Anticipating Outbreaks: Predictive Modeling to Improve Infectious Disease Surveillance

Anticipating Outbreaks: Predictive Modeling to Improve Infectious Disease Surveillance

A DISSERTATION PRESENTED
BY
SARAH MCGOUGH
TO
THE DEPARTMENT OF GLOBAL HEALTH & POPULATION
HARVARD T.H. CHAN SCHOOL OF PUBLIC HEALTH

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY
IN THE SUBJECT OF
POPULATION HEALTH SCIENCES

HARVARD UNIVERSITY
CAMBRIDGE, MASSACHUSETTS
MAY 2019

Dissertation Advisor: Nicolas Menzies Sarah McGough

Anticipating Outbreaks: Predictive Analytics to Improve Infectious Disease Surveillance

ABSTRACT

Rapid and effective responses to disease outbreaks require the ability to accurately detect and anticipate changing dynamics of an outbreak over time. However, disease surveillance is frequently undermined by extended delays between symptom onset and official case reports, often due to complex and multi-tiered disease reporting and communication systems interacting at national, state, and city levels. Timeliness of reporting and response may be further exacerbated in settings that experience resource constraints.

Digital data streams that are available in real- or near-real-time have the potential to complement or improve traditional disease surveillance by quickly and continuously capturing signals of population health that may be meaningful for disease tracking and forecasting. In addition, digital data are trending towards being made freely and publicly available through public servers and APIs, which remove barriers to data access and open up avenues for predictive modeling independent of resource level. Further, methodologies that focus on data-driven and self-adaptive learning can yield flexible and readily-implementable models for the public health sector.

Focusing on a collection of inputs, including Google search trends, Twitter, news reports, and satellite weather data, and employing statistical and machine learning methodologies to process, synthesize, and analyze these data, I present

several applications of disease detection and forecasting models, which are developed as real-time decision support tools. Each project uses, as a case study, a mosquito-borne disease outbreak, which requires anticipation on the scale of weeks or months to effectively interrupt transmission. Across case studies, I describe flexible models functional at both large (*e.g.* national) and small (*e.g.* city-level) spatial scales. Over the course of this thesis, I move towards increasingly more generalizable modeling techniques such that learning from input data becomes more autonomous, requires less human input, and can be applied to a wider range of systems (*e.g.* surveillance bodies, diseases). In all cases, I show how predictions can fill a critical time gap between case onset and case reporting, with the goal of supporting early warnings and outbreak anticipation within public health surveillance systems.

Contents

Listing of Figures

DEDICATED TO MY FAMILY.

Acknowledgments

It is difficult to summarize (and express sufficient gratitude for) the countless people who have made my degree, and this body of work, possible. Here is my attempt: I would like, firstly, to broadly thank the rich learning environment provided by Harvard University, the T.H. Chan School of Public Health, the Population Health Sciences program, and the Departments of Global Health & Population and Epidemiology. It is a joy to be surrounded by interesting scholars both within and outside my field, and the academic excellence across disciplines was an important motivation to study at this university.

I am truly grateful for the incredible mentorship and support I have received from all three of my dissertation committee members: Nick Menzies, Mauricio Santillana, and Marc Lipsitch. All three have provided dedicated and personalized advising as I have navigated both familiar and unfamiliar territory in my thesis projects. I am grateful to my advisor, Nick, for providing great direction across several projects over the years, for being extremely available and willing to provide support, even at off hours, and for basing his advising around supporting my personal and professional goals. I also thank Mauricio Santillana for introducing me to the world of predictive modeling, for challenging me with interesting problems, for endless research and life chats, and for opening up many new research avenues, collaborations, and opportunities for me to pursue. And lastly, I would like to thank Marc Lipsitch for taking me on quite early into my research career, for countless sessions working through this interesting, “chewy” nowcast problem together, and for generosity with his time to meet and go through ideas over the

years. It has been a highlight of my PhD to work on such a close level with each of these individuals.

I would also like to thank a few key collaborators: John Brownstein, for bringing me into the HealthMap and IDHA groups so early into my research career and exposing me to the exciting world of digital health; and Michael Johansson and Nathan Kutz for putting forth interesting and important ideas to run with. I am grateful for the administrative support of the Department of Global Health and Population, in particular Barbara Heil and Allison Gallant for their cheer, love, and support.

An additional thanks to other supporters I have met along the way: Don Goldmann, Ken McIntosh. Thank you to my Dudley House family: Jim and Doreen Hogle, Susan Zawalich, Jeffrey Shenette, and my fellow Dudley Fellows.

Lastly, because a PhD is not possible to achieve (sanely) without love, fun, and friendship:

I would like to thank Kate Grode, my lifelong best friend- for just getting it. I would like to thank RBF: MK Downer, Barbra Dickerman, and Krystal Cantos- for so much necessary emotional support and laughter. I would like to thank Tom Brady- for showing me that it's always possible to convert on 3rd and 10. To BDS: you know who you are- for making this a really rad 5 years. To my family (McGoughs and Domels)- for endless love and support; especially my mother, Arleen McGough, for surprises in the mail, and my father, Mark McGough, for always making the time to visit, and to both, for instilling in me a love of learning and education. And lastly, to August Domel- for everything.

1

Introduction

Disease surveillance is a critical input that informs public health action. Rapid and effective responses to disease outbreaks rely on timely and accurate case reports to assess risks, prioritize public health threats, allocate resources across multiple health sectors, and deploy interventions to interrupt disease transmission. However, rarely are cases reported into the surveillance system the moment they occur, for administrative, biological, and logistical reasons [1]. This compromises the role of surveillance in detecting changes in disease transmission in real-time. Predictive modeling offers one solution to this problem, by providing estimates of current (nowcast) and future (forecast) disease activity using a wide range of data streams and signal mining techniques.

In the past decade, the near real-time availability of novel and disparate internet-based data sources has motivated the development of complementary methodologies to track the incidence and spread of diseases. These approaches exploit information from internet search engines[2–5], news reports[6–8], clinicians' search engines[9], crowd-sourced participatory disease surveillance systems[10–12], Twitter microblogs[13–16], Electronic Health Records[17], and satellite images[18] to estimate the presence of a disease in a given location. Exploiting these relationships to quickly and continuously capture signals of

population health activity can generate accurate, prospective disease forecasts that complement traditional surveillance. Further, the majority of these data streams have the advantage being freely and digitally accessible, and thus may be leveraged independent of resource level.

A challenge in predictive modeling is developing sufficiently generalizable and adaptive models; that is, models that demonstrate successful application to different contexts (e.g. locations, diseases) and that are capable of learning from new information, in addition to standard concerns for overfitting and out-of-sample performance. Particularly for the public health system, it is useful to develop flexible and readily-implementable models that can adapt to a wide range of surveillance problems. Thus, methodologies that focus on data-driven and self-adaptive learning are especially of interest.

In this thesis, I leverage different methodologies and data streams for the purpose of infectious disease forecasting, focusing on two important predictive modeling goals: generalizability and reproducibility. Statistical and machine learning algorithms are used to create flexible, adaptive disease tracking and forecasting models that show promise in a wide range of applications (e.g. location, disease) and that rely on digital and open-access data, alongside data readily available in the surveillance system. Specifically, in the absence of access to real-time government-reported Zika case counts during the 2016 outbreak in Latin America, I demonstrate the ability of Internet-based data sources to track and predict the outbreak in five countries, through a collection of dynamic, multivariable models (Chapter 2). These models use readily-available and freely-

accessible digital data to fill a critical time gap for decision-makers ahead of the release of official case documentation. Then, focusing on the important but complex relationship between climate and mosquito-borne disease dynamics, I present a data-driven, machine learning approach capable of identifying, at a high spatial resolution, potentially useful weather patterns to predict dengue fever epidemics in a diverse set of municipalities in Brazil (Chapter 3). Specifically, this project exploits ensemble learning with support vector machines to generate dynamic models that are self-adaptive, i.e. require no human input to detect and learn from highly predictive patterns in the data. Beyond learning from weather patterns, this project additionally incorporates mechanistic knowledge of dengue outbreak cycles in order to improve model forecasts and align with understood transmission dynamics. Finally, in the last chapter of this thesis, I introduce a simple and flexible model capable of producing accurate nowcasts in a multitude of disease settings and temporal ranges (Chapter 4). The model requires no disease-specific parameterization, learning only from historical cases and reporting delays, which allows the model to function well in very different disease settings.

The models and applications presented here serve as proof-of-concept or pilot-implemented tools for real-time disease tracking and forecasting.

1. Jajosky RA, Groseclose SL. Evaluation of reporting timeliness of public health surveillance systems for infectious diseases. BMC Public Health. 2004;4: 29.

2. Polgreen PM, Chen Y, Pennock DM, Nelson FD. Using internet searches for influenza surveillance. Clin Infect Dis. 2008;47: 1443–1448.

3. Ginsberg J, Mohebbi MH, Patel RS, Brammer L, Smolinski MS, Brilliant L. Detecting influenza epidemics using search engine query data. Nature. 2009;457: 1012–1014.

4. Yuan Q, Qingyu Y, Nsoesie EO, Benfu L, Geng P, Rumi C, et al. Monitoring Influenza Epidemics in China with Search Query from Baidu. PLoS One. 2013;8: e64323.

5. Althouse BM, Ng YY, Cummings DAT. Prediction of Dengue Incidence Using Search Query Surveillance. PLoS Negl Trop Dis. 2011;5: e1258.

6. Majumder MS, Kluberg S, Santillana M, Mekaru S, Brownstein JS. 2014 ebola outbreak: media events track changes in observed reproductive number. PLoS Curr. 2015;7. doi:10.1371/currents.outbreaks.e6659013c1d7f11bdab6a20705d1e865

7. Majumder MS, Santillana M, Mekaru SR, McGinnis DP, Khan K, Brownstein JS. Utilizing Nontraditional Data Sources for Near Real-Time Estimation of Transmission Dynamics During the 2015-2016 Colombian Zika Virus Disease Outbreak. JMIR Public Health Surveill. 2016;2: e30.

8. Brownstein JS, Freifeld CC, Reis BY, Mandl KD. Surveillance Sans Frontières: Internet-based emerging infectious disease intelligence and the HealthMap project. PLoS Med. 2008;5: e151.

9. Santillana M, Nsoesie EO, Mekaru SR, Scales D, Brownstein JS. Using clinicians' search query data to monitor influenza epidemics. Clin Infect Dis. 2014;59: 1446–1450.

10. Smolinski MS, Crawley AW, Baltrusaitis K, Chunara R, Olsen JM, Wójcik O, et al. Flu Near You: Crowdsourced Symptom Reporting Spanning 2 Influenza Seasons. Am J Public Health. 2015;105: 2124–2130.

11. Paolotti D, Carnahan A, Colizza V, Eames K, Edmunds J, Gomes G, et al. Web-based participatory surveillance of infectious diseases: the Influenzanet participatory surveillance experience. Clin Microbiol Infect. 2014;20: 17–21.

12. Dalton C, Durrheim D, Fejsa J, Francis L, Carlson S, d'Espaignet ET, et al. Flutracking: a weekly Australian community online survey of influenza-like illness in 2006, 2007 and 2008. Commun Dis Intell Q Rep. 2009;33: 316–322.

13. Paul MJ, Dredze M, Broniatowski D. Twitter improves influenza forecasting. PLoS Curr. 2014;6. doi:10.1371/currents.outbreaks.90b9ed0f59bae4ccaa683a39865d9117

14. Broniatowski DA, Paul MJ, Dredze M. National and local influenza surveillance through Twitter: an analysis of the 2012-2013 influenza epidemic. PLoS One. 2013;8: e83672.

15. Nagar R, Yuan Q, Freifeld CC, Santillana M, Nojima A, Chunara R, et al. A case study of the New York City 2012-2013 influenza season with daily geocoded Twitter data from temporal and spatiotemporal perspectives. J Med Internet Res. 2014;16: e236.

16. Signorini A, Segre AM, Polgreen PM. 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. 2011;6: e19467.

17. Santillana M, Nguyen AT, Louie T, Zink A, Gray J, Sung I, et al. Cloud-based Electronic Health Records for Real-time, Region-specific Influenza Surveillance. Sci Rep. 2016;6: 25732.

18. Nsoesie EO, Patrick B, Naren R, Mekaru SR, Brownstein JS. Monitoring Disease Trends using Hospital Traffic Data from High Resolution Satellite Imagery: A Feasibility Study. Sci Rep. 2015;5: 9112.

2

Forecasting Zika Incidence in the 2016 Latin America Outbreak Combining Traditional Disease Surveillance with Search, Social Media, and News Report Data

2.0 ABSTRACT

Over 400,000 people across the Americas are thought to have been infected with Zika virus as a consequence of the 2015-2016 Latin American outbreak. Official government-led case count data in Latin America are typically delayed by several weeks, making it difficult to track the disease in a timely manner. Thus, timely disease tracking systems are needed to design and assess interventions to mitigate disease transmission.

We combined information from Zika-related Google searches, Twitter microblogs, and the HealthMap digital surveillance system with historical Zika suspected case counts to track and predict estimates of suspected weekly Zika cases during the 2015-2016 Latin American outbreak, up to three weeks ahead of the publication of official case data. We evaluated the predictive power of these data and used a dynamic multivariable approach to retrospectively produce predictions of weekly suspected cases for five countries: Colombia, El Salvador, Honduras, Venezuela, and Martinique. Models that combined Google (and Twitter data where available) with autoregressive information showed the best out-of-sample

predictive accuracy for 1-week ahead predictions, whereas models that used only Google and Twitter typically performed best for 2- and 3-week ahead predictions.

Given the significant delay in the release of official government-reported Zika case counts, we show that these Internet-based data streams can be used as timely and complementary ways to assess the dynamics of the outbreak.

2.1 INTRODUCTION

The rapid spread of Zika virus has led to more than 400,000 suspected cases across the Americas since its introduction to Brazil in 2014, and has triggered alerts around the globe[1]. This event has led to diverse interventions and travel warnings to affected areas, underscoring the importance of proactive disease surveillance. While cases of sexual transmission of Zika virus have been documented[2], the virus is primarily transmitted through the bite of the Aedes aegypti mosquito and causes nonspecific flu-like symptoms and skin rashes[3,4]. Of particular concern is the possible link between Zika virus and neurological disorders such as microcephaly, a birth defect in which babies of infected pregnant women are born with abnormally small heads[5–8]. Over 1800 cases of Zika-related microcephaly and central nervous system disorders in newborns have been reported since the beginning of the epidemic, and the virus has spread to 70 countries globally[9]. In February 2016, the World Health Organization declared Zika a global public health emergency[10]. With no existing vaccinations or treatment for Zika infections, control of the Aedes aegypti mosquito is critical to curb the spread of the virus, as has been observed in dengue fever studies[11,12]. This requires continuous and up-to-date surveillance of cases to drive vector control interventions accordingly[13].

In countries with now autochthonous transmission, the surveillance of Zika infections is predominantly passive; cases are identified on the basis of hospitalizations and clinical symptom reports. The Pan American Health Organization (PAHO) currently streamlines reports from ministries of health, and reports weekly confirmed and suspected cases of Zika by country[14]. The release of these reports and those produced by the ministries, however, is typically delayed by three or more weeks due to systematic processing and data collection. As a consequence, the changing dynamics of Zika are frequently hard to be assessed in a timely manner, and thus, the availability of current data on Zika to the public and public health officials is limited.

In the past decade, the near real-time availability of novel and disparate internet-based data sources has motivated the development of complementary methodologies to track the incidence and spread of diseases. These approaches exploit near real-time information from internet search engines[15–18], news reports[19–21], clinician's search engines[22], crowd-sourced participatory disease surveillance systems[23–25], Twitter microblogs[26–29], Electronic Health Records[30], and satellite images[31] to estimate the presence of a disease in a given location.

Some of the biases and errors observed when using these alternative data sources as individual indicators of disease incidence have been recently mitigated by using ensemble approaches that combine information from multiple data sources to produce a more robust disease estimate[32]. In parallel, multiple improvements have been proposed to disease tracking methodologies based on Google

searches[33–38]. Finally, it has been shown that in the absence of information from traditional government-lead disease reporting, the combined use of news reports and Google's search activity of the word "zika" in Colombia led to reasonable estimates of cumulative cases of Zika[20]. To the best of our knowledge, however, no attempts have been made to date to harness these and other digital data sources for near-real time weekly forecasting of Zika infections.

Here we assess the feasibility of using Zika-related Google search queries, Zika-related Twitter microblogs, and information from news reports collected by the web-based surveillance system HealthMap[16], in the prospective monitoring of Zika in five countries: Colombia, El Salvador, Honduras, Venezuela, and Martinique. In addition, we evaluate the ability of a collection of multivariable models that use information from these three data sources as input, to dynamically track and forecast the incidence of Zika virus up to 3 weeks ahead of the release of reports from PAHO, using multiple evaluation metrics.

2.2 METHODS

2.2.1 DATA

EPIDEMIOLOGIC DATA. We obtained weekly reports from the Pan American Health Organization (PAHO) that document the number of laboratory-confirmed and suspected cases of Zika in the Americas from the website (http://ais.paho.org/phip/viz/ed_zika_epicurve.asp) and from weekly epidemiological updates[39]. In the absence of this information, we obtained suspected and lab-confirmed Zika cases from epidemiological bulletins produced by the national Ministries of Health (MOH) of Colombia and Martinique[40,41]. Throughout the manuscript, we refer to these data as "official case count". Due to the lack of robust diagnostic capabilities across the Americas and the estimated large number of asymptomatic cases[4,42], the present study focuses on predicting suspected Zika cases, which can be used as a proxy for potential hospital visits in each locality. This information could be useful for public health decision-makers when designing resource allocation plans. Under PAHO criteria, cases were classified as suspected if the patient presented a rash and two or more of the following symptoms: fever, conjunctivitis, arthralgia, myalgia, and peri-articular edema[43]. The time series of suspected cases spans the entire epidemic period of each country, beginning with the earliest reported cases through the last available epidemiologic week in the data (last accessed August 3, 2016). Data profiles for each country can be seen in Table 1.

Table 2.1. Data profile for countries.

	Colombia	Venezuela	Martinique	Honduras	El Salvador
Cumulative cases	92891	51043	33925	22705	11779
Number of search terms	26	15	8	11	12
Weeks of data	46	38	30	26	37
Week of first cases	8/9/15	10/11/15	12/27/15	12/13/15	9/20/15
Week of last accessible cases	7/10/16	6/26/16	7/17/16	5/29/16	5/29/15
Number of training weeks (G+T, AR / AGO+T / ARGO+TH)	20, 17	15, 12	12, 9	12, 9	17, 14

GOOGLE SEARCH QUERIES. The selection process of potentially useful search terms to track Zika avoided forward-looking bias and was performed via the Google Correlate and Google Trends tools (https://www.google.com/trends/correlate/; https://www.google.com/trends/). We identified the most highly correlated terms with the time evolution of Zika incidence in Colombia and Venezuela on Google Correlate within the time period of May 2015 to Jan 2016, and used Google Trends to identify search terms related to the term "Zika" for all five countries. The time window for the selection of these terms did not exceed the training period of each model. Because the output of Google Trends and Google Correlate consists of country-specific search terms, these are different for each country. All highly correlated terms to the query "Zika" were selected as model inputs without discrimination, including some potential misspellings of the disease such as "sika" and "sica". We obtained weekly fractions of all identified Google search terms using the Google Trends website. The selected search terms were used as independent variables in the models and are shown in Table A.1.

TWITTER MICROBLOGS. We leveraged a custom script to access the free Twitter Public API to collect the maximum allowed number of tweets (up to 1% total Twitter volume) with any geographical coordinates. We then searched these tweets by country, using Twitter's assigned country code and restricting to tweets in which this parameter was present, for the weekly volume of Twitter micro-blogs containing any of the words "Zika", "microcephaly", and "microcefalia", but only Colombia and Venezuela had relevant Zika-related tweets, within the weeks of the epidemic outbreak, to merit the inclusion of Twitter data in our models. The fraction of tweets containing the Zika-related words when compared to the total number of tweets for each country was computed for every week and used as an independent variable in the models.

HEALTHMAP DIGITAL SURVEILLANCE. We obtained cumulative reported case counts of Zika virus disease in all countries via the HealthMap digital disease surveillance system (www.healthmap.org), which reports non-governmental media alerts of infections[16]. From these alerts, we calculated the weekly incidence of Zika infection for use as an independent variable in the models.

RELATIONSHIP BETWEEN CASES AND INTERNET-BASED DATA. In order to assess whether the selected Google search terms, Twitter microblogs, and HealthMap-reported cases could be useful for weekly prediction of Zika incidence, we computed the Pearson's correlation between each predictor and the official Zika case count, first

for the training period of each country and later for the entire time series. In addition, we evaluated the autocorrelation of the signal itself (as lag-1, lag-2, and lag-3 terms). To determine the optimal linear relationship between the predictors and cases, we applied a series of simple transformations to these data and selected the transformation which produced the highest Pearson's correlation. The results of this preliminary analysis was used for variable selection and to inform the dynamic transformation of variables process within the model, detailed below.

2.2.2 MODELS

A collection of multivariable models, inspired by those introduced in the Flu prediction literature[30,37], were considered to estimate and forecast weekly suspected cases of Zika in the aforementioned five countries. These models used as input the weekly Google search frequencies of Zika-related terms, the fraction of Zika-related Twitter microblogs, cumulative Zika case counts as recorded by the HealthMap disease surveillance system, and the available historical official case count data at a given point in time. For consistency and comparability, all models (i) automatically select the most relevant search terms for prediction, (ii) incorporate new information on Zika cases as reports are released every week, and (iii) identify the best functional relationship between each input variable and the outcome variable, every week.

The selection of the most predictive input variables was performed using a penalized Least Absolute Shrinkage and Selection Operator (LASSO) regression approach as described in[44]. While avoiding the use forward-looking information, we incorporated the most recently available information on Zika cases every week

by dynamically expanding the time window of the training set of the models. Finally, at each week, we analyzed whether transforming each input variable would increase its correlation with the output variable. If this were the case, then the transformed value of the input variable producing the highest correlation with case data would be used as input for the model. As more epidemiological information becomes available, this dynamic transformation process allows the model to recursively recalibrate and incorporate changes in the relationships between the input variables and the case count information observed so far. The transformations we considered were not exhaustive and included the log(x), x2, and sqrt(x).

In addition to the models that used the aforementioned data streams as input, we built a collection of baseline models for comparison and context. We considered models that only used historical observation of Zika cases to predict cases on the subsequent weeks and models that incorporated information from these various data streams. Given the success of Google search terms in tracking other diseases as observed in [27,28], our models utilized Google search as a central predictor, and we explored the additions of Twitter and HealthMap data for the improvement of model predictions. Specifically, we considered (i) AR: a baseline lag-3 autoregressive model that used only Zika surveillance information from the prior 3 weeks to predict suspected cases, (ii) G+T: a model which used only Google search and Twitter (if available) data for prediction as introduced in[33] (iii) ARGO+T: a model which used autoregressive information and Google and Twitter (if available) data, adapted from[37], and (iv) ARGO+TH: a model which combined all data streams (Twitter if available, Google, HealthMap) with lag-3 autoregressive terms.

For the two countries (Colombia and Venezuela) which had available Twitter data, we also constructed identical models (ii - iv) without this data source; that is, using Google and HealthMap data only. Our models are described by the following equation

$$\hat{y}_t = \alpha_t + \sum_{i=1}^{N} \gamma_i \, y_{(t-i)} + \sum_{j=1}^{K} \beta_j X_{j,t} + \tau T_t + \eta H_t + \varepsilon_t \qquad \varepsilon_t \sim N(0, \sigma^2)$$

where we expand an autoregressive model of lag N with the inclusion of the fraction of Google search frequency X for each term j, the fraction of Twitter volume T, and HealthMap-reported cases H. As described in[37], autoregressive terms generally help maintain predictions within a reasonable range, while Google and Twitter information help the models to respond more rapidly to sudden changes in the dynamics. Due to the novelty of the Zika outbreak, stationarity was not used as a way to assess the appropriateness of using autoregressive models as a baseline; instead, we relied on the observed high autocorrelation of the signal with recent time lags of case counts and evidence of similar mosquito-borne outbreaks modeling approaches[45,46].

At each week, we used our models to generate predictions for 1, 2, and 3 weeks ahead of current time. To avoid future-looking bias in our predictions, forecasts were made using only the information available to each model at each week t; and for each time horizon our case count estimate was obtained using a different model. For instance, all models with autoregressive terms are restricted, in further week-ahead predictions, from accessing weeks of case data that have not yet occurred relative to week t. Thus, 3-week ahead (t+3) forecasts for model (i)

were generated using only the lag-3 term (AR3) of official cases from 3 weeks prior to t+3: that is, using the observed cases available exactly at week t. 1-week ahead (t+1) forecasts for model (i), meanwhile, utilized all three AR1, AR2, and AR3 terms, which contain information on reported cases from the strictly observable weeks t, t-1, and t-2. In other words, data that would be unavailable in real-time for predictions - in our case, data on future infections - are excluded from each model. This same rule applies to models (iii) and (iv), which also include autoregressive information. Reflecting the delay in the release of case reporting, the models do access future weeks (relative to week t of case reporting) of Google searches, Twitter microblogs, and HealthMap-reported cases, since these digital streams are available closer to real-time than are official case data.

All models were trained through the same week in the time series and evaluated over the same time window, although the number of training weeks differed based on the information required in each model. Models containing autoregressive information began training 4 weeks into each epidemic, as opposed to training from the first week of reported cases, in order to necessarily inform the one-, two-, and three-week lag terms. A summary of dates and data used by country is shown in Table 1.

Models were fit as multiple generalized linear models with the glmnet package[47] in R v3.2.4[48], validated using k-fold cross validation, and evaluated for their out-of-sample predictive performance. For each model, we report three evaluation metrics: root mean square error (RMSE), the relative RMSE (rRMSE), and the Pearson correlation of predictions with observed cases, as detailed in[32].

Equations for each metric can be found in Equations A.1.

2.3 RESULTS

In order to evaluate the feasibility of using Zika-related Google searches, Twitter microblogs, HealthMap news reports, and historical official case counts to track Zika, we calculated the Pearson correlation between (a) the observed suspected case counts and each input variable, and (b) the observed suspected case counts and three transformations: log(x), x2, and sqrt(x), for each input variable. These transformations were observed to sometimes lead to better correlation values than the original raw variables for different time periods. Figures A.1-5 displays in each country the best transformation of each input variable and suspected Zika case counts. From the multiple panels for each country, it can be seen that at least a subset of these (transformed) variables showed potential to be useful to track Zika. Indeed, correlations ranged from 0.93 to 0.56 in Colombia; 0.90 to 0.18 in Honduras; 0.39 to 0.29 in Venezuela; 0.69 to 0.13 in Martinique; and 0.92 to 0.41 in El Salvador. The lowest-correlation predictors tended to be the lag-3 autoregressive term, HealthMap-reported cases, and non-specific Google search terms like "Virus."

For each country, we produced out-of-sample predictions for the one, two, and three-week ahead time-horizons with the four models introduced in the previous section. We evaluated models according to the maximum number of data sources available, and thus assessed all models with Twitter data, where available (Colombia and Venezuela). In addition, we evaluated models with and without the

inclusion of Twitter data. Plots comparing model predictions with the official Zika case count, by time horizon and country, are shown in Figures 2.1-2.3. Table 2.2 summarizes the out-of-sample predictive performance of the four models for each of the three week-ahead time horizons and for all countries, as captured by the three evaluation metrics. Note that while some model predictions showed high correlation values with official case counts, their predictions showed large discrepancies with the data. As a consequence, we relied on the relative RMSE (rRMSE) to establish the quality of model prediction given the short time span of the outbreaks. The rRMSE provides an estimate of the prediction error relative to the number of true cases observed in each week over the evaluation period, and, from our perspective, allows for better comparisons across models and time horizons. We henceforth judge model performance using this metric.

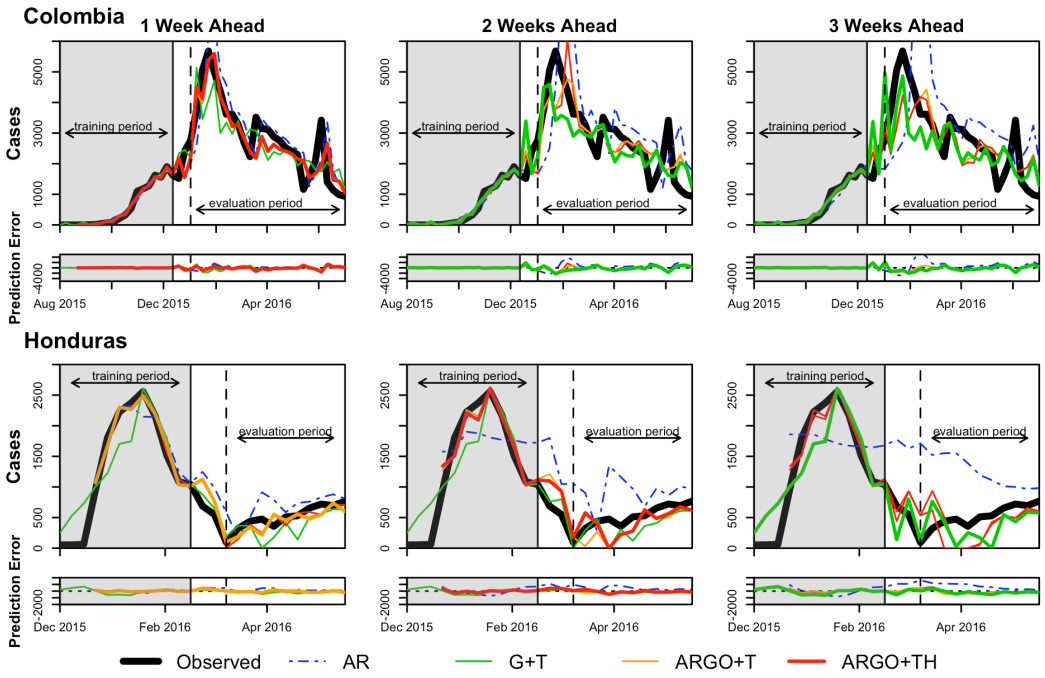

Figure 2.1. Prediction results for (a) Colombia and (b) Honduras. In each country, the weekly estimations of AR (dotted blue), G+T (green), ARGO+T (orange), and ARGO+TH (red) models are compared to the official case counts (black). Models include Twitter data where available (Colombia). The best model performance (lowest relative RMSE) in each time series by country is shown as a bolded line.

