National and Regional Influenza-Like-Illness Forecasts for the USA

Michal Ben-Nun[1,*], Pete Riley[1], James Turtle[1], David P. Bacon[2], and Steven Riley[1,3]

1 Predictive Science Inc., San Diego, CA, USA

2 Leidos, McLean, VA, USA

3 MRC Centre for Outbreak Analysis and Modelling, Imperial College London, UK

* mbennun@predsci.com

Abstract

There is demand from national health planners for forecasts of key metrics associated with influenza-like-illness (ILI); near-term weekly incidence, week of season onset, week of peak, and intensity of peak. Here, we describe our participation in a weekly prospective ILI forecasting challenge for the United States for the 2016/17 season and a subsequent evaluation of our performance.

We implemented a meta-population model framework with 32 specific model variants. Model variants differed from each other in their assumptions about: the force-of-infection (FOI); use of uninformative priors; the use of discounted historical data for not-yet-observed time points; and the treatment of regions as either independent or coupled. Outputs from model variants were chosen subjectively as the the basis for our weekly forecasts. Coupled models were only available part way thought the season.

Most frequently, during the 2016-17 season, we chose; FOI variants with both school vacations and humidity terms; uninformative priors; the inclusion of discounted historical data for not-yet-observed time points; and coupled regions (when available). Our near-term weekly forecasts substantially over-estimated early incidence when coupled models were not available. However, our forecast accuracy improved substantially once coupled solutions were available. We were able to forecast onset and

peak timing and peak intensity with reasonable accuracy but without a long lead time. When we conducted retrospective forecasts for six previous seasons for which data were available, we found that the 2016/17 season was not typical: on average, coupled models performed better when fit without historically augmented data.

We were able to substantially improve accuracy during a prospective forecasting exercise by coupling dynamics between regions. Although the reduction of subjectivity should be a long-term goal, some degree of human intervention is likely to improve forecast accuracy in the medium-term.

Author summary

Influenza typically infects approximately 4 million people each year, resulting in 400,000 or more deaths. Influenza-like-Illness (ILI) is a practical way for health-care workers to easily estimate likely influenza cases. The Centers for Disease Control (CDC) collects and disseminates this information, and has, for the last several years, run a forecasting challenge (the CDC Flu Challenge) for modelers to predict near-term weekly incidence, week of season onset, week of peak, and intensity of peak. We have developed a modeling framework that accounts for a range of mechanisms thought to be important for influenza transmission, such as climatic conditions, school vacations, and coupling between different regions. In this study we describe our forecast procedure for the 2016/2017 season and highlight which aspects resulted in better or worse forecasts. Most notably, we found that when the dynamics of different regions are coupled together, the forecast accuracy improves. We also found that the most accurate forecasts required some level of forecaster interaction, that is, the procedure could not be completely automated without a reduction in accuracy.

Introduction

Infectious pathogens with short generation times pose public health challenges because they generate substantial near-term uncertainty in the risk of disease. This uncertainty is most acute and shared globally during the initial stages of emergence of novel human pathogens such as SARS [1], pandemic influenza [2], or Zika virus [3]. However, at

national and sub-national levels, uncertainty arises frequently for epidemic pathogens such as seasonal influenza, dengue, RSV and rotavirus; causing problems both for health planners and at-risk individuals who may wish to change their behavior to mitigate their risk during peak periods.

Seasonal influenza affects populations in all global regions and is forecast annually in temperate populations, either implicitly or explicitly. Peak demand for both outpatient and inpatient care is driven by peak incidence of influenza in many years [4]. Therefore, the efficient provision of elective procedures and other non-seasonal health care can be improved by accurate forecasts of seasonal influenza. Implicitly, most temperate health systems use knowledge of historical scenarios with which to plan for their influenza season. The current situation is then assessed against the deviation from the historical averages and worst-cases as observed in their own surveillance system.

The United States Centers for Disease Control (CDC) has sought to formalize regional and national forecasts by introducing an annual competition. Each week, participating teams submit weekly estimates of incidence for the next four weeks, season onset, and timing and intensity of the peak. Methods used by teams include purely statistical models, mechanistic models, machine learning and hybrid approaches. Expert-opinion surveys have also been used and performed well. Some teams augment their forecasts of the official ILI data with the use of potentially faster or less-noisy datasets such as google flu trends [5].

