Pancasting: forecasting epidemics from provisional data

Logan Brooks

CMU-CS-20-100

February 2020

School of Computer Science
Carnegie Mellon University
Pittsburgh, PA 15213

Thesis Committee:
Roni Rosenfeld, Chair
Ryan Tibshirani
Zico Kolter
Jeffrey Shaman, Columbia University

Submitted in partial fulfillment of the requirements for the degree of Doctor of Philosophy.

This research was sponsored by the National Institute of Health under grant number U54GM088491, the National Science Foundation under grant number DGE1252522, the Defense Threat Reduction Agency under grant number HDTRA118C0008, the Centers for Disease Control and Prevention under grant number U01IP001121, and Uptake Technologies. The views and conclusions contained in this document are those of the author and should not be interpreted as representing the official policies, either expressed or implied, of any sponsoring institution, the U.S. government or any other entity.

Keywords: epidemiological forecasting, data revisions, kernel conditional density estimation, quantile regression, influenza

Abstract

Infectious diseases remain among the top contributors to human illness and death worldwide [Murray et al., 2012, World Health Organization]. While some infectious disease activity appears in consistent, regular patterns within a population, many diseases produce less predictable epidemic waves of illness. Uncertainty and surprises in the timing, intensity, and other characteristics of these epidemics stymies planning and response of public health officials, health care providers, and the general public. Accurate forecasts of this information with well-calibrated descriptions of their uncertainty can assist stakeholders in tailoring countermeasures, such as vaccination campaigns, staff scheduling, and resource allocation, to the situation at hand, which in turn could translate to reductions in the impact of a disease.

Domain-driven epidemiological models of disease prevalence can be difficult to fit to observed data while incorporating enough details and flexibility to explain the data well. Meanwhile, more general statistical approaches can also be applied, but traditional modeling frameworks seem ill-suited for irregular bursts of disease activity, and focus on producing accurate single-number estimates of future observations rather than well-calibrated measures of uncertainty on more complicated functions of the data. The first part of this work develops variants of simple statistical approaches to address these issues, and a way to incorporate features from certain domain-driven models.

Epidemiological surveillance systems commonly incorporate a data revision process, whereby each measurement may be updated multiple times to improve accuracy as additional reports and test results are received and data is cleaned. The second part of this work discusses how this process impacts proper forecast evaluation and visualization. Additionally, it extends the models above to "backcast" how existing measurements will be revised, which in turn can be used to improve forecast accuracy. These models are then expanded further to include auxiliary data from other surveillance systems.

The preceding sections describe several prediction algorithms, and many more are available in existing literature and deployed in operational systems. The final part of this work demonstrates one method to combine output from multiple such prediction systems with consideration of the domain, which on average tends to match or outperform its best individual component.

Acknowledgments

First, I would like to thank my advisor, Roni Rosenfeld, for his kind and patient guidance throughout the program. His vision and drive for operational, impactful research benefiting the social good in consultation with stakeholders has been an inspiration academically, as his altruism, amiability, and leadership have been personally. Without his eye for detail and ability to connect people, none of this work would have been possible.

During the PhD program, I have also been fortunate to receive the mentorship of Ryan Tibshirani. Roni and Ryan formed the Delphi Research Group at Carnegie Mellon University (CMU) in 2013, fostering a welcoming, congenial environment and discussion of ideas.

I am grateful to have been able to work and exchange ideas with many other members of the Delphi Research Group and for their cordiality. Their contributions are too many to enumerate in full, but a few are particularly relevant in the context of this document. David Farrow constructed data scraping, querying, and visualization tools, along with other utilities, that have been essential in implementing many of these models. David, as well as Sangwon Hyun and Aaron Rumack, have contributed to the design, implementation, analysis, deployment, and monitoring of many or all of these systems. Many members, including Shannon Gallagher, Daren Wu, Zirui Wang, and Nuoyu Li have provided ideas, in-depth study, and paths to improvement of particular models herein, while others have provided essential comparisons to other approaches, collection and investigation of additional data, and improvement of the ILI-Nearby nowcasting system and Delphi's server infrastructure upon which these downstream models rely.

This project has been enabled and shaped by efforts of U.S. government agencies to advance the science and practice of epidemic forecasting, particularly in the context of influenza, beginning with the "Predict the Influenza Season Challenge". I am grateful for the Centers for Disease Control and Prevention personnel who hosted

this challenge, those who paved the way for this idea, and those who carried it forward into later years. In particular, without the efforts of Matthew Biggerstaff and Michael Johansson and others involved in the Epidemic Prediction Initiative for establishing and maintaining quantitative forecast comparisons and dialogue between external forecasters and stakeholders. Thanks also go to those at the Council of State and Territorial Epidemiologists who have enabled discussions with additional stakeholders and paths forward to more detailed and widely useful forecasts. Many other organizations and individuals, within and beyond the U.S. government, have also contributed to these types of efforts via funding research, as well as preparing, and providing, and sharing useful data resources.

CMU has a remarkable number of dedicated staff members and initiatives that work to eliminate troubles beyond the latest overfitting estimator. In particular, I am thankful for the lightning-fast response time, enduring patience, and kind outreach of Christy Melucci, Deborah Cavlovich, and Catherine Copetas.

I would like to thank the committee for contributing their time and expertise, and for their flexibility in arranging presentations and tolerance reviewing some rather coarse rough drafts and preliminary work. Their feedback has called attention to surprising trends in the analysis, and identified promising alternative approaches and opportunities for future work.

Finally, I would like to extend my gratitude to past mentors, teachers, family, and friends, who have guided me toward enriching and rewarding graduate studies and research, and have encouraged and supported me throughout this endeavor.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Much of this chapter is based on material from Brooks et al. [2018].

1.0.1 Infectious diseases and the motivation for forecasting

Despite modern medical advances, infectious diseases remain among the top causes of human illness and death worldwide, and pose major threats even in high-income countries [Lozano et al., 2012, Murray et al., 2012, World Health Organization]. Within the scope of infectious diseases, leading contributors include lower respiratory infections (e.g., with pneumonia or influenza) and diarrheal diseases (e.g., from foodborne bacteria and viruses) [El Bcheraoui et al., 2018, Lozano et al., 2012, Murray et al., 2012, World Health Organization]. Some infectious disease activity occurs in consistent, regular "endemic" patterns within a population, but many diseases produce less predictable "epidemic" waves of illness. Uncertainty and surprises in the timing, intensity, and other characteristics of these epidemics stymies planning and response of public health officials, health care providers, and the general public, and contributes to a high health and economic burden.

For instance, in the United States and other temperate regions, lower respiratory infection activity various classes of respiratory and circulatory disease, such as lower respiratory infections, present fairly uniform "baseline" patterns repeating each year, punctuated by sharp spikes in prevalence often associated with influenza epidemics [Kyeyagalire et al., 2014, Serfling, 1963, Thompson et al., 2010, 2003, Zhou et al., 2012]. Influenza epidemics typically occur once a year during the "influenza

season" (roughly from October to May in the Northern Hemisphere), but vary in timing, intensity, and other traits; these "seasonal" epidemics are associated with an estimated 250 000 to 500 000 annual deaths worldwide [World Health Organization, 2016], with a range of 3000 to 56 000 deaths in the US alone [Centers for Disease Control and Prevention, National Center for Immunization and Respiratory Diseases (NCIRD), 2016a, Rolfes et al., 2016, Thompson et al., 2010]. Additionally, influenza "pandemics", which are rare global outbreaks of especially novel influenza viruses [Centers for Disease Control and Prevention, National Center for Immunization and Respiratory Diseases (NCIRD), 2016b, 2017b], can cause deaths on even greater scales [Johnson and Mueller, 2002, Viboud et al., 2010]. Potential countermeasures include [Niska and Shimizu, 2011] adjusting scheduling and providing on-site child care for health workers to better handle increased patient loads; canceling or rebooking less urgent medical appointments and procedures, admitting emergency department patients to inpatient hallway beds and reconfigured or alternative spaces; transferring patients to other facilities to avoid or reduce overcrowding; producing and tuning composition of vaccines; manufacturing, allocating, and redistributing antiviral medication, respirators, and other resources; inducing insurance companies to pay for more expensive brands of the same medications when availability of cheaper alternatives is limited; and launching or modifying campaigns to promote vaccination, effective hand-washing practices, wearing face masks [Aiello et al., 2008, Cowling et al., 2009, Rabie and Curtis, 2006, Simmerman et al., 2011, Suess et al., 2012, Talaat et al., 2011], and other beneficial behaviors, targeted to sick individuals, their close contacts, or health workers, in order to curtail the spread and consequences of infections. The design and effectiveness of these efforts depends on the range of expectations for and ultimate reality of an epidemic's size, timing, and other characteristics.

Accurate and reliable forecasts of this information could provide early warning, bolster situational awareness, and assist in designing countermeasures, which in turn may reduce the overall impact of infectious disease. While the idea of epidemic modeling and forecasting is not new, recent years have seen growing interest driving government initiatives that standardize datasets, tasks, and metrics to improve forecast usability, address decision-maker needs, attract and assist external modelers, and allow for rigorous evaluation and comparison. These efforts include the U.S. government's Dengue Forecasting project, CHIKV (Chikungunya virus) Challenge, and a series of influenza forecast comparisons. This document will focus on these influenza forecasting testbeds and corresponding surveillance systems in the US.

The Centers for Disease Control and Prevention (CDC) monitors influenza preva-

lence with several well-established surveillance systems [Centers for Disease Control and Prevention, 2013]; the recurring nature of seasonal epidemics and availability of historical data provide promising opportunities for the formation, evaluation, and application of statistical models. Starting with the 2013/2014 "Predict the Influenza Season Challenge" [Biggerstaff et al., 2016] and continuing each season thereafter as the Epidemic Prediction Initiative's FluSight project [Biggerstaff et al., 2018], CDC has solicited and compiled forecasts of influenza-like illness (ILI) prevalence from external research groups and worked with them to develop standardized forecast formats and quantitative evaluation metrics. Targets of interest include disease prevalence in the near future, as well as features describing the timing and overall intensity of the disease activity in the season currently underway. Policymakers desire not only in point predictions of these quantities, but full distributional forecasts; recent initiatives solicit both types of estimates, but base evaluation on customized log scores of distributional forecasts.

1.1 Models of disease dynamics

Various approaches to influenza epidemic forecasting are summarized in literature reviews [Chretien et al., 2014, Nsoesie et al., 2014, Unkel et al., 2012] and descriptions of the CDC comparisons [Biggerstaff et al., 2016, 2018]. Some common approaches are described below, with references to work applicable to the current FluSight project and related seasonal dengue forecasting tasks, emphasizing more recent work that may not be listed in the above three literature reviews:

- **Mechanistic models:** describe the disease state and interaction between individuals with causal models, as well as the surveillance data generation process.
    - **Compartmental models** (e.g., [Hickmann et al., 2015, Kandula et al., 2017, Shaman and Karspeck, 2012a, Shaman et al., 2013a, Zhang et al., 2017]) break down the population into a number of discrete "compartments" describing their characteristics (e.g., age, location) and state (e.g., susceptible to, infectious with, or recovered from a particular disease), and describe how the occupancy of these compartments changes over time, either deterministically or probabilistically. In many of these models, this division describes solely the state with respect to a single disease, ignoring details regarding age, spatial dynamics, and mixtures of ILI diseases, but keeping the number of parameters to infer low. Methods to fit

these models to data include variants of particle and ensemble Kalman filters [Yang et al., 2014], naïve importance sampling [Brooks et al., 2015a], iterative augmented-state filtering [Ionides et al., 2006, 2015, Lindström et al., 2012], general Bayesian frameworks [Osthus et al., 2019] (using JAGS [Plummer, 2003], Stan [Carpenter et al., 2017], etc.), filtering using linear noise approximation [Zimmer et al., 2017, 2018], and Gaussian process approximations [Buckingham-Jeffery et al., 2018].

- **Agent-based models** (e.g., [Deodhar et al., 2015, Nsoesie et al., 2014]), also known as individual-based models, these approaches use more detailed descriptions of disease state and/or individual characteristics and behavior, which are not easily simplified into a compartmental form, typically studied using computation-heavy simulations. These approaches usually include many more parameters than compartmental models, which may be set based on heuristics or additional data sources and studies, or, alternatively, inferred based on the surveillance data, often by using Markov chain Monte Carlo (MCMC) procedures. Developing effective inference and prediction techniques is an active area, with scalability to large populations still on the frontier of research, requiring special techniques and/or likelihood approximations [O'Neill, 2010].

- **Phenomenological models:** also referred to as statistical models, these approaches describe the surveillance data without directly incorporating the epidemiological underpinnings.

  - **Direct regression models** (e.g., [Brooks et al., 2015a, Chakraborty et al., 2014, Ray et al., 2017, Viboud et al., 2003]) attempt to estimate future prevalence or targets of interest using various types of regression, including nonparametric statistical approaches and alternatives from machine learning literature.

  - **Time series models** (e.g., [Generous et al., 2014, Höhle et al., 2017, Johansson et al., 2016, Lampos et al., 2015, Lowe et al., 2013, Martinez et al., 2011, Paul et al., 2014a, Yang et al., 2015, 2017]) represent the expected value of (transformations of) observations and/or underlying latent state at a particular time as (typically linear) functions of these quantities at previous times and additional covariates, paired with Gaussian, Poisson, negative binomial, or other noise distributions. This category includes linear dynamical systems and frameworks such as SARIMAX.

Complicated mechanistic approaches such as agent-based models are often too

complex to efficiently fit to surveillance data and are instead less strictly "calibrated" based on summary measures, which may not produce a close match to the surveillance observations. Instead, mechanistic forecasting approaches have focused on simpler compartmental models and frameworks for fitting them to surveillance data. However, oversimplified compartmental models often cannot tightly match the surveillance data for an entire season simultaneously [Brooks et al., 2015a], which can degrade prediction quality. Some degree of mismatch can be attributed to observation models that do not reflect important details of the surveillance system, which are discussed in the next section. Another contributor is the rigidity of compartmental models with deterministic state transitions, the full mixing assumption, and shallow-tailed observational noise, leading to overconfident forecasts; some paths forward are to incorporate variance inflation factors [Shaman and Karspeck, 2012a], overdispersed observational noise [Lowe et al., 2013], stochastic variants or process noise [King et al., 2015], random walk discrepancy terms [Osthus et al., 2019], error breeding procedures [Pei and Shaman, 2017], more complex models with appropriate filtering algorithms [Pei et al., 2018], or use of improper conditioning procedures to combat overconfidence.

Phenomenological models, on the other hand, offer a wide range of general-purpose methods designed around efficient, straightforward predictions. Univariate response models are extremely flexible, but seem inappropriate when the target of interest is a function of a surveillance time series. The most popular statistical time series methods, falling within "alphabet soup" frameworks such as SARIMAX and GARCH, directly model the time series, but sacrifice some flexibility by focusing on linear dynamics and Gaussian noise.

Chapter 2 expands the phenomenological front and moves toward the mechanistic one, presenting methods that incorporate the flexibility of univariate response models into ARI-type time series models, and ways to tailor these models to epidemiological settings to resemble a compartmental model. Concurrent work similar expands flexibility of time series models using an alternative copula approach [Ray et al., 2017], and other work incorporates additional aspects of epidemiological data within a Bayesian optimization framework [Osthus and Moran, 2019]; many additional alternatives are listed in [Biggerstaff et al., 2018] and [Biggerstaff et al., 2016].

1.2 Models of observations

Most epidemic modeling work, including much of the epidemic forecasting literature, tends to focus on the disease transmission dynamics, with the nature of surveillance system modeled with a very simple observational noise term. However, recognizing some details of surveillance systems is essential when performing retrospective forecast preparation and evaluation, and using these details to inform models can improve forecast accuracy. For example, surveillance data may contain spikes around holidays, which may be explained by differences in health care seeking behavior producing artifacts in the surveillance system, and/or due to changes in disease transmission behavior. Chapter 2 touches on some model settings that can be used to acknowledge holiday effects.

A more fundamental issue is that the ground truth from traditional surveillance systems used for evaluation is not available in real-time for use in forecasts. It takes time for symptoms to be recorded, diagnoses to be made, lab tests to complete; for health care workers to prepare and submit reports; and for public health officials to compile, clean, summarize, and publish the data. Furthermore, a case might only be reported after recovery or death, but recorded with a time closer to the onset of symptoms. In short, there is a trade-off between the accuracy of an observation and its timeliness. Traditional surveillance systems often address this problem by publishing an initial observation for a given time once the level of reliability is deemed acceptable, then later reporting a revised value or sequence of revised values that improve the expected accuracy. After some time, the observation may be finalized in the surveillance system or considered stabilized enough to be interchangeable with the finalized value and used as ground truth for forecast evaluation. Within this document, the distinction between these two cases will be ignored, and "finalized" will be used to refer to the data used for forecast evaluation. Chapter 3 discusses how the revision process impacts proper retrospective forecast evaluation, and how forecast accuracy can be improved by modeling the revision process.

In recent years, a number of novel digital surveillance sources and derived estimators have been prepared using internet search query data [Ginsberg et al., 2009], social media activity [Dredze et al., 2014], web page hits [Farrow, 2016, Hickmann et al., 2015], self-reported illness [Smolinski et al., 2015], internet-integrated monitoring and testing devices [Farrow, Accessed 2017-04-26, Miller et al., 2018], electronic health records and derived statistics [Santillana et al., 2015], insurance claims [Jahja et al., 2018, Viboud et al., 2014], or some combination along with traditional surveillance data (discussed further in Section 4.1). These estimates are not used as ground

truth for evaluation, but may have better timeliness and temporal, geographic, or demographic resolution, and offer opportunities for improved forecasting. Some of these sources undergo a similar revision process as more traditional surveillance data. For truly real-time systems, initial estimates for a given time interval may be available before the end of that time interval (e.g., an initial estimate for cases for an entire Sunday-to-Saturday week may be available based on data from Sunday through Tuesday). Chapter 4 discusses how to incorporate these additional data sources within the modeling framework presented in the previous chapters.

1.3 Epidemiological surveillance data

Epidemiological surveillance data exhibit a number of behaviors which are problematic for traditional time series methods:

- **Rare or one-time events** cause major shifts in reported disease prevalence, including
  - **invasions:** introduction of a disease into an area that has not encountered it before;
  - **novel strain pandemics:** epidemics with wider geographical spread or high incidence, often occurring at unseasonal times of year, caused by mutations in a strain of a disease that result in more effective transmission;
  - **mass vaccination and eradication campaigns:** coordinated efforts by public health officials to drastically increase the proportion of the population that is vaccinated against a disease; and
  - **sudden shifts in reporting practices or suitability:** changes in reporting requirements; the type or number of reporting health care providers (in a passive surveillance system (define)); disease definitions, testing procedures, testing equipment, or testing sensitivity to prevalent disease strains; reporting frequency, geographical and temporal scope and resolution, disease specificity; among other changes;
- **Seasonality in transmissibility** which results in irregular seasonal behavior in case counts: epidemic waves of varying heights and times that usually occur with some wide "on-season" time window (in addition to more predictable background seasonality for which sinusoidal or seasonal autoregressive terms and Gaussian-like noise seem more appropriate)

- **Nonadditive holiday effects** on health care seeking, reporting, or disease transmission rates;
- **Data revisions** to past surveillance data are common, as the reporting delay for cases may vary based on the attending health care provider and duration of illness, suspected cases of a disease may be included in early estimates but ruled out later, and, for rapidly available datasets, the time window for data aggregation may include times in the future (e.g., later days of the current week) which are necessarily not observed yet; and
- **Ragged data availability** , used here to refer to differences among surveillance signals in geotemporal and demographic resolution, availability, and reliability patterns; timeliness of release; and underlying stimuli, complicate the creation and use of models incorporating multiple signals simultaneously.

1.3.1 The ILINet surveillance system

This subsection reproduces or incorporates content from Brooks et al. [2018].

One example of a traditional surveillance system designed to monitor influenza activity is the U.S. Outpatient Influenza-like Illness Surveillance Network (ILINet). Recording every case of influenza is not practicable; infections are often asymptomatic [Leung et al., 2015] or symptomatic but not clinically attended [Hayward et al., 2014], laboratory testing may not be performed for clinically attended cases or give false negative results, and reporting of lab-confirmed cases is not mandatory in most instances. Instead, estimates of true influenza activity are often based on syndromic clinical surveillance data from ILINet [Brammer et al., 2013, Centers for Disease Control and Prevention, 2013], a group of health care providers that voluntarily report statistics regarding ILI, where ILI is defined as a 100 °F (37.8 °C) fever with a cough and/or sore throat without a known cause other than influenza. CDC aggregates these reports and estimates the weekly percentage of patients seen that have ILI, %ILI, across all health care providers using a measure called weighted %ILI (wILI).

- **Geographical resolution:** CDC reports wILI for each of the 10 U.S. Department of Health & Human Services (HHS) Regions, as well as for the nation as a whole; the wILI for each of these locations is a weighted average of the ILINet %ILI for state-level units based on population.

- **Temporal resolution:** wILI is available on a weekly basis; weeks begin on Sunday, end on Saturday, and are numbered according to the epidemiological week (epi week) convention in the United States.

- **Timeliness:** Initial wILI estimates for a given week are typically released on Friday of the following week; additional reports and revisions from participating health care providers are incorporated in updates throughout the season and after it ends.

- **Specificity:** Influenza is just one of many potential causes of ILI. Laboratory testing data [Centers for Disease Control and Prevention, 2013] suggest that influenza is responsible for a significant portion of ILI cases during the flu season, especially for weeks when wILI is high, but only for a very small fraction of cases in the typical flu off-season. Much of the variance and "peakiness" in wILI can be associated with influenza epidemics, but wILI trajectories do not taper off to near-zero values as one might expect in a direct measurement of influenza prevalence.

- **Influence of non-ILI cases:** Since wILI depends on records of both ILI cases and total cases, patterns in non-ILI cases can impact wILI trajectories. We discuss one such pattern in Section 2.5.

CDC hosts the latest ILINet report and other types of surveillance data through FluView Interactive, a collection of web modules [Centers for Disease Control and Prevention, National Center for Immunization and Respiratory Diseases (NCIRD), 2017a]; the Delphi Group at Carnegie Mellon University provides current and historical ILINet reports and some other data sources through our `delphi-epidata` API [The Delphi Group at Carnegie Mellon University, Accessed 2017-04-26] and `epivis` visualizer [Farrow, Accessed 2017-04-26]. Figure 1.1 and Figure 1.2 show wILI data at the national level from one such report.

Forecasting targets

Starting with the 2013/2014 "Predict the Influenza Season Challenge" [Biggerstaff et al., 2016] and continuing each season thereafter as the Epidemic Prediction Initiative's FluSight project [Biggerstaff et al., 2018], CDC has solicited and compiled forecasts of ILI prevalence from external research groups and worked with them to develop standardized forecast formats and quantitative evaluation metrics. The

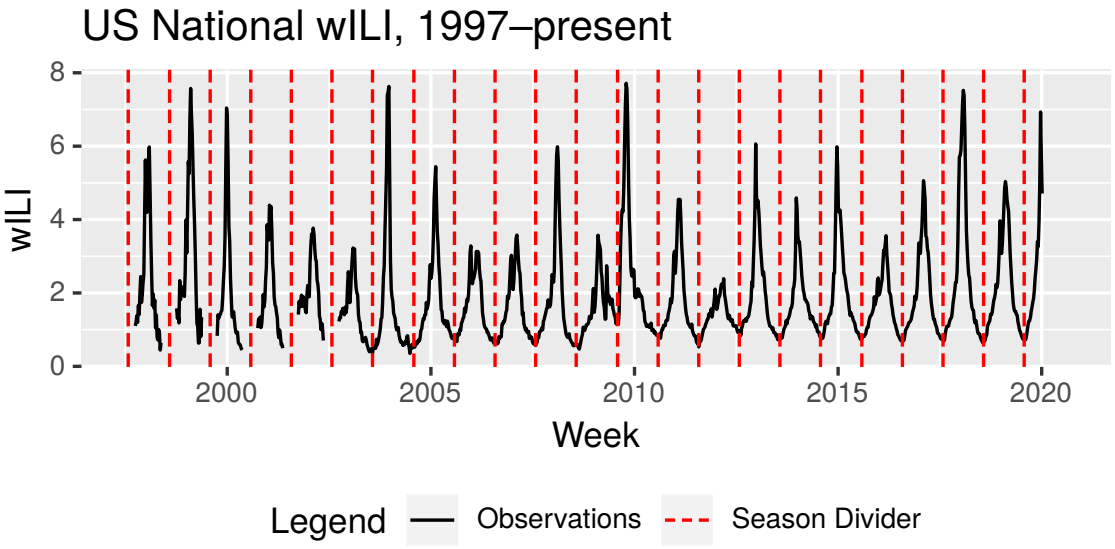

Figure 1.1: **Snapshot of national-level wILI data from a FluView report.** Red dashed lines mark calendar week 31, roughly the beginning of August, emphasizing the annual seasonality of ILI in the US. The maximum wILI value obtained within each season varies widely.

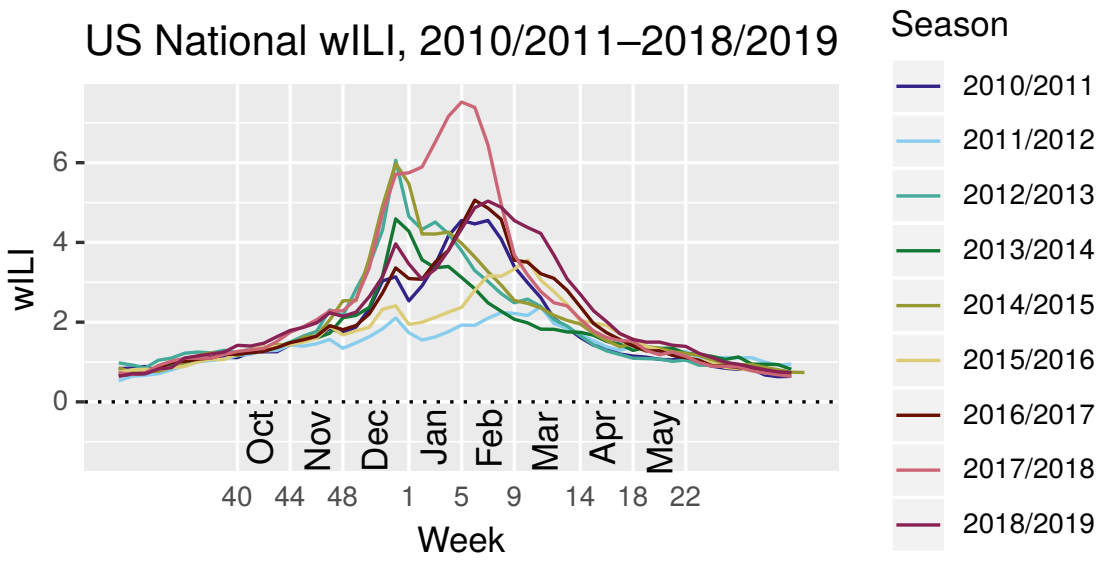

Figure 1.2: **Snapshot of recent national-level wILI data from a FluView report, cut into year-long seasons and superimposed.** Not only does the maximum wILI value obtained within each season vary widely, but the timing of this peak varies as well. Holiday effects are evident around winter holidays and Thanksgiving.

FluSight project focuses on in-season distributional forecasts and point predictions of key targets of interest to public health officials:

- **Short-term wILI:** the four wILI values following the last available observation (incorporating all data revisions through some week well after the season's end)
- **Season onset:** the first week in the first run of at least three consecutive weeks with wILI values above a location- and season- specific baseline wILI level set by CDC [Centers for Disease Control and Prevention, 2013], or "none" if no such runs exist; describes whether and when an influenza epidemic started in a given season
- **Season peak percentage:** the maximum of all wILI values for a given season
- **Season peak week:** the week or weeks in which wILI takes on its maximum value, or "none" if there was no onset in the 2015/2016 comparison

When making distributional forecasts, wILI values are discretized into CDC-specified bins and a probability assigned to each bin, forming a histogram over possible observations. The width of the bins was set at 0.5 %wILI for the 2015/2016 comparison and 0.1 %wILI for the 2016/2017, 2017/2018, and 2019/2020 comparisons; we use a width of 0.1 %wILI for cross-validation analysis. CDC typically presents wILI values rounded to a resolution of 0.1 %wILI; some targets and evaluations are based on these rounded values.

1.4 Evaluation metrics

This section reproduces or incorporates content from Brooks et al. [2018].

We focus on three metrics for evaluating performance of a forecast for a given target:

- **Unibin log score:** $\log \hat{p}_i$, where $\hat{p}_i$ is the probability assigned to i, the bin containing the observed value. This scoring rule is illustrated in Figure 1.3. We use this score for ensemble weight selection and most internal evaluation as it has ties to maximum likelihood estimation, and is "proper score" [Hendrickson and Buehler, 1971]. A score for a (reported) distributional prediction $\hat{p}$ is

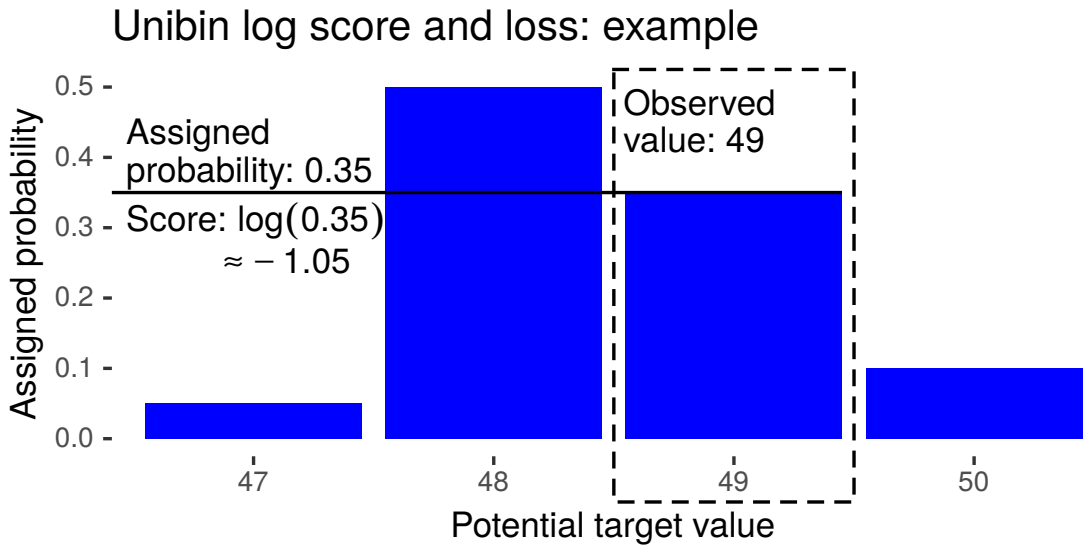

Figure 1.3: **Illustration of a distributional forecast and unibin log score calculation.** The potential target values displayed would often reasonable selections for the "Season onset" target in the FluSight forecasting context, although actual forecasts for this target will assign probabilities to 33 or 34 possibilities rather than just 4.

called "proper" if its expected value according to any (internal) distributional prediction $\hat{q}$ is maximized when reporting $\hat{p} = \hat{q}$, i.e., forecasters can maximize their expected scores by reporting their true beliefs. We refer to the "unibin log score" simply as the "log score" except for when comparing it with the multibin log score, which is defined next. The exponentiated mean unibin log scoreis the (geometric) average probability assigned to events that were actually observed. The exponentiated difference in the mean log scores of method A and method B is an estimate of the (geometric) expected winnings of unit-sized bets of the form "this bin will hold the true value" when bets are placed optimally according to the forecasts of A, and (relative) prices are set optimally according to the forecasts of B. A distributional prediction assigning a probability of zero to an event that actually occurred will be assigned a "raw" unibin log score of $-\infty$, and cause the corresponding forecaster to receive an overall raw log score of $-\infty$; in some analyses, we will threshold individual unibin log scores at $-10 \approx \log 0.0000454$ to enable more meaningful comparisons; in others, we will use raw log score, but only compare variants of methods that avoid infinite and extremely low individual scores.

- **Multibin log score:** $\log \sum_{i \text{ near observed value}} \hat{p}_i$, where the i's considered are typically bins within 0.5 %wILI of observed values for a wILI target, or within 1 week for a timing target. Similarly to the unibin log score, the multibin log score may be thresholded, e.g., at a minimum of $-10 \approx \log 0.0000454$, to limit the sensitivity of the mean score on any individual score. A thresholded multibin log score was designed by FluSight hosts in consultation with participants, and the judgment "near observed value" was selected as a level of error that would not significantly impact policymakers' decisions. The exponentiated mean multibin log score, or "skill score", is the (geometric) average amount of mass a forecaster placed within this margin for error of observed target values.

- **Absolute error:** $|\hat{y} - y|$, where $\hat{y}$ is the point prediction and y is an observed value. (In the case of onset, we consider point predictions for the value of onset conditioned on the fact that an onset actually occurs. We do not consider absolute error for onset in instances where no onset occurred. Some methods considered would sometimes fail to produce such conditional onset point predictions when they were confident that there was no onset, but these methods are not included in any of the figures containing absolute errors.)

The FluSight 2015/2016, 2016/2017, and 2018/2019 forecast comparison overall eval-

uations were based solely on the thresholded multibin log score [Epidemic Prediction Initiative, 2016], while the 2019/2020 comparison is set to use the unibin log score.

1.5 Overview

This work focuses on building models for disease data that accommodate seasonality, holiday effects, data revisions, and ragged data availability. Chapter 2 focuses on building nonparametric univariate time series models that factor in holiday effects and seasonality in transmissibility, ignoring the fact that data revisions occur and additional surveillance signals may be available. Chapter 3 deals with the modeling of data revisions. Chapter 4 discusses incorporation of additional data sources with differing availability patterns. Chapter 5 describes a method to combine pancasts from multiple methodologies leveraging information about their behavior from retrospective forecasts.

Chapter 2

Probabilistic forecasting of the spread of epidemics

Stakeholders desire accurate and reliable forecasts of disease prevalence in the next few weeks, and of summary statistics about the timing and intensity of epidemics. The goal is to improve situational awareness and decision-making regarding, for example, hospital staffing and scheduling impacting readiness for surges in the number of inpatients, or the timing of a vaccination campaign. Each of the prediction targets could be handled separately: one model could be built to forecast disease prevalence next week, another to forecast the week when prevalence is highest, and so on. However, we focus on a more unified approach: first, forecasting the distribution of the disease prevalence trajectory for the entire season, then extracting the corresponding distributions for the targets of interest. This chapter discusses methods of forecasting the future of a trajectory given observed values of this trajectory in the past. These methods fall into two categories: "entire-trajectory models", which treat $Y_{1..T}$ as a vector, and "chained one-ahead models", which break it into scalars using Markov-like assumptions. Section 2.2, Section 2.3, and Section 2.4 lay out specific modeling frameworks falling within one of these two categories. Subsection 2.4.2 discusses how to incorporate mechanistic model-inspired covariates into one of these frameworks. Finally, Section 2.5 discusses approaches to incorporating holiday effects in these frameworks. All of the methods described in this chapter are "revision-ignorant", assuming that these past values do not undergo a revision process; this aspect of surveillance data is addressed in Chapter 3 and Chapter 4.

2.1 Revision-ignorant forecasting task

Given past observations $Y_{1..t}$ of a univariate surveillance time series $Y_{1..T}$ for a semi-regular seasonal epidemic, we want to estimate the distribution of future trajectories, $Y_{t+1..T}$. The distributional aspect of the forecast is important: many time series methods focus on conditional mean estimation, and incorporate Gaussian observational and/or process noise as a matter of convenience to obtain a generative model; we seek a more flexible noise model able of capturing heavy tails and multi-modality. Furthermore, we prefer the conditional mean estimates that are produced to have a flexible, nonparametric flavor. Being able to produce a sample from the distribution for $Y_{t+1..T}$ is sufficient; we do not need an explicit representation of the model.

The forecasting methodologies described below are based on two general approaches satisfying the above criteria:

- **Entire-trajectory models:** characterize the set or distribution of possible trajectories $Y_{1..T}$, and use conditioning or inference techniques to obtain a related conditional distribution $Y_{t+1..T} \mid Y_{1..t}$.
- **Chained one-ahead models:** construct many "one-ahead" conditional density models $Y_u \mid Y_{1..u-1}$ and piece them together to form (a sample from) an estimate of the conditional distribution $Y_{t+1..T} \mid Y_{1..t}$.

Hybrid approaches are also possible, combing conditional future mean curves $\mathbb{E}[Y_{t+1..T} \mid Y_{1..t}]$ from some entire-trajectory model with a chained one-ahead model for the residuals $(Y_{t+1..T} - \mathbb{E}[Y_{t+1..T} \mid Y_{1..t}]) \mid Y_{1..t}$.

2.1.1 Entire-trajectory models:

Entire-trajectory models directly characterize the full set or distribution of trajectories $Y_{1..T}$, often as some latent mean curve $Z_{1..T}$ plus noise $Y_{1..T} - Z_{1..T}$. This entire-trajectory characterization is transformed into (a sample from) a related estimate of $Y_{t+1..T} \mid Y_{1..t}$ either via techniques to condition on $Y_{1..t}$ (such as importance sampling) or via maximum likelihood or maximum a posteriori estimation of $Z_{1..t}$ or $Z_{1..T}$ (e.g., involving regression to fit means of $Y_{1..t}$). This document will describe one such method, an empirical Bayes style approach that fits a library of mean curves and noise models and performs importance sampling to extract a posterior over related parameters and a sample from a predictive distribution for $Y_{t+1..T} \mid Y_{1..t}$.

2.1.2 Chained one-ahead models:

Chained one-ahead models can borrow from well-known, flexible methods for univariate regression and density estimation and repurpose them for time series estimation. A simple procedure allows us to sample from an estimate of $Y_{t+1..T} \mid Y_{1..t}$ based on samplers for estimates of one-step-ahead conditional distributions $Y_{t+1} \mid Y_{1..t}$, $Y_{t+2} \mid Y_{1..t+1}$, $\ldots$, $Y_T \mid Y_{1..T-1}$:

- Draw $Y_{t+1}^{\text{sim}} \sim Y_{t+1} \mid Y_{1..t}$
- Draw $Y_{t+2}^{\text{sim}} \sim Y_{t+2} \mid Y_{1..t}, Y_{t+1} = Y_{t+1}^{\text{sim}}$ (using model for $Y_{t+2} \mid Y_{1..t+1}$)
- Draw $Y_{t+3}^{\text{sim}} \sim Y_{t+3} \mid Y_{1..t}, Y_{t+1,t+2} = Y_{t+1,t+2}^{\text{sim}}$ (using model for $Y_{t+3} \mid Y_{1..t+2}$)
- $\ldots$
- Draw $Y_T^{\text{sim}} \sim Y_T \mid Y_{1..t}, Y_{t+1..t} = Y_{t+1..T-1}^{\text{sim}}$ (using model for $Y_T \mid Y_{1..T-1}$)
- Record $Y_{t+1..T}^{\text{sim}}$ and repeat this process to obtain additional simulated futures.

There are essentially no restrictions on the models selected for $Y_u \mid Y_{1..u-1}$ for each u. It would be more faithful to condition, e.g., in the first step, on the random variable Y_{t+1}^{sim} rather than on the condition $Y_{t+1} = Y_{t+1}^{\text{sim}}$, but also more algorithmically and computationally challenging.

One natural approach to building the conditional distributions above is to first directly estimate the conditional distribution $\Psi^{[u]} \mid \mathbf{\Phi}^{[u]}$, where $\Psi^{[u]}$ is a (potentially u-specific) function of $Y_{1..u}$ from which Y_u can be recovered given $Y_{1..u-1}$ (e.g., $\Psi^{[u]} = \Delta Y_u = Y_u - Y_{u-1}$ or $\Psi^{[u]} = \log Y_u$), and $\mathbf{\Phi}^{[u]}$ is a (potentially u-specific) vector of features derived from $Y_{1..u-1}$.[1] During simulation, $Y_{1..u-1}^{\text{sim}}$ will be used to calculate corresponding simulated feature values $\mathbf{\Phi}^{[u],\text{sim}}$, which are used to draw a simulated transformed value $\Psi^{[u],\text{sim}}$, from which a corresponding simulated value Y_u^{sim} can be recovered. Nonparametric methods along these lines include:

- **Kernel delta density,** which draws ΔY_u^{sim} from an estimate of the conditional density for $\Delta Y_u \mid \mathbf{\Phi}^{[\text{KDD},u]}$ based on smoothing kernel methods with some heuristic modifications, where $\mathbf{\Phi}^{[\text{KDD},u]}$ is a vector of heuristically constructed and weighted features for time u derived from $Y_{1..u-1}$, and

[1] Note that Δ denotes a backward difference rather than a forward difference throughout this document.

- **Quantile autoregression** using locally linear quantile regression and optional post-processing noise, which selects $\Psi^{[u],\text{sim}}$ as the sum of a random estimated conditional quantile and (optionally) some smoothing noise, where the conditional quantile is estimated for $\Psi^{[u]} \mid \mathbf{\Phi}^{[\text{QARlinear},u]}, \mathbf{\Phi}^{[\text{QARkernel},u]}$ as a linear function of covariates $\mathbf{\Phi}^{[\text{QARlinear},u]}$, with training data weighted with a smoothing kernel on covariates $\mathbf{\Phi}^{[\text{QARkernel},u]}$.

The entire-trajectory empirical Bayes framework and chained one-ahead kernel delta density and quantile autoregression approaches are described and studied in more detail in their own sections, and the latter two are extended in subsequent chapters.

2.2 Empirical Bayes framework

This section reproduces or incorporates content from Brooks et al. [2015a].