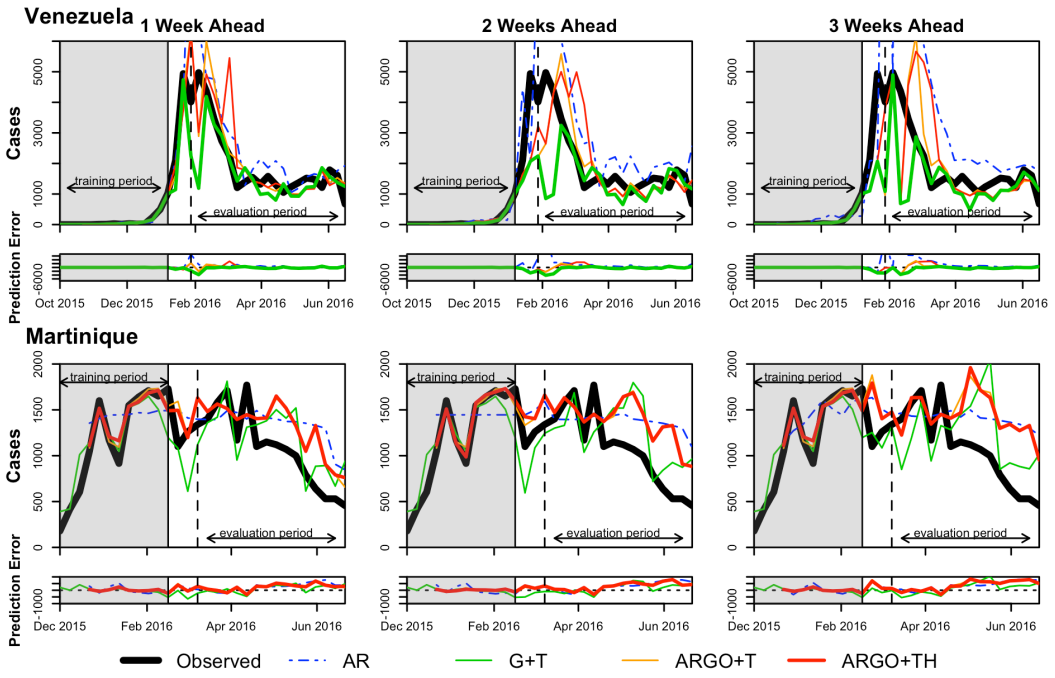

Figure 2.2. Prediction results for (a) Venezuela and (b) Martinique. In each country, the weekly estimations of AR (dotted blue), G+T (green), ARGO+T (orange), and ARGO+TH (red) models are compared to the official case counts (black). Models include Twitter data where available (Venezuela). The best model performance (lowest relative RMSE) in each time series by country is shown as a bolded line.

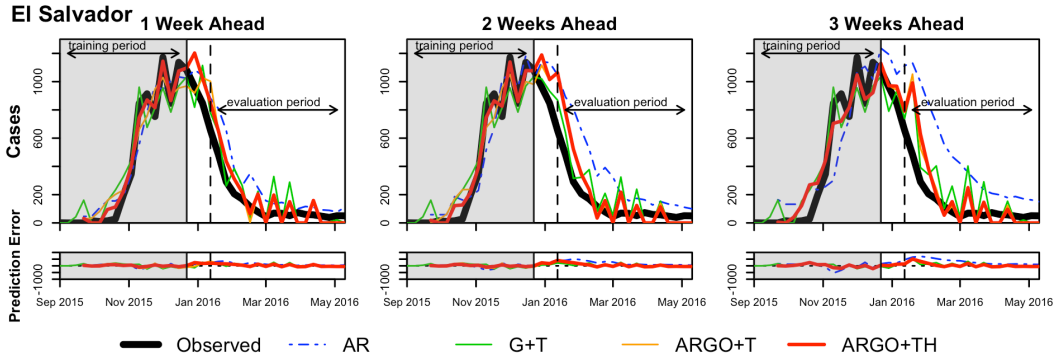

Figure 2.3. Prediction results for El Salvador. The weekly estimations of AR (dotted blue), G+T (green), ARGO+T (orange), and ARGO+TH (red) models are compared to the official case counts (black). The best model performance (lowest relative RMSE) in each time series is shown as a bolded line.

Table 2.2. RMSE, rRMSE, and Pearson's correlation coefficient (ρ) for 1-, 2-, and 3-week ahead out-of-sample predictions. Models include Twitter data where available (Colombia and Venezuela). The best fit metric for each week-ahead prediction is show in bold.

	Colombia								
	1 week			2 week			3 week		
Model	RMSE	rRMSE	ρ	RMSE	rRMSE	ρ	RMSE	rRMSE	ρ
AR	801.313	40.462	0.821	1484.018	66.829	0.539	2057.483	83.900	0.284
G+T	823.149	34.450	0.764	857.490	**37.300**	0.752	995.311	**41.903**	0.634
ARGO+T	621.673	30.076	0.870	**775.786**	39.583	**0.780**	914.643	44.233	0.679
ARGO+TH	**617.795**	**29.888**	**0.871**	848.968	40.153	0.731	**903.155**	42.440	**0.698**
	Venezuela								
	1 week			2 week			3 week		
	RMSE	rRMSE	ρ	RMSE	rRMSE	ρ	RMSE	rRMSE	ρ
AR	1665.733	68.542	0.822	4196.484	117.444	**0.834**	10349.050	259.699	**0.665**
G+T	972.937	**35.336**	0.626	1277.588	**39.813**	0.283	**1226.614**	**39.953**	0.475
ARGO+T	**892.063**	38.780	**0.831**	**927.343**	41.946	0.701	1372.884	48.249	0.486
ARGO+TH	1036.760	46.497	0.771	1148.229	67.028	0.626	1459.830	75.513	0.528
	Martinique								
	1 week			2 week			3 week		
	RMSE	rRMSE	ρ	RMSE	rRMSE	ρ	RMSE	rRMSE	ρ
AR	397.204	59.298	0.678	460.931	73.935	0.617	477.638	78.409	**0.744**
G+T	**302.038**	**40.123**	0.721	**376.475**	**47.758**	0.586	**450.635**	**53.835**	0.384

Table 2.2 (Continued). RMSE, rRMSE, and Pearson's correlation coefficient (ρ) for 1-, 2-, and 3-week ahead out-of-sample predictions. Models include Twitter data where available (Colombia and Venezuela). The best fit metric for each week-ahead prediction is show in bold.

ARGO+T	336.375	42.998	**0.800**	425.005	61.420	0.701	510.691	73.822	0.492
ARGO+TH	342.577	44.923	0.799	424.417	61.382	**0.710**	506.310	73.423	0.482

	Honduras								
	1 week			**2 week**			**3 week**		
	RMSE	**rRMSE**	**ρ**	**RMSE**	**rRMSE**	**ρ**	**RMSE**	**rRMSE**	**ρ**
AR	262.701	167.009	0.546	538.930	330.114	-0.068	886.701	555.937	-0.903
G+T	213.788	53.909	0.675	222.045	51.993	**0.740**	**292.718**	**64.733**	**0.355**
ARGO+T	144.327	**30.436**	0.784	222.278	55.670	0.736	323.089	158.377	0.243
ARGO+TH	**132.675**	41.605	**0.853**	**203.616**	**51.874**	0.584	335.778	163.436	0.085

	El Salvador								
	1 week			**2 week**			**3 week**		
	RMSE	**rRMSE**	**ρ**	**RMSE**	**rRMSE**	**ρ**	**RMSE**	**rRMSE**	**ρ**
AR	159.185	126.486	**0.961**	261.119	234.615	0.929	379.797	350.656	0.888
G+T	120.979	166.901	0.881	**124.338**	152.882	0.911	180.282	187.945	0.855
ARGO+T	122.995	112.516	0.960	151.654	103.649	**0.976**	170.130	115.720	**0.923**
ARGO+TH	**100.318**	**110.603**	0.957	149.407	**103.143**	0.975	**166.552**	**113.459**	0.920

As seen in the evaluation metric values, no single model performed best across metrics, time horizons, and countries. Based on the rRMSE, models that combined Google (and Twitter data where available) with autoregressive information showed better predictive accuracy for 1-week ahead predictions. Meanwhile, models that only used Google (and Twitter where available) typically performed best for two and three-week ahead predictions.

The ARGO+T or ARGO+TH models outperformed all other models in 1-week forecasts for all countries with the exception of Venezuela and Martinique. In Venezuela and Martinique, the ARGO+T model (rRMSE = 38.8 and 43.0, respectively) slightly underperformed relative to the G+T model (rRMSE = 35.3 and 40.1, respectively), with a difference in rRMSE of about 3 percent points. In Colombia and El Salvador, the difference in rRMSE was less than 2% between the ARGO+TH and the ARGO+T models, with both models improving the rRMSE substantially compared to the G+T model.

In further week-ahead predictions, the Google and Twitter only (G+T) model outperformed models that also incorporated autoregressive information, exhibiting the lowest rRMSE in 3 of 5 countries for 2-week forecasts, and in 4 of 5 countries for 3-week forecasts.

Across models, prediction accuracy decreased as predictions were made further into the future, resulting in increases in rRMSE (and RMSE) and declines in model correlations across time horizons. Of all countries studied, Colombia had the best model performance in each week-ahead horizon for every model, with the exception of 3-week G+T forecasts; of all time horizons, the 1-week ahead

predictions performed best in each country and model. In most cases, the autoregressive model over-predicted Zika incidence and underperformed all other models.

Table A.2 shows the performance of additional versions of these models (i.e., the ARGO+T model with and without Twitter data). It can be seen that the inclusion of Twitter microblog data into our models improved or was comparable to (within 0.2 rRMSE) the performance of all models lacking Twitter data in Colombia (range of rRMSE reduction: -0.13, 1.6), and of the ARGO+T and ARGO+TH models in Venezuela (range of rRMSE reduction: 8.14, 125.1), for all time horizons. Conversely, incorporating HealthMap digital cases improved the rRMSE by no more than 3.8 points, or 7% (range: 0.06%, 6.8%) across models, time horizons, and countries, but worsened the rRMSE by up to 25.1 points, or 60% (range: 1.4%, 59.8%). The relative predictive power of each variable, as given by their standardized model coefficients, at each week in the out-of-sample predictions, is displayed in a collection of heatmaps in Figure A.2.

2.4 Discussion and Conclusion

We have shown that Internet-based data sources can be used to track and forecast estimates of suspected weekly Zika cases, weeks ahead of the publication of official case counts. Models that rely exclusively on Google searches have among the lowest error (rRMSE) of all models, indicating that Google search terms alone have the potential to track Zika cases. The heatmaps shown in Figure A.6 confirm that Google search terms have significant predictive power in most

countries and time horizons.

In Colombia and Venezuela, where robust Twitter data were available, we found that Twitter improved predictions compared to models that lacked the data source. Meanwhile, though HealthMap news reports have been found to be good estimators of Zika cumulative incidence[20], the effect of incorporating HealthMap news reports into our models was marginal across countries and generally did not reduce prediction error in any of the weeks-ahead forecasts; where it did reduce prediction error, in El Salvador, it did by less than 2% compared to the next-best model lacking HealthMap data. We noted early evidence of HealthMap's weak predictive power in its low correlation with official case counts, as shown in Figures A1-5. Likewise, the heatmaps of Figure A.6 reveal that news reports data generally had low influence in models after the first several weeks of out-of-sample predictions. We noted, however, in a post-hoc analysis, that news of Zika infections were 2-3 weeks delayed with respect to the time when cases had occurred. This fact suggests that in the absence of official case count reports, one may use (a potentially lagged version) of news reports to track Zika activity as found previously by[20]. In the future, we would expect to improve model predictions by incorporating HealthMap data lagged back in time by 2-3 weeks.

As seen in flu forecasting studies[32], the quality of predictions decreased as the time horizon of prediction increased. Specifically, for one-week predictions, we found that the model that uses Google (and Twitter where available) combined with autoregressive terms (the ARGO+T model) performs best in most countries, and its performance is better than or comparable to the equivalent model that lacks

autoregressive information. Thus, the use of historical case information (autoregressive terms) improves predictions in the near future, a finding that has been documented in prior studies[26,30,37]. However, for 2-3 week-ahead predictions, models that use exclusively data from Google and Twitter (G+T), without autoregressive terms, perform best. This is likely because the 2-3 week old official case information is no longer crucial to refine the accuracy of predictions, and changes in Google search and Twitter activity better respond to fluctuations in Zika dynamics. Consequently, relying on historical case data becomes less useful in making predictions further into the future. This is also observed in the low relevance of lag terms in the 2- and 3-week heatmaps of all models (Figure A.6). Additionally, as automatically identified by our term selection methodology (LASSO), the predictive power of Google search terms is stronger in 1 week-ahead predictions than in 2 and 3 week-ahead predictions. This can be observed in the heatmaps shown in Figure A.6. This finding confirms the appropriateness of using a real-time hidden Markov process as a modeling framework, as discussed in [37]. From this perspective, people affected by Zika will search for Zika-related terms when affected by the virus or when they may suspect risk of exposure to it. This population search behavior suggests that monitoring search activity may help track disease incidence. The decreased relevance of search activity in 2 and 3 week-ahead predictions may suggest that autoregressive case count information may have a stronger role in future occurrences.

Our models improve upon prior methodologies[32,33,38] that use internet-based data sources to track flu by adding an internal dynamic variable

transformation process to reassess the relationship of all input variables with the official Zika case count each week. Indeed, the heatmaps of variable coefficients show that model forecasts depended on an ensemble of terms whose predictive power changed magnitude and direction week by week. Given that Google queries were selected on the basis of their relationship to case data or to the term "Zika" exclusively in the training period, it is likely that these relationships change and perhaps even weaken in later weeks. We thus emphasize the importance the need for dynamic transformation of the input variables to recursively reassess these relationships and readjust predictors to their best linear fit with the data.

Some of the limitations of our approach include, for example, the inherent population biases of Internet search engines and Twitter microblog users. Internet searches patterns may also reflect media coverage and situational awareness that may not coincide with the dynamics of the disease being tracked. Also, different countries and locations frequently have distinct news reporting practices. Local media in regions with endemic mosquito-borne diseases may react differently to outbreaks than regions where these diseases are less frequent. Media attention thus has the potential to dramatically influence our weekly predictions. The dynamic reassessment of the predictive power of each input variable, via LASSO and the dynamic transformation approach discussed earlier, is built in our model to mitigate these events. Terms that may peak during a week of high media attention can be thrown out of the influence of the model for the subsequent week of prediction if their relationship with case count information has weakened. Only the terms with high predictive power are selected by the LASSO. In this way, our models are self-

correcting. Nonetheless, we note that since our predictions rely largely on user search and media activity, our work is meaningful only in time periods when the population is aware of the disease; to this point, it has been demonstrated that Zika virus was introduced to Brazil and the Americas at least one year before the epidemic was recognized by health ministries and the public at large[49].

Another important consideration is the time lag between peaks in Zika virus incidence and microcephaly, of up to 5 months[50,51]. Our models capture search activity surrounding the Zika epidemic, and thus end up using search terms like "microcephaly" as input. These terms may be related to broader awareness of Zika activity. Given the estimated lag, however, evaluating microcephaly-related queries synchronously with cases has the potential to introduce a bias in the model. Further work must explore the effect of lagging these terms compared to our synchronous use of them.

As mentioned in the Methods section, Twitter data was not sufficient for use in the models for all countries. To improve upon this, future work could explore keyword queries that incorporate symptoms of Zika infection. In addition, to increase the total volume of tweets we plan to collect historical data based on these new query strings and explore ways to geocode the data ourselves, instead of relying on the current Twitter-generated subset of tweets with coordinate information.

Another challenge lies in the prediction of very low case numbers. In several weeks of the countries studied, official case counts of Zika fell below 50 suspected cases per week; this is very low relative to the thousands of cases experienced per

week at the height of the epidemics. We observe that the quality of predictions decreases during time periods with low case numbers, and the model tends to under-predict cases. Our prediction approaches worked best in locations with highest Zika incidence, independently of Internet penetration. This tendency was also observed in the assessment of the Google Dengue Trends system in[38][45].

Limitations on the use of official suspected case counts from PAHO as our prediction goal include under-reporting. Indeed, Zika has been observed to be asymptomatic in at least 80% of infected persons[42]. As a consequence, our models likely underestimate the true number of Zika infections that exist, while reasonably estimating the actual number of suspected cases that seek medical attention. Unfortunately, no surveillance system has yet reported estimates of asymptomatic Zika infections, and it is unclear whether asymptomatic infections can result in the same consequences of birth and neurological defects as do symptomatic infections.

The predictions of our model should be compared to those of SIR-type models and epidemiologic models that evaluate Zika incidence in the context of important, known drivers of Zika, such as climate and ecological factors. In this paper, we explore whether digital data streams are viable estimators of Zika cases. In future inquiry, we believe that these methods could be incorporated into, and enhance, traditional epidemiologic methods to track the virus.

Given the need of early interventions to curb mosquito-borne disease transmission, our model predictions fill a critical time-gap in existing Zika surveillance since official case count reports will, most likely, continue being

published multiple weeks after the occurrence of Zika cases. Moreover, access to real-time and likely future estimates of Zika activity provide an opportunity for health and government officials to allocate resources differently when potential changes in Zika dynamics are likely to occur, even ahead of official case documentation. The models presented here show promise to be expanded to any country at any time to track Zika cases and signal changes in transmission for public health decision-makers. Our models currently predict Zika activity at the country level, which we feel is useful for national decision-makers and surveillance purposes; however, our methodology can be extended to finer spatial units, such as the regional or municipal level. Performing predictions with higher spatial resolution will allow more targeted interventions and allocation of resources to the areas with the greatest projected burden of disease.

To produce these predictions in a publicly available and timely manner, we will work to create a website that displays Zika estimates for multiple countries continuously updated in real-time, similar to content published on www.healthmap.org/flutrends and www.healthmap.org/denguetrends.

2.5 REFERENCES

[1] PAHO. Cumulative Zika suspected and confirmed cases reported by countries and territories in the Americas, 2015-2016 [Internet]. 8 Nov 2016. Available: http://ais.paho.org/phip/viz/ed_zika_epicurve.asp

[2] Hills SL, Kate R, Morgan H, Charnetta W, Oster AM, Marc F, et al. Transmission of Zika Virus Through Sexual Contact with Travelers to Areas of Ongoing Transmission — Continental United States, 2016. *MMWR Morb Mortal Wkly Rep.* 2016;65. doi:10.15585/mmwr.mm6508e2er

[3] CDC. Zika: Transmission & Risks [Internet]. 25 Jul 2016. Available: http://www.cdc.gov/zika/transmission/index.html

[4] CDC. Zika: Symptoms [Internet]. 28 Jun 2016. Available:

http://www.cdc.gov/zika/symptoms/symptoms.html

[5] de Oliveira CS, da Costa Vasconcelos PF. Microcephaly and Zika virus. *J Pediatr* . 2016;92: 103–105.

[6] Dyer O, Owen D. US agency says Zika virus causes microcephaly. *BMJ*. 2016; i2167.

[7] Vogel G, Gretchen V. Zika virus discovered in infant brains bolsters link to microcephaly. *Science*. 2016; doi:10.1126/science.aaf4040

[8] Hu Y-F, Wu J, Huang D-Y, Ma J-T, Ma Y-H. Available Evidence of Association between Zika Virus and Microcephaly. *Chin Med J* . 2016;129: 2347.

[9] WHO. Zika Virus Situation Report [Internet]. 8 Sep 2016. Available: http://apps.who.int/iris/bitstream/10665/250049/1/zikasitrep8Sep16-eng.pdf?ua=1

[10] WHO. WHO Director-General summarizes the outcome of the Emergency Committee regarding clusters of microcephaly and Guillain-Barré syndrome [Internet]. 1 Feb 2016. Available: http://www.who.int/mediacentre/news/statements/2016/emergency-committee-zika-microcephaly/en/

[11] Achee NL, Gould F, Perkins TA, Reiner RC Jr, Morrison AC, Ritchie SA, et al. A critical assessment of vector control for dengue prevention. *PLoS Negl Trop Dis*. 2015;9: e0003655.

[12] Scott TW, Morrison AC. Vector Dynamics and Transmission of Dengue Virus: Implications for Dengue Surveillance and Prevention Strategies. *Current Topics in Microbiology and Immunology*. 2009. pp. 115–128.

[13] Olkowski S, Stoddard ST, Halsey ES, Morrisson AC, Barker CM, Scott TW. Sentinel versus passive surveillance for measuring changes in dengue incidence: Evidence from three concurrent surveillance systems in Iquitos, Peru [Internet]. 2016. doi:10.1101/040220

[14] PAHO. Cumulative Zika suspected and confirmed cases reported by countries and territories in the Americas, 2015-2016 [Internet]. 8 Nov 2016. Available: http://ais.paho.org/phip/viz/ed_zika_epicurve.asp

[15] Polgreen PM, Chen Y, Pennock DM, Nelson FD. Using internet searches for influenza surveillance. *Clin Infect Dis*. 2008;47: 1443–1448.

[16] Ginsberg J, Mohebbi MH, Patel RS, Brammer L, Smolinski MS, Brilliant L. Detecting influenza epidemics using search engine query data. *Nature*. 2009;457: 1012–1014.

[17] Yuan Q, Qingyu Y, Nsoesie EO, Benfu L, Geng P, Rumi C, et al. Monitoring Influenza Epidemics in China with Search Query from Baidu. *PLoS One*. 2013;8: e64323.

[18] Althouse BM, Ng YY, Cummings DAT. Prediction of Dengue Incidence Using Search Query Surveillance. *PLoS Negl Trop Dis*. 2011;5: e1258.

[19] Majumder MS, Kluberg S, Santillana M, Mekaru S, Brownstein JS. 2014 ebola outbreak: media events track changes in observed reproductive number. *PLoS Curr.* 2015;7. doi:10.1371/currents.outbreaks.e6659013c1d7f11bdab6a20705d1e865

[20] Majumder MS, Santillana M, Mekaru SR, McGinnis DP, Khan K, Brownstein JS. Utilizing Nontraditional Data Sources for Near Real-Time Estimation of Transmission Dynamics During the 2015-2016 Colombian Zika Virus Disease Outbreak. *JMIR Public Health Surveill*. 2016;2: e30.

[21] Brownstein JS, Freifeld CC, Reis BY, Mandl KD. Surveillance Sans Frontières: Internet-based emerging infectious disease intelligence and the HealthMap project. *PLoS Med.* 2008;5: e151.

[22] Santillana M, Nsoesie EO, Mekaru SR, Scales D, Brownstein JS. Using clinicians' search query data to monitor influenza epidemics. *Clin Infect Dis*. 2014;59: 1446–1450.

[23] Smolinski MS, Crawley AW, Baltrusaitis K, Chunara R, Olsen JM, Wójcik O, et al. Flu Near You: Crowdsourced Symptom Reporting Spanning 2 Influenza Seasons. *Am J Public Health.* 2015;105: 2124–2130.

[24] Paolotti D, Carnahan A, Colizza V, Eames K, Edmunds J, Gomes G, et al. Web-based participatory surveillance of infectious diseases: the Influenzanet participatory surveillance experience. *Clin Microbiol Infect.* 2014;20: 17–21.

[25] Dalton C, Durrheim D, Fejsa J, Francis L, Carlson S, d'Espaignet ET, et al. Flutracking: a weekly Australian community online survey of influenza-like illness in 2006, 2007 and 2008. *Commun Dis Intell Q Rep*. 2009;33: 316–322.

[26] Paul MJ, Dredze M, Broniatowski D. Twitter improves influenza forecasting. *PLoS Curr.* 2014;6. doi:10.1371/currents.outbreaks.90b9ed0f59bae4ccaa683a39865d9117

[27] Broniatowski DA, Paul MJ, Dredze M. National and local influenza surveillance through Twitter: an analysis of the 2012-2013 influenza epidemic. *PLoS One.* 2013;8: e83672.

[28] Nagar R, Yuan Q, Freifeld CC, Santillana M, Nojima A, Chunara R, et al. A case study of the New York City 2012-2013 influenza season with daily geocoded Twitter data from temporal and spatiotemporal perspectives. *J Med Internet Res.* 2014;16: e236.

[29] Signorini A, Segre AM, Polgreen PM. 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.* 2011;6: e19467.

[30] Santillana M, Nguyen AT, Louie T, Zink A, Gray J, Sung I, et al. Cloud-based Electronic Health Records for Real-time, Region-specific Influenza Surveillance. *Sci Rep.* 2016;6: 25732.

[31] Nsoesie EO, Patrick B, Naren R, Mekaru SR, Brownstein JS. Monitoring Disease

Trends using Hospital Traffic Data from High Resolution Satellite Imagery: A Feasibility Study. *Sci Rep.* 2015;5: 9112.

[32] Santillana M, Mauricio S, Nguyen AT, Mark D, Paul MJ, Nsoesie EO, et al. Combining Search, Social Media, and Traditional Data Sources to Improve Influenza Surveillance. *PLoS Comput Biol*. 2015;11: e1004513.

[33] Santillana M, Zhang DW, Althouse BM, Ayers JW. What can digital disease detection learn from (an external revision to) Google Flu Trends? *Am J Prev Med*. 2014;47: 341–347.

[34] Cook S, Samantha C, Corrie C, Fowlkes AL, Mohebbi MH. Assessing Google Flu Trends Performance in the United States during the 2009 Influenza Virus A (H1N1) Pandemic. *PLoS One.* 2011;6: e23610.

[35] Olson DR, Konty KJ, Paladini M, Viboud C, Simonsen L. Reassessing Google Flu Trends data for detection of seasonal and pandemic influenza: a comparative epidemiological study at three geographic scales. *PLoS Comput Biol.* 2013;9: e1003256.

[36] Soules A, Aline S. I Google, You Google, We Google. Against the Grain. 2013;20. doi:10.7771/2380-176x.2734

[37] Yang S, Santillana M, Kou SC. Accurate estimation of influenza epidemics using Google search data via ARGO. *Proc Natl Acad Sci U S A.* 2015;112: 14473–14478.

[38] Yang S, Kou SC, Lu F, Brownstein J, Santillana M. Advances in the use of Google searches to track dengue in Mexico, Brazil, Thailand, Singapore, and Taiwan. arXiv:161202812. 2016; Available: https://arxiv.org/abs/1612.02812

[39] PAHO. Regional Zika Epidemiological Update [Internet]. 2016. Available: http://www.paho.org/hq/index.php?option=com_content&id=11599&Itemid=41691

[40] Agence Régionale de Santé. Points épidémiologiques Zika [Internet]. 2016. Available: http://www.martinique.pref.gouv.fr/Publications/Dossiers/L-epidemie-de-Zika-en-Martinique2/Points-epidemiologiques-Zika

[41] Instituto Nacional de Salud. Reporte de análisis de Zika [Internet]. 2016. Available: http://www.ins.gov.co/noticias/paginas/zika.aspx#.V7HrsZMrKYU

[42] Duffy MR, Chen T-H, Hancock WT, Powers AM, Kool JL, Lanciotti RS, et al. Zika virus outbreak on Yap Island, Federated States of Micronesia. *N Engl J Med.* 2009;360: 2536–2543.

[43] PAHO. Zika Virus Case Definitions [Internet]. 18 Apr 2016. Available: http://www.paho.org/hq/index.php?option=com_content&view=article&id=11117&Itemid=41532&lang=en

[44] Tibshirani R, Robert T. Regression shrinkage and selection via the lasso: a retrospective. *J R Stat Soc Series B Stat Methodol.* 2011;73: 273–282.

[45] Gluskin RT, Johansson MA, Santillana M, Brownstein JS. Evaluation of Internet-based dengue query data: Google Dengue Trends. *PLoS Negl Trop Dis.* 2014;8: e2713.

[46] Johansson MA, Reich NG, Hota A, Brownstein JS, Santillana M. Evaluating the performance of infectious disease forecasts: A comparison of climate-driven and seasonal dengue forecasts for Mexico. *Sci Rep.* In press;

[47] Friedman J, Hastie T, Tibshirani R. Regularization Paths for Generalized Linear Models via Coordinate Descent. *J Stat Softw.* 2010;33: 1–22.

[48] R Core Team. R: A Language and Environment for Statistical Computing [Internet]. Vienna, Austria: R Foundation for Statistical Computing; 2016. Available: http://www.R-project.org

[49] Faria NR, Azevedo R do S da S, Kraemer MUG, Souza R, Cunha MS, Hill SC, et al. Zika virus in the Americas: Early epidemiological and genetic findings. *Science.* 2016;352: 345–349.

[50] Paploski IAD, Prates APPB, Cardoso CW, Kikuti M, Silva MMO, Waller LA, et al. Time Lags between Exanthematous Illness Attributed to Zika Virus, Guillain-Barré Syndrome, and Microcephaly, Salvador, Brazil. *Emerg Infect Dis.* 2016;22: 1438–1444.

