{\theta_m\}^M$ being M draws from the posterior distribution of θ where $\theta_m = (\mu_m, \sigma_m)$, the QM posterior predictive distribution was simulated by repeatedly sampling a θ_m^*, then sampling X_m from a distribution with CDF $F_{\theta_m^*}$. Repeating this for $M = 50,000$ times gives a QM posterior predictive sample. For the CDF F_θ where θ is the true parameter, $\xi_m = F_\theta(X_m)$ was calculated and the empirical CDF $\hat{F}_\xi$ was calculated from the sample $\{\xi_m\}^M$. F_ξ in (3.16) was then replaced by $\hat{F}_\xi$ to calculate UWD1. For SPL and KDE $F_\xi(x)$ was replaced by the respective estimated CDFs. The `integrate` function in `R` was used for calculating the UWD1 in (3.16).

To estimate TV and KLD, rather than sampling from the posterior distribution, the marginal means for the parameters were calculated from the posterior distribution samples. These marginal means are $\bar{\theta}_M = (\bar{\mu}_M, \bar{\sigma}_M)$ where $\bar{\mu}_M$ and $\bar{\sigma}_M$. To calculate TV, f in (3.17) is replaced by $f_{\bar{\theta}_M}$ and g is replaced by f_θ. The TV calculation was done using the `integrate` function in `R`. The KLD was estimated also using $f_{\bar{\theta}_M}$ and f_θ, but was done via importance sample. That is (3.18) was estimated by

$$\widehat{KLD}(f_\theta || f_{\bar{\theta}_M}) = \frac{1}{S} \sum_{s=1}^{S} \log \left(\frac{f_\theta(Y_s)}{f_{\bar{\theta}_M}(Y_s)} \right)$$

where $Y_s \overset{iid}{\sim} F_\theta$. For both SPL and KDE, $f_{\bar{\theta}_M}$ is the estimated PDF.

The right side of figure 3.3 shows the UWD1, TV, and KLD metrics averaged over the 500 simulation replicates as n increases for the five QM methods. The QGP, ORD, and IND are nearly indistinguishable and outperform the SPL and KDE in almost every case, though KDE tends to improve as K increases, performing similar to the parametric methods for $K = 50$. For UWD1, the SPL performs similarly to the parametric QM methods, but for the TV and KLD, the SPL performs much worse, especially for smaller sample sizes.

This study shows the ability of the QGP QM model to estimate and perform inference on model parameters where the true distribution family is known. It also shows the QGP's ability to

produce a posterior predictive distribution which closely matches a true distribution, according to several metrics, relative to other QM methods. Between parameter inference and predicting the true distribution, QGP is superior to all methods we compared it with except for the ORD model which has similar results to the QGP. A similar simulation study where the data is simulated from an exponential family distribution is in appendix 3.A. The results of that study are similar to those of the normal distribution family study.

When fitting the Bayesian models both in the normal and in the exponential setting, the IND model tended to fit the fastest, followed by the ORD models, and finally the QGP models. The time to fit however was very fast, usually no more than seven seconds except for the case where $K = 50$ quantiles, in which case the QGP model where n was unknown often took around 20 seconds to fit. However, with $K = 50$ the QGPN model continued to fit very rapidly, fitting in under seven seconds.

The analysis above show the usefulness in using the QGP model both in parameter estimation and in QM when given sample quantiles. In the normal case, the QGP showed improved parameter inference over the IND model and allows for making inference on the sample size n. Compared to the SPL and KDE QM methods, the QGP fits tend to be closer in distance to the true distribution. The QGP model, however, performs about the same as the ORD model in terms of inference and QM. The section below outlines a situation where the QGP model may be a better option for QM than the ORD model.

3.5.1.2 Quantile defined distributions

The normal and exponential families both have CDF and PDF functions which are either available in closed form or are easy to evaluate with software which is widely accessible. The next study we performed was done to compare the QGP and ORD models where the distribution family was a quantile defined distribution which lacks a CDF that is easy to evaluate. The purpose of this study was to show an example where one may prefer to perform QM using the QGP rather than the ORD.

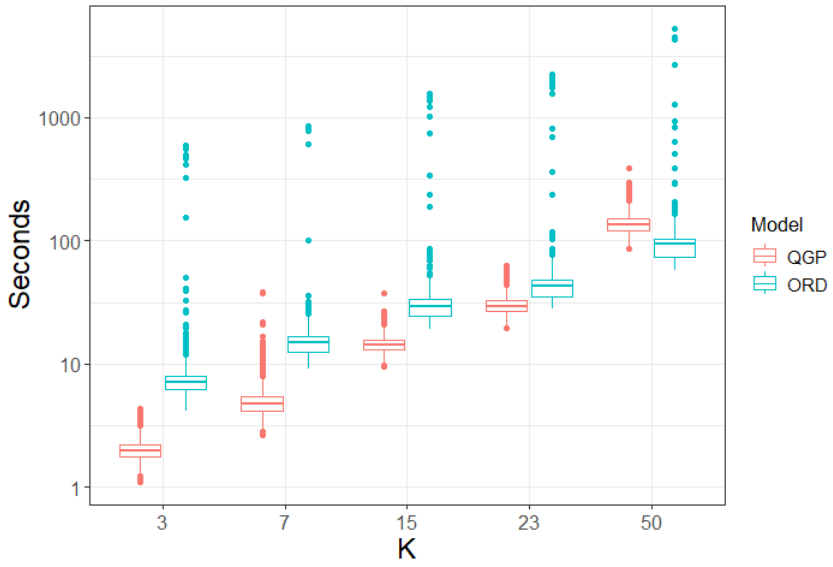

Figure 3.4: Boxplots of time to fit a model in seconds (y-axis) for the 500 replicates in the simulation study for QM of the generalized lambda distribution (GLD) for $K \in \{3, 7, 15, 23, 50\}$ (x-axis). Boxplots are separated by QM methods QGP (red) and ORD (blue). The y-axis is on the $\log_{10}$ scale.

The GLD is reviewed in section 3.3.1.4, and a simulation carried out similarly to the previous studies was performed with GLD as the true distribution. However, QM was only done by the QGP and ORD models. Because of the relative difficulty of evaluating the CDF of the GLD and the ease of evaluating the QF, the CDF transformed QGP model in (3.9) instead of (3.12) was used. For the ORD model, evaluating the CDF was required for modeling. Evaluating the CDF of the GLD requires using an algebraic solver, and it took longer to code the ORD model in `Stan` due to there not being readily available software which we were aware of. Once working, however, both QGP and ORD models fit the estimated quantiles well, however, the ORD usually required more time to fit than the QGP. Figure 3.4 shows boxplots for the 500 replicates for different values of K quantiles. In all cases except for $K = 50$ the QGP tends to fit much faster than ORD. Note that the y-axis in figure 3.4 is on the $\log_{10}$ scale to make visualization easier. Even at $K = 50$ some outliers for the ORD model required thousands of seconds to fit whereas the longest required time for fitting a QGP model was a few hundred seconds.

The MLD was also reviewed in section 3.3.1.4. We were able to fit a QGP model for the MLD which appeared to fit sample quantiles well, but we were unable to fit a working ORD model. The

algebraic solver struggled to evaluate the CDF, and posterior sampling in `Stan` was extremely slow. When fitting finally finished, the result was a very poor fit. Perhaps with more time and effort we could have made a working model, but the QGP worked well enough that we did not feel it was worth the effort.

3.5.2 Matching unknown distributions

The simulation studies in this section are for the situation where the true distribution family is unknown, unimportant, or too complex to be practically evaluated but where one still wishes to recover or approximate the distribution. We selected three distributions each with a different shape to analyze how well the distributions can be approximated by QM using the QGP and the competing methods.

With the true distribution family being unknown, we made the modeling decision to set F in the QGP model of (3.12) to be a mixture of normals distribution taking the form of (3.11). The number of component distributions was chosen to be $C = 4$ to provide enough flexibility to approximate the different shapes of quantile functions. This decision was based on the common claim that any continuous distribution may be approximated by a finite mixture of normal distributions (McLachlan and Peel, 2000; Nguyen and McLachlan, 2019; Nguyen et al., 2020). This claim may be optimistic, but we found that a four component normal mixture distribution can approximate the non-normally shaped distributions reasonably well. However, any function which meets the technical definition of a CDF would be appropriate to use for F, see Gasthaus et al. (2019) for an example.

The three distributions from which data were simulated and sample quantiles estimated and to which QM methods were fit were the extreme value (EV) distribution with location 0 and scale 1 $EV(0, 1)$, Laplace (LA) with location 0 and scale 1 $La(0, 1)$, and a two component normal mixture (MIX) $wN(-1, 0.9) + (1 - w)N(1.2, 0.6)$ where $w = 0.35$. From the three distributions, 500 replicates of K quantiles were simulated for each of $K \in \{9, 13, 23, 50\}$ and sample size $n \in \{50, 150, 500, 1{,}000, 2{,}000, 5{,}000\}$. QM for each replicate of simulated quantiles was done using

QGP, ORD, IND, SPL, and KDE. For the QGP, ORD, and IND models, $\theta = \{\theta_1, \theta_2, \theta_3, \theta_4\}$ where $\theta_c = (\mu_c, \sigma_c)$ for $c \in \{1, 2, 3, 4\}$. The prior distributions assigned were as below where $\nu = n$ for the QGP and ORD models and $\nu = 1/\sigma_\rho$ for the IND model.

$$\mu_c \overset{ind}{\sim} N(5, 7^2)$$
$$\sigma_c \overset{ind}{\sim} N(0, 6^2)\mathbb{1}\{\sigma_c > 0\}$$
$$w \sim Dir(\boldsymbol{\alpha}), \quad \alpha_c = 1$$
$$\nu \sim N(0, 3000^2)\mathbb{1}\{\nu > 0\}$$

Often when implementing posterior sampling for a mixture distribution model, one deals with an issue called the label-switching problem. This is where for a mixture distribution parameter $\theta = \{\theta_1, ..., \theta_C\}$, the model likelihood is the same for different permutations of θ. This lack of identifiability for elements of θ makes parameter inference useless, but the predictive distribution may still be close to the true distribution (Stephens, 2000). Because of this, the HMC posterior sampling chain was run longer than in the studies in the previous section. The chain was run for 80,000 steps where the first 20,000 draws were discarded as a burn-in. For this study, we were concerned only with QM and not with parameter inference, so the fact that the parameters are unidentifiable was not considered critical. However, when assessing the TV and KLD between the QM and true distributions, variational Bayes (VB) in `Stan` was used instead of HMC for fitting the models (Kucukelbir et al., 2015). This decision was made because of the lack of parameter identifiability and because the marginal parameter means are used to estimate the PDF. While MCMC methods perform better on parameter inference, VB methods have been shown to have predictive performance comparable to MCMC (Blei et al., 2017).

