RESEARCH ARTICLE

Accurate influenza forecasts using type-specific incidence data for small geographic units

James Turtle[1]*, **Pete Riley**[1], **Michal Ben-Nun**[1], **Steven Riley**[1,2]

1 Infectious Disease Group, Predictive Science Inc., San Diego, California, United States, **2** MRC Centre for Outbreak Analysis and Modelling, Department of Infectious Disease Epidemiology, School of Public Health, Imperial College London, United Kingdom

* jturtle@predsci.com

Abstract

Influenza incidence forecasting is used to facilitate better health system planning and could potentially be used to allow at-risk individuals to modify their behavior during a severe seasonal influenza epidemic or a novel respiratory pandemic. For example, the US Centers for Disease Control and Prevention (CDC) runs an annual competition to forecast influenza-like illness (ILI) at the regional and national levels in the US, based on a standard discretized incidence scale. Here, we use a suite of forecasting models to analyze type-specific incidence at the smaller spatial scale of clusters of nearby counties. We used data from point-of-care (POC) diagnostic machines over three seasons, in 10 clusters, capturing: 57 counties; 1,061,891 total specimens; and 173,909 specimens positive for Influenza A. Total specimens were closely correlated with comparable CDC ILI data. Mechanistic models were substantially more accurate when forecasting influenza A positive POC data than total specimen POC data, especially at longer lead times. Also, models that fit subpopulations of the cluster (individual counties) separately were better able to forecast clusters than were models that directly fit to aggregated cluster data. Public health authorities may wish to consider developing forecasting pipelines for type-specific POC data in addition to ILI data. Simple mechanistic models will likely improve forecast accuracy when applied at small spatial scales to pathogen-specific data before being scaled to larger geographical units and broader syndromic data. Highly local forecasts may enable new public health messaging to encourage at-risk individuals to temporarily reduce their social mixing during seasonal peaks and guide public health intervention policy during potentially severe novel influenza pandemics.

Author summary

Forecasting influenza is a public health priority because it allows better planning by both policy makers and healthcare facilities. Most seasonal influenza forecasting is currently undertaken for large geographical areas and is based on the number of people who have the signs and symptoms of influenza. Different types of models can provide roughly

equivalent forecasts for these data: models that use the opinion of experts, statistical models, and models that reflect the underlying biology. Newly developed data streams are resolving ever smaller spatial scales, but the impacts of coupling and smaller resolution on type-specific forecasting have not been well established.

In addition to the use of local point-of-care type-specific influenza data, we assess the effects of county-to-county coupling and specific humidity on forecast outcomes. Here, we show that if the data are available for smaller geographical areas and are from actual tests for influenza, they can be forecasted accurately by models that represent the underlying biological mechanisms. Further, we add to growing evidence that mechanistic models perform better on type-specific data than they do on symptom-based data, which can be triggered by a number of respiratory viruses and/or influenza types.

Introduction

Influenza infections cause substantial morbidity and mortality across all geographical areas and sociodemographic groups [1]. Forecasting pipelines—of data and associated analytics—can create knowledge of current and likely future incidence and therefore have high potential value to individuals and health planners [2], potentially leading to at-risk individuals choosing to modify their behaviour to reduce their individual risk when incidence is high [3]. Some health officials are already working to incorporate forecasts into decision making [4], and health systems could more efficiently manage scarce resources such as intensive care services that are impacted by local influenza incidence but also utilized by many different healthcare pathways [5]. More generally, seasonal influenza is an ideal modelling case study for respiratory virus pandemics. Recent pandemics (influenza—2009, SARS-CoV-2—2020) have demonstrated the difficult health and socioeconomic decisions that must be made during a novel virus outbreak, making forecasting a key priority to support rapid investment decisions and optimization of available resources [6, 7].

Current forecasting pipelines for influenza focus on the use of syndromic data for large geographically diffuse populations [8]. For example, the US Centers for Disease Control and Prevention (CDC) forecasting challenge produces public forecasts of influenza-like illness rates at regional and national scales for the US, based on a long-running national syndromic surveillance system that has proven robust during a pandemic [9]. Influenza forecasters use a variety of methods that can be categorized by their underlying model: mechanistic, statistical, and crowd-sourced. Mechanistic models use epidemiological first principles to approximate disease transmission dynamics. In practice this often takes the form of compartmental differential equations describing a metapopulation [10–14]. Statistical models [15–19] use patterns and tendencies in the data-stream history to predict future data values. This includes many forms of implementation including: various types of regression, machine-learning, and time-series methods. Crowd-sourced models present participants with current data and ask them to manually construct a forecast. In practice, the 'crowd' generally consists of volunteer experts who update their forecasts weekly [20, 21]. Of course, complete model descriptions are much more nuanced with many mechanistic models being housed in a probabilistic framework [10–14], crowd-sourced model selection by machine learning [22], combination mechanistic and statistic models [23], ensembles of automated models [24, 25], multi-team ensembles [2], and models that incorporate social media metrics [19, 26] to name a few.

In the United States, syndromic influenza-like illness (ILI) data is available at the national, regional, and state levels [27] in a consistent and computer-readable format. At smaller spatial

scale, influenza data is often publicly available for large cities and/or counties, but in inconsistent data-types (outpatient ILI, emergency-room ILI, laboratory positive tests, etc) and often not in a computer-quantifiable format (i.e. PDF). One group has collected influenza data from the public health departments of many counties and cities in the U.S. [28], but presumably at significant cost. Possibly due to this lack of availability and coverage, there have been relatively few studies that attempt to forecast at the county [12, 28, 29] or city [26, 30–32] level. Of those, only [28] applies a model broadly without limiting focus to one or two areas per country.

Unlike existing data sources, the dataset used here provides access to spatial resolution as precise as zip code. The Quidel point of care dataset has been used for one other study [33], which focused on U.S. regional level nowcasting and forecasting. Point-of-care diagnostic results are becoming increasingly available to researchers. BioFire Diagnostics recently announced a similar dataset [34] that includes influenza sub-type results in addition to type data.

Here, we consider the possibility that alternate data streams may be much more closely related to the underlying biology and hence be inherently more forecastable using mechanistic models. Specifically, that more-local and pathogen-specific data may permit more accurate forecasting pipelines with mechanistic models than are currently possible using more geographically aggregated syndromic data [8, 35] or low resolution type-specific data cross referenced with local syndromic proxy data [28]. We apply a set of previously validated models [11] to incidence data from geolocated point-of-care diagnostic devices [36], and use a skill scoring system that is analogous to that used thus far [8, 37] for the CDC prospective competition to compare the forecastability of alternate data streams.

Materials and methods

Deployment of test machines and capture of test data

Our data are gathered from a near-real time network of benchtop Sofia and Sofia 2 POC diagnostic machines (Quidel, San Diego, USA). At the start of the study period in July 2016, the network consisted of 1,676 machines located in physician offices (880), hospitals (257), urgent care clinics (427) and other/not reported (112) settings. At the end of the study period (July 2019) the network had grown to 12,345 machines with 5,736 in physician offices, 1,206 in hospitals, 1,194 in urgent care clinics, and 4,209 in other/not reported settings. The standard kit tests for influenza A and B. Data are automatically transmitted to Quidel. Reporting cadence can vary depending on device connectivity, but most devices report results within a day of testing.

Selection of counties and clusters

Individual strain tests were sorted into weekly totals for each county represented in the dataset. Of the approximately 3,240 counties and county equivalents in the United States, 1,258 are represented in this dataset (see Fig 1A). However, to omit sparse data streams, counties that did not report for at least 40 weeks in both the 2016–17 and 2017–18 seasons were discarded. Next, counties with less than 250 specimens tested in any single season of 2016–17, 2017–18, or 2018–19 were discarded. These requirements reduced the number of considered counties from 1,258 to 136. This county filtering process was based on data reported up to and including week 3 of 2019. Throughout this manuscript, weeks are numbered according to the Morbidity and Mortality Weekly Report (MMWR) definition [38].