The empirical Bayes framework was initially designed by Ryan Tibshirani and Roni Rosenfeld; other authors in the original paper and later collaborators contributed to the implementation, extension, application, and study of the method. Below, certain details of the model are selected with a certain disease and surveillance system in mind, but can (and have [van Panhuis et al., 2014]) been adjusted when applied in other contexts.

The forecasting framework is composed of five major procedures:

1. Model past seasons' epidemic curves as smoothed versions plus i.i.d. Gaussian noise.
2. Construct a prior for the current season's epidemic curve by considering sets of transformations of past seasons' curves.
3. Set a point estimate for the ground truth values for the recent past given provisional values and auxiliary data sources.
4. Repeatedly sample whole-season trajectories and assign them weights, such that the product of the sampling frequency and assigned weights is proportional to the posterior probability: the prior probability times the likelihood of the point estimates for ground truth in the recent past. (E.g., by sampling trajectories from the prior and assigning weights based on the likelihood.)

5. Use the sampled trajectories and their associated weights to calculate a posterior distribution over any desired forecasting target.

The first two steps only need to be executed once, at the beginning of the current season. As additional data becomes available throughout the season, we generate forecasts using steps 3–5.

We perform predictions for each geographical unit — the US as a whole or individual HHS Regions — separately. Historically, surveillance has focused on influenza activity between epidemiological weeks 40 and 20, inclusive. We define seasons as epidemic weeks 21 to 39, the "preseason", together with weeks 40 to 20. During the 2013/2014 competition, data was available for 15 historical seasonal influenza epidemics. We excluded the 2009/2010 season from the data since it included non-seasonal behavior from the 2009 pandemic in the preseason. Additionally, there was partial data available for the 2013/2014 season (updated weekly).

Data model

We view wILI trajectories for a geographical unit r as the sum of some underlying ILI curve plus noise:

$$y_i^{r,s} = f^{r,s}(i) + \epsilon_i^{r,s}, \qquad \epsilon_i^{r,s} \sim \mathcal{N}(0, \tau^{r,s}), \qquad \text{for each week } i, \tag{2.1}$$

where $y_i^{r,s}$ is the wILI observation for the ith week of season s, $f^{r,s}$ is the underlying curve, and ϵ_i^s is (independent) zero-mean normally distributed noise. We estimate the underlying ILI curve $\hat{f}^{r,s}$ from the wILI curve $y^{r,s}$ with quadratic trend filtering [Tibshirani, 2014] for each historical season s. This method smooths out fluctuations in the wILI data, producing a new set of points that lie on a piecewise quadratic curve. We use the `cv.trendfilter` [Arnold and Tibshirani, 2014] method to select an appropriate amount of smoothness for each curve, then estimate the corresponding noise level $\hat{\tau}^{r,s}$:

$$(\hat{\tau}^{r,s})^2 = \operatorname*{avg}_i [y_i^{r,s} - \hat{f}^{r,s}(i)]^2.$$

The quadratic trend filtering procedure produces one point for each available wILI observation, i.e., 33 or 34 for the first six seasons, where only data from the flu season is available, and 52 or 53 for the rest. We fill in the curve on the rest of the real line — the missing off-season observations in the first six seasons, plus any additional values requested from any season's curve due to time shift transformations described below — by copying the first available wILI value at earlier times, copying the last measurement at later times, and using linear interpolation at non-integer values. These

filled-in values are later used to construct variants of these curves that are shifted and/or stretched along the time axis in order to obtain a wider library of curves in the prior. Trend filtering seems better suited for epidemic data with than the more common smoothing spline fit because it is more "locally adaptive", responding better to varying levels of smoothness in data [Tibshirani, 2014], e.g., relatively sharp peaks mixed with smoother, flatter, less active regions. Figure 2.1 compares trend filtering, SIR, and smoothing spline fits for two fairly representative wILI trajectories. A reviewer for this work identified Bayesian nonparametric covariance regression [Fox and Dunson, 2011] as another alternative for fitting curves and noise models, which can incorporate heteroscedasticity and spatial relationships. More realistic versions of each of these smoothing methods would explicitly incorporate holiday effects on both the trend and on the noise; from Figure 2.1 and similar plots, we observe some potential issues with fits of the first three methods: (a) smoothing out and removing the holiday effects (which sometimes leads to inappropriately low trend estimates around the peak week), (b) undersmoothing of non-holiday weeks due to the impact of the holiday on smoothness parameter selection, and/or (c) oversmoothing of the entire curve, each accompanied by a related effect on estimates of noise levels.

Prior

The key assumption of the framework is that the current season will resemble one of the past seasons, perhaps with a few changes.

- **Shape:** The general shape f^r of the underlying curve is taken from one of the past seasons. We select each of the historical shapes with equal probability: $f^r \sim \mathsf{Unif}\{\hat{f}^{r,s} : \text{historical season } s\}$.

- **Noise:** The standard deviation of the normally distributed noise at each week is assumed to take on values from the past years' candidates with equal probability: $\sigma \sim \mathsf{Unif}\{\hat{\tau}^{r,s} : \text{historical season } s\}$.

- **Peak height:** The distribution of underlying peak heights is drawn from a continuous uniform distribution: $\theta \sim U[\theta_m, \theta_M]$. We use an unbiased estimator [Lehmann and Casella, 1998, Chapter 2] for θ_m and θ_M based on past seasons' trend filtered curves. The resulting curve is $f_2^r(i) = b^r + \frac{\theta^r - b^r}{\max_j f^r(j) - b^r}(f^r(i) - b^r)$, where b^r is the current year's CDC baseline wILI level (i.e., the onset threshold) for the selected geographical region r, e.g., 2% for the US as a nation for the 2013/2014 flu season.

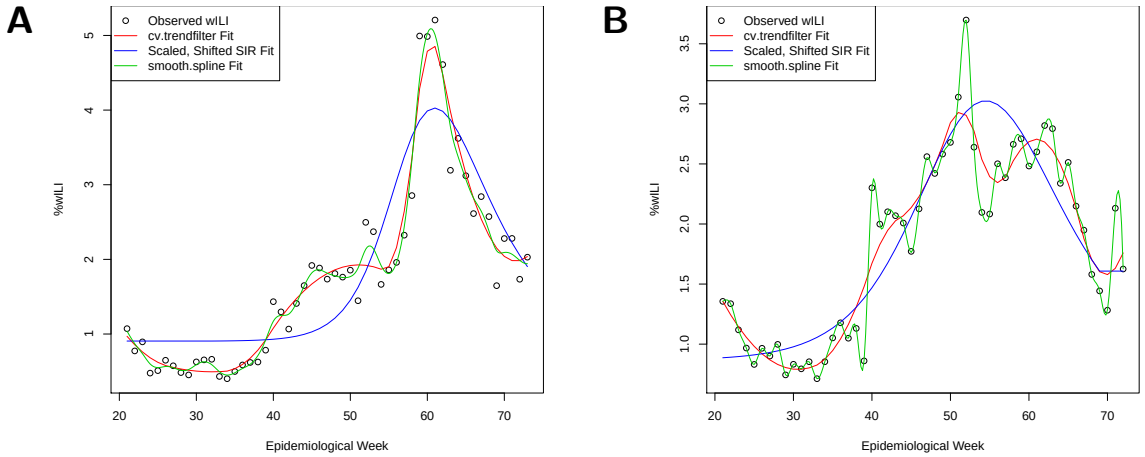

Figure 2.1: **Trend filtering, SIR, and smoothing spline fits for HHS region 3 for two seasons.** The quadratic trend filtering fit was performed with the `cv.trendfilter` [Arnold and Tibshirani, 2014] method, which automatically selects an base level of smoothness to use, and adapts to differences in smoothness in different parts of the trajectory. The cubic natural smoothing spline fit was produced by `smooth.spline` [R Core Team, 2015], which also automatically selects a level of smoothness, but by different criteria. (A) 2008/2009 season: trend filtering and smoothing splines both smooth out the holiday effects. The smoothing spline appears to overfit to noise in the preseason and early flu season. (B) 2006/2007 season: in addition to holiday effects, there is a large jump in wILI at week 40, which coincides with the beginning of the influenza season and a large jump in the number of reporting providers (from about 30 to over 100). The trend filtering procedure has trouble matching the beginning-of-season and holiday effects, attributing most of these effects to noise and smoothing them out. The `smooth.spline` procedure selects a level of smoothness that essentially duplicates the observed wILI and would produce a noise estimate near 0, which does not seem appropriate. Alternative methods of selecting a level of smoothness may produce looser fits and avoid these near-0 noise estimates, though. Beginning-of-season and holiday effects can be incorporated in both of the smoothing procedures, and would likely improve the resulting fits. Regional wILI dynamics are generally not tightly fit by the described SIR model.

- **Peak week:** The distribution of underlying peak weeks is formed in a similar manner to the peak height distribution; we find unbiased estimators μ_m, μ_M for uniform distribution bounds, but restrict the distribution to integral output: $\mu \sim \mathsf{Unif}\{i \in \{1..53\} : \mu_m \leq i \leq \mu_M\}$. The resulting curve is $f_3^r(i) = f_2^r(i - \mu^r + \arg\max_j f_2^r(j))$.

- **Pacing:** We allow for variations in the "pace" of an epidemic by incorporating a time scale that stretches the curve about the peak week; the distribution of time scale factors is $\nu \sim U[0.75, 1.25]$. The resulting curve is $f_4^r(i) = f_3^r\left(\frac{i - \arg\max_j f_3^r(j)}{\nu} + \arg\max_j f_3^r(j)\right)$.

To generate a possible curve for the current season, i.e., to sample from the prior, we independently sample a shape, noise level, peak height, peak week, and pacing parameter from the above distributions, then generate the corresponding wILI curve. We have also developed and are investigating an alternative "local" transformation prior [van Panhuis et al., 2014] that does not use information from other historical curves when transforming a particular historical curve f, but instead reuses the noise level for f and makes smaller *changes* to the peak week and height of f, which are restricted to a smaller, predefined range; this is more appropriate for surveillance data with less regular seasonal behavior, such as dengue case counts in Brazil.

In total, we model the underlying curve $f^{r,\,s_{\text{curr}}}$ for the current season as the curve generated by a randomly sampled parameter configuration $\langle f^r, \sigma^r, \nu^r, \theta^r, \mu^r \rangle$, using the following equation:

$$f^{r,\,s_{\text{curr}}}(i) = f_4^r(i) = b^r + \frac{\theta^r - b^r}{\max_j f^r(j) - b^r}\left[f^r\left(\frac{i - \mu^r}{\nu^r} + \arg\max_j f^r(j)\right) - b^r\right].$$

Figure 2.2 illustrates the peak week, peak height, and pacing transformations, and different levels of noise that could be considered. The data model for the current season's wILI values $y^{r,\,s_{\text{curr}}}$ is the same as that for historical seasons, shown in Equation 2.1.

Sampling from the posterior

We use importance sampling [Liu, 2008] to obtain a large set of curves from the posterior weighted by how closely they match the epidemic curve so far, beginning with week 40. More concretely, we obtain a single weighted sample from the posterior by (i) sampling a historical smoothed curve f, noise level σ, and transformation

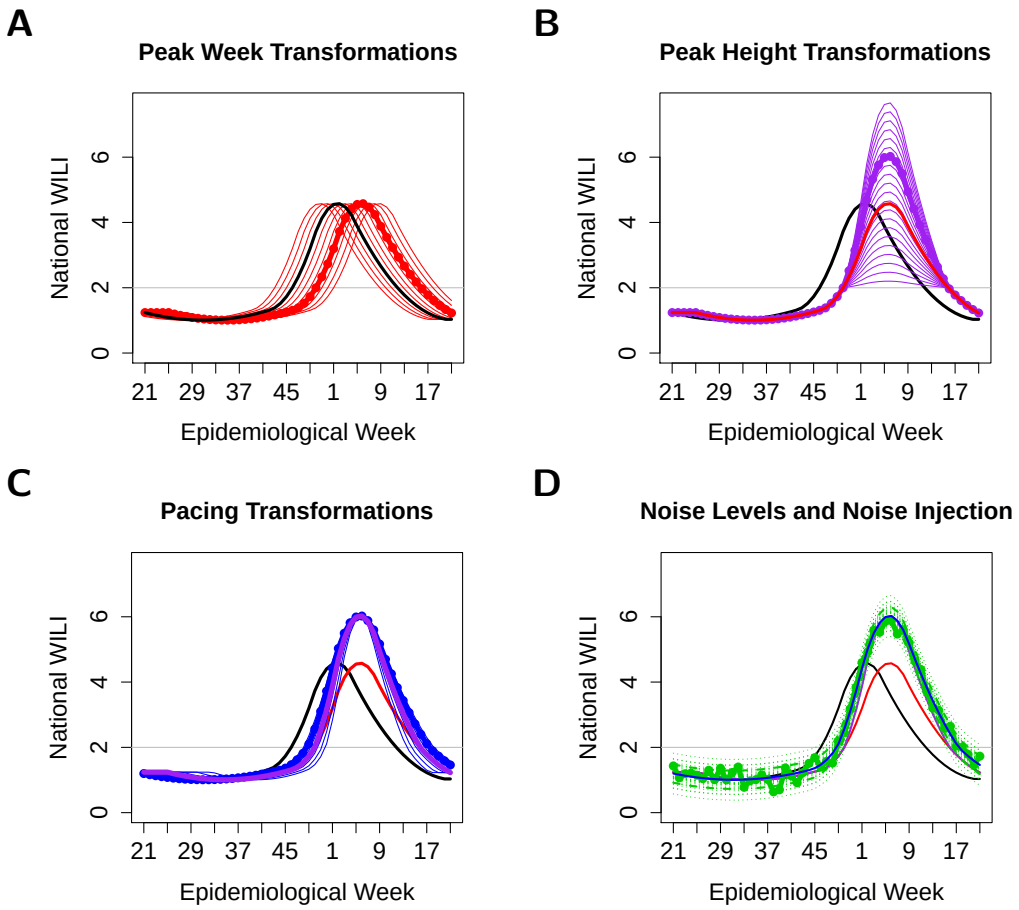

Figure 2.2: **Examples of possible peak week, peak height, and pacing transformations, and different noise levels.** Thick black, original curve; red, possible peak week transformations; thick red, a random peak week transformation; purple, possible peak height transformations; thick purple, a random peak height transformation; blue, possible pacing transformations; thick blue, a random pacing transformation; dotted green, 5th and 95th (pointwise) percentiles of noise distribution for possible noise levels; dashed green, percentiles for a random noise level; thick green, one possible trajectory for the selected transformations and noise level. (A) Peak week transformations. Peak weeks of historical smoothed curves occurred between weeks 51 and week 10 of the next year, so we limit transformations to give peak weeks roughly within this range. (B) Peak height transformations. Peak heights of historical smoothed curves were between 2% and 8%, so we limit transformations to give peak heights roughly within this range. (C) Pacing transformations. We stretch the curve by a factor between 75% and 125% about the peak week. (D) Noise levels. We randomly select one of 15 noise levels from the fitting procedure and add this level of Gaussian noise to the transformed curve.

parameters ν, θ, and μ from the prior; (ii) applying the peak height, peak week, and pacing transformations; (iii) assigning the curve an "importance weight" or "likelihood" based on how well it matches existing observations for the current seasons; and (iv) drawing noisy wILI observations around the curve for the rest of the season. We apply this procedure many times to obtain a collection of possible wILI trajectories and associated weights, forming a probability distribution over possible futures for the current season.

Sampling algorithm Outlined below is a simpler version of the sampling algorithm that does not use importance sampling.

Algorithm 1: One weighted sampling procedure for empirical Bayes model

Data: $y^{r,s_{\text{curr}}}$, the wILI observations so far; $z^{r,s_{\text{curr}}}$, a version of $y^{r,s_{\text{curr}}}$ with two extra points estimated from GFT; prior distributions of wILI curves, noise levels, and transformations

Result: weighted collection of curves

Let $\phi(x;\mu,\sigma)$ be the normal pdf;

for *a large number of times* **do**

  Randomly draw f^r, σ, ν, θ, and μ from the corresponding priors;

  Let $f^{r,s_{\text{curr}}}(i) = f_4^r(i) = b^r + \frac{\theta^r - b^r}{\max_j f^r(j) - b^r}\left[f^r\left(\frac{i-\mu^r}{\nu^r} + \arg\max_j f^r(j)\right) - b^r\right]$;

  Calculate weight $w = \prod_{i=1}^{\text{length}(z^{r,s_{\text{curr}}})} \phi\left(z; f^{r,s_{\text{curr}}}(i), \sigma\right)$;

  Let v be a 53-length vector, a possible curve for this season;

  for i *in* 1..length$(y^{r,s_{\text{curr}}})$ **do**

    $v_i \coloneqq y_i^{r,s_{\text{curr}}}$;

  end

  for i *in* (length$(y^{r,s_{\text{curr}}}) + 1$)..53 **do**

    $v_i \coloneqq f^{r,s_{\text{curr}}}(i)$;

  end

  Add curve v with weight w to the collection of possibilities for this season (the posterior estimate)

end

To improve computational efficiency, we also use an importance sampling technique that first divides up the possible values of f^r, σ, ν, θ, and μ into bins and estimates the average weight of $f^{r,s_{\text{curr}}}$'s in each bin using a single configuration from that bin. By sampling values of f^r, σ, ν, θ, and μ more frequently from the higher-weighted bins (and compensating appropriately for this decision in the weight calculation), we are able to construct a collection of curves with a high total weight

more quickly than the version above.

Forecasting targets

For the initial CDC ILINet forecasting challenge, we were interested in four forecasting targets: the epidemic's onset, peak week, peak height, and duration. These features were already used to summarize epidemic curves and perform retrospective analysis, and CDC selected them as forecasting targets for the competition, as accurate predictions of these milestones would assist policy makers in planning vaccination campaigns, resource allocation, and messages to the public.

- **Onset:** The first week that the wILI curve is above a specified CDC baseline wILI level, and remains there for at least the next two weeks. For example, the 2013/2014 national baseline wILI level was 2%, so the onset was the first in at least three consecutive weeks with wILI levels above 2%.

- **Peak Week:** The week(s) in which the wILI curve attains its maximum value.

- **Peak:** The maximum observed wILI value in a season.

- **Duration:** Roughly, how many weeks the wILI level remained above the CDC baseline since the onset. We defined this more rigorously as the sum of the lengths of all periods of three or more consecutive weeks with wILI levels above the CDC baseline.

We generate distributions for each of these targets by repeatedly (i) sampling a possible wILI trajectory and associated weight from the posterior, (ii) calculating the four forecasting targets for that trajectory, and (iii) storing these four values along with the trajectory's weight. We represent these forecasting target posterior distributions with histograms, and generate point estimates by taking the posterior mean for each target.

Sample empirical Bayes forecasts

Figure 2.3 illustrates some forecasts made by a version of the system described above for the same location at different times throughout the same influenza season.

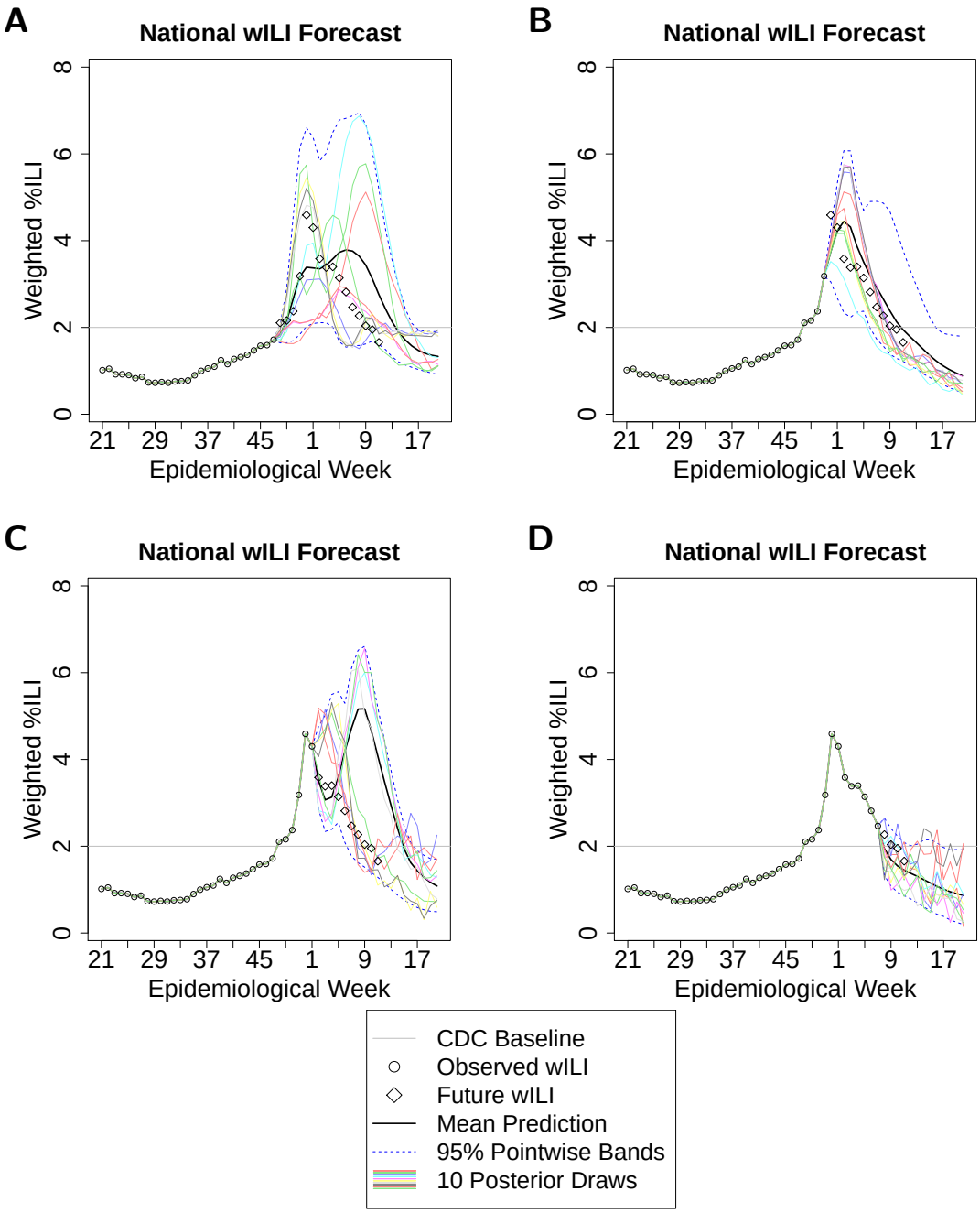

Figure 2.3: **2013/2014 national forecast, using empirical Bayes framework, retrospectively, using the final revisions of wILI values, using revised wILI data through epidemiological weeks (A) 47, (B) 51, (C) 1, and (D) 7.**

Cross-validation point prediction performance

Figure 2.4 shows the cross-validated error for national point predictions of our current empirical Bayes framework, as well a few other approaches, for each for the above four forecasting targets. The methods for predicting $\text{tar}_j^r(y^{r,s_{\text{cv}}})$ are summarized below.

- **Baseline (Mean of Other Seasons):** takes the average target value across the 14 other seasons, completely ignoring any data from the current season; provides an idea of whether other forecasters provide reasonable levels of error at the beginning of the season, and how much they benefit from incorporating data from the season they are forecasting.
- **Pinned Baseline (Mean of Other Seasons, Conditioned on Current Season to Date):** constructs 14 possible wILI trajectories for the current season by using the available observations for previous weeks and other historical curves for future weeks; reports the mean target value across these 14 trajectories; this is another very generic baseline that allows us to see the effect of using more complex wILI models and forecasting methods.
- **Pointwise Percentile (P2014) [van Panhuis et al., 2014]:** Constructs a single possible future wILI trajectory using the pointwise qth quantile from other seasons; estimates an appropriate value of q from the observed data so far, trying to match more recent observations more closely than less recent ones.
- k **Nearest Neighbors (knn):** Uses a method similar to existing systems for shorter-term prediction [Viboud et al., 2003] to identify k sections of other seasons' data that best match recent observations, and uses them to construct and weight k possible future wILI trajectories.
- **Empirical Bayes (Transformed Versions of Other Seasons' Curves):** Our current framework, using transformed versions of other seasons' curves to form the prior.
- **Empirical Bayes (SIR Curves):** Our current framework, using scaled and shifted SIR curves rather than other seasons' curves to form the prior; this is a somewhat similar approach to the SIRS-EAKF method used by the contest winner [Shaman and Karspeck, 2012a]. Figure 2.1 shows two fits to regional data.

Figure 2.4 indicates that, for all forecasting targets and most weeks, the average point prediction error for the EB method is similar (overlapping error bars) or lower

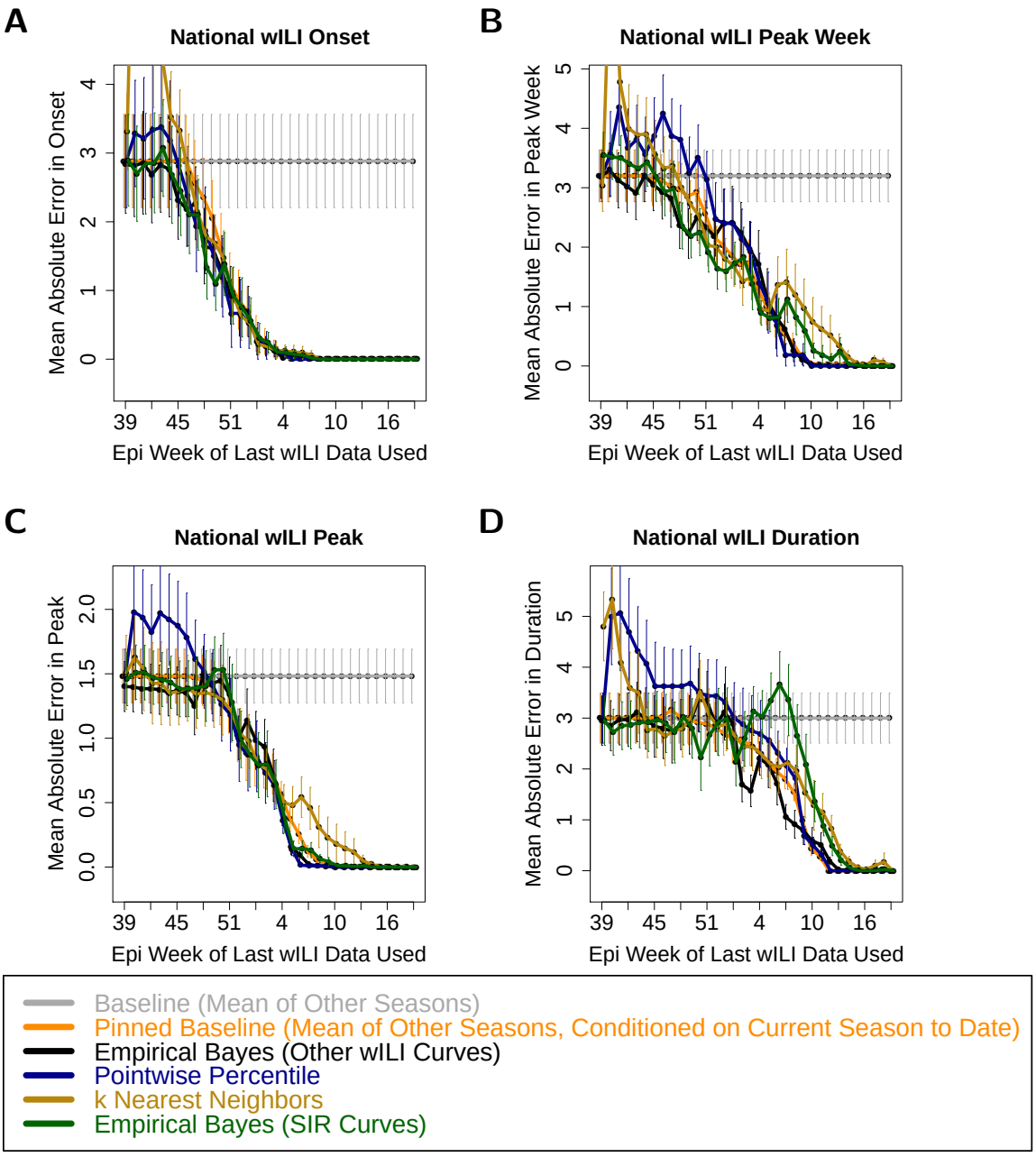

Figure 2.4: **Cross-validated mean absolute error estimates and standard error bars for point predictions for (A) onset, (B) peak week, (C) peak height, and (D) duration.** (The onset and duration were defined based on the 2% national threshold set by CDC for the 2013/2014 season.)

than the average error for the best predictor for that target and week. An important feature of this approach is that it provides a smooth distribution over possible curves and target values, rather than just a single point. From this distribution, we can calculate point predictions to minimize some expected type of error or loss, build credible intervals, and make probabilistic statements about future wILI and target values.

2.3 Kernel delta density

This section reproduces or incorporates content from Brooks et al. [2018].

Kernel density estimation and kernel regression use smoothing kernels to produce flexible estimates of the density of a random variable (e.g., $f_{Y_{t+1..T}}$) and the conditional expectation of one random variable given the value of another (e.g., $\mathbb{E}[Y_{t+1..T} \mid Y_{1..t}]$), respectively; we can combine these two methods to obtain estimates of the conditional density of one random variable given the value of another. One possible approach would be to use the straightforward entire-trajectory model

$$\hat{f}_{Y_{t+1..T}|Y_{1..t}}(y_{t+1..T} \mid y_{1..t}) = \frac{\sum_{s=1}^{S} I^{[1..t]}(y_{1..t}, Y_{(1..t)+(\Delta t)_s}) O^{[t+1..T]}(y_{t+1..T}, Y_{(t+1..T)+(\Delta t)_s})}{\sum_{s=1}^{S} I^{[1..t]}(y_{1..t}, Y_{(1..t)+(\Delta t)_s})},$$

where $\{1..S\}$ is the set of fully observed historical training seasons, and $I^{[1..t]}$ and $O^{[t+1..T]}$ are smoothing kernels describing similarity between "input" trajectories and between "output" trajectories, respectively. However, while basic kernel smoothing methods can excel in low-dimensional settings, their performance scales very poorly with growing dimensionality. During most of the season, neither $Y_{1..t}$ nor $Y_{t+1..T}$ is low-dimensional, and the current season's observations are extremely unlikely to closely match any past $Y_{(1..t)+(\Delta t)_s}$ or $Y_{(t+1..T)+(\Delta t)_s}$. This, in turn, can lead to kernel density estimates for $Y_{t+1..T}$ based almost entirely on the single season s with the closest $Y_{(1..t)+(\Delta t)_s}$ when conditioning on $Y_{1..t}$, and unrealistic density estimates for $Y_{t+1..T}$ even without conditioning on $Y_{1..t}$. The high-dimensional output issue is readily resolved by the chained one-ahead approach, combining univariate conditional density estimates for each observation conditioned on previous observations: $f_{\Delta Y_u|Y_{1..u-1}}$ for each u from $t+1$ to T, where $\Delta Y_u = Y_u - Y_{u-1}$. Estimating single-dimensional densities requires relatively little data. However, this reformulation exacerbates the high-dimensional input problem since we are conditioning on $Y_{1..u-1}$, which can be considerably longer than $Y_{1..t}$. We address the high-dimensional

input problem by approximating $f_{\Delta Y_u|Y_{1..u-1}}$ with $f_{\Delta Y_u|\mathbf{\Phi}^{[\mathrm{KDD},u]}}$ where $\mathbf{\Phi}^{[\mathrm{KDD},u]}$ is some low-dimensional vector of features derived from $Y_{1..u-1}$. The straightforward conditional density estimation method described above for $Y_{t+1..T} \mid Y_{1..t}$ can be applied to the chained distributions $\Delta Y_u \mid \mathbf{\Phi}^{[\mathrm{KDD},u]}$, although literature indicates that this approach is suboptimal [Hansen, 2004].

We developed the conditional estimates above based on combining kernel regression and univariate kernel density estimation techniques, it can also be understood as sampling from a joint kernel density estimate over input and output variables using a product kernel. A slightly more complicated take on the former viewpoint has been found to yield faster theoretical and simulated statistical convergence rates [Hansen, 2004]. The latter interpretation offers additional alternatives such as deriving results from a joint density estimate based on a kernel that is not the product of an input and output kernel, as well as copula techniques. These approaches have been incorporated in a separate epidemiological forecasting system working directly with the higher-dimensional inputs and outputs rather than the one-step-ahead approach [Ray et al., 2017]. A host of work on kernel conditional density estimation offers avenues to improving these kernel delta density approaches, as well as resolving the original issues regarding high dimensionality.

We use two sets of choices for the approximate conditional density function and summary features to form two versions of the method.

- **Markovian delta density:** approximates the conditional density of ΔY_u given $Y_{1..u-1}$ with its conditional density given just the previous (real or simulated) observation, Y_u:

$$\begin{aligned}
\hat{f}_{Y_{t+1..T}|Y_{1..t}}(y_{t+1..T} \mid y_{1..t}) &= \prod_{u=t+1}^{T_2} \hat{f}_{\Delta Y_u|Y_{1..u-1}}(\Delta y_u \mid y_{1..u-1}) \\
&= \prod_{u=t+1}^{T_2} \hat{f}_{\Delta Y_u|Y_{u-1}}(\Delta y_u \mid y_{u-1}) \\
&= \prod_{u=t+1}^{T_2} \frac{\sum_s I^{[u]}(y_{u-1}, Y_{u-1+(\Delta t)_s}) \cdot O^{[u]}(\Delta y_u, \Delta Y_{u+(\Delta t)_s})}{\sum_s I^{[u]}(y_{u-1}, Y_{u-1+(\Delta t)_s})},
\end{aligned}$$

  where $I^{[u]}$ and $O^{[u]}$ are Gaussian smoothing kernels. The first equality corresponds to the chain rule of probability on the actual (not estimated) densities; the second incorporates the Markov assumption (i.e., selects $\mathbf{\Phi}^{[u]} = [Y_{u-1}]$); and the third gives our choice of estimators for the conditional densities $\hat{f}_{\Delta Y_u|Y_{u-1}}$ for each u. The bandwidth of each $I^{[u]}$ and $O^{[u]}$ is chosen separately using

bandwidth selection procedures for regular kernel density estimation of Y_{u-1} and ΔY_u, respectively. (Specifically, we use the `bw.SJ` function from the `R`[R Core Team, 2015] built-in `stats` package, with `bw.nrd0` as a fallback in the case of errors. These functions do not accept weights for the inputs; it may be possible to improve forecast performance by incorporating these weights or by using other approaches to select the bandwidths.) Note that density estimates for ΔY_u are based on data from past seasons on week u only, allowing the method to incorporate seasonality and holiday effects (for holidays that consistently occur at the same time of year).

Forecasts are based on Monte Carlo simulations of $Y_{t+1..T} \mid Y_{1..t}$ using the chained one-step-ahead procedure described in the previous section. This process is illustrated in Figure 2.5. Repeating this procedure many times yields a sample from the model for $Y_{t+1..T} \mid Y_{1..t}$; stopping at 2000 draws seems sufficient for use in our ensemble forecasts, while at least 7000 are needed to smooth out noise when displaying distributional target forecasts for the delta density method in isolation. Any negative simulated wILI values in these trajectories are clipped off and replaced with zeroes.

- **Extended delta density:** approximates the conditional density of ΔY_u given $Y_{1..u-1}$ with its conditional density given four features:
  - the previous wILI value, Y_{u-1};
  - the sum of the previous k^u wILI values, roughly corresponding to the sum of wILI values for the current season;
  - an exponentially weighted sum of the previous k^u wILI; values, where the weight assigned to time u' is $0.5^{t'-u'}$; and
  - the previous change in wILI value, ΔY_{u-1}.

  The approximate conditional density assigns each of these features a weight (0.5, 0.25, 0.25, and 0.5, respectively) in order to reduce overfitting and emphasize some relative to the others, and incorporates data from other weeks close to u (specifically, within l^u weeks; the choice of l^u is discussed in a later sectionadd holiday section somewhere and reference here) with a truncated Laplacian kernel. We selected these weights and other settings, such as kernel bandwidth selection rules, somewhat arbitrarily based on intuition and experimentation on out-of-sample data; a cross-validation subroutine could be used to make the selection as well, but would multiply the amount of computation required. In case the resulting product of Gaussian and Laplacian kernels is

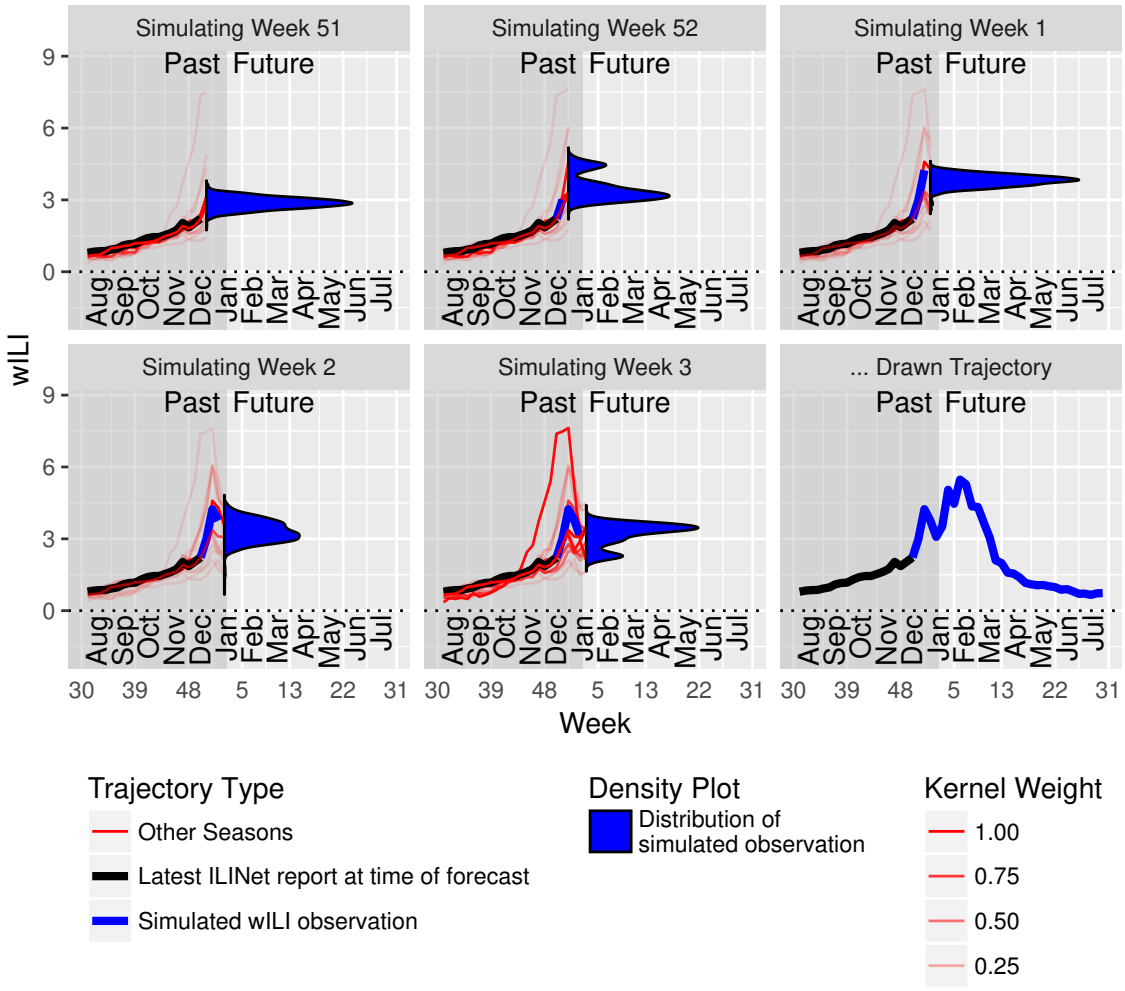

Figure 2.5: **The delta density method conditions on real and simulated observations up to week $u-1$ when building a probability distribution over the observation at week u.** This figure demonstrates the process for drawing a single trajectory from the Markovian delta density estimate. The past data $Y_{1..t}$, which incorporates observations through week 48, is shown in black. Kernel smoothing estimates for future values at times u from $t+1$ to T_2 are shown in blue, as are simulated observations drawn from these estimates. Past seasons' trajectories are shown in red, with alpha values proportional to the weight they are assigned by the kernel I^u.

too narrow, we mix its results with a wide boxcar kernel which evenly weights all data from time $u - l^u$ to $u + l^u$:

$$\hat{f}_{\Delta Y_u | Y_{1..u-1}}(\Delta y_u \mid y_{1..u-1})$$
$$= 0.9 \cdot \frac{\sum_s \sum_{u'=u-l^u}^{u+l^u} 0.7^{|u'-u|} \left[I_1^u(y_{u-1}, Y_{(u'-1)+(\Delta t)_s}) \right]^{0.5} \cdots O^u(\Delta y_u, \Delta Y_{(u')+(\Delta t)_s})}{\sum_s \sum_{u'=u-l^u}^{l^u} 0.7^{|u'-u|} \left[I_1^u(y_{u-1}, Y_{(u'-1)+(\Delta t)_s}) \right]^{0.5} \cdots \left[I_4^u(\Delta y_{u-1}, \Delta Y_{(u'-1)+(\Delta t)_s}) \right]^{0.5}}$$
$$+ 0.1 \cdot \frac{\sum_s \sum_{u'=u-l^u}^{u+l^u} O^u(\Delta y_u, \Delta Y_{(u')+(\Delta t)_s})}{\sum_s \sum_{u'=u-l^u}^{u+l^u} 1}.$$

Using data from $u' \neq u$ incorporates additional reasonable outcomes for Δy_u by incorporating past wILI patterns with different timing, but risks including some very unreasonable possibilities produced by repeatedly drawing from the same u' rather than following seasonal trends with increasing u''s. For example, when a portion of a past season that is more similar to itself with a slight time shift than to any other past season, it may be selected for multiple consecutive u's and produce an unreasonable trajectory. This could potentially occur when drawing data from the relatively flat regions of wILI trajectories of many seasons, or when incorporating observations around an unusually early, late, high, or low peak. To prevent this possibility, we combine the natural estimate for Y_u arising from the density estimate for ΔY_u with a random draw $Y_{\text{uncond}u}$ from the unconditional density estimate for Y_u (using a Gaussian kernel and only data from week u):

$$Y_u^{\text{sim}} = 0.9 \cdot \left(Y_{u-1} + \Delta Y_u^{\text{sim}} \right) + 0.1 \cdot Y_u^{\text{uncond}}.$$

2.4 Quantile autoregression

Locally linear quantile regression offers an alternative approach to modeling $Y_u \mid Y_{1..u-1}$ offering greater flexibility in covariate relationships and better anticipated behavior with larger numbers of covariates; post-processing its output with additional random noise is one way to address potential issues with discrete outputs that do not cover the entire support of Y_u. Basic linear quantile regression estimates the τth conditional quantile of some variable Y given covariates X as a linear function of X; locally linear quantile regression additionally allows for weighting of training instances based on a smoothing kernel on another set of covariates X' (potentially overlapping with X). Additionally, the same types of transformations can be applied on the output and covariates as in the kernel smoothing case. A specification of a simple locally linear quantile autoregression approach could consist of:

- $\Psi^{[u]}$: a transformation of Y_u from which we can recover Y_u (potentially using information from $Y_{1..u-1}$),
- $\mathbf{\Phi}^{[\text{QARlinear},u]}$: a set of features (derived from $Y_{1..u-1}$) to use in the linear combination estimating some quantile of Y_u,
- $\mathbf{\Phi}^{[\text{QARkernel},u]}, K^{\mathbf{\Phi}^{[\text{QARkernel},u]}}$: a set of features (derived from $Y_{1..u-1}$) and corresponding smoothing kernel (or "weighted" smoothing kernel as used in extended delta density) that assigns weights to training instances, and
- $K^{\Psi^{[u]}}$, a smoothing kernel that defines the distribution of additive post-processing noise.