Figures 3.5, 3.6, and 3.7 each show examples of fits of $K = 23$ quantiles simulated from the EV, LA, and MIX distributions respectively for different sample sizes n. Included QM methods are KDE, SPL, IND, and QGP. ORD was excluded to make visualization easier, but the fits are very similar to the QGP fits. The KDE and SPL fits provide no uncertainty estimation of the quantiles, and return only a continuous function. The SPL fits show a lot of wiggle with the

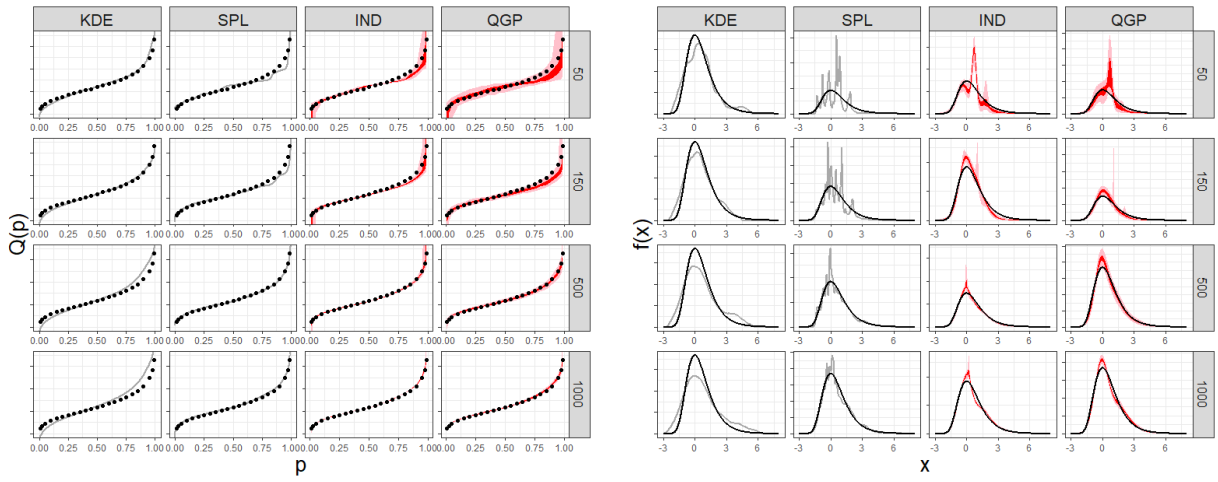

Figure 3.5: QM fits of $K = 23$ quantiles by KDE, SPL, IND, and QGP for $n \in \{50, 150, 500, 1{,}000\}$. The quantiles were sampled from the extreme value distribution $Ev(0, 1)$. The quantile fits (left) show the true quantiles (black) with either the QM fit line (grey) or the credible intervals of 50% (red) and 95% (pink). The estimated PDF plots (right) show the true PDF (black) with either a the QM estimated PDF (grey) or the credible intervals of 50% (red) and 95% (pink).

wiggle decreasing as n increases, whereas the KDE fits do not show as much wiggle, but the KDE fits the distribution tails poorly even as n increases. The IND and QGP models provide uncertainty in the fits, but the QGP is much more conservative with wider intervals which provide superior coverage of the quantiles and the PDFs.

Figure 3.8 shows the simulated coverage of the true quantiles for four values of K for QGP, ORD, and IND fits. The percent coverage is averaged over all K quantiles and all 500 simulated replicates. The two figures show the 50% and 95% coverage for the three models EV, LA and MIX. The plots show that the QGP and ORD models are largely able to meet and exceed the nominal coverage where the IND model falls very short. Because of the extended time it takes to evaluate the QF when calculating coverage, calculation was done on a thinned sample of the posterior distribution where only every 100^{th} HMC draw was kept.

Figure 3.9 shows the UWD1, TV, and KLD distances between QM and true distributions averaged over 500 simulation replicates for the five QM methods. In the UWD1 plots, the SPL appears to match the true distribution only slightly better than the QGP, ORD, and IND for $K = 9$, but the difference is almost unnoticeable for larger K. For TV, SPL appears to perform

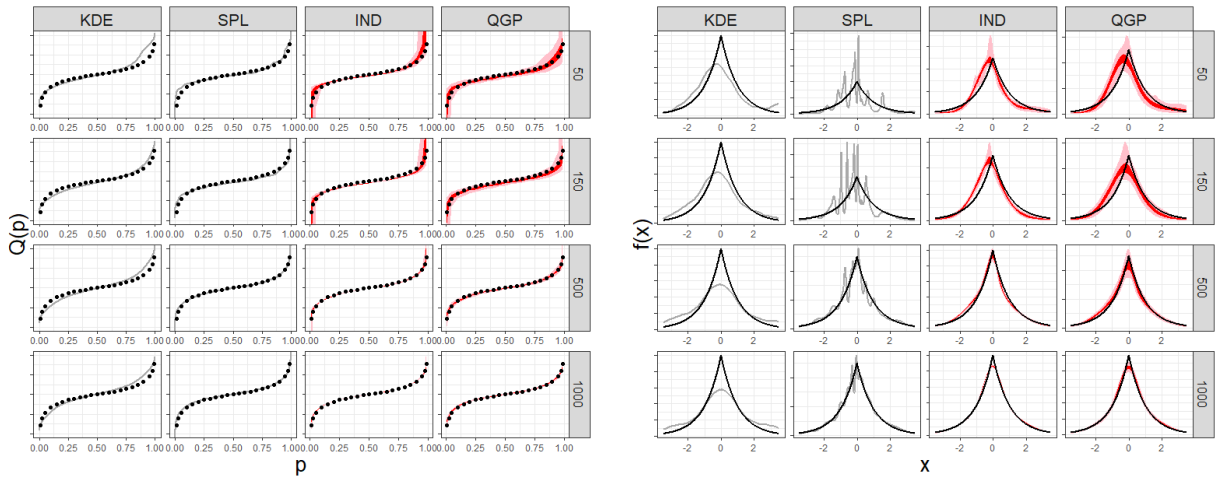

Figure 3.6: QM fits of $K = 23$ quantiles by KDE, SPL, IND, and QGP for $n \in \{50, 150, 500, 1{,}000\}$. The quantiles were sampled from the Laplace distribution $La(0, 1)$. The quantile fits (left) show the true quantiles (black) with either the QM fit line (grey) or the credible intervals of 50% (red) and 95% (pink). The estimated PDF plots (right) show the true PDF (black) with either a the QM estimated PDF (grey) or the credible intervals of 50% (red) and 95% (pink).

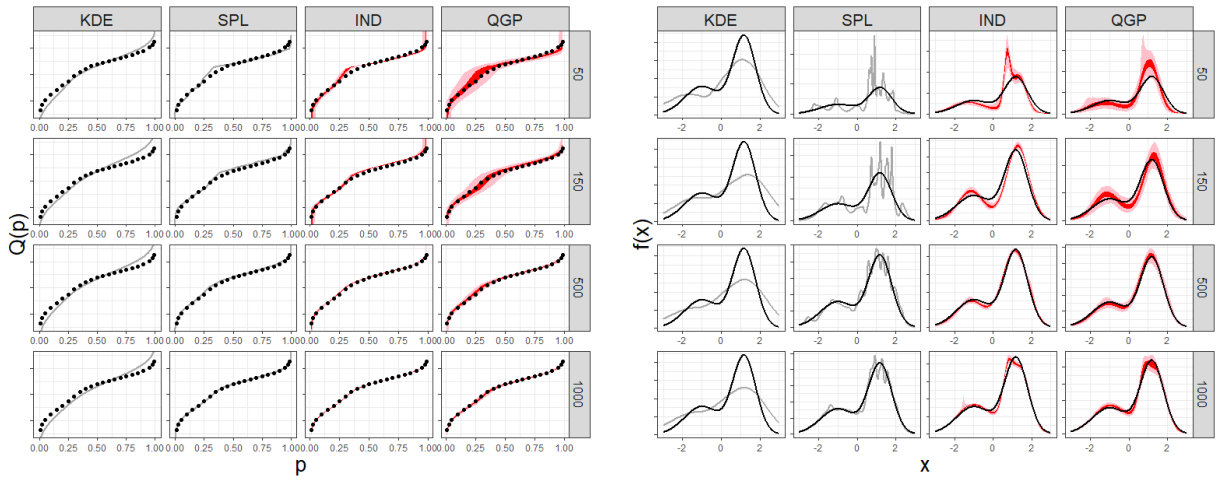

Figure 3.7: QM fits of $K = 23$ quantiles by KDE, SPL, IND, and QGP for $n \in \{50, 150, 500, 1{,}000\}$. The quantiles were sampled from the two component normal mixture distribution $wN(-1, 0.9)+(1-w)N(1.2, .6)$ where $w = 0.35$. The quantile fits (left) show the true quantiles (black) with either the QM fit line (grey) or the credible intervals of 50% (red) and 95% (pink). The estimated PDF plots (right) show the true PDF (black) with either a the QM estimated PDF (grey) or the credible intervals of 50% (red) and 95% (pink).

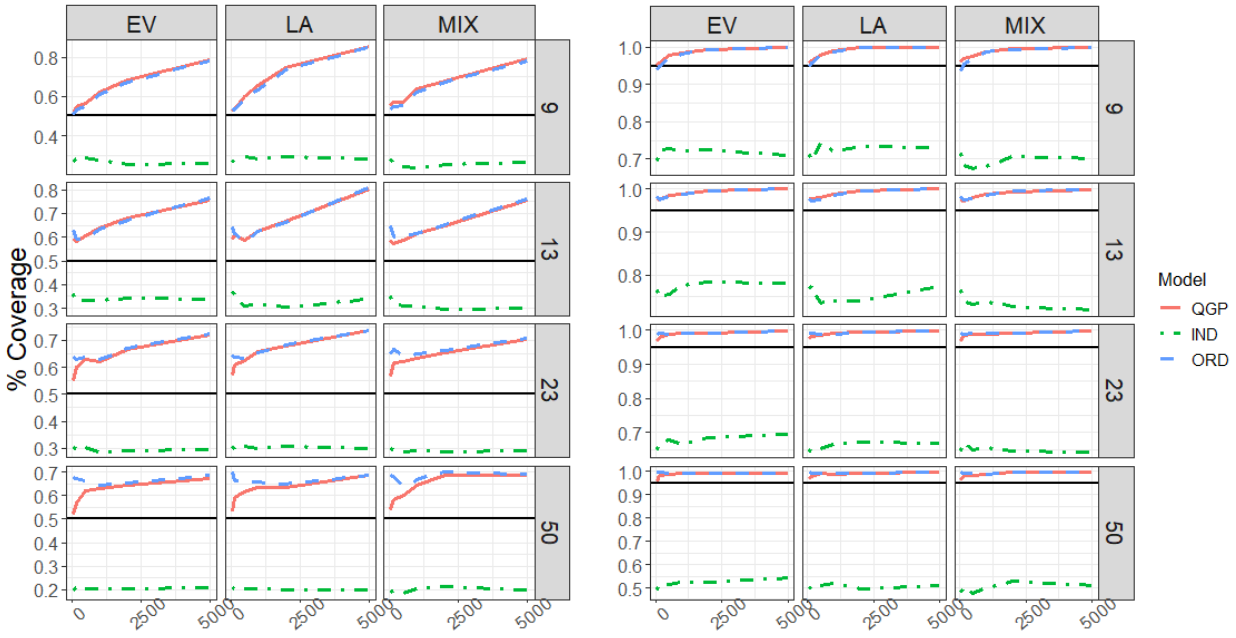

Figure 3.8: Percent coverage of the true quantile values for select quantiles for 500 simulated replicates of sample quantiles averaged over all quantiles. Faceted by the true EV, La, and normal mixture distributions and for $K \in \{9, 13, 23, 50\}$. The plots shows percent coverage for the 50% credible intervals (left) and 95% credible intervals (right). The models included are QGP, ORD, and IND. The vertical bar (black) shows the nominal coverage level.

the best except for the cases where $K \in \{23, 50\}$ where the QGP, ORD, and IND perform slightly better. There is also more separation between the performance of the QGP, ORG, and IND under TV. In most cases, the QGP and ORD perform similarly, though where the true model is a normal mixture distribution, QGP performs better. QGP and ORD both also outperform IND in a few cases. The results for KLD are similar to those for TV.