To identify localized groups within the 136 qualifying counties, we applied a k-means clustering algorithm for the number of clusters k. Cluster distances were based on county population centroids c (lat/lon). For a single cluster a_j with N counties and geographic cluster
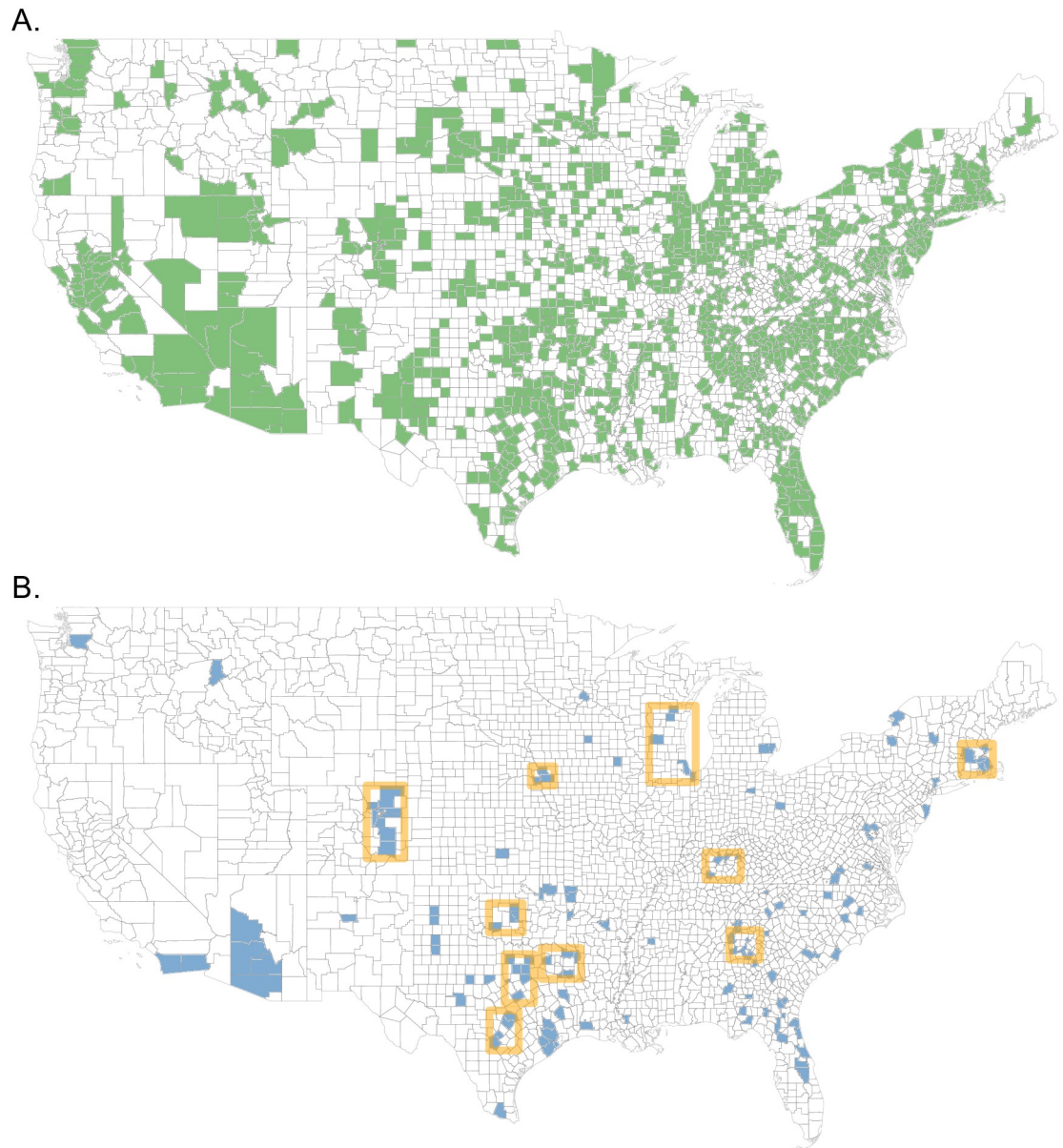
Fig 1. County filtering and clustering. (A) All counties with minimum one test result from week 27, 2016 to week 26, 2019. (B) Counties with at least 250 specimens tested for the seasons 2016–17, 2017–18, and 2018–19 and 40 weeks of non-zero reporting for the seasons 2016–17 and 2017–18 (as calculated on week 3, 2019). Square orange borders outline the 10 county clusters considered in this study. The map base layer was sourced from the U.S. Census Bureau [39]

https://doi.org/10.1371/journal.pcbi.1009230.g001

centroid b_j, we define the intra-cluster distance as the sum of squares

$$a_j = \sum_{i=1}^{N} \|c_i - b_j\|^2$$

where c_i is the centroid of the i-th county of cluster j. The k-means algorithm was applied while varying k over the interval [2, 60]. The total intra-cluster sum of squares (over all clusters) stopped decreasing at approximately $k = 38$. So we chose $k = 38$. Smaller, sparsely spaced clusters were removed by requiring $N >= 4$ and $a_j/N <= 0.5$.

The one exception to these rules is the Atlanta cluster. Most Atlanta counties do not pass the incidence minimums for the 2016–17 season. However, county coverage for 2017–18 and 2018–19 is the best of any large metropolitan area in the United States, so Atlanta was included for these seasons.

This process resulted in 10 county clusters, covering 57 counties as described in S2 Table. The 136 filtered counties and resulting county clusters are illustrated in Fig 1B.

The cutoff values for county inclusion were chosen to insure a sampling of county clusters with which to test the intra-county coupling across all three seasons. These decisions reduced the total number of counties (1258) drastically to only 136. However, it should be noted that the number of counties meeting these cutoffs (250 specimens and 40 reporting weeks per season) has increased significantly each season (see Table 1) from 136 to 541. A map of eligible counties by season appears in S1 Fig. Furthermore, the population encompassed by filtered counties increased from 76 million in 2016–17 to 181 million in 2018–19, providing representative data for over half of the country's population.

For each county of the 10 clusters, data was compiled to weekly totals for both Total Specimens (TS) and specimens that were Positive for A (PA). Weekly cluster data resulted from a sum of the constituent counties.

Mechanistic modelling framework

The following sections offer a summary of the model framework, covariate data collection, and fitting procedure used for this work. These concepts were executed using the software package Dynamics of Interacting Community Epidemics (DICE). DICE is a publicly available R-package available for download with documentation and examples [40].

The foundation of our model is the common Susceptible-Infectious-Recovered (SIR) set of equations [41]

$$\frac{\mathrm{d}S_i}{\mathrm{d}t} = -\lambda_i(t)S_i,$$

$$\frac{\mathrm{d}I_i}{\mathrm{d}t} = \lambda_i(t)S_i - \frac{I_i}{T_g},$$

$$\frac{\mathrm{d}R_i}{\mathrm{d}t} = \frac{I_i}{T_g}.$$

Here the state variables are number of susceptibles S, infectious I, and recovered R in the ith physical region. The equations are completed by generation time T_g and force of infection λ. The model is applied to a single season with the assumption that the entire population is susceptible at start of season. Included in λ are terms to modulate the force of infection for specific

Table 1. Filtered county totals by season.

Season	No. counties	Population (millions) encompassed
2016–17	136	76
2017–18	281	109
2018–19	541	181

Total number of counties with at least 250 specimens tested and 40 weeks reported in a season. The right column is the total population of the filtered counties. These totals were calculated following the end of the 2018–19 season.

https://doi.org/10.1371/journal.pcbi.1009230.t001

humidity and county coupling

$$\lambda_i(t) = \sum_{j=1}^{D} \beta_j(t) m_{ij} \frac{\sum_{l=1}^{D} m_{lj} I_l}{\sum_{p=1}^{D} m_{pj} N_p}.$$

The transmitting-contact rate $\beta_j(t)$ pertains to transmission dynamics internal to county j, where D is the total number of coupled counties. The mobility matrix m_{ij} describes the probability that a contact made by an individual in i was with an individual from j [42]. The probability of inter-county contacts is proportionate to the mobility kernel

$$\kappa(r_{ij}) = \frac{1}{1 + (r_{ij}/s_d)^{\gamma}}$$

that is a variation of the offset power function. Here, the distance r_{ij} is calculated as the great-circle distance between counties' population centroids. Saturation distance s_d and the exponent γ are fit parameters. The mobility matrix is population weighted and row-normalized

$$m_{ij} = N_j \kappa(r_{ij}) \frac{1}{\sum_k N_k \kappa(r_{ik})}.$$

Following the functional form in [43], the transmitting-contact rate in region i becomes time-dependent when modulated by specific humidity $q_i(t)$

$$\beta_i(t) = \beta_{0i}(1 + \Delta_{Ri} \cdot e^{-a_i \cdot q_i(t)}).$$

The basic reproduction number $R_0 = T_g \beta_{0i}$, specific humidity scaling Δ_{Ri}, and specific humidity power a_i are fit parameters. When $\Delta_{Ri} \to 0$, the transmitting-contact rate reverts to a constant value β_{0i}.

All results presented here pertain to forecasting cluster weekly incidence, where cluster incidence is the sum of the counties incidence. The spatial aspect of the model is characterized in three ways: direct, uncoupled, and coupled. The self-describing 'direct' fit matches the model directly to cluster totals. When applying the uncoupled model to a cluster, each constituent county is fit independently. The individual county forecasts are then aggregated to produce a cluster forecast. The coupled approach simultaneously fits the coupled model to all constituent counties.

Specific humidity data was taken from the National Centers for Environmental Prediction (NCEP)/National Center for Atmospheric Research (NCAR) Reanalysis-2 project [44]. This provides climate model evaluations 1979-present on a 1.875˚ spatial grid. The reanalysis-2 spatial grid is layered with GADM (the Database of Global Administrative Areas) county definitions [45] and the SocioEconomic Data and Applications Center (SEDAC) population densities [46] to produce weekly population-weighted averages for specific humidity in each county.