The corresponding sampling procedure for Y_u^{sim} is:

1. Draw quantile level $\tau \sim U[0,1]$.
2. Compute estimate $\hat{q}$ of the level τ quantile of $\Psi^{[u]} \mid \Phi^{[\text{QARlinear},u]}, \Phi^{[\text{QARkernel},u]}$ using locally linear quantile regression.
3. Draw $\epsilon \sim K^{\Psi^{[u]}}$ from post-processing noise distribution.
4. Let $\Psi^{[u],\text{sim}} = \hat{q} + \epsilon$.
5. Let Y_u^{sim} be the value of Y_u given by $\Psi^{[u]} = \Psi^{[u],\text{sim}}$ and $Y_{1..u-1}$.

Quantile autoregression has already been formulated and studied from a theoretical perspective and applied to economic datasets [Koenker and Xiao, 2006]. A recent application to flu forecasting [Wang, 2016] studied different data weighting approaches based on time of season. Similarly, quantile autoregression can be applied to epidemiological data and customized based on domain knowledge.

2.4.1 Connection to smoothing kernel approaches

The family of locally linear quantile autoregression approaches above subsumes the considered delta density approaches after mirroring any heuristic modifications to the kernel conditional density estimates. Consider a kernel conditional density estimate of $\Delta Y_u \mid \mathbf{\Phi}^{[\text{KDD},u]}$ using covariate kernel $K^{\mathbf{\Phi}^{[\text{KDD},u]}}$. If the response kernel $K^{\Delta Y_u}$ is replaced with the degenerate Dirac delta distribution, the resulting kernel conditional "density" estimates are just weighted empirical distributions. The corresponding

quantiles are weighted sample quantiles of ΔY_u with weights based on $K^{\mathbf{\Phi}^{[\text{KDD},u]}}$; this coincides with the estimated quantiles of the locally linear/constant quantile regression model with the same $\Psi^{[u]}$, $\mathbf{\Phi}^{[\text{QARlinear},u]} = (1)$ (the model only fits an "intercept"), $\mathbf{\Phi}^{[\text{QARkernel},u]} = \mathbf{\Phi}^{[\text{KDD},u]}$, and $K^{\mathbf{\Phi}^{[\text{QARkernel},u]}} = K^{\mathbf{\Phi}^{[\text{KDD},u]}}$. The sampling procedures also coincide: drawing from a weighted empirical distribution function gives an equivalent distribution to selecting a weighted sample quantile with a level randomly distributed on the unit interval. ("Sample quantile" here is restricted to quantiles of the type outputted by quantile regression; for a finite number of quantile levels, there will not be a unique associated sample quantile and the one selected may vary across implementations, but these levels are drawn with probability 0. For other types of quantiles, e.g., from continuous quantile functions [Hyndman and Fan, 1996], this is normally not the case.) Using $K^{\Delta Y_u}$ instead of the Dirac distribution is equivalent to just adding additional noise to a draw from the weighted empirical distribution; thus, the smoothing kernel approach can be completely mimicked by a locally linear quantile regression approach using the same $K^{\Delta Y_u}$ as the post-processing noise distribution.

2.4.2 Incorporating covariates inspired by mechanistic models

While quantile regression can be restricted and post-processed to match the output of the kernel conditional density method, it is natural to favor use of $\mathbf{\Phi}^{[\text{QARlinear},u]}$ covariates not only in appeal to more general statistical arguments regarding scaling with higher dimensionality inputs and boundary bias, but also due to similarities with domain-driven mechanistic models when incorporating autoregressive terms. Furthermore, additional covariates can be constructed to strengthen this resemblance while maintaining the flexibility of quantile modeling and smoothing kernel weighting.

Epidemiological compartmental models are a popular class of mechanistic model that divides a population into a fixed number of "compartments" and considers all individuals within each compartment to behave identically. System dynamics are characterized by the manner in which individuals are added, removed, or flow between different compartments. For example, "SIRS" compartmental models represent population state by the number or proportion of individuals in each of three states: those

- Susceptible to infection with some disease,

- Infectious and spreading the disease, and
- Recovered from the infectious stage of a disease and currently immune to future reinfection;

Susceptible individuals can become Infectious by interacting with Infectious individuals, Infectious individuals transition to Recovered over time, and Recovered individuals can become Susceptible again due to waning immunity or mismatches of antibodies with currently circulating strains of a pathogen; these possible transitions are the basis for the initialism "SIRS". A simple deterministic, continuous-time, proportion-based SIRS model can be specified with the following system of differential equations:

$$\begin{aligned}
s'(t) &= -s(t) \cdot \beta i(t) + r(t) \cdot \mu \\
i'(t) &= +s(t) \cdot \beta i(t) - i(t) \cdot \gamma \\
r'(t) &= +i(t) \cdot \gamma - r(t) \cdot \mu \\
s(0) + i(0) + r(0) &= 1, s(0) \geq 0, i(0) \geq 0, r(0) \geq 0,
\end{aligned}$$

where

- $s(t)$, $i(t)$, and $r(t)$ are the proportions of the population in the Susceptible, Infectious, and Recovered states, respectively, at time t;
- β is the rate at which any individual experiences contact with another person in which the latter could potentially spread an infection to the former (assumed to be the same across all pairs of individuals, regardless of their current state), potentially modulated by the current weather (i.e., $\beta(\mathbf{w})$ where $\mathbf{w}$ is a vector of weather variables) or other data;
- μ is the rate at which recovered individuals become susceptible again;
- γ is the rate at which infectious individuals recover; and
- the conditions on the state at $t = 0$ are preserved as invariants for all other t.

The underlying proportions $s(t)$, $i(t)$, and $r(t)$ are latent; a simple noiseless observation model assumes that infectious individuals produce some kind of reported health care events at a steady rate, with no false positives from the other compartments:

$$y(t) = i(t) \cdot N\rho,$$

where

- $y(t)$ is the number of reported health care events at time t,
- N is the population size, and
- ρ is the rate at which infectious individuals generate reported health care events;

we sometimes refer to SIRS models incorporating observations of medically "Attended" cases as a "SIRSA" models. Already, this formulation suggests the use of models with linear autoregressive terms, as changes to compartment occupancy depend linearly or quadratically on the current occupancy, and the observations depend linearly on compartment occupancy. However, the latent dynamics and quadratic terms complicate the relationship; fortunately, a few manipulations will allow us to fully characterize the dynamics of $y(t)$ without any reference to latent state, revealing a very direct relationship with linear autoregressive and additional auxiliary terms. (These types of manipulations and others have been developed before [Heffernan et al., 2005, Hethcote and Tudor, 1980], but such analyses often do not include an observation model, and focus on system behavior and parameter inference rather than prediction.) The techniques used are likely more widely familiar in the context of differential equations than discrete-time difference equations, so we examine the former first then establish parallels in the latter.

Our ultimate goal is to express $y'(t)$ as a causal function of $y(t)$ (i.e., a function depending only on $y(\tau)$ for $\tau \leq t$). First, note that

$$\begin{aligned} y'(t) &= i'(t) \cdot N\rho \text{ and} && \text{(derivatives are linear)} \\ i(t) &= \frac{1}{N\rho} y(t) && \text{(scale both sides of } y(t) \text{ definition)} \end{aligned}$$

so we can instead seek to express $i'(t)$ as a causal function of $i(t)$ and quickly obtain $y'(t)$ as a causal function of $y(t)$. Next, observe that

$$\begin{aligned} i'(t) &= \beta s(t) i(t) - \gamma i(t) \\ &= \beta[1 - i(t) - r(t)] i(t) - \gamma i(t), && \text{(proportions sum to 1)} \end{aligned}$$

so we just need to express $r(t)$ as a causal function of $i(t)$. Rearranging the equation

for $r'(t)$ and applying an integrating factor approach, we find that

$$\begin{aligned} r'(t) &= i(t) \cdot \gamma - r(t) \cdot \mu \\ \mu r(t) + r'(t) &= \gamma i(t) \\ \mu e^{\mu t} r(t) + e^{\mu t} r'(t) &= \gamma e^{\mu t} i(t) \\ e^{\mu t} r(t) &= \gamma \int_{t_0}^{t} e^{\mu \tau} i(\tau) \, \mathrm{d}\tau + C \\ r(t) &= \gamma \int_{t_0}^{t} e^{-\mu(t-\tau)} i(\tau) \, \mathrm{d}\tau + C e^{-\mu t}, \end{aligned}$$

for

- a time t_0 which is arbitrary for this derivation, but which we must select to be in the range of times for which observations are available, to ensure the integral involves only observed values of its argument, and
- a constant of integration $C \geq 0$ determining the initial conditions;

thus, $r(t)$ can be represented as a scaled exponential moving average of $i(t)$ (a causal function of $i(t)$) plus an exponential decay term. Applying the earlier observations gives

$$\begin{aligned} i'(t) &= \beta[1 - i(t) - r(t)]i(t) - \gamma i(t) \\ &= \beta[1 - i(t) - \gamma \int_{t_0}^{t} e^{-\mu(t-\tau)} i(\tau) \, \mathrm{d}\tau - C e^{-\mu t}]i(t) - \gamma i(t) \\ &= (\beta - \gamma) \left[i(t) \right] - \beta \left[i^2(t) \right] - \beta\gamma \left[\int_{t_0}^{t} e^{-\mu(t-\tau)} i(\tau) \, \mathrm{d}\tau \cdot i(t) \right] - \beta C \left[e^{-\mu t} i(t) \right] \end{aligned}$$

and

$$\begin{aligned} y'(t) &= i'(t) \cdot N\rho \\ &= N\rho(\beta - \gamma) \left[i(t) \right] - N\rho\beta \left[i^2(t) \right] - N\rho\beta\gamma \left[\int_{t_0}^{t} e^{-\mu(t-\tau)} i(\tau) \, \mathrm{d}\tau \cdot i(t) \right] - N\rho\beta C \left[e^{-\mu t} i(t) \right] \\ &= (\beta - \gamma)[y(t)] - \frac{\beta}{N\rho} \left[y^2(t) \right] - \frac{\beta\gamma}{N\rho} \left[\int_{t_0}^{t} e^{-\mu(t-\tau)} y(\tau) \, \mathrm{d}\tau \cdot y(t) \right] - \beta C \left[e^{-\mu t} y(t) \right]. \end{aligned}$$

The discrete-time analogues of the key equations above and some additional trans-

formations follow:

$$
\begin{aligned}
s_{t+1} &= s_t - \beta s_t i_t + \mu r_t \\
i_{t+1} &= i_t + \beta s_t i_t - \gamma i_t \\
r_{t+1} &= r_t + \gamma i_t - \mu r_t \\
s_0 + i_0 + r_0 &= 1, s_0 \geq 0, i_0 \geq 0, r_0 \geq 0 \\
y_t &= N\rho i_t \\
\Delta y_{t+1} = y_{t+1} - y_t &= (\beta - \gamma)\left[y_t\right] - \frac{\beta}{N\rho}\left[y_t^2\right] - \frac{\beta\gamma}{N\rho}\left[\sum_{t_0}^{t-1}(1-\mu)^{t-1-\tau} y_\tau \cdot y_t\right] - \beta C\left[(1-\mu)^{t-1} y_t\right] \\
y_{t+1} &= (1 + \beta - \gamma)\left[y_t\right] - \frac{\beta}{N\rho}\left[y_t^2\right] - \frac{\beta\gamma}{N\rho}\left[\sum_{t_0}^{t-1}(1-\mu)^{t-1-\tau} y_\tau \cdot y_t\right] - \beta C\left[(1-\mu)^{t-1} y_t\right] \\
\frac{\Delta y_{t+1}}{y_t} &= (\beta - \gamma)\left[1\right] - \frac{\beta}{N\rho}\left[y_t\right] - \frac{\beta\gamma}{N\rho}\left[\sum_{t_0}^{t-1}(1-\mu)^{t-1-\tau} y_\tau\right] - \beta C\left[(1-\mu)^{t-1}\right].
\end{aligned}
$$

The last few equations motivate the use of the bracketed quantities on the right as covariates in a regression for the response variable given on the left. However, the parameter μ is unknown, so the last two bracketed quantities in each of these equations cannot be formed as stated. Fortunately, additional manipulations of the last equation resolve this issue:

$$
\begin{aligned}
\frac{\Delta y_{t+1}}{y_t} &= (\beta - \gamma)\left[1\right] - \frac{\beta}{N\rho}\left[y_t\right] - \frac{\beta\gamma}{N\rho}\left[\sum_{t_0}^{t-1}(1-\mu)^{t-1-\tau} y_\tau\right] - \beta C\left[(1-\mu)^{t-1}\right] \\
&= (\beta - \gamma)\left[(1-\mu)\cdot 1 + \mu \cdot 1\right] - \frac{\beta}{N\rho}\left[(1-\mu)y_{t-1} - (1-\mu)y_{t-1} + y_t\right] \ldots \\
&\ldots - \frac{\beta\gamma}{N\rho}\left[(1-\mu)\sum_{t_0}^{t-2}(1-\mu)^{t-2-\tau} y_\tau + y_{t-1}\right] - \beta C\left[(1-\mu)\cdot(1-\mu)^{t-2}\right] \\
&= (1-\mu)\left[\frac{\Delta y_t}{y_{t-1}}\right] + (\beta - \gamma)\left[\mu \cdot 1\right] - \frac{\beta}{N\rho}\left[-(1-\mu)y_{t-1} + y_t\right] - \frac{\beta\gamma}{N\rho}\left[y_{t-1}\right] \\
&= (1-\mu)\left[\frac{\Delta y_t}{y_{t-1}}\right] + (\beta - \gamma)\mu\left[1\right] - \frac{\beta}{N\rho}\left[y_t\right] - \frac{\beta}{N\rho}(\gamma + \mu - 1)\left[y_{t-1}\right] \\
\Delta\left[\frac{\Delta y_{t+1}}{y_t}\right] &= -\mu\left[\frac{\Delta y_t}{y_{t-1}}\right] + (\beta - \gamma)\mu\left[1\right] - \frac{\beta}{N\rho}\left[y_t\right] - \frac{\beta}{N\rho}(\gamma + \mu - 1)\left[y_{t-1}\right] \\
\Delta y_{t+1} &= (1-\mu)\left[\frac{y_t}{y_{t-1}}\Delta y_t\right] + (\beta - \gamma)\mu\left[y_t\right] - \frac{\beta}{N\rho}\left[y_t^2\right] - \frac{\beta}{N\rho}(\gamma + \mu - 1)\left[y_t y_{t-1}\right] \\
y_{t+1} &= (1-\mu)\left[\frac{y_t}{y_{t-1}}\Delta y_t\right] + (\beta\mu - \gamma\mu + 1)\left[y_t\right] - \frac{\beta}{N\rho}\left[y_t^2\right] - \frac{\beta}{N\rho}(\gamma + \mu - 1)\left[y_t y_{t-1}\right].
\end{aligned}
$$

The bracketed quantities contain no latent state and no unknown parameters; they can be incorporated as covariates in a nonlinear autoregression framework.

Parameter and latent state inference

The primary goal of this effort is to inform construction of a higher-quality, easily fit model for y_t; any interpretation regarding the latent state is suspect in such a simplistic model, particularly causal or counterfactual reasoning, and especially if some parameters are nonidentifiable. Still, it is notable that we can recover (estimates of) μ, γ, β, $N\rho$, and the latent state time series, at least in this particular SIRS model (identifiability is discussed in the next heading). Performing the linear regression suggested by the equation for $\Delta\left[\frac{\Delta y_{t+1}}{y_t}\right]$ above gives an estimate for μ: the negation of the coefficient fit for $\frac{\Delta y_{t+1}}{y_t}$. Then, treating μ as given in the equation for $\frac{\Delta y_{t+1}}{y_t}$ involving exponentially weighted moving averages and performing the linear regression suggested there, we obtain fit coefficients $\theta_1..\theta_4$, which can be used to find

the rest of the parameters and latent state mentioned above:

$$
\begin{aligned}
\theta_1 &= \beta - \gamma \\
\theta_2 &= -\frac{\beta}{N\rho} \\
\theta_3 &= -\frac{\beta\gamma}{N\rho} \\
\theta_4 &= -\beta C \\
\gamma &= \frac{\theta_3}{\theta_2} && = \frac{-\beta\gamma/(N\rho)}{-\beta/(N\rho)} && \text{(using definitions of } \theta_2, \theta_3) \\
\beta &= \theta_1 + \frac{\theta_3}{\theta_2} && = \theta_1 + \gamma && \text{(using definition of } \theta_1) \\
N\rho &= -\frac{\theta_1}{\theta_2} - \frac{\theta_3}{\theta_2^2} && = -\frac{\beta}{\theta_2} && \text{(using definition of } \theta_2) \\
C &= -\frac{\theta_4}{\theta_1 + \frac{\theta_3}{\theta_2}} && = -\frac{\theta_4}{\beta} && \text{(using definition of } \theta_4) \\
i_t &= \frac{y_t}{N\rho} \\
r_t &= \gamma \sum_{t_0}^{t-1} (1-\mu)^{t-1-\tau} i_\tau + C(1-\mu)^{t-1} && && \text{(parallel of continuous-time result)} \\
s_t &= 1 - i_t - r_t.
\end{aligned}
$$

The parameter estimates for key parameters γ, β, and $N\rho$ contain θ_2 as a divisor and may be sensitive to errors in its estimates; one point of interest for inference work is whether estimates of these key parameters can be improved by alternative formulations involving $1/theta_2$ or $\log\theta_2$ and/or fitting techniques other than ordinary linear regression. Extending this approach to a probabilistic formulation likely will likely involve some of these transformations and alternative regression techniques as well, e.g., log-transformed parameters and multinomial regression.

Identifiability

At least for some selections of parameter values and initial state, the quantities above are uniquely identifiable. For example, this appears to be the case with β = 0.3, γ = 0.15, μ = 0.001, $N\rho$ = 300, s_0 = 0.80, i_0 = 0.01, and r_0 = 0.19, and with observations starting at t_0 = 1, where parameter and latent state estimates from the

procedure described above exactly match the true values given a sufficient number of observations (six observations gives relative absolute errors below 10^{-4} for μ and below a standard threshold around $1.5 \cdot 10^{-8}$ for all quantities but μ; at sixteen or more observations, relative absolute error for μ also appears to stabilize below this narrower threshold). However, in this SIRS model, starting from the above parameter values, the latent state and observed values will approach a fixed point called the endemic equilibrium, shown in Figure 2.6. If the first observation t_0 was instead after the system had already stabilized at the endemic equilibrium, or if i_0 was instead selected to be 0, then none of the above quantities can be fully identified (save for $i_t = 0$ for all t in the latter case, when assuming $N\rho > 0$). The inference procedure will also encounter issues due to multicollinearity among features. Still, this inference procedure may be of interest in a pandemic scenario where a steady state has not been reached and this simple SIRS model is deemed appropriate. A more important observation is that the modeled flat-line asymptotic behavior does not line up with real-world disease dynamics that would be of interest to predict; randomness, seasonality, and a host of other details have been omitted. Adding these features to a mechanistic framework and performing similar derivations and identifiability analysis is of interest for future work; here, though, we will rely on the one-step-ahead conditional distribution framework to compensate for some of these omissions.

Discussion, future directions, and variants on derivations

The derivations above present some exciting possibilities for fitting compartmental models using standard regression routines, which may scale more readily than particle filter and MCMC approaches. While the derivations are based on a deterministic model, various types of regression models, such as quantile regression and generalized linear models, can be applied to the covariate-response combinations suggested by the above equations, providing a way to introduce noise into the model. The noise introduced seems to correspond to a type of process noise in i_t, but not process noise in s_t and r_t nor observational noise in y_t; the latter is especially important when dealing with noisy signals so momentary fluctuations are not mistaken for trends. We explore performance of the simplest method along these lines — just adding the bracketed quantities listed immediately preceding Section 2.4.2 into a quantile autoregression alongside existing covariates, with no re-derivations, no constraints on the fit coefficients, etc. — in Chapter 4. Unfortunately, this approach is unsuccessful in improving overall performance due at least in part to superexponential growth in some simulated trajectories. Further investigation could ascertain the cause or

Figure 2.6: **Simple SIRS models' state approaches a fixed point.** Shown here are the values of the latent state of a simple deterministic SIRS model where the latent state approaches a fixed point called an endemic equilibrium, where the proportion of individuals who are infectious is approximately a constant greater than zero. Another possible fixed point, approached given $i_0 = 0$ and/or certain other disease parameters configurations leading to disease extinction, is a fully susceptible population.

causes of superexponential growth — which could include the simple manner in which the covariates were included in the quantile autoregression framework, issues with the deterministic SIRSA model used to derive the covariates (discussed in the next paragraph), and/or numerical computation issues with the simulation procedure when paired with these covariates (hinted as a possibility below in Figure 2.7) — and could develop a derivation, fitting, and simulation framework that avoids such superexponential growth.

Many details have been oversimplified or omitted entirely in the SIRSA model considered above: there is no process noise nor observational noise (which is addressed only partially by use of derived covariates in an autoregressive framework), resistance to a disease is either complete immunity or susceptibility, and is lost instantaneously after an exponentially-distributed amount of time following recovery from infection; the infectious contact rate is constant, ignoring seasonal trends, weather, school calendars, and holidays; the population is treated as homogeneous and interactions fully mixed, without breakdowns into geographical, demographic, and occupational groups, nor any social network structure; multiple diseases, types, subtypes, and strains and their interaction are not incorporated; public health responses are not included; only a single surveillance source has been considered, with a case reporting rate that neglects holiday, media, geographic, demographic, and disease strain-based effects, and no false positives; the population is constant with no birth, death, or migration; importation of cases from other locations or animal reservoirs is excluded; and so on. Incorporating some of these details will require generalization of the derivations applied to the simplistic model studied above. Existing work which could help toward this end includes: (a) spectral methods for predictive linear dynamical models (e.g., hidden Markov models (HMMs) [Hsu et al., 2012], kernelized HMMs [Song et al., 2010], and linear dynamical systems (Kalman filtering) [Boots, 2012]), for fitting models with multivariate observational and process noise in a tractable, deterministic manner, (b) differential equation and control theory machinery such as the state-transition matrix and Peano–Baker series [Baake and Schlaegel, 2011] for generalizing the integral and exponential-moving-sum reformulations to the multivariate case; (c) linear algebra tools such as the Schur complement for further multivariate manipulations, and (d) quantile regression extensions such as the multiple quantile graphical model [Ali et al., 2016] and quantile filtering [Johannes et al., 2009] for maintaining a nonparametric noise model. This type of analysis may benefit from existing work or find additional applications in other modeling domains, including population ecology, chemical rate equations, and general agent-based modeling.

The next few headings provide auxiliary derivations with more interpretable and extensible formulations for the same deterministic SIRS model, and work towards application in more complicated setups.

More interpretable formulation without reliance on sum constraint By applying an integrating factor approach to re-express both $r(t)$ and $s(t)$ in terms of $i(t)$ and more carefully expressing constants of integration, we arrive at a more interpretable formulation which also does not rely on the constraint $s(t)+i(t)+r(t) = 1$, providing a path forward to cases incorporating changes in population.

A similar integrating factor approach is used throughout many derivations. Note that, given some time series $x(t)$, $a(t)$, $b(t)$, if

$$x'(t) = a(t)x(t) + b(t)$$

for all t, then:

$$\begin{aligned}
x'(t) - a(t)x(t) &= b(t) \\
e^{-\int_{t_0}^{t} a(\tau)\,\mathrm{d}\tau}\left[x'(t) - a(t)x(t)\right] &= e^{-\int_{t_0}^{t} a(\tau)\,\mathrm{d}\tau} b(t) \\
e^{-\int_{t_0}^{t} a(\tau)\,\mathrm{d}\tau} x(t) &= x(t_0) + \int_{t_0}^{t} e^{-\int_{t_0}^{\nu} a(\tau)\,\mathrm{d}\tau} b(\nu)\,\mathrm{d}\nu \\
x(t) &= x(t_0) e^{\int_{t_0}^{t} a(\tau)\,\mathrm{d}\tau} + \int_{t_0}^{t} e^{\int_{t_0}^{t} a(\tau)\,\mathrm{d}\tau - \int_{t_0}^{\nu} a(\tau)\,\mathrm{d}\tau} b(\nu)\,\mathrm{d}\nu \\
&= x(t_0) e^{\int_{t_0}^{t} \mathrm{d}\tau\, a(\tau)} + \int_{t_0}^{t} \mathrm{d}\nu\, b(\nu) e^{\int_{\nu}^{t} \mathrm{d}\tau\, a(\tau)}.
\end{aligned}$$

This manipulation may be hard to interpret in the abstract form above, but becomes clear as it is applied to transform the SIRS differential equations. Recall that

$$\begin{aligned}
s'(t) &= -s(t) \cdot \beta i(t) + r(t) \cdot \mu \\
i'(t) &= +s(t) \cdot \beta i(t) - i(t) \cdot \gamma \\
r'(t) &= +i(t) \cdot \gamma - r(t) \cdot \mu.
\end{aligned}$$

Using the manipulation above with $x(t) = r(t)$, $a(t) = -\mu$, and $b(t) = i(t) \cdot \gamma$, we obtain:

$$\begin{aligned}
r(t) &= r(t_0) e^{\int_{t_0}^{t} (-\mu)\,\mathrm{d}\tau_{rr}} + \int_{t_0}^{t} \mathrm{d}\nu_{ir}\, i(\nu_{ir}) \gamma e^{\int_{\nu_{ir}}^{t} \mathrm{d}\tau_{rr}\,(-\mu)} \\
&= r(t_0) e^{-(t-t_0)\mu} + \int_{t_0}^{t} \mathrm{d}\nu_{ir}\, i(\nu_{ir}) \gamma e^{-\mu(t-\nu_{ir})}.
\end{aligned}$$

Multiplying through by N gives an easily read interpretation:

$$Nr(t) = Nr(t_0)e^{-(t-t_0)\mu} + \int_{t_0}^{t} \mathrm{d}\nu_{ir}\, Ni(\nu_{ir})\gamma e^{-\mu(t-\nu_{ir})};$$

that is, the number of recovered agents at time t is the sum of:

- the number of recovered agents at time t_0 that never lost immunity (as of time t),
- the number of infectious agents that recovered at some time ν_{ir} between t_0 and t and did not subsequently lose immunity before time t.

The same manipulation can be used to re-express $s(t)$ and the result combined with the above expression for $r(t)$:

$$\begin{aligned}
s'(t) &= -\beta i(t)s(t) + \mu r(t) \\
s(t) &= s(t_0)e^{-\int_{t_0}^{t} \mathrm{d}\tau_{ss}\,\beta i(\tau_{ss})} + \int_{t_0}^{t} \mathrm{d}\nu_{rs}\, r(\nu_{rs})\mu e^{-\int_{\nu_{rs}}^{t} \mathrm{d}\tau_{ss}\,\beta i(\tau_{ss})} \\
&= s(t_0)e^{-\int_{t_0}^{t} \mathrm{d}\tau_{ss}\,\beta i(\tau_{ss})} + \int_{t_0}^{t} \mathrm{d}\nu_{rs} \left(r(t_0)e^{-\mu(\nu_{rs}-t_0)} + \int_{t_0}^{\nu_{rs}} \mathrm{d}\nu_{ir}\, i(\nu_{ir})\gamma e^{-\mu(\nu_{rs}-\nu_{ir})} \right) \mu e^{-\int_{\nu_{rs}}^{t} \mathrm{d}\tau_{ss}\,\beta i(\tau_{ss})} \\
&= s(t_0)e^{-\int_{t_0}^{t} \mathrm{d}\tau_{ss}\,\beta i(\tau_{ss})} + \\
&\ldots r(t_0) \int_{t_0}^{t} \mathrm{d}\nu_{rs}\, e^{-\mu(\nu_{rs}-t_0)} \mu e^{-\int_{\nu_{rs}}^{t} \mathrm{d}\tau_{ss}\,\beta i(\tau_{ss})} + \\
&\ldots \int_{t_0}^{t} \mathrm{d}\nu_{rs} \int_{t_0}^{\nu_{rs}} \mathrm{d}\nu_{ir}\, i(\nu_{ir})\gamma e^{-\mu(\nu_{rs}-\nu_{ir})} \mu e^{-\int_{\nu_{rs}}^{t} \mathrm{d}\tau_{ss}\,\beta i(\tau_{ss})};
\end{aligned}$$

multiplying the equations through by N yields a statement that the number of susceptible agents at time t is the sum of:

- the number of susceptible agents at time t_0 that were never infected (as of time t),
- the number of recovered agents at time t_0 that lost immunity between times t_0 and t but were never subsequently infected (as of time t), and
- the number of infectious agents that recovered at some time ν_{ir} between t_0 and t that subsequently lost immunity at some time ν_{rs} and remained susceptible until time t.

Plugging the above expressions into the differential equation for $i'(t)$ yields:

$$\begin{aligned}
i'(t) &= +s(t) \cdot \beta i(t) - i(t) \cdot \gamma \\
&= s(t_0) e^{-\int_{t_0}^{t} \mathrm{d}\tau_{ss}\, \beta i(\tau_{ss})} i(t) \dots \\
&\dots + r(t_0) \int_{t_0}^{t} \mathrm{d}\nu_{rs}\, e^{-\mu(\nu_{rs} - t_0)} \mu e^{-\int_{\nu_{rs}}^{t} \mathrm{d}\tau_{ss}\, \beta i(\tau_{ss})} \beta i(t) \dots \\
&\dots + \int_{t_0}^{t} \mathrm{d}\nu_{rs} \int_{t_0}^{\nu_{rs}} \mathrm{d}\nu_{ir}\, i(\nu_{ir}) \gamma e^{-\mu(\nu_{rs} - \nu_{ir})} \mu e^{-\int_{\nu_{rs}}^{t} \mathrm{d}\tau_{ss}\, \beta i(\tau_{ss})} \beta i(t) \dots \\
&\dots - i(t) \cdot \gamma.
\end{aligned}$$

Here, $i'(t)$ is expressed using only $i(t)$, $i(\cdot)$ at previous times, and the initial conditions of $s(t_0)$ and $r(t_0)$. Similarly, $y'(t)$ can be similarly expressed as a causal function of $y(t)$ and initial latent state alone, as in the original derivation. The integrating factor technique could also potentially be used to re-express some or all instances of $i(\cdot)$ in terms of itself at previous times, and lead to even more ways to express $y'(t)$ under the same constraints. This reformulation does not rely on the fact that $s(t) + i(t) + r(t) = 1$, and so this approach could provide opportunities to derive covariates from models incorporating birth and death events. It does, however, rely on the fact that in the graph of possible state transitions for individual agents, the subgraph for inexactly observed or unobserved states — s and r in this derivation — is acyclic, in order for the number of integrals in the final expression for $i(t)$ to be finite, or, in the discrete-time case, for the number of summations in the expression for i_t to not grow with $t - t_0$. Such a situation may introduce additional obstacles in later steps requiring more powerful techniques to resolve.

Posynomial reformulation of discrete-time SIRS model A "posynomial" is a function of the form

$$\sum_{k=1}^{K} c_k x_1^{a_{1k}} x_2^{a_{2k}} \dots x_n^{a_{nk}},$$

where $c_k, x_1, x_2, \dots, x_n > 0$ and $a_{jk} \in \mathbb{R} \forall j \in \{1..n\}, k \in \{1..K\}$; recognizing or reformulating optimization problems (such as model fitting) in terms of posynomials sometimes enables the use of certain efficient, reliable algorithms for "geometric programming" [Boyd and Vandenberghe, 2004, sec. 4.5]. The latent state and observations in the simple discrete-time SIRS model used in this chapter can be quickly rewritten as posynomial function of transformed parameters and quantities from the

previous time step, assuming they are all nonzero:

$$\begin{aligned}
s_{t+1} &= s_t - \beta s_t i_t + \mu r_t && = (s_t + \bar{\beta} i_t + r_t) s_t + \mu r_t; \\
i_{t+1} &= i_t + \beta s_t i_t - \gamma i_t && = \bar{\gamma} i_t + \beta s_t i_t; \\
r_{t+1} &= r_t + \gamma i_t - \mu r_t && = \bar{\mu} r_t + \gamma i_t; \\
y_{t+1} &= N \rho i_{t+1} && = N \rho \bar{\gamma} i_t + N \rho \beta s_t i_t; \text{ where}
\end{aligned}$$

where $\bar{\beta} = 1 - \beta$, $\bar{\gamma} = 1 - \gamma$, and $\bar{\mu} = 1 - \mu$ are treated as additional parameters; the first equation can be interpreted as stating that susceptible individuals at time $t + 1$ are either:

- susceptible individuals from time t that potentially interacted with:
  - a susceptible individual,
  - an infectious individual, but were not infected, or
  - a recovered individual, and
- recovered individuals that lost immunity between times t and $t + 1$.

One issue with this reformulation is that constraining $\beta + \bar{\beta} = 1$, $\gamma + \bar{\gamma}$, $\mu + \bar{\mu}$, and $s_t + i_t + r_t = 1$ cannot be implemented directly in geometric programming (although relaxations replacing $=$ with $<$ are possible); however, this issue may actually not apply in the context of more complicated, believable SIRS models which incorporate birth and death. This posynomial reformulation can be applied recursively to rewrite the state and observations at time $t + 1$ as a posynomials of parameters and initial state, or perhaps used in combination with the manipulations from the heading above to rewrite Ns_{t+1}, Ni_{t+1}, and Nr_{t+1} as posynomials of parameters, initial state, and observations $y_{1..t}$. This strategy may transfer to probabilistic SIRS models using binomial random variables, leading to statements such as

$$\begin{aligned}
Y_{t+k} \mid S_0, I_0, R_0, Y_{1..t} \sim B(S_0, g_{\mathrm{S},0,t+k}(\theta)) + B(I_0, g_{\mathrm{I},0,t+k}(\theta)) + B(R_0, g_{\mathrm{R},0,t+k}(\theta)) + \\
\sum_{\tau=1}^{t} B(Y_\tau, g_{\mathrm{Y},\tau,t+k}(\theta)),
\end{aligned}$$

for count random variables $S_0, I_0, R_0, Y_{1..t+k}$, and posynomial functions $g_{\cdot,\cdot}(\cdot)$ which may have special forms.

However, experiments with the deterministic SIRS models also reveals a potential hazard with rewriting the original system equations: superexponential numerical

error growth. Figure 2.7 shows values of y_t over time given the same initial state and same parameters using the original SIRS equations and using the posynomial reformulation. The first plot shows that y_t in the posynomial formulation does indeed appear to coincide with y_t from the original equations for some time, supporting the assertion that the reformulation is mathematically valid with exact arithmetic. However, the second plot, which shows the same time series over an expanded time frame, demonstrates superexponential error growth in the posynomial formulation. Shortly after the end of the second plot, at time $t = 106$, y_t in the posynomial reformulation becomes infinity when using double-precision floating-point arithmetic. Clearly, the invariant $s_t + i_t + r_t = 1$ has been violated, and $s_t + i_t + r_t$ in the posynomial reformulation grows superexponentially as well. The exponential-moving-sum and other reformulations may or may not exhibit this same type of numerical error growth issue when applied in a purely mechanistic setting with known initial state and parameters, but naïvely adding the derived covariates to the quantile autoregression framework, it does demonstrate superexponential growth to infinity in some of its simulations; some performance metrics are shown in Chapter 4.

Deterministic SIRS; discrete types; one infection/immunity at a time We can extend the model above to incorporate multiple disease types in certain ways while avoiding the extra complication of cycles within the latent state transition graph. There are many schemes for building multi-strain models, partially driven by an explosion in the number of possible states and parameters in straightforward, faithful approaches [Kucharski et al., 2016]. We take a simple approach intended to readily scale to several pathogens or strains by allowing for only a single infection at a time, and forgetting information about all but the most recent infection of each individual. When modeling M different strains, the model includes $2M + 1$ possible agent states:

- susceptible to all strains;
- infectious with strain k, for $k \in \{1..M\}$;
- recovered from strain l, for $l \in \{1..M\}$.

Cross-protection and differing infectious contact, recovery, reporting, and waning immunity rates are captured by a quadratic number of parameters:

- β_k, $k \in \{1..M\}$: infectious contact rates for interactions of fully-susceptible individuals with strain-k-infectious individuals;

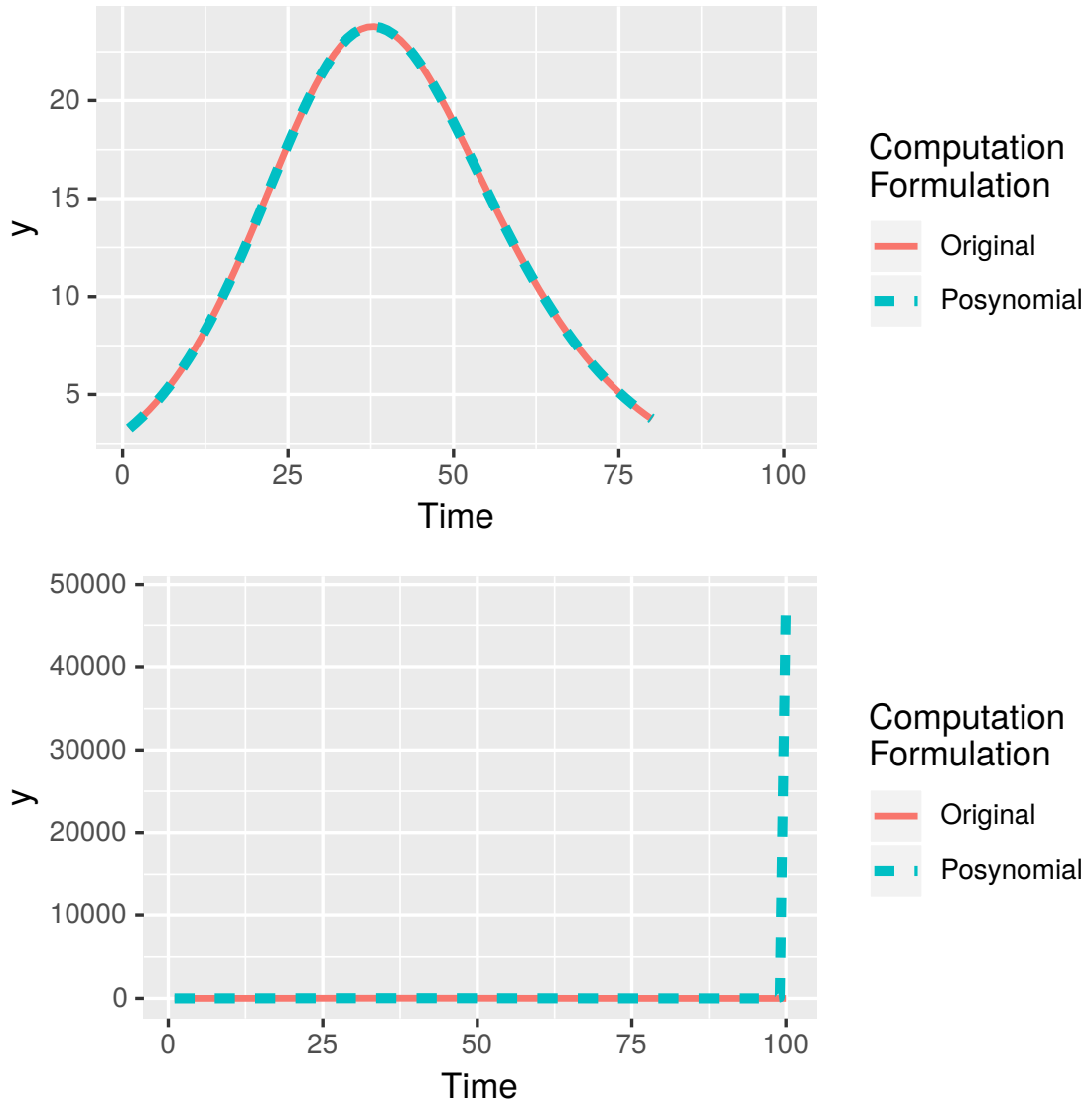

Figure 2.7: **A posynomial reformulation of SIRS equations appears to coincide numerically with the original equations for some time, but subsequent time steps demonstrate superexponential error growth.** Both of the plots shows y_t versus t as calculated by the original SIRS equations and the posynomial reformulation; the first plot covers $t \in \{1..80\}$, while the second covers $t \in \{1..100\}$. The source of the errors evident in the second plot is entirely numerical; the parameters and initial state are given exactly, rather than estimated.