[51] Reefhuis J, Gilboa SM, Johansson MA, Valencia D, Simeone RM, Hills SL, et al. Projecting Month of Birth for At-Risk Infants after Zika Virus Disease Outbreaks. *Emerg Infect Dis.* 2016;22: 828–832.

3

Combining weather patterns and cycles of population susceptibility to forecast dengue fever epidemic years in Brazil: a dynamic, ensemble learning approach

3.0 ABSTRACT

A major challenge in disease forecasting is developing algorithms that can learn from complex disease dynamics, identifying patterns and signals with minimal human input. Focusing on the important but complex relationship between dengue fever outbreaks and (a) weather patterns (temperature, rainfall) and (b) the empirically observed 3-4 year disease burden cycles, we present a data-driven, machine learning approach capable of autonomously and continuously identifying weather patterns and cycles in population susceptibility to predict dengue fever outbreak years in Brazilian municipalities. Specifically, our approach is dynamic, adaptable to multiple and heterogeneous study areas with high spatial resolution, and leverages publically available data sources, including a globally-available meteorological data source. We produce annual retrospective, out-of-sample epidemic forecasts months ahead of the historically-observed seasonal onset of dengue epidemics, and show that using just two simple weather inputs can yield good forecast accuracy, and further improve when combined with learned cycles of dengue fever outbreaks, a proxy for population infection susceptibility.

3.1 INTRODUCTION

The last decade has seen enormous advances in the way data is generated and collected, resulting in large volumes of complex information known as "Big Data." The recent availability of these data has opened up new avenues for epidemic monitoring, with data streams such as satellite imagery[1,2], Google searches[3,4], mobile phones [5,6], genomics[7,8], and disease surveillance databases[9,10] providing rich sources of information on the causes and outcomes of disease, population behaviors, environmental conditions, and other potential signals of population health. Exploiting these relationships to generate accurate, prospective forecasts would benefit health systems by allowing early mobilization of resources for the prevention of morbidities and deaths in the face of public health threats. However, a major challenge in disease forecasting is developing algorithms that can autonomously and continuously learn from these complex, dynamic systems, identifying patterns and signals with minimal human input.

One such complex system is the interplay of human, climate, and mosquito dynamics that give rise to the transmission of mosquito-borne diseases. Dengue fever, a viral mosquito-borne disease transmitted predominately by the *Aedes aegypti* and *Aedes albopictus* mosquitoes, infects an estimated 390 million people per year, with nearly half the world's population living at risk of infection[11]. The global burden of dengue has doubled every 10 years over the last 3 decades[12], and the disease is projected to expand its latitude range as global temperatures increase and create new habitats for the *Aedes* mosquitoes among previously-unexposed human populations[13]. Climate, in particular temperature and

precipitation, creates favorable conditions for the breeding and survival of *Aedes* mosquitoes as well as for the transmission of the dengue fever virus. Distinct ranges of temperature and precipitation have been observed to have an influence on the extrinsic incubation period[14,15], mosquito maturation rate[16], length of larval hatch time[17], survival rate[18], and biting rate[19]. However, the relationships that govern these parameters and give rise to dengue transmission are complex and dynamic, changing over time and across geographies. Moreover, multi-year cycles of dengue fever outbreaks, caused by one or more circulating dengue fever serotypes (DENV I, II, III, IV) and short-term immunity conferred after infection, add an important layer of complexity to prediction[20].

The dengue forecasting literature lacks a systematic, self-adaptive, and reproducible approach capable of identifying weather patterns that may be predictive of dengue fever outbreaks, particularly at a local level. Vector-borne diseases commonly exhibit spatial heterogeneity, a result of spatial variation in vector habitat, weather patterns, and human control actions[21–24]. For developing forecast systems, this feature implies a trade-off between model generalizability and spatial resolution. As a consequence, most studies to date focus on producing predictions for a single location, ranging from the national- to the city-level[25–27], while others build and evaluate multiple modeling strategies per study site in efforts to manually identify relationships between weather patterns and dengue incidence over different geographies and temporal windows[28,29]. Both approaches highlight the difficulty in producing forecast models that are viable in diverse settings. In contrast, data-driven techniques demonstrate promise by

learning from multi-scale, complex systems and automatically adapting to new information. A recent study showed the promise of a data-driven approach in identifying weather patterns with meaningful signals for dengue fever outbreaks, but not in an out-of-sample fashion[30].

3.1.1 Our Contribution

Focusing on the important but complex relationship between dengue incidence and (a) weather patterns and (b) the empirically observed 3-4 year disease burden cycles, we present a data-driven, machine learning approach capable of autonomously and continuously identifying weather patterns and cycles in population susceptibility to predict dengue fever outbreak years in Brazilian municipalities. Specifically, our approach is dynamic, adaptable to multiple and heterogeneous study areas with high spatial resolution, and leverages publicly available data sources, including a globally-available meteorological data source. We produce annual retrospective, out-of-sample epidemic forecasts at the city-level, months ahead of the historically-observed seasonal onset of dengue epidemics. We assess the feasibility of this autonomous learning approach using two simple weather inputs (temperature, rainfall) in 20 Brazilian cities with diverse microclimates, and we attempt to characterize the conditions that yield successful forecasts.

3.2 Results

3.2.1 Exploiting weather signals to create a data-driven forecast system

We obtained data on annual dengue fever cases (Brazilian Ministry of Health) for 2001-2017 and on daily temperature and precipitation (GMAO-NASA)

for 2000-2016, for 20 dengue-endemic municipalities spanning large geographic and population ranges (Fig. B.1, Table B.1). Weather patterns were extracted and analyzed across hundreds of partially-overlapping time intervals collectively spanning the last 7 months of a given year, a time period that typically precedes the onset of epidemic outbreaks in Brazil. These patterns were then assessed for their ability to predict an outbreak year (defined as a year in which the number of cases exceeds 100 per 100,000 persons) for the subsequent year. Retrospective, out-of-sample forecasts trained on a yearly expanding window were produced for 10 years (2008-2017) and for each time interval using support vector machines, a binary classifier (Fig. 3.1A-C). Every year, the top predictive time intervals were automatically selected to participate in an ensemble voting system based on historical out-of-sample prediction accuracy (Fig. 3.1D-E). In order to accrue enough out-of-sample prediction years to input to the ensemble voting system, we used the first 4 years of out-of-sample predictions to inform ensemble model selection, and produced ensemble-based predictions for the remaining 6 years (2012-2017).

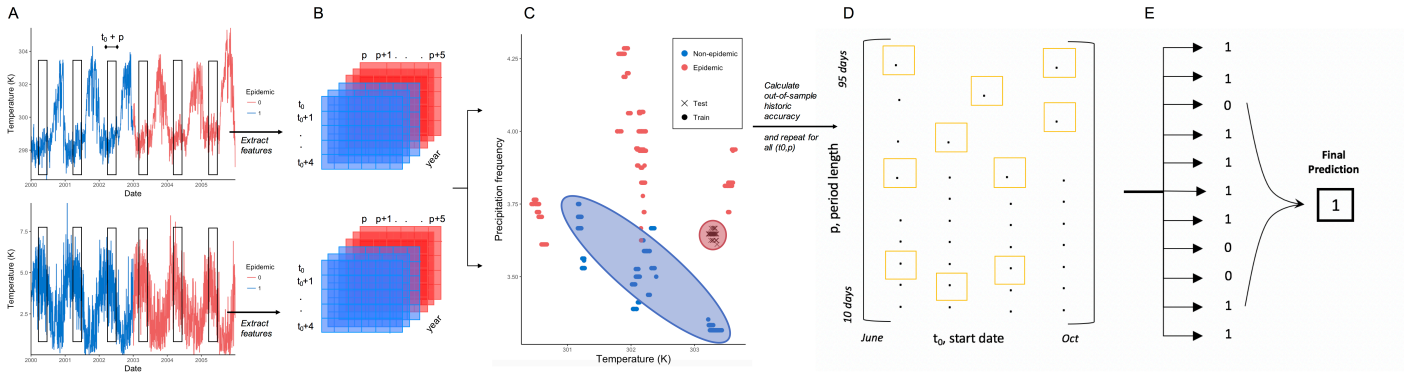

Figure 3.1. Ensemble forecast workflow. (A) To predict next year's epidemic status, we extract features from a daily time series of temperature (K) and precipitation (mm) over a defined (t0, p) time interval and for each year in the training period. (B) We produce an array of features corresponding to the mean value of temperature and precipitation over the (t0, p) interval, and (C) train a support vector machine to classify next year's epidemic status. (D) This process is repeated for all 432 (t0, p) intervals, and the top 11 models are automatically selected to (E) contribute to a majority voting system based on historic out-of-sample accuracy.

This system, which autonomously identifies and exploits the predictions of multiple time windows during the calendar year, makes it possible to identify temporally similar regions of highly predictive periods of the year preceding dengue outbreaks, here referred to as "weather signatures." Weather signatures represent similar time windows during the year that show consistently high out-of-sample prediction accuracy. We observed that cities with higher ensemble (2012-2017) prediction accuracy tended to have clear weather signatures, while cities with poor performance exhibited no specific tendencies (Figs. 3.2, 3.3A). Further, strong weather signatures in a city often corresponded to or preceded important tropical seasons, such as the rainy and dry seasons.

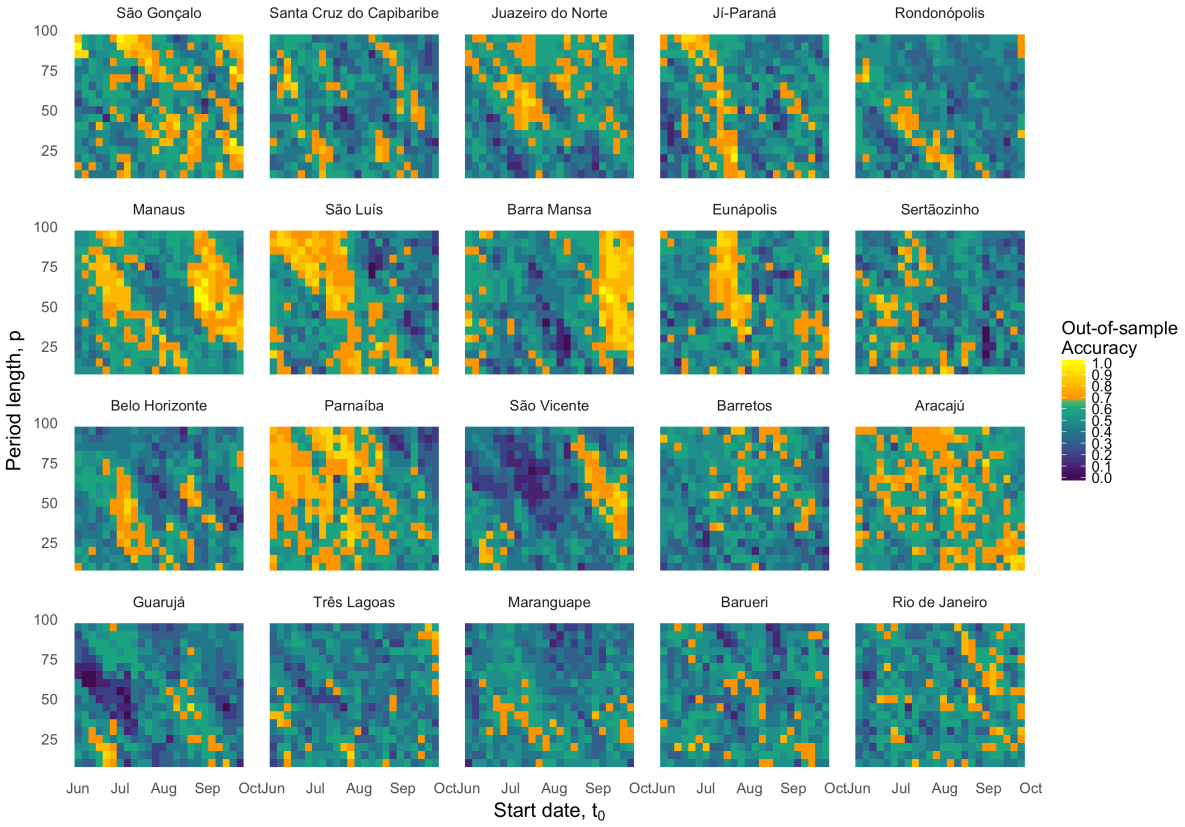

Figure 3.2. The 10-year (2008-2017) out-of-sample forecast accuracy (%) for each time window of temperature and precipitation, by municipality. The x-axis (t0) indicates the start date of the time interval, and the y-axis (p) indicates the length of the time interval from which weather data were gathered (10-95 days). Models achieving at least 7/10 correct out-of- sample forecasts are shown in shades of yellow. Municipalities are ordered by decreasing ensemble prediction accuracy; that is, the proportion of years correctly forecasted by the ensemble method over the years 2012-2017.

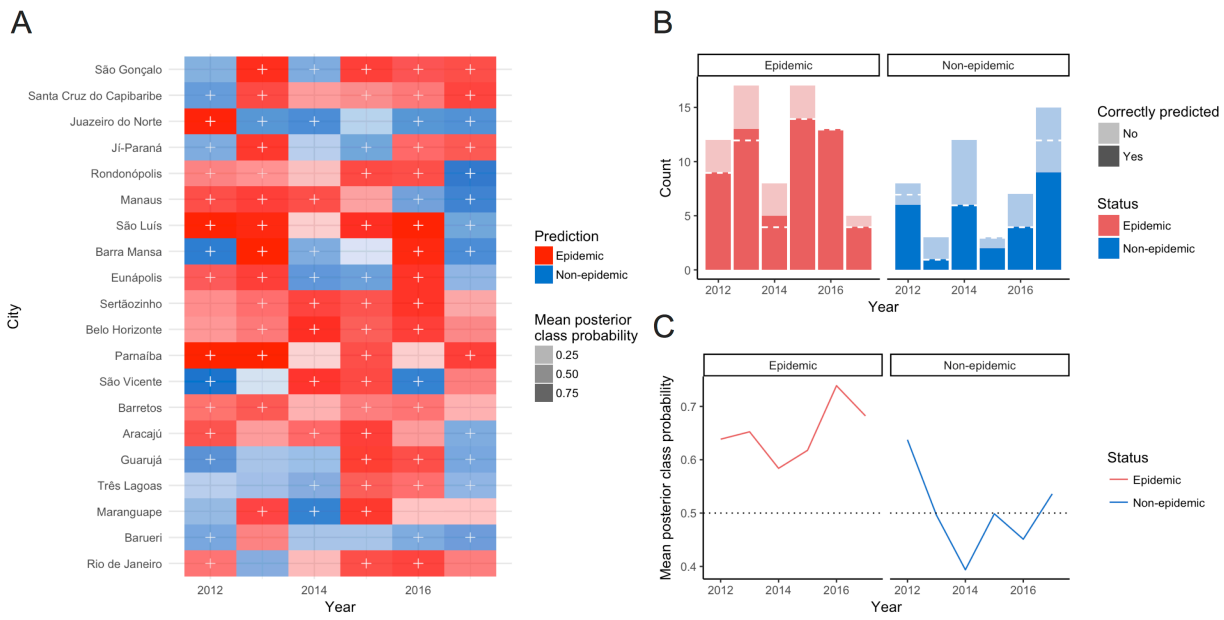

Figure 3.3. Weather-based prediction results for 120 municipality-years.

A) Annual out-of-sample forecasts of outbreak status (epidemic/non-epidemic) for 20 Brazilian municipalities from 2012-2017, shaded by mean posterior probability (MPP) of the true outbreak status. Correct forecasts are indicated by a plus (+) sign, and cells with light shading indicate that the model predicted the correct class with low probability. Municipalities are ordered by decreasing ensemble prediction accuracy; that is, the proportion of years correctly forecasted by the ensemble method over the years 2012-2017.

B) The number of total epidemic and non-epidemic years correctly forecasted across 20 municipalities, by year. The dashed white line indicates the number correctly forecasted after incorporation of empirically-observed dengue cycles.

C) The mean posterior class probability across municipalities, by year and epidemic status.

3.2.2 WEATHER-BASED FORECASTING PERFORMANCE

Using weather data (temperature and precipitation) alone to predict annual dengue outbreaks, our approach accurately forecasted 81% of all epidemic years across 20 municipalities in Brazil between 2012-2017 (Table 3.1, Fig. 3.3). The approach only identified 58% of non-epidemic years correctly. This resulted in an *overall accuracy* of approximately 72%. For reference, the frequency of epidemic

and non-epidemic years was 60% and 40%, thus, a naive approach that predicts that all years are epidemic (the class majority) would achieve an accuracy of 60%. Our approach significantly exceeded (p=0.005) the predictive power of a naive predictor.

Table 3.1. Performance of weather-based out-of-sample forecasts across 120 municipality-years in Brazil, with and without consideration for DENV susceptibility cycles.

Evaluation Metric	Weather	Weather + DENV Cycle
Accuracy	71.70%	75%
Hit rate (Sensitivity)	81%	78%
Non-epidemic detection rate (Specificity)	58%	71%
No-information rate	60%	60%
P(Accuracy > No-Information Rate)	p = 0.005	p=0.0004

3.2.3 INCORPORATING DENGUE SUSCEPTIBILITY CYCLES

Our weather-based ensemble approach remains ignorant to the specific relationship between weather patterns and dengue outbreaks, instead allowing the data to drive model selection and predictions. However, endemic transmission of dengue fever is typically distinguished by periodic outbreak cycles of around 3-4 years. These outbreak cycles are thought to occur as a result of 1) an exhaustion of the susceptible population after an outbreak, and 2) and short-term cross-immunity to other circulating DENV serotypes after infection[20]. Both factors result in a depletion of the population vulnerable to infection, and act as barriers to subsequent outbreaks. Independent of climate variability over the years, we expect some preservation of these susceptibility cycles.

Inspired by this phenomenon, we implemented a post-hoc decision rule incorporating empirical information on 3- and 4-year dengue fever cycles observed in endemic municipalities in Brazil. We computed the probability of transitioning between outbreak states (epidemic/non-epidemic) after 2 and 3 consecutive years as the mean second- and third-order Markov transition probabilities, respectively, across municipalities meeting endemic selection criteria (B.1.1 Supplemental Materials and Methods: "Study Sites"). This Markov transition matrix was computed using the first 11 years of data preceding the first ensemble out-of-sample predictions. For each prediction year, a winner-takes-all decision rule overturned the ensemble prediction if the probability of a specific transition to one class exceeded the percent of model votes for the opposite class.

3.2.4 DENGUE CYCLES IMPROVE UPON WEATHER-BASED FORECASTS

Compared to the exclusively weather-based approach, incorporating these empirically-observed dengue cycles improved the ability to predict non-epidemic years by approximately 20% (specificity=69%) and increased overall accuracy to 74.2% (Table 1). This improvement is the consequence of the decision rule's role in identifying and overturning specifically *epidemic* forecasts following a sequence of observed epidemic years. The decision rule replaced 7 epidemic forecasts with non-epidemic forecasts, of which 5 were correct (Fig. 3.3B). A majority of these overturns belonged to cities which had experienced 3 consecutive epidemic years leading up to the prediction.

The decision rule is intended to serve as an "expert opinion" for situations in which there is strong evidence for the transition between outbreak states after multiple consecutive years of one state. Our specific finding - that the dengue cycles were used exclusively to overturn epidemic forecasts - suggests that while the weather conditions in those locations and years were identified to be conducive to an outbreak, there was stronger evidence that the population may have had low susceptibility to infection (thus avoiding an outbreak), based on multiple consecutive preceding years of high disease incidence.

3.2.5 Model performance by year

The success of epidemic forecasts varied by year, suggesting that certain years were better suited for weather-based outbreak predictions, while other years may have been outliers for either dengue activity or weather conditions and thus were more difficult to predict. During the last three years of the time series (2015-2017), epidemics were predicted by the weather-only models with at least 80% accuracy, with 100% of the 13 outbreaks in 2016 correctly forecasted (Fig. 3.3B,C). Conversely, non-epidemic years during 2013-2014 were particularly difficult to predict, with only one-third and one-half of cities correctly forecasting non-epidemics for these years, respectively. The most successful non-epidemic predictions occurred in 2012, for which 6 out of 8 non-epidemics (75%) were predicted correctly. Overall, 2015 and 2016 were the most successfully classified years, with 80% and 85% of municipalities correctly classified as epidemics or non-

epidemics, respectively, while 2014 and 2017 were the most difficult years to predict, with 45% and 35% of municipalities misclassified, respectively.

Incorporating information on the dengue cycles helped detect an additional non-epidemic in 2012 and 2015, and an additional 3 non-epidemics in 2017 (Fig. 3.3B).

3.2.6 QUANTIFYING THE STRENGTH OF PREDICTIONS

Because our forecast system produces deterministic binary predictions (epidemic/non-epidemic year) using support vector machine classifiers, a natural question is how to quantify the strength of each prediction. In particular, this issue is important to identify the conditions under which predictions are made with strong conviction, and whether the strength of a prediction corresponds to its accuracy. For instance, misclassification of epidemics based on weather conditions may be the consequence of several factors, including: anomalous weather years, non-weather factors that contribute more strongly to epidemic status, and poor distinction (separability) between epidemic and non-epidemics in the training data, possibly the result of a limited time series. Understanding the strength with which incorrect predictions were made is thus uniquely of interest. We explored ways to characterize the strength of predictions based on both the historic strength of the selected ensemble generating the prediction, as well as the strength of the weather-based classifiers themselves.

As a reminder, the ensemble that predicts epidemic status for each city and for each year is composed of multiple time windows (each its own model) that have

consistently exhibited the highest out-of-sample prediction performance compared to the rest of the calendar year. The consistency of time window selection into the ensemble is represented in Fig. 3.4. In our framework, time windows are automatically selected into the final ensemble as a function of both 1) their own historic out-of-sample performance and 2) the historic performance of their calendar neighbors, that is, models representing temporally similar windows. Consequently, we computed a metric of ensemble strength that captures both of these elements (B.1.5), and observed that the strongest ensembles belonged to cities with clear temporal patterns (as shown in Figs. 3.3, 3.4), which in turn were among the best-performing cities (Fig. B.2). In other words, we found that cities that perform better tend to make predictions based on similar periods of high-performing windows (consistency), rather than based on several temporally disparate but high-performing windows (inconsistency). Still, ensemble strength only represents the historic performance of time windows (how well time windows have predicted in previous years), which is an incomplete characterization of prediction strength; for instance, models with good historic prediction performance may still fail if the weather data for the upcoming year are not separable and/or provide evidence for the wrong outbreak state.

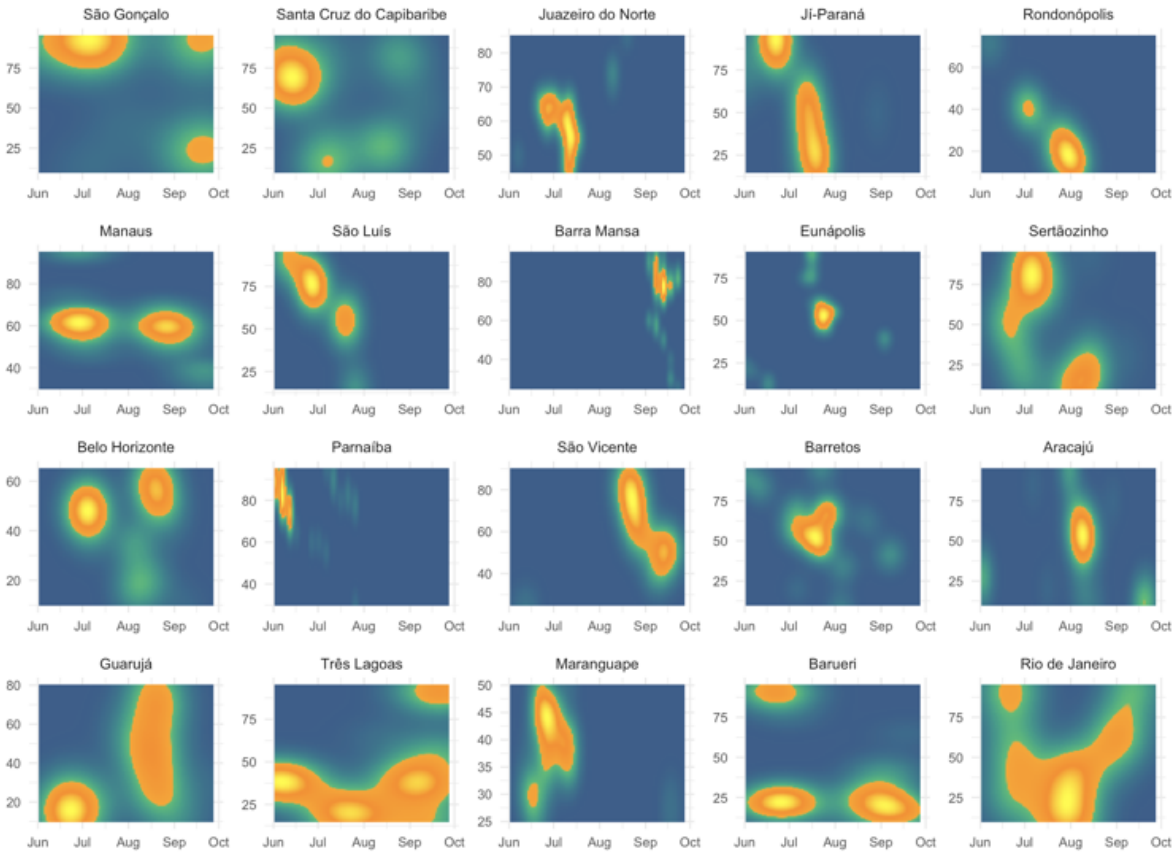

Figure 3.4. Periods of the year selected into the ensemble forecast model for 2012-2017, by municipality. The x-axis (t0) indicates the start date of the time interval, and the y-axis (p) indicates the length of the time interval from which weather data were gathered (10-95 days). Municipalities with smaller and brighter yellow centers are those which exhibit the highest consistency in the predictive performance of weather patterns. Municipalities are ordered by decreasing ensemble prediction accuracy; that is, the proportion of years correctly forecasted by the ensemble method over the years 2012-2017.

Consequently, to assess the extent to which available weather evidence were able to distinguish between epidemic vs. non-epidemic years for different cities and years (the strength of the classifier), we extracted calibrated posterior probabilities of each SVM model using Platt's scaling[31]. The posterior probability reflects the distance to the separation boundary distinguishing epidemic and non-epidemic years on the basis of weather. Thus, a higher probability

represents how strongly the weather patterns of the prediction year aligned with those experienced by prior outbreak or non-outbreak years. We observed that in general, the probabilities were well calibrated, i.e. roughly 80% of predictions made with 0.8 probability were epidemics (Fig. B.3); however, the small sample (120 out-of-sample predictions) somewhat limited the ability to assess this feature.

Over the 11 models composing each ensemble SVM, we computed the mean of the posterior probabilities (MPP, Fig. 3.3A) and compared these to the accuracy of the prediction. We found that a majority (75%) of municipality-years were predicted with moderate (0.6-0.8) or strong (0.8-1.0) mean posterior predicted-class probabilities (i.e. P(Epidemic) for predicted epidemics and 1-P(Epidemic) for predicted non-epidemics), with over 70% of these moderate or strong predictions correct (Fig. B.4). However, we also found that many misclassifiers "failed silently," that is, outputted incorrect answers with high confidence[32]. Our results show that over half of missed true epidemics were predicted to be epidemics with less than 0.3 MPP, and likewise, over half of missed non-epidemics were classified as epidemics with over 0.7 MPP. In general, this implies that incorrect predictions were typically the result of strong model conviction against true outbreak status; that is, based on historical climate patterns, these municipality-years had periods of weather conditions conducive to either outbreaks or low dengue activity, but experienced the opposite. In a few cities that showed no strong evidence of weather signatures (i.e. Barueri, Rio de Janeiro; Fig. 3.4), mean posterior probabilities were more borderline (0.4-0.6), suggesting that the climatic distinction between epidemic and non-epidemic years may have been

low in those locations, resulting in low separability in the data and higher occurrence of incorrect predictions.

We therefore endeavor to apply a loose typology to instances of misclassification, in order to better understand the limitations of the present modeling framework. For municipality-years whose epidemic status was misclassified with strong conviction (mean posterior predicted-class probability ≥ 0.8), it is likely that the dengue activity that year was highly anomalous given what had been experienced in historically-similar weather conditions (Fig. B.5). For municipality-years whose epidemic status was misclassified with borderline conviction (i.e. mean posterior predicted-class probability close to 0.5), the error is more likely to be a consequence of insufficient data to discriminate between epidemic and non-epidemic years on the basis of weather patterns alone; that is, the models were not well suited to make this distinction in the first place (Fig. B.5).

Both cases highlight separate limitations of our approach. First, we expect that both a greater variety of environmental variables (e.g. humidity, vegetation, standing water) and non-environmental variables (e.g. human activity and public health interventions) will contribute to more accurate predictions by considering broader factors that contribute to dengue fever activity in a given location. Second, the robustness of our predictions was limited by a short time series of annual information, which may not be adequate to detect true differences in epidemic and non-epidemic years on the basis of weather alone. Nonetheless, our reproducible modeling framework accommodates additional predictors and longer time series

with no additional effort, and thus we highlight these as limitations of only the present analysis, with potential for improved performance in other data settings.

3.2.7 Combining ensemble and classifier strengths

We observed that models with stronger posterior probabilities and weaker ensembles achieved higher accuracy compared to those with weaker probabilities and stronger ensembles, suggesting that the previous year's weather data played a larger role than historic model performance in ultimately determining whether a prediction would be correct (Fig. B.4). There was no clear hierarchy, however, in the ordering of the combined strength of the classifiers and the ensembles; that is, models with strong classifier probabilities and strong ensembles did not necessarily outperform models with weaker classifier probabilities and weaker ensembles.