Model fitting

Here we define incidence as the population count that was ILI symptomatic (TS) or influenza infectious (PA) and presented to a reporting clinic for testing. Model equations are applied to the full population, however model incidence-reported IRm is returned from the model as a

proportion of weekly new infectious as

$$\text{IRm} = B + p_C \int_{t_{i-1}}^{t_i} \lambda(t)S(t)dt.$$

The proportion clinical p_C scaling accounts for both the portion of the population covered by reporting clinics and the proportion of infectious that sought medical evaluation/treatment. In this framework t_i is defined as the Saturday that ends week i, and t_{i-1} is the previous Saturday. A baseline incidence B is introduced as a phenomenological parameter that accounts for year-round testing and, particularly, early-season weeks where the incidence trend is not yet exponential. B and p_C are both fit parameters.

In the fitting procedure, the objective function is a Poisson based, Log-Likelihood (LLK). For the i-th week of data, we calculate the probability that the data incidence reported IRd_i is a Poisson expression of model incidence reported IRm_i

$$P(x = \text{IRd}_i) = \text{Pois}(\text{IRm}_i).$$

The model likelihood is then a product over N weeks of $P(x = \text{IRd}_i)$, and results in an LLK of

$$\text{LLK} = \sum_i^N w_i(\text{IRd}_i \cdot \log(\text{IRm}_i) - \text{IRm}_i - \log(\text{IRd}_i)).$$

Here, w_i is used to give zero weight to weeks where the data has been designated 'no data'. The following is a summary description of how the spatial aspect of the model effects the fitting procedure. For the direct model, a single LLK fits the model to cluster totals. For the uncoupled model, we independently fit the model to each county and then aggregate model results to cluster forecasts. For the coupled model, cluster LLK is calculated as a population-weighted average of the individual county LLKs, and thus the counties are fit simultaneously in a single fit procedure.

The fit procedure is a Metropolis-Hastings type Markov Chain Monte Carlo (MCMC) process [47] with adaptive step size to map the likelihood minima in parameter space. This procedure was implemented in Fortran and is accessed through the R wrapper, all of which is included in the linked GitHub repository. For each MCMC fit, we processed three randomly-initialized chains, each for 2×10^6 steps, and returned 10^4 parameter sets (evenly distributed from the second half of the chain). Of the three chains, we chose the one with the lowest LLK minimum. By using 3 chains starting in different parts of parameter space, we avoid the rare event where a chain gets 'stuck' in a local minima of the objective function. Due to the lack of sufficient historic data and resulting model-fit posteriors, we do not attempt to use informed parameter priors. Therefore, all parameter priors are uniform.

For each of the three seasons (2016–17, 2017–18, 2018–19) we fit the data from week 27 to x, where x covered the weeks [36, ..., 52, 1, ..., 18], and projected the model forward 1 to 10 weeks. This resulted in 1- to 10-week forecasts for each of the evaluated weeks [46, ..., 52, 1, ..., 19]. At each week we applied 6 model variants stemming from combinations of spatial organization (direct, uncoupled, coupled) and β-modulation (fixed β, humidity-modulated β).

Forecast scoring

For each forecast target, our model produces a distribution of predicted outcomes. This distribution is binned for evaluation based on the scoring system used by the CDC Influenza Challenge for the 2016–17, 2017–18 and 2018–19 seasons [37]. The maximum range is separated

into 130 equally sized bins with a 131st bin for all values over the maximum previously observed. In this scoring system, higher magnitude seasons tend to score lower. As a simple correction for this, we re-scaled the bins for TS and PA so that both metrics peak in the same bin as CDC %ILI for the encompassing region (see 'CDC_Parent' in S2 Table). Although each cluster was generally much smaller than its encompassing CDC state/region, we suggest that CDC %ILI is the most spatially consistent and comprehensive dataset to perform this scaling with. Re-scaled bins for the San Antonio/Austin cluster, 2016–17 season appear as an example in S2 Fig. Forecast probabilities for the observed bin (red shading) as well as the 5 pre- and pro-ceeding bins (yellow shading) are summed to produce forecast skill. Forecast score is the natural log of skill, with scores below -10 set to -10. For example, if we assigned a total probability of 20% to bins in the skill window, then the score would be ln(0.2) = −1.6. Note that due to the work of Bracher [48] and the response [49], starting with the 2019–20 influenza season, the CDC Influenza challenge defines forecast skill as only the probability in the observed bin. Here we primarily show results from the 11-bin skill window for consistency with matching Influenza Challenge seasons. However, some single bin scores are included in the supplementary material for comparison.

Null models

For a given season, cluster, week, and data metric, Null models were evaluated using data from the same cluster, metric, and week, but different seasons. The most basic Null model 'N.pt' is a point forecast calculated as the mean of the Null data. The 'N.d' model is a distribution fit to the same data. For intensity targets we use a log normal distribution and for timing targets a normal distribution. However, both of these models suffer from the same problem: due to the small number of seasons in the dataset, the model is typically fit to only 2 data points. The third Null model 'N.d5' was designed to mitigate this problem. For a given forecast week, 'N.d5' uses the historic data from a 5-week window including the forecast week as well as the 2 pre- and pro-ceeding weeks to fit a log normal distribution. As an example, for epidemic week 4 of 2017 the data from epidemic weeks 2, 3, 4, 5, and 6 for years 2018 and 2019 are fit to a log normal. This distribution is then intensity-binned to serve as the N.d5 model forecast.

Results

We obtained 3,258,166 specimen results from 12,227 point-of-care (POC) machines in the United States starting from week 27, 2016 through week 26, 2019. These data were down-selected to obtain clusters of nearby counties with at least 250 tested specimens in any season, resulting in 10 clusters, capturing: 57 counties (S2 Table); 1,061,891 total specimens; and 173,909 specimens positive for Influenza A (Fig 2). Where no data was reported and there was no temporal overlap of devices, we designated that week as 'no data' rather than a zero. A cluster containing metropolitan Atlanta was just below the average season requirements for 2016–17 but had very high coverage for 2017–18 and 2018–19 and was thus included.

Total specimens (TS) and specimens positive for influenza A (PA) both showed broadly similar dynamics, however, there were some intriguing differences between them (Fig 2). As would be expected, the more specific PA data had a clearer difference between off-season and in-season. Also, regardless of noise during periods of low activity, the county-level PA time series produced sharper epidemics, with 6.05 weeks average duration at or above half maximum incidence compared with 8.84 weeks for the TS data. For clusters, the peaks became slightly wider with a width at half peak for PA averaging 6.48 weeks and TS averaging 9.57 weeks. The TS data showed a high correlation with influenza-like illness data for the enclosing
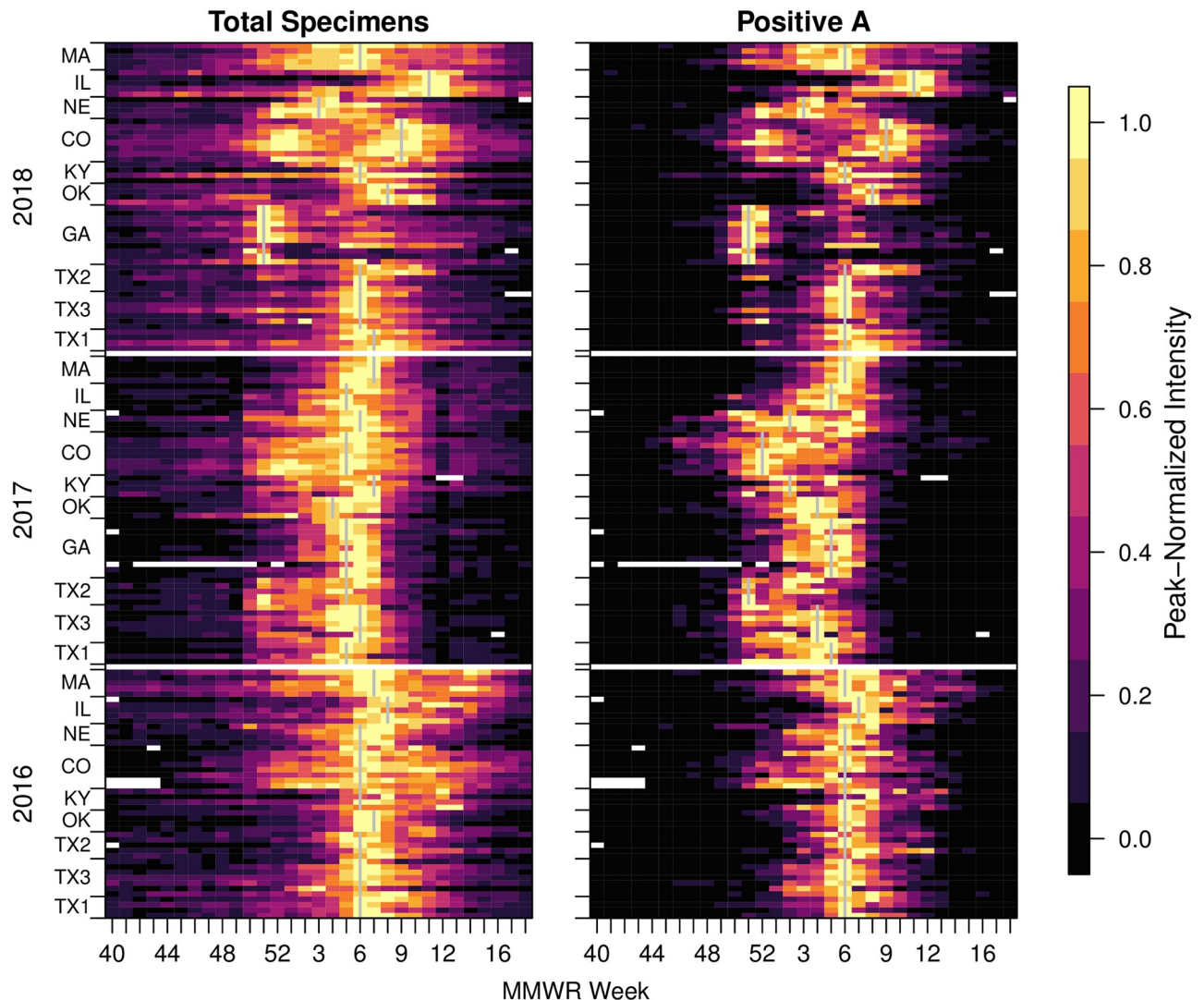
Fig 2. Incidence for each cluster for syndromic (total specimens, LHS) and specimens positive for A (RHS). Each row of rectangles represents a single county with counties grouped into clusters (bounded by tick marks) and then into seasons. Seasons are indicated by start year on the y-axis (with a blank row between years). Within seasons, clusters are ordered by latitude and within clusters, counties are also ordered by latitude. Cluster peak weeks are marked with a gray vertical line. In this figure, the incidence for each week is normalized for each county and year using the peak value for that year (Legend, RHS). White indicates 'no data'.