- β_{kl}, $k, l \in \{1..M\}$, $k \neq l$: infectious contact rates for interactions of strain-l-recovered individuals with strain-k-infectious individuals;
- γ_k, $k \in \{1..M\}$: recovery rate for strain k; and
- ρ_k, $k \in \{1..M\}$: reporting rate for strain k; and
- μ_l, $l \in \{1..M\}$: waning immunity rate for strain l.

The differential equations are still quite similar to the single-strain case:

$$\begin{aligned}
s'(t) &= -\sum_k \beta_k s(t) i_k(t) + \sum_l \mu_l r_l(t) \\
i_k'(t) &= \beta_k i_k(t) s(t) + \sum_{l \neq k} \beta_{kl} i_k(t) r_l(t) - \gamma_k i_k(t) \\
r_l'(t) &= \gamma_l i_l(t) - \sum_{k \neq l} \beta_{kl} i_k(t) r_l(t) - \mu_l r_l(t) \\
s(t) + \sum_k i_k(t) + \sum_l r_l(t) &= 1 \\
y_k(t) &= N \rho_k i_k(t)
\end{aligned}$$

Furthermore, manipulations similar to the single-strain case appear to be possible, e.g., rewriting the number of recovered individuals in terms of the number of infectious individuals:

$$\begin{aligned}
r_l'(t) &= \gamma_l i_l(t) - \sum_{k \neq l} \beta_{kl} i_{k \neq l}(t) r_l(t) - \mu_l r_l(t) \\
r_l(t) &= \int_{t_0}^{t} e^{\left(\sum_{k \neq l} \beta_{kl} \int_{t_0}^{\nu} i_k(\tau)\,\mathrm{d}\tau + \mu_l(\nu - t_0)\right) - \left(\sum_{k \neq l} \beta_{kl} \int_{t_0}^{t} i_k(\tau)\,\mathrm{d}\tau + \mu_l(t - t_0)\right)} \gamma_l i_l(\nu)\,\mathrm{d}\nu + \\
&\quad \ldots r_l(t_0) e^{-\sum_{k \neq l} \beta_{kl} \int_{t_0}^{t} i_l(\tau)\,\mathrm{d}\tau - \mu_l(t - t_0)} \\
r_l(t) &= \int_{t_0}^{t} \gamma_l i_l(\nu) e^{-\sum_{k \neq l} \beta_{kl} \int_{\nu}^{t} i_l(\tau)\,\mathrm{d}\tau - \mu_l(t - \nu)}\,\mathrm{d}\nu + \\
&\quad \ldots r_l(t_0) e^{-\sum_{k \neq l} \beta_{kl} \int_{t_0}^{t} i_l(\tau)\,\mathrm{d}\tau - \mu_l(t - t_0)}
\end{aligned}$$

However, additional difficulties and complications are expected when fitting coherent exponential moving average parameters or working around them.

2.5 Incorporating holiday effects

This subsection reproduces or incorporates content from Brooks et al. [2018].

Holidays can impact the spread, observation, and impact of a disease. For example, reduced school and workplace contact may reduce disease transmission, patients may not seek or may delay medical care for less serious issues, and some health care providers may not be open or operate with reduced staffing. The delta density methods described above attempt to match holiday behaviors by restricting training windows around major holidays to focus on data from the same, or nearby, weeks of the year. This reduction in the amount of training data might actually degrade performance. A more direct model of the holiday effects may allow a model to match holiday behavior with less data, and simultaneously remove the perceived need for narrow training windows.

For example, CDC's wILI measure is an estimate of the proportion of health care visits in an area that are due to ILI. Sharp rises and drops in wILI are common from early or mid-December to early January (roughly coinciding with a four week period beginning with epi week 50), with either the season's peak or a lower, secondary peak commonly occurring on epi week 52. This pattern appears to arise from at least two factors:

- spikes downward in the number of non-ILI visits during the holiday season (corresponding to increases in wILI), perhaps caused by patients choosing not to visit the doctor for less serious issues on holidays, and

- decreasing slope of the average ILI visit curve during the holidays (changing from its highest positive value to a slightly negative value), perhaps due to "deceleration" in the true incidence of ILI resulting from a decreased average infectious contact rate during holidays, which partially counteracts the above increases in wILI due to health care seeking behavior changes during the holidays, but also accentuates this spike visually due to the negative slope at the end of the holiday period.

Similarly, there are spikes or minor blips downward in the average number of non-ILI visits (which can result in small increases in wILI) associated with Thanksgiving Day; Labor Day; Independence Day; Memorial Day; Birthday of Martin Luther King, Jr.; Washington's Birthday; Columbus Day; and perhaps other holidays. Spikes upward in wILI at Thanksgiving can push wILI unexpectedly over the onset threshold, and holiday effects may help explain the surprising frequency at which peaks occur on epi week 7 but not neighboring weeks. Figure 2.8 summarizes these effects using average ILI, non-ILI, and wILI trends for all age groups. Additional age-specific patterns may be obscured by this analysis of overall trends.

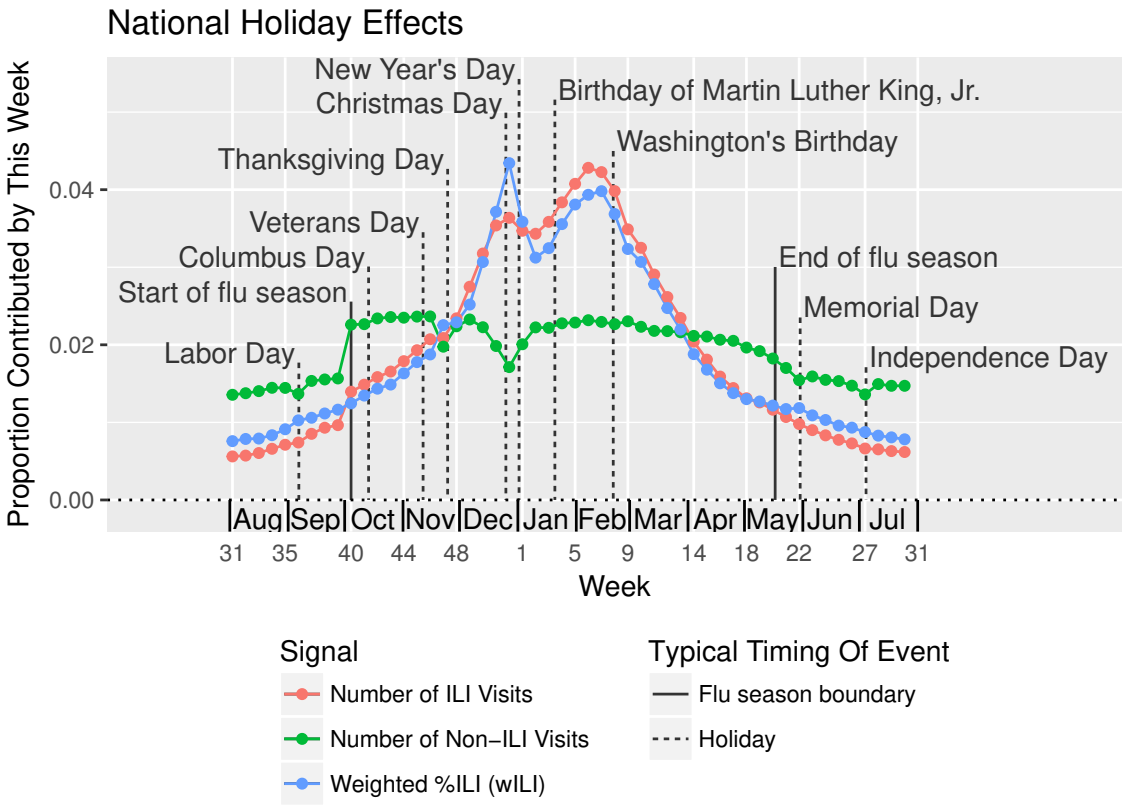

Figure 2.8: **On average, wILI is higher on holidays than expected based on neighboring weeks.** Weekly trends in wILI values, as expressed by the contribution of a each week to a sum of wILI values from seasons 2003/2004 to 2015/2016, excluding 2008/2009 and 2009/2010 (which include portions of the 2009 influenza pandemic), show spikes and bumps upward on and around major holidays. (U.S. federal holidays are indicated with event lines.) The number of non-ILI visits to ILINet health care providers spikes downwards on holidays (disproportionately with any drops in the number of ILI visits), contributing to higher wILI. The number of ILI visits generally declines in the second half of the winter holiday season, causing winter holiday peaks to appear even higher relative to nearby weeks. In addition to holiday effects, we see that average ILINet participation jumps upward on epi week 40, and gradually tapers off later in the season and in the off-season.

Approaches to incorporating holiday effects include:

- in entire-trajectory models, adding constant or random multiplicative effects to each trajectory in the set or distribution of possible $Y_{1..T}$'s;
- in chained one-ahead models, weighting training instances based on whether they share the same holiday status as the test instance or come from the same time of year; and
- in chained one-ahead models, adding holiday indicator variables and/or interactions of holiday indicator variables with other quantities of interest as covariates in the quantile autoregression framework, and/or modeling the number of ILI and non-ILI visits instead of %wILI.

Preliminary performance analysis suggests that:

- The first approach applied within the empirical Bayes framework produces trajectories in the prior that are qualitatively more similar to actual surveillance trajectories, but doesn't lead to better forecast performance. This phenomenon arose when fitting a constant (across elements of the prior) or randomly scaled multiplier patterns to weeks of the season typically associated with Thanksgiving and with winter holidays, which was fit in conjunction with the trend filtering procedure.
- Each U.S. federal holiday above occurs at roughly the same time of year every year, falling on one of two possible epi week numbers. Thus, models that predict behavior at a given epi week by prioritizing or focusing solely on past behavior at that given epi week will automatically perform a rough adjustment for holiday effects. This factor informs our decision to use historical data only from corresponding weeks in the Markovian delta density method, and a truncated Laplacian kernel with narrower width near winter holidays in the extended delta density method. Specifically, for the extended delta density method, we choose the half-width of the kernel to be $l^u = \min\{10, \max\{0, |u - 22| - 1\}\}$, which assigns $l^u = 0$ for u within one week of epi week 52, and larger l^u's the farther u is from this time period, up to a maximum value of 10. However, the extended delta density method actually exhibits a large degree of bias in ground truth estimates around these holidays; this bias or error seems attributable to other details of the extended delta density model, though, as it is present even in the week with a kernel width of 0, which causes the truncated Laplacian kernel to match the Markovian delta density kernel.

- One method in the third category is to add week-of-season or time-relative-to-holiday indicator variables and/or interactions with other quantities as covariates in the quantile autoregression framework; initial performance results suggest this method is surprisingly unsuccessful despite matching well with the intuition for the multiplicative nature of surveillance system holiday effects in this setting.

Chapter 3

Modeling surveillance data revisions

The above discussion assumed that, when forecasting future measurements of disease prevalence, we have access to these same desired measurements for all times in the past. In reality, this exact data is not immediately accessible, as accurate measurements may take weeks or years to be completed. However, to enable decisionmakers to quickly assess and respond to a situation, epidemiological surveillance systems often publish a sequence of provisional estimates of each complete measurement, with later versions more accurate on average. The existence of multiple versions of measurements has significant implications for proper forecast evaluation and analysis, and explicitly accounting for the revision process can improve model forecasts:

- **Faithful retrospective validation:** When estimating the performance of a proposed model by mimicking the forecasts it would have made in the past, it is important that we input the version of each measurement that would have been available at the time of each forecast; otherwise, accuracy estimates will almost surely be too high since the evaluation was based on higher accuracy input data.

- **Faithful forecast visualization:** Visualizing past forecasts together with completed measurements can cause confusion when the version of the measurements fed into the forecast has significant error; plotting the available version alongside the complete measurements and forecast can eliminate this confusion.

- **Forecast improvement:** Forecast performance can potentially be improved by

modeling the data revision process in addition to future observations, especially when a small change in past observations can cause a large change in the prediction target or associated forecast evaluations (as is sometimes the case for some timing and overall intensity targets), or when there is a high degree of error in earlier versions of measurements.

Section 3.1 describes in more detail the nature of some provisional data, and Section 3.2 provides notation to discuss such data. Section 3.3 describes methods for distribution prediction of finalized data given the partial, provisional data that are available in real time. Chapter 4 describes additional extensions and studies the performance of different approaches.

3.1 Examples of provisional data

This section reproduces or incorporates content from Brooks et al. [2018].

The use of provisional estimates is commonplace in epidemiological surveillance, and is also observed in other contexts; for example:

- ILINet is a network of health care providers that voluntarily submit reports to CDC, which cleans and aggregates the data. Providers may differ in timeliness and frequency of reporting, and new providers may enter the system and might provide a chunk of data, and the aggregate measure of ILI prevalence is updated as additional providers submit or revise their data. CDC adjusts for the fact that different versions will be based on different numbers of providers by reporting the *proportion* of visits due to ILI, but earlier versions can still be biased, as slower or less frequent reporters may serve different populations with higher or lower typical ILI proportions than earlier reporters. The revisions may also be correlated across time, as a lower frequency or slower huge provider or group of similar providers may report a chunk of multiple weeks at the same time. CDC may also perform data cleaning, which can affect the entire season at the same time; for example, they may remove all data from a particular provider.
- The Influenza Hospitalization Surveillance Network (FluSurv-NET) is a surveillance network for laboratory-confirmed influenza hospitalizations. Many of the issues above still are applicable; for example, differences in types of laboratory

test used, testing location, testing capacity, hospital administration, etc., can contribute to differences in timeliness of reporting between hospitals. Reporting may not take place until after a patient is discharged, which spreads reports apart further based on uncontrollable factors regarding duration of patients' illnesses. Additionally, reports may be revised after cases are ruled out as additional tests are performed. The combined effect is that the initially reported hospitalization rates are always or nearly always lower than the finalized figures, and half of the time are $\approx 50\%$ of the finalized value or less, while later versions have a growing chance of overestimating the finalized value but are closer to it on average.

- Gross domestic product (GDP) and gross national product (GNP) estimates can also be revised over time. Previous work has named different types of updates and addressed the task of forecasting these updates in the context of Kalman filtering [Aruoba, 2008, Jacobs and Van Norden, 2011, Julio et al., 2011, Mankiw and Shapiro, 1986].

Details of which provisional data are released and the nature of their revisions vary by setting; some specifics for the ILINet system are included below.

ILINet versioning process Recall that ILINet is a network of outpatient health care providers providing statistics which are compiled, processed, and published by CDC. These health care providers vary in size, types of care provided, administrative resources available, and nature of their recordkeeping and reporting systems. Not all providers transmit statistics on a weekly basis in time for inclusion in the quickest CDC estimates of ILI activity for every week. After this first deadline has passed, ILINet members provide "backfill" statistics or revisions for past weeks and CDC continually considers data cleaning operations of incoming and priorly submitted statistics; wILI observations are updated accordingly. Forecasting performance can be improved by modeling and "backcasting" these updates, accounting for the following sources of error:

- **Biased early reports:** earlier wILI versions are generally biased downwards early in the in-season, and upwards towards the end of the in-season, which may lead to forecasts of lower, later peaks early in the season, and of longer epidemic duration later in the season;
- **Overconfident short-term distributional forecasts:** since updates in wILI can cause "observed" data, e.g., of the wILI at the presumed peak week, to

shift, ignoring backfill may lead to "thin", overconfident forecast distributions;

- **Revisions of "observed" seasonal targets:** wILI updates sometimes cause large changes in the apparent onset week or peak week when there are bumps or multiple peaks in the trajectory: wILI updates can cause a measurement to change from above the CDC baseline to below (or vice versa), or for an earlier, lower peak to rise above a later peak (or vice versa); ignoring backfill updates can cause models to completely miss some possibilities when these targets appear to be determined. A similar type of error can arise from revisions to the peak height value (regardless of whether the peak week changes); even small updates can result in large unibin log score penalties.

3.2 Notation

In Chapter 2, the goal was to estimate the distribution of future observations of a time series of interest, $Y_{t+1..T}$, as a function of past observations of that time series, $Y_{1..t}$. However, as described above, we do not have access to $Y_{1..t}$ itself in real time, but instead a sequence of provisional reports, $Y_1^{(1)}, Y_{1..2}^{(2)}, \ldots, Y_{1..t-1}^{(t-1)}, Y_{1..t}^{(t)}$, each adding a new (provisional) observation and revising previous values. Our goal now is to build a distributional forecast of the entire, finalized time series of interest, $Y_{1..T_2}$, effectively leveraging information from provisional measurements $Y_1^{(1)}, Y_{1..2}^{(2)}, \ldots, Y_{1..t}^{(t)}$ and completed measurements $Y_{1..T_1}$ (where $T_1 \leq t$ and $T_1 < T_2$). That is, we want to jointly "backcast" (a.k.a. "backforecast", "back-forecast") $Y_{T_1+1..t}$ and forecast $Y_{t+1..T_2}$, and append the results to observations $Y_{1..T_1}$. Figure 3.1 depicts the prediction targets together with the observations available as of report t, which visually form a "provisional data triangle".

3.3 Nonparametric one-ahead backcasting and forecasting methods

There is an approach to backcasting and forecasting in this setting which is very similar in nature to the chained one-ahead future trajectory simulation procedure from Subsection 2.1.2. We can simulate a random trajectory $Y_{1..T_2}^{\text{sim}}$ from the distribution of $Y_{1..T_2}$ given all provisional data $Y^{(1..t)}$ by chaining together $T_2 - T_1$ 1-step-ahead simulations:

Time	$t-4$	$t-3$	$t-2$	$t-1$	t	$t+1$	$t+2$	$t+3$	$t+4$
Simulation targets									
Finalized data	Y_{t-4}	Y_{t-3}	Y_{t-2}	Y_{t-1}	Y_t	Y_{t+1}	Y_{t+2}	Y_{t+3}	Y_{t+4}
Issue t	$Y_{t-4}^{(t)}$	$Y_{t-3}^{(t)}$	$Y_{t-2}^{(t)}$	$Y_{t-1}^{(t)}$	$Y_t^{(t)}$	Latest available report; to be updated each issue from $(t+1)$ onward			
Issue $t-1$	$Y_{t-4}^{(t-1)}$	$Y_{t-3}^{(t-1)}$	$Y_{t-2}^{(t-1)}$	$Y_{t-1}^{(t-1)}$					
Issue $t-2$	$Y_{t-4}^{(t-2)}$	$Y_{t-3}^{(t-2)}$	$Y_{t-2}^{(t-2)}$						
Issue $t-3$	$Y_{t-4}^{(t-3)}$	$Y_{t-3}^{(t-3)}$							
Issue $t-4$	$Y_{t-4}^{(t-4)}$								

Figure 3.1: **Initial surveillance data values and subsequent revisions form a "provisional data triangle".** The top shaded region of this diagram represents some of the backcasting and forecasting forecasting targets, $Y_{1..T_2}$: the finalized values of the surveillance data after many revisions. (The exact versions to be considered "finalized" may be explicitly specified by stakeholders soliciting forecasts.) The second shaded region corresponds to the most recent surveillance report or "issue", $Y_{1..t}^{(t)}$, containing (a) $Y_{1..t-1}^{(t)}$, revisions of values from the previous report, and (b) $Y_t^{(t)}$, the initial estimate of Y_t. The lines below the second shaded region correspond to a few older reports, $Y_{1..t-4}^{(t-4)}..Y_{1..t-1}^{(t-1)}$; moving from older reports below to newer reports above, each report contains an observation for one additional time interval and revisions for the rest, giving rise to the triangular arrangement of the diagram.

- Let $Y^{\text{sim}}_{1..T_1} = Y_{1..T_1}$
- Draw $Y^{\text{sim}}_{T_1+1} \sim Y_{T_1+1} \mid Y^{(1..t)}, Y_{1..T_1} = Y^{\text{sim}}_{1..T_1}$
- Draw $Y^{\text{sim}}_{T_1+2} \sim Y_{T_1+2} \mid Y^{(1..t)}, Y_{1..T_1+1} = Y^{\text{sim}}_{1..T_1+1}$
- Draw $Y^{\text{sim}}_{T_1+3} \sim Y_{T_1+3} \mid Y^{(1..t)}, Y_{1..T_1+2} = Y^{\text{sim}}_{1..T_1+2}$
- ...
- Draw $Y^{\text{sim}}_{T_2} \sim Y_{T_2} \mid Y^{(1..t)}, Y_{1..T_2-1} = Y^{\text{sim}}_{1..T_2-1}$

That is, we simulate the first latent observation Y_{T_1+1}, then feed that simulated value $Y^{\text{sim}}_{T_1+1}$ into a model for Y_{T_1+2}, then feed the resulting value $Y^{\text{sim}}_{T_1+2}$ along with $Y^{\text{sim}}_{T_1+1}$ into a model for Y_{T_1+3}, and so on. The model selected for $Y_u \mid Y^{(1..t)}, Y_{1..u-1}$ is once again arbitrary, but it is often convenient to consider direct models of $\Psi^{[u]} \mid \mathbf{\Phi}^{[u]}$, where $\Psi^{[u]}$ can now depend on $Y^{(1..t)}, Y_{1..u}$ such that Y_u is recoverable, and $\mathbf{\Phi}^{[u]}$ is a feature vector prepared from $Y^{(1..t)}, Y_{1..u-1}$. For example, a natural choice for $\mathbf{\Phi}^{[u]}$ is a small selection of provisional data for times intervals u and nearby time intervals. One drawback of this choice paired with the procedure above is that it leads to algorithms akin to Kalman filtering or fixed-lag smoothing rather than comprehensive Kalman smoothing; that is, the u-th simulated value, Y^{sim}_u, will either ignore available data for time intervals after u ("filtering"), or will ignore available data for time intervals after $u + k$ for some fixed $k > 0$ ("fixed-lag smoothing"), rather than considering data from all time intervals (comprehensive smoothing). For lower-noise data sets, these omissions may not be too harmful, as enough signal is already present in nearby (and far past) time intervals to backcast accurately. The kernel delta density and locally linear quantile autoregression approaches have analogues in this setting: kernel residual density and quantile ARX (autoregression with additional covariates treated as exogenous):

- **Kernel residual density:** uses kernel smoothing methods to estimate the conditional distribution (a) of residuals $Y_u - \hat{Y}_u$ given some covariates and an initial estimate $\hat{Y}_u$ when $u \le t$, and (b) of deltas $Y_u - Y_{u-1}$ when $u > t$.
- **Quantile ARX:** uses quantile regression to estimate the conditional distribution of Y_u given a selection of features from $Y_{1..u-1}$ and $Y^{(1..t)}$.

3.3.1 Kernel residual density

This subsection reproduces or incorporates content from Brooks et al. [2018].

The kernel residual density method chains together draws from conditional density estimates of $(Y_u - \hat{Y}_u) \mid \mathbf{\Phi}^{[u]}$ for u from $T_1 + 1$ to t and of $\Delta Y_u \mid \mathbf{\Phi}^{[u]}$ for u from $t + 1$ to T_2, where $\mathbf{\Phi}^{[u]}$ is a function of $Y_{1..u-1}$ and $Y^{(1..t)}$ and $\hat{Y}_u$ is some initial estimate of Y_u:

- The data revision-ignorant delta density method can be seen as a special case where $\hat{Y}_{T_1+1..t} = Y_{T_1+1..t}$ (i.e., past values $Y_{T_1+1..t}$ are all treated as known and are simply duplicated in the simulated trajectories) and $\hat{Y}_{t+1..T_2} = Y_{t..T_2-1}$ (i.e., the estimator $\hat{Y}_u$ when $u \geq t + 1$ is the lagged value Y_{u-1}, available in simulated trajectories from previous simulation steps). Each later residual $Y_u - Y_{u-1}$ corresponds to a delta in the delta density approach, ΔY_u.
- A data revision-aware variant is obtained by using $\hat{Y}_{T_1+1..t} = Y^{(t)}_{T_1+1..t}$ (i.e., the latest provisional value for every time interval for which provisional data is available) while keeping $\hat{Y}_{t+1..T_2} = Y_{t..T_2-1}$.

The performance of these kernel residual density approaches is studied alongside additional variants in Section 4.2 (the revision-unaware approach corresponding to "Ground truth, no nowcast" and "Real-time data, no nowcast", and the revision-aware approach above corresponding to "Backcast, no nowcast").

3.3.2 Quantile ARX

Another candidate is a regularized locally linear quantile regression on a subset of the conditioning covariates. One option is to simulate quantiles of Y_u as a linear function of the following covariates, along with a data weighting kernel and optional extra post-processing noise: Figure 3.2 visualizes this availability-dependent selection with a template for a roughly corresponding Bayes net. This Bayes net description is inexact, as the sampling procedure described in the introduction to this chapter models the dependence of each node on its ancestors, but not its descendants. Usually, we will start simulating with u's where most of this data is available, but at higher u some of the covariates will be excluded due to unavailability. For example, when simulating Y_{t+1}, the above covariate set would incorporate only $Y^{(t)}_t$ and $Y^{\text{sim}}_{t-3..t}$. Training instances for the quantile regression model map these test covariates to the following training covariates:

Name	Type	Description	Notation
Stable@u-4	Input	Stable/simulated value 4 weeks before	$Y_{u-4}/Y_{u-4}^{\text{sim}}$
Stable@u-3	Input	Stable/simulated value 3 weeks before	$Y_{u-3}/Y_{u-3}^{\text{sim}}$
Stable@u-2	Input	Stable/simulated value 2 weeks before	$Y_{u-2}/Y_{u-2}^{\text{sim}}$
Stable@u-1	Input	Stable/simulated value 1 weeks before	$Y_{u-1}/Y_{u-1}^{\text{sim}}$
Latest@u-1	Input	Latest value for 1 week before	$Y_{u-1}^{(t)}$
Latest@u	Input	Latest value for given week	$Y_{u}^{(t)}$
Latest@u+1	Input	Latest value for 1 week after	$Y_{u+1}^{(t)}$
Second-Latest@u	Input	Second-latest value for given week	$Y_{u}^{(t-1)}$
Stable@u	Output	Stable/simulated value for given week	Y_u/Y_u^{sim}

Table 3.1: One potential choice of $\mathbf{\Phi}^{[u],\text{QARXlinear}}$ and $\Psi^{[u]}$.

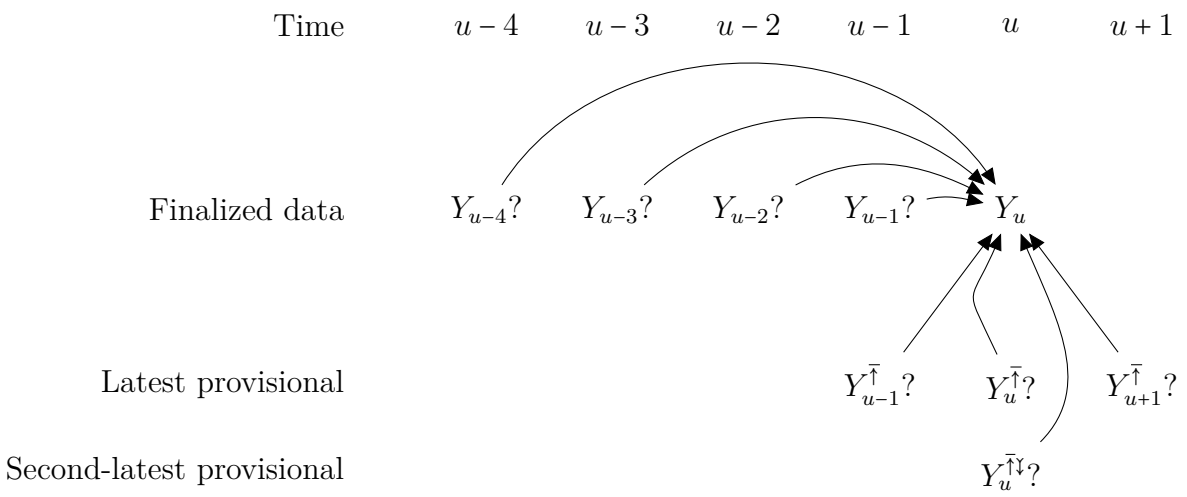

Figure 3.2: Bayes net template related to earlier covariate table. Here, u could refer to a past, present, or future week, not just the current week. Question marks denote covariates that are included if available (observed/simulated) at test/application time. The $\bar{\uparrow}$ symbol refers to the latest version of a wILI measurement available at test time (if there are any versions available), while $\bar{\uparrow}\downarrow$ refers to the second-latest version of a wILI measurement available at test time (if there are ≥ 2 versions available).

- $Y^{(t+\Delta t)}_{(u-1..u+1)+\Delta t}$ corresponding to available $Y^{(t)}_{u-1..u+1}$
- $Y^{(t-1+\Delta t)}_{u+\Delta t}$ corresponding to available $Y^{(t-1)}_{u}$
- $Y_{(u-4..u-1)+\Delta t}$, corresponding to available $Y^{\text{sim}}_{u-4..u-1}$ or $Y_{u-4..u-1}$

The training set is limited to those instances where all of the above covariates are available. Weights can be assigned to training instances to encourage use of data from similar times of year and similar values of the covariates. Regularization is incorporated to prevent overfitting and remain robust in the face of collinearities. (Collinearities can arise, e.g., when the training set used fills in holes in records for $Y^{(1..t)}$ with other values from $Y^{(1..T)}$ or $Y_{1..T}$.) The performance of this approach is studied in Subsection 4.3.1 (approaches "T:B" and "T:BNF").

Consider a more restrictive set of covariates: $Y^{(t)}_u$, if available, and Y^{sim}_{u-1} (or Y_{u-1} if available). Then the above process draws from conditional distributions that resemble a state space filter; for example, $Y_{T_1+1} \mid Y^{(t)}_{T_1+1}, Y_{T_1}$, using natural Markov assumptions, would be equivalent to $Y_{T_1+1} \mid Y^{(t)}_{T_1+1}, Y_{1..T_1}$, but would not consider information from observations for subsequent epiweeks such as $Y^{(t)}_{T_1+1..T_2}$. Since dependencies between data updates to observations for nearby weeks, we may want to ensure that this information is included. One simple way is to simply add more elements from $Y^{(t)}_{T_1+1..T_2}$ as covariates when available, but this might lead to issues with fitting too many parameters, e.g., at $T_1 + 1$. An alternative would be to add a backward pass that parallels a state space smoothing algorithm; this approach may not be feasible when using complicated transformations or data weights. Yet another path is to add subsequent values such as Y_{u+1} to the conditioning covariates and perform fitting and sampling using algorithms for the Multiple Quantile Graphical Model [Ali et al., 2016].

Chapter 4

Incorporating additional surveillance sources into pancasters

The previous two chapters discuss models for a single source of surveillance data that reports (multiple versions of) a single measurement for each time in the past for a particular location. This approach forgoes useful information available from additional traditional surveillance sources and a number of novel digital surveillance sources such as search query volume, social media activity, page hits, illness self-reporting, internet-integrated monitoring and testing devices, electronic health records, and insurance claims. We can generalize the above approaches to forecast multiple data sources and/or locations at once, incorporating information from multiple auxiliary data streams. This chapter describes a joint modeling and simulation approach that incorporates dependencies across sources using a domain-informed dependence graph. This task is often referred to as "nowcasting" or "nearcasting" when "predicting" ground truth data for times u at or around the time that the predictions are made. Even the most quickly released traditional surveillance data are not as timely as novel digital sources, so this typically entails estimating stable ground truth data Y_u given lower-latency external data X_u for the same time period and provisional data $Y_{1..t}^{(t)}$, $t < u$, which does not contain an estimate for time interval u. Performing joint distribution prediction for $Y_{T_1..T_2}$ covering times before, near, and after t combines the tasks of backcasting, nowcasting, and forecasting, and we will henceforth refer to it as "pancasting".

Section 4.1 discusses the task of nowcasting, some available nowcasting systems for wILI, and the general tactic that will be taken to incorporate these in distributional trajectory predictions. Section 4.2 pairs this tactic with the kernel residual density backcasting approach and various forecasters. Section 4.3 applies the same tactic in the quantile ARX modeling framework, with the option of performing the entire pancast within the quantile ARX model, which appears competitive with any of the piecemeal approaches combining separate backcasters, nowcasters, and forecasters.

4.1 Latency of initial wILI values and "nowcasting"

This section reproduces or incorporates content from Brooks et al. [2018].

The initial ILINet wILI value for a given "target" week (from Sunday to Saturday) is typically released on Friday of the following week. Data sources with lower latency and higher temporal resolution can be used to prepare wILI estimates ("nowcasts") earlier in the following week or even during the target week itself. More generally, auxiliary data for past and current weeks can improve not only models of disease activity in these weeks but also forecasts of future disease activity. Given a backcaster that simulates finalized data for past weeks $Y_{1..t}$ given observed ILINet and auxiliary data, a nowcaster that simulates Y_{t+1} given these observations and (a simulated) $Y_{1..t}$, and a forecaster that simulates $Y_{t+2..T}$ given these observations and (a simulated) $Y_{1..t+1}$, we can sample from an enhanced model of $Y_{1..T}$ (given the latest wILI observations $Y^t_{1..t}$, previous versions of wILI, and auxiliary data) using the following procedure:

1. Repeatedly draw a random value $Y^{\text{sim}}_{1..T}$ for $Y_{1..T}$ by:
   (a) drawing a random value $Y^{\text{sim}}_{1..t}$ for $Y_{1..t}$ conditioned on the observed data, using the backcaster, then
   (b) drawing a random value Y^{sim}_{t+1} for Y_{t+1} conditioned on the observed data and $Y_{1..t} = Y^{\text{sim}}_{1..t}$, using the nowcaster, then
   (c) drawing a random value $Y^{\text{sim}}_{t+2..T}$ for $Y_{t+2..T}$ conditioned on the observed data and $Y_{1..t+1} = Y^{\text{sim}}_{1..t+1}$, using the forecaster, then

(d) combining $Y^{\text{sim}}_{1..t}$, Y^{sim}_{t+1}, and $Y^{\text{sim}}_{t+2..T}$ into a single (random) trajectory $Y^{\text{sim}}_{1..T}$, and

2. Collect these individual, randomly drawn trajectories into a list (i.e., a random sample).

As with the earlier method of combining backcasts and forecasts without a nowcaster, this procedure may be too computationally expensive for some implementations of some forecasters; we use these steps exactly with the delta density methods, for example, but consider modifications and approximations for some other forecasters.

This methodology can be applied in conjunction with one of many available nowcasters. We focus on ILI-Nearby [Farrow, 2016, Farrow et al., 2019], which produces nowcasts for wILI by fusing together several "sensors" using another type of stacked generalization, where each sensor is also a nowcast of wILI data; we reproduce a list of references from [Farrow, 2016] on other methodologies for nowcasting and incorporating auxiliary data here [Achrekar et al., 2011, Araz et al., 2014, Broniatowski et al., 2013, Culotta, 2010, Dredze et al., 2014, Dugas et al., 2013, Eysenbach, 2006, Generous et al., 2014, Ginsberg et al., 2009, Hickmann et al., 2015, Hulth et al., 2009, McIver and Brownstein, 2014, Paul et al., 2014b, Polgreen et al., 2008, Preis and Moat, 2014, Ritterman et al., 2009, Santillana et al., 2014, 2015, Shaman and Karspeck, 2012b, Shaman et al., 2013b, Signorini et al., 2011, Soebiyanto et al., 2010, Yang et al., 2015] along with some more recent work [Johansson et al., 2016, Lampos et al., 2015, Yang et al., 2017], with special note of other work using multiple auxiliary data sources [Yang et al., 2017] or nowcasters [Santillana et al., 2015]. We consider four distributional nowcasters:

- Y^{sim}_{t+1} produced by the forecaster, i.e., not using separate nowcasts at all — the basis for all kernel residual density performance estimates unless otherwise noted, as no nowcasts were incorporated into the Delphi-Stat forecasts for the 2015/2016 season;

- Y^{sim}_{t+1} following a normal distribution with mean and standard deviation given by the ILI-Nearby nowcasting system (ignoring the backcaster's output);

- Y^{sim}_{t+1} following a Student's t distribution with two degrees of freedom, centrality parameter set to the ILI-Nearby point estimate, and scale parameter set to the ILI-Nearby standard deviation estimate, intended to be a wide-tailed variant of the above (ignoring the backcaster's output);

- an ensemble of first and third approaches, with associated weights (probabilities) of 15% and 85% respectively. (The choice of weights was inherited from a similar approach that mixed "1 wk ahead" delta density forecasts with nowcasts, rather than ensemble forecasts (including a uniform component) based on these two approaches; a nowcast weight of 85% was selected on a limited amount of out-of-sample (preseason) forecasts to maximize log score.).

Section 4.3 discusses a special case of the first approach where backcasts, nowcasts, and forecasts are all unified under the same modeling framework, which treats provisional ILINet data and ILI-Nearby as potentially missing inputs.

4.2 Backcasting and nowcasting wILI using kernel residual density and ILI-Nearby

This section reproduces or incorporates content from Brooks et al. [2018].

We first consider pancasting by stitching together the forecasts of separate backcasting, nowcasting, and forecasting systems. We estimate the distribution of backfill updates using the residual density method described in Subsection 3.3.1, with $t_1 = 0$, $t_2 = t$, $X_{1..t} = Y^t_{1..t}$ the latest version of wILI available, $Y_{1..t}$ the corresponding final revisions, and $\mathbf{\Phi}^u = [Y_{u-1}]$. The weight given to a historical nonfinal-to-final residual is based on three factors:

- **Lag amount:** later revisions of wILI values tend to be closer to the final revision than earlier revisions are; thus, when estimating the distribution of n-week-old wILI to finalized wILI residuals, only n-week-old wILI to finalized wILI data is considered; backfill data for other lags is ignored (i.e., has zero weight);
- **The current season's nonfinal wILI value:** historical backfill updates with nonfinal wILI values closer to the nonfinal wILI value from the current season are given greater weights according to a Gaussian kernel (with bandwidth based on a rule for kernel density estimation of the historical nonfinal wILI values);
- **Epi week of observation:** since the backfill pattern changes throughout a season, historical backfill updates corresponding to nearby epi weeks are weighted more highly than those from a different time of the season, using a Laplacian kernel (with an arbitrarily selected bandwidth).

The bandwidth of the density estimate is based on a kernel density estimate of the nonfinal-to-final residuals.

The backcasting method is modular and can combine with any forecaster expecting ground truth wILI as input. The straightforward approach is to sample a few hundred or thousand trajectories from the backfill simulator, feed each of these into the forecaster to obtain a trajectory or a distribution over targets, and aggregate the results. Some forecasting methods considered do not have a simple way to quickly generate single-trajectory forecasts, so we also use alternative approaches to reduce computation, such as randomly pairing backcasts and trajectory forecasts, where the trajectory forecasts are efficiently generated in batch, based on the pointwise mean of the backcasts. Figure 4.1 shows sample forecasts over wILI trajectories generated by each of these approaches and compares them to some alternatives described in Appendix B. Note that the pancast trajectory distributions express some uncertainty (vertical spread among black lines) even for time intervals where there is a provisional estimate (yellow line) available. Naturally, uncertainty is generally greater for later time periods; e.g., greater around the last time period for which a provisional data estimate is available (right tip of yellow line) than the thoroughly revised data from August (leftmost part of yellow line).

Figure 4.2 shows cross-validation performance estimates for the extended delta density method based on the following input data:

- **Ground truth, no nowcast:** the ground truth wILI for the left-out season up to the forecast week is provided as input, resulting in an optimistic performance estimate;
- **Real-time data, no nowcast:** the appropriate wILI report is used for data from the left-out season, but no adjustment is made for possible updates; this performance estimate is valid, but we can improve upon the underlying method;
- **Backcast, no nowcast:** the appropriate wILI report is used for data from the left-out season, but we use a residual density method to "backcast" updates to this report; this performance estimate is valid, and the backcasting procedure significantly improves the log score;
- **Backcast, Gaussian nowcast:** same as "Backcast, no nowcast" but with another week of simulated data added to the forecast, based on a Gaussian-distributed nowcast; and
- **Backcast, Student t nowcast:** same as "Backcast, Gaussian nowcast" but using a Student t-distributed nowcast in place of the Gaussian nowcast.