3.3 Discussion

Here we have demonstrated a novel method to forecast dengue fever outbreak years in Brazil at the smallest administrative unit, using a single, flexible modeling framework and only two simple weather inputs. Our approach automatically learns from the patterns of any inputted series and leverages the best historic predictions to generate an ensemble forecast. Further, by integrating our statistical approach with observed cycles of dengue fever outbreaks as a proxy for population susceptibility, our models achieve higher accuracy and improve substantially in predicting non-epidemic years. These forecasts provide timely information on dengue fever activity to policymakers months ahead of outbreak

seasons. Further, our entirely data-driven models show an ability to learn from complex relationships between dengue epidemics and climatic conditions and identify, in vastly different locations, important weather patterns with potential biological significance. Importantly, these models can be immediately extended to other locations, requiring no location-specific manipulation or inputs aside from a globally-available time series of daily temperature and precipitation.

Using weather information only, our models seek to characterize and exploit the predictive ability of distinct weather patterns preceding outbreak years. Because our framework automatically identifies the time periods for which weather patterns produce strong signals, it was possible to identify weather signatures in multiple locations with vastly different ecosystems and geographic locations. For this, we observed that cities with better overall prediction accuracy had stronger weather signatures, with some biological consistency. For example, the southeastern municipality of Barra Mansa (5 of 6 ensemble years predicted correctly) exhibited strong signals from time windows spanning the first half of the city's rainy season, in October through December of each year. Farther north, the hot, wet, and humid municipality of Manaus (5 of 6 ensemble years predicted correctly), situated at the mouth of the Amazon, appeared to show two distinct weather signatures straddling the driest month of the year, August. These patterns, generated from 10 years of out-of-sample model predictions, suggest that in different regions of Brazil, weather may affect dengue transmission differently and at different times of the year. However, in locations where weather-based predictions were less successful, these signatures were not clear; for instance, Rio

de Janeiro (3 out of 6 ensemble years predicted correctly) showed no clear temporal trend. In cities such as these, we might expect to see a lower influence of weather patterns on transmission compared to other predictors (e.g. policy, behavior, land use). We did not find clear relationships between prediction accuracy and city characteristics such geography, population density, or municipality size. We believe this work may catalyze important research both on the local influence of weather patterns on dengue outbreaks as well as the extent to which other, non-weather factors drive outbreaks in these locations.

Even weather conditions that appear highly suitable for an outbreak (or none), based on historical information, may be challenged by other factors that limit (or encourage) transmission of dengue. A key strength of our approach is the incorporation of empirically-observed information on dengue fever susceptibility cycles, to correct for potential short-term immunity that results from previous exposure to the dengue virus. We found that these susceptibility cycles were critical to the performance of models, particularly those which identified weather patterns suitable for a dengue outbreak in a year with potentially low population susceptibility to infection. For instance, this approach correctly identified 3 additional non-epidemics in 2017 compared to weather patterns alone, supporting the discourse on the unusually low dengue activity seen in Brazil in 2017[33]. Still, our models missed half (6/12) of non-epidemics in 2014, which was predicted by experts to be a low transmission year due to immunity provided by a large 2013 outbreak with no changes in circulating DENV serotypes[34,35]. Thus, incorporating information on specific circulating serotypes could be used to better

detect changes in population immunity and enhance our approach, though this surveillance information is more challenging to routinely acquire. Regardless, here we highlight the importance of incorporating mechanistic processes of disease transmission into data-driven approaches that may be otherwise blinded to them.

Because dengue transmission is driven by multiple complex socioecological and biological factors, we expect our models to capture only a portion of the epidemiologic triangle. Here we show the performance of two simple and relevant weather indicators of dengue fever, but the incorporation of additional weather features (i.e. humidity, vegetation, soil water absorption) combined with a feature selection step may lead to improved accuracy of forecasts, by considering more complex weather conditions preceding dengue outbreaks. Further, weather- and susceptibility-based models can contribute valuable information to larger ensemble approaches leveraging a collection of mobility, sociodemographic, epidemiologic, climatic, and biological information.

Our approach also showcases the feasibility (and limitations) of predicting in a "small data" setting, wherein only 17 outcome data points were available (each representing annual outbreak status between 2001-2017). We chose a short training period (initial 7 years) to maximize the number of out-of-sample predictions, but ultimately it is difficult to establish strong climatic distinctions between outbreak and non-outbreak years in the data with so few samples. Thus, we anticipate improvement in performance for settings that have multiple decades of data, which would allow for longer training periods, improved separability in the data, and more

stable identification of dengue susceptibility cycles, all improving the quality and robustness of predictions.

Ultimately, this framework provides a simple, reproducible method of predicting dengue fever outbreak years in a wide range of locations. Given that the global and economic burden of dengue is placed at an estimated 390 million infections and $8.9 billion per year[11,36], optimizing resource allocation for the disease prevention is critical. However, control of the Aedes mosquito requires weeks or months before effects are seen on the vector population, so predicting dengue outbreaks up to several months of their onset is ideal. Our reproducible approach, which uses of globally-available data with daily resolution, is intended to serve as an unsupervised learning framework to produce early outbreak warnings in any desired context, resulting in more efficient resource mobilization, budgeting, and prevention campaigns. Developing transparent early warning systems at the local level is emerging as a top global health priority, making our contribution both timely and impactful.

3.4 Materials and Methods

We developed a single, flexible modeling framework capable of identifying potentially useful weather patterns to predict dengue fever, and used this to forecast annual outbreak status (epidemic / non-epidemic).

Our workflow, outlined in Fig. B.1, combines elements from signal processing/spectral analysis, machine learning, and ensemble modeling to achieve robust, data-driven epidemic forecasts that do not require any prior knowledge of

the system (i.e. climatic influences on dengue transmission). Our research question is inherently one of time series classification, to forecast epidemic vs. non-epidemic years of dengue fever. The workflow begins with a time series of hourly and daily weather information, which serve as inputs to a collection of classifiers that contribute to ensemble-based epidemic predictions. Our approach can be described in 5 steps:

1. *Signal preprocessing*: for a time series of weather data, define time intervals of varying sizes (10-95 days across the last 7 months of the calendar year), and use a windowing technique to include information within several days of the interval
2. *Time series feature extraction*: extract summary measures for 2 weather variables with known influence on mosquito-borne disease dynamics, temperature and precipitation
3. *Independent model training and prediction*: train a collection of independent support vector machine (SVM) classifiers on historical information from each unique time interval, and generate an out-of-sample epidemic prediction for the following year
4. *Model selection*: choose the best 11 models, representing strongly predictive periods of the year preceding outbreaks, based on a) historical out-of-sample prediction accuracy and b) out-of-sample performance of neighboring time intervals
5. *Ensemble prediction*: determine a final out-of-sample epidemic forecast by majority vote of the selected top models

To potentially enhance the performance of this exclusively weather-based approach, we implemented a post-hoc step incorporating empirical information on 3- and 4-year dengue fever cycles as a proxy for population susceptibility to infection.

6. *Dengue cycles*: implement a decision rule governed by the second- and third-order Markov transition probabilities, reflecting the transition between consecutive sequences of epidemic and non-epidemic states

We applied our approach to 20 cities in Brazil spanning large geographic and population ranges (Fig. 3.1, Table B.1). We used as input a historical time series spanning 17 years and consisting of information on dengue case reports (number, annual) and 2 weather variables: 2-meter air temperature (Kelvin, daily) and precipitation (kg/m^2, hourly). We describe data sources, acquisition, and processing in the Supporting Information. After an initial training period of 7 years, we generated 10 years of out-of-sample epidemic predictions for each of the independent models using a one-year expanding training window (Step 2). We used the first 4 years of out-of-sample predictions to inform ensemble model selection (Step 4), and produced ensemble-based predictions for the remaining 6 years (Step 5).

3.4.1 SIGNAL PREPROCESSING

Using a daily time series of weather data to forecast dengue fever epidemic status requires identifying the most predictive period(s) of the calendar year during

which weather information contains a strong signal for subsequent dengue fever outbreaks. In order to construct a single framework that can automatically identify important weather signals in multiple different locations with vastly different ecosystems and weather patterns, we allow the data to inform the choice of time intervals. Our algorithm achieves this by scanning over multiple, partially-overlapping time intervals across the calendar year, and building hundreds of models on these different intervals in order to select those with the strongest signals.

Each time interval is defined by a start date, *t0*, between early June and late September, and a period length, *p*, of between 10 and 95 days. The combination of each (*t0, p*) produces multiple, partially-overlapping intervals spanning the last 7 months of the calendar year.

Borrowing from spectral analysis and wavelet decomposition, we use a windowing-inspired approach to better capture signals within the time intervals. Windowing is typically used to improve signal clarity, and here we apply a rectangular "range" as described in[30] to incorporate information in the days both within and around each time interval. We define a rectangle of 5 x 6, indicating that, for every defined (*t0, p*) time interval, the algorithm collects information from 5 consecutive start dates, *t0, t0+1, … , t0+4*, spanning 6 consecutive period lengths, *p, p+1, … , p+5*. Each time interval and weather variable, then, is summarized by 30 data points, each capturing slightly different temporal slices from the time series. This process effectively adds a bit of redundant information to the model building process - to which our learning algorithm, the support vector machine, is in general

robust - in order to pick up signals in the data that may not be captured by applying an arbitrary "start" and "end" cutoff to the data.

3.4.2 TIME SERIES FEATURE EXTRACTION

Time series data must be transformed into appropriate inputs in order to be used in supervised learning models. This process, called time series feature extraction, involves computing summary features of the time series, which can range from simple means to complex wavelet transforms. To test the feasibility of our approach using only simple summary features, we extracted the following features within each (*t0, p*) time interval based on the findings of[30]: 1) the arithmetic mean of daily temperature, and 2) mean precipitation frequency, with frequency defined as the time interval (in days) between peaks (local maxima) of daily precipitation.

3.4.3 INDEPENDENT MODEL TRAINING AND PREDICTION

The goal of our independent model building step is to identify dynamically, through the continually-updating performance of a collection of models, the periods of the year that are most predictive of annual dengue outbreaks, in order to exploit a small number of them to generate forecasts.

To forecast outbreak years, we trained a collection of support vector machine (SVM) classifiers on an initial 7 year training period, and produced annual forecasts incorporating the most recently available weather information using a dynamic, one-year expanding training window. A unique SVM was trained for each of the (*t0, p*) time intervals, resulting in a total of 432 independent models trained

per year. Each model generated out-of-sample predictions for the remaining 10 years of data. Predictions were made by classifying the 30 out-of-sample data points corresponding to the weather information preceding the target year, and taking a majority vote. In order to handle highly nonlinear relationships between weather variables, both radial basis function (RBF) and sigmoid kernels were used and evaluated for performance, and show results for the best respective kernel in each city. We tuned model parameters (gamma, soft margin cost function, and coefficient) using 10-fold cross-validation.

Support vector machines, a supervised learning method for classification, were used because of their flexibility in the face of complex, nonlinear decision boundaries and their robustness to overfitting and outliers. The property that underpins these advantages is known as the "large-margin classifier." SVMs are also known for their good performance in high-dimensional feature space, which is advantageous for the scale-up of the model to include dozens more predictors.

3.4.4 MODEL SELECTION

From the resulting collection of 432 models, the best-performing models (n=11) were selected each year based on a) historical out-of-sample prediction accuracy (% outbreak forecasts correct) and b) out-of-sample prediction accuracy of neighboring models (representing similar time intervals). These models thus represent strongly predictive periods of the year preceding outbreaks, and the algorithm rewards the high performance of similar temporal windows over the high performance of a time window whose neighbors exhibit poor prediction tendencies.

Because the model building process is dynamic, resulting in a new collection of models each year with continually-updating performance measures, the selection of the 11 models changes from year to year.

In order to get a sense of the out-of-sample performance of the 432 models, we allowed all models to generate 4 years of out-of-sample predictions before the top 11 models were selected based on this prediction accuracy. As a result, the ensemble approach, which exploited the predictions of the top 11 models, was used for the final 6 years of out-of-sample predictions.

3.4.5 Ensemble prediction

Ensemble learning helps improve machine learning algorithms by combining the results of multiple trained predictors in order to generate a single, robust prediction. In our approach, we combine the results from the strongest-performing models, which represent the most highly predictive time periods preceding dengue outbreaks. While there are an abundance of ensembling methods in machine learning, we use a simple majority vote of the 11 models to decide a single forecast. These single forecasts were produced for the last 6 years of the 17-year dataset, representing the culmination of a prediction process that involves: 7-year initial training period, 4-year out-of-sample model calibration period, and 6-year out-of-sample ensemble prediction period. Across 20 Brazilian municipalities, this scheme produced 120 municipality-years of out-of-sample ensemble predictions.

3.4.6 Dengue cycles

Our weather-based ensemble approach remains ignorant to the relationship between weather patterns and dengue outbreaks, instead allowing the data to drive model selection and predictions. However, endemic transmission of dengue fever is typically distinguished by periodic outbreak cycles of around 3-4 years. These outbreak cycles are thought to occur as a result of 1) an exhaustion of susceptibles after an outbreak, and 2) and short-term cross-immunity to other circulating DENV serotypes after infection[20]. Both factors result in a depletion of the population vulnerable to infection, and act as barriers to subsequent outbreaks. Independent of climate variability over the years, we expect some preservation of these cycles.

Consequently, we implemented a "decision rule" in the model based on the observed transitions between epidemic- and non-epidemic years across 51 Brazilian municipalities meeting endemic inclusion criteria (Supplemental Information). Across these municipalities, we computed the mean second- and third-order Markov transition probabilities, representing the probability of transition from one outbreak state (epidemic/non-epidemic) to the opposite outbreak state (non-epidemic/epidemic) after 2 and 3 consecutive years, respectively. Thus, we obtained the transition probabilities governing the following 3- and 4-year cycles: 001, 110, 0001, and 1110 (0= non-epidemic year, 1= epidemic year). Transition probabilities were computed based only on the first 11 years of data; that is, the years preceding the 6 out-of-sample ensemble predictions.

Our decision rule acts as a surrogate "expert opinion," overturning the ensemble prediction if the probability of a specific transition exceeded the percent of model votes (out of 11 votes). For example, if the ensemble predicts an epidemic

year to succeed 2 epidemic years with 7 votes, the corresponding "strength" of that vote is 63% (7/11), which is weaker than the corresponding observed second-order transition probability for a non-epidemic year to follow 2 epidemic years (0.71). In this case, the model vote would be overridden to predict a non-epidemic year instead of an epidemic year.

We compared the performance of predictions based solely on weather patterns to those which incorporate additional empirical data from outbreak cycles.

3.5 REFERENCES

1. Ford TE, Colwell RR, Rose JB, Morse SS, Rogers DJ, Yates TL. Using Satellite Images of Environmental Changes to Predict Infectious Disease Outbreaks. *Emerg Infect Dis.* 2009;15: 1341–1346.

2. Sewe MO, Tozan Y, Ahlm C, Rocklöv J. Using remote sensing environmental data to forecast malaria incidence at a rural district hospital in Western Kenya. *Sci Rep.* 2017;7: 2589.

3. McGough SF, Brownstein JS, Hawkins JB, Santillana M. Forecasting Zika Incidence in the 2016 Latin America Outbreak Combining Traditional Disease Surveillance with Search, Social Media, and News Report Data. *PLoS Negl Trop Dis*. 2017;11: e0005295.

4. Yang S, Santillana M, Kou SC. Accurate estimation of influenza epidemics using Google search data via ARGO. *Proc Natl Acad Sci U S A*. 2015;112: 14473–14478.

5. Bengtsson L, Gaudart J, Lu X, Moore S, Wetter E, Sallah K, et al. Using mobile phone data to predict the spatial spread of cholera. *Sci Rep.* 2015;5: 8923.

6. Kramer AM, Pulliam JT, Alexander LW, Park AW, Rohani P, Drake JM. Spatial spread of the West Africa Ebola epidemic. *R Soc Open Sci.* 2016;3: 160294.

7. Zhu Z, Chan JF-W, Tee K-M, Choi GK-Y, Lau SK-P, Woo PC-Y, et al. Comparative genomic analysis of pre-epidemic and epidemic Zika virus strains for virological factors potentially associated with the rapidly expanding epidemic. *Emerg Microbes Infect.* 2016;5: e22.

8. Dudas G, Carvalho LM, Bedford T, Tatem AJ, Baele G, Faria NR, et al. Virus genomes reveal factors that spread and sustained the Ebola epidemic. *Nature.* 2017;544: 309–315.

9. Reich NG, Brooks LC, Fox SJ, Kandula S, McGowan CJ, Moore E, et al. A collaborative multiyear, multimodel assessment of seasonal influenza forecasting in the United States. *Proc Natl Acad Sci U S A.* 2019;116: 3146–3154.

10. Buczak AL, Baugher B, Moniz LJ, Bagley T, Babin SM, Guven E. Ensemble method for dengue prediction. *PLoS One.* 2018;13: e0189988.

11. Bhatt S, Gething PW, Brady OJ, Messina JP, Farlow AW, Moyes CL, et al. The global distribution and burden of dengue. *Nature.* 2013;496: 504–507.

12. Stanaway JD, Shepard DS, Undurraga EA, Halasa YA, Coffeng LE, Brady OJ, et al. The global burden of dengue: an analysis from the Global Burden of Disease Study 2013. *Lancet Infect Dis*. 2016;16: 712–723.

13. Morin CW, Comrie AC, Ernst K. Climate and dengue transmission: evidence and implications. *Environ Health Perspect*. 2013;121: 1264–1272.

14. Tjaden NB, Thomas SM, Fischer D, Beierkuhnlein C. Extrinsic Incubation Period of Dengue: Knowledge, Backlog, and Applications of Temperature Dependence. *PLoS Negl Trop Dis*. 2013;7: e2207.

15. Rohani A, Wong YC, Zamre I, Lee HL, Zurainee MN. The effect of extrinsic incubation temperature on development of dengue serotype 2 and 4 viruses in Aedes aegypti (L.). *Southeast Asian J Trop Med Public Health.* 2009;40: 942–950.

16. Liu Z, Zhang Z, Lai Z, Zhou T, Jia Z, Gu J, et al. Temperature Increase Enhances Aedes albopictus Competence to Transmit Dengue Virus. Front Microbiol. *Frontiers*; 2017;8. doi:10.3389/fmicb.2017.02337

17. Byttebier B, De Majo MS, Fischer S. Hatching Response of Aedes aegypti (Diptera: Culicidae) Eggs at Low Temperatures: Effects of Hatching Media and Storage Conditions. *J Med Entomol.* Oxford University Press; 2014;51: 97–103.

18. Barry W. Alto DB. Temperature and Dengue Virus Infection in Mosquitoes: Independent Effects on the Immature and Adult Stages. *Am J Trop Med Hyg.* The American Society of Tropical Medicine and Hygiene; 2013;88: 497.

19. Scott TW, Morrison AC, Lorenz LH, Clark GG, Strickman D, Kittayapong P, et al. Longitudinal studies of Aedes aegypti (Diptera: Culicidae) in Thailand and Puerto Rico: population dynamics. *J Med Entomol.* 2000;37: 77–88.

20. Adams B, Holmes EC, Zhang C, Mammen MP Jr, Nimmannitya S, Kalayanarooj S, et al. Cross-protective immunity can account for the alternating epidemic pattern of dengue virus serotypes circulating in Bangkok. *Proc Natl Acad Sci U S A*. 2006;103: 14234–14239.

21. Mbogo CM, Mwangangi JM, Nzovu J, Gu W, Yan G, Gunter JT, et al. Spatial and temporal heterogeneity of Anopheles mosquitoes and Plasmodium falciparum transmission along the Kenyan coast. *Am J Trop Med Hyg.* 2003;68: 734–742.

22. Acevedo MA, Prosper O, Lopiano K, Ruktanonchai N, Caughlin TT, Martcheva M, et al. Spatial heterogeneity, host movement and mosquito-borne disease transmission. *PLoS One*. 2015;10: e0127552.

23. Torres-Sorando L, Rodríguez DJ. Models of spatio-temporal dynamics in malaria. *Ecol Modell*. 1997;104: 231–240.

24. Teurlai M, Menkès CE, Cavarero V, Degallier N, Descloux E, Grangeon J-P, et al. Socio-economic and Climate Factors Associated with Dengue Fever Spatial Heterogeneity: A Worked Example in New Caledonia. *PLoS Negl Trop Dis*. 2015;9: e0004211.

25. Descloux E, Mangeas M, Menkes CE, Lengaigne M, Leroy A, Tehei T, et al. Climate-based models for understanding and forecasting dengue epidemics. *PLoS Negl Trop Dis*. 2012;6: e1470.

26. Guo P, Liu T, Zhang Q, Wang L, Xiao J, Zhang Q, et al. Developing a dengue forecast model using machine learning: A case study in China. *PLoS Negl Trop Dis.* Public Library of Science; 2017;11: e0005973.

27. Chuang T-W, Chaves LF, Chen P-J. Effects of local and regional climatic fluctuations on dengue outbreaks in southern Taiwan. *PLoS One.* 2017;12: e0178698.

28. Johansson MA, Reich NG, Hota A, Brownstein JS, Santillana M. Evaluating the performance of infectious disease forecasts: A comparison of climate-driven and seasonal dengue forecasts for Mexico. *Sci Rep.* 2016;6: 33707.

29. Lauer SA, Sakrejda K, Ray EL, Keegan LT, Bi Q, Suangtho P, et al. Prospective forecasts of annual dengue hemorrhagic fever incidence in Thailand, 2010–2014. *Proceedings of the National Academy of Sciences.* 2018;115: E2175–E2182.

30. Stolerman L, Maia P, Kutz JN. Data-Driven Forecast of Dengue Outbreaks in Brazil: A Critical Assessment of Climate Conditions for Different Capitals [Internet]. 2016. Available: http://arxiv.org/abs/1701.00166

31. Platt J. Probabilistic Outputs for Support Vector Machines and Comparisons to Regularized Likelihood Methods. Advances in large margin classifiers. 1999;10: 61–74.

32. Hendrycks D, Gimpel K. A Baseline for Detecting Misclassified and Out-of-Distribution Examples in Neural Networks.

33. Lopes TRR, Silva CS, Pastor AF, Silva JVJ, Júnior. Dengue in Brazil in 2017: what happened? *Rev Inst Med Trop Sao Paulo.* Instituto De Medicina Tropical De Sao Paulo; 2018;60. doi:10.1590/S1678-9946201860043

34. van Panhuis WG, Hyun S, Blaney K, Marques ETA Jr, Coelho GE, Siqueira JB Jr, et al. Risk of Dengue for Tourists and Teams during the World Cup 2014 in Brazil. *PLoS Negl Trop Dis.* Public Library of Science; 2014;8: e3063.

35. Massad E, Wilder-Smith A, Ximenes R, Amaku M, Lopez LF, Coutinho FAB, et al. Risk of symptomatic dengue for foreign visitors to the 2014 FIFA World Cup in Brazil. *Mem Inst Oswaldo Cruz*. 2014;109: 394–397.

36. Shepard DS, Undurraga EA, Halasa YA, Stanaway JD. The global economic burden of dengue: a systematic analysis. *Lancet Infect Dis.* 2016;16: 935–941.

37. Gelaro R, McCarty W, Suárez MJ, Todling R, Molod A, Takacs L, et al. The Modern-Era Retrospective Analysis for Research and Applications, Version 2 (MERRA-2). *J Clim.* 2017;30: 5419–5454.

4

Nowcasting by Bayesian Smoothing: A flexible, generalizable model for real-time epidemic tracking

4.0 ABSTRACT

Delays in case reporting are common to disease surveillance systems. One major consequence is that disease activity is not fully known until days or weeks later, giving surveillance bodies an incomplete picture of current disease activity at a given moment in time. Nowcasting, or "predicting the present," offers a solution to the issue of reporting delays. Here, we introduce Nowcasting by Bayesian Smoothing (NobBS), a simple and flexible Bayesian model for nowcasting infectious diseases in different settings. Specifically, we show the performance of this approach in weekly nowcasts of dengue fever cases in Puerto Rico and influenza-like illness (ILI) cases in the United States over multiple years, requiring no disease-specific parameterization despite being very different diseases (directly transmitted vs. vector-borne) and exhibiting substantially different reporting delays. This method allows for both uncertainty in the delay distribution and the time evolution of the epidemic curve, producing smooth, time-correlated estimates of cases. We test NobBS against an established Bayesian nowcast method(9) and find that NobBS outperforms this benchmark for both diseases and over multiple time periods. In particular, we show that while point estimates of the models are similar when time-to-report distributions are relatively fixed over time, NobBS

improves the estimation of uncertainty and accommodates temporal variation in delay probabilities.

4.1 INTRODUCTION

Effective public health action relies on disease surveillance that is timely and accurate, especially in disease outbreaks(1, 2). Specifically, surveillance provides the information required to assess risks, prioritize and allocate resources to public health threats, deploy and discontinue interventions to interrupt disease transmission, and monitor the impact of those interventions. Ideally, disease surveillance systems should closely track the often fast-changing circumstances of outbreaks, distinguishing true changes in the dynamics from artifacts of reporting.

Despite the importance of timely surveillance data, substantial challenges exist to collect and report case information in real time. Multiple features of the disease and surveillance system contribute to reporting delays, including: delays in symptoms onset after infection; delays in medical care-seeking after onset; delays in providers obtaining and reporting diagnostic information; level of awareness of disease activity influencing care-seeking and reporting; and system-level processing delays, a result of complex and multi-tiered disease reporting and communication systems interacting at multiple administrative levels(3). Reporting delays can be further exacerbated in resource-constrained settings. As a consequence, surveillance data are typically not complete until weeks or months after infections have actually occurred, providing an incomplete picture of current disease activity.

Nowcasting, or "predicting the present," is an approach to mitigate the impact of reporting delays. With origins in the insurance claims and actuarial literature(4, 5), nowcast models aim to estimate the number of occurred-but-not-yet-reported events (e.g. insurance claims, disease cases) at any given time based on an incomplete set of reports. In public health settings, nowcasting approaches have been explored for AIDS in the 1980s and 1990s(6–8) as a consequence of the long incubation period from HIV infection until development of AIDS. More recently, nowcasting has been applied to infectious disease outbreaks such as foodborne illness outbreaks(9, 10). These studies draw principally on survival analysis and actuarial techniques to model the reporting delay and draw inferences based on historical patterns. Infectious disease nowcast models have largely focused on specific applications, not the common challenges that exist across many different diseases. These studies have strictly focused on modeling the reporting delay distribution—a legacy of the actuarial techniques giving rise to many of these approaches—and generally neglect a key feature of outbreaks: that future cases are intrinsically linked to past reported cases, a fact that creates potentially strong autocorrelation in the true number of cases over short time intervals. In other words, the infectious disease transmission process provides an additional signal of the number of cases to be expected in the near future. Lastly, previous models have largely focused on providing point estimates of the number of cases. Point estimates are useful, but quantifying the uncertainty in those estimates may provide critical context for users of surveillance data.

Here, we introduce Nowcasting by Bayesian Smoothing (NobBS), a simple and flexible generalized Bayesian model for nowcasting infectious diseases in different settings. Specifically, NobBS allows for both uncertainty in the delay distribution and the time evolution of the epidemic curve, producing smooth, time-correlated estimates of cases. We show the performance of NobBS in weekly nowcasts of dengue fever cases in Puerto Rico and influenza-like illness (ILI) cases in the United States over multiple years, requiring no disease-specific parameterization despite being very different diseases (directly transmitted vs. vector-borne) and exhibiting substantially different reporting delays. We test NobBS against an established Bayesian nowcast method(9) and find that NobBS outperforms this benchmark for both diseases and over multiple time periods. In particular, we show that while point estimates of the models are similar when time-to-report distributions are relatively fixed over time, NobBS improves the estimation of uncertainty and accommodates temporal variation in delay probabilities.

4.2 RESULTS

We developed a Bayesian approach to nowcast total case numbers using incomplete, time-stamped reported case data based on an estimated delay distribution, intrinsic autocorrelation from the transmission process, and historical case data. Generally, the approach learns from historical information on cases reported at multiple delays (e.g. no delay, 1-week delay, 2-week delay, etc.) from the date of case onset to estimate the reporting delay probability at each delay and

the relationship between case counts from week-to-week, and uses this to predict the number of not-yet-reported cases in the present. We tested this approach, NobBS, using two different infectious disease surveillance data sources: dengue fever surveillance in Puerto Rico, and national notifications of influenza-like illness (ILI) in the United States. Using all of the available data on case reporting delays up to the point of prediction, weekly dengue nowcasts were estimated for the time period December 23, 1991 through November 29, 2010 (989 weeks), and weekly ILI nowcasts were produced over the period June 30, 2014 through September 25, 2017 (170 weeks). For comparison, we generated weekly nowcasts over the same periods using an existing Bayesian approach (9). To access a large amount of historical data relative to the length of each time series, dengue fever models used a 104-week (approximately 2-y) moving window while the ILI models used a 27-week (approximately 6-mo) moving window. Our primary outcome metric to assess nowcast performance was the logarithmic score, a proper score that evaluates the probability assigned to the observed outcome rather than error associated with a point prediction. For purposes of discussion, we reported the exponentiated form of the mean logarithmic score (the geometric mean of the assigned probabilities) to provide a metric on the scale of 0 (no certainty of the outcome) to 1 (complete certainty of the outcome). In addition, we estimated other metrics describing the performance of point estimates (mean absolute error (MAE), root mean square error (RMSE), and relative root mean square error (rRMSE)) and the prediction interval (95% prediction interval (PI) coverage), and of these, focus on comparing the rRMSE and 95% PI coverage across approaches.