This section shows the ability of the QGP model to accurately perform parameter inference and QM when given sample quantiles. Although the inference for QGP is asymptotic, the QGP performance in inference and in QM is similar to that of ORD where the inference is exact. There are also circumstances where the QGP may outperform the ORD as shown in cases where the QF is easily evaluated but where the CDF is difficult to evaluate. Another case where the QGP outperformed the ORD was in QM of the normal mixture distribution. An advantage that the

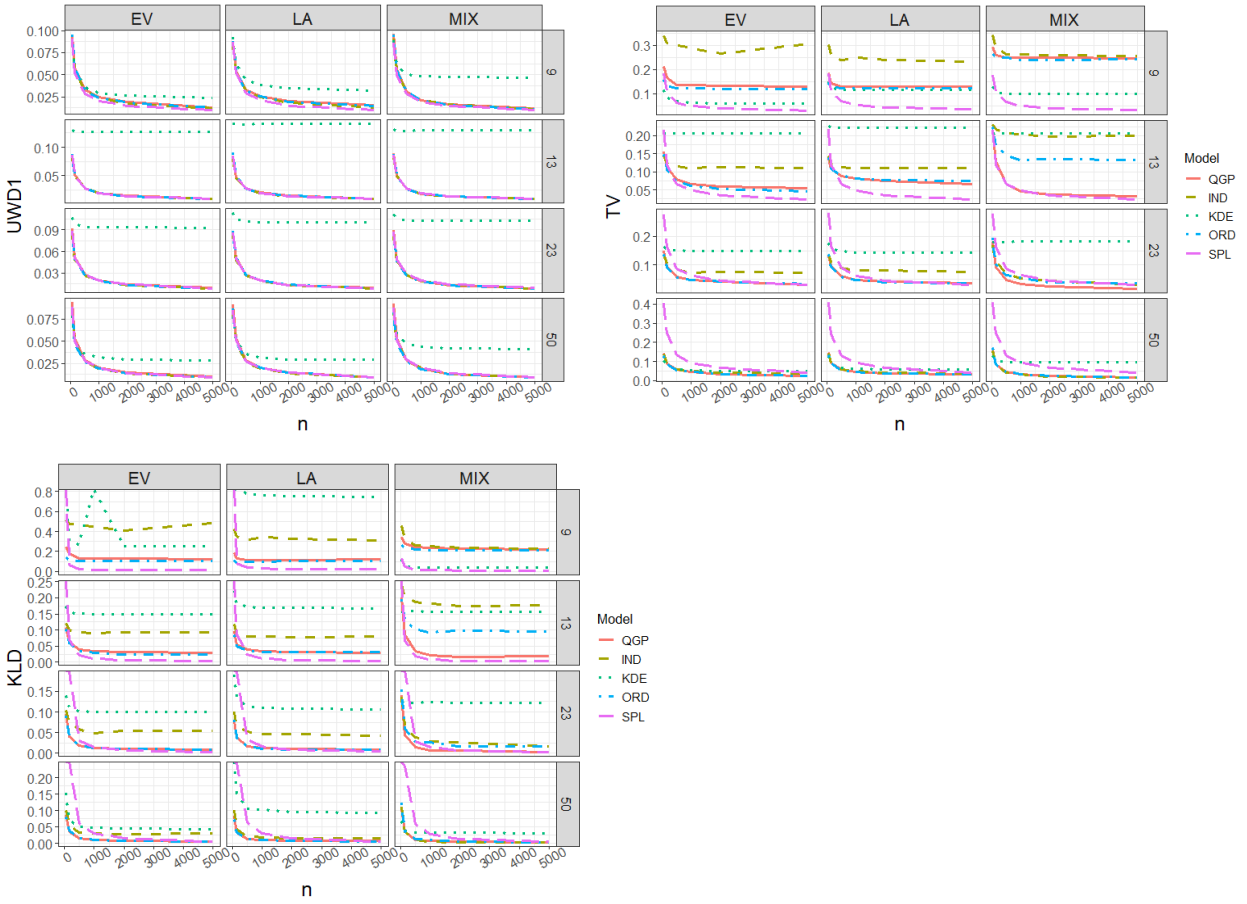

Figure 3.9: Distance between the true distribution and the QM estimated distribution averaged over 500 simulation replicates and measured by UWD1 (top left), TV (top right), and KLD (bottom). Plots are faceted by number of quantiles $K \in \{9, 13, 23, 50\}$ and distribution with the distributions being extreme value (EV), Laplace (LA), and two component normal mixture (MIX). Lines are colored and shaped by QM method, and are drawn by increasing sample size n.

QGP has over SPL and KDE is that it provides accurate uncertainty estimation of quantile uncertainty where SPL and KDE provide no uncertainty estimation.

3.6 CDC flu forecasts analysis

In this section, the QGP model is used for QM of quantile forecasts targeting US flu hospitalizations. Beginning in 2013, the CDC began hosting a yearly forecast competition of the influenza outbreak in the US. This competition is known as FluSight. The flu epidemic typically begins in the fall and ends in late spring the following year, and the forecast competition lasts

around 30 weeks starting in October and ending in May. FluSight involves several academic and industry research teams who each independently develop forecasts every week for predicting certain flu targets for future weeks (Biggerstaff et al., 2016). During the 2022-23 and 2023-24 seasons, the forecast targets were the 1, 2, 3, and 4-week ahead hospitalizations as reported by Health and Human Services (HHS) for the 50 US states, Puerto Rico, the District of Columbia, and the nation as a whole. The data for weekly hospitalizations of flu patients may be found at HealthData.gov (2024). The official guidelines of the most 2023-24 FluSight and all submitted quantile forecasts are publicly available at Github (2024) (Mathis et al., 2024).

Participating teams were free to create forecasts however they pleased, but forecasts for each target were required to consist of 23 quantiles for the probability levels $\boldsymbol{p} = (0.01, 0.025, 0.5, 0.1, ..., 0.95, 0.975, 0.99)$, which were given by FluSight. Figure 3.10 shows examples of quantile forecasts of the same target from 12 participating teams. Each plot shows the log forecast for hospitalizations in the US for the week of January 13, 2024, and it is clear that there are major distributional differences between forecasts. We denote a quantile forecast for flu hospitalizations as $\hat{Q}^{(H)}(\boldsymbol{p}) = (\hat{Q}^{(H)}(0.01), ..., \hat{Q}^{(H)}(0.99))$. The general quantile forecast representation format allowed for forecasts to be compared by the same metric, the weighted interval score (WIS), and to be easily combined into a multi-model ensemble forecast (Mathis et al., 2024). However, under the quantile representation, the tools for scoring forecasts and for building ensemble models are limited as many of the existing tools for scoring forecasts and constructing ensembles require CDFs or PDFs (Wadsworth et al., 2023; Ranjan and Gneiting, 2010). Because of these limitations, it may be desirable to approximate continuous distributions from the quantile forecasts to allow for more flexibility in scoring or ensemble building. In this section, we fit the QGP model for QM to every forecast of hospitalizations submitted to the FluSight during the 2023-24 season, and an analysis is made to compare the scoring of the quantile forecasts to QGP estimated forecasts using proper scoring rules.

Before fitting a QGP model to a quantile forecast we transformed the quantiles $\hat{Q}^{(H)}(\boldsymbol{p})$ to $\log(\hat{Q}^{(H)}(\boldsymbol{p}) + 1)$ so that forecasts for all states were on a similar scale. The model fit was the

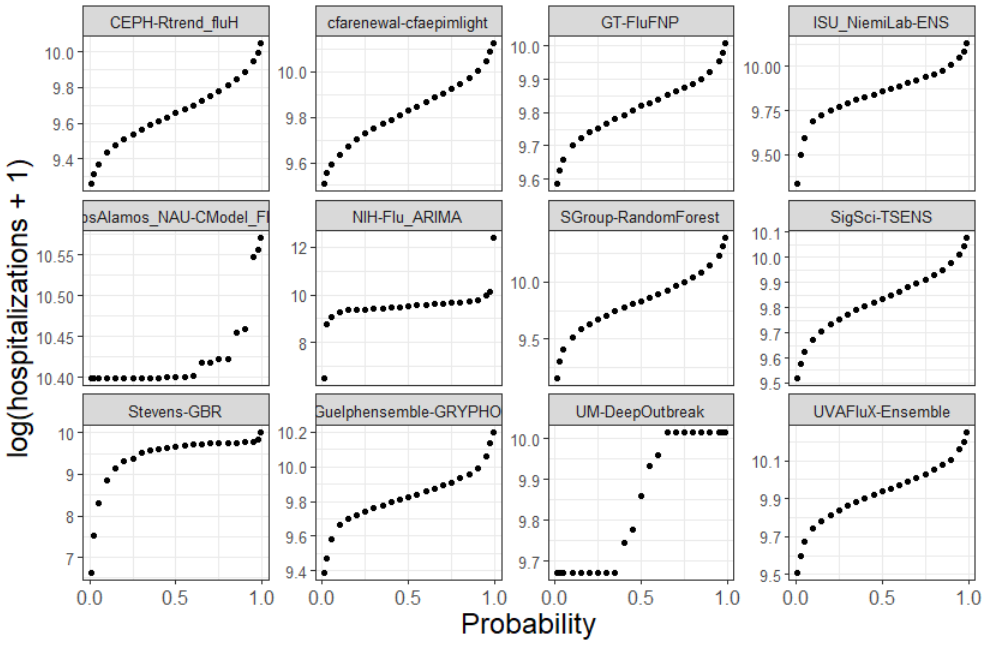

Figure 3.10: 1 week ahead log flu hospitalization quantile forecasts from 12 teams who participated in the 2023-24 CDC flu forecast competition. Forecasts are for the national level during the week of January 13, 2024.

same as the QGP fit in section 3.5.2, including the same prior distribution assignments. According to the official forecast competition rules, no forecast could include quantiles that were less than 0. As a result, there were forecasts with one or many quantiles equal to 0. When fitting the QGP model, these instances of 0 values were removed so that some forecasts had $K < 23$ quantiles. Between 39 competing forecast teams, 53 locations, 29 forecast dates, and 4 horizons, there were 180,312 quantile forecasts to which the QGP was fit. For each forecast, the WIS was calculated for the quantile forecast and the continuous ranked probability score (CRPS) was calculated over the predictive distribution of the fit QGP model.

The WIS and CRPS are both proper scoring rules, which is a class of scoring rule defined so as to keep a forecaster honest in their forecasts (Gneiting and Raftery, 2007; Gneiting and Katzfuss, 2014). The definition for the WIS is in (4.17) and is the same as that in Bracher et al. (2021). The WIS consists of the sum of multiple intervals scores (IS), the definition of which is in (4.18). Here α is the nominal level of an interval with l and r being the respective lower and

upper bounds of the interval. R is the number of intervals in the quantile forecast, α_r is the nominal level for the r^{th} interval, and y^* is the observed value which one is attempting to forecast.

$$WIS_{0,R}(F, y^*) = \frac{1}{R + 1/2} \times (w_0 \times |y^* - median| + \sum_{r=1}^{R} \{w_r \times IS_{\alpha_r}(F, y^*)\}) \tag{3.19}$$

$$IS_\alpha(l, r; y^*) = (r - l) + \frac{2}{\alpha}(l - y^*)\mathbb{1}\{y^* < l\} + \frac{2}{\alpha}(y^* - r)\mathbb{1}\{y^* > r\} \tag{3.20}$$

The CRPS is a widely used scoring rule for forecasting and is a function of the CDF of a continuous distribution thus making it unavailable for scoring quantile functions. The CRPS is defined in (4.12) where F is the CDF of a forecast and y^* is the observed event one attempts to forecast. The definition is the same as in Gneiting and Katzfuss (2014).

$$\text{CRPS}(F, y^*) = \int_{-\infty}^{\infty} (F(x) - \mathbb{1}(y^* \leq x))^2 dx \tag{3.21}$$

Fitting continuous distributions to the quantile forecasts would allow for forecast comparison using the CRPS. The CRPS assesses a forecast across an entire distribution, including the tails, which the WIS is unable to do, giving more reason why one may want to perform QM on a quantile forecast. We fit the QGP model in (3.12) to the forecast competition forecasts from the 2023-24 season and compare the results of scoring the given quantile forecasts by the WIS and the CRPS calculated from the posterior predictive of the QGP model. Posterior predictive samples from the QM forecasts allow for approximate calculation of the CRPS. To calculate the WIS, we use the `evalcast` `R` package (McDonald et al., 2023), and to calculate the CRPS we use the `scoringutils` package (Jordan et al., 2019).

Figures 3.11 and 3.12 show results of comparing the WIS from quantile forecasts and the CRPS from the QGP fitted forecasts. Figure 3.11 has in the x-axis the WIS and CRPS is in the y-axis. Each point is for one forecast for any participating team, season week, and state. The plot is faceted by horizon. Clearly the relationship between the quantile forecasts and the QGP forecasts are very close with high correlation near 0.9 for each horizon.