https://doi.org/10.1371/journal.pcbi.1009230.g002

CDC region (overall average Pearson coefficient 0.96, S1 Table). Additionally, S2 Fig serves as an example illustration of this similarity.

When averaging across all clusters and seasons, at each forecast week; we found substantial evidence that PA data could be more accurately forecast than could TS data (Fig 3 and S3–S5 Figs). We assessed accuracy using a similar scoring system as that used for influenza by the US CDC [8] during the same seasons. The incidence scale was separated into 131 bins and we used our model to assign a probability to every possible category. In the 2019–20 season, the CDC influenza competition moved to a single-bin scoring system. For completeness, we have reproduced S3 and S4 Figs with single-bin scoring as S6 and S7 Figs respectively. The best model for 2-week ahead forecasts scored an average of -1.41 for PA data compared with -2.55 for TS data. Although the accuracy of the forecast did vary over the season, the superiority of
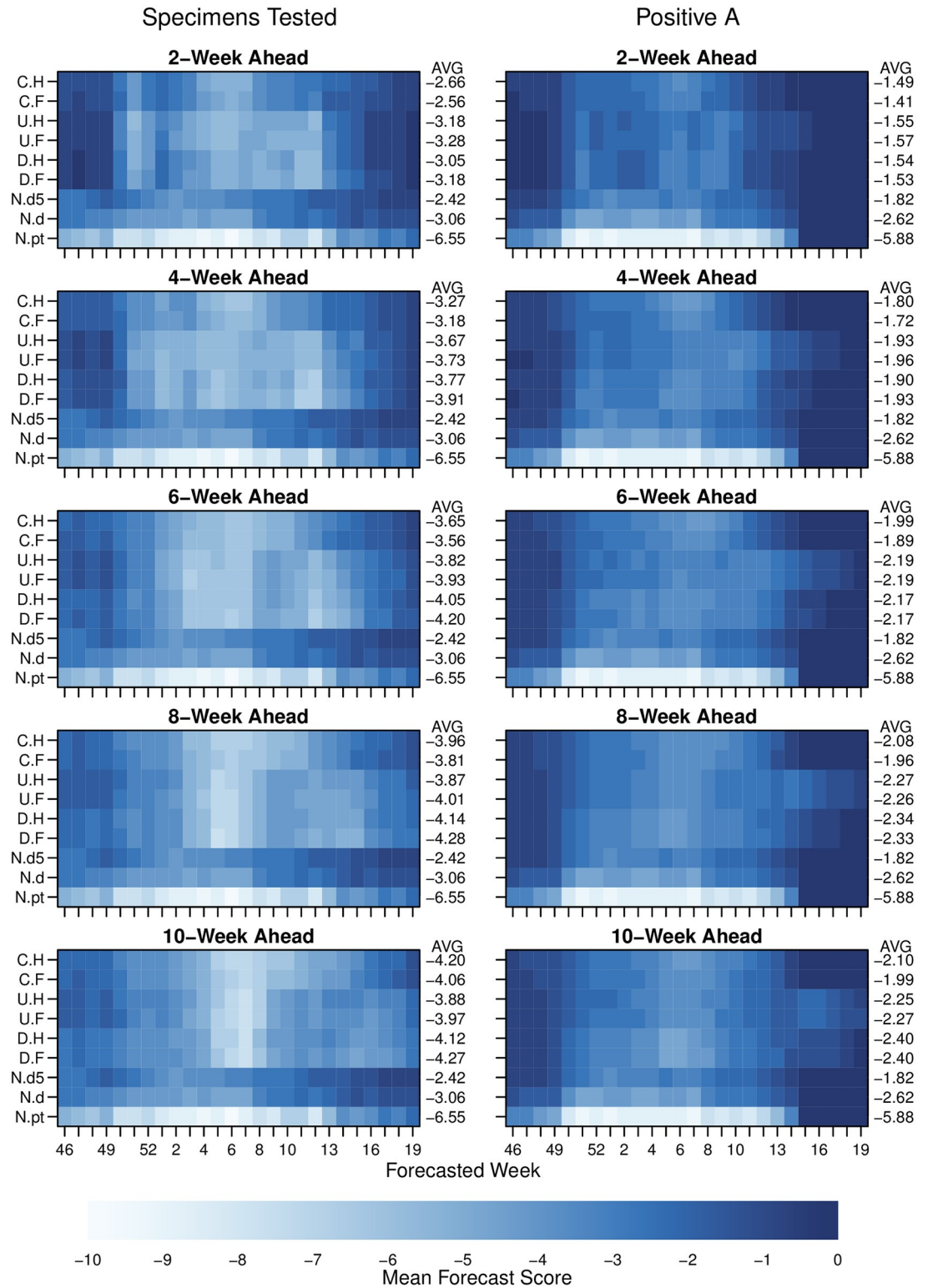
Fig 3. Forecast scores for 2- to 10-weeks ahead lead times, and comparing different data metrics. Pixel colour shows forecast score (see main text) for a given observation week averaged across clusters and seasons, e.g. the 4-week ahead forecast for week 48 was made using data only up to week 44. Averages across all weeks for a given model are printed on the RHS of each row of pixels. Model type is shown on LHS y-axis tick labels: C.H, coupled model with humidity modulated contact rate; C.F, coupled model with fixed contact rate; U.H, uncoupled with humidity; U.F, uncoupled with fixed contact rate; D.H, model directly fitted to cluster with humidity term; D.F, model directly fitted to cluster with fixed contact rate; N.pt, null model made from simple

model of that week for all other seasons; N.d, null model made from fitting a log normal to all observations for that week from other years; and N.d5, null model made from fitting log-normal to the observation week, two weeks prior and two weeks following for all other years (see main text). Models are ordered approximately from least complex on the bottom rows to most complex on the top row. See S2 and S3 Figs for complete 1–10 weeks ahead forecast results.

https://doi.org/10.1371/journal.pcbi.1009230.g003

best fits to PA data over TS data was maintained across the whole season. In fact, the accuracy of the best fit model for PA data with a 10-week lead time was still better than that of the TS data at a 2-week lead time.

We considered specific examples to understand how and why the models were better able to fit these data in different scenarios. For example, after testing all forecast models at all forecast times for the Colorado cluster for 2016–17, the best fitting model was better able to capture the rapid acceleration of incidence and the shape of the peak PA data stream than was possible for the best fitting model for the TS data (Fig 4A and 4B). Despite not being able to accurately fit the decay phase of the epidemic, the best fitting model for the PA data still achieved a substantially better score of -0.51 than the score of -0.82 that was possible for the corresponding TS data. We note that a score of -0.51 corresponds to 60% of forecasts falling within the skill window. However, for the same cluster for the following year, the same did not apply. The best model fits to both datasets were substantially worse than the year before, largely due to the cluster exhibiting a double peak (Fig 4C and 4D). The average single-peak line through the slightly diffuse TS data performed slightly better than did a similar line through the more sharply peaked PA data. We note that our data and models do not distinguish between subtypes of influenza A and that both H1N1 and H3N2 were circulating during similar periods in this season. Similar panels are available for all models and seasons in S12 Fig. Looking at the seasons more broadly, S8 Fig contains scores averaged over cluster and 1- to 4-weeks-ahead forecasts to show model performance by season and data type. Here we see relatively high scores for the 2016–17 season, likely a result of semi-synchronized H3N2 and B peaks. The 2017–18 season was characterized by an early H3N2 peak followed by a late B peak. Thus, the model performed well on PA data, but struggled with the broad peaks in TS data. A--very unusual--early H1N1 peak followed by late H3N2 peak caused 2018–19 to score the worst for both TS and PA data.