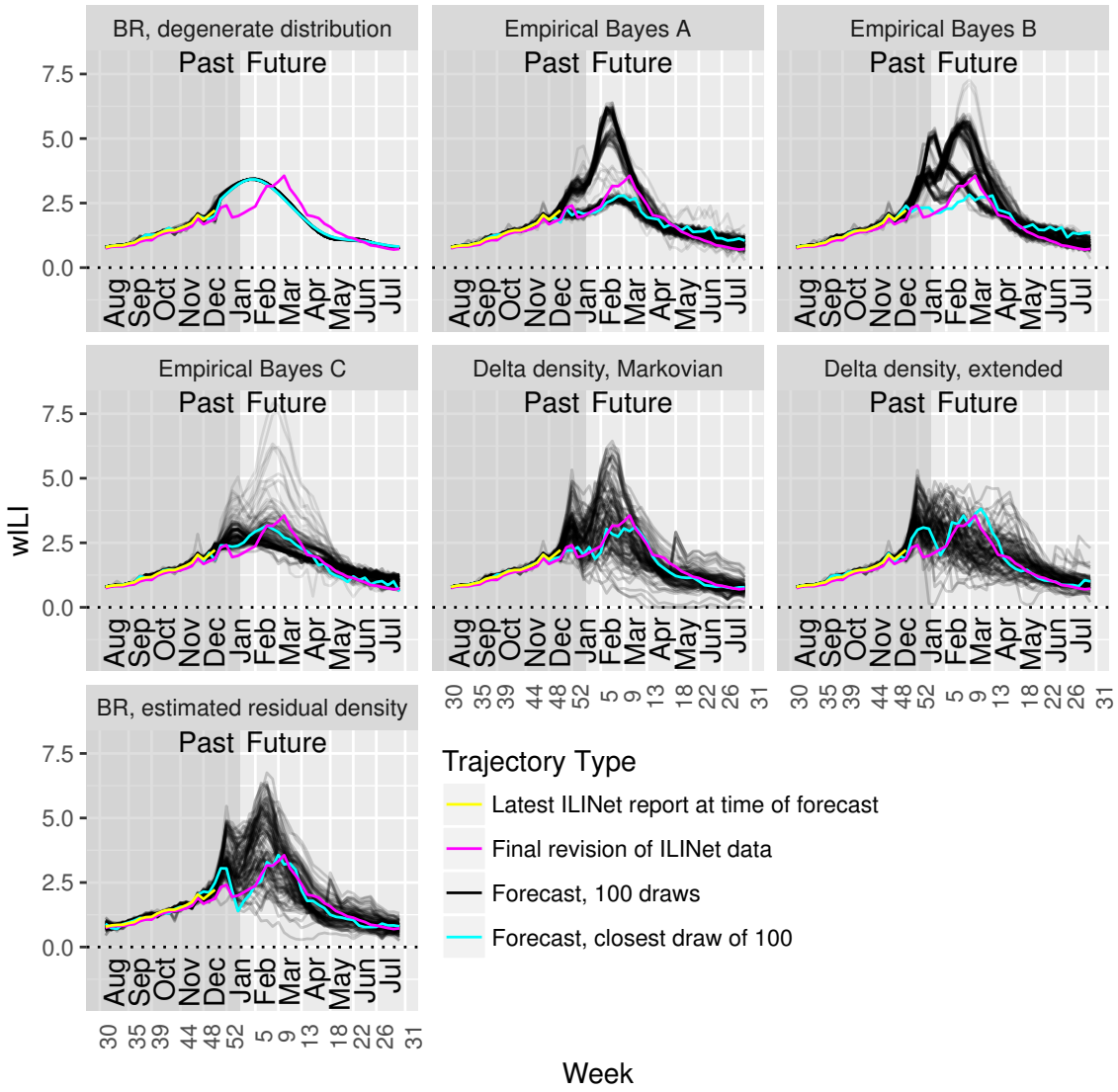

Figure 4.1: **Delta and residual density methods generate wider distributions over trajectories than methods that treat entire seasons as units.** These plots show sample forecasts of wILI trajectories generated from models that treat seasons as units (BR, Empirical Bayes) and from models incorporating delta and residual density methods. Yellow, the latest wILI report available for these forecasts; magenta, the ground truth wILI available at the beginning of the following season; black, a sample of 100 trajectories drawn from each model; cyan, the closest trajectory to the ground truth wILI from each sample of 100.

- **Backcast, ensemble nowcast:** same as the previous two but using the ensemble nowcast (which combines "no nowcast" with "Student t nowcast").

For every combination of target and forecast week, using ground truth as input rather than the appropriate version of these wILI observations produces either comparable or inflated performance estimates.

Using the "backcasting" method to model the difference between the ground truth and the available report helps close the gap between the update-ignorant method. The magnitude of the performance differences depends on the target and forecast week. Differences in mean scores for the short-term targets are small and may be reasonably explained by random chance alone; the largest potential difference appears to be an improvement in the "1 wk ahead" target by using backcasting. More significant differences appear in each of the seasonal targets following typical times for the corresponding onset or peak events; most of the improvement can be attributed to preventing the method from assigning inappropriately high probabilities (often 1) to events that look like they must or almost certainly will occur based on available wILI observations for past weeks, but which are ultimately not observed due to revisions of these observations. The magnitude of the mean log score improvement depends in part on the resolution of the log score bins; for example, wider bins for "Season peak percentage" may reduce the improvement in mean log score (but would also shrink the scale of all mean log scores). Similarly, the differences in scores may be reduced but not eliminated by use of multibin scores for evaluation or ensembles incorporating uniform components for forecasting.

Using the heavy-tailed Student t nowcasts or nowcast ensemble appears to improve on short-term forecasts without damaging performance on seasonal targets. The Gaussian nowcast has a similar effect as the other nowcasters except on the "1 wk ahead" target that it directly predicts: its distribution is too thin-tailed, resulting in lower mean log scores than using the forecaster by itself on this target.

4.3 Unified quantile ARX-based pancast filtering model

A limited number of additional data sources with sufficient temporal availability and matching resolution can be easily and directly added to the kernel residual density and quantile ARX models by treating these external data sources as exogenous covariates, plus a mechanism for handling missingness which takes into account the fact

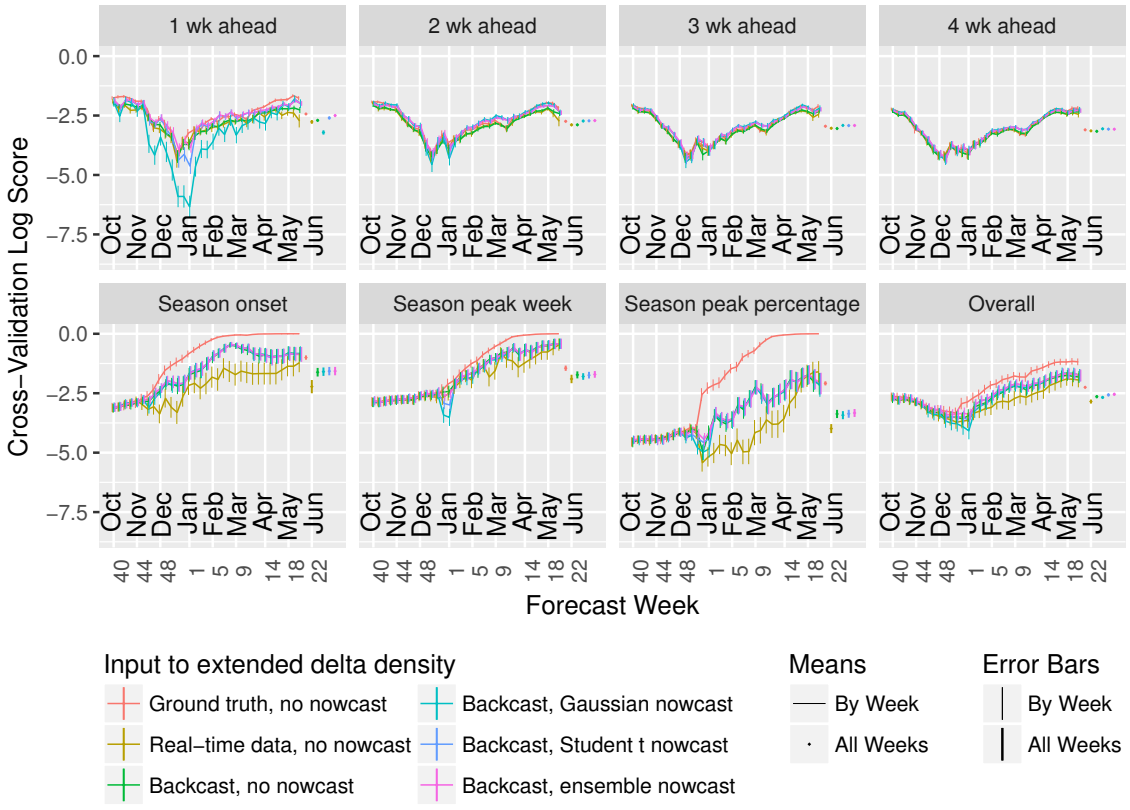

Figure 4.2: **Using finalized data for evaluation leads to optimistic estimates of performance, particularly for seasonal targets, "backcasting" improves predictions for seasonal targets, and nowcasting can improve predictions for short-term targets.** Mean log score of the extended delta density method, averaged across seasons 2010/2011 to 2015/2016, all locations, all targets, and forecast weeks 40 to 20, both broken down by target and averaged across all targets ("Overall"). Rough standard error bars for the mean score for each target (or overall) appear on the right, in addition to the error bars at each epi week.

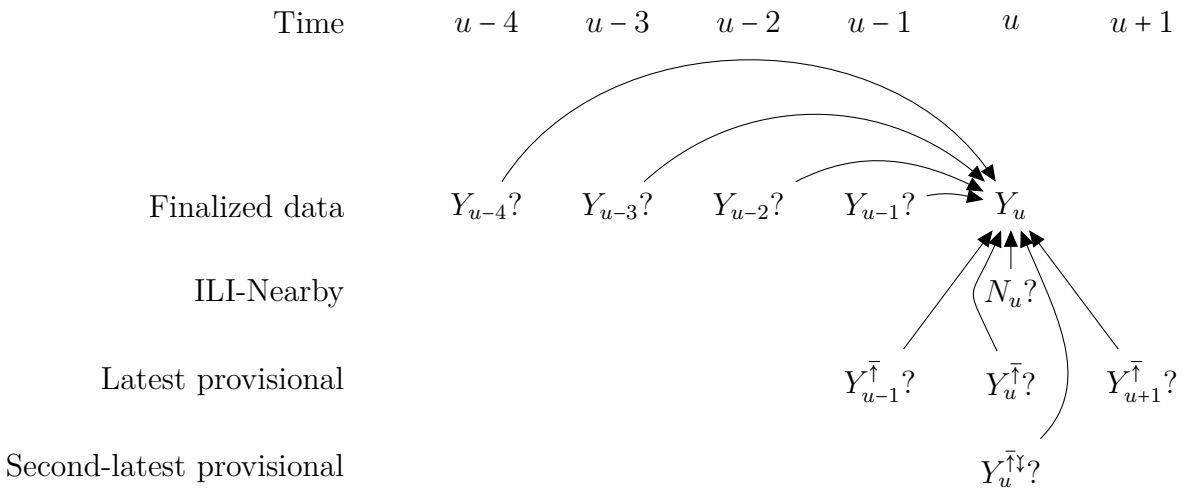

Figure 4.3: Bayes net template corresponding to earlier covariate table. Here, u could refer to a past, present, or future week, not just the current week. Question marks denote covariates that are included if available (observed/simulated) at test/application time. The $\bar{\uparrow}$ symbol refers to the latest version of a wILI measurement available at test time (if there are any versions available), while $\bar{\uparrow}\downarrow$ refers to the second-latest version of a wILI measurement available at test time (if there are ≥ 2 versions available).

that the data sources involved are streaming with different latencies. In the quantile autoregression framework, the ILI-Nearby data can be treated in the same way as a provisional data point: added to the selection of covariates used in the quantile ARX pancasting routine. Figure 4.3 visualizes this availability-dependent selection with a Bayes net for a single location.

4.3.1 Unified quantile autoregression pancast performance

Figure 4.5 compares the average unibin log score of various forecasting methods given preliminary data, partial pancasts, or full pancasts as input. Only the delta density methods leave nontrivial input from the pancaster untouched, with other forecasting methods ignoring pancasting input or performing alterations for computational tractability that hurt unibin log score: the Uniform and EmpiricalTrajectories baselines completely ignores pancaster input, BasisRegression resamples and EmpiricalFutures sparsely resamples the pancast (partial) trajectories, and EmpiricalBayes and EmpiricalBayes_Cond4 average over the pancast (partial) trajectories and treat the

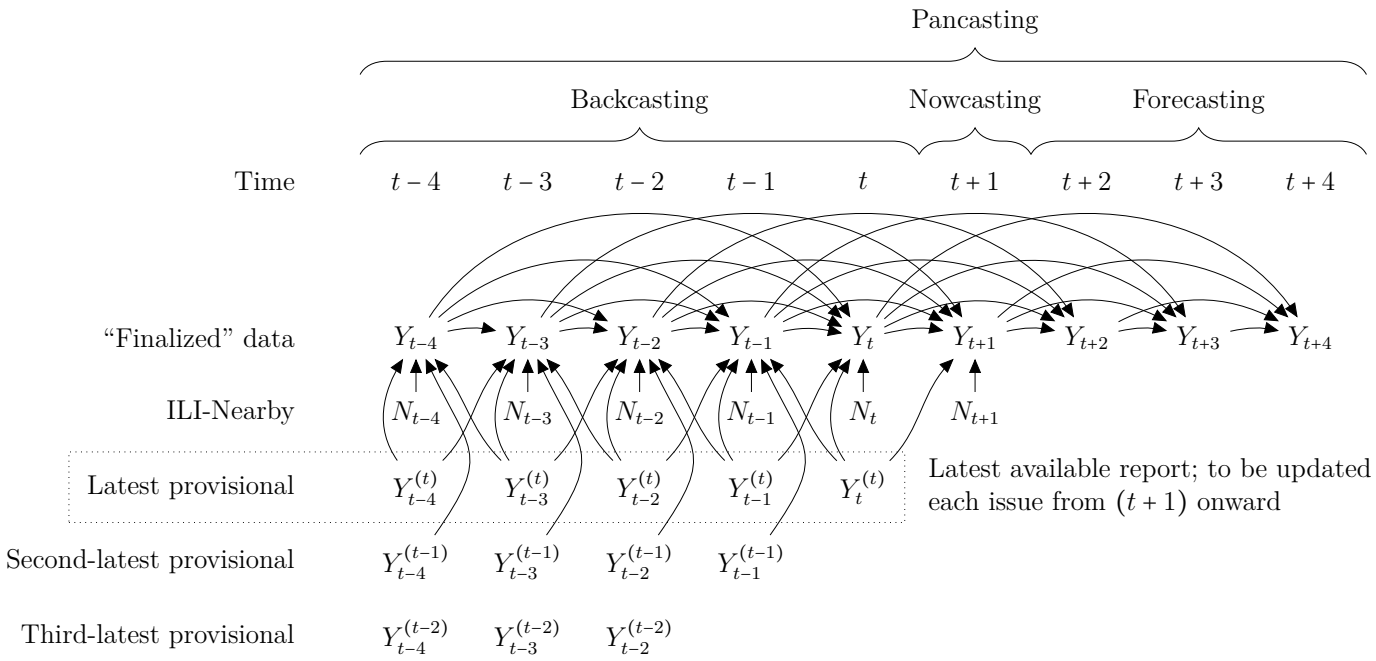

Figure 4.4: Expansion of the Bayes net template above for a short trajectory.

result as an observed (partial) trajectory with no uncertainty. The output of each pancaster-forecaster pair other than those for the Uniform baseline is a weighted sample of complete trajectories, which is transformed into a weighted sample over forecasting target values, which are in turn smoothed using uniform pseudocounts and kernel density estimation (differences between EmpiricalBayes and Empirical-Bayes_Cond4 given a full pancast as input are due differences in the amount of smoothing due to differing sample sizes output from these forecasters). We observe that:

- The pancaster-forecaster pairs with the highest overall scores are completely based on chained one-ahead models (quantile ARX pancasting or kernel delta density models for every observation) rather than entire-trajectory approaches.
- Differences in performance between different forecasting methodologies are more obvious than those between backcasting and nowcasting methodologies.
- Pairs that faithfully incorporate distributional backcasts have higher overall scores than those that do not; those that heavily resample or replace the backcast distribution with a singleton have lower overall unibin scores.

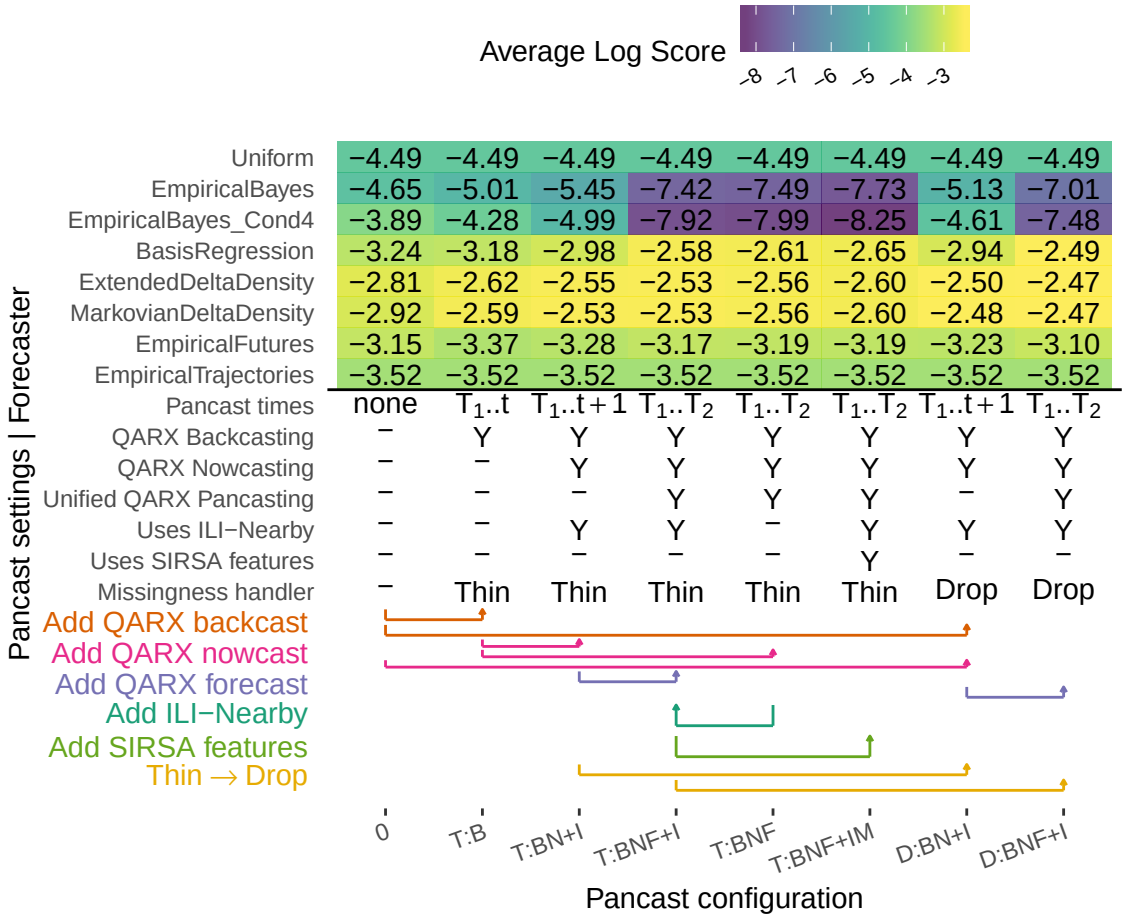

Figure 4.5: **Quantile ARX pancasting variants of the best forecasting approaches considered have higher cross-validation log scores for ILINet national and regional forecasts compared to forecasting-only variants.** This table shows the average cross validation unibin log score in the ILINet national and regional forecasting setting for various types of forecasters and pancasting configurations. The first column corresponds to scores of forecasting methods when they are provided preliminary data as if they were finalized. Subsequent columns correspond to sampling partial or full trajectories from a particular type of quantile ARX pancaster as input to the forecasting method. "Pancast times" and the next three rows both describe how much of the trajectory is modeled using the pancaster: none (revision-ignorant), times $T_1..t$ (backcasting), times $T_1..t+1$ (backcasting & nowcasting), or times $T_1..T_2$ (unified pancasting). "Uses ILI-Nearby" indicates whether auxiliary data from ILI-Nearby is used by the pancaster in addition to preliminary ILINet data. "Uses SIRSA features" indicates whether the pancaster incorporates SIRS-inspired covariates, without addressing issues with superexponential trajectory growth. "Missingness handler" is "Thin" when the pancaster uses covariate missingness indicator variables and a thin SVD approach to avoid issues with (near-)singular training covariate matrices, and is "Drop" when using an ad-hoc method to select a subset of covariates and training instances without missingness or singularity, incorporating lasso regularization when near-singularity does not interfere with the fitting routine.

- Pairs that incorporate distributional nowcasts faithfully have higher overall scores than those that do not.
- Pairs that incorporate auxiliary data from ILI-Nearby have comparable or higher overall scores than comparable pairs excluding them.
- Pairs that incorporate SIRS-inspired covariates have lower overall scores than comparable pairs excluding them; this suboptimal performance is due at least in part to superexponential growth in some simulated trajectories which may be avoidable with better fitting and/or simulation approaches.
- Pairs that use the "Drop" missingness handler in the pancast stage have higher overall scores than comparable pairs with the "Thin" missingness handler. (However, the "Drop" missingness handler does not avoid all (near-)singular matrix issues in the fitting routine used for other pancasting configurations tested, causing analysis for these configurations to fail and be excluded; the current "Thin" implementation is more robust in an operational sense.)

Figure 4.6 shows a cross-validation overall *multibin* log score analysis with similar overall conclusions, except that some methods that incorporate distributional backcasts and nowcasts "unfaithfully" can exhibit potential increases in overall scores rather than drops. Figure 4.7 and Figure 4.8 show cross validations score breakdowns by target and by location for national and regional ILINet forecasting for the same set of pancasting configurations combined with the ExtendedDeltaDensity forecaster; the best performing configurations from the overall scores have consistent top-ranking performance across all targets and locations.

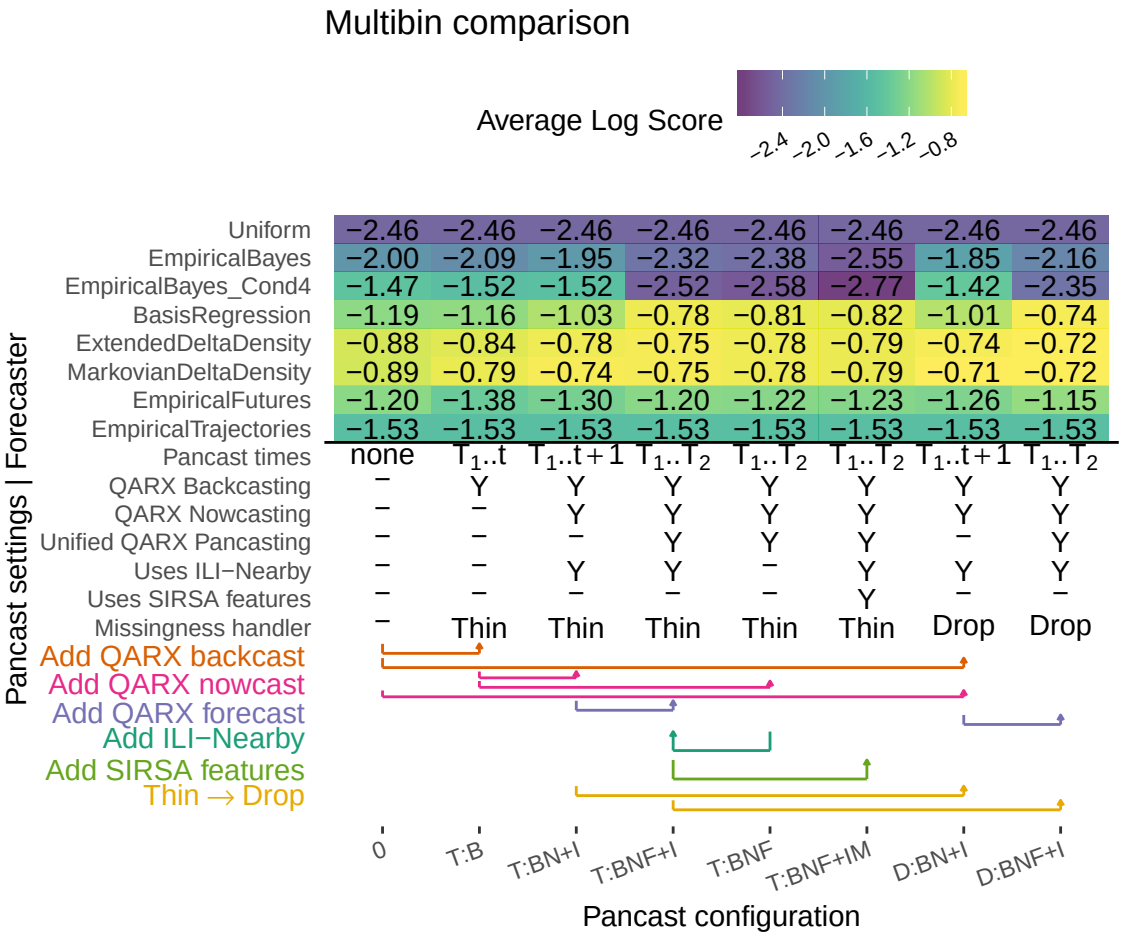

Figure 4.6: **Cross-validation overall multibin log scores for national and regional ILINet forecasts.**

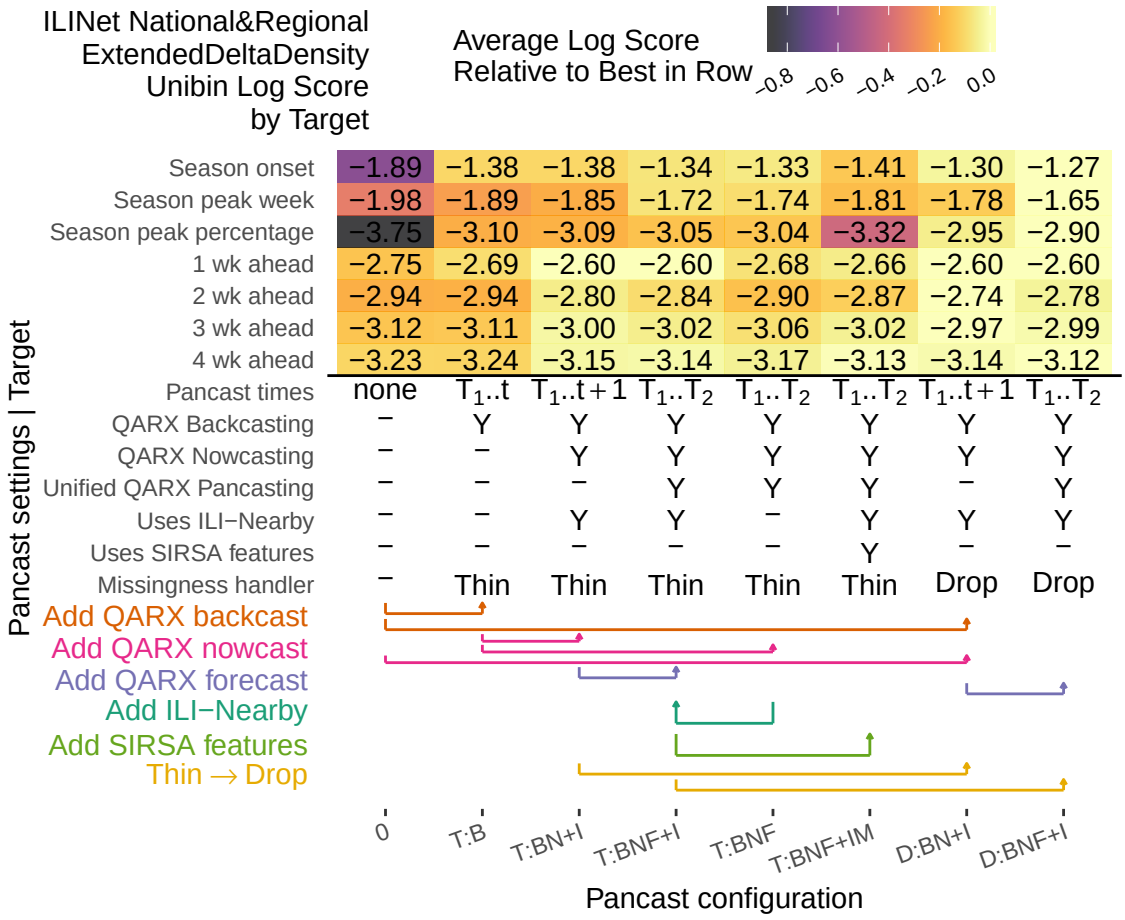

Figure 4.7: **Incorporating backcasts and nowcasts into ILINet forecasts from the extended delta density method yields higher cross-validation log scores for all targets; the unified pancaster has mixed results across targets relative to the two-step backcast&nowcast-forecast approach.**

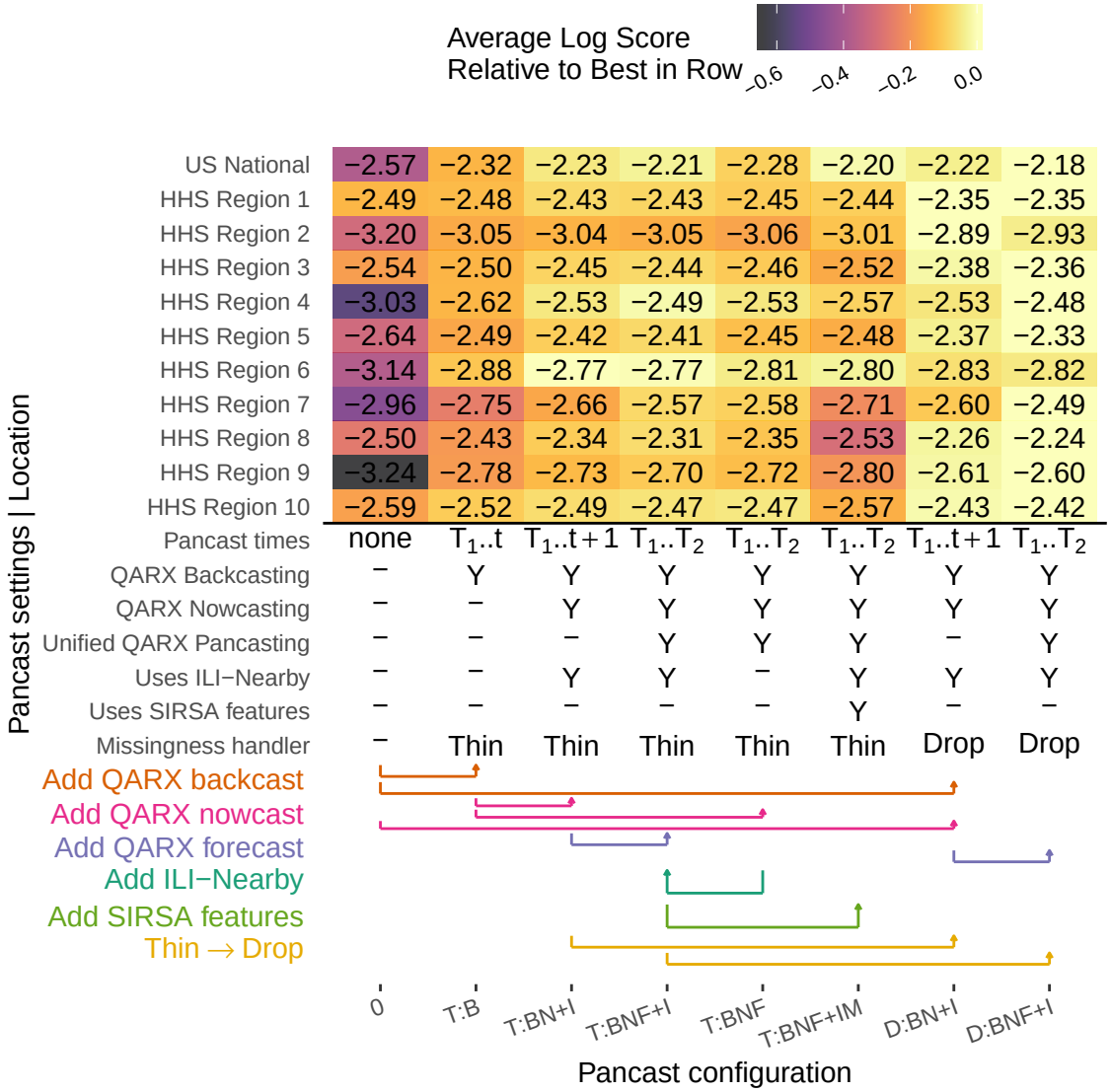

Figure 4.8: **Incorporating backcasts and nowcasts into ILINet forecasts from the extended delta density method yields higher cross-validation log scores for all locations, and the unified pancaster has similar or still better average scores.** Different locations have different %wILI scales and exhibit different revision patterns. Estimated performance gains from backcasting&nowcasting are higher in those particularly large (in absolute %wILI terms) revisions on average, but not all high performance gains correspond to locations with large revisions on average; see Figure 4.9.

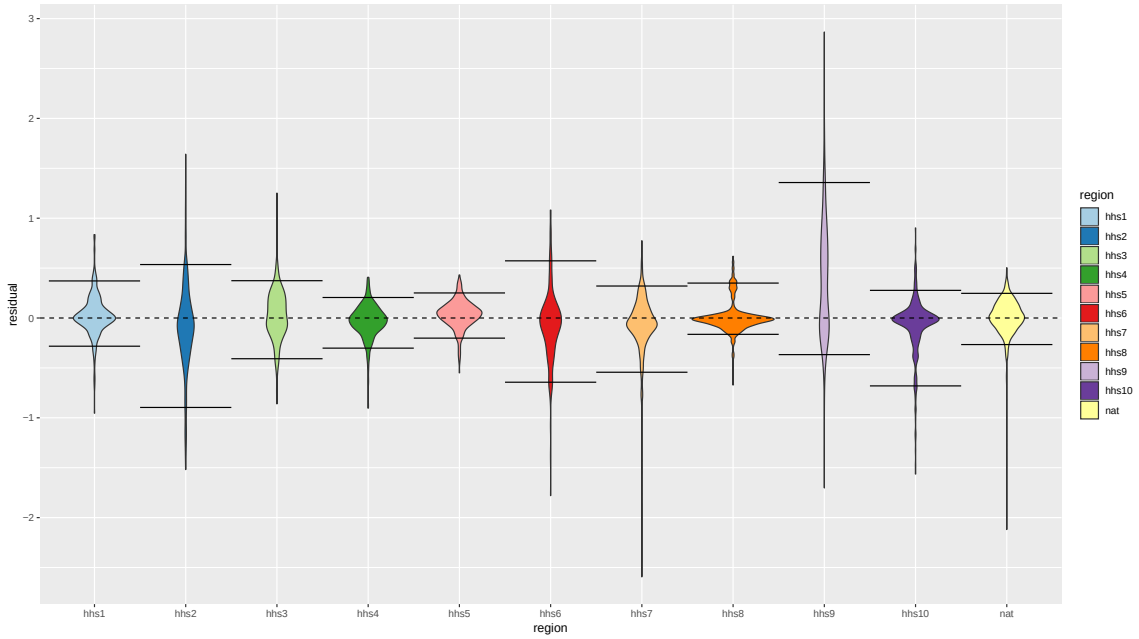

Figure 4.9: **ILINet absolute-scale revision distributions by location.** Violins represent density estimates, while horizontal lines mark the 5th and 95th empirical percentiles. The "residual" here is $Y_t - Y_t^{(t)}$ (the negation of the error of the first available provisional estimate for each time period), and is not normalized to account for differences in individual or average values of Y_t across locations; these absolute residuals may be indicative of possible impacts on the average log scores for forecasters, which are measured based on an absolute scale, but may not give the best picture of revision patterns themselves. Differences in scale and revision patterns may be explained by differences in the mixes of types, behavior, technology, and communication protocols of the healthcare providers participating in ILINet in different location.

Chapter 5

Combining multiple methods: stacking approach to model averaging

Much of this chapter is based on material from Brooks et al. [2018].

Forecasting systems that select effective combinations of predictions from multiple models can improve on the performance of the individual components, as demonstrated by their successful application in many domains. For each probability distribution and point prediction in a forecast, we treat the choice of an effective combination as a statistical estimation problem, and base each decision on the models' behavior in leave-one-out cross-validation forecasts. Additional cross-validation analysis indicates that this approach achieves performance comparable to or better than the best individual component.

Section 5.1 motivates the use of ensemble in this problem domain. Section 5.2 describes a "stacking" approach to combining distributional predictions into an ensemble leveraging historical or retrospective predictions generated from each of the ensemble's component models. Section 5.3 studies the performance of this stacking approach.

5.1 Background, motivation for combining forecasts

Methods that combine the output of different models, called "ensembles", "multi-model ensembles", "super-ensembles", "model averages", or various other names based on the domain and type of approach, have been applied successfully in many problem settings, improving upon the results of the best individual model. An ensemble approach is motivated in the context of seasonal epidemic forecasting by factors such as:

- **Model misspecification and overconfidence in distributional forecasts:** Many methods overlook the possibility of a significant proportion of observed outcomes, or assign otherwise inappropriate probabilities. These omissions and other mistakes are not identical across models; the gaps left by one component can be filled in by another.
- **Leveraging partially correlated errors in point predictions:** The point prediction errors of individual methods can vary in magnitude and are often only partially correlated with each other, allowing ensemble methods to improve performance, e.g., by highly weighting more accurate predictors, or by reducing the variance when combining multiple unbiased estimators.
- **Strengths and weaknesses in different targets:** Some methods may work well for certain forecasting targets, but have poor performance or fail to produce predictions for others; model averages can be smoothly adjusted to account for different behaviors for different targets.
- **Changes in performance within seasons:** Making predictions at the beginning, middle, and end of a season can be seen as different tasks, and the relative performance characteristics of the components may change based on the time of season (or whether it is around a holiday). Just as ensemble methods can account for distinct patterns based on forecasting target, they can be tailored to account for changes in behavior within a season.

We developed an adaptively weighted model average that consistently outperforms the best individual component. Other teams submitting forecasts to the FluSight comparison have concurrently developed other data-driven ensemble systems and found similar success Ray and Reich [2017], Yamana et al. [2017]; less data-driven ensemble techniques applied to forecasts of multiple research groups also exhibit

strong performance, including a wisdom-of-crowds Morgan et al. [2018] and simple average approaches McGowan et al. [2019]. Our ensemble framework is distinguished from these other methods in that it very directly estimates the best model average weights for a given location, time, target, and evaluation metric. This framework has been reapplied with slightly different settings by the FluSight Network, which uses historical and ongoing component forecasts from multiple research groups Reich et al. [2019a] to form a highly-ranked ensemble forecast Reich et al. [2019c].

5.2 A stacking approach to model averaging

For each location l, week t, target i, and evaluation metric e, we choose a (weighted) model average as the final prediction: an ensemble forecast of the form $\mathbf{Xw}$, where

- $\mathbf{X}$ is the output of the m ensemble components — either (a) a row vector of point predictions with m entries, or (b) a matrix of distributional predictions with m columns — and
- $\mathbf{w} \in [0,1]^m$ is a (column) vector of weights, one per component, with $\sum_{j=1}^{m} w_j = 1$.

Variants of the same models — or methods based on related approaches or assumptions — may at times produce similar forecasts that commit the same errors, while producing a misleading impression of consensus; a successful ensemble may need to consider not only the performance of each individual component, but also the relationships between the raw output of the components. To this end, we use a "stacking generalization" approach Breiman [1996], Wolpert [1992], treating the selection of weights $\mathbf{w}$ for the current season, $S+1$, as the task of frequentist estimation of the risk-optimal weight vector,

$$\mathbf{w}^* = \underset{\substack{\mathbf{w}\in[0,1]^m \\ \sum_{j=1}^{m} \mathbf{w}_j = 1}}{\arg\max}\, \mathbb{E}\left[\text{Score}(\mathbf{w}, S+1, l, t, i, e)\right],$$

based on leave-one-season-out cross-validation:

$$\hat{\mathbf{w}} = \mu \mathbf{e}_{\text{uniform}} + (1-\mu) \underset{\substack{\mathbf{w}\in[0,1]^m \\ \sum_{j=1}^{m} \mathbf{w}_j = 1}}{\arg\max} \sum_{\substack{s'\in\{1..S\}, \\ l',t',i',e'}} \begin{array}{l}\text{RelevanceWeight}(s', l', t', i', e'; S+1, l, t, i, e) \\ \cdot \text{CrossValidationScore}(\mathbf{w}, s', l', t', i', e'),\end{array}$$

where μ is an inflation factor that gives addition weight to the uniform component ($\mathbf{e}_{\text{uniform}}$ is a vector containing a 1 in the position corresponding to the uniform distribution component, and 0 in every other position). We changed the RelevanceWeight

function used for real-time forecasts throughout the 2015/2016 season, but study only the following RelevanceWeight function in the cross-validation analysis of the adaptively weighted ensemble:

$$\text{RelevanceWeight}(s, l, t, i, e; s', l', t', i', e') = \begin{cases} 1, & |t - t'| \leq 4,\ i = i',\ e = e' \\ 0, & \text{otherwise.} \end{cases}$$

A larger collection of cross-validation data can be considered by assigning relevance weights of 1 to additional training instances; relevance weights can also be gradually decreased for less similar data rather than jumping down to zero.

When e is the unibin or multibin log score:

- Using the rule of three Jovanovic and Levy [1997] to estimate the frequency of events that we haven't seen before, we chose $\mu = \frac{3}{S \cdot L}$ for most submissions. (Prior to the submission for 2015 EW43, we used a constant $\mu = 0.01$ to guarantee a certain minimum log score.)
- The optimization problem is equivalent to fitting a mixture of distributions, and we can use the degenerate EM algorithm Rosenfeld [Accessed 2017-03-21] to efficiently find the weights; convex optimization techniques such as the logarithmic barrier method are also appropriate.

When e is mean absolute error:

- We choose $\mu = 0$ (and further, exclude the uniform distribution method from the ensemble entirely).
- This optimization problem is referred to as least absolute deviation regression or median regression, with linear inequality and equality constraints on the coefficients; we reformulate the problem as a linear program and use the `lpSolve` package Berkelaar and others [2015] to find a solution.