Figure 3.12 shows the ranking of teams under the quantile forecasts' WIS vs. the QGP forecasts' CRPS. Here we say a team is considered the best performing by having the lowest score

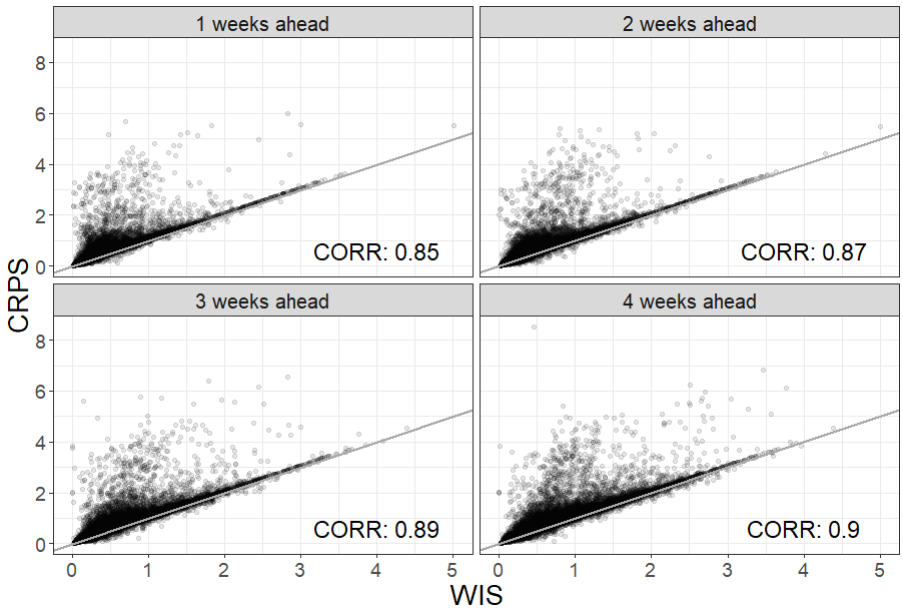

Figure 3.11: Scatterplots of WIS and CRPS values for forecasts from 39 different teams for 2023-24 CDC flu forecast competition. The plots are faceted by forecast horizon. Each point represents scores for a flu forecast for one week during the season, one state, and one competing team. Points are transparent to show where more scores tended to be. Overall linear correlation is also given in the corner of each plot.

averaged across all season weeks, states, and horizons. The x-axis indicates where one of the 39 teams ranks by WIS and the y-axis where they rank by CRPS. If a block is on the diagonal line, the team's ranking was the same under the WIS as under the CRPS. The darkness of the block indicates how correlated that team's WIS is to the CRPS over the whole season. Of the 39 teams represented, 16 have the same ranking for both scores, and for over half of all teams the ranking difference between WIS and CRPS is within 3 positions. For most teams the linear correlation between WIS and CRPS is very high with only five teams having a correlation below 0.85.

The results in this section show the close relationship between the quantile forecasts of the CDC flu forecast competition and the more complete forecast distributions estimated via QGP. The implications of these results include being able to score forecasts according to scoring rules used for continuous distributions, including more problem specific rules, and being able to combine forecasts into an ensemble according to sums of CDFs or PDFs. Increased flexibility in

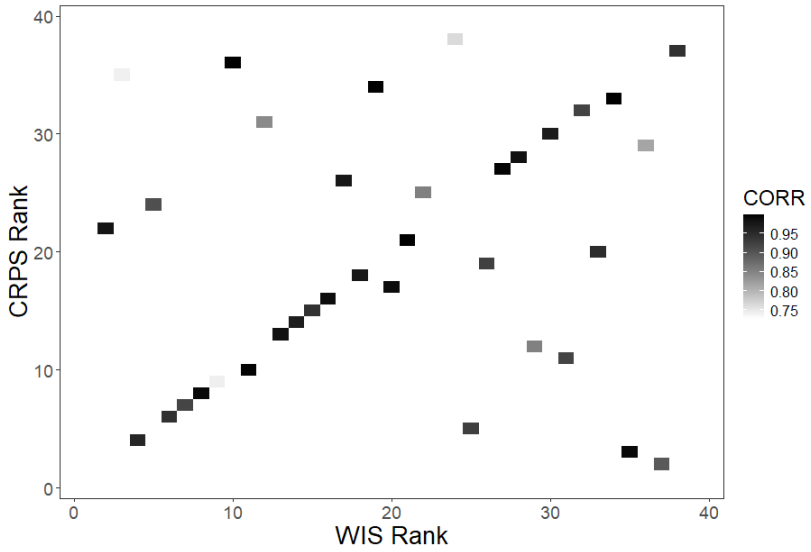

Figure 3.12: Blocks showing the overall season ranking of 39 competing forecast teams for WIS (x-axis) and CRPS (y-axis). Blocks on the diagonal line have the same ranking under both scores. Blocks are shaded to show correlation between WIS and CRPS for all forecasts submitted by that team.

scoring and ensemble construction could lead to improved decision making as problem specific evaluation becomes more accessible.

3.7 Conclusion

In this paper we consider the situation where quantiles with known probability levels are given as data or is given to represent a distribution rather than raw data or a fully defined distribution. We review basic properties of QFs, sample quantiles, and resulting CLTs which motivate a new model, the QGP model, for matching a set of quantiles to a continuous distribution. The QGP model is assessed for parameter estimation and inference, distribution approximation, and it is compared with other QM methods found in the literature. The simulation studies show that the QGP model does well in estimating parameters and allows for accurate measurement of uncertainty. It provides a way to perform QM for distributions defined by CDFs as well as quantile distributions where the CDF is difficult to evaluate. An application of the QGP for QM of quantile forecasts from the 2023-24 FluSight project shows how QGP matched distributions are closely related to the original quantile forecasts in terms forecast

performance among competing models. QGP forecasts, however, may be scored using a variety of scoring rules unavailable to quantile forecasts, and independent forecasts may be combined using methods made only for combining distributions by CDF or PDF functions. QGP has already been applied in chapter 4 of Wadsworth and Niemi (2024) who used the QGP predictive distributions fit in section 3.6 to construct ensemble forecasts.

Approximating continuous distributions given only a set of estimated quantiles is not a new idea, but QGP provides a method of doing so while also accurately accounting for asymptotic uncertainty in estimated or predicted quantiles. The CLTs in this paper are limited to the case where the quantiles are estimated given a distribution sample, however similar results have been shown for the quantile regression case where quantiles are estimated in the presence of covariates (Kocherginsky et al., 2005; Koenker and Bassett Jr, 1978). These results could lead to additional QM modeling using the QGP model where covariates are given, and this may be more akin to the QM done by Sgouropoulos et al. (2015). Additional research for the QGP model may be in how the required distribution function is selected. Where a true model is unknown, we elected to model a distribution as a normal mixture distribution. Of course this has its limits, and using other non parametric functions, similar to Gasthaus et al. (2019), may provide needed flexibility. As long as the selected function is continuous and once differentiable, the theory herein applies.

The application in this paper suggests QM being useful for forecast hubs. Gerding et al. (2023) found QM a necessity in order to evaluate forecasts by a new scoring rule with COVID-19 specific applications. The extra work that comes with QM may or may not be worth the effort for any given forecast hub, but in some contexts it is something worth considering.

3.8 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.

3.9 References

Alvarez, L. and Orestes, V. (2023). Quantile mixture models: Estimation and inference. Technical report, Working paper.

Baran, S. and Lerch, S. (2018). Combining predictive distributions for the statistical post-processing of ensemble forecasts. *International Journal of Forecasting*, 34(3):477–496.

Belgorodski, N., Greiner, M., Tolksdorf, K., and Schueller, K. (2017). *rriskDistributions: fitting distributions to given data or known quantiles*. R package version 2.1.2.

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.

Blei, D. M., Kucukelbir, A., and McAuliffe, J. D. (2017). Variational inference: A review for statisticians. *Journal of the American Statistical Association*, 112(518):859–877.

Bogner, K., Liechti, K., and Zappa, M. (2017). Combining quantile forecasts and predictive distributions of streamflows. *Hydrology and Earth System Sciences*, 21(11):5493–5502.

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.

Casella, G. and Berger, R. L. (2002). *Statistical Inference*. Duxbury Advanced Series.

CDC (2022). CDC Extended BMI-for-age growth charts. https://www.cdc.gov/growthcharts/extended-bmi.htm. Accessed: 2024-10-22.

Chernozhukov, V., Fernández-Val, I., and Galichon, A. (2010). Quantile and probability curves without crossing. *Econometrica*, 78(3):1093–1125.

Chow, W. C. (2009). Brownian bridge. *Wiley Interdisciplinary Reviews: Computational Statistics*, 1(3):325–332.

Chung, Y., Neiswanger, W., Char, I., and Schneider, J. (2021). Beyond pinball loss: Quantile methods for calibrated uncertainty quantification. *Advances in Neural Information Processing Systems*, 34:10971–10984.

Cox, D. R. and Snell, E. J. (1968). A general definition of residuals. *Journal of the Royal Statistical Society: Series B (Methodological)*, 30(2):248–265.

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. (2022a). The United States COVID-19 forecast hub dataset. *Scientific Data*, 9(1):462.

Cramer, E. Y., Ray, E. L., Lopez, V. K., Bracher, J., Brennen, A., Castro Rivadeneira, A. J., Gerding, A., Gneiting, T., House, K. H., Huang, Y., et al. (2022b). Evaluation of individual and ensemble probabilistic forecasts of COVID-19 mortality in the United States. *Proceedings of the National Academy of Sciences*, 119(15):e2113561119.

Cramér, H. (1951). *Mathematical Methods of Statistics*. Princeton University Press.

Dilger, M., Schneider, K., Drossard, C., Ott, H., and Kaiser, E. (2022). Distributions for time, interspecies and intraspecies extrapolation for deriving occupational exposure limits. *Journal of Applied Toxicology*, 42(5):898–912.

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

Gasthaus, J., Benidis, K., Wang, Y., Rangapuram, S. S., Salinas, D., Flunkert, V., and Januschowski, T. (2019). Probabilistic forecasting with spline quantile function RNNs. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 1901–1910. PMLR.

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.

Gerding, A., Reich, N. G., Rogers, B., and Ray, E. L. (2023). Evaluating infectious disease forecasts with allocation scoring rules. *arXiv preprint arXiv:2312.16201*.

Gibbs, A. L. and Su, F. E. (2002). On choosing and bounding probability metrics. *International Statistical Review*, 70(3):419–435.

Gilchrist, W. (2000). *Statistical Modelling with Quantile Functions*. Chapman and Hall/CRC.

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

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., Wolffram, D., Resin, J., Kraus, K., Bracher, J., Dimitriadis, T., Hagenmeyer, V., Jordan, A. I., Lerch, S., Phipps, K., et al. (2023). Model diagnostics and forecast evaluation for quantiles. *Annual Review of Statistics and Its Application*, 10:597–621.

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.

He, X. (1997). Quantile curves without crossing. *The American Statistician*, 51(2):186–192.

He, Y., Xu, Q., Wan, J., and Yang, S. (2016). Short-term power load probability density forecasting based on quantile regression neural network and triangle kernel function. *Energy*, 114:498–512.

HealthData.gov (2024). COVID-19 reported patient impact and hospital capacity by state (raw). https://healthdata.gov/dataset/COVID-19-Reported-Patient-Impact-and-Hospital-Capa/6xf2-c3ie/about_data. Accessed: 2024-10-22.

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.

(https://stats.stackexchange.com/users/20519/zhanxiong), Z. (2024). Upper bound for 1-Wasserstein distance between standard uniform and other distribution on $[0, 1]$. Cross Validated. URL:https://stats.stackexchange.com/q/645854 (version: 2024-04-26).

Hu, Y. and Scarrott, C. (2018). evmix: An R package for extreme value mixture modeling, threshold estimation and boundary corrected kernel density estimation. *Journal of Statistical Software*, 84(5):1–27.

Hyndman, R. J. and Fan, Y. (1996). Sample quantiles in statistical packages. *The American Statistician*, 50(4):361–365.

Joiner, B. L. and Rosenblatt, J. R. (1971). Some properties of the range in samples from Tukey's symmetric lambda distributions. *Journal of the American Statistical Association*, 66(334):394–399.