Among the mechanistic models, coupled models generally performed better than uncoupled models or models that fit the cluster incidence directly (Fig 3). Also, the accuracy of mechanistic models relative to non-mechanistic average-based models compared for TS and PA data suggests that mechanistic models are able to capture more information from the PA data than a simple baseline (null) model. For the TS data, the 5-week rolling historical average model outperformed the best mechanistic models for all time scales from 2- to 10-weeks. However, for PA data, the best mechanistic model outperformed the 5-week average historical model in the 1- to 4-week ahead range. We note that there were few years available to the historical models in these data, therefore the performance of historical-average models would likely improve over time as the training data set grows.

Patterns in the posterior densities for key mechanistic parameters also suggest that when the models are fit to the PA data they are capturing more of the underlying biology (Fig 5). The posterior median for the basic reproductive number R_0 was consistent from year to year and location to location when fit to the PA data. For coupled models, cluster R_0s were higher (mean 1.32; 95% CI [1.31, 1.33]) than for uncoupled models (mean 1.18; 95% CI [1.17, 1.18]). Given that transmissibility is reflected by $R_0 - 1$, this difference is significant and likely reflects the need for uncoupled models to have longer durations (resulting from lower R_0) because the overall cluster incidence was not driven by epidemics in subpopulations taking off at different times. This theory is supported by an inverse relationship for the estimated proportion of
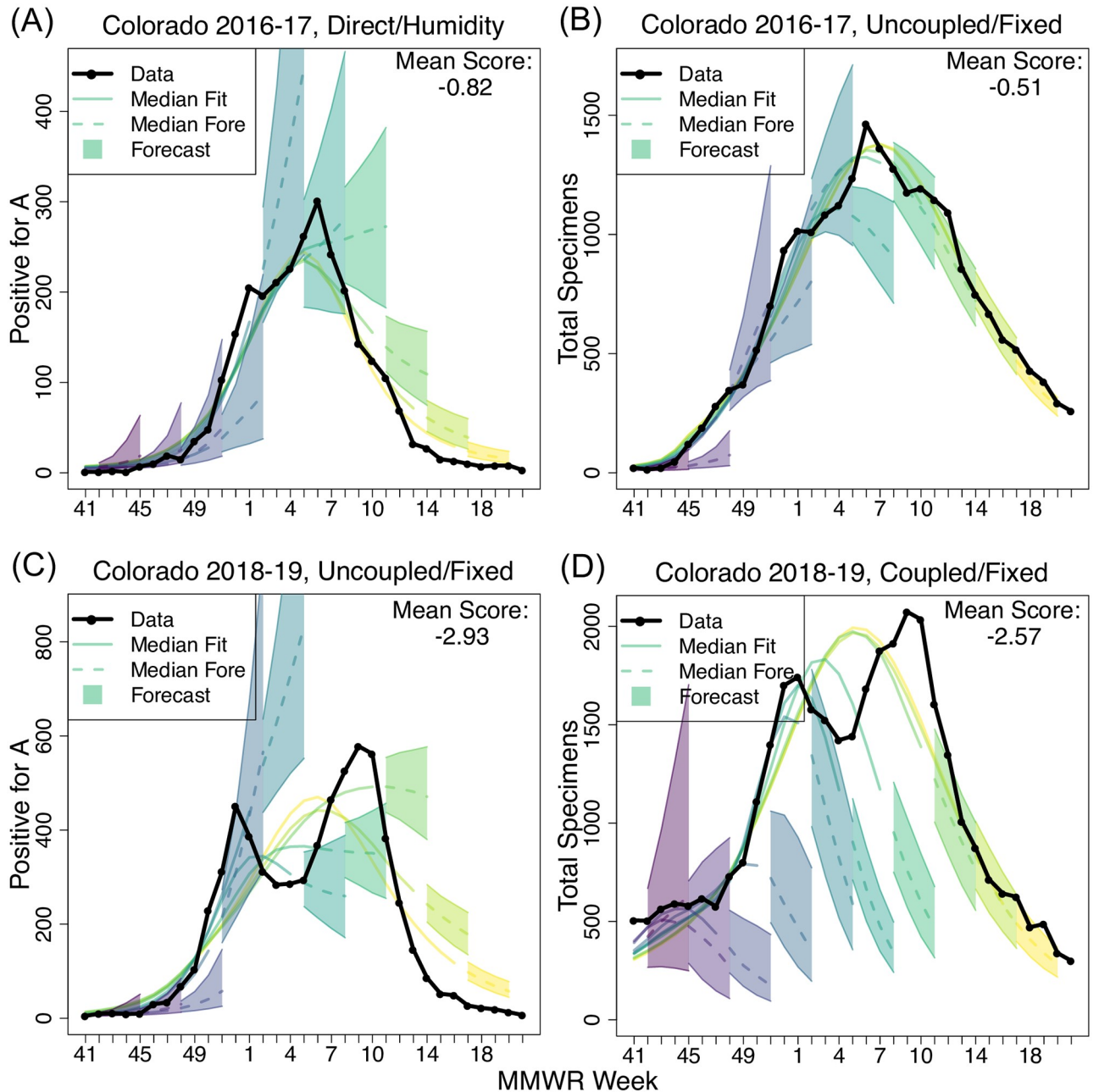
Fig 4. Illustrative high-scoring and lower-scoring forecasts for the Colorado cluster. Here the incidence data is shown in black and shaded cones depict the central 50% forecast windows for 1–4 week ahead. Forecasts were generated for every week, but only every third forecast is shown here for clarity. Colour progression from left to right indicates time of forecast. Solid lines show fit to data at the point the forecast was assumed to have been made. Dashed lines show forecasts beyond the time at which data were assumed to be available. Mean score is the average score across MMWR forecasted weeks and 1 to 4 weeks ahead targets.

https://doi.org/10.1371/journal.pcbi.1009230.g004

infections which result in clinical cases p_C. Estimates of p_C were lower for coupled models than for uncoupled models because subpopulations with higher R_0s require fewer of their infections to result in cases to achieve the same level of incidence as those from the uncoupled model. Despite the coupled model outperforming the uncoupled model, estimates for the parameters that determined coupling strength were not well constrained when fit to the data (S9 Fig).
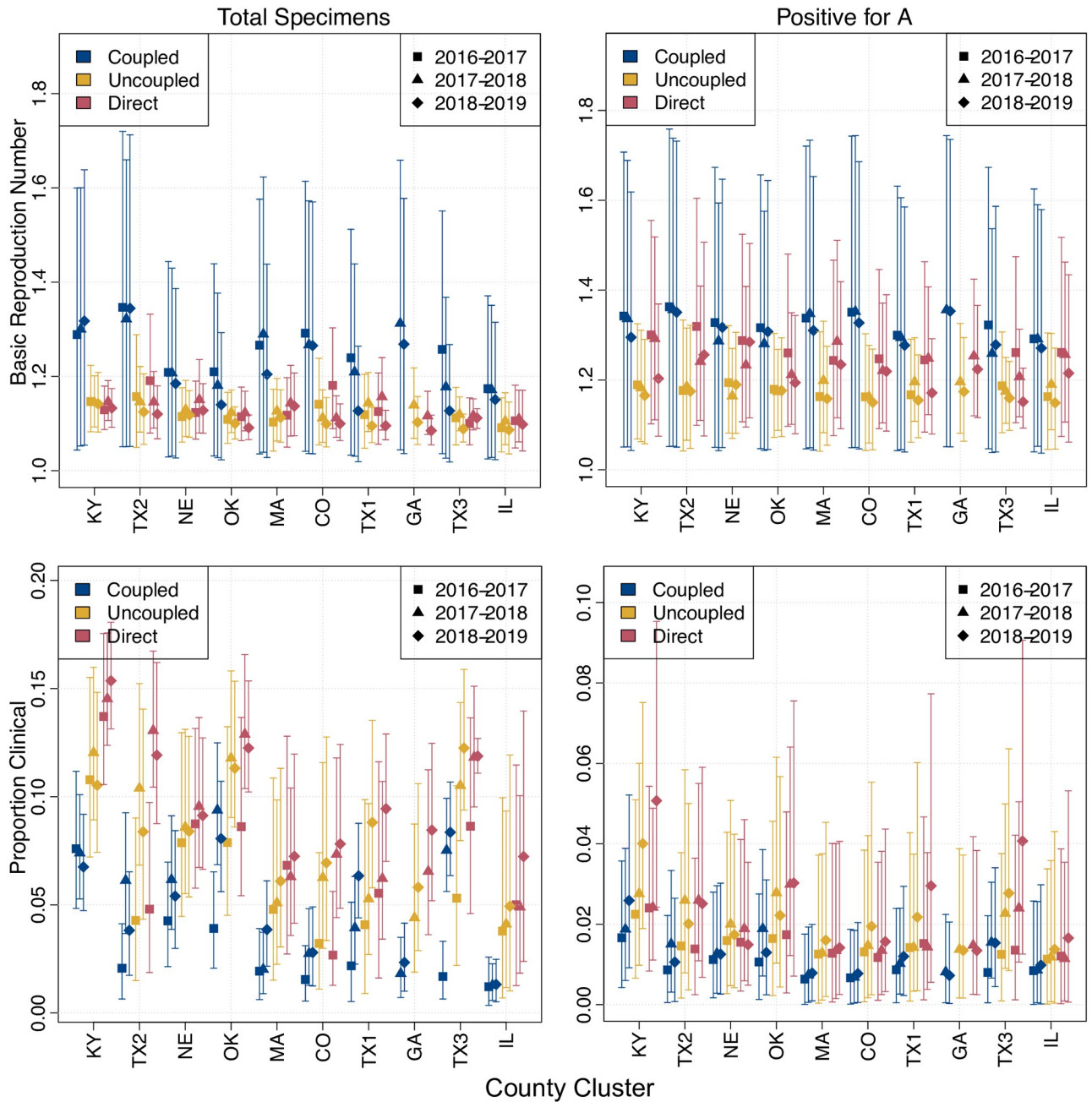
Fig 5. Posterior parameter densities by season, cluster and model type. Bars show central 95% of posterior credibility interval with marker showing posterior mean. Colour indicates model type and symbol shape indicates season. Values for each cluster are grouped together as per x-axis tick labels.