4.2.1 Performance in forecasting weekly dengue and influenza incidence

Figs. 4.1-4.2 show weekly dengue fever and ILI nowcasts for NobBS and the benchmark approach over multiple seasons for both diseases. Table 4.1 summarizes the point and probability-based accuracy metrics for each, where higher accuracy is indicated by lower MAE, RMSE, and rRMSE, higher average scores, and lower distance from 0.95 for the 95% prediction interval (PI) coverage.

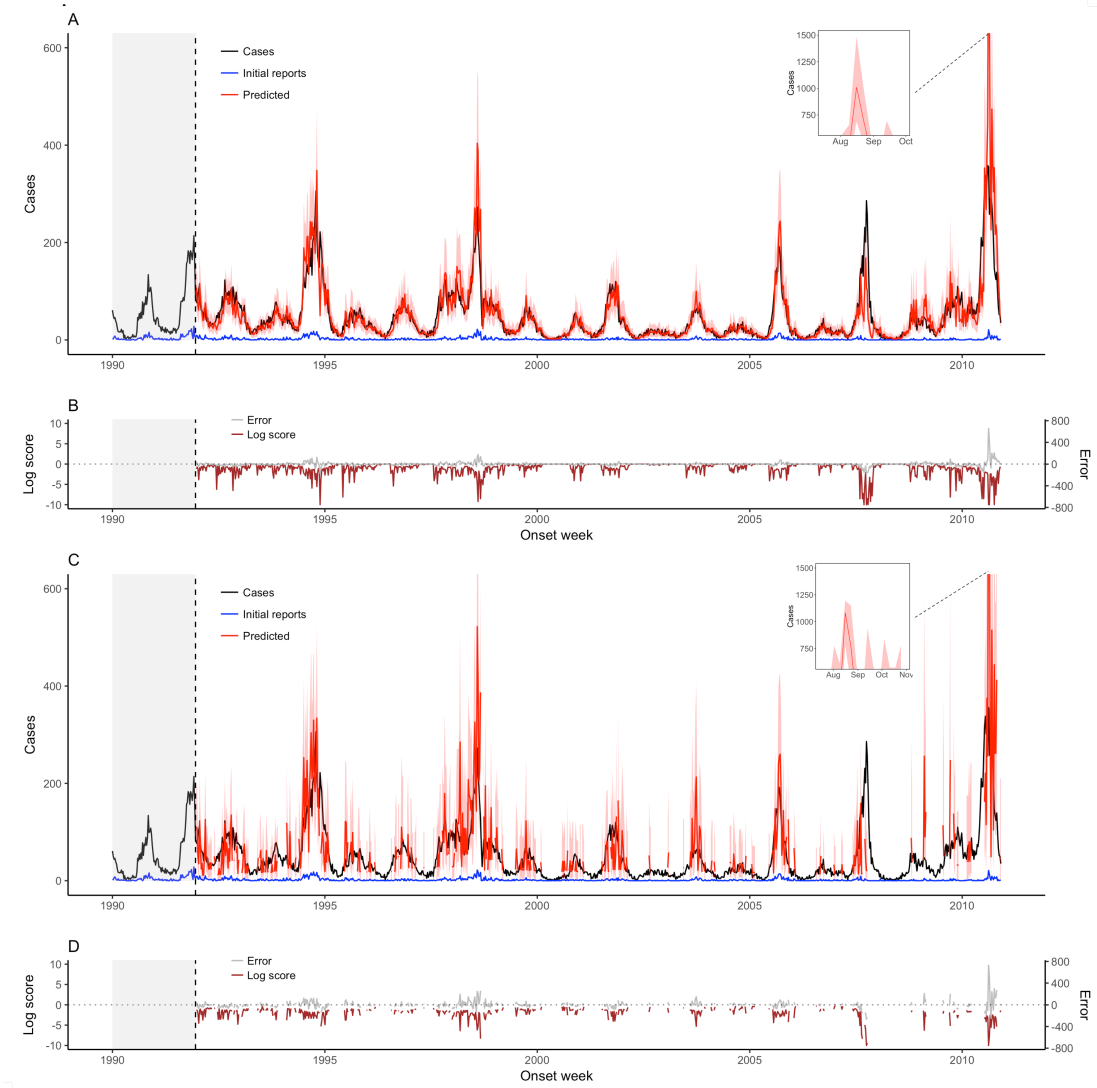

Figure 4.1. Weekly dengue fever nowcasts for December 23, 1991 through December 25, 2000 using a 2-year moving window. (A) NobBS nowcasts along with (B) point estimate and uncertainty accuracy, as measured by the log score and the prediction error, are compared to (C) nowcasts by the benchmark approach with (D) corresponding log scores and prediction errors. For nowcasting, the number of newly-reported cases each week (blue line) are the only data available in real-time for that week, and help inform the estimate of the total number of cases that will be eventually reported (red line), shown with 95% prediction intervals (pink bands). The true number of cases eventually reported (black line) is known only in hindsight and is the nowcast target. Historical information on reporting is available within a 104-week moving window (grey shade) and used to make nowcasts. The log score (brown line) and the difference between the true and mean estimated number of cases (grey line) are shown as a function of time.

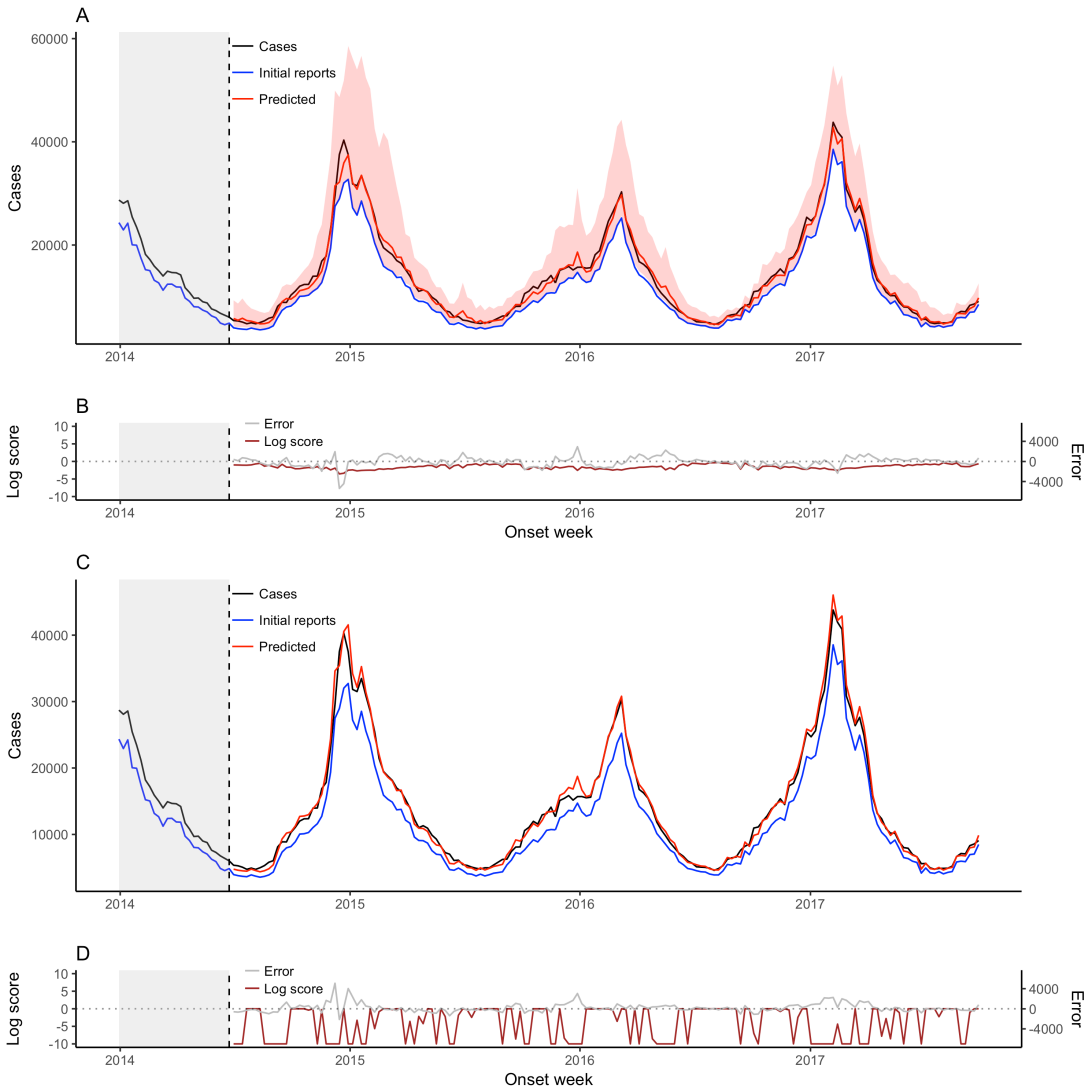

Figure 4.2. Weekly ILI nowcasts for June 30, 2014 through September 25, 2017 using a 6-month moving window. (A) NobBS nowcasts along with (B) point estimate and uncertainty accuracy, as measured by the log score and the prediction error, are compared to (C) nowcasts by the benchmark approach with (D) corresponding log scores and prediction errors. For nowcasting, the number of newly-reported cases each week (blue line) are the only data available in real-time for that week, and help inform the estimate of the total number of cases that will be eventually reported (red line), shown with 95% prediction intervals (pink bands). For the benchmark approach, the 95% prediction intervals are very narrow and are thus difficult to see. The true number of cases eventually reported (black line) is known only in hindsight and is the nowcast target. Historical information on reporting is available within a 27-week moving window (grey shade) and used to make nowcasts. The log score (brown line) and the difference between the true and mean estimated number of cases (grey line) are shown as a function of time.

Table 4.1. Performance measures for each nowcast approach by disease (mean % reported with no delay).

Disease	Model	Period	% of weeks predicted	Average Score	MAE	RMSE	rRMSE	95% PI coverage
Dengue	NobBS	Full time period*	100%	0.349	16	37.6	0.600	0.87
(4%)		Weeks in which at least 1 case was reported in the first week	--	0.274	21	46.6	0.464	0.85
	Benchmark (ref. 9)	Full time period*	55%	--	32	57.4	1.14	--
		Weeks in which at least 1 case was reported in the first week	--	0.161	37	68.1	1.24	0.91
Influenza (82%)	NobBS	Full time period*	100%	0.218	693	987.8	0.074	1.00
	Benchmark (ref. 9)	Full time period*	100%	0.017	609	916.2	0.062	0.00

*Full time period for: dengue fever (12/23/1991-11/29/2010) and ILI (6/30/2014-9/25/2017)

Because the NobBS model accounts for both under-reporting and the autocorrelated progression of transmission across successive weeks, it makes predictions even in weeks when there are no cases reported for the week. Conversely, the benchmark model does not make nowcasts for weeks in which there are no initial case reports (common in the dengue Puerto Rico data), hence the nowcasts in Figs. 4.1C and 4.2C appear as discontinuous lines. To account for these differences, we report accuracy metrics between NobBS and the benchmark approach for both (1) the full time series of the data and (2) weeks when at least one case was reported in the first week, i.e. the subset of weeks for which both models could make predictions (Table 4.1). To compare the full time series of nowcasts across approaches, even in the absence of predictions by the benchmark model, we assigned missing estimates a point prediction of 0 but did not calculate a penalized log score for those weeks.

The benchmark approach made predictions in only 55% of weeks in the dengue fever time series (Table 4.1). In this subset of weeks, the NobBS approach achieved relatively smooth and accurate tracking of the dengue fever time series (rRMSE = 0.464, average score = 0.274) despite low proportions of cases reported on the week of onset (Fig. 4.1A-B). The 95% prediction interval (PI) coverage, defined as the proportion of times the 95% PI included the true number of cases, was 0.85. In comparison, the benchmark approach produced less accurate point estimates and slightly broader uncertainty intervals (rRMSE = 1.24, average score = 0.161, 95% PI coverage = 0.91) with greater fluctuation in nowcasts from week-to-week (Fig. 4.1C-D). Because many weeks in the dengue data were low

incidence, assigning a prediction of 0 to the benchmark approach's missing nowcasts improved its rRMSE to 1.14 in the full time series, though NobBS still far exceeded the accuracy of this and all other metrics (Table 4.1).

Nowcast point estimates tracked the ILI time series well for both approaches, though with greater error in point estimates by all measures for the NobBS approach (NobBS rRMSE = 0.074 vs. benchmark rRMSE = 0.062; see Table 1 for other metrics). However, the NobBS approach produced considerably wider prediction intervals (Figs. 4.1C, 4.2C) resulting in both higher log scores (NobBS average score = 0.218 vs. benchmark average score = 0.017) and 100% coverage by the 95% prediction intervals compared to 0% coverage for the benchmark (Table 4.1).

To quantify the smoothness of the predictions of NobBS, particularly in the dengue time series, we calculated the 1-week lagged autocorrelation of predictions (ρ_a) and compared this to the 1-week lagged autocorrelation of cases (ρ_c). In addition, we computed metrics reflecting the accuracy of the approaches in capturing the change in cases from week-to-week: the mean absolute error of the change (MAEΔ) and the RMSE of the change (RMSEΔ) (Table 4.2). The formulae for these additional metrics are provided in Materials and Methods. Because some weeks experienced no change in case numbers from the previous week, we did not calculate the rRMSE. Comparing the full time series, the nowcasts produced by NobBS exhibited high autocorrelation for both diseases (ρ_a = 0.876 for dengue, 0.973 for ILI) while the benchmark approach yielded lower autocorrelation for dengue fever nowcasts, comparatively (ρ_a = 0.631 for dengue, 0.970 for ILI).

Further, the autocorrelation of NobBS nowcasts was closer to that of the true cases for both diseases ($\rho_c = 0.958$ for dengue fever and $\rho_c = 0.972$ for ILI). For dengue, over the weeks in which at least 1 case was initially reported, the NobBS approach achieved both lower mean absolute difference between predicted and observed changes in cases (NobBS MAEΔ = 23 vs. benchmark MAEΔ = 50) and lower RMSE of the change (NobBS RMSEΔ = 35.8 vs. benchmark RMSEΔ = 64.6). In addition, NobBS outperformed the benchmark approach over the full time series of dengue cases (Table 4.2). For ILI, however, the metrics for the weekly change were similar for the two approaches (Table 4.2). For reference, dengue cases changed in absolute value by on average 9.79 cases/week and ILI by 1,312.6 cases/week.

Table 4.2. Performance measures for estimates of the change in disease incidence from the previous week.

Model	Period	Dengue (mean cases/week=48)				Influenza (mean cases/week=14,000)			
		MAEΔ	RMSEΔ	ρ_a	ρ_c	MAEΔ	RMSEΔ	ρ_a	ρ_c
NobBS	Full time period*	17	35.8	0.876	0.958	669	1027.1	0.973	0.972
	Weeks in which at least 1 case was reported in the first week	23	45.2	--	--	--	--	--	
Benchmark (ref. 9)	Full time period*	34	64.6	0.631	0.958	612	1004.2	0.970	0.972
	Weeks in which at least 1 case was reported in the first week	50	88.2	--	--	--	--	--	

*Full time period for: dengue fever (12/23/1991-11/29/2010) and ILI (6/30/2014-9/25/2017)

4.2.2 Reporting delays impact nowcast performance

The delay distributions between the reporting systems are strikingly different (Figs. 4.1, 4.2, C.1). In the case of the dengue fever surveillance system, which includes specimen collection and laboratory testing, only approximately 4% of cases were processed during the week of onset, on average. In contrast, the U.S. Outpatient Influenza-like Illness (ILI) Surveillance Network (ILINet) captures only syndromic data reported electronically, with over 80% of ILI cases reported, on average, the same week they present (i.e. with no delay). Overall, we observed that the accuracy of nowcast point estimates (rRMSE) was higher for the ILI data compared to dengue, which may be related to the high proportion of cases reported with 0-weeks delay in these data. In addition, in several weeks of the time series we observed that the error of model predictions was larger when there were larger absolute changes in the number of cases initial case reports – a finding that is especially true for dengue fever, which experienced high fluctuations in the number of initial reports over time (Table C.1, Fig. C.2). Note that because of the difference in predictive distribution bin widths based on the number of cases that accrue for influenza vs. dengue fever (*Materials & Methods*), average scores are not comparable across diseases.

4.2.3 NobBS improves nowcasting with varying reporting delays

Dengue fever and ILI also exhibit differences in the *trends* of reporting delay probabilities *over time.* For dengue fever, we observe a noisier, more time-varying probability of reporting for cases, with more extreme fluctuations in the

proportion of initial reports compared to ILI cases, which show more constant (tighter ranges of) reporting probabilities from week-to-week (Fig. C.3). Independent of the initial proportion of cases reported (high vs. low), we hypothesized that these trends (relatively constant vs. time-varying) are particularly impactful on the performance of the nowcast, and that relatively constant reporting probabilities, as seen in the ILI data, may be linked to the higher accuracy of these predictions.

To test the robustness of the model, we simulated ILI data using the final counts from the true dataset, but imposing a time-varying delay distribution; specifically, with faster initial reporting during weeks of high incidence (described in *Materials and Methods*). Using these simulated data, we found that NobBS was relatively robust to changes in reporting delays (Fig. C.4, Table C.2). In the context of stable reporting delays (original ILI data), NobBS performed comparably to the benchmark model (Fig. 4.2, Table 4.1). However, NobBS outperformed the benchmark in terms of point estimates (NobBS rRMSE = 0.302 vs. benchmark rRMSE = 0.621), uncertainty estimates (NobBS average score = 0.06 vs. benchmark average score ≈ 0), and accuracy of the predicted change (Table C.3) in the presence of more time-varying reporting delays (simulated ILI data), a reality in many epidemics(11).

4.2.4 PERFORMANCE BY YEAR

The performance of ILI nowcasts across accuracy measures was relatively consistent by year, but there were fluctuations in the year-to-year performance of

both approaches applied to dengue data (Table 4.3). In general, nowcast point estimates for dengue were more accurate during the first half of the series (1992-2000) compared to the second half (2001-2010), with rRMSE's nearly doubling after 2000. In addition, average scores tended to be high in years that experienced a very low number of dengue cases (e.g. 2000, 2002, 2004, 2006). The model was particularly effective at identifying periods of low incidence, with high probabilities assigned to the correct outcome bin (width = 25 cases, details in *Materials and Methods*) when the number of cases eventually reported was low (Fig. C.5). On the other hand, during periods of high dengue activity, lower probabilities were assigned to the correct bin, a feature of the bin size (fixed at width = 25 cases) containing a smaller fraction of the predictive distribution. Overall, NobBS outperformed the benchmark approach on all performance measures across individual years (Table 4.3).

Both approaches had their lowest accuracy on three high incidence dengue seasons: 1994, 2007, and 2010 (Table 4.3; Fig. 4.1). The average scores for these years range between 0.041 and 0.17 across the NobBS and benchmark approaches, falling clearly below the rest of the years in performance. These scores not only reflect unusually poor point estimate predictions as judged by rRMSE, but also the finding that the predictive distribution for weeks in these years for both approaches rarely included the true value of interest (a consequence of dramatic over- or underestimates), resulting in many estimates being assigned log scores of -10.

Table 4.3. Annual performance measures for each nowcast model, by disease. All predicted weeks for each model are compared.

Disease	Year	Cases	NobBS				Benchmark (ref. 9)			
			MAE	rRMSE	RMSE	Average Score	MAE	rRMSE	RMSE	Average Score
Dengue	1992	3,570	15	0.271	19.7	0.262	27	0.473	33.2	0.154
	1993	2,044	10	0.325	13.0	0.436	20	0.559	23.5	0.237
	1994	5,455	29	0.356	45.9	0.171	50	0.690	63.3	0.108
	1995	2,075	13	0.450	16.2	0.330	28	1.035	38.4	0.178
	1996	1,856	8	0.520	11.0	0.472	17	0.617	21.9	0.270
	1997	2,413	12	0.375	16.2	0.402	20	0.625	26.7	0.228
	1998	5,334	33	0.448	47.8	0.129	65	0.801	89.9	0.072
	1999	1,823	9	0.389	11.9	0.493	18	0.897	23.5	0.250
	2000	766	4	0.359	6.1	0.720	17	2.225	20.2	0.304
	2001	2,274	11	0.487	16.6	0.437	26	0.492	37.7	0.189
	2002	821	5	0.522	5.7	0.834	16	1.101	23.0	0.352
	2003	1,422	6	0.471	9.5	0.590	32	1.412	47.5	0.193
	2004	911	6	0.599	7.2	0.610	13	2.088	17.0	0.368
	2005	2,543	14	0.998	21.4	0.407	32	1.150	42.0	0.178
	2006	734	4	0.891	6.3	0.770	13	1.211	15.8	0.395
	2007	3,290	30	0.675	55.4	0.102	55	0.632	93.6	0.066
	2008	843	8	1.032	12.8	0.629	38	4.145	50.7	0.191
	2009	2,448	19	0.667	26.7	0.225	57	2.405	81.9	0.092
	2010	6,820	71	0.583	132.4	0.055	121	0.854	198.7	0.041
Influenza	2014	726,312	1052	0.085	1565.9	0.188	958	0.091	1482.0	0.004
	2015	679,850	685	0.086	890.3	0.203	624	0.069	848.0	0.019
	2016	704,020	696	0.072	861.8	0.224	376	0.043	480.1	0.063
	2017	632,353	551	0.046	712.9	0.258	659	0.047	934.2	0.008

4.2.5 MOVING WINDOW SIZES

To leverage a large number of historical training weeks while also considering the number of available weeks in the time series, we chose moving windows of 104 weeks (approx. 2-y) for dengue fever (a longer time series) and 27 weeks (approx. 6-mo) for ILI (a shorter time series). On the one hand, moving windows allow for a more stable estimation of the recent delay distribution, as information from very old and potentially less relevant weeks are forgotten. On the other hand, the size of the moving window reflects judgment on how quickly and smoothly changes in the data should be realized by the model: longer moving windows tend to produce smoother estimates, but the model may be less sensitive to abrupt changes in the data (e.g. changes in how quickly cases are reported during an outbreak) or shorter-interval secular trends, e.g. seasonality.

While we chose long moving windows to capitalize on data availability, these considerations may affect the choice of moving window size and nowcast performance, depending on the data. In light of this, we experimented with moving windows of different lengths to assess the impact on nowcast performance with dengue fever data. We tested moving windows of 5, 12, and 27 weeks (approx. 6 months) and found that accuracy metrics were similar for moving windows of 12 weeks or longer (range in rRMSE: 0.6-0.655; average score: 0.35-0.37) (Table C.4; Fig. C.6). A 5-week moving window, however, produced substantially lower accuracy nowcasts (rRMSE = 7.381) with several steep case overestimates in 2007-08 and 2010 (Fig. C.6A).

4.3 DISCUSSION

We introduce a new approach for Bayesian nowcasting and demonstrate its application in two disease contexts with different reporting systems, outperforming an existing method in terms of point estimate (reduced RMSE) and probabilistic (higher logarithmic score) predictive performance. In particular, NobBS performs well even when the delays in case reporting change over time. Lacking any disease-specific parameterization, and relying only on historical trends of case reporting as input, this approach can be immediately adapted in a variety of disease settings.

Across diseases, NobBS outperformed the benchmark approach on accuracy of uncertainty estimates, and produced comparable or better point estimates. For the subset of weeks in which both models could produce forecasts (week with at least one case initially reported), point estimates for NobBS were substantially more accurate than the benchmark model for dengue cases (rRMSE improved by 300%) and slightly less accurate for ILI cases (rRMSE decreased by 19%). However, analysis of the probability distributions of the nowcasts revealed a much more substantial difference; the average score for NobBS was approximately twice as high for dengue and more than 10 times as high for ILI cases (Table 4.1). This indicates that the NobBS approach assigned much higher probability to the actual outcome, even at the cost of some point accuracy for the ILI cases.

While utilizing a similar modeling structure on case reporting delays as in ref.(9), NobBS introduces a simple dependency between case counts over time;

that is, changes in case counts between weeks are assumed to be related via a first-order random walk process on the logarithmic scale. This feature is critical in the context of infectious disease transmission, where the number of true infections in a given week mechanistically depends in part on the number of true infections in previous weeks due to the infectious process, whether the pathogen is transmitted directly or by vectors (12). Hence, variations of autoregressive models are common in disease forecasting(13, 14). When reporting delays are time-varying, as is often the case in epidemics(11), we show that the NobBS approach is less accurate, but still shows improvement over the benchmark approach likely because the NobBS approach is informed by the number of cases experienced in previous weeks, not just the delay distribution, making it more robust to larger fluctuations.

The accuracy of predictions is related at least in part to the number of cases reported to the surveillance system in week 0. When a larger proportion of cases were reported with no delay, as was the case for ILI compared to dengue, the point estimate accuracy was higher. This is not surprising, as a large fraction of true cases reported initially leaves fewer cases left to predict.

We also observe greater volatility in the nowcasts when the initial number of cases reported increases suddenly from low values. Two weeks in the dengue time series highlight this: August 3, 1998 and August 16, 2010. In those weeks, the number of cases initially increased by 16 and 17, respectively, from the previous week. Over the previous 10 weeks, for comparison, the average absolute change in initial reports was, respectively, 2.6 and 1.8. Because this increase is an

outlier in the historically-observed distribution of reporting delays, in particular for delay d=0, the model substantially overestimated the true number of cases before correcting the following week. We observed that shorter moving windows either exacerbated this issue (e.g. in 2010) or did not produce much change (e.g. in 1998) (Fig. C.6), potentially because the number of cases initially reported with delay d=0 is small to begin with relative to the total number of cases reported across delays in previous weeks. While the smooth, autocorrelated relationship fit in the NobBS model helps reduce the effect of week-to-week variability in early reporting, it remains a challenge.

Beyond supporting real-time disease tracking by public health officials, NobBS can complement existing disease forecast efforts by providing more accurate nowcasts to forecasting teams in the place of real-time reporting underestimates. It is common for teams in the CDC Epidemic Prediction Initiative (e.g. FluSight, Aedes Challenge) to experience poorer forecasts when using unrevised, surveillance data as inputs without accounting for reporting delays(15), and thus NobBS can help fill this time gap to improve prospective estimates as well.

4.4 Materials and Methods

4.4.1 Surveillance Data

We collected data on approximately 53,000 cases of dengue fever in Puerto Rico and 2.77 million cases of ILI in the United States over a 21-year (1092 weeks) and 3.75-year (196 weeks) period, respectively. Time-stamped

weekly dengue data for laboratory-confirmed cases of dengue in Puerto Rico were collected by the Puerto Rico Department of Health and Centers for Disease Control and Prevention. The times used for the analysis were the time of onset as reported by the reporting clinician and the time of laboratory report completion. ILI data originated from the U.S. Outpatient Influenza-like Illness Surveillance Network (ILINet), which consolidates information from over 2000 outpatient healthcare providers in the United States who report to the CDC on the number of patients with ILI. The times used for the analysis were the week of ILI-related care seeking and the week when those cases were posted online in FluView (https://gis.cdc.gov/grasp/fluview/fluportaldashboard.html) as collected in the DELPHI epidemiological data API (https://github.com/cmu-delphi/delphi-epidata). ILI data with delays of more than 6 months occasionally had irregularities, so we restricted the analyses to delays of up to 6 months.

4.4.2 SIMULATED ILI DATA

To simulate ILI data with a time-varying probability of reporting delay d=0, we drew, for each week, Pr(d=0) from Unif(0.2, 0.9) for all weeks in which the total number of eventually-observed cases exceeded the approximate mean of the ILI series (14,000 cases), and from Unif(0, 0.65) for all weeks in which the total observed case count was less than or equal to 14,000. This probability was used to calculate the simulated number of cases that would be observed with d=0, out of the total number of cases that would be eventually observed for that week. The remaining cases were distributed to other delays ranging from 1-52 weeks

using NB(0.9,0.4). This produced a rough approximation for a hypothetical scenario in which cases are reported faster (higher probability of d=0) during weeks with higher disease activity (more cases).

4.4.3 Reporting Triangle

Delays in reporting are often structurally decomposed into a (T x D) dimensional "reporting triangle," where T is the most recent week ("now") and D is the maximum reporting delay, in weeks, observed in the data. The data are right-truncated, since at any given week t, delays longer than $T - t$ cannot be observed. For example, at week $t=T$, only the cases reported with delay d=0 are observable; cases reported with longer delays (i.e. 1- or 2-week delays, $d=1$ or $d=2$) will be known in future weeks. In Table C.5, we present an example of the reporting triangle using ILI data.

For each week t, the goal of nowcasting is to produce estimates for the total number of cases eventually reported, N_t, based on an incomplete set of observed cases with delay d, $n_{t,d}$. Since not every $n_{t,d}$ is observed for a delay d, but will be observed at some unknown time point in the future, $N_t = \text{sum}(n_{t,d})$.

The NobBS approach is motivated by modeling the marginal cell counts of the reporting triangle, $n_{t,d}$, in an adaptation of the loglinear chain ladder method developed in actuarial literature (16).

4.4.4 NOWCASTING BY BAYESIAN SMOOTHING (NOBBS)

Let $n_{t,d}$ be the number of cases reported for week t with delay d. We assume that the underlying cases occur in a Poisson process such that

$$n_{t,d} \sim Pois(\lambda_{t,d}).$$

We also allow for extra-Poisson variation, that is, when the variance is larger than the mean and a negative binomial process (of which the Poisson is a special case) is more appropriate. We apply this in the case of the influenza data:

$$n_{t,d} \sim NB(r, p_{t,d}), \text{ where}$$

$$p_{t,d} = r/(r + \lambda_{t,d}).$$

We then model the mean, $\lambda_{t,d}$, as a simple log-linear equation

$$log\left(\lambda_{t,d}\right) = \alpha_t + log\left(\beta_d\right),$$

where α_t represents the true epidemiologic signal for week t and β_d as the probability of reporting with delay=d. In other words, NobBS contains random effects for week t and the reporting delay d. Exponentiating both sides of the equation, it is clear to see that $\lambda_{t,d} = e^{\alpha_t} * \beta_d$.