We compare the "adaptive" weighting scheme above to two alternatives:

- **Fixed-weightset-based stacking:** the same approach as above, with the same μ selections but a different RelevanceWeight function:

  $$\text{RelevanceWeight}(s, l, t, i, e; s', l', t', i', e') = \begin{cases} 1, & e = e' \\ 0, & \text{otherwise;} \end{cases}$$

  and

- **Equal weights:** does not use the above stacking scheme; instead, for every prediction, assigns each component the same weight in the ensemble, $\frac{1}{m}$ (replacing $\hat{\mathbf{w}}$ with $\frac{1}{m}\mathbf{1}$).

The ensemble and each of its components forecast the targets $\mathbf{Z}^{(t)}$ given a point or distributional estimate for $Y_{1..T}$. The fixed-weightset-based stacking matches the real-time approach in Yamana et al. [2017], while the adaptive approach considered offers a way to condition on "time of season" that is possible in real time. The adaptive approach considered is less flexible and likely suffers more from the curse of dimensionality more than the gradient tree boosting approach in Ray and Reich [2017], but has more easily inspected weight selections and does not require a softmax transformation.

5.3 Ensemble performance

5.3.1 Cross-validation of ensemble and its components

ILINet forecasting, U.S. national and HHS regional geographies

Figure 5.1 shows the distribution of cross-validation log scores for several forecasting methods, described earlier in the text and in Appendix B, and the three ensemble approaches specified earlier in the text, in the context of the FluSight forecasting comparison on U.S. national and HHS regional ILINet data. Except for the uniform distribution and ensembles, all forecasting methods miss some possibilities completely, reporting unreasonable probabilities less than $\exp(-10) \approx 0.0000454$ for events that actually occurred. In these situations, the log score has been increased to the cap of −10 (as CDC does for multibin log scores, although multibin scores are in general less negative already). Delta and residual density forecasting methods (Delta density, Markovian; Delta density, extended; and BR, residual density) are less likely to commit these errors than other non-ensemble, non-uniform approaches, and have higher average log scores. Ensemble approaches combine forecasts of multiple components, missing fewer possibilities, and ensuring that a reasonable log score is obtained by incorporating the uniform distribution as a component. For the full Delphi-Stat ensemble, the main advantage of the ensemble over its best component appears to be successfully filling in possibilities missed by the best component with other models to avoid -10 and other low log scores appears, while for ensembles of subsets of the forecasting methods, there are other benefits; Appendix A shows the

impact of these missed possibilities and the log score cap.

Figure 5.1 also includes estimates of the mean log score for each method and rough error bars for these estimates. We expect there to be strong statistical dependence across evaluations for the same season and location, and weaker dependencies between different seasons and locations; thus, the most common approaches to calculating standard errors, confidence intervals, and hypothesis test results will be inappropriate. Properly accounting for such dependencies and calibrating intervals and tests is an important but difficult task and is left for future investigation. We use "rough standard error bars" on estimates of mean evaluations: first, the relevant data (e.g., all cross-validation evaluations for a particular method and evaluation metric) is summarized into one value for each season-location pair by taking the mean of all evaluations for that season-location pair; we then calculate the mean and standard error of the mean of these season-location values using standard calculations as if these values were independent. Under some additional assumptions which posit the existence of a single underlying true mean log score for each method, these individual error bars — or rough error bars for the mean difference in log scores between pairs of methods — suggest that the observed data is unlikely to have been recorded if the true mean log score of the extended delta density method were greater than that of the adaptively weighted ensemble, or if the true mean log score of the "Empirical Bayes A" method were greater than the extended delta density method.

Methods that model wILI trajectories and "pin" past wILI to its observed values have a large number of log scores near 0 because they are often able to confidently "forecast" many onsets and peaks that have already occurred; ensemble methods also have a large number of log scores near 0. Note that these scores are closer to 0 for ensembles that optimize weighting of different methods than for the ensemble with uniform weights. For this particular set of forecasting methods, targets, and evaluation seasons:

- the equally-weighted ensemble has lower average log score than the best individual component (extended delta density),
- using the stacking approach to assign weights to ensemble components improves ensemble performance significantly and gives higher average log score than the best individual component,
- the adaptive weighting scheme does not provide a major benefit over a fixed-weight scheme using a single set of weights for each evaluation metric.

When given subsets of these forecasting methods as input, with regard to average performance:

- the equally-weighted ensemble often outperforms the best individual, but is sometimes slightly (≈ 0.1 log score) worse;
- the stacking approach improves upon the performance of the uniformly weighted ensemble; and
- the adaptive weighting scheme's performance is equal to or better than that of the fixed-weight scheme, sometimes improving on the log score by ≈ 0.1. The adaptive weighting scheme's relative performance appears to improve with more input seasons, fewer ensemble components, and increased variety in underlying methodologies and component performance. These trends suggest that using wider RelevanceWeight kernels, regularizing the component weights, or considering additional data from 2003/2004 to 2009/2010, for which ground truth wILI but not weekly ILINet reports are available, may improve the performance of the adaptive weighting scheme. In addition to these avenues for possible improvement in ensemble weights for the components presented in Figure 5.1, the adaptive weighting scheme provides a natural way of incorporating forecasting methods that generate predictions for only a subset of all targets, forecast weeks, or forecast types (distributional forecast or point prediction). For example, in the 2015/2016 season, we incorporated a generalized additive model that provided point predictions (and later, distributional forecasts) for peak week and peak height given at least three weeks of observations from the current season.

Figure 5.2 shows a subset of the cross-validation data used to form the ensemble and evaluate the effectiveness of the ensemble method, for two sets of components: one using all the components of Delphi-Stat, and the other incorporating three of the lower-performance components and a uniform distribution for distributional forecasts.[1] The Delphi-Stat ensemble near-uniformly dominates the best component, extended delta density, in terms of log score, and has comparable mean absolute

[1]The specific forecasting methods selected in the subset were "Empirical Bayes B", "BR, degenerate", and "Targets, conditional density", plus "Targets, uniform" for distributional forecasts. This subset was selected to examine ensemble performance on a subset of lower-performance methods based on different methodologies. We find roughly similar results on random subsets of methods, but performance gains are (a) lessened when there are less obvious difference in performance trends among the included components, and (b) limited when the highest-scored individual components (the delta density methods, especially the extended version).

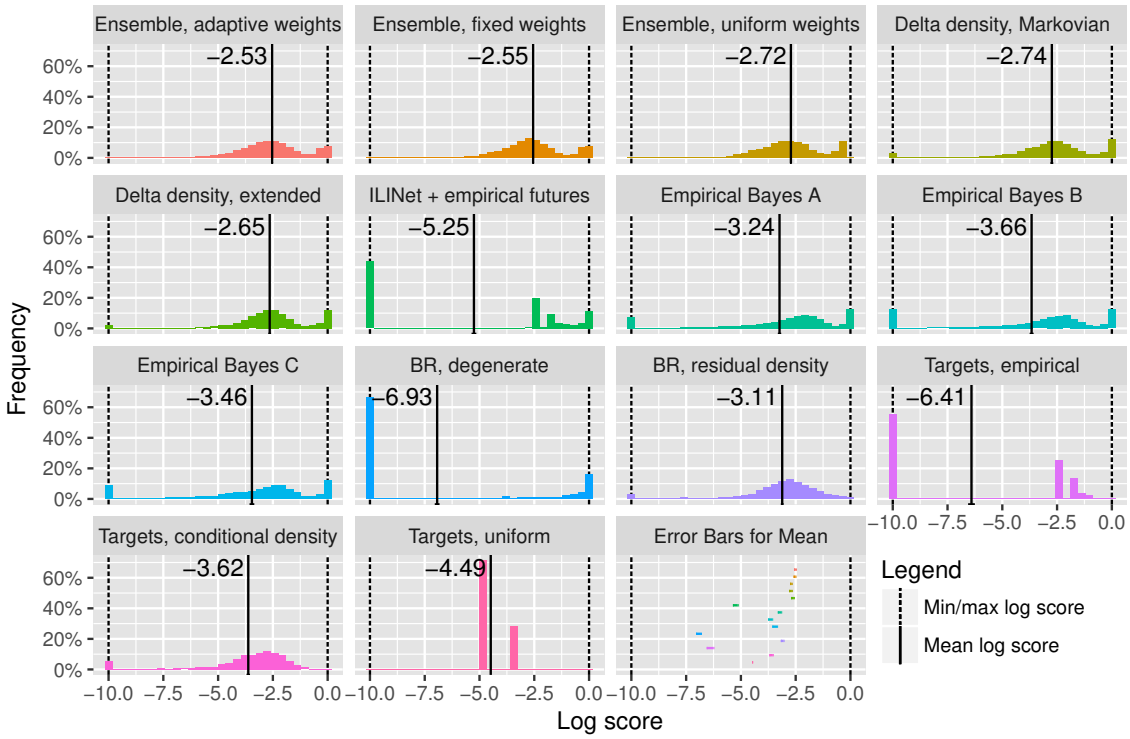

Figure 5.1: **Delta and residual density methods cover more observed events and attain higher average log scores than alternatives operating on seasons as a unit; ensemble approaches can eliminate missed possibilities while retaining high confidence when justified.** This figure contains histograms of cross-validation log scores for a variety of forecasting methods, averaged across seasons 2010/2011 to 2015/2016, all locations, forecast weeks 40 to 20, and all forecasting targets. A solid black vertical line indicates the mean of the scores in each histogram, which we use as the primary figure of merit when comparing forecasting methods; a rough error bar for each of these mean scores is shown as a colored horizontal bar in the last panel, and as a black horizontal line at the bottom of the corresponding histogram if the error bar is wider than the thickness of the black vertical line. Log scores near 0 typically correspond to forecasts of seasonal targets when most of the season is over.

Estimate	Median percent signed error
Initial report	−40% (underestimate)
⇝ 1 wk ahead ML forecast	−42% (underestimate)
Initial backcast	−3% (≈ unbiased)
⇝ 1 wk ahead ML forecast	−6% (minor bias)

Table 5.1: **Backcasting stable hospitalization data removes almost all bias in initial reports; this in turn removes most bias in ensemble 1 wk ahead point predictions.** Performance metrics were computed using cross-validation on data for the entire FluSurv-NET as a whole, and the "Overall" age group only. Bias and absolute error are calculated based on relative deviations (normalized by the stable values) for better interpretability, and use the median to give finite metric values in the presence of some stable values of zero.

error overall. The ensemble approach produces greater gains for the smaller subset of methods, surpassing not only its best components, but all forecasting methods in the wider Delphi-Stat ensemble except for the delta density approaches.

FluSurv-NET hospitalization forecasting

Table 5.1 shows that the backcasting method described in Chapter 3 (without the use of ILI-Nearby, which is designed to predict ILINet data) successfully removes bias and error in initial FluSurv-NET reports (taken as an estimate of stable values). Other cross-validation results (not in this table) indicate that incorporating backcasting reduces median relative absolute error as well. Figure 5.3 and Figure 5.4 show some similar cross-validation unibin log score tables as shown for the national and regional ILINet forecasts earlier. In the FluSurv-NET setting, faithfully incorporating distributional backcasts and nowcasts appears to have a larger impact than switching between the different forecasting methodologies considered, in contrast with the national and regional ILINet setup. This observation is not a surprise, as revisions are relatively larger on average and more biased on average than in the national and regional ILINet data.

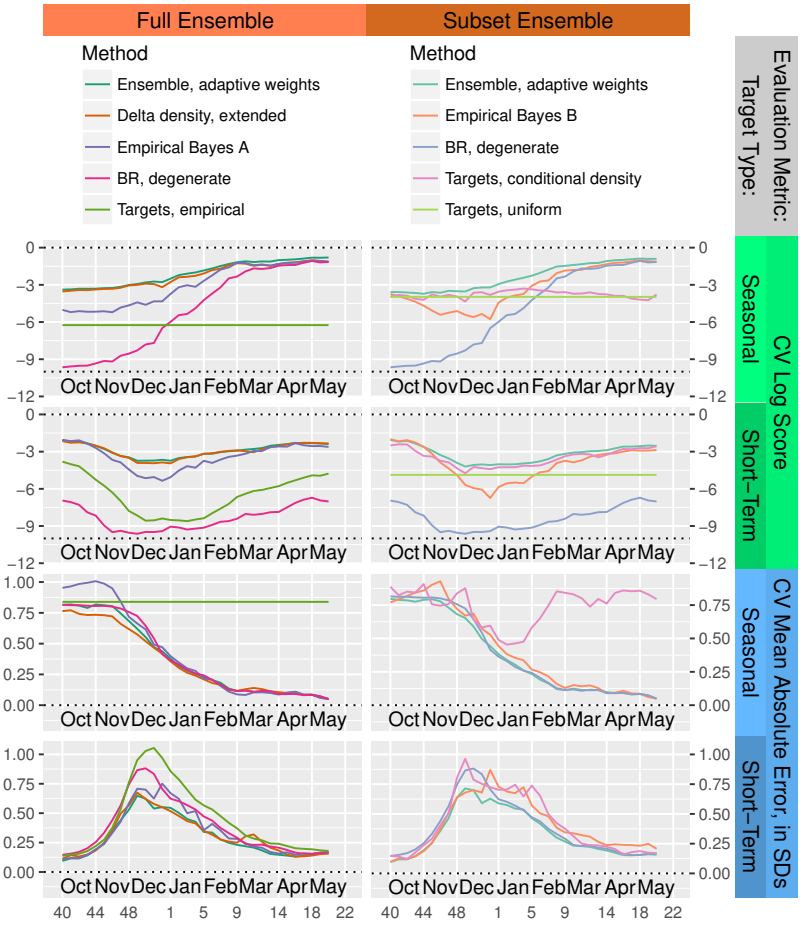

Figure 5.2: **The ensemble method matches or beats the best component overall, consistently improves log score across all times, and, for some sets of components, can provide significant improvements in both log score and mean absolute error.** These plots display cross-validation performance for two ensembles and some components broken down by evaluation metric, target type, and forecast week; each point is an average of cross-validation evaluations for all 11 locations, seasons 2010/2011 to 2015/2016, and all targets of the given target type; data from the appropriate ILINet reports is used as input for the left-out seasons, while finalized wILI is used for the training seasons. Top half: log score evaluations (higher is better); bottom half: mean absolute error, normalized by the standard deviation of each target (lower is better). Left side: full Delphi-Stat ensemble, which includes additional methods not listed in the legend; right side: ensemble of the three methods listed in the legend, with the uniform distribution component incorporated only in distributional forecasts. Many components of the full ensemble are not displayed. The "Targets, uniform" method is excluded from any mean absolute error plots as it was not incorporated into the point prediction ensembles.

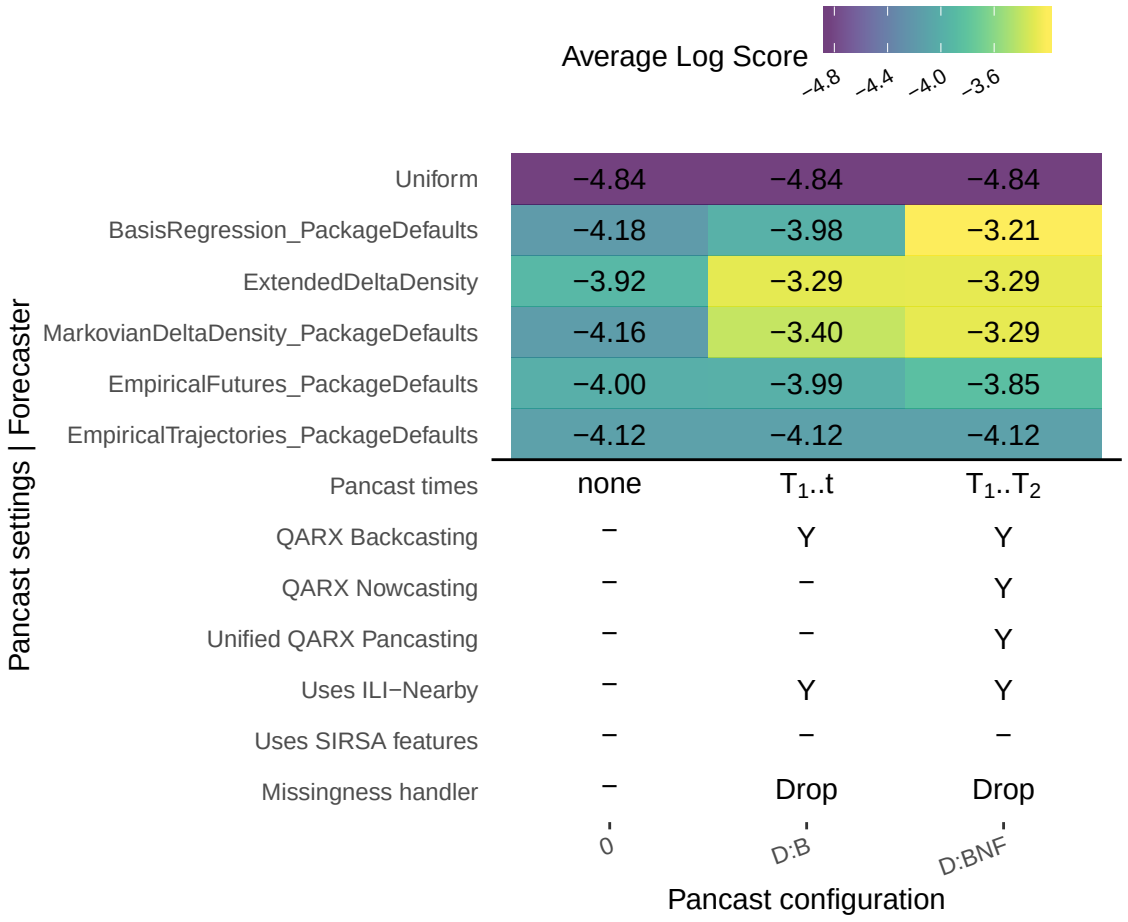

Figure 5.3: **Cross-validation overall unibin log scores for FluSurv-NET forecasts for a few pancaster-forecaster pairs.**

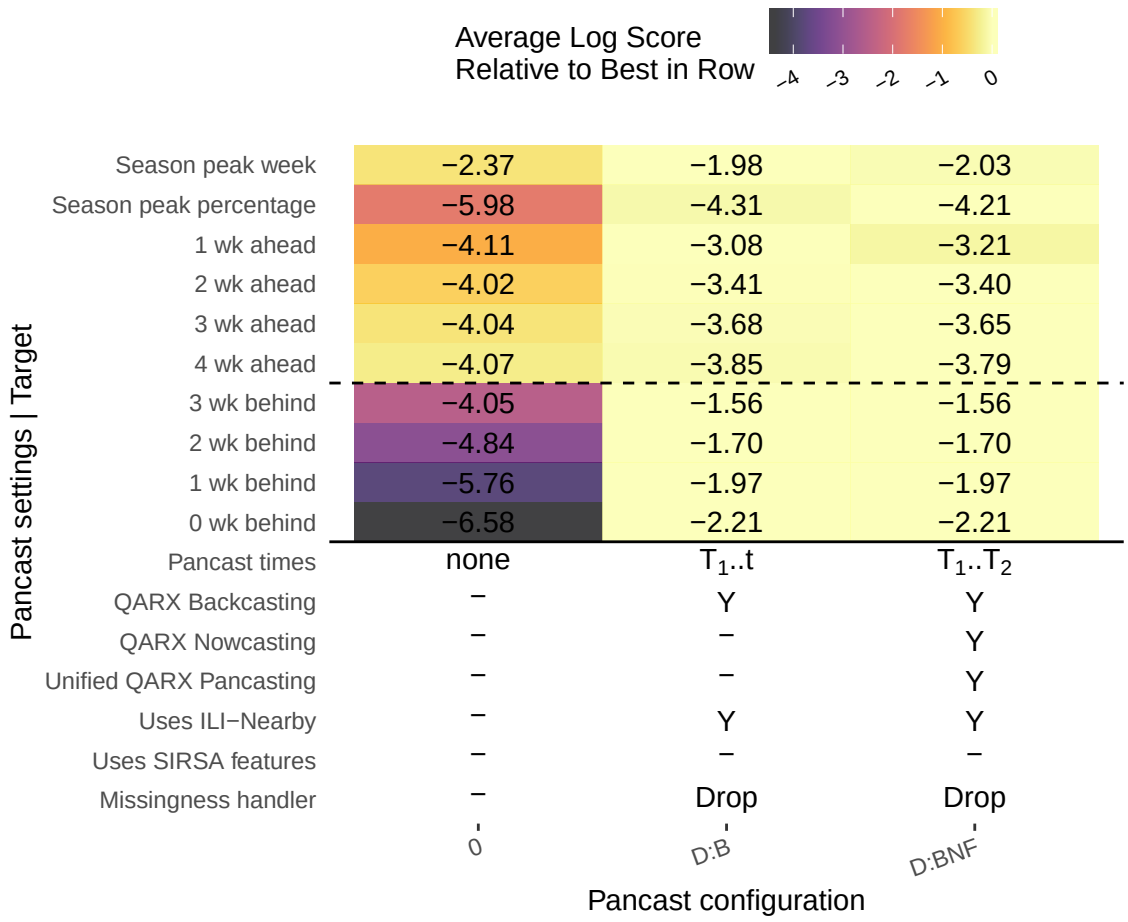

Figure 5.4: **Cross-validation overall unibin log scores for FluSurv-NET forecasts broken down by target and pancaster, using the ExtendedDeltaDensity forecaster; "x wk behind" targets are included for additional information and not considered in overall score comparisons.**

5.3.2 External, prospective evaluation

The ensemble methodology set forth above has been externally and prospectively evaluated as part of CDC's Epidemic Prediction Initiative's ongoing comparisons of forecasting methodologies developed by multiple research groups. Table 5.2 summarizes the overall multibin log scores for a few systems in these comparisons:

- Delphi-Stat: the ensemble framework above, using one of the three weighting schemes described earlier, or a similar scheme (selected based on cross validation unibin log score), applied to the collection of methods described in Appendix B, Figure 4.5, or some similar set; at least some form of backcasting was incorporated since midway through the 2015/2016 season.
- EB, Spline: early stand-alone versions of the empirical Bayes and basis regression methods described in Chapter 2 and Appendix B.
- Delphi-Epicast Farrow [2016], Farrow et al. [2017]: a wisdom-of-crowds approach to forecasting which is has relatively good performance and consistent application across seasons and forecasting targets. (Epicast2 is Epicast with a submission error corrected. Epicast-Mturk is a variant incorporating forecasts crowdsourced from Amazon Mechanical Turk in place of the typical volunteer pool.)
- FSNet: FluSight-Network Reich et al. [2019b]: the ensemble framework above, using one of a few similar weighting schemes (selected based on pseudoprospective multibin log score), applied to a collection of methods from multiple research groups Reich et al. [2019a], organized, run, monitored, and maintained by researchers in the Reich Lab at the University of Massachusetts Amherst.
- EqWts: an ensemble of all CDC comparison entries, with each entry assigned equal weight.
- PPFST, PPFST2 Morgan et al. [2018]: a wisdom-of-crowds approach to ensemble forecasts, incorporating methods from multiple research groups.
- HistAvg: a baseline method based on kernel density estimation that does not factor in any observations from the season for which forecasts are being made.

The ensemble framework described in this chapter is incorporated in two systems (Stat and FSNet) with consistently high rankings in these comparisons.

— 2014/2015 — US National		— 2015/2016 — US + Regions		— 2016/2017 — US + Regions		2017/2018 US + Regions		States		Hospitalizations		2018/2019 US + Regions		States		Hospitalizations	
Rank	System	Rank	System	Rank	System	Rank	System	Rank	System	Rank	System	Rank	System	Rank	System	Rank	System
1	Epicast	1	**Stat**	1	Epicast	1	Epicast	1	**Stat**	1	**Stat**	⋮	2 others	1	.	1	**Stat**
2	EB	2	Epicast	2	**Stat**	2	FSNet	2	.		EqWts	3	FSNet	2	**Stat**	2	.
	HistAvg	3	Archefilter		EqWts	⋮	3 others	3	.	2	.		PPFST2		PPFST2		EqWts
3	.	4	.	3	.		EqWts		EqWts	3	Epicast		PPFST	3	.		HistAvg
4	.	5	.	4	.		PPFST	4	.		HistAvg	4	**Stat**		EqWts	3	Epicast
5	.	6	.	⋮	8 others	⋮	2 others	5	.	4	.	5	.		PPFST	4	.
6	Spline	7	.	13	.	8	**Stat**		HistAvg				EqWts	⋮	2 others	5	.
7	.	8	.	14	.	9	.	6	.			6	.		HistAvg		
		9	.	15	.	⋮	6 others	7	.			7	Epicast2	6	.		
		10	.		HistAvg	16	.	8	.			⋮	7 others	7	.		
		11	.	16	.		HistAvg	9	.			15	Epicast	8	Epicast-Mturk		
		12	.	17	.	17	.	10	.			⋮	6 others	9	.		
		13	.	⋮	8 others	18	.	11	.				HistAvg	10	.		
		14	.	26	.	⋮	9 others	12	.			22	.	⋮	2 others		
				27	.	28	.					⋮	11 others	13	.		
				28	.	29	.					33	.	14	.		

Styling key:
Delphi-Stat and early stand-alone versions of components (this work)
Delphi-Epicast and variants (wisdom-of-crowds approach)
FluSight-Network ensemble (same stacking framework with components from multiple research groups)
Other multi-group ensembles (unranked; prepared on different schedule from other systems)
HistAvg baseline (target distribution from other seasons; does not use data from current season; unranked)
Other forecasting systems (denoted by "." for single systems or "⟨n⟩ others" for multiple)

Table 5.2: **Delphi-Stat consistently attains high ranks in comparisons organized by CDC's Epidemic Prediction Initiative.** The FluSight-Network multi-group ensemble, which uses the same framework, but includes component models from multiple groups and considers different weighting schemes. The 2013/2014 national and regional ILINet forecasting challenge is not included in this table as there was not a comprehensive ranking published; forecasts based on the empirical Bayes methodology were submitted but did not rank first.

Appendix A

"Missed possibilities" and -10 log score threshold

This appendix reproduces or incorporates content from Brooks et al. [2018].

Individual unibin and multibin log scores below -10 have been increased to a minimum of -10 (as if a probability of ≈ 0.0000454 — still very small for the selected bin sizes — had been assigned) in all analysis. This threshold operation can be interpreted as adding up to a certain amount of probability mass to a distributional forecast, which normally has a total probability mass of 1; the maximum number of bins in any target's distributional forecast is 131, so the threshold operation cannot increase the amount of probability mass to more than ≈ 1.006. This threshold was implemented by CDC for forecast comparisons so that submissions would not be assigned very low mean log scores (e.g., $-\infty$) for assigning a few events extremely low (e.g., 0) probabilities to events that actually occurred. We also use it when comparing individual methods in the ensemble. Without such a threshold, each FluSight submission or ensemble component would need to ensure that no possibilities are missed and assigned extremely low probabilities, e.g., by mixing model forecasts with a uniform distribution (which bears similarity to the threshold operation) using the rule of three to determine the mixing weights. Thresholded log scores are no longer proper scores, as forecasters may expect to benefit by reporting probabilities of 0 for any bin with a modeled probability less than the exponentiated threshold, and using the difference in mass to increase probabilities assigned to other bins; with a threshold of -10, there is not much expected benefit (at most ≈ 0.006 mass would

be reassigned), but at higher thresholds, this impropriety may be problematic. The stacking-based ensembles presented in the main text, and in this appendix unless otherwise noted, use weight selections intended to maximize mean unibin log score without thresholding.

For the full Delphi-Stat ensemble, the main advantage of the ensemble over its best component appears to be successfully filling in possibilities missed by the best component with other models to avoid -10 and other low log scores appears, while for ensembles of subsets of the forecasting methods, there are other benefits. We investigate changes to this log score threshold, and experiment with removing the lowest $p\%$ of log scores instead. As the log score threshold or p is increased, the relative performance of an ensemble over the best component declines and becomes negative when the ensemble is still tuned to optimize non-thresholded log score. Tailoring the optimization criterion to better match modified evaluation criteria can help restore the ensemble's superior or competitive performance compared to its best component.

A.1 Analysis of full Delphi-Stat ensemble

Figure 5.1 shows histograms of the cross validation log scores of the Delphi-Stat components and full ensemble with the original $-10 \approx \log(0.0000454)$ threshold; compared with the extended delta density method, the adaptively weighted ensemble:

- has higher mean log score;
- eliminates all -10 log scores;
- has less log scores of 0, but more right below 0; and
- smoother and wider tails about the mode of the histogram near the mean log score.

Figure A.2 shows the same histograms using a threshold of $-7 \approx \log(0.000912)$; the four points above still hold, but the difference in mean log score between the two forecasters is notably smaller.

Figure A.3 and Figure A.4 show that the adaptively weighted ensemble and extended delta density are surpassed by other methods for thresholds from -3 to 0. However, Figure A.4 also shows that a threshold of $-3 \approx \log(0.0498)$ already changes

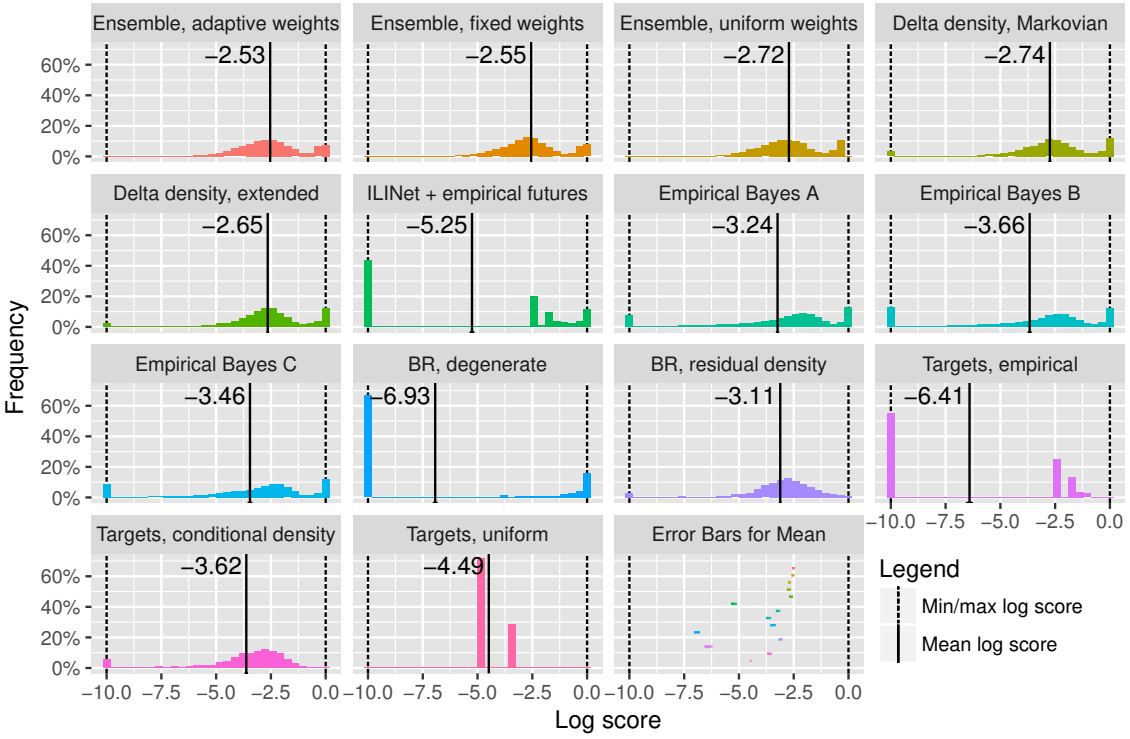

Figure A.1: **Figure 5.1 from the main text: log score means and histograms for each method using a log score threshold of -10, and ensemble weights trained ignoring the log score threshold.** This figure contains histograms of cross-validation log scores for a variety of forecasting methods, averaged across seasons 2010/2011 to 2015/2016, all locations, forecast weeks 40 to 20, and all forecasting targets. The solid black vertical lines indicate the mean of the scores in each histogram, which we use as the primary figure of merit when comparing forecasting methods; a rough error bar for each of these mean scores is shown as a colored horizontal bar in the last panel, and as a black horizontal line at the bottom of the corresponding histogram if the error bar is wider than the thickness of the black vertical line.

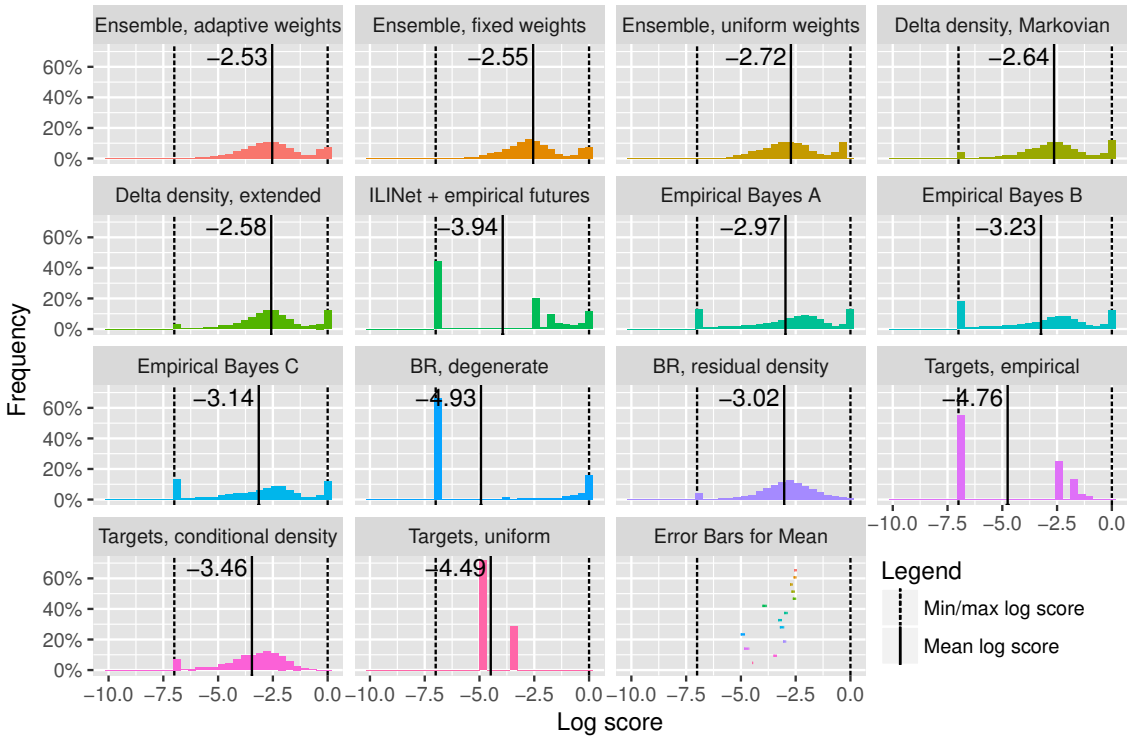

Figure A.2: **Log score means and histograms for each method using a log score threshold of -7 and ensemble weights trained ignoring the log score threshold.**

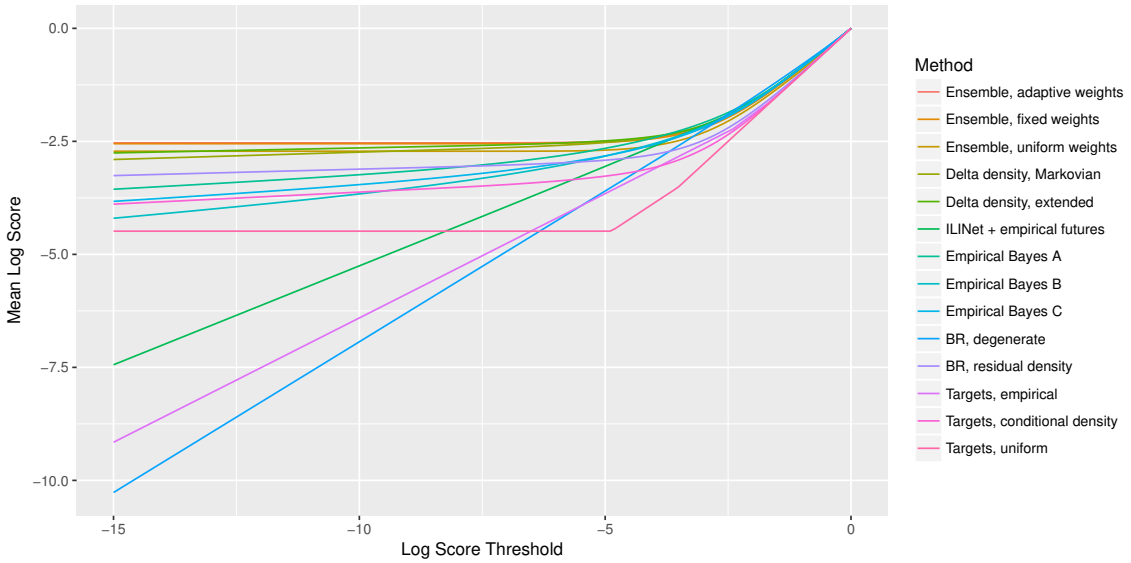

Figure A.3: **Thresholded mean log scores for each method and thresholds from -15 to 0.**

from 25% to over 50% of the log scores for each method, which seems inappropriate. Nevertheless, ensemble methods could still be useful in this case, but the weight selection objective must be updated to better match the evaluation metric; Figure A.5 shows that the ensemble score can be improved significantly by solving a relaxation (approximation) of the thresholded log score optimization problem. The relative trends are similar when throwing away the lowest $p\%$ of log scores for a method rather than imposing a minimum log score threshold; Figure A.6 shows that, when p is high enough to discard all $-\infty$ log scores for delta density methods, their performance is similar to that of the ensemble. Again, optimizing the ensemble weights to these modified error metrics could potentially result in performance improvements.

A.2 Analysis of a subset of presented methods

Figure A.7 shows log score histograms for a subset of the methods above and ensembles using only those methods. The best component in this subset is "Targets, conditional density", which is completely missing the spike in log scores near 0 present in "Empirical Bayes B" and "BR, degenerate" (which model trajectories and calculate target distributions from these trajectory distributions), but still has higher mean

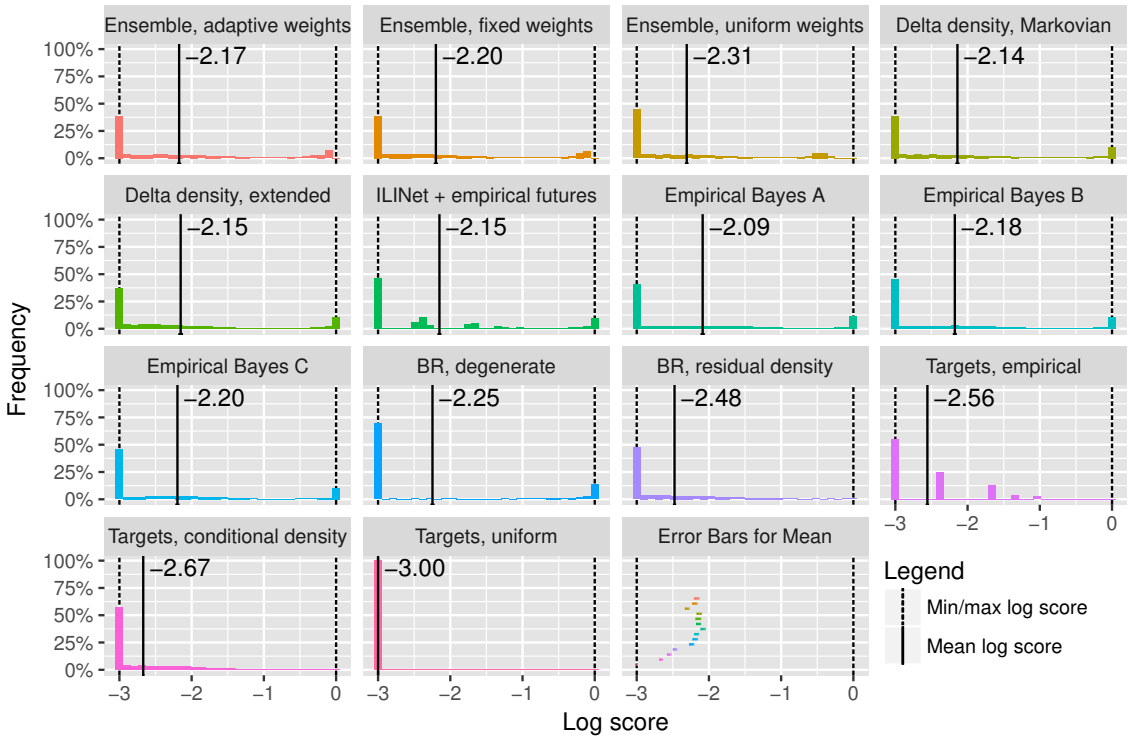

Figure A.4: **Log score means and histograms for each method using a log score threshold of -3 and ensemble weights trained ignoring the log score threshold.** Note that the ranges of values shown along both axes differ from the ranges used for similar figures for the -10 and -7 thresholds.