Jordan, A., Krüger, F., and Lerch, S. (2019). Evaluating probabilistic forecasts with scoringRules. *Journal of Statistical Software*, 90(12):1–37.

Keelin, T. W. (2016). The metalog distributions. *Decision Analysis*, 13(4):243–277.

Kocherginsky, M., He, X., and Mu, Y. (2005). Practical confidence intervals for regression quantiles. *Journal of Computational and Graphical Statistics*, 14(1):41–55.

Koenker, R. (2017). Quantile regression: 40 years on. *Annual Review of Economics*, 9(1):155–176.

Koenker, R. and Bassett Jr, G. (1978). Regression quantiles. *Econometrica*, 46(1):33–50.

Kucukelbir, A., Ranganath, R., Gelman, A., and Blei, D. (2015). Automatic variational inference in Stan. *Advances in Neural Information Processing Systems*, 28.

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.

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.

McDonald, D., Bien, J., O'Brien, M., Grabman, J., Colquhoun, S., Narasimhan, B., and Tibshirani, R. (2023). *evalcast: Tools For Evaluating COVID Forecasters*. https://cmu-delphi.github.io/covidcast/evalcastR/, https://github.com/cmu-delphi/covidcast.

McLachlan, G. and Peel, D. (2000). *Finite Mixture Models*. John Wiley & Sons, Inc.

Nguyen, H. D. and McLachlan, G. (2019). On approximations via convolution-defined mixture models. *Communications in Statistics-Theory and Methods*, 48(16):3945–3955.

Nguyen, T. T., Nguyen, H. D., Chamroukhi, F., and McLachlan, G. J. (2020). Approximation by finite mixtures of continuous density functions that vanish at infinity. *Cogent Mathematics & Statistics*, 7(1):1750861.

Nirwan, R.-S. and Bertschinger, N. (2020). Bayesian quantile matching estimation. *arXiv preprint arXiv:2008.06423*.

Panaretos, V. M. and Zemel, Y. (2019). Statistical aspects of Wasserstein distances. *Annual Review of Statistics and its Application*, 6(1):405–431.

Parzen, E. (2004). Quantile probability and statistical data modeling. *Statistical Science*, 19(4):652–662.

Perepolkin, D., Goodrich, B., and Sahlin, U. (2023). The tenets of quantile-based inference in bayesian models. *Computational Statistics & Data Analysis*, 187:107795.

Pohle, M.-O. (2020). The murphy decomposition and the calibration-resolution principle: A new perspective on forecast evaluation. *arXiv preprint arXiv:2005.01835*.

Ramberg, J. S. and Schmeiser, B. W. (1974). An approximate method for generating asymmetric random variables. *Communications of the ACM*, 17(2):78–82.

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

Ray, E. and Gerding, A. (2024). *distfromq: Reconstruct a Distribution from a Collection of Quantiles*. R package version 1.0.4.

Sgouropoulos, N., Yao, Q., and Yastremiz, C. (2015). Matching a distribution by matching quantiles estimation. *Journal of the American Statistical Association*, 110(510):742–759.

Shandross, L., Howerton, E., Contamin, L., Hochheiser, H., Krystalli, A., of Infectious Disease Modeling Hubs, C., Reich, N. G., and Ray, E. L. (2024). hubEnsembles: Ensembling methods in R. *medRxiv*, pages 2024–06.

Sherratt, K., Gruson, H., Johnson, H., Niehus, R., Prasse, B., Sandmann, F., Deuschel, J., Wolffram, D., Abbott, S., Ullrich, A., et al. (2023). Predictive performance of multi-model ensemble forecasts of COVID-19 across European nations. *Elife*, 12:e81916.

Simpson, M., Holan, S. H., Wikle, C. K., and Bradley, J. R. (2023). Interpolating population distributions using public-use data: An application to income segregation using American Community Survey data. *Journal of the American Statistical Association*, 118(541):84–96.

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

Staudte, R. G. (2017). The shapes of things to come: Probability density quantiles. *Statistics*, 51(4):782–800.

Stephens, M. (2000). Dealing with label switching in mixture models. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 62(4):795–809.

Tukey, J. W. (1960). The practical relationship between the common transformations of percentages of counts and of amounts. *Statistical techniques research group technical report*, 36.

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*.

Walker, A. (1968). A note on the asymptotic distribution of sample quantiles. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 30(3):570–575.

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.

Wilkinson, D. J. (2018). *Stochastic Modelling for Systems Biology*. Chapman and Hall/CRC.

Yang, L. (2024). Double probability integral transform residuals for regression models with discrete outcomes. *Journal of Computational and Graphical Statistics*, 33:1–17.

3.A Additional simulation study for exponential QGP

Here we include a simulation study similar to the first study in section 3.5. However, instead of simulating data from a normal distribution, data is simulated from an exponential distribution with parameter $\lambda = 4$. The same prior distribution for λ was used in fitting QGP, ORD, and IND models, and the same prior distribution was assigned to n in QGP and ORD models and $1/\sigma_\rho$. The prior for λ was $\pi(\lambda) \sim N(0, 7^2)\mathbb{1}\{\lambda > 0\}$, the prior for n was $\pi(n) \sim N(0, 3000^2)\mathbb{1}\{n > 0\}$, and the prior for $1/\sigma_\rho$ was $\pi(1/\sigma_\rho) \sim N(0, 3000^2)\mathbb{1}\{1/\sigma_\rho > 0\}$.

Figure 13 shows examples of posterior sample density plots for λ and n. QGP and ORD show similar distributions which tend to be a bit wider than the posterior distributions from the IND model. The left side of figure 14 shows the percent coverage of the 500 simulation replicates for λ and n. Here, the IND model coverage is much closer to the nominal level than in the normal distribution case of the previous section. The right side of figure 14 shows the average UWD1, TV, and KLD distances for the five QM methods. Again the parametric models tend to perform better than the non-parametric SPL and KDE methods with the SPL and KDE performing better as n and K increase.

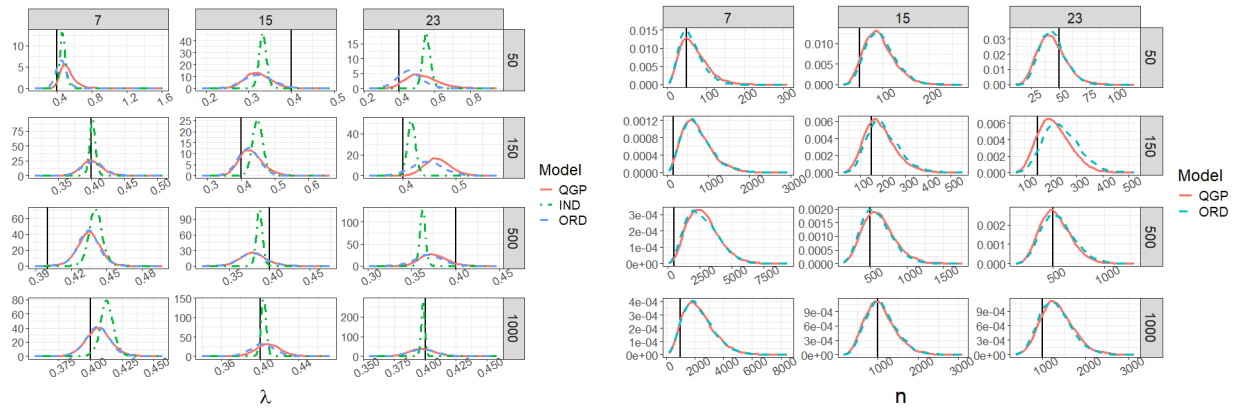

Figure 13: Density plots of posterior distribution samples for the exponential parameters by QM for QGP, ORD, and IND models. QM was done on estimated quantiles from a exponential distribution with parameter $\lambda = 4$. The posterior densities are for λ (left) and sample size n (right). Plots are faceted by true sample size $n \in \{50, 150, 500, 1{,}000\}$ (y-axis) and number of quantiles $K \in \{7, 15, 23\}$ (x-axis). Vertical lines (black) show the value of the true parameter.

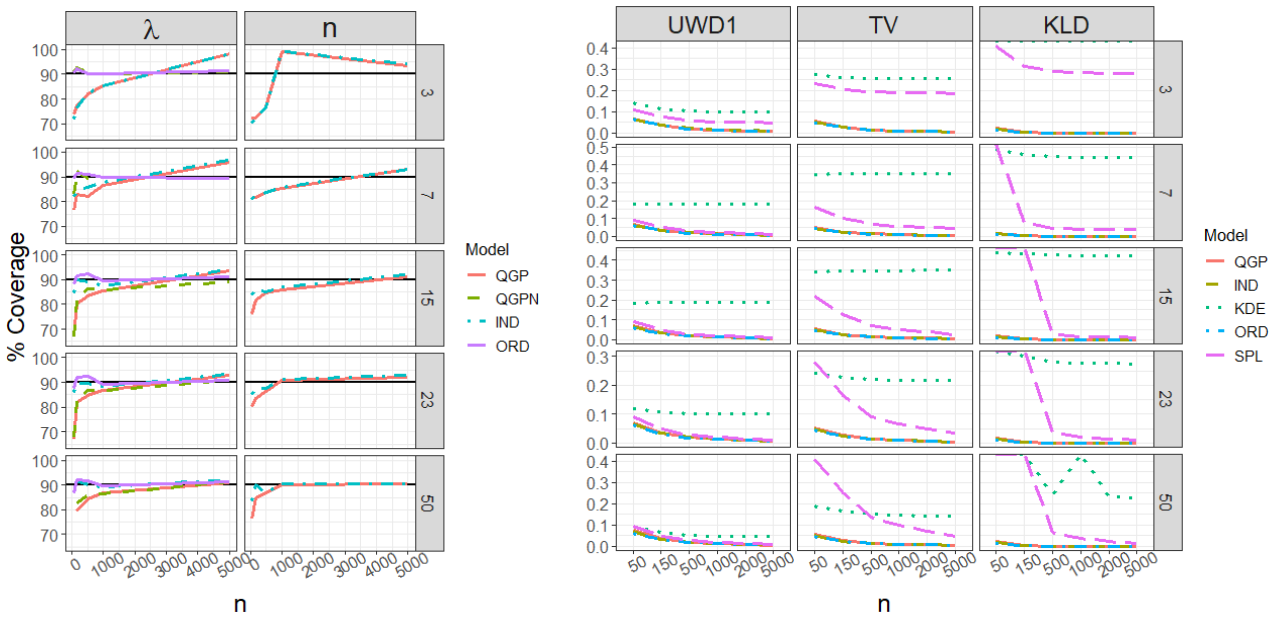

Figure 14: Posterior coverage (left) calculated as the percentage of times the true parameter fell within the modeled 90% credible interval over the 500 replications. Coverage is faceted by the exponential parameter λ and n with $K \in \{3, 7, 15, 23, 50\}$, and by increasing sample size (x-axis). The five models QGP, ORD, QGPN, ORDN, and IND are colored as shown the legend. The horizontal line (black) is at the nominal 90% level. Only QGP and ORD appear for the parameter n as they are the only two which estimate an unknown n. Distance between the true distribution and the estimated QM predictive distribution (right) averaged over the 500 replications. Distances include UWD1, TV, and KLD for $K \in \{3, 7, 15, 23, 50\}$, and by increasing sample size (x-axis).

CHAPTER 4. BAYESIAN STACKING VIA PROPER SCORING RULE OPTIMIZATION USING A GIBBS POSTERIOR

Spencer Wadsworth and Jarad Niemi

Department of Statistics, Iowa State University, Ames, IA 50011

4.1 Abstract

Combining probabilistic forecasts into a single ensemble forecast has become standard practice in collaborative forecast projects in many fields with linear pooling and quantile averaging being the most commonly used methods. A common idea is that weight selection methods should be tailored to the specific research question, and this has led to the use of selecting weights via optimization of proper scoring rules. Bayesian predictive synthesis has also emerged as a model probability updating scheme which provides a Bayesian solution to selecting model weights which is much more flexible than standard Bayesian model averaging. The various existing methods may or may not improve forecasting for any given dataset, and room for additional methodology may always exist. In this manuscript, we introduce a Gibbs posterior on stacked model weights based on minimizing the continuous ranked probability score, a popular proper scoring rule. The Gibbs posterior extends model stacking into a more probabilistic framework by allowing for uncertainty quantification of weights and for optimal solutions to be influenced by a prior distribution. We provide a result on the posterior asymptotic consistency of the stacked Gibbs posterior under the independent and identically distributed data assumption. We compare ensemble forecast performance with two versions of 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 flu forecasting competition. In both the simulation studies and the

real data analysis, the stacked Gibbs posterior produces ensemble forecasts which perform better than the ensemble forecasts constructed by the other methods considered.