https://doi.org/10.1371/journal.pcbi.1009230.g005

Although coupling posteriors showed large spreads, there were subtle differences from cluster to cluster and between the data metrics. Plotting the posterior coupling kernel κ as a function of distance (see S10 and S11 Figs) shows differing patterns of coupling fall-off.

Discussion

We have shown for a large set of mechanistic models, using a skill scoring system analogous to current best practise, that cluster-level influenza incidence-data based on type-confirmed test

results generate substantially better forecasts than those based only on the volume of respiratory samples being tested. We have also shown a close correlation between the weekly total number of respiratory samples in each cluster and the pattern of influenza-like illness reported to the national surveillance system for the geographical region enclosing each cluster. Among these mechanistic models, cluster forecasts were more accurate in those that fit to subpopulation county data than those that do not. Also, the relative difference between the mechanistic models and simple historical average models is much greater when forecasting type-specific data than when forecasting total samples.

A limitation of our study is that we compared only forecasts from our suite of mechanistic and historical average models against the two datasets. Therefore, it is possible that other approaches may have been able to forecast the TS data more effectively than the PA data. However, the difference in our ability to forecast the two types of data was considerable and significantly greater than the difference in performance between our approach and other high performing participants in the CDC challenge study, which is measured on a similar scale [8]. Also, intuitively, the shape of epidemics in the PA data was more consistent with a short-generation-time infectious disease epidemic: the degree of noise prior to the start of the season was low and the epidemic itself was more peaked, as measured by width at half maximum, suggesting that the substantial improvement in forecastability using a mechanistic approach is consistent with the PA data being much more representative of the underlying biological process.

A previous study compared the forecastability of pathogen-specific data and found evidence of slightly increased forecastability of local type- and subtype-specific data, albeit using different accuracy metrics [28]. However, the virological data available to the prior study represented considerably larger spatial units than the separate ILI data. More local sub-type epidemics were inferred by combining regional laboratory-confirmed positive proportions with Google Flu Trends [50] estimates of state and municipality ILI data, implicitly assuming that the temporal distribution of types and subtypes was uniform for large geographical regions. A similar study [29] looked at influenza positive laboratory data at small spatial scale combined with several other data types, but did not address subpopulation coupling or type-specific data. We suggest that aggregated positive tests at the geographical and temporal scales of analysis contain maximal information that is not present in similar inferred data streams.

Our results suggest a number of additional refinements to our forecasting workflow that may further improve model forecast accuracy. We assumed that each cluster consisted only of the sub-populations for which data were available. It may be that the inclusion of the intervening and surrounding populations that were not observed could increase accuracy. This would substantially increase the computational complexity of the approach, but could potentially fill-in surveillance gaps at the same time as increasing the accuracy and robustness of the forecasts. Additionally, tightly clustered counties with low reporting could be combined into a single super-county. This would allow us to include incidence data that is currently being lost in the county filtering process. A separate issue is that we have treated these as weekly data, which reflects their current reporting pattern. However, accuracy for now-casting and short term forecasting could be improved substantially by looking at daily incidence which are available from this surveillance network. Although some work would be required to make a day-of-week adjustment [51], accuracy of 1-week ahead and 2-week ahead forecasts would likely increase substantially over and above the gains described here.

Public health authorities may wish to consider the routine use of type- and subtype-specific point-of-care data for influenza surveillance in addition to traditional ILI surveillance to enable more accurate local influenza forecasts during seasonal and pandemic periods. Our results were for 10 geographical clusters based only on three seasons of data (so a maximum of

two training years). Very soon, the network of machines from which these data were obtained and other similar networks will have much more complete geographical and temporal coverage [52]. Similar networks are being established in many other populations.

Other studies have shown that aggregating local forecasts can improve accuracy over direct forecasts for larger geographical units, but have not focused on the potential accuracy at small spatial scales [11, 53]. Also, more descriptive ecological analyses have demonstrated the potential for improved local forecasts by characterizing how humidity and population density are correlated with average properties of epidemic curves [54]. Building on these prior studies, the smaller spatial unit and temporal accuracy of the results presented here could be used to guide public health interventions and give individuals the opportunity to modify their behaviour over short periods of time, based on very local information. In particular, people at high risk of severe disease may choose to reduce their social mixing during short periods where the local risk of infection can be accurately predicted to be high.

Supporting information

S1 Fig. County filtering by season. Counties with at least 250 specimens and 40 reporting weeks per season. (A) 2016–17 season—187 counties, (B) 2017–18 season—281 counties, and (C) 2018–19 season—541 counties. The map base layer was sourced from the U.S. Census Bureau [39]
(PDF)

S2 Fig. Peak-normalization of scoring bins across data metrics, illustrated for 2016–2017 season in the San Antonio / Austin area cluster. Bins were scaled to each metric such that all peaks fall in the same bin. The skillful bins for the peak week are shown in yellow shading. Forecast probabilities in these bins, in addition to the observed bin (red shading), are summed to calculate skill. The Centers for Disease Control and Prevention percentage ILI (CDC %ILI) is for the enclosing CDC region (state of Texas).
(PDF)

S3 Fig. Mean scores for total specimens forecasts 1–10 weeks ahead. Pixel colour shows forecast score (see main text) for a given observation week averaged across clusters and seasons, e.g. the 4-week ahead forecast for week 48 was made using data only up to week 44. Averages across all weeks for a given model are printed on the RHS of each row of pixels. Model type is shown on LHS y-axis tick labels: C.H, coupled model with humidity modulated contact rate; C.F, coupled model with fixed contact rate; U.H, uncoupled with humidity; U.F, uncoupled with fixed contact rate; D.H, model directly fitted to cluster with humidity term; D.F, model directly fitted to cluster with fixed contact rate; N.pt, null model made from simple model of that week for all other seasons; N.d, null model made from fitting a log normal to all observations for that week from other years; and N.d5, null model made from fitting log-normal to the observation week, two weeks prior and two weeks following for all other years (see main text). Models are ordered approximately from least complex on the bottom rows to most complex on the top row.
(PDF)

S4 Fig. Mean scores for positive-for-A forecasts 1–10 weeks ahead. Pixel colour shows forecast score (see main text) for a given observation week averaged across clusters and seasons, e.g. the 4-week ahead forecast for week 48 was made using data only up to week 44. Averages across all weeks for a given model are printed on the RHS of each row of pixels. Model type is shown on LHS y-axis. Models are ordered approximately from least complex on the bottom

rows to most complex on the top row.
(PDF)

S5 Fig. Peak target mean forecast scores for total specimens and positive for A. The panels are divided by target (top—Peak Week, bottom—Peak Intensity) and data metric (left—total specimens tested, right—specimens positive for A). Forecast scores for each square are averaged across all clusters and all seasons. The mean value of each row appears in the AVG column on the right. Model type is shown on LHS y-axis.
(PDF)

S6 Fig. Mean single-bin scores for total specimens forecasts 1–10 weeks ahead. Pixel colour shows forecast single-bin score for a given observation week averaged across clusters and seasons, e.g. the 4-week ahead forecast for week 48 was made using data only up to week 44. Averages across all weeks for a given model are printed on the RHS of each row of pixels. Model type is shown on LHS y-axis.
(PDF)