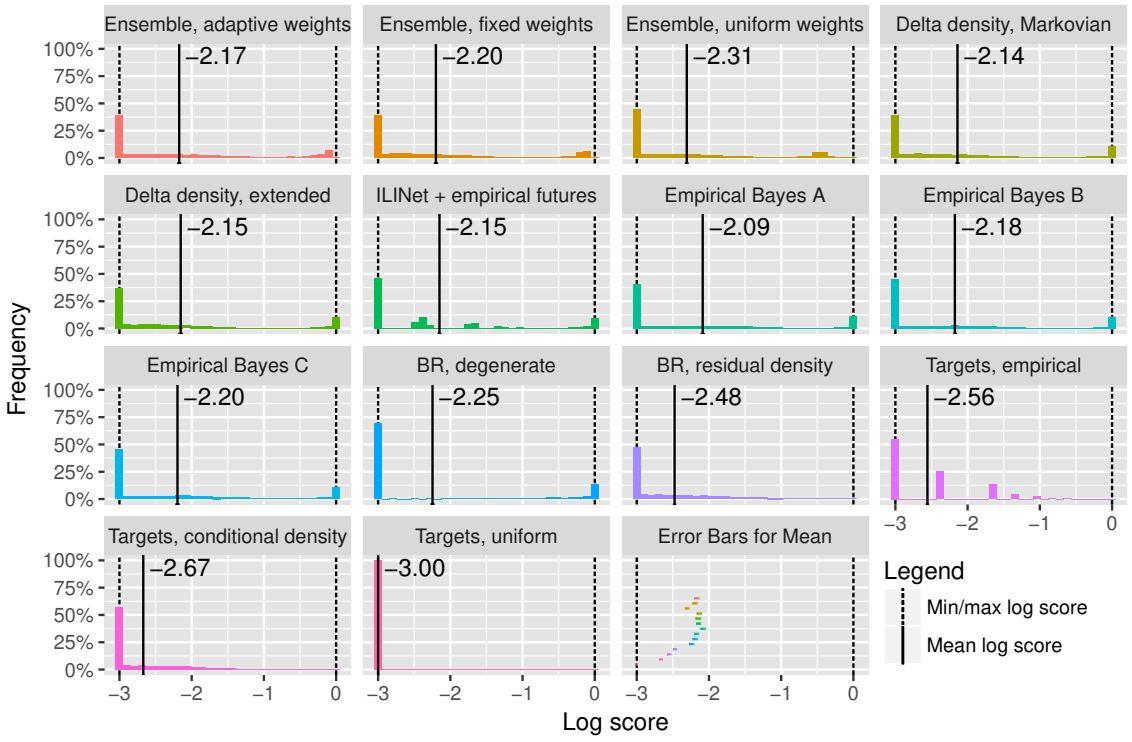

Figure A.5: **Log score means and histograms for each method using a log score threshold of -3 and ensemble weights trained using a concave relaxation of thresholded log score.** Note that the ranges of values shown along both axes differ from the ranges used for similar figures for the -10 and -7 thresholds.

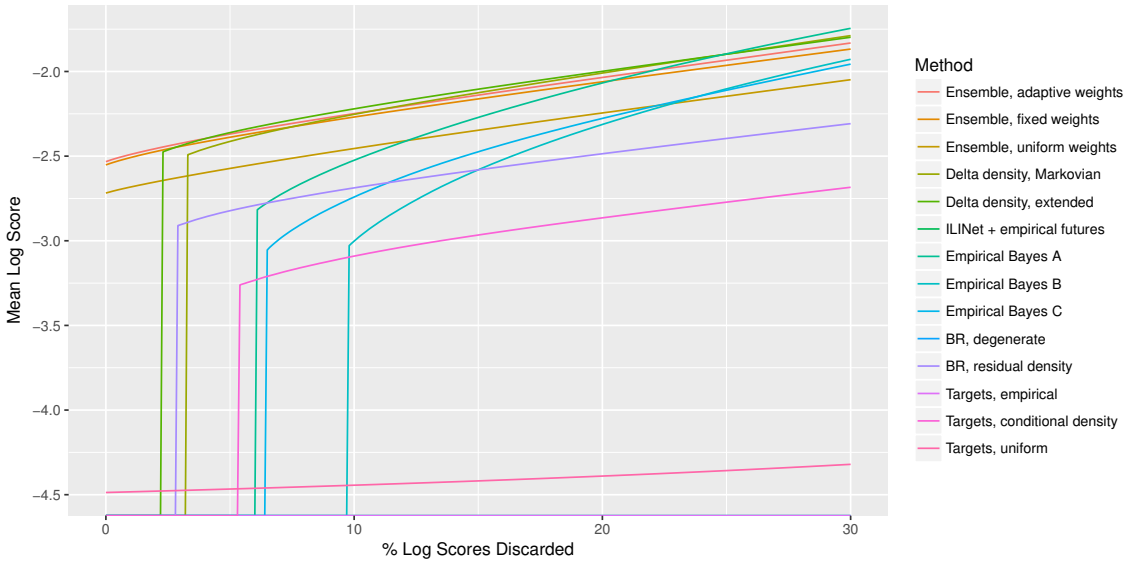

Figure A.6: **Mean log score for each method in the full ensemble, with no thresholding but throwing out the lowest p percent of log scores for each method for various values of p.**

log score than these two due to less scores of -10 and a higher concentration of scores from -5 to -1. The ensemble is able to combine the strengths of these models and the uniform distribution, avoiding any scores of -10 (or even -8), incorporating a spike in log scores near 0, and concentrating the rest of its log scores on the higher end of the -8 to -1 range. "Empirical Bayes B" is a close second to "Targets, conditional density", but the ensemble approach provides additional benefit besides just avoiding its missed possibilities; Figure A.8 shows that, even when ignoring the lowest 10% of log scores for each method (which removes all scores of $-\infty$ for "Empirical Bayes B"), the adaptively weighted ensemble provides a large improvement in log score. This benefit vanishes and "Empirical Bayes B" starts to perform better as higher percentages (20% to 30%) of log scores are ignored; again, it may be possible to construct a successful ensemble in these cases by choosing an optimization criterion more similar to the evaluation criterion.

Figure A.7: **Log score means and histograms for a subset of methods (the same as the subset in Chapter 5 of the main text) using a log score threshold of -10, and ensemble weights trained ignoring the log score threshold.**

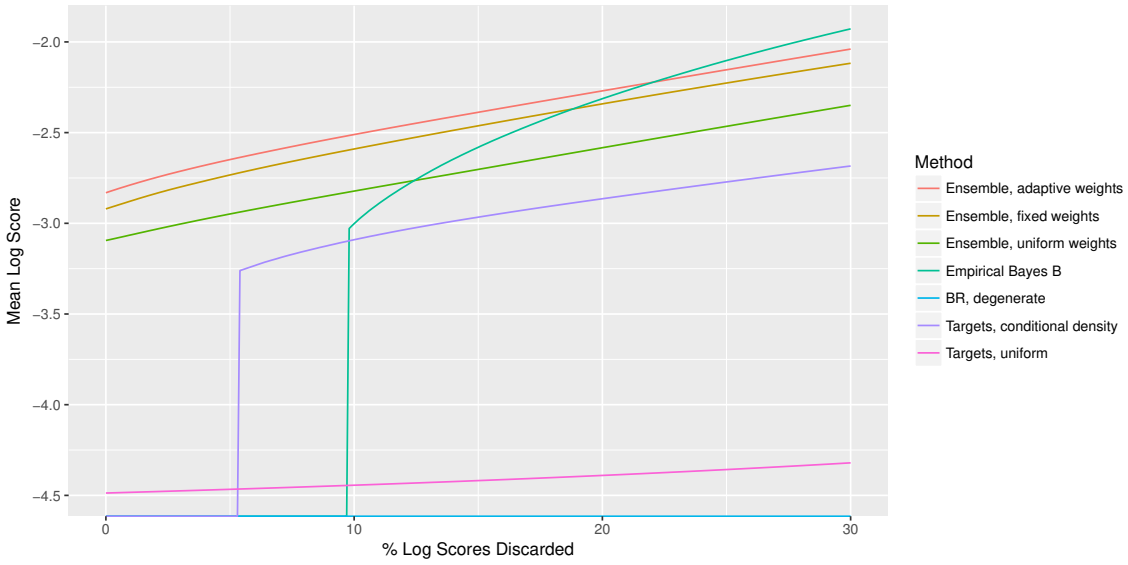

Figure A.8: **Mean log score for each method in the subset ensemble, with no thresholding but throwing out the lowest p percent of log scores for each method for various values of p.**

Appendix B

Description of all ensemble components in the 2015/2016 Delphi-Stat forecasting system

This appendix reproduces or incorporates content from Brooks et al. [2018].

This appendix describes details of the components of the Delphi-Stat ensemble system following the end of the 2015/2016 season, the version used for the performance analysis in Section 4.2 and Section 5.3.1. Changes made to Delphi-Stat throughout the 2015/2016 season are described in Appendix C; additional changes, such as a switch to the quantile regression pancasting framework, were implemented in subsequent seasons. Our past and ongoing forecasts, as well as `Python` [Van Rossum and Drake, 2003] and `R` [R Core Team, 2015] code for components of the systems used to generate them, are publicly available online [Brooks et al., 2015b, Carnegie Mellon University Delphi group, Accessed 2017-04-26, Reichek and Gao, Accessed 2017-04-26].

B.1 Ensemble components

Delphi-Stat incorporated 10 individual forecasting methods in the 2015/2016 season based on diverse methodologies to forecast the targets of interest $\mathbf{Z}^{(t)}$ conditioned on the finalized wILI values up to time t, $Y_{1..t}$. When producing prospective forecasts,

we do not have access to the finalized values $Y_{1..t}$, but rather the t-th report for the current season, $Y^t_{1..t}$; we discuss a method for distributional estimates of $Y_{1..t}$ based on $Y^t_{1..t}$ in the main text. All methods produce distributional forecasts for the targets of interest; their point predictions are the medians of the corresponding distributional forecasts. Most methods, rather than directly producing forecasts for the targets, first estimate the distribution of the entire wILI trajectory $Y_{1..T}$ based on the available data, then calculates the corresponding distribution over the targets. (Since the data are at a weekly resolution, the number of wILI values in the current season, T, is either 52 or 53; we present the methods here as if all seasons were of the same length T, omitting all details dealing with mismatches between the length of a training season and the length of a test season.)

B.1.1 Methods based on delta density

Markovian delta density

Described in the main text.

Extended delta density

Described in the main text.

B.1.2 Methods based on empirical distribution of curves

Another class of methods are based on using and expanding the empirical distribution of wILI trajectories.

Empirical distribution of wILI trajectories for future times

Consider all $Y^s_{t+1..T}$, $s \in \{1 .. S\}$, equally likely to reoccur. Observations from the current season are used for times up to t.

Empirical Bayes procedure on wILI trajectories

Model $Y_{1..T}$ as some underlying curve, $F_{1..T}$, plus i.i.d. Gaussian observational noise. Estimate $F^s_{1..T}$ and a noise level for each $s \in \{1 .. S\}$ using a trend filtering proce-

dure. Build a distribution for $F_{1..T}$ and the noise level using these estimates, plus a probability distribution over ways to shift and scale these curves to produce a wider range of possibilities for $Y_{1..T}$. The resulting distribution describes our prior beliefs about the distribution of $Y_{1..T}$ before seeing any observations from the current season; calculate the corresponding posterior distribution, $Y_{1..T} \mid Y_{1..t}$, describing our beliefs after seeing the available observations, using importance sampling techniques [Liu, 2008].

Implements the empirical Bayes method as described in [Brooks et al., 2015a], with a few modifications:

- Only the time-shift and wILI-scale transformations are used.
- The time-shift is a "local" transformation: rather than having a distribution of peak weeks determine the shift amount, we directly choose a distribution over shift amounts. Specifically, we use a discrete uniform distribution over integers centered at zero, width equal (ignoring rounding) to twice the bin width of a histogram of the historical peak weeks using Sturges' rule.
- The wILI-scale is a "local" transformation: rather than having a distribution of peak heights determine the scale amount, we directly choose a distribution over scale amounts. Specifically, we use a log-uniform distribution centered at 0 in the log-scale with log-scale width equal tot twice the bin width of a histogram of the logarithms of the historical peak heights, using Sturges' rule. Note that this behavior can significantly bias the mean of the prior for the peak heights, but does not significantly affect the median of the prior for the peak heights. Another difference from the scaling transformation in the paper is that, instead of scaling the wILI trajectory above and about the CDC baseline, we scale from 0, and also multiply the noise associated with each observation based on how much it was scaled.
- Instead of randomly mixing and matching smooth curve shapes and noise levels, these two parameters are linked together: a given noise level estimate is always paired with the corresponding smoothed curve.
- We add a "reasonable future" term to the posterior log-likelihood (given observations in past weeks) of each proposed trajectory, proportional to the average log-likelihood of the 3 most similar historical curves in future weeks.
- We condition on a maximum of 5 observations from the current season; if more than 5 observations are available for the current season, we use only the most

recent 5.

- We use the `glmgen` package [Arnold et al., 2014] to rapidly perform trend filtering for smoothing past seasons' trajectories.

We form two other versions of the empirical Bayes forecaster by using subsets of these changes and other parameter settings; these variants were used in the 2016/2017 ensemble but not the 2015/2016 ensemble.

B.1.3 Basis regression approach

Estimates the mean curve $\mathbb{E}Y_{1..T}$ with elastic net regression from a collection of basis functions to a trajectory of "pseudo-observations" $\tilde{Y}_{1..T}$ which is the concatenation of (a) the available observations $Y_{1..t}$, and (b) the pointwise mean of $Y^s_{t+1..T}$ for $s \in \{1 \,..\, S\}$. We chose a B-spline basis, which produces a variation on smoothing spline estimation of $\mathbb{E}Y_{1..T}$. The `glmnet` package [Friedman et al., 2010] was used to perform the elastic net regression, with evenly weighed L^1 and L^2 regularization (the default setting, $\alpha = 0.5$), and to automatically select the overall regression penalty coefficient λ using random 5-fold cross-validation on weeks of the current season, seeing how well the smoothed estimate for $\mathbb{E}Y_{1..T}$ is able to predict left-out pseudo-observations from $\tilde{Y}_{1..T}$.

Basis regression with degenerate distributional forecast

Forecasts that $Y_{1..T}$ will be equal to the basis regression estimate for $\mathbb{E}Y_{1..T}$ with probability 1. There is a small amount of randomness in the basis regression estimation procedure itself arising from the default method for selecting λ, so we actually take a sample by calling the procedure many times, forming a very narrow distribution.

Basis regression with residual density distributional forecast

Constructs a distributional forecast for $Y_{1..T}$ by applying the residual density method with $X_{1..T}$ equal to the basis regression estimate for $\mathbb{E}Y_{1..T}$ and other settings the same as in the Markovian delta density method.

B.1.4 No-trajectory approaches

These approaches form a forecast for $\mathbf{Z}^{(t)}$ from an estimate of $Y_{1..t}$ without first constructing a forecast for the entire trajectory $Y_{1..T}$.

Empirical distribution of target values

Consider all $Y^s_{1..T}$, $s \in \{1 \,..\, S\}$, equally likely to reoccur, ignoring and overriding the available observations from the current season ($Y_{1..t}$). For each target, the distributional forecast is its empirical distribution, and the point prediction is the corresponding median.

Direct target forecasts with kernel smoothing

Uses the kernel smoothing method used in the delta density method to estimate the distribution of $\mathbf{Z}^{(t)}$ conditioned only on (an estimate of) Y_t.

Direct target forecasts with generalized additive model

Uses a generalized additive model to predict the expected value of a subset of the targets, and assumes a normal distribution for the residuals when making distributional forecasts. Provided by Shannon Gallagher. This method was used in the 2015/2016 ensemble, but not the 2016/2017 ensemble nor the cross-validation analysis.

Uniform distribution

Outputs the same probability for each bin, regardless of the input data. The corresponding point predictions are excluded from the ensemble.

Appendix C

Log of changes to Delphi-Stat throughout the 2015/2016 season and for cross-validation analysis

This appendix reproduces or incorporates content from Brooks et al. [2018].

C.1 Initial description (2015 EW42)

The Delphi-Stat system is an ensemble of several baselines and statistical forecasting methods. Its forecasts are a linear combination of the forecasts of these individual systems, with a separate set of coefficients determined for each epi week, geographical area (nation + 10 HHS regions), metric (MAE or log score), and target. The methods are outlined below. Note that the term "past epiweeks" refers to a set of epi week numbers in any season — specifically, epi weeks 21 up to the forecast week; "future epiweeks" is used in a similar fashion. (Nonnegative coefficients summing to 1 are calculated for point predictions using constrained LAD regression (implemented using the linear programming package `lpSolve` [Berkelaar and others, 2015]), and for distributional predictions with the degenerate EM algorithm [Rosenfeld, Accessed 2017-03-21].)

- **Empirical prior:** ignores all data from the current season, and considers each

training season — 2003/2004 to 2014/2015, excluding the pandemic — as equally likely to reoccur.

- **Pinned baseline:** uses the available observations for the current season for previous epi-weeks; for future epi weeks, each training curve is considered equally likely to reoccur.

- **Basis regression:**

    1. Aligns training curves with the current season by shifting in time and scaling weighted ILI values until the maximum of each training curve in past epiweeks is the same as that of the current season. (Scaling is performed only above the CDC baseline; if a curve is entirely below the CDC baseline, it is not scaled at all.)
    2. Fits a smooth curve to the observed data in past epiweeks and the mean of the aligned training curves in future epiweeks. (The smooth curve is a spline: specifically, a linear combination of B-splines selected with elastic net using the glmnet package [Friedman et al., 2010], with a trade-off penalty between the importance of matching past and future epiweeks.)
    3. Uses observations from the current season in past epiweeks; considers this single curve as the only possibility for future weeks.

- **Basis regression with noise:**

    1. Generates the spline curve above.
    2. Considers the spline as estimating the change in weighted ILI from one week to the next; for each epi week, estimates the distribution of errors at that epi week using the training curves. (Distributions are estimated using weighted kernel density estimation: when adding noise to a simulated 2015/2016 curve at some future epiweek, training curves that more closely resemble the simulated curve in previous epiweeks contribute more to the result.)
    3. Generates many simulated 2015/2016 curves by taking the observations from the current season so far, and at each week, adding the estimated change from the spline curve, then drawing a value from the estimated error distribution.

- **Time-parameterized weighted kernel density estimation:** Follows the same process as the basis regression with noise; however, it directly estimates the

distribution of changes in weighted ILI values, rather than the corresponding distribution of errors in the spline estimate.

- **Empirical Bayes:** We use the procedure described in this document [Brooks et al., 2015a], with a few modifications: a smoothed (trend-filtered [Tibshirani, 2014]) curve is never paired with a noise estimate from another smoothed curve, scaling and shifting is performed only in small amounts resulting in "local" transformations, an additional component is added to the likelihood to encourage reasonable predictions at future weeks (by penalizing simulated curves if they deviate too much from all of the training curves), and incorporating a random inflation in the noise parameter to prevent forecast "overconfidence".

- **Uniform prior:** Considers each cell in the spreadsheet to be equally likely. (This component only produces distributional forecasts.) Additional weight is added to this component after the coefficients for each method are determined via cross-validation to prevent any 0 or near-0 probability forecasts.

C.2 Changes, 2015 EW43

- **Mixing coefficients between methods:** a set of weights for each of the forecasting methods is determined for each epi week, metric (MAE or log score), and target, but are tied across areas (nation + 10 HHS regions); thus, any method will receive the same weight in all areas (for the same epi week, metric, and target). For distributional forecasts, the weight assigned to the uniform distribution is increased by approximately 2.5% (based on the rule of three), and weight taken away evenly from all methods to make the weights again sum to 1. This is accomplished by changing the RelevanceWeight function from

$$\text{RelevanceWeight}(s,l,t,i,e;s',l',t',i',e') = \begin{cases} 1, & l = l',\ t = t',\ i = i',\ e = e' \\ 0, & \text{otherwise} \end{cases}$$

to

$$\text{RelevanceWeight}(s,l,t,i,e;s',l',t',i',e') = \begin{cases} 1, & t = t',\ i = i',\ e = e' \\ 0, & \text{otherwise,} \end{cases}$$

and setting μ as described in the main text. These changes motivated on two hypotheses:

    - The previous weight vector calculations, which previously only considered 11 training instances at a time (one per season from 2003/2004 to 2014/2015, excluding 2009/2010), were based on much too little data, and considering training instances from other locations would be beneficial (even though training data from other locations seems less relevant than training data from the same location).
    - The μ value from the rule of three will be more appropriate than an μ value selected to ensure an arbitrary minimum log score value, and will automatically update based on the amount of training data available.

- **New method added to ensemble:** direct target density estimation: uses the same weighted kernel density estimation approach as two existing methods to directly forecast each of the targets without constructing, rather than constructing a distribution of flu curves and extracting the target values from these curves. Adjustments to the output are made so that all predicted possible values are integers when appropriate and lie in the correct range.

C.3 Changes and clarifications, 2015 EW44

- **New method added to ensemble:** modified time-weighted kernel density estimation: this version changes the weighting criteria used for matching simulated data for this year to past seasons; attempts to make simulated trajectories more closely resemble past seasons' data; and considers a wider range of past data. When constructing trajectories, this version weights past seasons based on the previous week's wILI value; the sum of previous wILI values in the season; a weighted sum of wILI values stressing more recent weeks; and a weighted sum of the week-to-week changes in wILI stressing more recent times. With low probability, these weights are ignored and a random change in wILI is selected from historical data. The simulated data values are also pushed towards randomly selected historical data by a small amount. When simulating data at epi week t, instead of just looking at other seasons at week t, also considers nearby weeks, unless t is a time near the end of year holidays.

- **Clarification:** older kernel density estimation method, direct target density estimation: only weight data based on the previous wILI value.

C.4 Changes, 2015 EW46

- **Backfill forecasting:** we now use backfill forecasting in combination with almost all of the forecasting methods in the Delphi-Stat ensemble. For each nonfinal wILI value in the current season, we estimate a distribution for its final revised value. The distribution is based on historical revisions of wILI with the same lag (e.g., the latest measurement vs. the second most recent measurement), and is formed using weighted kernel density estimation, with weights depending on the epiweek to which the measurement corresponds, and the nonfinal wILI value itself.

C.5 Changes, 2016 EW03

Another statistical method has been added to the Delphi-Stat ensemble:

- **Target forecast:** We use an additive model to create predictions that are target specific using the past 3 values observed.

C.6 Changes, for cross-validation analysis

- **Changes to ensemble weight training data:** ensemble weights are selected using cross-validation component forecasts based on the version of test season data that would have been available at the forecast time, rather than ground truth; since regional back issues are available starting only in late 2009, cross-validation analysis is performed on seasons 2010/2011 to 2015/2016 as described in the main text.

- **Changes to RelevanceWeight function:** the RelevanceWeight function still seems like it will lead to ensemble weight vectors based on too little training data, especially considering the reduction in the number of training seasons, so we use the RelevanceWeight function specified in the text, which considers cross-validation component evaluations from forecast weeks within 4 weeks of t when setting weights for forecast week t (chosen to include many additional weeks while keeping early-season evaluations from influencing late-season weights, and late-season evaluations from influencing early-season weights).

- **Changes to methods in ensemble:** the additive model was removed from the ensemble to ease system maintenance, and the two Empirical Bayes variants were added to compare cross-validation forecast behavior and potentially improve the ensemble performance.

Appendix D

Additional details on selected elements of pancasting system

Source code for the Delphi-Stat system is available online as an R package [Brooks et al., 2015b]. This appendix describes certain elements of the system used for the performance analysis in Subsection 4.3.1, and the current approach to ensemble forecasting.

D.1 Ensemble forecasting

In the ILINet and FluSurv-NET forecasting settings, retrospective forecasts are made for each season and week from some set, and contain predictions for each "epigroup" (location for ILINet, age group for FluSurv-NET), target (onset week, peak week, etc.), and forecast type (point or distribution). When preparing ensemble forecasts, we consider "instances" to be component forecasts prepared:

- in season s,
- using data from issue week w (of season s),
- for epigroup g,
- for target t,
- with type m,

- with pancasting configuration b, and
- with forecaster f.

Retrospective component forecasts are prepared and evaluated in one of four modes:

- **Leave-one-season-out (LOSOCV) v1:**
  - **Component forecasts' training data:** all revisions of measurements made for other seasons
  - **Component forecasts' test/conditioning data:** $Y_{1..t}^{(t)}$ when available; missing revision data are filled in with the latest values as of when the analysis was run (i.e., with values from $Y_{1..t}^{(\text{analysis time})}$)
  - **Evaluation data:** the latest version of the data available ($Y_{1..\text{analysis time}}^{(\text{analysis time})}$)
  - **Ensemble training data:** LOSOCV v1 component forecasts for other seasons
  - **Ensemble selection data:** LOSOCV v1 ensemble forecasts for other seasons
- **Leave-one-season-out (LOSOCV) v2:**
  - **Component forecasts' training data:** all prior issues ($Y^{(1)}..Y^{(t)}$) plus the parts of issues from future seasons that do not contain measurements for the test season
  - **Component forecasts' test/conditioning data:** all recorded data from $Y^{(1)}..Y^{(t)}$, plus, wherever values of $Y_{1..t}^{(t)}$ are not available: nothing, if there is an older version available ($Y_u^{(v)}$, $v < t$), otherwise the earliest later version $Y_u^{(v)}$ with $v - u \leq 52$, otherwise the latest version as of when the analysis was run ($Y_u^{(\text{analysis time})}$)
  - **Evaluation data:** the latest version of the data available ($Y_{1..\text{analysis time}}^{(\text{analysis time})}$)
  - **Ensemble training data:** LOSOCV v2 component forecasts for other seasons
  - **Ensemble selection data:** LOSOCV v2 ensemble forecasts for other seasons
- **Pseudoprospective v1:**

    - **Component forecasts' training and test data:** same as LOSOCV v2 test/conditioning data.
    - **Evaluation data:** the latest version of the data available ($Y_{1..\text{analysis time}}^{(\text{analysis time})}$)
    - **Ensemble training data:** pseudoprospective v1 forecasts for prior seasons
    - **Ensemble selection data:** pseudoprospective v1 ensemble forecasts for prior seasons

- **Pseudoprospective v2:**
    - **Component forecasts' training and test data:** same as LOSOCV v2 test/conditioning data. all recorded data from $Y^{(1)}..Y^{(t)}$, plus, wherever values of $Y_{1..t}^{(t)}$ are not available: nothing, if there is an older version available ($Y_u^{(v)}$, $v < t$), otherwise the earliest later version $Y_u^{(v)}$
    - **Evaluation data:** the data as of issue week 28 immediately following the end of the test season, filling in any missing records in the same manner as missing values from $Y_{1..t}^{(t)}$ in the component forecasts' training and test data
    - **Ensemble training data:** pseudoprospective v1 or v2 forecasts for prior seasons
    - **Ensemble selection data:** pseudoprospective v2 ensemble forecasts for prior seasons

- **Hybrid LOSOCV-pseudoprospective v1:**
    - **Component forecasts' training data:** all prior issues ($Y^{(1)}..Y^{(t)}$) plus the parts of issues from future seasons up to some issue $I_{\text{LOSOCV end}}$ that do not contain measurements for the test season
    - **Component forecasts' test/conditioning data:** same as LOSOCV v1
    - **Evaluation data:** same as LOSOCV v1
    - **Ensemble training data:** hybrid LOSOCV-pseudoprospective v1 forecasts for other seasons up to some season $S_{\text{LOSOCV end}}$
    - **Ensemble selection data:** hybrid LOSOCV-pseudoprospective v1 ensemble forecasts for other seasons up to some season $S_{\text{LOSOCV end}}$

Ensemble performance statistics were prepared using LOSOCV v1, with training data from 2003/2004 to 2015/2016, excluding 2009/2010, and evaluation data from 2010/2011 to 2015/2016, epi weeks 44 to 17. Pancaster-forecaster pairs were analyzed with hybrid LOSOCV v2, with training data from 1997/1998 to 2018/2019, and evaluation data from 2010/2011 to 2018/2019, model weeks 40 to 73. The FluSight-Network ensemble is currently prepared using pseudoprospective v2 with evaluation data from 2010/2011 to 2018/2019, epi weeks 40 to 20; component forecasts from the Delphi-Stat system submitted to the network were prepared using hybrid LOSOCV-pseudoprospective v1 with training data from 2003/2004 to 2018/2019, and LOSOCV end issue 201039 and season 2009/2010.

Weighting schemes similar to the following are considered (see Chapter 5 for the ones used in the ensemble performance study):

- **Target&metric-based:** a different weightset is fit for each target and metric, based on all ensemble training data for that target and metric
- **Target&metric&3time-based:** a different weightset is fit for each target, metric, and week based on all ensemble training data for that target and metric, and model weeks within 1 week of the target model week
- **Target&metric&9time-based:** a different weightset is fit for each target, metric, and week based on all ensemble training data for that target and metric, and model weeks within 4 weeks of the target model week
- **Coherent log-score-9time-based:** a different weightset is fit for each week based on all ensemble training data for model weeks within 4 weeks of the target model week, based on log score evaluations (even for point predictions) — this scheme is "coherent" in that it uses the same weightset for all targets and types of predictions made at the same time.

The weighting scheme is selected based on the "ensemble selection data" described above. Thus, the general procedure of generating ensemble forecasts has the following steps:

1. Generate retrospective component forecasts for each s, w, g, t, m selected to generate the ensemble training data, using every pancaster-forecaster pair b, f.
2. Generate retrospective ensemble forecasts for the same values of s, w, g, t, m to generate the ensemble selection data, using every ensemble weighting scheme e.

3. Select the best ensemble weighting scheme based on the ensemble selection data $\hat{e}$.

4. Generate prospective component forecasts and combine them using weighting scheme $\hat{e}$.

D.2 Quantile pancasting framework

The quantile regression pancaster generates 200 simulated trajectories $Y^{\text{sim } 1}_{T_1+1..T_2}..Y^{\text{sim } 200}_{T_1+1..T_2}$ as follows:

1. For each node Y_u to simulate, in order:

   (a) Handle test data missingness: select covariates to consider in the model for Y_u using a Bayes net template (which specifies what actions to take when potential covariates are missing in test data, and allows for conditioning on previous simulations).

   (b) Translate test covariates and responses into node "characterizations": characterize each test covariate as $Source^{(u+lag)}_{u+shift}$, for some *Source*, *lag*, and *shift*.

   (c) Populate training data set by forming training instances (from the same location) with covariates and responses with characterizations matching those of the test instance; include all training instances with non-missing response data, even if training covariate data are missing. Assign training instances weights based on $\mathbf{\Phi}^{[\text{QARXkernel},u]}$ and the associated smoothing kernel — in all cases studied, just a boxcar kernel on the model week for each instance, which effectively just limits the training data to instances corresponding to model weeks within 4 weeks of the model week of the test instance.

   (d) Use the training covariate missingness handler and a quantile regression method to add simulated values of Y_u onto each simulation (i.e., to generate $Y_u^{\text{sim } 1..200}$).

These simulated trajectories $Y^{\text{sim } 1}_{T_1+1..T_2}..Y^{\text{sim } 200}_{T_1+1..T_2}$ are then fed into each forecaster to generate a point and distributional prediction for each target. For all forecasters but the uniform-distribution baseline, the distributional prediction $\mathbf{p}$ is used to form $\mathbf{p}' = \frac{M}{M+3}\mathbf{p} + \frac{3}{M+3}\mathbf{u}$, where $\mathbf{u}$ is the uniform distribution over bins, where M is the

number of simulations produced by the forecaster. The final distributional forecast for the pancaster-forecaster pair is $\mathbf{p}''$, a kernel-smoothed version of $\mathbf{p}'$ prepared by

1. pairing the average valid value for each bin with the corresponding "weight" from $(M+3)\mathbf{p}'$, and
2. calling a weighted version of `bw.nrd0` on this "weighted sample".

The purpose of this baseline-combination and smoothing step is to account for the fact that M simulations are just a noisy version of the true distribution associated with the given pancaster-forecaster pair, and to ensure that every pancaster-forecaster pair assigns nonzero probability to every possible value of every target, so that they can more readily be compared using non-thresholded log scores. Additional smoothing may be beneficial, as this step does not account for cases where the pancaster produces less simulations than the forecaster and requires resampling, nor for model misspecification, overconfidence, and overfitting; we rely on the ensemble method to account or compensate for some of these issues.

D.3 Handling missing data in quantile regression framework

Predictions for "response nodes" in the Bayes net template expansion — i.e., nodes with incoming arrows — are made under the assumption that they will be observed before or at evaluation time; "missing" is never a correct prediction. The training data selection algorithm accounts for this assumption by forming training instances only from times where the response variable is nonmissing. However, the candidate covariates described by the Bayes net template (the nodes with arrows pointing to a given response node) are allowed to be missing in both test and training instances. Missingness of candidate covariates in test instances is handled by the Bayes net template specification itself; template resolution for a given response node only selects covariates that are non-missing in the test instance. Missingness of the selected covariates in training instances is addressed by one of two handlers, "Thin" or "Drop", designed to build upon quantile regression routines that do not allow missingness or near-singularity in training data.

Algorithm 2 describes the "Thin" approach, which uses missingness indicator covariates (currently with no interaction covariates) and zero-filling to resolve missingness, and an SVD to detect and correct for near-singularity. Algorithm 3 describes

Algorithm 2: Fitting procedure using "Thin" missingness and near-singularity handler

Data:

$\mathbf{\Phi} \in (\mathbb{R} \cup \{\text{NA}\})^{n,p}$: selected covariates

$\boldsymbol{\psi} \in \mathbb{R}^n$: response values (nonmissing)

$\boldsymbol{\tau} \in [0,1]^m$: quantile levels

$\mathbf{w} \in \{\text{none}, \mathbb{R}_+^p\}$: instance weights (optional)

`dtolrelmaxconstant`: threshold determining when to drop left-singular and right-singular vectors in an SVD when fitting regression coefficients, used in a way to attempt to prevent (near-)singular matrix errors in `quantreg::rq` [Koenker, 2015] routines; default is 10^{-6}; lower than $10 \cdot \epsilon_{\text{machine}}$ likely risks errors

Result:

$\mathbf{B} \in \mathbb{R}^{1+p,m}$: fitted coefficient matrix including intercepts

Construct $\mathbf{M} \in \{0,1\}^{n,p}$ with $m_{ij} \coloneqq \begin{cases} 1, & \phi_{ij} \text{ is NA} \\ 0, & \text{otherwise} \end{cases}$;

Construct $\mathbf{C} \in \mathbb{R}^{n,p}$ with $c_{ij} \coloneqq \begin{cases} 0, & \phi_{ij} \text{ is NA} \\ \phi_{ij}, & \text{otherwise} \end{cases}$;

Let $\mathbf{\Phi}' \in \mathbb{R}^{n,1+2p} \coloneqq [\mathbf{1}\ \mathbf{M}\ \mathbf{C}]$;

Let $\mathbf{U}'\mathbf{D}'\mathbf{V}'^T \coloneqq \mathbf{\Phi}'$ be an SVD (with $\mathbf{U}'$ an $n \times (1+2p)$ orthogonal matrix and such that $\mathbf{D}$ is a $(1+2p) \times (1+2p)$ (diagonal) matrix with all non-negative entries);

Let $\texttt{dtolrelmax} \coloneqq \max\{n, 1+2p\} \cdot \texttt{dtolrelmaxconstant}$;

Set $\mathbf{U}'' \coloneqq \mathbf{U}_{\cdot,\mathbf{k}}$, $\mathbf{D}'' \coloneqq \mathbf{D}'_{\mathbf{k},\mathbf{k}}$, and $\mathbf{V}'' \coloneqq \mathbf{V}'_{\mathbf{k},\cdot}$ to be a thin SVD corresponding to singular values — previously at indices $\mathbf{k}$ — greater than or equal to $\texttt{dtolrelmax} \cdot \max_l \mathbf{D}'_{l,l}$;

Let $\mathbf{\Phi}''' \coloneqq \mathbf{U}''\mathbf{D}''$;

Let $\mathbf{A} \in \mathbb{R}^{1+2p,m}$ be the coefficient matrix obtained from a (potentially weighted) quantile regression routine on $\mathbf{\Phi}'''$, $\boldsymbol{\psi}$, $\boldsymbol{\tau}$, $\mathbf{w}$;

Let $\mathbf{B}^{\text{full}} \coloneqq \mathbf{V}''\mathbf{A}$;

Let $\mathbf{B} \coloneqq \mathbf{B}^{\text{full}}_{\{1..1+p\},\cdot}$ (dropping missingness indicator coefficients that will be multiplied by zero in test instance);

Return $\mathbf{B}$.

the "Drop" approach, which employs a heuristic selection algorithm to select a set of covariates that appear together without missingness in at least a certain number or proportion of training instances, favoring selection of covariates that appear first in the training data matrix; it uses a QR decomposition with pivoting to remove some near-singularities, injects noise in training data to attempt to remove others, and utilizes fallback quantile regression algorithms to address any near-singularity errors that still occur. As noted in Chapter 4, it appears to have higher performance than "Thin" for covariate selections where it consistently avoids errors from the quantile regression routines, but as it sometimes fails to avoid all near-singular matrix issues, it is not as operationally robust.

D.4 Delta density forecasters

Both delta density forecasters share the same first few pre-processing steps, outlined below: When preparing Gaussian kernel density estimates for covariates or response variables, both variants share the same method to select bandwidths: they try the `bw.SJ` [R Core Team, 2015] method on the training data for that covariate or response variable, falling back on `bw.nrd0` [R Core Team, 2015] in the case of errors. (This approach is computationally convenient, but is unlikely to be statistically optimal.)

Algorithm 5 descirbes the Markovian delta density method, which performs conditioning effectively using the product of (a) a zero-width boxcar kernel on the week of the season, and (b) a Gaussian kernel over the previous measurement. Section 2.3 details the differences between the Markovian delta density method and the extended delta density variant.

Algorithm 3: Fitting procedure using "Drop" missingness and near-singularity handler

Data:

$\mathbf{\Phi} \in (\mathbb{R} \cup \{\text{NA}\})^{n,p}$: selected covariates

$\boldsymbol{\psi} \in \mathbb{R}^n$: response values (nonmissing)

$\boldsymbol{\tau} \in [0,1]^m$: quantile levels

$\mathbf{w} \in \{\text{none}, \mathbb{R}_+^p\}$: instance weights (optional)

`tol`: tolerance, e.g, 10^{-3}, for near-singularities in the QR decomposition

`relsigma`: scaling factor, e.g, 10^{-3}, for standard deviation when jittering training data

Result:

$\mathbf{B} \in \mathbb{R}^{1+p,m}$: fitted coefficient matrix including intercepts (or, for certain covariate selections, near-singularity errors that were not successfully avoided or addressed)

Let `min.nrow` $\coloneqq \max\{10, n/10\}$;

Initialize mutable $\mathbf{\Phi'} \leftarrow \mathbf{\Phi}$;

for $j \in \{1..p\}$, *sequentially,* **do**

  if *current* $\mathbf{\Phi'}_{\cdot,j}$ *has at least* ***`min.nrow`*** *nonmissing values* **then**

    Set $\mathbf{\Phi'} \leftarrow \mathbf{\Phi'}_{\mathbf{i},\cdot}$ where $\mathbf{i}$ are the indices of the nonmissing $\mathbf{\Phi'}_{\cdot,j}$ entries

  else

    Set $\mathbf{\Phi'} \leftarrow \mathbf{\Phi'}_{\cdot,-j}$ (i.e., drop column j from $\mathbf{\Phi'}$), maintaining the same column indices for columns $j+1..p$ rather than shifting them

  end

end

Let $\mathbf{\Phi''} \coloneqq \mathbf{\Phi'}_{\cdot,\mathbf{j}}$, where $\mathbf{j}$ are the indices of features that are not dropped linear regression of $\boldsymbol{\psi}$ on $\mathbf{\Phi'}$ with an intercept, where the linear regression utilizes QR decomposition with pivoting with collinearity tolerance `tol` ;

Let $\mathbf{\Phi'''} \coloneqq \mathbf{\Phi''} + \mathbf{E}$, where $\mathbf{E}$ is a "jitter" matrix of independent draws from Gaussian noise variables, with $\mathbf{E}_{i,j} \sim \mathcal{N}(0, \hat{\sigma}_j)$, where $\hat{\sigma}_j$ is the sample standard deviation of $\mathbf{\Phi''}_{\cdot,j}$;

Let $\mathbf{A} \in \mathbb{R}^{1+p''',m}$ be the coefficient matrix obtained from a (potentially weighted) quantile regression routine on $[\mathbf{1} \;\; \mathbf{\Phi'''}]$, $\boldsymbol{\psi}$, $\boldsymbol{\tau}$, $\mathbf{w}$, where the routine tries: (a) a lasso-based fit if desired, then (b) a non-lasso fit using the Frisch-Newton algorithm if the lasso routine failed due to unhandled near-singularities, aborting if the latter fails as well;

Construct $\mathbf{B} \in \mathbb{R}^{1+p,m}$ with $\mathbf{B}_{1,\cdot} \coloneqq \mathbf{A}_{1,\cdot}$ and

$$\mathbf{B}_{1+j,\cdot} \coloneqq \begin{cases} \mathbf{A}_{1+j,\cdot}, & \text{column } j \text{ from } \mathbf{\Phi} \text{ was not dropped when forming } \mathbf{\Phi'} \\ \mathbf{0}, & \text{otherwise} \end{cases};$$

Return $\mathbf{B}$.