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

4.2 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 is 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 4.1 shows

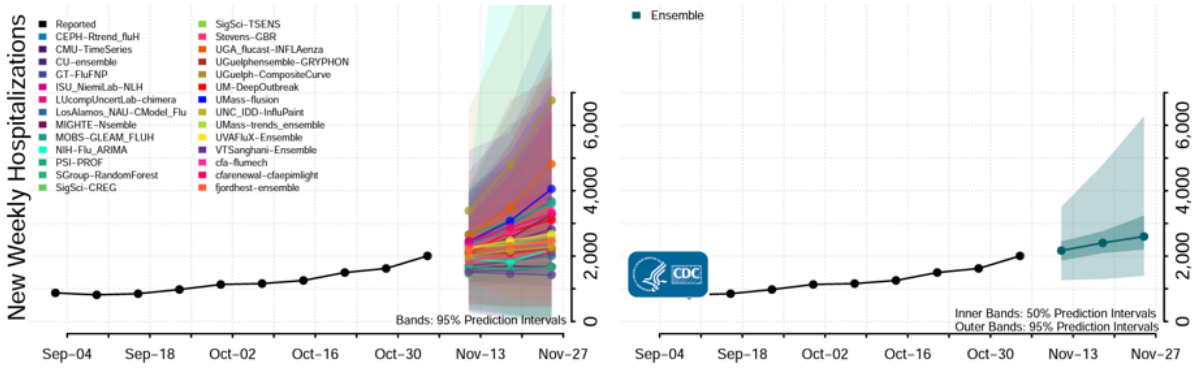

Figure 4.1: United States flu hospitalization forecasts from multiple competing forecast teams (left) for 1-4 week ahead horizons from the week of November 6, 2023 and an ensemble forecast (right) made by combining all competing forecasts into one. Image downloaded from https://www.cdc.gov/flu/weekly/flusight/flu-forecasts.htm

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.

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. These methods, however, do not provide uncertainty quantification of the estimated weights. Weights in an ensemble may also be selected by maximizing a likelihood (Raftery et al., 2005) or via a Bayesian posterior distribution, but the resulting ensemble forecasts may not be suited for the problem specific criteria under which forecasts are assessed, i.e. the minimization of a problem specific loss function. Bayesian model probabilities may be used as weights, but selecting weights this way may also not be well suited to problem specific needs. An important case of linear pooling is the simple average of participating 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 4.3 includes a brief review of probabilistic forecasts and linear pool combinations of forecasts. In section 4.4 the SGP is developed, and important details of the Gibbs posterior and proper scoring rules are covered. Section 4.5 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 4.6 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 4.7.

4.3 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 (Wadsworth et al., 2023; 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 $\boldsymbol{y}_t$. 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).

4.3.1 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 (4.1). 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 $\sum_{c=1}^{C} w_c = 1$.

$$\bar{P}_{\omega}(\tilde{y}_{t+h}) = \sum_{c=1}^{C} w_c P_c(\tilde{y}_{t+h}) \tag{4.1}$$

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 (4.1) 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 estimator for combination weights consists of leave-one-out 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 4.4. 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 (4.1) 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

$$p_{BMA}(\mathcal{M}_c|\boldsymbol{y}_t) = \frac{p(\boldsymbol{y}_t|\mathcal{M}_c)p(\mathcal{M}_c)}{\sum_{j=1}^{C} p(\boldsymbol{y}_t|\mathcal{M}_j)p(\mathcal{M}_j)} \tag{4.2}$$

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 (4.1) where the c^{th} forecast model is $P_c(y_{t+h}) = P(y_{t+h}|\mathcal{M}_c)$ 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 4.5 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

$$p_{AVS}(\mathcal{M}_c|\boldsymbol{y}_t) \propto \exp\{-\eta S(\mathcal{M}_c)\}p(\mathcal{M}_c) \tag{4.3}$$

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 4.5 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 (4.1) 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 been recommended to use an EQW average as a baseline by which other methods should be compared (Li et al., 2023; Clemen, 1989).

4.4 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 (4.1). The SGP is defined in (4.4) 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 4.4.1 and 4.4.2.

$$\pi_n^{(\eta)}(\omega) \propto \exp\{-\eta n S_n(\omega)\}\pi(\omega) \tag{4.4}$$

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 (4.2) and (4.3). 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.

4.4.1 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 (4.5) where $\in$ indicates that θ^* may not be unique.

$$\theta^* \in \underset{\theta \in \Theta}{\operatorname{argmin}} R(\theta) \tag{4.5}$$

Since $R(\theta)$ is inaccessible, an empirical risk function $R_n(\theta)$ based on the given data is formulated as in (4.6). Then $\hat{\theta}_n$, an estimate of θ^*, is estimated by the estimator in (4.7).

$$R_n(\theta) = \frac{1}{n} \sum_{i=1}^{n} l_\theta(y_i) \tag{4.6}$$

$$\hat{\theta}_n = \underset{\theta \in \Theta}{\operatorname{argmin}} R_n(\theta) \tag{4.7}$$

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 (4.8) where R_n is the empirical risk from (4.6). 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*.

$$\pi_n^{(\eta)}(\theta) \propto \exp\{-\eta n R_n(\theta)\}\pi(\theta) \tag{4.8}$$

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 (4.4), 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 on two loss functions. The first of these functions is in (4.9), 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.

$$S_n(\omega) = G_n(\omega) = \frac{1}{n}\sum_{i=1}^{n} G(\bar{P}_\omega, y_i). \tag{4.9}$$

Here $\bar{P}_\omega$ is the CDF of an ensemble forecast with C competing forecast models from (4.1). 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 4.4.2). A second empirical risk, used in a time dynamic setting is given in (4.10). Here we have an observed time series $y_1, ..., y_T$, and we want to construct an ensemble model for forecasting the future observation $\tilde{y}_{T+h}$.

$$S_T(\omega) = G_T(\omega) = \frac{1}{T}\sum_{t=1}^{T} \alpha^{T-t} G(\bar{P}_{t,\omega}, y_t) \tag{4.10}$$

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 4.5, the risk in (4.10) is used where competing models are fixed in time whereas in the analysis of flu forecasts in section 4.6, the forecasts vary in time.

4.4.2 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 (4.11) and evaluates a distribution by its PDF p.

$$\text{LogS}(p, y) = -\log(p(y)) \tag{4.11}$$

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 (4.12) and evaluates a distribution through its CDF P.

$$\text{CRPS}(P, y) = \int_{-\infty}^{\infty} (P(x) - \mathbb{1}(y \leq x))^2 dx \tag{4.12}$$

For P belonging to certain distribution families, there may be a closed form solution to (4.12) (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 (4.13).

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

$$\begin{aligned} \alpha_{c,c'} =& -\frac{1}{\sqrt{2\pi}} \sqrt{\sigma_c^2 + \sigma_{c'}^2} \exp\left(\frac{(\mu_c - \mu_{c'})^2}{2(\sigma_c^2 + \sigma_{c'}^2)}\right) \\ & - \frac{\mu_c - \mu_{c'}}{2} \left[2\Phi\left(\frac{(\mu_c - \mu_{c'})}{\sqrt{\sigma_c^2 + \sigma_{c'}^2}}\right) - 1\right] \\ \beta_c =& \sqrt{\frac{2}{\pi}} \sigma_c \exp\left(-\frac{(\mu_c - y)^2)}{2\sigma_c^2}\right) \\ & + (\mu_c - y)\left[2\Phi\left(\frac{\mu_c - y}{\sigma_c}\right) - 1\right] \end{aligned} \tag{4.13}$$

The CRPS has some general forms other than (4.12) including (4.14) which was shown by Gneiting and Raftery (2007).

$$\text{CRPS}(P, y) = E|X - y| - \frac{1}{2} E|X - X'| \tag{4.14}$$

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 (4.14). We also note that if P takes the form of an ensemble forecast where all components are continuous distributions, then (4.14) can be further reduced to (4.15). A simple proof for this is

in appendix 4.A.1.

$$CRPS(\bar{P}, y) = \sum_{c=1}^{C} w_c E|X_c - y| - \frac{1}{2}\sum_{c=1}^{C}\sum_{c'=1}^{C} w_c w_{c'} E|X_c - X_{c'}| \tag{4.15}$$

Li et al. (2019) showed this result for all mixture components being Gaussian, but it is 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 (4.4), we let the function $G(\cdot, \cdot) = CRPS(\cdot, \cdot)$ and G_n take the form of either (4.9) or (4.10).

4.4.3 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 4.1 which is the same definition used by Martin and Syring (2022).

Definition 4.1. *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 (4.4) with the empirical risk function in (4.9) 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 (4.9), posterior consistency of the SGP proposed in (4.4) is affirmed by theorem 4.1.

Theorem 4.1. *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., nor does it include results for forecast models which are updated given more 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.

4.5 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.

4.5.1 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

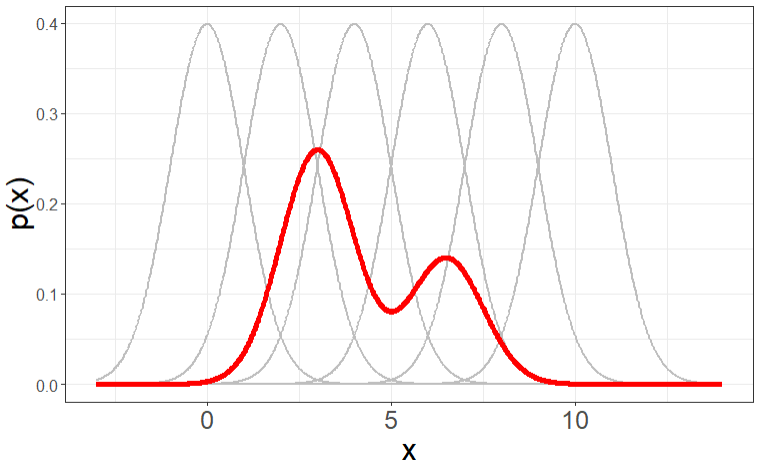

Figure 4.2: Mixture distribution density function from which data are simulated (red) and 6 candidate normal predictive densities (grey). The mixture distribution is $\nu \times N(3,1) + (1-\nu) \times N(6.5,1); w = 0.65)$. The candidate predictive distributions are each normal with variance 1 and means 0, 2, 4, 6, 8, and 10.

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 (4.5) is used, and the consistency results from theorem 4.1 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 4.2 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 4.3 shows an example of how the CRPS varies for different values of η for the AVS and SGP. In the case of AVS, leave-one-out 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 4.4 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

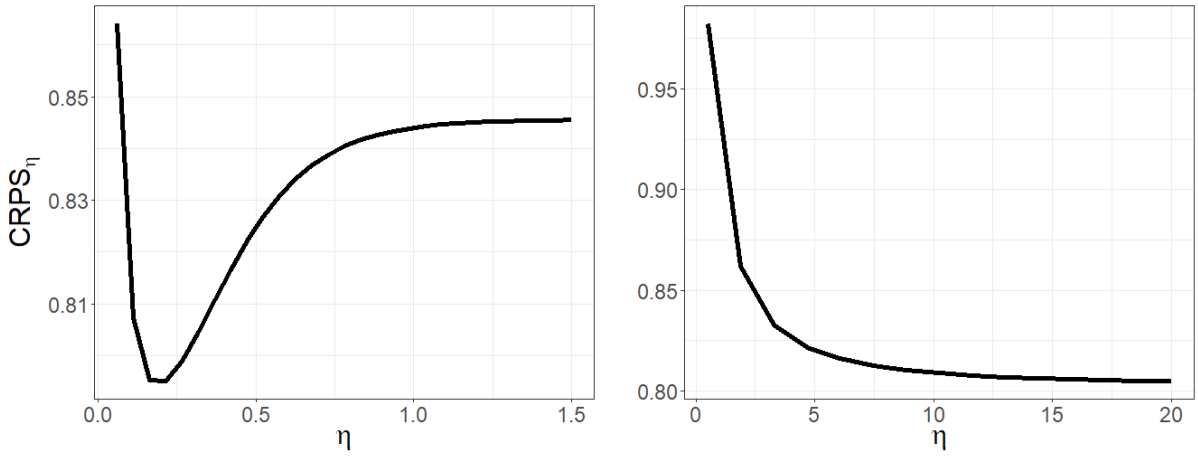

Figure 4.3: An example of CRPS values for different values of η after optimizing competing model weights for AVS (left) and for SGP (right).