S7 Fig. Mean single-bin scores for positive-for-A forecasts 1–10 weeks ahead. Pixel colour shows forecast single-bin score for a given observation week averaged across clusters and seasons, e.g. the 4-week ahead forecast for week 48 was made using data only up to week 44. Averages across all weeks for a given model are printed on the RHS of each row of pixels. Model type is shown on LHS y-axis.
(PDF)

S8 Fig. Forecast scores for 1- to 4-week-ahead targets by season. Pixel color shows multi-bin forecast score (see main text) for a given observation week averaged across clusters and 1–4 week ahead target. Averages across all weeks of the season for a given model are printed on the RHS of each row of pixels. Model type is shown on LHS y-axis.
(PDF)

S9 Fig. Coupling parameter mean and central 80% interval. Results were compiled from MCMC chains across all models and forecast weeks for a given cluster and season. Coupling is modeled with the offset power law relationship described in (11) using free parameters for saturation distance and power. Briefly, the probability of coupling occurring across space is determined by a function with a plateau followed by a power law drop-off. Saturation distance is the width of the plateau and 'power' is the power coefficient of the drop-off.
(PDF)

S10 Fig. Spatial kernel posterior by cluster for positive-for-A. Spatial kernel κ posterior as a function of distance. Posterior distributions are across all season-years and β-modulation variants (fixed, humidity) for each cluster. Posteriors are taken from late-season fits using data through EW 18. Median kernel values appear as a black line and the 25% to 75% interval is shaded blue. For each unique county pair, there appears a vertical orange line indicating the great-circle distance (km) between population centroids.
(PDF)

S11 Fig. Spatial kernel posterior by cluster for total specimens. Spatial kernel κ posterior as a function of distance. Posterior distributions are across all season-years and β-modulation variants (fixed, humidity) for each cluster. Posteriors are taken from late-season fits using data through EW 18. Median kernel values appear as a black line and the 25% to 75% interval is shaded blue. For each unique county pair, there appears a vertical orange line indicating the

great-circle distance (km) between population centroids.
(PDF)

S12 Fig. Illustrative forecasts for all models. Each page contains results from a single model. Rows are divided first by data metric and then by season. Columns are divided by cluster. In the panels incidence data is shown in black and shaded cones depict the central 50% forecast windows for 1–4 week ahead. Forecasts were generated for every week, but only every third forecast is shown here for clarity. Colour progression from left to right indicates time of forecast. Solid lines show median fit to data at the point the forecast was assumed to have been made. Dashed lines show forecasts beyond the time at which data were assumed to be available. Mean score is the average score across MMWR forecasted weeks and 1 to 4 weeks ahead targets.
(PDF)

S1 Table. Pearson correlations between data metrics. Correlations are calculated for each cluster and season, and comparing total specimens, positive for, A and CDC %ILI. From the resulting distributions, we have extracted the median, 20%, and 80% quantiles.
(PDF)

S2 Table. County cluster information.
(PNG)

Acknowledgments

Author Contributions

Conceptualization: James Turtle, Pete Riley, Michal Ben-Nun, Steven Riley.

Data curation: James Turtle.

Formal analysis: James Turtle, Michal Ben-Nun.

Funding acquisition: Pete Riley.

Investigation: James Turtle.

Methodology: James Turtle, Pete Riley, Michal Ben-Nun, Steven Riley.

Project administration: Pete Riley.

Software: James Turtle, Michal Ben-Nun.

Supervision: Pete Riley, Steven Riley.

Validation: James Turtle, Steven Riley.

Visualization: James Turtle.

Writing – original draft: James Turtle, Steven Riley.

Writing – review & editing: James Turtle, Pete Riley, Michal Ben-Nun, Steven Riley.

References

1. Iuliano AD, Roguski KM, Chang HH, Muscatello DJ, Palekar R, Tempia S, et al. Estimates of global seasonal influenza-associated respiratory mortality: a modelling study. Lancet. 2018; 391(10127):1285–1300. https://doi.org/10.1016/S0140-6736(17)33293-2

2. 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(8):3146–3154. https://doi.org/10.1073/pnas.1812594116 PMID: 30647115

3. Cowling BJ, Ng Diane M W, Ip DKM, Liao Quiyan, Lam Wendy W T, Wu JT, et al. Community Psychological and Behavioral Responses through the First Wave of the 2009 Influenza A(H1N1) Pandemic in Hong Kong. The Journal of infectious diseases. 2010; 202(6):867–876. https://doi.org/10.1086/655811 PMID: 20677945

4. Lutz CS, Huynh MP, Schroeder M, Anyatonwu S, Dahlgren FS, Danyluk G, et al. Applying infectious disease forecasting to public health: a path forward using influenza forecasting examples. BMC Public Health. 2019; 19(1):1659. https://doi.org/10.1186/s12889-019-7966-8 PMID: 31823751

5. Kitler ME, Gavinio P, Lavanchy D. Influenza and the work of the World Health Organization. Vaccine. 2002; 20:S5–S14. https://doi.org/10.1016/S0264-410X(02)00121-4 PMID: 12110248

6. Gandon S, Day T, Metcalf CJE, Grenfell BT. Forecasting Epidemiological and Evolutionary Dynamics of Infectious Diseases. Trends Ecol Evol. 2016; 31(10):776–788. https://doi.org/10.1016/j.tree.2016.07.010 PMID: 27567404

7. World Health Organization. WHO public health research agenda for influenza 2017 update; 2017. Available from: https://www.who.int/influenza/resources/research/en/.

8. McGowan CJ, Biggerstaff M, Johansson M, Apfeldorf KM, Ben-Nun M, Brooks L, et al. Collaborative efforts to forecast seasonal influenza in the United States, 2015-2016. Sci Rep. 2019; 9(1):683. https://doi.org/10.1038/s41598-018-36361-9 PMID: 30679458

9. Biggerstaff M, Alper D, Dredze M, Fox S, Fung ICH, Hickmann KS, et al. Results from the centers for disease control and prevention's predict the 2013-2014 Influenza Season Challenge. BMC Infect Dis. 2016; 16:357. https://doi.org/10.1186/s12879-016-1669-x PMID: 27449080

10. Zimmer C, Yaesoubi R, Cohen T. A likelihood approach for real-time calibration of stochastic compartmental epidemic models. PLoS computational biology. 2017; 13(1):e1005257. https://doi.org/10.1371/journal.pcbi.1005257 PMID: 28095403

11. Ben-Nun M, Riley P, Turtle J, Bacon DP, Riley S. Forecasting national and regional influenza-like illness for the USA. PLoS Comput Biol. 2019; 15(5):e1007013. https://doi.org/10.1371/journal.pcbi.1007013 PMID: 31120881

12. Moss R, Zarebski A, Dawson P, McCaw JM. Retrospective forecasting of the 2010–2014 Melbourne influenza seasons using multiple surveillance systems. Epidemiology & Infection. 2017; 145(1):156–169. https://doi.org/10.1017/S0950268816002053 PMID: 27671159

13. Osthus D, Hickmann KS, Caragea PC, Higdon D, Del Valle SY. Forecasting seasonal influenza with a state-space SIR model. The annals of applied statistics. 2017; 11(1):202. https://doi.org/10.1214/16-AOAS1000 PMID: 28979611

14. Yang W, Karspeck A, Shaman J. Comparison of filtering methods for the modeling and retrospective forecasting of influenza epidemics. PLoS Comput Biol. 2014; 10(4):e1003583. https://doi.org/10.1371/journal.pcbi.1003583 PMID: 24762780

15. Brooks LC, Farrow DC, Hyun S, Tibshirani RJ, Rosenfeld R. Flexible Modeling of Epidemics with an Empirical Bayes Framework. PLoS Comput Biol. 2015; 11(8):e1004382. https://doi.org/10.1371/journal.pcbi.1004382

16. Yang S, Santillana M, Kou SC. Accurate estimation of influenza epidemics using Google search data via ARGO. Proceedings of the National Academy of Sciences. 2015; 112(47):14473–14478. https://doi.org/10.1073/pnas.1515373112 PMID: 26553980

17. Brooks LC, Farrow DC, Hyun S, Tibshirani RJ, Rosenfeld R. Nonmechanistic forecasts of seasonal influenza with iterative one-week-ahead distributions. PLOS Computational Biology. 2018; 14(6):1–29. https://doi.org/10.1371/journal.pcbi.1006134 PMID: 29906286

18. Adhikari B, Xu X, Ramakrishnan N, Prakash BA. Epideep: Exploiting embeddings for epidemic forecasting. In: Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining; 2019. p. 577–586.

19. Zhang Q, Gioannini C, Paolotti D, Perra N, Perrotta D, Quaggiotto M, et al. Social data mining and seasonal influenza forecasts: the fluoutlook platform. In: Joint European Conference on Machine Learning and Knowledge Discovery in Databases. Springer; 2015. p. 237–240.