We place prior distributions on α_t and β_d reflecting properties of each parameter. Since β_d represents a probability vector containing delays = 0, …, D, we place on it a Dirichlet prior of length D:

$$\beta_d \sim Dir(\theta)$$

$$\theta = (\theta_0, \ldots, \theta_D).$$

The maximum delay D can be identified as the maximum observable delay in the data, which may change as the time series extends, or can be fixed at some value D thought to represent a very long delay. In the latter case, θ_D can be modeled as

the probability of delay ≥ *D*. For dengue fever, we choose to fix *D* at 10 weeks, since over 99% of the cases observed in the first two years (prior to producing out-of-sample nowcasts) were reported within 10 weeks. For influenza, we chose *D* to be the longest possible delay within the 27-week moving window, or *D*=26. To be precise, the implications of choosing a maximum delay *D* within a moving window of *W* weeks means that the nowcast will be a slight underestimate, as the estimate for delays greater than or equal to *D* is based on partially-observed information from previous weeks, and delays longer than *W* are unobserved by the model (see the reporting triangle in Table C.5). Technically speaking, then, NobBS produces an underestimate for the number of cases that will eventually be reported within the moving window. However, since the vast majority (>99%) of cases are typically reported within delay *D*, we feel these model constraints are negligible.

We place weakly informative priors on θ representing a small number of hypothetical total cases (10) distributed across delay bins, loosely representing the probability of reporting delays for each delay *d* observed in the first two years of data for dengue fever and the first 6 months of data for ILI (training periods).

We allow a dependency between successive α_t's to capture the time evolution and autocorrelation of cases from week-to-week, commonly exhibited by epidemic curves. We therefore model α_t as a first-order random walk:

$$\alpha_{t=1} = Normal(0, c_\alpha^2)$$

$$\alpha_{t>1} \sim Normal(\alpha_{t-1}, \tau_\alpha^2)$$

Because α_t is in natural log form, this constitutes a geometric random walk.

We place weakly informative priors on the precisions of the Normal distribution, c_{α}^{2}=0.001 and $\tau_{\alpha}^{2} \sim Gamma(0.01, 0.01)$. For the negative binomial stopping-time parameter, r, we place an informative Gamma(60,20) prior to reflect belief that the process deviates moderately from the Poisson.

Models were compiled in JAGS on R (v 3.3.2) using the package "rjags" producing 10,000 posterior samples after a burn-in period of 100 iterations.

4.4.5 NOWCAST ESTIMATES

We produced weekly nowcasts beginning with the 27th week (influenza) and 104th week (dengue fever) and through the final week of the series. This resulted in 989 weekly out-of-sample estimates of dengue fever cases and 170 weekly out-of-sample estimates of ILI.

We used a two-year moving window to estimate a stable delay distribution within the window. As a sensitivity, and to gauge the minimum amount of historical information required to produce accurate nowcasts, we also applied moving windows of 5, 12, and 27 weeks (approximately 6 months).

We used as a benchmark for comparison the "nowcast" function of the R package "surveillance" by Höhle and an der Heiden (described in ref. (9)) designed to produce Bayesian nowcasts for epidemics using a hierarchical model for $n_{t,\, d \leq T-t} \mid n_{t,d}$, or the observed cases conditional on the expected total number of cases. We applied the function assuming a time-homogenous delay distribution and recommended parameterization described by the authors in http://staff.math.su.se/hoehle/blog/2016/07/19/nowCast.html, and for

comparability, used the same moving window sizes (27 and 104 weeks) to produce nowcasts over the same time periods.

4.4.6 Model Performance Metrics

The mean absolute error (MAE), root mean square error (RMSE) and relative root mean square error (rRMSE) are defined, respectively, as:

$$MAE = \frac{1}{n}\sum_{i=1}^{n} abs(y_i - x_i)$$

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - x_i)^2}$$

$$rRMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(\frac{y_i - x_i}{y_i})^2}$$

and were used to quantify the accuracy of point estimates, x_i, compared to true case numbers, y_i, across the different models.

To quantify the accuracy of the point estimates in capturing the *change* in cases from week *t-1* to week *t*, we computed the mean absolute error of the change (MAEΔ) and the RMSE of the change (RMSEΔ):

$$MAE\Delta = \frac{1}{n-1}\sum_{i=2}^{n} abs((x_i - x_{i-1}) - (y_i - y_{i-1}))$$

$$RMSE\Delta = \sqrt{\frac{1}{n-1}\sum_{i=2}^{n}((y_i - y_{i-1}) - (x_i - x_{i-1}))^2}$$

To capture smoothness in predictions from week-to-week, we also calculated the lag-1 autocorrelation of predictions (ρ_a):

$$\rho_a = \frac{\sum_{i=2}^{n}(x_i - \bar{x})(x_{i-1} - \bar{x})}{\sum_{i=2}^{n}(x_i - \bar{x})^2}$$

The logarithmic scoring rule was used to quantify the accuracy of the posterior predictive distribution of the nowcast. Predictive distributions were assigned to a series of bins categorized across possible values of true case counts. We used bin widths of 25 cases for dengue fever and 1000 cases for influenza, allowing for a larger number of bins for ILI cases based on case ranges of approx. 0-400 for dengue fever and 4,000-40,000 for ILI. For a predictive distribution with binned probability pi for a given nowcast target, the logarithmic score was calculated as ln(pi). For example, there were 115 cases eventually observed for the week of January 20, 1992. The NobBS nowcast for this week, which assigned a probability of 0.4 to the bin [100,125), thus received a log score of ln(0.4) = -0.92. As in (15, 17), a very low log score of -10 was assigned for weeks in which the predictive distribution did not include the true case value and for weeks in which the bin probability ≤ e-10. This rule provides a lower limit (-10) to the score of highly inaccurate predictions.

The average log score across all prediction weeks was computed for all models to assess nowcast performance. The exponentiated average log score yields a nowcast score that can be interpreted as the average probability assigned to the bin corresponding to the true number of cases, and is a metric for model comparison purposes used in several other forecast contexts (15, 17). In this

paper, we present the exponentiated average log score and refer to this as the average score.

4.5 REFERENCES

1. Lipsitch M, et al. (2011) Improving the evidence base for decision making during a pandemic: the example of 2009 influenza A/H1N1. *Biosecur Bioterror* 9(2):89–115.

2. Thacker SB, Berkelman RL, Stroup DF (1989) The Science of Public Health Surveillance. *J Public Health Policy* 10(2):187.

3. Gikas A, et al. (2004) Prevalence, and associated risk factors, of self-reported diabetes mellitus in a sample of adult urban population in Greece: MEDICAL Exit Poll Research in Salamis (MEDICAL EXPRESS 2002). *BMC Public Health* 4:2.

4. Kaminsky KS (1987) Prediction of IBNR claim counts by modelling the distribution of report lags. *Insur Math Econ* 6(2):151–159.

5. Lawless JF (1994) Adjustments for reporting delays and the prediction of occurred but not reported events. *Can J Stat* 22(1):15–31.

6. Pagano M, Tu XM, De Gruttola V, MaWhinney S (1994) Regression Analysis of Censored and Truncated Data: Estimating Reporting- Delay Distributions and AIDS Incidence from Surveillance Data. *Biometrics* 50(4):1203.

7. Comiskey CM, Ruskin HJ (1992) AIDS in Ireland: the reporting delay distribution and the implementation of integral equation models. *Comput Appl Biosci* 8(6):579–581.

8. Cui J, Kaldor J (1998) Changing pattern of delays in reporting AIDS diagnoses in Australia. *Aust N Z J Public Health* 22(4):432–435.

9. Höhle M, an der Heiden M (2014) Bayesian nowcasting during the STEC O104:H4 outbreak in Germany, 2011. *Biometrics* 70(4):993–1002.

10. Salmon M, Schumacher D, Stark K, Höhle M (2015) Bayesian outbreak detection in the presence of reporting delays. *Biom J* 57(6):1051–1067.

11. Noufaily A, et al. (2015) Modelling reporting delays for outbreak detection in infectious disease data. *J R Stat Soc A* 178(1):205–222.

12. Wallinga J, Teunis P (2004) Different epidemic curves for severe acute respiratory syndrome reveal similar impacts of control measures. *Am J Epidemiol* 160(6):509–516.

13. Johansson MA, Reich NG, Hota A, Brownstein JS, Santillana M (2016) Evaluating the performance of infectious disease forecasts: A comparison of climate-driven and seasonal dengue forecasts for Mexico. *Sci Rep* 6:33707.

14. Yang S, Santillana M, Kou SC (2015) Accurate estimation of influenza epidemics using Google search data via ARGO. *Proc Natl Acad Sci U S A* 112(47):14473–14478.

15. Reich NG, et al. (2019) A collaborative multiyear, multimodel assessment of seasonal influenza forecasting in the United States. *Proc Natl Acad Sci U S A* 116(8):3146–3154.

16. Kremer E (1982) IBNR-claims and the two-way model of ANOVA. *Scandinavian Actuarial Journal* 1982(1):47–55.

17. McGowan CJ, et al. (2019) Collaborative efforts to forecast seasonal influenza in the United States, 2015-2016. *Sci Rep* 9(1):683.

5

Conclusion and Summary

In this thesis, I present a collection of flexible, self-adaptive prediction models. I demonstrate these models in a range of disease and geographic contexts, and assess their performance and ability to complement traditional public health surveillance efforts. I start by examining the role of digital data (Google, Twitter, and digital news reports) in providing estimates for weekly Zika incidence in five Latin American countries, up to three weeks ahead of the release of official surveillance reports. Across all models presented, I show that Google data plays a key role in capturing rapidly-changing signals of population health activity in real-time, and thus becomes more useful in producing further-ahead weekly predictions compared to historical case information. Importantly, I demonstrate that even outbreak features present in the 2015-6 Zika epidemic such as (1) novelty of the disease and (2) intense media coverage – factors which might make for a noisy signal coming from digital, search- and report-based data – can be accommodated by the model. While Internet penetration differed dramatically across study countries, these models show good performance across sites as well as the ability to adapt and learn from new information: important criteria for model generalizability.

In Chapters 3 and 4 I develop more flexible and generalizable models that perform well across multiple, finer spatial scales and across multiple disease and surveillance contexts. Again, I draw from data that can be freely and digitally collected, either from external bodies (NASA weather data) or intrinsic properties of surveillance data (time-stamped reporting data). I show that using a single modeling framework, weather patterns can be systematically and autonomously extracted, analyzed, and used to make forecasts of dengue fever outbreaks at the city-level in Brazil. By analyzing the consistency of weather patterns across years and cities, it is then possible to identify signatures of successful predictions as well as explain where predictions might fail: an important criterion for both model transparency and generalizability. I show that, for some cities that experience highly accurate annual epidemic predictions, there are clear and environmentally-significant time periods that produce strong signals for an outbreak, including rainy and dry seasons. Cities where predictions were less accurate tended to have no clear weather patterns, either suggesting the need for potentially more complex weather inputs or revealing the importance of other, non-weather factors such as behavior, policy, land use, or simply stochastic noise. While this model serves as a proof-of-concept using just two simple weather inputs, temperature and precipitation, the framework can be easily extended to multiple weather variables such as humidity, soil water absorption, and more, allowing for more complex patterns to emerge.

Finally, I construct an approach to generate more accurate real-time estimates of disease activity (nowcasts) to support public health decision-making. Combatting the common issue of reporting delays that yield real-time case

underestimates, I test a Bayesian approach to predict the number of not-yet-reported cases over multiple years for two different surveillance systems, dengue fever in Puerto Rico and influenza-like-illness in the United States. I show that the model is accurate across prediction contexts and time periods, and outperforms an established Bayesian nowcast model. Importantly, the approach models an autocorrelated, underlying case accrual process that improves predictions even when the reporting delay is time-varying. Because the model does not require inputs other than what is already collected by surveillance systems, this approach can be readily implemented in the public health system.

These projects, while diverse, all share one common goal: through generalizable learning models, they produce accurate and timely predictions of disease activity to complement traditional public health surveillance and curtail the problems faced by reporting delays and delayed action. In public health, timing can be everything: anticipating outbreaks makes the difference between effectively interrupting disease transmission and failing to catch a growing threat. In this thesis, I show that it is possible to rely on a wide range of data streams to make both actionable and timely predictions that improve disease surveillance.

A

Supporting Information for Chapter 2

A.1 EQUATIONS: MODEL PERFORMANCE METRICS

Equation A.1.1. Root Mean Square Error (RMSE)

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - x_i)^2} \quad (A.1.1)$$

Equation A.1.2. Relative Root Mean Square Error (rRMSE)

$$rRMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(\frac{y_i - x_i}{y_i})^2} \quad (A.1.2)$$

Equation A.1.3. Pearson Correlation

$$\rho = \frac{\sum_{i=1}^{n}(y_i - \bar{y})(x_i - \bar{x})}{\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}} \quad (A.1.3)$$

Table A.1. Google search terms used as input variables for each country.

Colombia	Venezuela	Honduras	El Salvador	Martinique
zika	zika	zika	zika	zika
sintomas	Guillain barre	sintomas zika	sintomas zika	zika martinique
zika sintomas	el zika	el zika	sintomas del zika	le zika
el zika	sintomas	sintomas del zika	zika enfermedad	zika symptomes
sintomas del zika	sintomas zika	zika virus	zika virus	zika symptome
virus zika	virus	que es zika	sintomas de zika	zika virus
virus	zika virus	sintomas de zika	sika	symptome du zika
zika colombia	que el zika	enfermedad zika	que es zika	symptomes zika
que es zika	sintomas del zika	zika en honduras	zika tratamiento	
virus del zika	que es zika	virus del zika	sica	
el sika	la zika	sika	guillain barre	
el zika sintomas	tratamiento zika		sica el salvador	
sintomas de zika	sintomas de zika			
zika embarazo	el zika virus			
zika microcefalia				
zika sintomas colombia				
zika fiebre				
sika				
sintomas del				
el sica				
zika repite				

Table A.2. Comparison of models in Colombia and Venezuela, with and without Twitter data. RMSE, rRMSE, and Pearson's correlation coefficient (ρ) are shown for 1-, 2-, and 3-week ahead out-of-sample predictive performance. The best rRMSE for each model pair with and without Twitter data (i.e. ARGO vs. ARGO+T) is shown in bold.

	Colombia								
	1 week			2 week			3 week		
Model	RMSE	rRMSE	ρ	RMSE	rRMSE	ρ	RMSE	rRMSE	ρ
AR	801.313	40.462	0.821	1484.018	66.829	0.539	2057.483	83.900	0.284
Google only	767.867	35.567	0.786	804.209	38.882	0.772	959.216	43.119	0.646
G+T	823.149	**34.450**	0.764	857.490	**37.300**	0.752	995.311	**41.903**	0.634
ARGO	628.096	30.181	0.866	798.808	40.176	0.763	930.665	**44.104**	0.660
ARGO+T	621.673	**30.076**	0.870	775.786	**39.583**	0.780	914.643	44.233	0.679
ARGO+H	631.882	30.262	0.864	892.063	41.189	0.707	953.619	43.558	0.649
ARGO+TH	617.795	**29.888**	0.871	848.968	**40.153**	0.731	903.155	**42.440**	0.698

	Venezuela								
	1 week			2 week			3 week		
	RMSE	rRMSE	ρ	RMSE	rRMSE	ρ	RMSE	rRMSE	ρ
AR	1665.733	68.542	0.822	4196.484	117.444	0.834	10349.050	259.699	0.665
Google only	413.265	**28.706**	0.952	694.306	**32.275**	0.855	659.727	**30.112**	0.896
G+T	972.937	35.336	0.626	1277.588	39.813	0.283	1226.614	39.953	0.475
ARGO	1629.280	50.795	0.829	3565.201	80.405	0.841	7325.554	173.308	0.659
ARGO+T	892.063	**38.780**	0.831	927.343	**41.946**	0.701	1372.884	**48.249**	0.486
ARGO+H	1509.605	54.637	0.808	2573.568	78.326	0.862	4628.385	115.996	0.740
ARGO+TH	1036.760	**46.497**	0.771	1148.229	**67.028**	0.626	1459.830	**75.513**	0.528

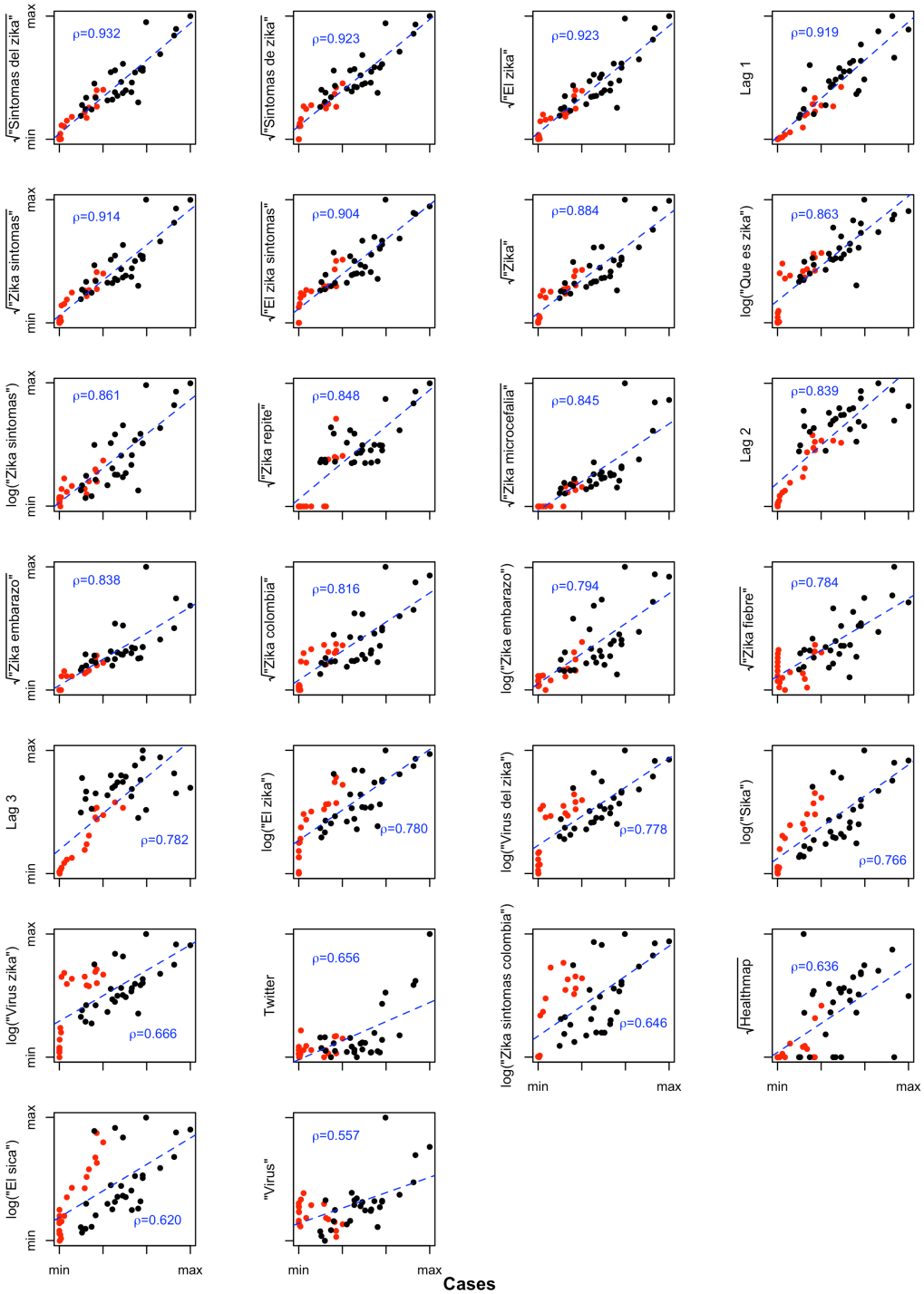

Figure A.1. Correlation of digital predictors with official suspected Zika case counts in Colombia. The transformation that produced the highest correlation with Zika cases for each variable is shown in each plot. Data points from weeks within the training period are distinguished in red.

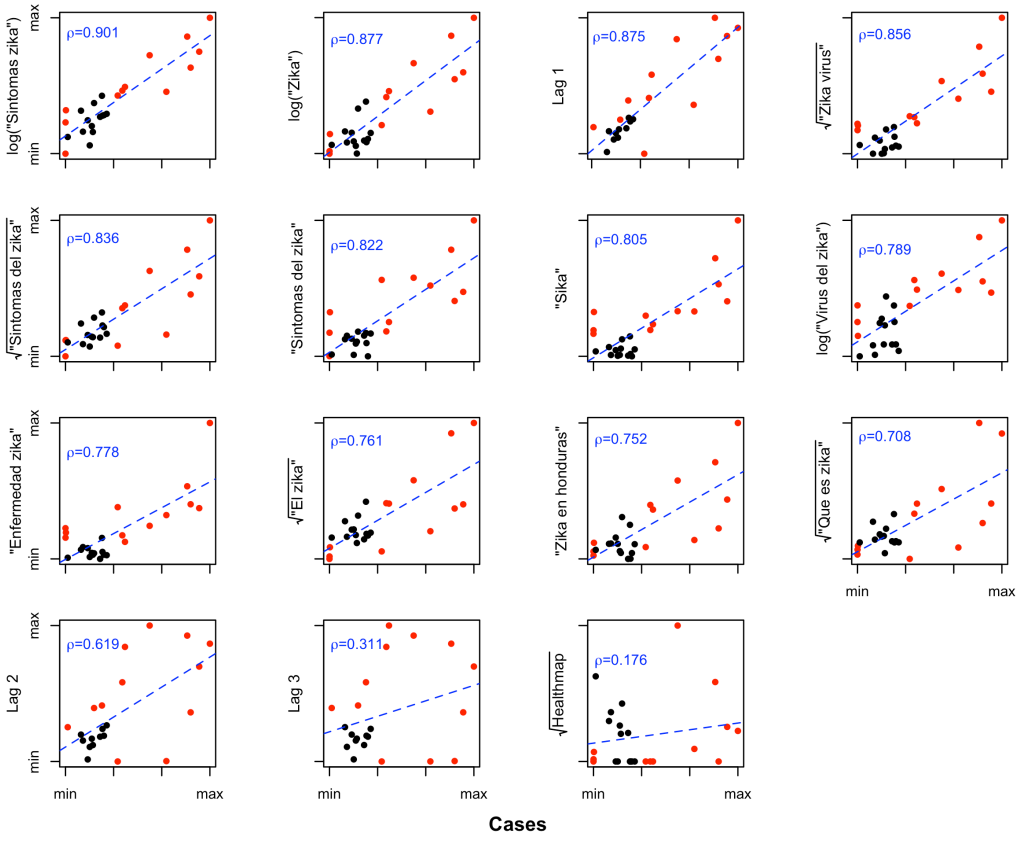

Figure A.2. Correlation of digital predictors with official suspected Zika case counts in Honduras. The transformation that produced the highest correlation with Zika cases for each variable is shown in each plot. Data points from weeks within the training period are distinguished in red.

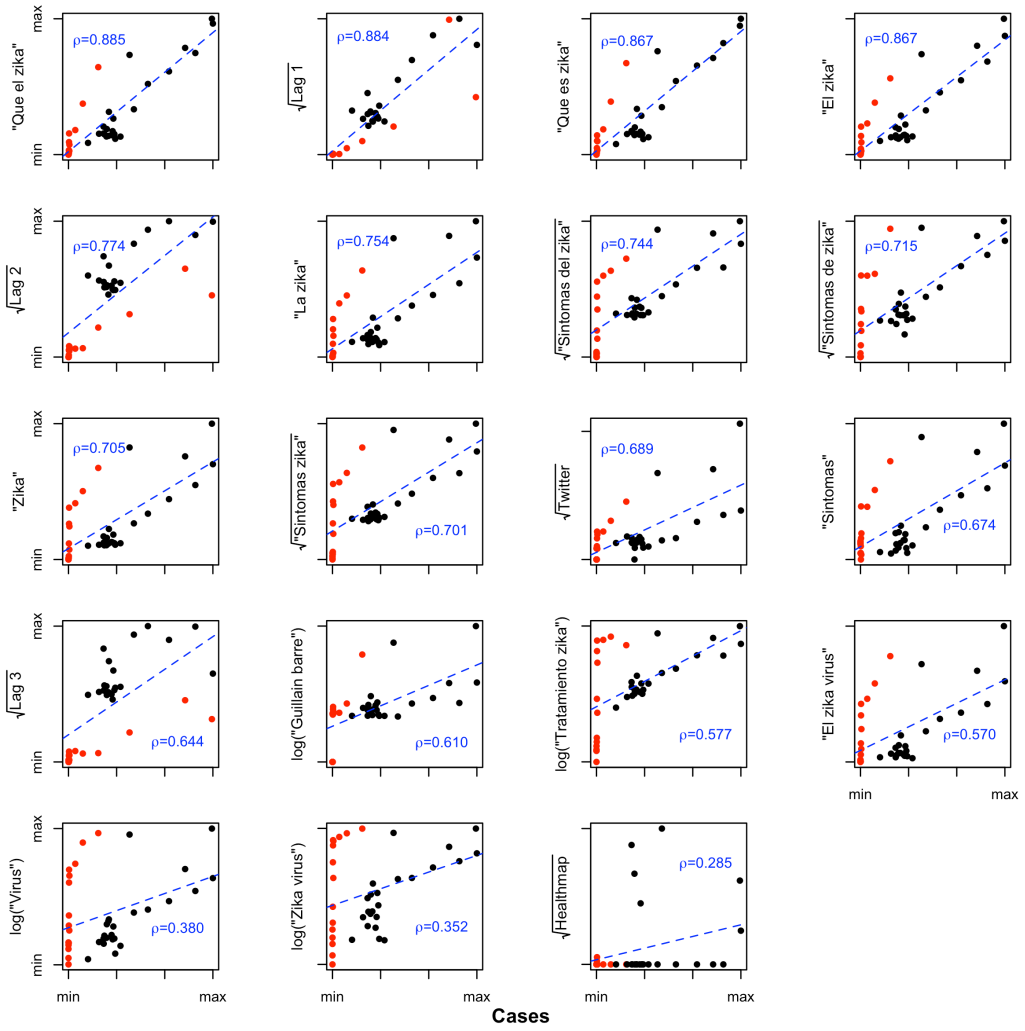

Figure A.3. Correlation of digital predictors with official suspected Zika case counts in Venezuela. The transformation that produced the highest correlation with Zika cases for each variable is shown in each plot. Data points from weeks within the training period are distinguished in red.

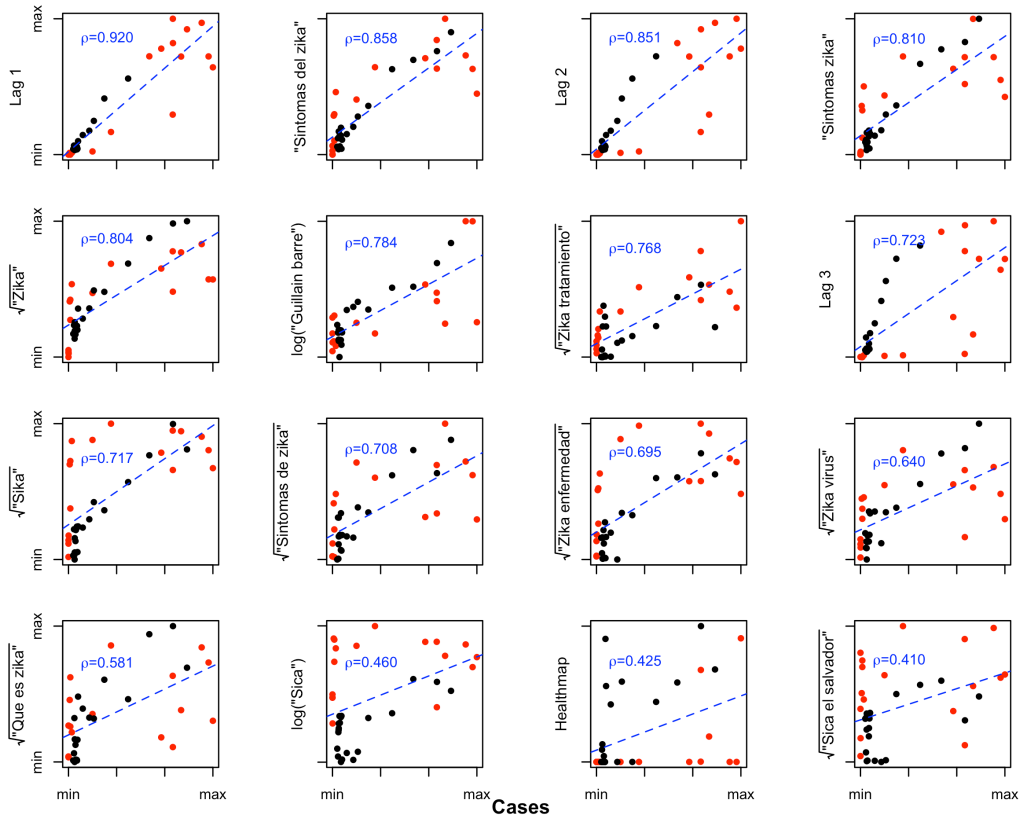

Figure A.4. Correlation of digital predictors with official suspected Zika case counts in El Salvador. The transformation that produced the highest correlation with Zika cases for each variable is shown in each plot. Data points from weeks within the training period are distinguished in red.