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 4.5 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 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.

4.5.2 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

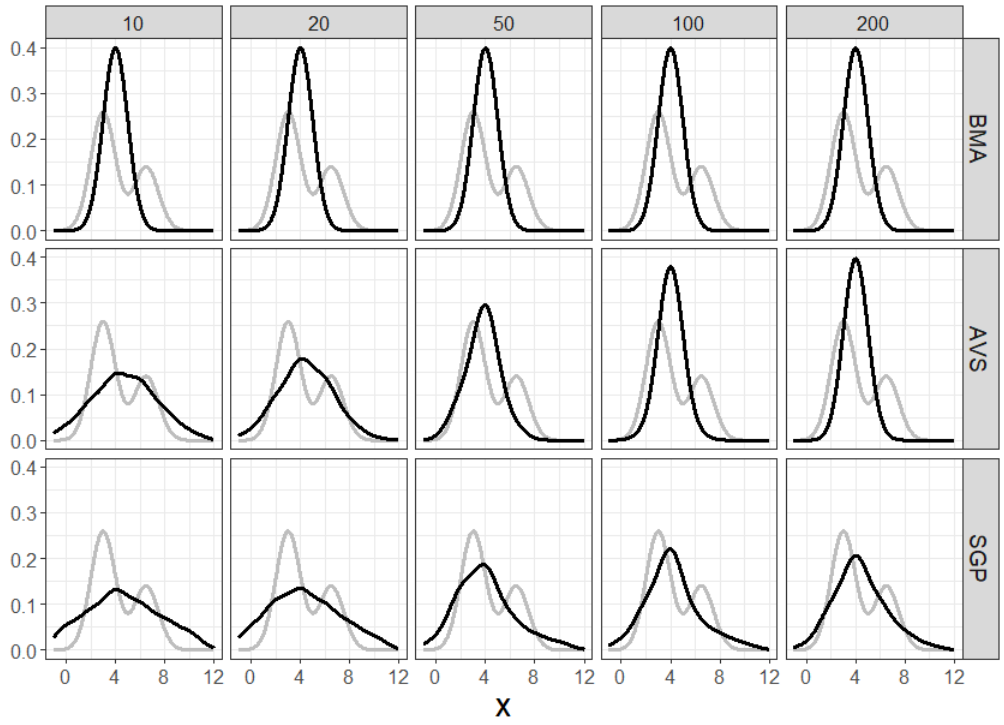

Figure 4.4: Examples of estimated densities after weighting of competing models faceted by weighting method and sample size. The estimated densities (black) overlay the true model density (grey).

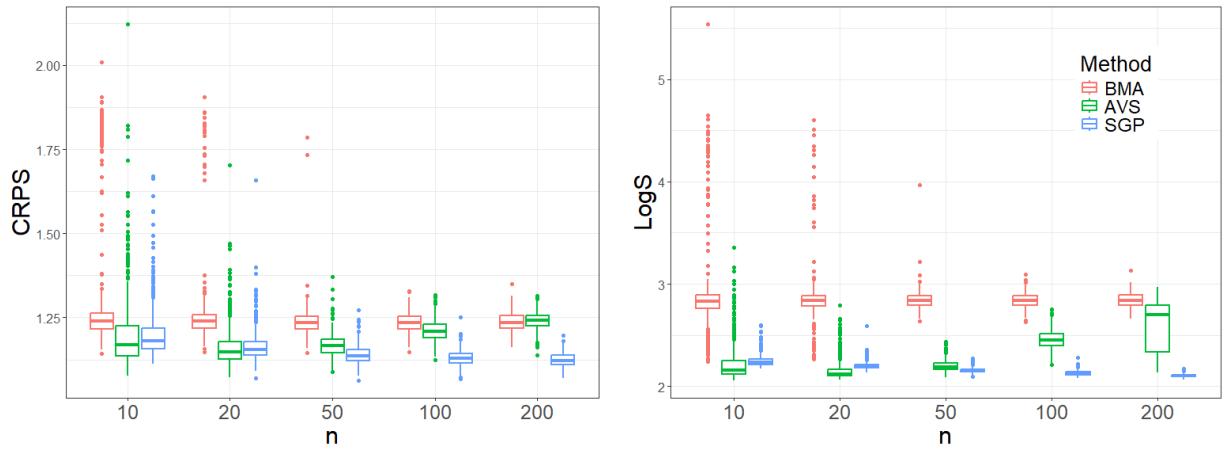

Figure 4.5: Boxplots of the mean of 1,000 Monte Carlo CRPSs (left) and LogSs (right) for 500 replicates. Plots are colored by weighting method.

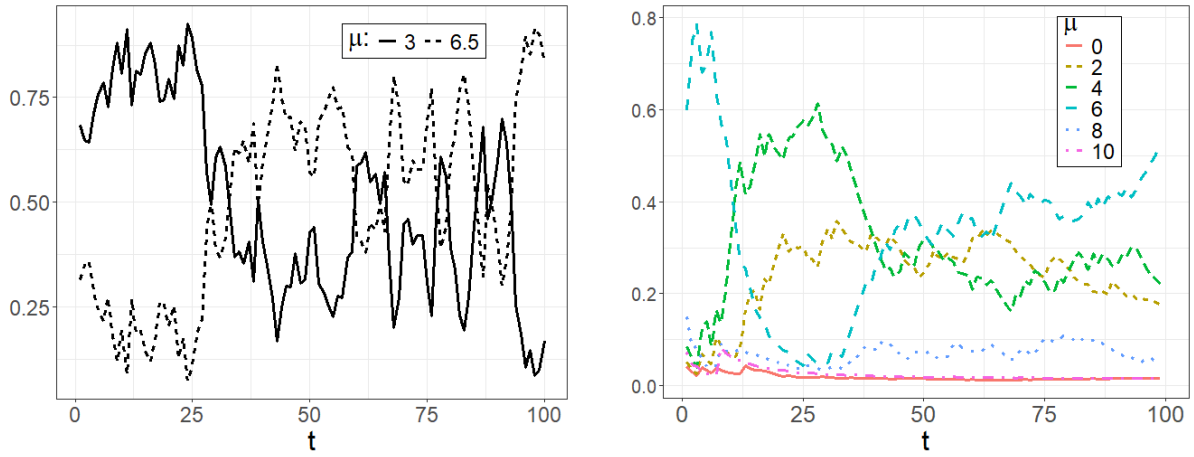

Figure 4.6: An example of simulated component weights for 100 time steps. The figure shows the simulated weights for the two components of the true distribution (left) and the weights optimized under SGP for the six competing component models (right).

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 (4.16). 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 point according to a random walk process.

$$\begin{aligned} Y_t &\sim \omega_{1,t} \times N(3,1) + \omega_{2,t} \times N(6.5,1) \\ \omega_{i,t} &= \frac{\exp(z_{i,t})}{\sum_a \exp(z_{i,t})}; \quad i \in \{1,2\} \\ \boldsymbol{z}_t &\sim N(\boldsymbol{z}_{t-1}, \sigma^2 I) \\ (\omega_{1,1}, \omega_{2,1}) &= (0.65, 0.35) \\ \sigma^2 &= 0.01 \end{aligned} \tag{4.16}$$

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 4.6 gives an example of the dynamic weights simulated from (4.16). 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

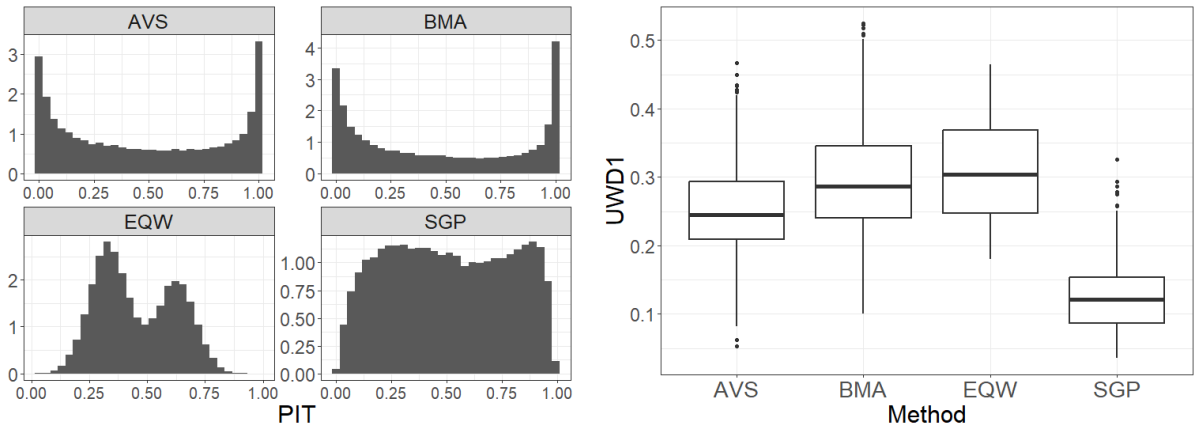

Figure 4.7: Plots for assessing calibration of one step ahead forecasts for the four weighting methods. The figure shows the PIT for all 99×500 simulated one-step-ahead forecasts for AVS, BMA, EQW, and SGP (left) and boxplots of the UWD1 distance between a standard uniform distribution and the empirical distribution of the PIT for the weighting schemes for 99 predictions (right).

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 4.7 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 4.8 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.

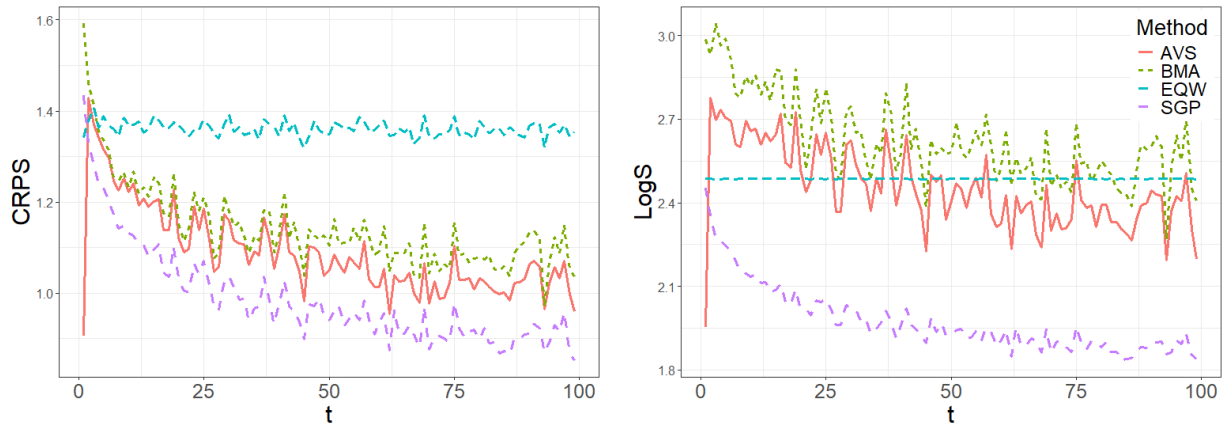

Figure 4.8: The mean CRPS (left) and LogS (right) over 500 replicates of one step ahead predictions at 99 time points colored by weighting methods.

The two studies in this section show the SGP's ability to construct accurate forecasts for 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.