20. Farrow DC, Brooks LC, Hyun S, Tibshirani RJ, Burke DS, Rosenfeld R. A human judgment approach to epidemiological forecasting. PLoS Comput Biol. 2017; 13(3):e1005248. https://doi.org/10.1371/journal.pcbi.1005248 PMID: 28282375

21. Morgan JJ, Wilson OC, Menon PG. The wisdom of crowds approach to influenza-rate forecasting. In: ASME International Mechanical Engineering Congress and Exposition. vol. 52026. American Society of Mechanical Engineers; 2018. p. V003T04A048.

22. Farrow DC, Brooks LC, Hyun S, Tibshirani RJ, Burke DS, Rosenfeld R. A human judgment approach to epidemiological forecasting. PLoS computational biology. 2017; 13(3):e1005248. https://doi.org/10.1371/journal.pcbi.1005248 PMID: 28282375

23. Osthus D, Gattiker J, Priedhorsky R, Del Valle SY, et al. Dynamic Bayesian influenza forecasting in the United States with hierarchical discrepancy (with discussion). Bayesian Analysis. 2019; 14(1):261–312. https://doi.org/10.1214/18-BA1117

24. Ray EL, Reich NG. Prediction of infectious disease epidemics via weighted density ensembles. PLoS computational biology. 2018; 14(2):e1005910. https://doi.org/10.1371/journal.pcbi.1005910 PMID: 29462167

25. Yamana TK, Kandula S, Shaman J. Individual versus superensemble forecasts of seasonal influenza outbreaks in the United States. PLoS computational biology. 2017; 13(11):e1005801. https://doi.org/10.1371/journal.pcbi.1005801 PMID: 29107987

26. Dugas AF, Jalalpour M, Gel Y, Levin S, Torcaso F, Igusa T, et al. Influenza forecasting with Google flu trends. PloS one. 2013; 8(2):e56176. https://doi.org/10.1371/journal.pone.0056176 PMID: 23457520

27. Centers for Disease Control and Prevention. FluView interactive; Accessed: January 2020. Available from: https://www.cdc.gov/flu/weekly/fluviewinteractive.htm.

28. Kandula S, Yang W, Shaman J. Type- and Subtype-Specific Influenza Forecast. Am J Epidemiol. 2017; 185(5):395–402. https://doi.org/10.1093/aje/kww211

29. Morita H, Kramer S, Heaney A, Gil H, Shaman J. Influenza forecast optimization when using different surveillance data types and geographic scale. Influenza Other Respi Viruses. 2018; 12(6):755–764. https://doi.org/10.1111/irv.12594 PMID: 30028083

30. Yang W, Olson DR, Shaman J. Forecasting influenza outbreaks in boroughs and neighborhoods of New York City. PLoS computational biology. 2016; 12(11):e1005201. https://doi.org/10.1371/journal.pcbi.1005201 PMID: 27855155

31. Su K, Xiong Y, Qi L, Xia Y, Li B, Yang L, et al. City-wide influenza forecasting based on multi-source data. In: 2018 IEEE International Conference on Big Data (Big Data). IEEE; 2018. p. 3930–3937.

32. Lu FS, Hou S, Baltrusaitis K, Shah M, Leskovec J, Hawkins J, et al. Accurate influenza monitoring and forecasting using novel internet data streams: a case study in the Boston Metropolis. JMIR public health and surveillance. 2018; 4(1):e4. https://doi.org/10.2196/publichealth.8950 PMID: 29317382

33. Leuba SI, Yaesoubi R, Antillon M, Cohen T, Zimmer C. Tracking and predicting US influenza activity with a real-time surveillance network. PLoS computational biology. 2020; 16(11):e1008180. https://doi.org/10.1371/journal.pcbi.1008180 PMID: 33137088

34. Meyers L, Ginocchio CC, Faucett AN, Nolte FS, Gesteland PH, Leber A, et al. Automated real-time collection of pathogen-specific diagnostic data: syndromic infectious disease epidemiology. JMIR public health and surveillance. 2018; 4(3):e59. https://doi.org/10.2196/publichealth.9876 PMID: 29980501

35. Choi SB, Kim J, Ahn I. Forecasting type-specific seasonal influenza after 26 weeks in the United States using influenza activities in other countries. PLoS One. 2019; 14(11):e0220423. https://doi.org/10.1371/journal.pone.0220423 PMID: 31765386

36. Dunn J, Obuekwe J, Baun T, Rogers J, Patel T, Snow L. Prompt detection of influenza A and B viruses using the BD Veritor System Flu A+ B, Quidel Sofia Influenza A+ B FIA, and Alere BinaxNOW Influenza A&B compared to real-time reverse transcription-polymerase chain reaction (RT-PCR). Diagn Microbiol Infect Dis. 2014; 79(1):10–13. https://doi.org/10.1016/j.diagmicrobio.2014.01.018 PMID: 24582581

37. Centers for Disease Control and Prevention. FluSight Influenza Forecasting Challenge 2018-2019 Guidelines; 2018. Available from: https://predict.cdc.gov/api/v1/attachments/flusight%202018%E2%80%932019/flu_challenge_2018-19_tentativefinal_9.18.18.docx.

38. Centers for Disease Control and Prevention. Morbitity and Mortality Weekly Report week-of-year definition; Accessed: 2020. Available from: https://wwwn.cdc.gov/nndss/document/MMWR_week_overview.pdf.

39. U S Census Bureau. Cartographic Boundary Files; 2018. Accessed: February 2021. Available from: https://www2.census.gov/geo/tiger/GENZ2018/shp/cb_2018_us_county_500k.zip.

40. Ben-Nun M, Riley P, Turtle J, Bacon DP, Riley S. Dynamics of Interacting Community Epidemics; 2020. [Computer Software]. Zenodo. https://doi.org/10.5281/zenodo.5090751.

41. Hethcote HW. The mathematics of infectious diseases. SIAM Rev. 2000; 42(4):599–653. https://doi.org/10.1137/S0036144500371907

42. Mills HL, Riley S. The Spatial Resolution of Epidemic Peaks. PLoS Computational Biology. 2014; 10(4):e1003561. https://doi.org/10.1371/journal.pcbi.1003561 PMID: 24722420

43. Shaman J, Pitzer VE, Viboud C, Grenfell BT, Lipsitch M. Absolute humidity and the seasonal onset of influenza in the continental United States. PLoS Biol. 2010; 8(2):e1000316. https://doi.org/10.1371/journal.pbio.1000316 PMID: 20186267

44. Kalnay E, Kanamitsu M, Kistler R, Collins W, Deaven D, Gandin L, et al. The NCEP/NCAR 40-Year Reanalysis Project. Bull Am Meteorol Soc. 1996; 77(3):437–472. https://doi.org/10.1175/1520-0477 (1996)077%3C0437:TNYRP%3E2.0.CO;2

45. Areas GA. GADM database of Global Administrative Areas, version 2.0; 2012. University of California, Berkeley. Available from: https://gadm.org/.

46. Lloyd CT, Sorichetta A, Tatem AJ. High resolution global gridded data for use in population studies. Sci Data. 2017; 4:170001. https://doi.org/10.1038/sdata.2017.1 PMID: 28140386

47. Gilks WR, Richardson S, Spiegelhalter D. Markov Chain Monte Carlo in Practice. CRC Press; 1995.

48. Bracher J. On the multibin logarithmic score used in the FluSight competitions. Proc Natl Acad Sci U S A. 2019; 116(42):20809–20810. https://doi.org/10.1073/pnas.1912147116 PMID: 31558612

49. Reich NG, Osthus D, Ray EL, Yamana TK, Biggerstaff M, Johansson MA, et al. Reply to Bracher: Scoring probabilistic forecasts to maximize public health interpretability. Proc Natl Acad Sci U S A. 2019; 116(42):20811–20812. https://doi.org/10.1073/pnas.1912694116 PMID: 31558611

50. 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(10):e1003256. https://doi.org/10.1371/journal.pcbi.1003256 PMID: 24146603

51. Riley P, Cost AA, Riley S. Intra-Weekly Variations of Influenza-Like Illness in Military Populations. Mil Med. 2016; 181(4):364–368. https://doi.org/10.7205/MILMED-D-15-00226 PMID: 27046183

52. Thornton HV, Turner KME, Harrison S, Hammond A, Hawcroft C, Hay AD. Assessing the potential of upper respiratory tract point-of-care testing: a systematic review of the prognostic significance of upper respiratory tract microbes. Clin Microbiol Infect. 2019; 25(11):1339–1346. https://doi.org/10.1016/j.cmi. 2019.06.024 PMID: 31254715

53. Pei S, Kandula S, Yang W, Shaman J. Forecasting the spatial transmission of influenza in the United States. Proc Natl Acad Sci U S A. 2018; 115(11):2752–2757. https://doi.org/10.1073/pnas.1708856115 PMID: 29483256

54. Dalziel BD, Kissler S, Gog JR, Viboud C, Bjørnstad ON, Metcalf CJE, et al. Urbanization and humidity shape the intensity of influenza epidemics in U.S. cities. Science. 2018; 362(6410):75–79. https://doi. org/10.1126/science.aat6030 PMID: 30287659