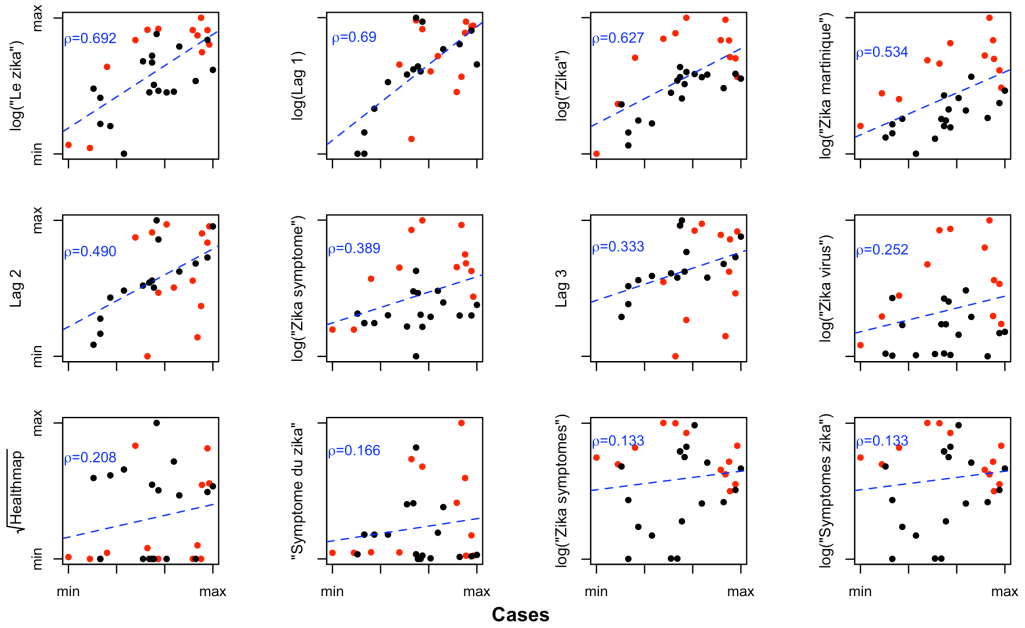

Figure A.5. Correlation of digital predictors with official suspected Zika case counts in Martinique. The transformation that produced the highest correlation with Zika cases for each variable is shown in each plot. Data points from weeks within the training period are distinguished in red.

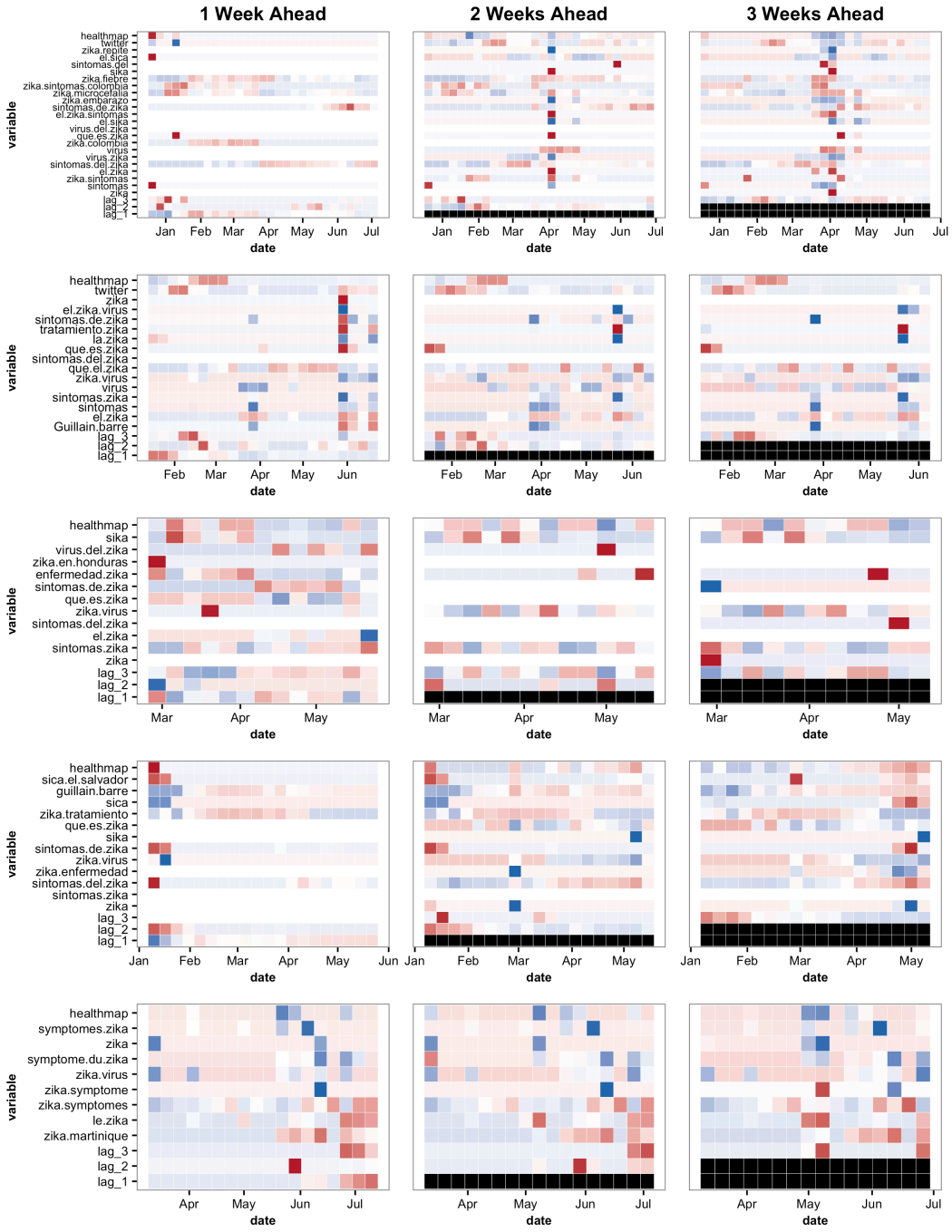

Figure A.6. Heatmaps showing the relative influence (positive: red; negative: blue) of all input variables on predictions of Zika cases in (a) Colombia, (b) Honduras, (c) Venezuela, (d) El Salvador, and (e) Martinique.

B

Supporting Information for Chapter 3

B.1 SUPPLEMENTAL MATERIALS & METHODS

B.1.1 STUDY SITES

We tested our model on a time series of 17 years (2001-2017) for 20 municipalities in Brazil meeting the following criteria: 1) having experienced between 7 and 10 epidemic years during the 17-year time period, with no more than 70% of epidemic years occurring in either half of the time period; 2) having a population over 100,000 by 2017. The first criterion captures a loose definition of a "dengue-endemic" location, which is convenient in two ways: first, locations which experience only epidemic or only non-epidemic years are likely experiencing disease dynamics unrelated to annual changes in weather patterns, and are thus not appropriate for our model; and second, it ensures that our model is able to train initially on both classes (epidemic and non-epidemic year) for each location before making out-of-sample predictions. In accordance with Brazilian Ministry of Health standards, we defined an epidemic year to be a year in which the number of confirmed cases of dengue fever exceeds 100 per 100,000 persons.

The municipalities included in the study span a wide geographic range (14 Brazilian states) and range in land area from 24 to 4000 mi^2, in starting population (population in 2001) of between 87,000 and 6 million, and in starting population density from 24 to 13,000 persons/mi^2 (Table S1).

B.1.2 EPIDEMIOLOGIC DATA

The number of confirmed cases of dengue fever are reported annually at the municipal level and made publicly available from the Brazilian Notifiable Disease

System (SINAN) for the years 2001-2012. We obtained weekly case numbers at the municipal level from the Brazilian MOH for the years 2013-2015, and from local municipal governments and epidemiologic reports for 2016 and 2017.

B.1.3 DEMOGRAPHIC DATA

Population estimates by year and municipality were obtained as publicly-available data from the Brazilian Institute of Geography and Statistics (IBGE).

B.1.4 WEATHER DATA

Globally modeled and assimilated weather data were obtained from the Modern Era Retrospective-analysis for Research and Applications, Version 2 (MERRA-2)[37]. The MERRA-2 data are publicly available through the Global Modeling and Assimilation Office (GMAO) at NASA Goddard Space Flight Center. We obtained daily temperature at 2 meters (mean, K) and hourly precipitation (kg/m2) at a native grid resolution of 0.5° x 0.625° and extracted these to municipalities by overlaying a spatial file of municipality boundaries and taking the weighted average of the grid cells covering municipal boundaries. We calculated the total accumulated rainfall in a day (mm) as the sum of hourly precipitation ($kg/m^2/hr$, which is equivalent to mm/hr) over the 24-hour period. We show the time series of mean daily temperature (K) and total precipitation (mm) in Fig. B.6.

B.1.5 ENSEMBLE STRENGTH

We computed a simple metric to quantify the strength of the 11-model ensemble used to make yearly forecasts for each municipality. For each model, the metric was the sum of (a) the historical out-of-sample forecast accuracy of the time

window, computed as the number of years correctly predicted divided by N prior years, and (b) the average historical out-of-sample accuracy of the time window and its (up to) 8 surrounding neighbors ($t0 \pm 5$, $p \pm 5$). Thus, the maximum possible strength for any given ensemble was 2.0 (1.0 + 1.0).

Table B.1. Population and land characteristics of 20 dengue-endemic study cities.

City	**Area (mi²)**	**Population***	**Population Density (pp/mi²)**
Rio de Janeiro	485	6320000	13030.93
Belo Horizonte	127.8	1433000	11212.83
Aracajú	70.22	571149	8133.71
São Luís	319	958,545	3004.84
Sertãozinho	155.6	101784	654.14
Manaus	4402	1793000	407.31
Rondonópolis	1608	144049	89.58
São Gonçalo	91.6	337273	3682.02
Barra Mansa	211.3	171125	809.87
Eunápolis	462	93413	202.19
Tres Lagoas	3941	96341	24.45
Barueri	24.78	240749	9715.46
SaoVicente	87.65	332445	3792.87
Juazeiro do Norte	95.97	249939	2604.35
Parnaiba	168.2	145705	866.26
SantaCruz	142.7	87582	613.75
Maranguape	228.1	113561	497.86
Barretos	604	112101	185.60
Ji-Paraná	2663	116610	43.79
Guaruja	55.44	290752	5244.44

*city proper. Source: Demographic Statistics Database, United Nations Statistics Division 2010

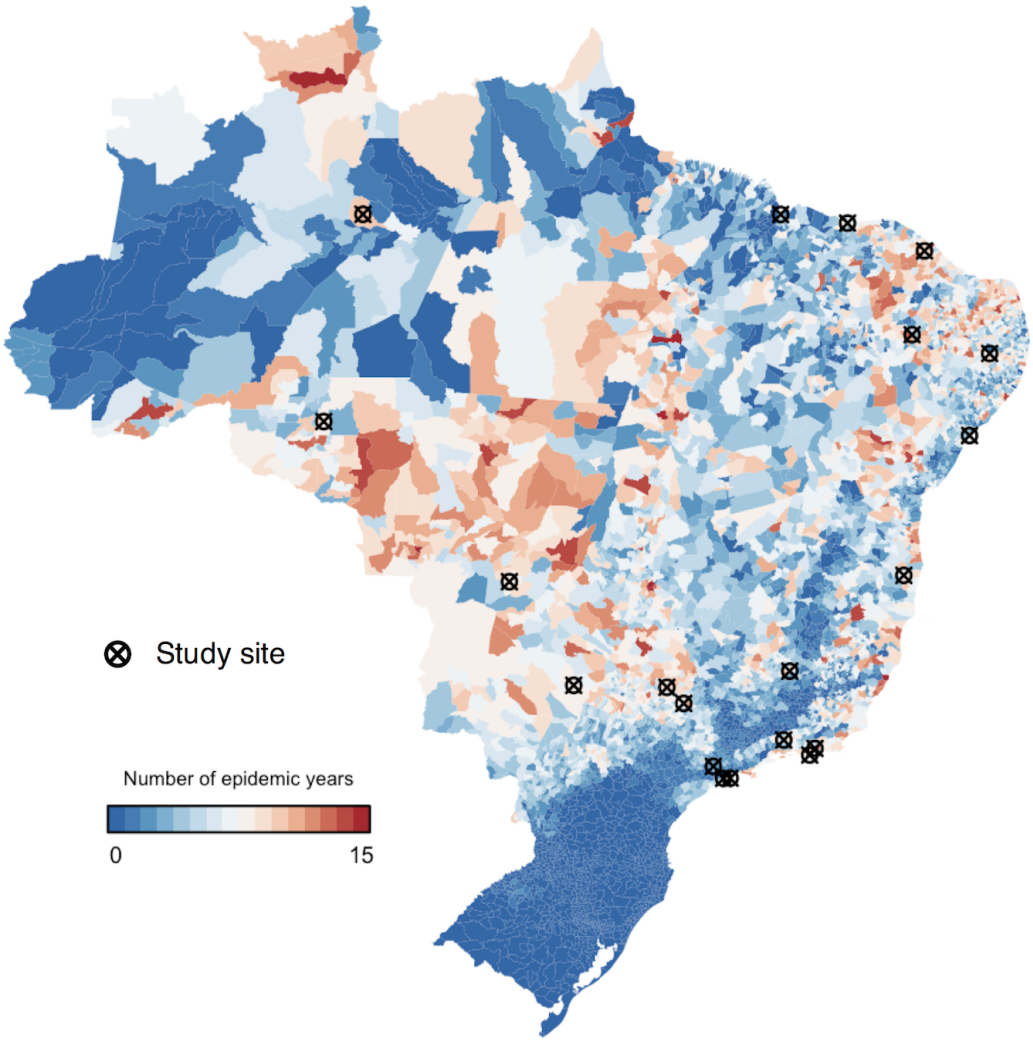

Figure B.1. Number of dengue fever epidemic years in Brazil, 2001-2015. Data on annual cases for all municipalities in Brazil were available through 2015, shown here, and we obtained data separately through 2017 for the 20 study municipalities (black crossed circles).

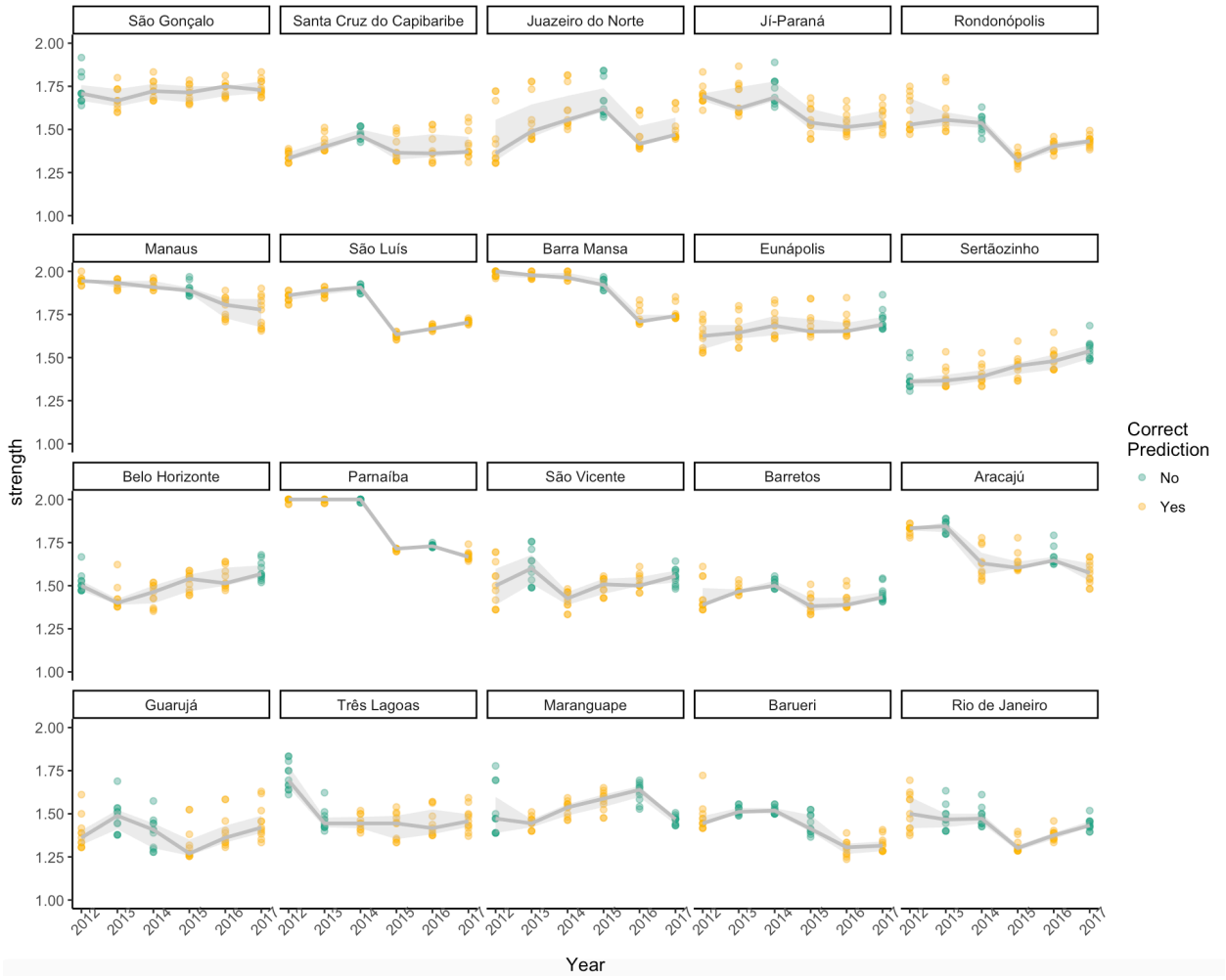

Figure B.2. Ensemble strength for 120 municipality-years (2012-2017). Ensemble strength is calculated for each of the 11 time windows selected into the ensemble each year, as a function of the historic out-of-sample accuracy of (a) the selected time window and (b) neighboring time windows (see Section B Materials & Methods). Municipalities are ordered by decreasing ensemble prediction accuracy; that is, the proportion of years correctly forecasted by the ensemble method over the years 2012-2017. Points are colored by prediction result (yellow=correct; green=incorrect).

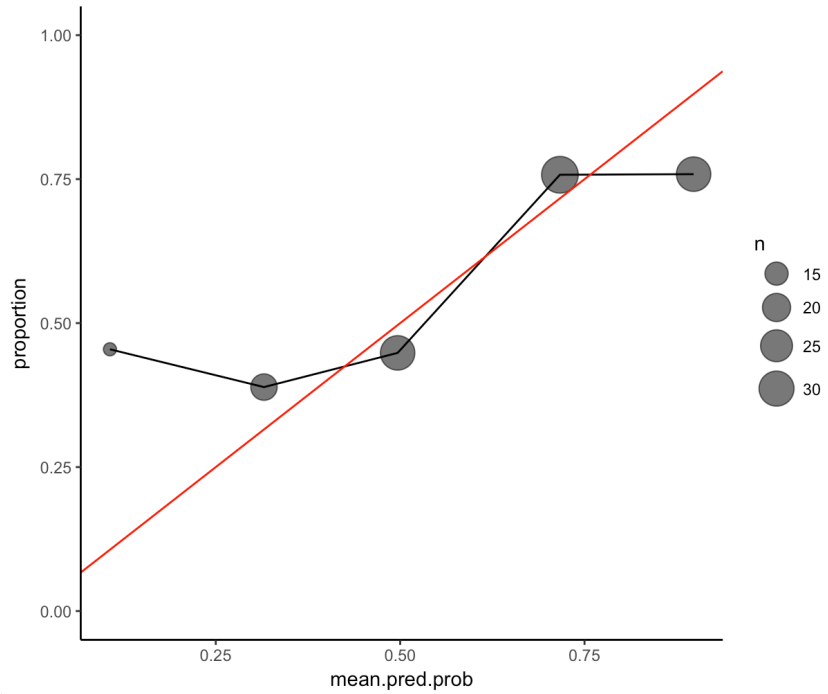

Figure B.3. Calibration curve of mean posterior probabilities over 120 municipality-years (2012-2017).

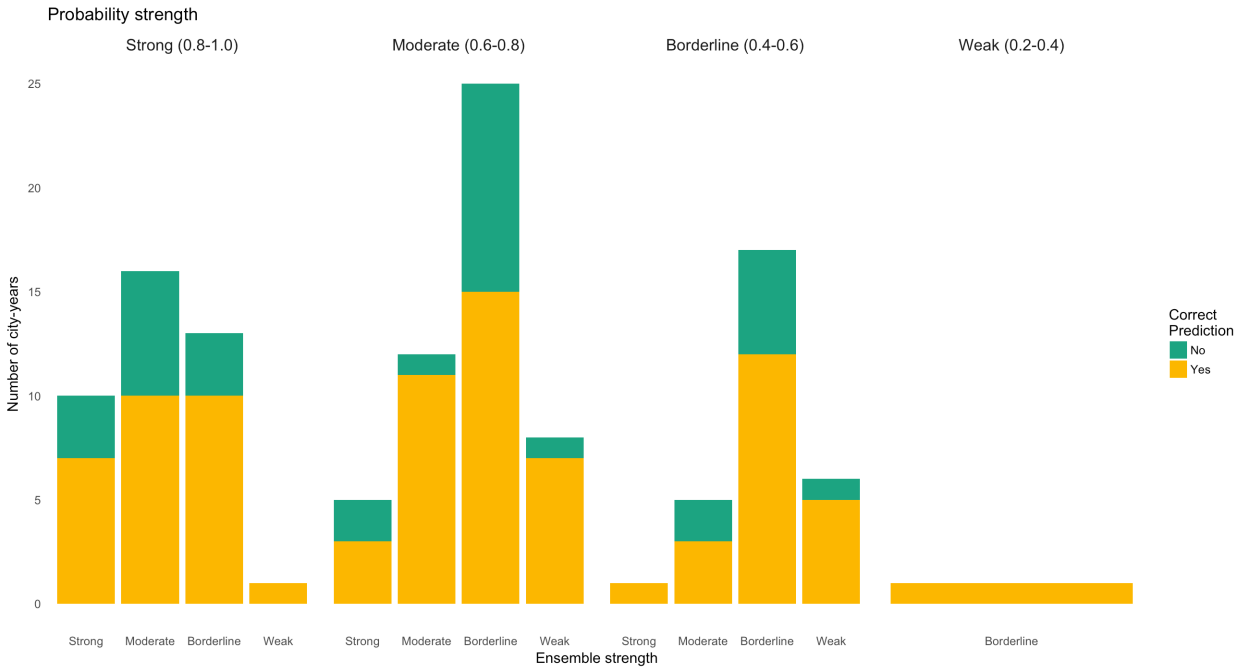

Figure B.4. Classifier and ensemble strengths. Categories of classifier probability strength (weak: 0.2-0.4, borderline: 0.4-0.6, moderate: 0.6-0.8, and strong: 0.8-1.0) and ensemble strength (weak: 1.2-1.4, borderline: 1.4-1.6, moderate: 1.6-1.8, and strong: 1.8-2.0). The classifier probability is the mean posterior class probability, computed as P(Epidemic) for predicted epidemics and 1-P(Epidemic) for predicted non-epidemics, averaged over the 11 models of the ensemble. See Supporting Information Materials & Methods for calculation of the ensemble strength metric. There were no instances of probabilities < 0.2 nor of ensemble strengths < 1.2.

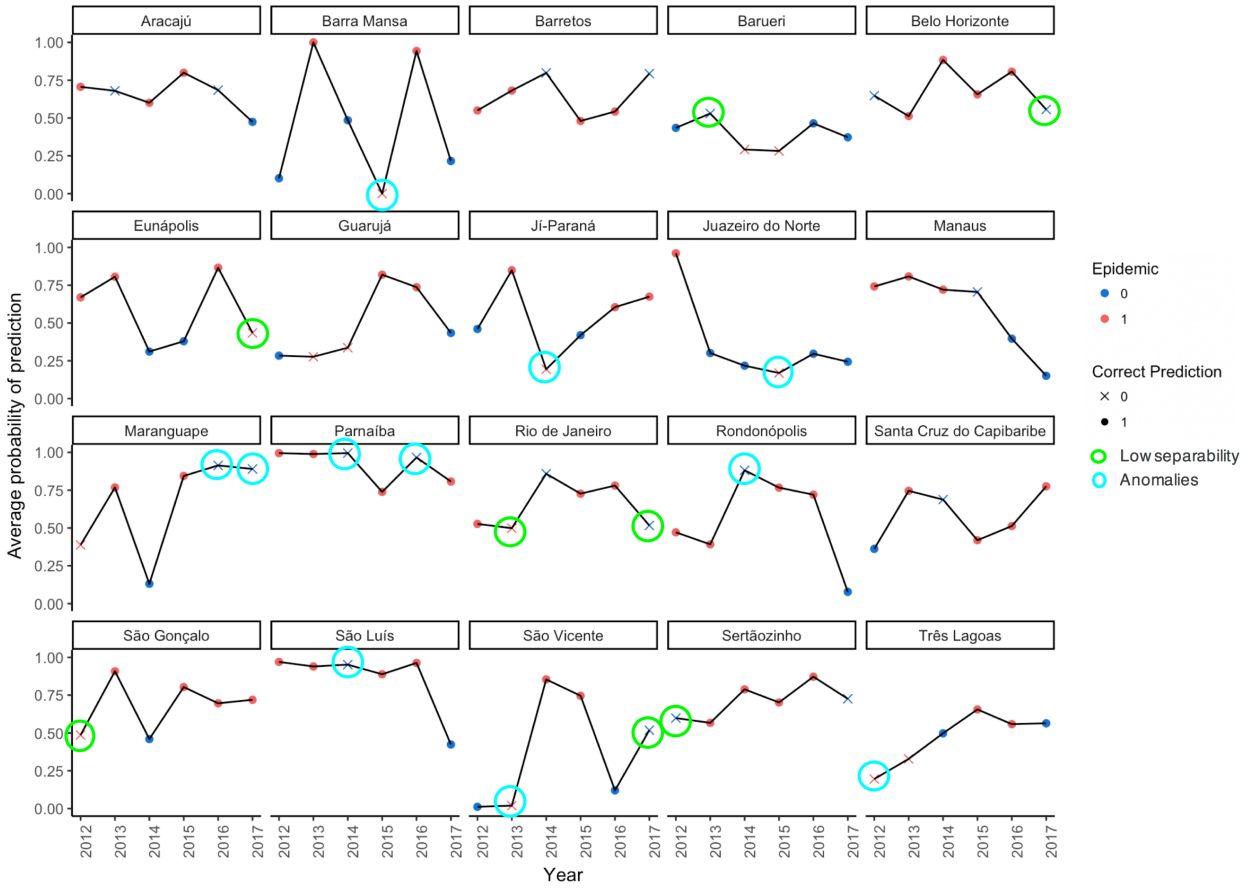

Figure B.5. Potentially anomalous or weakly-separable municipality-years for prediction. (A) Predicted probabilities by municipality and year for ensemble forecasts (2012-2017). Predictions are colored by their true epidemic status (red=epidemic, blue=non-epidemic) with point shape indicating accuracy (closed circle=correct, cross=incorrect). A cyan circle designates potentially anomalous years, defined as years that were incorrectly predicted with strong conviction (mean posterior predicted class probability ≥ 0.8). A bright green circle designates years potentially following periods with low separability, defined as years that were misclassified with borderline conviction (0.4 ≤ mean posterior predicted class probability < 0.6).

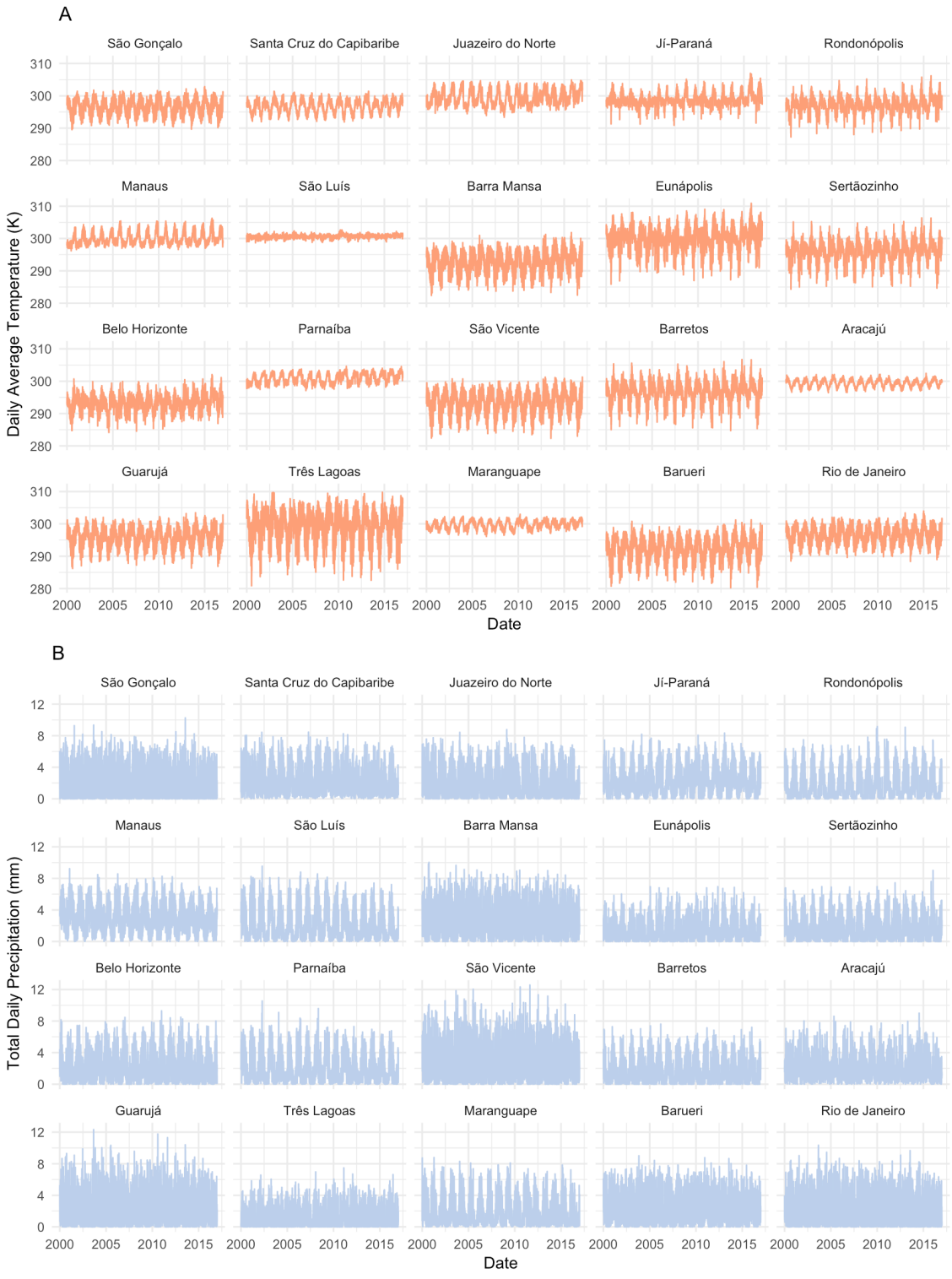

Figure B.6. Daily time series of weather inputs: 2000-2016 patterns of A) average temperature (K) and B) total precipitation (mm), by municipality.