4.6 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 (4.17) (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 (4.18), and y^* is the true observation targeted by P.

$$WIS_{0,I}(P, y^*) = \frac{1}{I + 1/2} \times (v_0 \times |y^* - median| + \sum_{k=1}^{K} \{v_i \times IS_{\alpha_i}(P, y^*)\}) \tag{4.17}$$

$$IS_\alpha(l, r; y^*) = (r - l) + \frac{2}{\alpha}(l - y^*)\mathbb{1}\{y^* < l\} + \frac{2}{\alpha}(y^* - r)\mathbb{1}\{y^* > r\} \tag{4.18}$$

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 (Wadsworth et al., 2023). In order to apply the SGP from (4.4) 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 (4.15) were calculated from the posterior predictive samples allowing for estimation of the CRPS for individual forecasts and ensemble forecasts. Then, the SGP was used to select weights and construct 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 (4.10) 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 $\bar{w}_{c,t} = M^{-1}\sum_{m=1}^{M} w_{c,t}^{(m)}$ 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 $\log(\eta)$ on an equally spaced grid of 20 values between -1 and 5. 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 evaluated at the log hospitalization forecast distribution 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 4.9 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 4.10 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

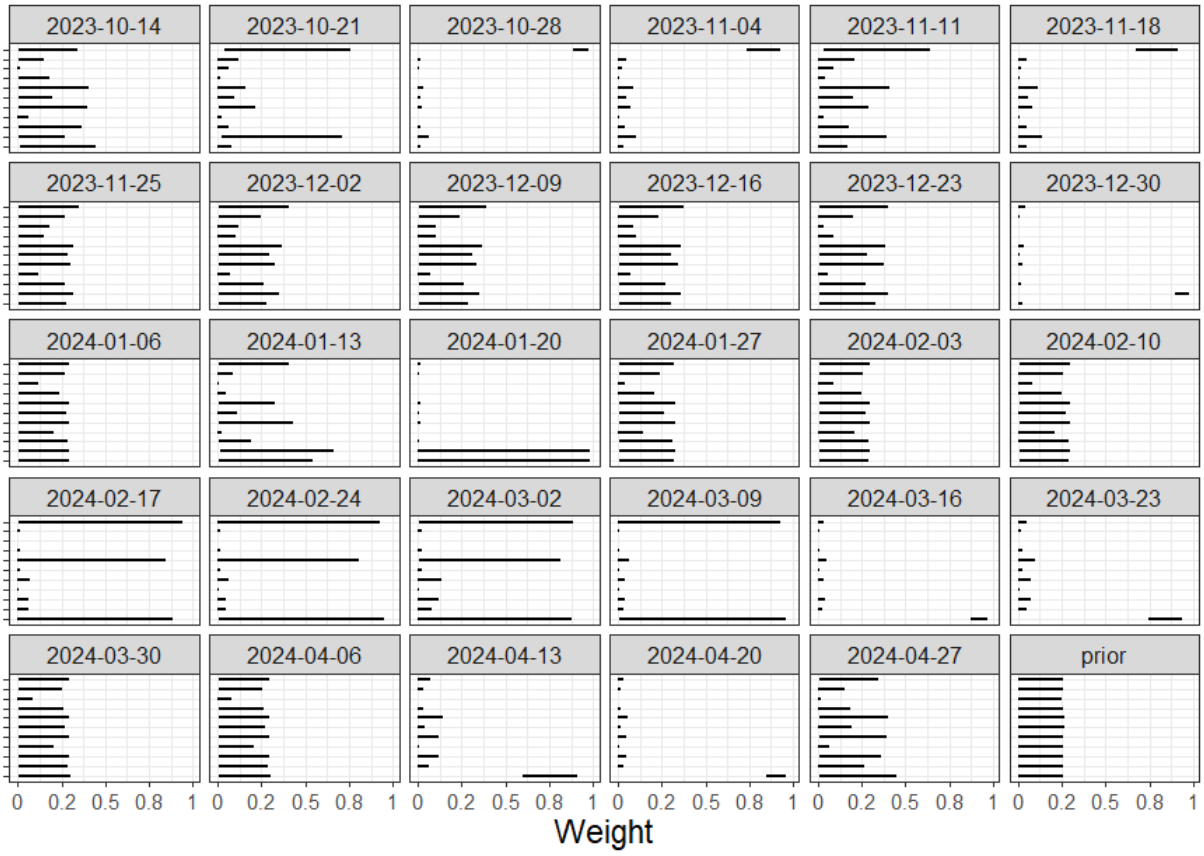

Figure 4.9: Posterior 90% credible intervals for forecast weights estimated via SGP for the 2023-24 CDC FluSite targetting flu hospitalizations at the national level of the United States. The intervals in the bottom right corner are those from the uninformative Dirichlet prior.

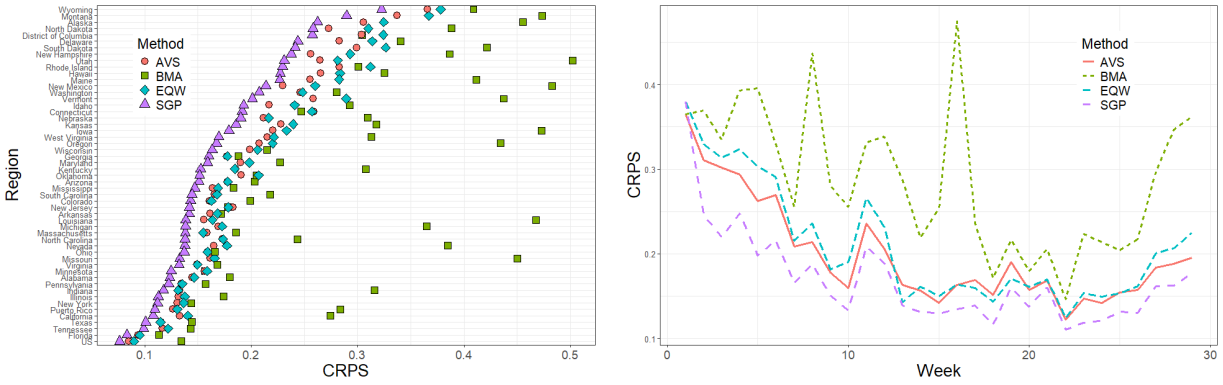

Figure 4.10: Mean over regions CRPS (left) for the four ensemble methods separated by shape and color, and mean over weeks CRPS (right) for the four methods

Table 4.1: 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

forecasts. Where the ensemble forecasts are evaluated by week, there is only one week where the SGP does not outperform the other methods.

Table 4.1 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 (4.10) 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 (4.15).

Of the methods compared, the SGP is arguably the most complicated, but in forecast performance it greatly outperformed the other methods.

4.7 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 (4.1) 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 4.1 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 4.6, 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.

4.8 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.

4.9 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., Gneiting, T., Balabdaoui, F., and Polakowski, M. (2005). Using Bayesian model averaging to calibrate forecast ensembles. *Monthly weather review*, 133(5):1155–1174.

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.

4.A Proofs of equation (4.15) and theorem 4.1

4.A.1 Proof of equation (4.15)

Proof. The random variables X and X' in (4.14) 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 (4.14) 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 (4.15) follows. □

4.A.2 Proof of theorem 4.1

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

ii) Ω is compact

iii) $CRPS(\bar{P}_\omega, y)$ is continuous at each $\omega \in \Omega$ with probability 1

iv) 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. □

CHAPTER 5. GENERAL CONCLUSION

In this dissertation, we introduced new methodology covering several aspects of probabilistic forecasting in collaborative hubs. All of the methodology was motivated by the CDC flu forecasting competition known as FluSight. The methodology touches on different aspects of collaborative forecast hubs including modeling for forecasting of flu hospitalizations, quantile matching for estimating continuous probabilities given sample quantiles, and optimally combining multiple forecasts into an ensemble forecast.

In chapter 2 we introduce a two component modeling framework applied to FluSight. The modeling framework allows for the incorporation of forecasts of ILI data for use in a linear model of flu hospitalizations. Several variations of a forecast model under this framework were analyzed including ILI models for nonlinear functions of the SIR compartmental model and the ASG function, modeling the discrepancy of ILI data, modeling the hospitalizations according to a linear or quadratic model with ILI as a predictive covariate, and modeling hospitalization error using different distribution families. A simulation study revealed the usefulness of modeling discrepancy during certain weeks of the flu season where there is a systematic peak not captured by the SIR or ASG functions. It also suggested that the ASG function may outperform the SIR more often than the SIR outperforms the ASG when used for forecasting in the US. The forecast modeling was applied to the 2023-24 FluSight data, and it produced reasonable forecasts, though in the real data analysis the SIR models appeared to outperform the ASG models.

In chapter 3 we introduced a statistical model for performing QM, or predictive distribution estimation of the distribution from which sample quantiles were estimated. The model introduced, QGP, is based on a central limit theorem for sample quantiles. QM is not a new idea, but many commonly used methods are either unable to accurately quantify the uncertainty of sample quantiles or they are limited to fitting certain distributions. The QGP depends on

selecting either a QF or a CDF, and in the case where the distribution family is unknown, we opted to use a normal mixture distribution for approximating a distribution via QM because of the flexibility of normal mixture distributions. A series of simulation studies showed the QGP's ability to accurately perform parameter inference and provide reasonable uncertainty quantification for sample quantiles. They also showed the QGP's ability to accurately approximate with uncertainty a true distribution via QM. The QGP was applied to all available forecasts of FluSight for the 2023-24 season.

In chapter 4 we introduced a new probabilistic forecast combination method, the SGP, for optimizing an ensemble forecast. The SGP is made to optimally select weights of a linear pooled forecast according to a proper scoring rule. Because the SGP is a Gibbs posterior, beyond selecting optimal combination weights, it also provides uncertainty quantification for model weights of the linear pool and allows for prior distribution information to be used in the optimization. Two simulation studies and an analysis on the 2023-24 FluSight hospitalization data and forecasts show the SGP's ability to create optimal forecasts and outperform model averaging and EQW methods for weighting selecting weights for forecasts.

The work herein adds to the tools that may be used in collaborative forecast hubs. All methods were motivated by the FluSight project, though the methods, especially those of chapters 3 and 4 may be applied more generally for non-disease outbreak problems. Going forward, the possibilities of research in the space of probabilistic forecasting in disease outbreak hubs are massive. Potential topics for future study in disease outbreaks include quantifying the forecastability of disease outbreak dynamics and assessing the information contained in various probabilistic forecast representations.

I had several ideas that I did not fully pursue, either due to time constraints or changes in direction. Though these ideas did not come to full fruition, I may pursue them in future research. While working on quantile matching in chapter 3, I considered treating quantiles or functions of quantiles as probability bins rather than as sample quantiles. I believe that by treating the quantiles as bins, the properties of the quantile function would allow for modeling the bins using a

censored likelihood which could also be used for distribution matching of bin distributions. An extension to chapter 3 that I have only briefly considered is performing quantile matching for correlated quantile forecasts. For multiple quantile forecasts that are ordered in time or space, perhaps modeling the distribution parameters as time-dynamic or as spatially correlated could lead to improved quantile matching in some applications.

While working on the Gibbs posterior in chapter 4, I considered forecast modeling situations where rather than simply being give the forecast distributions, forecast models would need to be fit, perhaps in a time-dynamic framework. This would make fitting the stacked Gibbs posterior much more difficult, and one of the challenges would be estimating the learning rate turning parameter. While considering this, I came across the idea of Pareto smoothed importance sampling used to approximate the leave-one-out cross validation error for Bayesian posterior distributions (Vehtari et al., 2024). This was used in Yao et al. (2018) when performing model stacking, but I think it could be applied to the stacked Gibbs posterior to make for efficient estimation of the learning rate and allow for using the Gibbs posterior for stacking in complicated scenarios. I am also excited about the possibility of using the Gibbs posterior for nonlinear model combination methods or under different scoring rules.

5.1 References

Vehtari, A., Simpson, D., Gelman, A., Yao, Y., and Gabry, J. (2024). Pareto smoothed importance sampling. *Journal of Machine Learning Research*, 25:1–57.

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.