Algorithm 4: Shared delta density setup

Data:
$Y^{\text{input sim } 1..n_{\text{input}}}_{T_1+1..T_2}$: partially observed and/or simulated trajectories from a backcaster or pancaster
$\mathbf{w}^{\text{input}} \in \mathbb{R}_+^{n_{\text{input}}}$: importance weights for the input simulations
T_2: time of last observation already observed or simulated
T_3: time of last observation to simulate
$\mathcal{D}$: set of time-shifted training trajectories covering roughly $T_1 + 1$ to T_3
n_{output}: number of fully simulated trajectories to produce
Result:
$Y^{\text{output sim } 1..n_{\text{output}}}_{T_1+1..T_2}$: potentially resampled version of $Y^{\text{input sim } 1..n_{\text{input}}}$, to be extended into fully simulated trajectories
$\mathbf{w}^{\text{output}} \in \mathbb{R}_+^{n_{\text{output}}}$: corresponding trajectory importance weights
if $n_{input} = n_{output}$ **then**
  Let $Y^{\text{output sim } 1..n_{\text{output}}}_{T_1+1..T_2} \coloneqq Y^{\text{output sim } 1..n_{\text{input}}}_{T_1+1..T_2}$;
  Let $\mathbf{w}^{\text{output}} = \mathbf{w}^{\text{input}}$
else
  Let $Y^{\text{output sim } 1..n_{\text{output}}}_{T_1+1..T_2}$ be a resampling of $Y^{\text{output sim } 1..n_{\text{input}}}_{T_1+1..T_2}$ using weights $\mathbf{w}^{\text{input}}$;
  Let $\mathbf{w}^{\text{output}} \in \mathbb{R}_+^{n_{\text{output}}} \coloneqq c\mathbf{1}$, where c is the mean of $\mathbf{w}^{\text{input}}$ if downsampling or $\mathbf{w}^{\text{input}} \cdot \mathbf{1}/n_{\text{output}}$ if upsampling
end
Let $\mathbf{D}$ be a matrix of training trajectories formed from $\mathcal{D}$ by dropping times which are not observed in every training trajectory in $\mathcal{D}$ (e.g., clipping off the 53rd week in some trajectories when working with year-long trajectories)

Algorithm 5: Markovian delta density

Data:
$Y_{T_1+1..T_2}^{\text{input sim } 1..n_{\text{input}}}$: partially observed and/or simulated trajectories from a backcaster or pancaster
$\mathbf{w}^{\text{input}} \in \mathbb{R}_+^{n_{\text{input}}}$: importance weights for the input simulations
T_2: time of last observation already observed or simulated
T_3: time of last observation to simulate
$\mathcal{D}$: set of time-shifted training trajectories covering roughly $T_1 + 1$ to T_3
n_{output}: number of fully simulated trajectories to produce
Result:
$Y_{T_1+1..T_3}^{\text{output sim } 1..n_{\text{output}}}$: fully simulated trajectories
$\mathbf{w}^{\text{output}} \in \mathbb{R}_+^{n_{\text{output}}}$: corresponding trajectory importance weights
Let $Y_{T_1+1..T_2}^{\text{output sim } 1..n_{\text{output}}}$, $\mathbf{w}^{\text{output}} \in \mathbb{R}_+^{n_{\text{output}}}$, and $\mathbf{D}$ be given by the shared delta density setup;
for $u \in \{T_2 + 1..T_3\}$, *sequentially,* **do**
  Let u' be the closest time index to u such that u' and $u' - 1$ are reported in $\mathbf{D}$, breaking any ties arbitrarily (expectation: $u' = u$ except maybe when $u = T_3$, and no ties occur);
  Draw $\Delta Y_u^{\text{sim } 1..n_{\text{output}}}$ from a Gaussian kernel conditional density estimate of ΔY based on portions of $\mathbf{D}$ at time u';
  Let $Y_u^{\text{output sim } 1..n_{\text{output}}} = Y_{u-1}^{\text{output sim } 1..n_{\text{output}}} + \Delta Y_u^{\text{sim } 1..n_{\text{output}}}$
end

Bibliography

H. Achrekar, A. Gandhe, R. Lazarus, Ssu-Hsin Yu, and B. Liu. Predicting Flu Trends using Twitter data. In *2011 IEEE Conference on Computer Communications Workshops (INFOCOM WKSHPS)*, pages 702–707, April 2011. doi: 10.1109/INFCOMW.2011.5928903. 4.1

Allison E Aiello, Rebecca M Coulborn, Vanessa Perez, and Elaine L Larson. Effect of hand hygiene on infectious disease risk in the community setting: a meta-analysis. *American journal of public health*, 98(8):1372–1381, 2008. 1.0.1

Alnur Ali, J Zico Kolter, and Ryan J Tibshirani. The multiple quantile graphical model. In *Advances in Neural Information Processing Systems*, pages 3747–3755, 2016. d, 3.3.2

Ozgur M. Araz, Dan Bentley, and Robert L. Muelleman. Using Google Flu Trends data in forecasting influenza-like–illness related ED visits in Omaha, Nebraska. *The American Journal of Emergency Medicine*, 32(9):1016–1023, September 2014. ISSN 0735-6757, 1532-8171. doi: 10.1016/j.ajem.2014.05.052. URL http://www.ajemjournal.com/article/S0735-6757(14)00421-5/fulltext. 4.1

Taylor Arnold, Veeranjaneyulu Sadhanala, and Ryan Tibshirani. *glmgen: Fast algorithms for generalized lasso problems*, 2014. R package version 0.0.3. B.1.2

Taylor B. Arnold and Ryan J. Tibshirani. *genlasso: Path algorithm for generalized lasso problems*, 2014. URL http://CRAN.R-project.org/package=genlasso. R package version 1.3. 2.2, 2.1

S Borağan Aruoba. Data revisions are not well behaved. *Journal of money, credit and banking*, 40(2-3):319–340, 2008. 3.1

Michael Baake and Ulrike Schlaegel. The peano-baker series. *Proceedings of the Steklov Institute of Mathematics*, 275(1):155–159, 2011. b

Michel Berkelaar and others. *lpSolve: Interface to 'Lp_solve' v. 5.5 to Solve Linear/Integer Programs*, 2015. URL http://CRAN.R-project.org/package=lpSolve. R package version 5.6.11. 5.2, C.1

Matthew Biggerstaff, David Alper, Mark Dredze, Spencer Fox, Isaac Chun-Hai Fung, Kyle S. Hickmann, Bryan Lewis, Roni Rosenfeld, Jeffrey Shaman, Ming-Hsiang Tsou, Paola Velardi, Alessandro Vespignani, and Lyn Finelli. Results from the centers for disease control and prevention's predict the 2013–2014 Influenza Season Challenge. *BMC Infectious Diseases*, 16:357, 2016. ISSN 1471-2334. doi: 10.1186/s12879-016-1669-x. 1.0.1, 1.1, 1.3.1

Matthew Biggerstaff, Michael Johansson, David Alper, Logan C Brooks, Prithwish Chakraborty, David C Farrow, Sangwon Hyun, Sasikiran Kandula, Craig McGowan, Naren Ramakrishnan, et al. Results from the second year of a collaborative effort to forecast influenza seasons in the united states. *Epidemics*, 2018. 1.0.1, 1.1, 1.3.1

Byron Boots. Spectral approaches to learning predictive representations. Technical report, CARNEGIE-MELLON UNIV PITTSBURGH PA SCHOOL OF COMPUTER SCIENCE, 2012. a

Stephen Boyd and Lieven Vandenberghe. *Convex optimization*. Cambridge university press, 2004. 2.4.2

Lynnette Brammer, Alicia P. Budd, and Lyn Finelli. *Seasonal and pandemic influenza surveillance*, chapter 12, pages 200–210. John Wiley & Sons Ltd, 2013. ISBN 9781118543504. doi: 10.1002/9781118543504.ch16. URL http://dx.doi.org/10.1002/9781118543504.ch16. 1.3.1

Leo Breiman. Stacked regressions. *Machine learning*, 24(1):49–64, 1996. 5.2

David A. Broniatowski, Michael J. Paul, and Mark Dredze. National and Local Influenza Surveillance through Twitter: An Analysis of the 2012-2013 Influenza Epidemic. *PLOS ONE*, 8(12):e83672, December 2013. ISSN 1932-6203. doi: 10.1371/journal.pone.0083672. URL http://journals.plos.org/plosone/article?id=10.1371/journal.pone.0083672. 4.1

Logan C Brooks, David C Farrow, Sangwon Hyun, Ryan J Tibshirani, and Roni Rosenfeld. Flexible modeling of epidemics with an empirical Bayes framework. *PLoS Computational Biology*, 11(8):e1004382, 2015a. 1.1, 2.2, B.1.2, C.1

Logan C. Brooks, David C. Farrow, Sangwon Hyun, Ryan J. Tibshirani, and Roni Rosenfeld. *epiforecast: Tools for forecasting semi-regular seasonal epidemic curves and similar time series*, 2015b. R package version 0.0.1. B, D

Logan C Brooks, David C Farrow, Sangwon Hyun, Ryan J Tibshirani, and Roni Rosenfeld. Nonmechanistic forecasts of seasonal influenza with iterative one-week-ahead distributions. *PLoS computational biology*, 14(6):e1006134, 2018. 1, 1.3.1, 1.4, 2.3, 2.5, 3.1, 3.3.1, 4.1, 4.2, 5, A, B, C

Elizabeth Buckingham-Jeffery, Valerie Isham, and Thomas House. Gaussian process approximations for fast inference from infectious disease data. *Mathematical biosciences*, 2018. 1.1

Carnegie Mellon University Delphi group. Delphi. https://github.com/cmu-delphi, Accessed 2017-04-26. B

Bob Carpenter, Andrew Gelman, Matthew D Hoffman, Daniel Lee, Ben Goodrich, Michael Betancourt, Marcus Brubaker, Jiqiang Guo, Peter Li, and Allen Riddell. Stan: A probabilistic programming language. *Journal of Statistical Software*, 76 (1), 2017. 1.1

Centers for Disease Control and Prevention. Overview of influenza surveillance in the united states, 2013. URL https://www.cdc.gov/flu/weekly/overview.htm. 1.0.1, 1.3.1, 1.3.1

Centers for Disease Control and Prevention, National Center for Immunization and Respiratory Diseases (NCIRD). Estimating seasonal influenza-associated deaths in the United States | seasonal influenza (flu) |CDC, 2016a. URL https://www.cdc.gov/flu/about/disease/us_flu-related_deaths.htm. 1.0.1

Centers for Disease Control and Prevention, National Center for Immunization and Respiratory Diseases (NCIRD). Pandemic Basics | Pandemic Influenza (Flu) | CDC. https://www.cdc.gov/flu/pandemic-resources/basics/index.html, 2016b. 1.0.1

Centers for Disease Control and Prevention, National Center for Immunization and Respiratory Diseases (NCIRD). FluView Interactive | Seasonal Influenza (Flu) | CDC. https://www.cdc.gov/flu/weekly/fluviewinteractive.htm, 2017a. 1.3.1

Centers for Disease Control and Prevention, National Center for Immunization and Respiratory Diseases (NCIRD). Pandemic Influenza | Pandemic Influenza (Flu) | CDC. https://www.cdc.gov/flu/pandemic-resources/, 2017b. 1.0.1

Prithwish Chakraborty, Pejman Khadivi, Bryan Lewis, Aravindan Mahendiran, Jiangzhuo Chen, Patrick Butler, Elaine O Nsoesie, Sumiko R Mekaru, John S Brownstein, Madhav V Marathe, et al. Forecasting a moving target: Ensemble models for ili case count predictions. In *Proceedings of the 2014 SIAM international conference on data mining*, pages 262–270. SIAM, 2014. 1.1

Jean-Paul Chretien, Dylan George, Jeffrey Shaman, Rohit A Chitale, and F Ellis McKenzie. Influenza forecasting in human populations: a scoping review. *PloS one*, 9(4):e94130, 2014. 1.1

Benjamin J Cowling, Kwok-Hung Chan, Vicky J Fang, Calvin KY Cheng, Rita OP Fung, Winnie Wai, Joey Sin, Wing Hong Seto, Raymond Yung, Daniel WS Chu, et al. Facemasks and hand hygiene to prevent influenza transmission in households: a cluster randomized trial. *Annals of internal medicine*, 151(7):437–446, 2009. 1.0.1

Aron Culotta. Towards Detecting Influenza Epidemics by Analyzing Twitter Messages. In *Proceedings of the First Workshop on Social Media Analytics*, SOMA '10, pages 115–122, New York, NY, USA, 2010. ACM. ISBN 978-1-4503-0217-3. doi: 10.1145/1964858.1964874. URL http://doi.acm.org/10.1145/1964858.1964874. 4.1

Suruchi Deodhar, Jiangzhuo Chen, Mandy Wilson, Manikandan Soundarapandian, Keith Bisset, Bryan Lewis, Chris Barrett, and Madhav Marathe. Flu caster: A pervasive web application for high resolution situation assessment and forecasting of flu outbreaks. In *Healthcare Informatics (ICHI), 2015 International Conference on*, pages 105–114. IEEE, 2015. 1.1

Mark Dredze, Renyuan Cheng, Michael J. Paul, and David Broniatowski. HealthTweets.org: A Platform for Public Health Surveillance Using Twitter. In *Workshops at the Twenty-Eighth AAAI Conference on Artificial Intelligence*, pages 593–596, June 2014. URL https://www.aaai.org/ocs/index.php/WS/AAAIW14/paper/view/8723. 1.2, 4.1

Andrea Freyer Dugas, Mehdi Jalalpour, Yulia Gel, Scott Levin, Fred Torcaso, Takeru Igusa, and Richard E. Rothman. Influenza Forecasting with Google Flu Trends. *PLOS ONE*, 8(2):e56176, February 2013. ISSN 1932-6203. doi: 10.1371/

journal.pone.0056176. URL http://journals.plos.org/plosone/article?id=10.1371/journal.pone.0056176. 4.1

Charbel El Bcheraoui, Ali H Mokdad, Laura Dwyer-Lindgren, Amelia Bertozzi-Villa, Rebecca W Stubbs, Chloe Morozoff, Shreya Shirude, Mohsen Naghavi, and Christopher JL Murray. Trends and patterns of differences in infectious disease mortality among us counties, 1980-2014. *JAMA*, 319(12):1248–1260, 2018. 1.0.1

Epidemic Prediction Initiative. Forecast evaluation, March 2016. URL https://predict.phiresearchlab.org/legacy/flu/evaluation.html. 1.4

Gunther Eysenbach. Infodemiology: Tracking Flu-Related Searches on the Web for Syndromic Surveillance. *AMIA Annual Symposium Proceedings*, 2006:244–248, 2006. ISSN 1942-597X. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1839505/. 4.1

David C Farrow. *Modeling the Past, Present, and Future of Influenza*. Phd thesis, Carnegie Mellon University, 2016. URL http://reports-archive.adm.cs.cmu.edu/anon/cbd/CMU-CB-16-101.pdf. 1.2, 4.1, 5.3.2

David C. Farrow. EpiVis. http://delphi.midas.cs.cmu.edu/epivis/epivis.html, Accessed 2017-04-26. 1.2, 1.3.1

David C Farrow, Logan C Brooks, Sangwon Hyun, Ryan J Tibshirani, Donald S Burke, and Roni Rosenfeld. A human judgment approach to epidemiological forecasting. *PLOS Computational Biology*, 13(3):e1005248, 2017. 5.3.2

David C Farrow, Maria Jahja, Roni Rosenfeld, and Ryan J Tibshirani. Kalman filter, sensor fusion, and constrained regression: Equivalences and insights. *arXiv preprint arXiv:1905.11436*, 2019. 4.1

Emily Fox and David Dunson. Bayesian nonparametric covariance regression. *arXiv preprint arXiv:1101.2017*, 2011. 2.2

Jerome Friedman, Trevor Hastie, and Robert Tibshirani. Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software*, 33(1):1–22, 2010. URL http://www.jstatsoft.org/v33/i01/. B.1.3, 2

Nicholas Generous, Geoffrey Fairchild, Alina Deshpande, Sara Y. Del Valle, and Reid Priedhorsky. Global Disease Monitoring and Forecasting with Wikipedia.

PLOS Computational Biology, 10(11):e1003892, November 2014. ISSN 1553-7358. doi: 10.1371/journal.pcbi.1003892. URL http://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1003892. 1.1, 4.1

Jeremy Ginsberg, Matthew H. Mohebbi, Rajan S. Patel, Lynnette Brammer, Mark S. Smolinski, and Larry Brilliant. Detecting influenza epidemics using search engine query data. *Nature*, 457(7232):1012–1014, February 2009. ISSN 1476-4687. doi: 10.1038/nature07634. URL https://www.nature.com/articles/nature07634. 1.2, 4.1

Bruce E Hansen. Nonparametric conditional density estimation. *Unpublished manuscript*, 2004. 2.3

Andrew C Hayward, Ellen B Fragaszy, Alison Bermingham, Lili Wang, Andrew Copas, W John Edmunds, Neil Ferguson, Nilu Goonetilleke, Gabrielle Harvey, Jana Kovar, et al. Comparative community burden and severity of seasonal and pandemic influenza: results of the flu watch cohort study. *The Lancet Respiratory Medicine*, 2(6):445–454, 2014. 1.3.1

Jane M Heffernan, Robert J Smith, and Lindi M Wahl. Perspectives on the basic reproductive ratio. *Journal of the Royal Society Interface*, 2(4):281–293, 2005. 2.4.2

Arlo D Hendrickson and Robert J Buehler. Proper scores for probability forecasters. *The Annals of Mathematical Statistics*, pages 1916–1921, 1971. 1.4

Herbert W Hethcote and David W Tudor. Integral equation models for endemic infectious diseases. *Journal of mathematical biology*, 9(1):37–47, 1980. 2.4.2

Kyle S. Hickmann, Geoffrey Fairchild, Reid Priedhorsky, Nicholas Generous, James M. Hyman, Alina Deshpande, and Sara Y. Del Valle. Forecasting the 2013–2014 Influenza Season Using Wikipedia. *PLOS Computational Biology*, 11(5):e1004239, May 2015. ISSN 1553-7358. doi: 10.1371/journal.pcbi.1004239. URL http://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1004239. 1.1, 1.2, 4.1

Daniel Hsu, Sham M Kakade, and Tong Zhang. A spectral algorithm for learning hidden markov models. *Journal of Computer and System Sciences*, 78(5):1460–1480, 2012. a

Anette Hulth, Gustaf Rydevik, and Annika Linde. Web Queries as a Source for Syndromic Surveillance. *PLOS ONE*, 4(2):e4378, February 2009. ISSN 1932-6203. doi: 10.1371/journal.pone.0004378. URL http://journals.plos.org/plosone/article?id=10.1371/journal.pone.0004378. 4.1

Rob J Hyndman and Yanan Fan. Sample quantiles in statistical packages. *The American Statistician*, 50(4):361–365, 1996. 2.4.1

Michael Höhle, Sebastian Meyer, and Michaela Paul. *surveillance: Temporal and Spatio-Temporal Modeling and Monitoring of Epidemic Phenomena*, 2017. URL https://CRAN.R-project.org/package=surveillance. R package version 1.13.1. 1.1

Edward L Ionides, C Bretó, and Aaron A King. Inference for nonlinear dynamical systems. *Proceedings of the National Academy of Sciences*, 103(49):18438–18443, 2006. 1.1

Edward L Ionides, Dao Nguyen, Yves Atchadé, Stilian Stoev, and Aaron A King. Inference for dynamic and latent variable models via iterated, perturbed bayes maps. *Proceedings of the National Academy of Sciences*, 112(3):719–724, 2015. 1.1

Jan PAM Jacobs and Simon Van Norden. Modeling data revisions: Measurement error and dynamics of "true" values. *Journal of Econometrics*, 161(2):101–109, 2011. 3.1

Maria Jahja, Ryan J Tibshirani, and Matthew Biggerstaff. Investigating ground truth measures of seasonal influenza via digital surveillance nowcasting. 2018. 1.2

Michael S Johannes, Nick Polson, and Seung M Yae. Quantile filtering and learning. 2009. d

Michael A. Johansson, Nicholas G. Reich, Aditi Hota, John S. Brownstein, and Mauricio Santillana. Evaluating the performance of infectious disease forecasts: A comparison of climate-driven and seasonal dengue forecasts for Mexico. *Scientific Reports*, 6:33707, September 2016. ISSN 2045-2322. doi: 10.1038/srep33707. URL https://www.nature.com/articles/srep33707. 1.1, 4.1

Niall PAS Johnson and Juergen Mueller. Updating the accounts: global mortality of the 1918-1920" spanish" influenza pandemic. *Bulletin of the History of Medicine*, pages 105–115, 2002. 1.0.1

Borko D Jovanovic and Paul S Levy. A look at the rule of three. *The American Statistician*, 51(2):137–139, 1997. 5.2

Juan Manuel Julio et al. Modeling data revisions. Technical Report 641, Subgerencia de Estudios Económicos del Banco de la República, 2011. In *Borrador de Economía*. 3.1

Sasikiran Kandula, Wan Yang, and Jeffrey Shaman. Type-and subtype-specific influenza forecast. *American journal of epidemiology*, page 1, 2017. 1.1

Aaron A King, Matthieu Domenech de Celles, Felicia MG Magpantay, and Pejman Rohani. Avoidable errors in the modelling of outbreaks of emerging pathogens, with special reference to ebola. *Proc. R. Soc. B*, 282(1806):20150347, 2015. 1.1

Roger Koenker. *quantreg: Quantile Regression*, 2015. URL `http://CRAN.R-project.org/package=quantreg`. R package version 5.11. 2

Roger Koenker and Zhijie Xiao. Quantile autoregression. *Journal of the American Statistical Association*, 101(475):980–990, 2006. 2.4

Adam J Kucharski, Viggo Andreasen, and Julia R Gog. Capturing the dynamics of pathogens with many strains. *Journal of mathematical biology*, 72(1-2):1–24, 2016. 2.4.2

Robert Kyeyagalire, Stefano Tempia, Adam L Cohen, Adrian D Smith, Johanna M McAnerney, Veerle Dermaux-Msimang, and Cheryl Cohen. Hospitalizations associated with influenza and respiratory syncytial virus among patients attending a network of private hospitals in south africa, 2007–2012. *BMC infectious diseases*, 14(1):694, 2014. 1.0.1

Vasileios Lampos, Andrew C. Miller, Steve Crossan, and Christian Stefansen. Advances in nowcasting influenza-like illness rates using search query logs. *Scientific Reports*, 5:12760, August 2015. ISSN 2045-2322. doi: 10.1038/srep12760. URL `https://www.nature.com/articles/srep12760`. 1.1, 4.1

Erich Leo Lehmann and George Casella. *Theory of point estimation*, volume 31. Springer Science & Business Media, 1998. 2.2

Nancy H Leung, Cuiling Xu, Dennis K Ip, and Benjamin J Cowling. Review article: The fraction of influenza virus infections that are asymptomatic: A systematic review and meta-analysis., 2015. 1.3.1

Erik Lindström, Edward Ionides, Jan Frydendall, and Henrik Madsen. Efficient iterated filtering. *IFAC Proceedings Volumes*, 45(16):1785–1790, 2012. 1.1

Jun S Liu. *Monte Carlo strategies in scientific computing*. Springer Science & Business Media, 2008. 2.2, B.1.2

Rachel Lowe, Trevor C Bailey, David B Stephenson, Tim E Jupp, Richard J Graham, Christovam Barcellos, and Marilia Sá Carvalho. The development of an early warning system for climate-sensitive disease risk with a focus on dengue epidemics in southeast brazil. *Statistics in medicine*, 32(5):864–883, 2013. 1.1

Rafael Lozano, Mohsen Naghavi, Kyle Foreman, Stephen Lim, Kenji Shibuya, Victor Aboyans, Jerry Abraham, Timothy Adair, Rakesh Aggarwal, Stephanie Y Ahn, et al. Global and regional mortality from 235 causes of death for 20 age groups in 1990 and 2010: a systematic analysis for the global burden of disease study 2010. *The lancet*, 380(9859):2095–2128, 2012. 1.0.1

N Gregory Mankiw and Matthew D Shapiro. News or noise? an analysis of gnp revisions, 1986. 3.1

Edson Zangiacomi Martinez, Elisângela Aparecida Soares da Silva, and Amaury Lelis Dal Fabbro. A sarima forecasting model to predict the number of cases of dengue in campinas, state of são paulo, brazil. *Revista da Sociedade Brasileira de Medicina Tropical*, 44(4):436–440, 2011. 1.1

Craig J. McGowan, Matthew Biggerstaff, Michael Johansson, Karyn M. Apfeldorf, Michal Ben-Nun, Logan Brooks, Matteo Convertino, Madhav Erraguntla, David C. Farrow, John Freeze, Saurav Ghosh, Sangwon Hyun, Sasikiran Kandula, Joceline Lega, Yang Liu, Nicholas Michaud, Haruka Morita, Jarad Niemi, Naren Ramakrishnan, Evan L. Ray, Nicholas G. Reich, Pete Riley, Jeffrey Shaman, Ryan Tibshirani, Alessandro Vespignani, Qian Zhang, Carrie Reed, and The Influenza Forecasting Working Group. Collaborative efforts to forecast seasonal influenza in the united states, 2015–2016. *Scientific reports*, 9(1):1–13, 2019. 5.1

David J. McIver and John S. Brownstein. Wikipedia Usage Estimates Prevalence of Influenza-Like Illness in the United States in Near Real-Time. *PLOS Computational Biology*, 10(4):e1003581, April 2014. ISSN 1553-7358. doi: 10.1371/journal.pcbi.1003581. URL http://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1003581. 4.1

Aaron C Miller, Inder Singh, Erin Koehler, and Philip M Polgreen. A smartphone-driven thermometer application for real-time population-and individual-level influenza surveillance. *Clinical Infectious Diseases*, 67(3):388–397, 2018. 1.2

Jeffrey J Morgan, Otto C Wilson, and Prahlad G Menon. The wisdom of crowds approach to influenza-rate forecasting. In *ASME 2018 International Mechanical Engineering Congress and Exposition*. American Society of Mechanical Engineers Digital Collection, 2018. 5.1, 5.3.2

Christopher JL Murray, Theo Vos, Rafael Lozano, Mohsen Naghavi, Abraham D Flaxman, Catherine Michaud, Majid Ezzati, Kenji Shibuya, Joshua A Salomon, Safa Abdalla, et al. Disability-adjusted life years (dalys) for 291 diseases and injuries in 21 regions, 1990–2010: a systematic analysis for the global burden of disease study 2010. *The lancet*, 380(9859):2197–2223, 2012. (document), 1.0.1

Richard W Niska and Iris Shimizu. Hospital preparedness for emergency response: United states, 2008. Technical Report 37, US Department of Health and Human Services, Centers for Disease Control and Prevention, National Center for Health Statistics, 2011. 1.0.1

Elaine O Nsoesie, John S Brownstein, Naren Ramakrishnan, and Madhav V Marathe. A systematic review of studies on forecasting the dynamics of influenza outbreaks. *Influenza and other respiratory viruses*, 8(3):309–316, 2014. 1.1

Philip D O'Neill. Introduction and snapshot review: relating infectious disease transmission models to data. *Statistics in medicine*, 29(20):2069–2077, 2010. 1.1

Dave Osthus and Kelly R Moran. Multiscale influenza forecasting. *arXiv preprint arXiv:1909.13766*, 2019. 1.1

Dave Osthus, James Gattiker, Reid Priedhorsky, and Sara Y Del Valle. Dynamic bayesian influenza forecasting in the united states with hierarchical discrepancy (with discussion). *Bayesian Analysis*, 14(1):261–312, 2019. 1.1

Michael J Paul, Mark Dredze, and David Broniatowski. Twitter improves influenza forecasting. *PLOS Currents Outbreaks*, 2014a. 1.1

Michael J. Paul, Mark Dredze, and David Broniatowski. Twitter Improves Influenza Forecasting. *PLoS Currents*, 6, October 2014b. ISSN 2157-3999. doi: 10.1371/currents.outbreaks.90b9ed0f59bae4ccaa683a39865d9117. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4234396/. 4.1

Sen Pei and Jeffrey Shaman. Counteracting structural errors in ensemble forecast of influenza outbreaks. *Nature communications*, 8(1):925, 2017. 1.1

Sen Pei, Sasikiran Kandula, Wan Yang, and Jeffrey Shaman. Forecasting the spatial transmission of influenza in the united states. *Proceedings of the National Academy of Sciences*, page 201708856, 2018. 1.1

Martyn Plummer. Jags: A program for analysis of bayesian graphical models using gibbs sampling. In *Proceedings of the 3rd International Workshop on Distributed Statistical Computing (DSC 2003)*, Vienna, Austria, 2003. 1.1

Philip M. Polgreen, Yiling Chen, David M. Pennock, Forrest D. Nelson, and Robert A. Weinstein. Using Internet Searches for Influenza Surveillance. *Clinical Infectious Diseases*, 47(11):1443–1448, December 2008. ISSN 1058-4838. doi: 10.1086/593098. URL https://academic.oup.com/cid/article/47/11/1443/282247. 4.1

Tobias Preis and Helen Susannah Moat. Adaptive nowcasting of influenza outbreaks using Google searches. *Royal Society Open Science*, 1(2):140095, October 2014. ISSN 2054-5703. doi: 10.1098/rsos.140095. URL http://rsos.royalsocietypublishing.org/content/1/2/140095. 4.1

R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2015. URL https://www.R-project.org/. 2.1, 2.3, B, D.4

Tamer Rabie and Valerie Curtis. Handwashing and risk of respiratory infections: a quantitative systematic review. *Tropical medicine & international health*, 11(3): 258–267, 2006. 1.0.1

Evan L Ray and Nicholas G Reich. Prediction of infectious disease epidemics via weighted density ensembles. *arXiv preprint arXiv:1703.10936*, 2017. 5.1, 5.2

Evan L Ray, Krzysztof Sakrejda, Stephen A Lauer, Michael A Johansson, and Nicholas G Reich. Infectious disease prediction with kernel conditional density estimation. *Statistics in medicine*, 36(30):4908–4929, 2017. 1.1, 2.3

Nicholas G Reich, Logan C Brooks, Spencer J Fox, Sasikiran Kandula, Craig J McGowan, Evan Moore, Dave Osthus, Evan L Ray, Abhinav Tushar, Teresa K Yamana, et al. A collaborative multiyear, multimodel assessment of seasonal influenza forecasting in the united states. *Proceedings of the National Academy of Sciences*, 116(8):3146–3154, 2019a. 5.1, 5.3.2

Nicholas G Reich, Craig J McGowan, Teresa K Yamana, Abhinav Tushar, Evan L Ray, Dave Osthus, Sasikiran Kandula, Logan C Brooks, Willow Crawford-Crudell, Graham Casey Gibson, et al. Accuracy of real-time multi-model ensemble forecasts for seasonal influenza in the us. *PLoS computational biology*, 15(11), 2019b. 5.3.2

Nicholas G Reich, Craig J McGowan, Teresa K Yamana, Abhinav Tushar, Evan L Ray, Dave Osthus, Sasikiran Kandula, Logan C Brooks, Willow Crawford-Crudell, Graham Casey Gibson, et al. A collaborative multi-model ensemble for real-time influenza season forecasting in the us. *bioRxiv*, page 566604, 2019c. 5.1

Kevin Reichek and Lisheng Gao. CMU Delphi forecasts. http://delphi.midas.cs.cmu.edu/forecast/, Accessed 2017-04-26. B

Joshua Ritterman, Miles Osborne, and Ewan Klein. Using prediction markets and Twitter to predict a swine flu pandemic. In *Proceedings of the 1st International Workshop on Mining Social Media*, volume 9, pages 9–17, 2009. 4.1

MA Rolfes, IM Foppa, S Garg, B Flannery, L Brammer, JA Singleton, and others. Estimated influenza illnesses, medical visits, hospitalizations, and deaths averted by vaccination in the United States. https://www.cdc.gov/flu/about/disease/2015-16.htm, 2016. 1.0.1

Roni Rosenfeld. The "degenerate EM" algorithm for finding optimal linear interpolation coefficients λ_i. http://www.cs.cmu.edu/~roni/11761/Presentations/degenerateEM.pdf, Accessed 2017-03-21. 5.2, C.1

Mauricio Santillana, D. Wendong Zhang, Benjamin M. Althouse, and John W. Ayers. What Can Digital Disease Detection Learn from (an External Revision to) Google Flu Trends? *American Journal of Preventive Medicine*, 47(3):341–347, September 2014. ISSN 0749-3797, 1873-2607. doi: 10.1016/j.amepre.2014.05.020. URL http://www.ajpmonline.org/article/S0749-3797(14)00238-4/fulltext. 4.1

Mauricio Santillana, André T. Nguyen, Mark Dredze, Michael J. Paul, Elaine O. Nsoesie, and John S. Brownstein. Combining Search, Social Media, and Traditional Data Sources to Improve Influenza Surveillance. *PLOS Computational Biology*, 11(10):e1004513, October 2015. ISSN 1553-7358. doi: 10.1371/journal.pcbi.1004513. URL http://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1004513. 1.2, 4.1

Robert E Serfling. Methods for current statistical analysis of excess pneumonia-influenza deaths. *Public health reports*, 78(6):494, 1963. 1.0.1

Jeffrey Shaman and Alicia Karspeck. Forecasting seasonal outbreaks of influenza. *Proceedings of the National Academy of Sciences*, 109(50):20425–20430, 2012a. 1.1, 2.2

Jeffrey Shaman and Alicia Karspeck. Forecasting seasonal outbreaks of influenza. *Proceedings of the National Academy of Sciences*, 109(50):20425–20430, December 2012b. ISSN 0027-8424, 1091-6490. doi: 10.1073/pnas.1208772109. URL http://www.pnas.org/content/109/50/20425. 4.1

Jeffrey Shaman, Alicia Karspeck, Wan Yang, James Tamerius, and Marc Lipsitch. Real-time influenza forecasts during the 2012–2013 season. *Nature communications*, 4, 2013a. 1.1

Jeffrey Shaman, Alicia Karspeck, Wan Yang, James Tamerius, and Marc Lipsitch. Real-time influenza forecasts during the 2012–2013 season. *Nature Communications*, 4:2837, December 2013b. ISSN 2041-1723. doi: 10.1038/ncomms3837. URL https://www.nature.com/articles/ncomms3837. 4.1

Alessio Signorini, Alberto Maria Segre, and Philip M. Polgreen. The Use of Twitter to Track Levels of Disease Activity and Public Concern in the U.S. during the Influenza A H1n1 Pandemic. *PLOS ONE*, 6(5):e19467, May 2011. ISSN 1932-6203. doi: 10.1371/journal.pone.0019467. URL http://journals.plos.org/plosone/article?id=10.1371/journal.pone.0019467. 4.1

James M Simmerman, Piyarat Suntarattiwong, Jens Levy, Richard G Jarman, Suchada Kaewchana, Robert V Gibbons, Ben J Cowling, Wiwan Sanasuttipun, Susan A Maloney, Timothy M Uyeki, et al. Findings from a household randomized controlled trial of hand washing and face masks to reduce influenza transmission in bangkok, thailand. *Influenza and other respiratory viruses*, 5(4):256–267, 2011. 1.0.1

Mark S Smolinski, Adam W Crawley, Kristin Baltrusaitis, Rumi Chunara, Jennifer M Olsen, Oktawia Wójcik, Mauricio Santillana, Andre Nguyen, and John S Brownstein. Flu Near You: crowdsourced symptom reporting spanning 2 influenza seasons. *American journal of public health*, 105(10):2124–2130, 2015. 1.2

Radina P. Soebiyanto, Farida Adimi, and Richard K. Kiang. Modeling and Predicting Seasonal Influenza Transmission in Warm Regions Using Climatological Parameters. *PLOS ONE*, 5(3):e9450, March 2010. ISSN 1932-6203. doi: 10.1371/journal.pone.0009450. URL http://journals.plos.org/plosone/article?id=10.1371/journal.pone.0009450. 4.1

Le Song, Byron Boots, Sajid M Siddiqi, Geoffrey J Gordon, and Alex Smola. Hilbert space embeddings of hidden markov models.(2010). 2010. a

Thorsten Suess, Cornelius Remschmidt, Susanne B Schink, Brunhilde Schweiger, Andreas Nitsche, Kati Schroeder, Joerg Doellinger, Jeanette Milde, Walter Haas, Irina Koehler, et al. The role of facemasks and hand hygiene in the prevention of influenza transmission in households: results from a cluster randomised trial; berlin, germany, 2009-2011. *BMC infectious diseases*, 12(1):26, 2012. 1.0.1

Maha Talaat, Salma Afifi, Erica Dueger, Nagwa El-Ashry, Anthony Marfin, Amr Kandeel, Emad Mohareb, and Nasr El-Sayed. Effects of hand hygiene campaigns on incidence of laboratory-confirmed influenza and absenteeism in schoolchildren, cairo, egypt. *Emerging infectious diseases*, 17(4):619, 2011. 1.0.1

The Delphi Group at Carnegie Mellon University. The Delphi Epidemiological Data API. https://github.com/cmu-delphi/delphi-epidata, Accessed 2017-04-26. 1.3.1

MG Thompson, DK Shay, H Zhou, CB Bridges, PY Cheng, E Burns, JS Bresee, and NJ Cox. Estimates of deaths associated with seasonal influenza — united states, 1976-2007. *Morbidity and Mortality Weekly Report*, 59(33):1057, 2010. URL https://www.cdc.gov/mmwr/preview/mmwrhtml/mm5933a1.htm. 1.0.1

William W Thompson, David K Shay, Eric Weintraub, Lynnette Brammer, Nancy Cox, Larry J Anderson, and Keiji Fukuda. Mortality associated with influenza and respiratory syncytial virus in the united states. *Jama*, 289(2):179–186, 2003. 1.0.1

Ryan J Tibshirani. Adaptive piecewise polynomial estimation via trend filtering. *The Annals of Statistics*, 42(1):285–323, 2014. 2.2, C.1

Steffen Unkel, C Farrington, Paul H Garthwaite, Chris Robertson, and Nick Andrews. Statistical methods for the prospective detection of infectious disease outbreaks: a review. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 175(1):49–82, 2012. 1.1

Willem G van Panhuis, Sangwon Hyun, Kayleigh Blaney, Ernesto TA Marques Jr, Giovanini E Coelho, João Bosco Siqueira Jr, Ryan Tibshirani, Jarbas B da Silva Jr, and Roni Rosenfeld. Risk of dengue for tourists and teams during the world cup 2014 in brazil. *PLoS Negl Trop Dis*, 8(7):e3063, 2014. 2.2, 2.2, 2.2

Guido Van Rossum and Fred L Drake. *Python language reference manual*. Network Theory, 2003. B

Cecile Viboud, Mark Miller, Donald R Olson, Michael Osterholm, and Lone Simonsen. Preliminary estimates of mortality and years of life lost associated with the 2009 a/h1n1 pandemic in the us and comparison with past influenza seasons. *PLoS currents*, 2, 2010. 1.0.1

Cécile Viboud, Vivek Charu, Donald Olson, Sébastien Ballesteros, Julia Gog, Farid Khan, Bryan Grenfell, and Lone Simonsen. Demonstrating the use of high-volume electronic medical claims data to monitor local and regional influenza activity in the us. *PloS one*, 9(7):e102429, 2014. 1.2

Cécile Viboud, Pierre-Yves Boëlle, Fabrice Carrat, Alain-Jacques Valleron, and Antoine Flahault. Prediction of the spread of influenza epidemics by the method of analogues. *American Journal of Epidemiology*, 158(10):996–1006, 2003. 1.1, 2.2

Daren Wang. Predicting seasonal influenza epidemics, 2016. Carnegie Mellon University Department of Statistics Advanced Data Analysis project. Advisors: Ryan Tibshirani, Wilbert Van Panhuis. 2.4

David H Wolpert. Stacked generalization. *Neural networks*, 5(2):241–259, 1992. 5.2

World Health Organization. The top 10 causes of death. URL `http://www.who.int/news-room/fact-sheets/detail/the-top-10-causes-of-death`. (document), 1.0.1

World Health Organization. WHO | Influenza (Seasonal), 2016. URL `http://www.who.int/mediacentre/factsheets/fs211/en/`. 1.0.1

Teresa K Yamana, Sasikiran Kandula, and Jeffrey Shaman. Individual versus superensemble forecasts of seasonal influenza outbreaks in the united states. *PLoS computational biology*, 13(11):e1005801, 2017. 5.1, 5.2

Shihao Yang, Mauricio Santillana, and S. C. Kou. Accurate estimation of influenza epidemics using Google search data via ARGO. *Proceedings of the National Academy of Sciences*, 112(47):14473–14478, November 2015. ISSN 0027-8424, 1091-6490. doi: 10.1073/pnas.1515373112. URL `http://www.pnas.org/content/112/47/14473`. 1.1, 4.1

Shihao Yang, Mauricio Santillana, John S. Brownstein, Josh Gray, Stewart Richardson, and S. C. Kou. Using electronic health records and Internet search information for accurate influenza forecasting. *BMC Infectious Diseases*, 17:332, May 2017. ISSN 1471-2334. doi: 10.1186/s12879-017-2424-7. URL https://doi.org/10.1186/s12879-017-2424-7. 1.1, 4.1

Wan Yang, Alicia Karspeck, and Jeffrey Shaman. Comparison of filtering methods for the modeling and retrospective forecasting of influenza epidemics. *PLoS computational biology*, 10(4):e1003583, 2014. 1.1

Qian Zhang, Nicola Perra, Daniela Perrotta, Michele Tizzoni, Daniela Paolotti, and Alessandro Vespignani. Forecasting seasonal influenza fusing digital indicators and a mechanistic disease model. In *Proceedings of the 26th International Conference on World Wide Web*, pages 311–319. International World Wide Web Conferences Steering Committee, 2017. 1.1

Hong Zhou, William W Thompson, Cecile G Viboud, Corinne M Ringholz, Po-Yung Cheng, Claudia Steiner, Glen R Abedi, Larry J Anderson, Lynnette Brammer, and David K Shay. Hospitalizations associated with influenza and respiratory syncytial virus in the united states, 1993–2008. *Clinical infectious diseases*, 54 (10):1427–1436, 2012. 1.0.1

Christoph Zimmer, Reza Yaesoubi, and Ted Cohen. A likelihood approach for real-time calibration of stochastic compartmental epidemic models. *PLoS computational biology*, 13(1):e1005257, 2017. 1.1

Christoph Zimmer, Sequoia I Leuba, Ted Cohen, and Reza Yaesoubi. Accurate quantification of uncertainty in epidemic parameter estimates and predictions using stochastic compartmental models. *Statistical methods in medical research*, page 0962280218805780, 2018. 1.1