C

Supporting Information for Chapter 4

Table C.1. NobBS average error for moderate and high absolute changes in initial case reports, compared to the previous week.

		Average error (predicted - actual) for weeks where:	
Disease	**Threshold**	**Δ initial reports > threshold**	**Δ initial reports <= threshold**
Dengue	Moderate: 5 cases	61.6	-1.8
	High: 10 cases	277.0	-0.2
ILI	Moderate: 1,000 cases	-92.8	-38.4
	High: 2,500 cases	-198.9	-37.7

Table C.2. Comparing model performance on influenza reports with constant and non-constant delay distributions.

Model	Influenza: Constant delay						Influenza: Non-constant (time-varying) delay				
	Period	MAE	rRMSE	RMSE	Average Score	95% PI coverage	MAE	rRMSE	RMSE	Average Score	95% PI coverage
NobBS	06/30/2014 - 03/14/2016	777.9	0.081	1135.2	0.172	1.00	3476.5	0.302	4622.7	0.06	0.93
Benchmark (ref. 9)	06/30/2014 - 03/14/2016	689.9	0.072	15559.2	0.016	0.00	7315	0.621	10300.4	8.71E-05	0.57

Table C.3. Performance measures for estimates of the change in ILI incidence from the previous week, comparing constant and non-constant ILI delay distributions.

Model	Influenza: Constant delay					Influenza: Non-constant (time-varying) delay			
	Period	MAEΔ	RMSEΔ	ρ_a	RMAΔ	MAEΔ	RMSEΔ	ρ_a	RMAΔ
NobBS	06/30/2014 - 03/14/2016	804	1268.3	0.96	1.01	3559.8	6745.3	0.78	3.77
Benchmark (ref. 9)	06/30/2014 - 03/14/2016	758	1252.7	0.96	1.08	8518.1	14169.3	0.60	7.63

Table C.4. Select performance measures for dengue fever nowcast model with different moving window sizes.

Model	Moving window size	rRMSE	Average Score	Correlation
NobBS	5 weeks	7.381	0.368	0.275
	12	0.634	0.370	0.760
	27 weeks (approx. 6 months)	0.655	0.369	0.806
	104 weeks (approx. 12 years)	0.600	0.349	0.84

Table C.5. Reporting triangle for June 8, 2015. The *TxD* reporting triangle decomposes the number of cases, $n_{t,d}$, reported for each week *t* (rows) and each delay *d* (column). Here we show a reporting triangle containing reports up to a delay of $D = 18$ weeks. The goal of nowcasting is to predict the missing (NA) $n_{t,d}$'s.

	Delay *d* (weeks)																		
Week *t*	**0**	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**	**16**	**17**	**18**
2/23/2015	15911	1955	504	306	340	104	7	4	78	31	15	8	0	23	3	0	NA	NA	NA
3/2/2015	15294	2037	345	528	103	18	16	163	1	7	5	1	14	11	3	NA	NA	NA	NA
3/9/2015	14863	1574	659	151	361	43	24	91	11	3	2	11	18	1	NA	NA	NA	NA	NA
3/16/2015	13708	2110	171	441	90	56	111	47	0	1	2	41	43	NA	NA	NA	NA	NA	NA
3/23/2015	13772	1372	606	76	186	73	115	0	0	0	8	39	NA	NA	NA	NA	NA	NA	NA
3/30/2015	12147	1886	275	252	95	159	62	23	2	16	47	NA	NA	NA	NA	NA	NA	NA	NA
4/6/2015	11688	1178	387	111	124	57	18	49	68	37	NA	NA	NA	NA	NA	NA	NA	NA	NA
4/13/2015	9612	1158	291	159	164	23	43	44	40	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
4/20/2015	9092	1161	208	136	6	31	30	29	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
4/27/2015	9060	993	140	82	172	39	36	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
5/4/2015	8702	924	226	213	45	35	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
5/11/2015	7558	1182	241	131	37	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
5/18/2015	7015	1311	482	46	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
5/25/2015	6934	799	92	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
6/1/2015	5802	642	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
6/8/2015	4708	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA

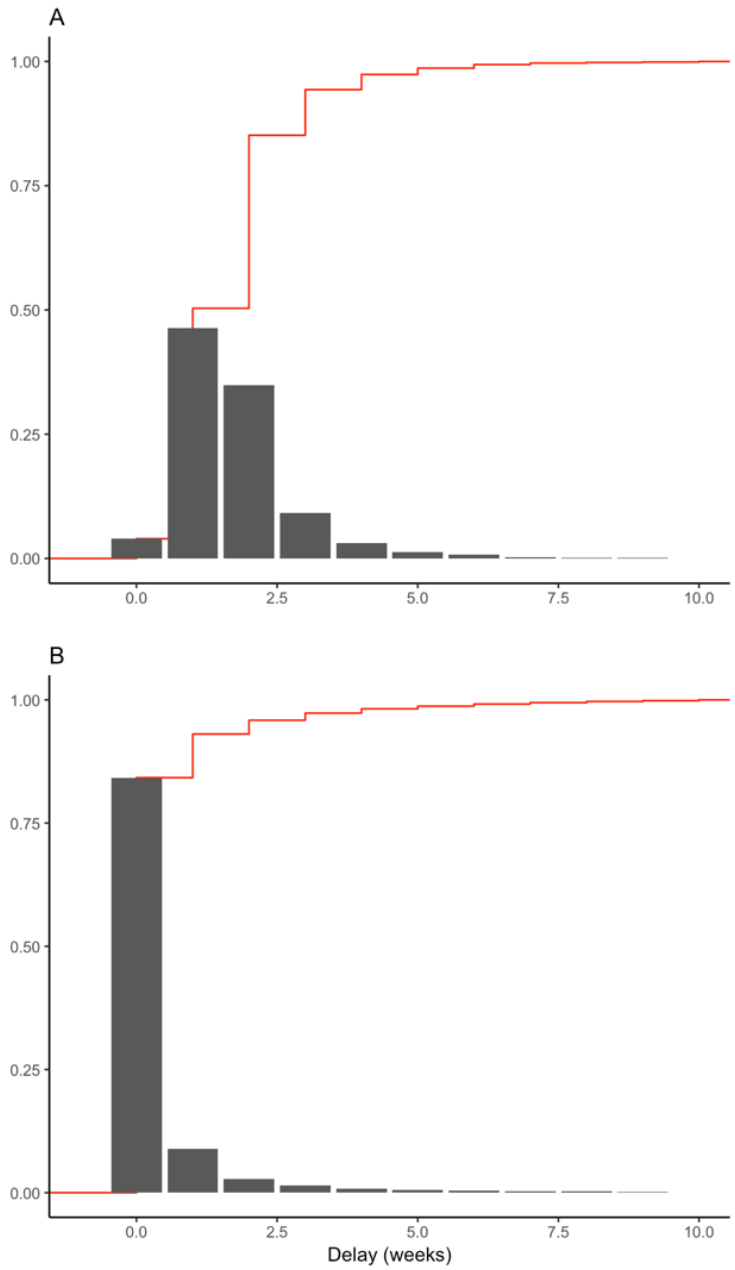

Figure C.1. The delay distribution (grey) and cumulative distribution (red), in weeks, over the full time series for (A) dengue fever and (B) influenza-like illness (ILI) cases.

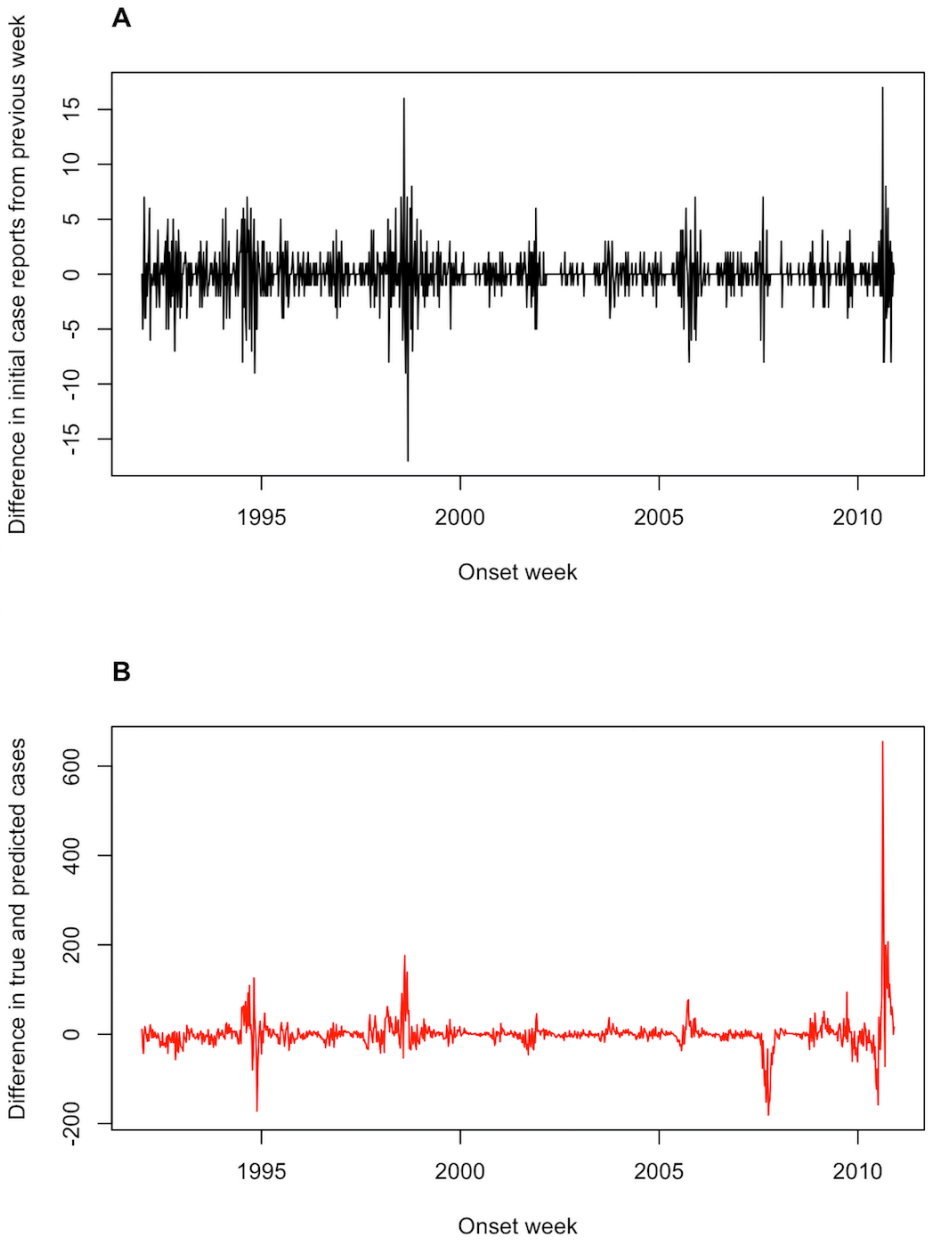

Figure C.2. Comparing (A) the change in initial case reports (from previous week) to (B) the error of NobBS for dengue fever.

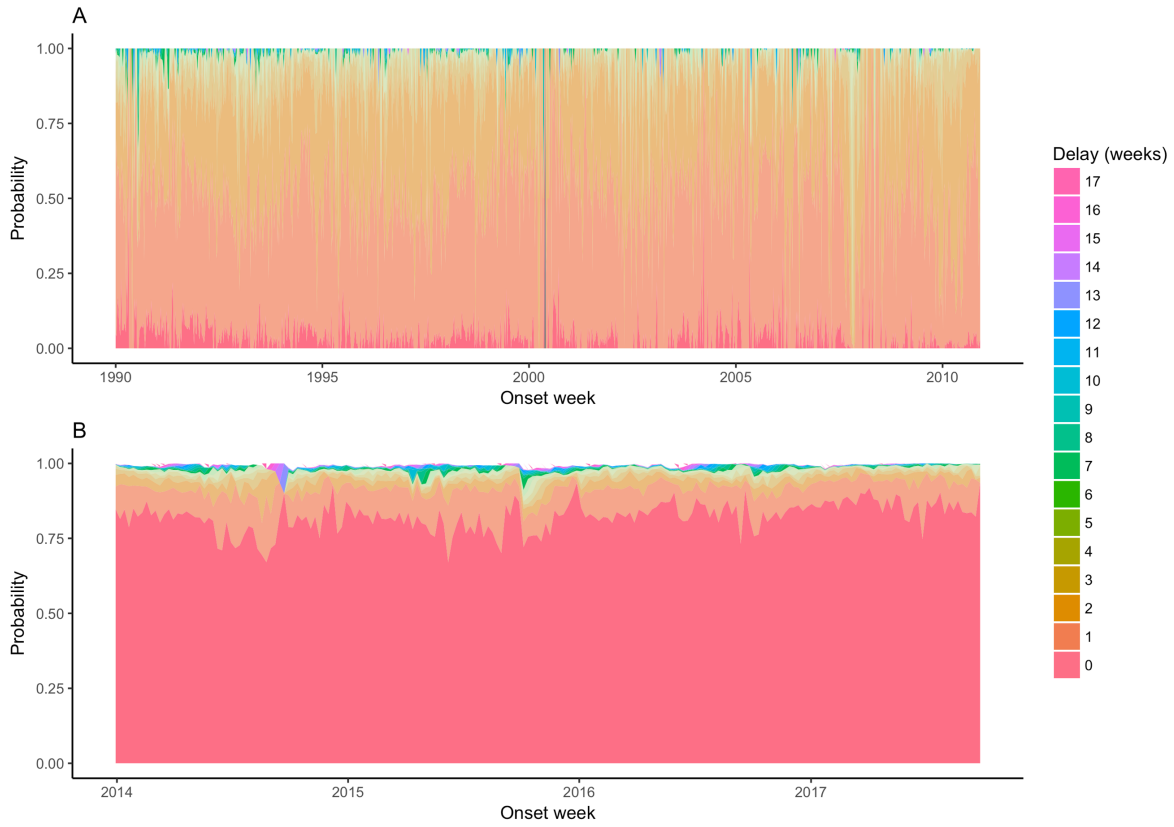

Figure C.3. Weekly reporting delay probabilities for delays up to 17 weeks for (A) dengue fever from 1990-2010 and (B) influenza-like illness from 2014-2017.

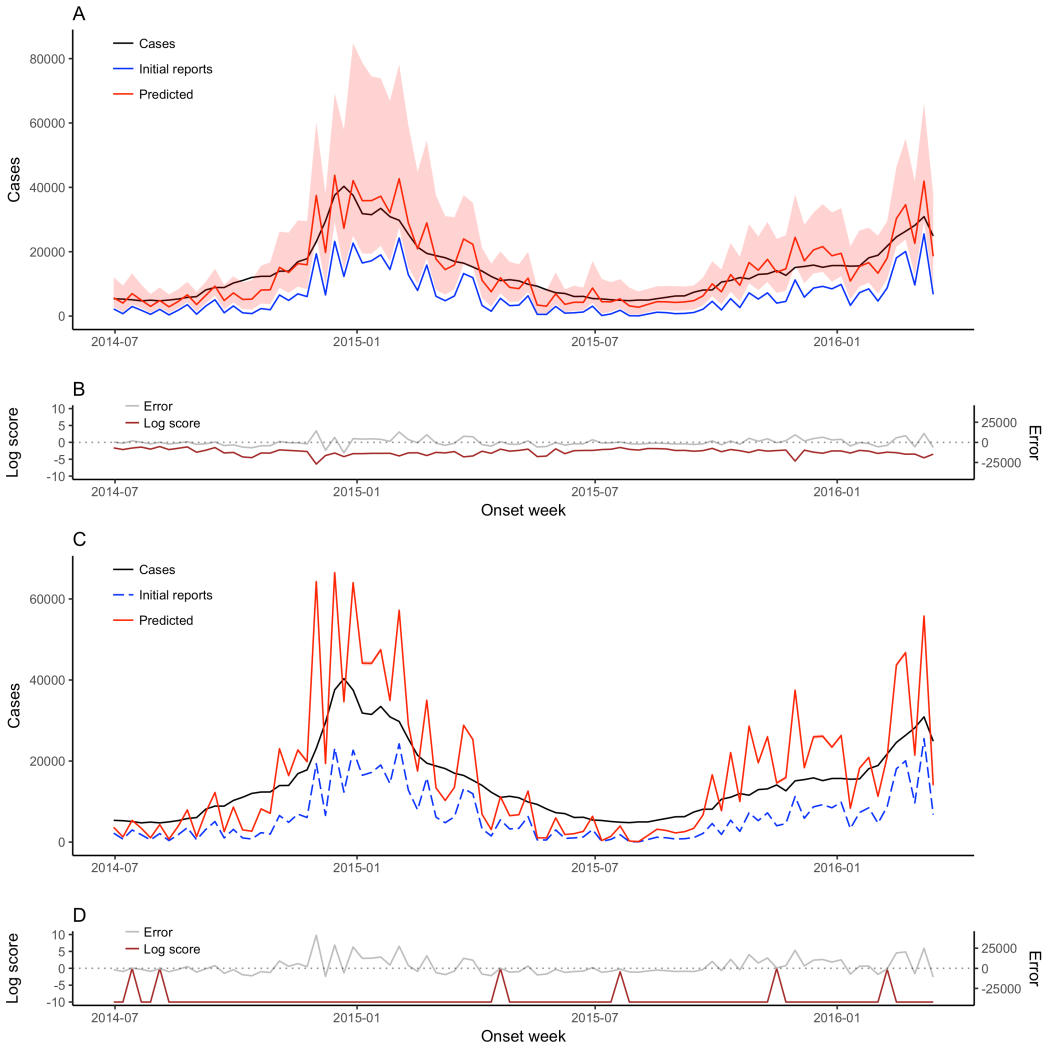

Figure C.4. Weekly ILI nowcasts for June 30, 2014 through March 14, 2016 using a non-constant (time-varying) delay distribution and 2-year moving window. (A) NobBS nowcasts along with (B) point estimate and uncertainty accuracy, as measured by the log score and the prediction error, are compared to (C) nowcasts by the benchmark approach with (D) corresponding log scores and prediction errors. For nowcasting, the number of newly-reported cases each week (blue line) are the only data available in real-time for that week, and help inform the estimate of the total number of cases that will be eventually reported (red line), shown with 95% prediction intervals (pink bands). For the benchmark approach, the 95% prediction intervals are very narrow and are thus difficult to see. The true number of cases eventually reported (black line) is known only in hindsight and is the nowcast target. The log score (brown line) and the difference between the true and mean estimated number of cases (grey line) are shown as a function of time.

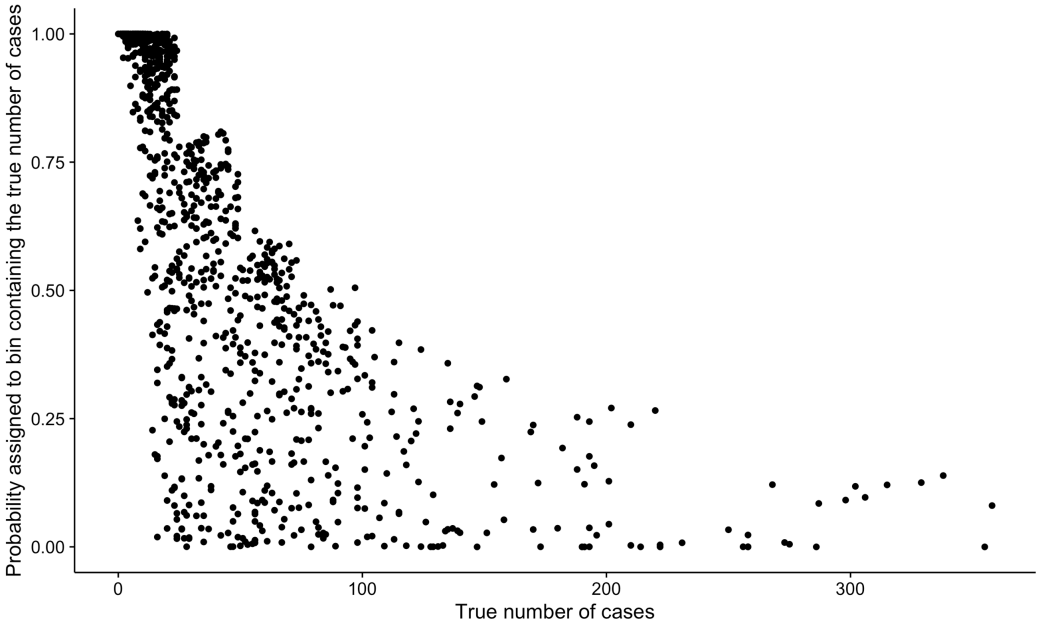

Figure C.5. Comparing the probability assigned to the bin containing the true number of cases (y-axis) to the true number of cases (x-axis), for weekly dengue fever nowcasts using NobBS.

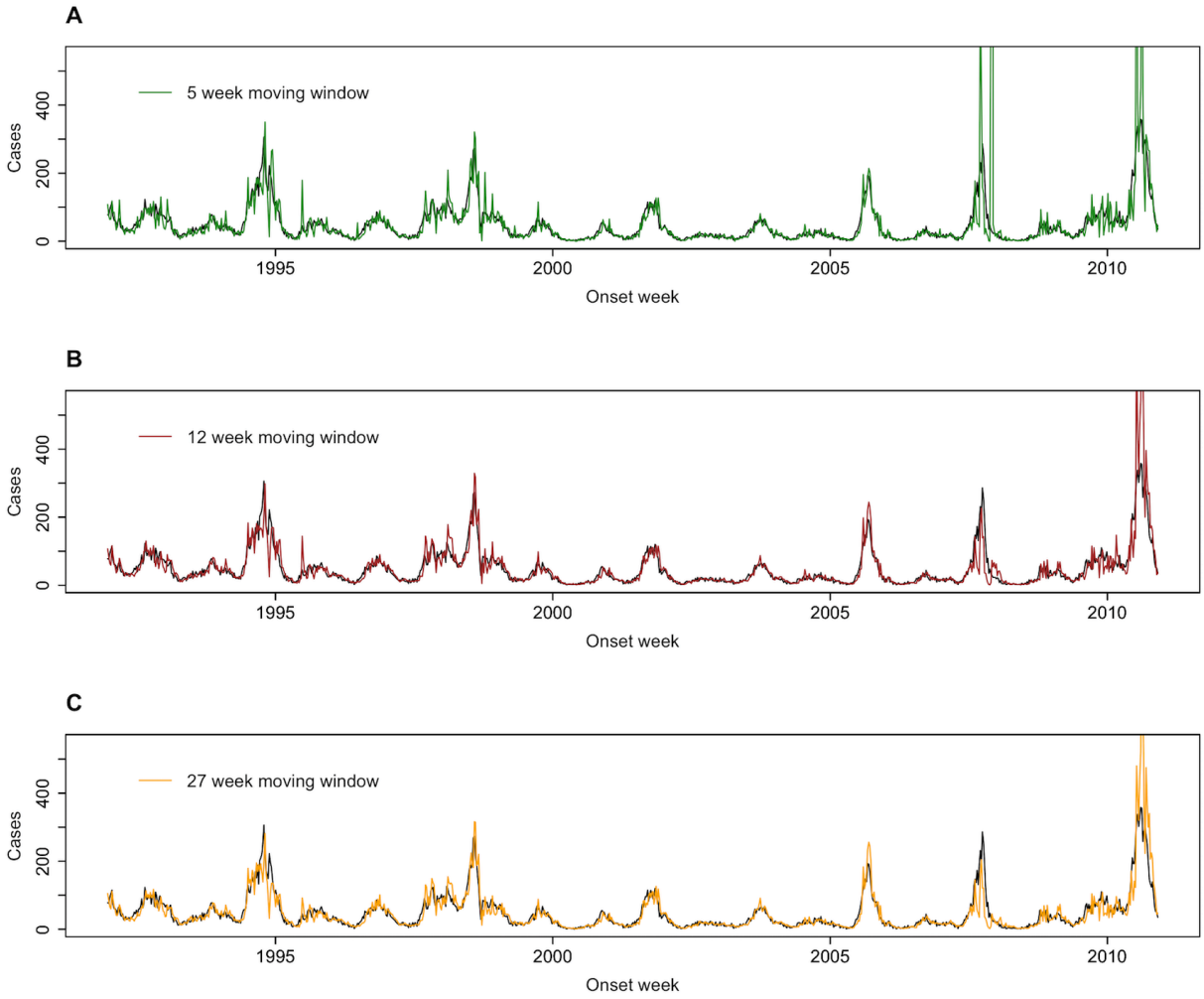

Figure C.6. Weekly NobBS dengue fever nowcasts using (A) 5-week moving window, (B) 12-week moving window, and (C) 27-week (approx. 6 month) moving window. Plots are zoomed in the y-axis to show the details of prediction.

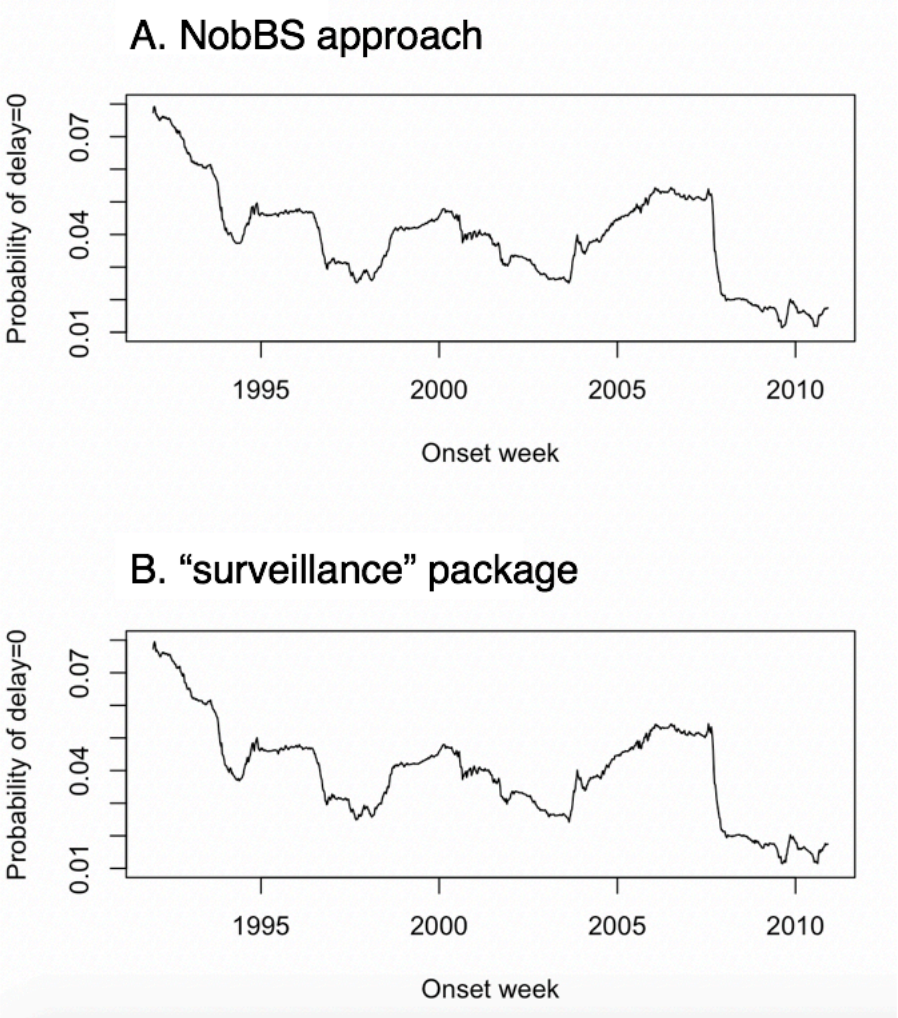

Figure C.7. Comparing the estimated reporting probability of delay 0 from (A) NobBS and (B) the nowcast model in ref. (9).