Here we describe a our mechanistic-model-supported participation in the 2016/17 CDC influenza forecasting challenge, as an example of a disease forecasting process. We emphasize a subjective human component of this process. We also describe a retrospective evaluation of the models for the previous six seasons. All the models described are implemented in the R package Dynamics of Interacting Community Epidemics (DICE, `https://github.com/predsci/DICE`).

Methods

Data

The CDC Influenza-like Illness Surveillance Network (ILINet) Human and Health Services (HHS) region and national data were downloaded from the CDC-hosted web application FluView [6] and used to create a historic database of ILI cases. Fig. 1 shows which states are grouped into each HHS region. Because we require an absolute number of cases per week, the CDC ILINet data is converted from percent ILI cases per patient to ILI cases. We estimate the absolute number of weekly ILI cases by dividing the weighted percent of ILI cases in the CDC data by 100 and multiplying it by the total weekly number of patients. We assume two outpatient visits per person per year so that the total weekly number of patients is estimated as: (total regional population)*(2 outpatient visits per person per year)*(1 year/52 weeks).

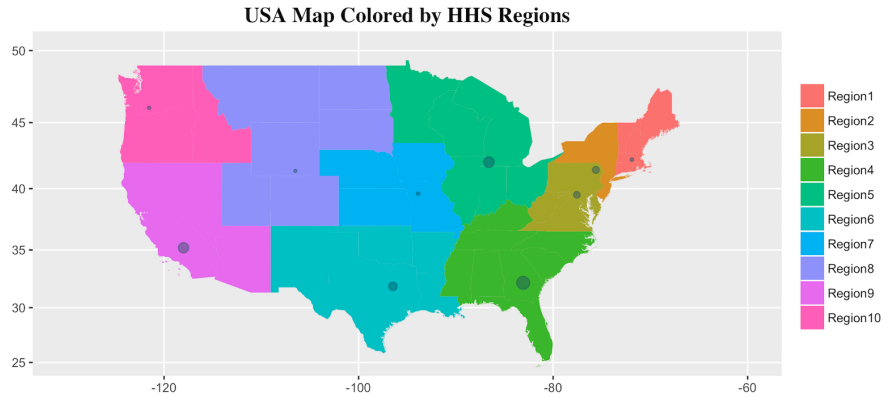


Fig 1. A map of the continental US colored by the ten HHS regions. The green circle in each HHS region denotes the population density weighted location of the centroid of the region, and the radius of each circle is proportional to the weight of the region which is determined by its relative population. Regions are sometimes referred to by the city in which their HHS office is located: 1, Boston; 2, New York; 3, Philadelphia; 4, Atlanta; 5, Chicago; 6, Dallas; 7, Kansas City; 8, Denver; 9, San Fransisco; 10, Seattle.

Specific humidity (SH) is included in the DICE for this time period and uses Phase-2 of the North American Land Data Assimilation System (NLDAS-2) data base provided by NASA [7–9]. The NLDAS-2 data base provides hourly specific humidity (measured 2-meters above the ground) for the continental US at a spatial grid of 0.125° which we average to daily and weekly values. The weekly data is then spatially-averaged for the states and CDC regions. For the three states outside the continental US (Alaska, Hawaii, and Puerto Rico), and all other countries, we obtain the SH data from NOAA's

NCEP-NCAR Reanalysis project [10] (see for example [11]) which provides daily (again 2-meter above ground) SH data on a spatial grid of 2.5° for the entire world. These data are averaged and interpolated using the same procedure as for the NLDAS-2 dataset. When running in a forecast mode (as was done during this CDC flu challenge), future SH data are provided using the average of historic data.

School vacation schedules were collected for the 2014-2015 and 2015-2016 academic years for every state. For each state, a school district was identified to represent each of the three largest cities. Vacation schedules were then collected directly from the district websites. These three school vacation schedules were first processed to a weekly schedule with a value of 0 indicating class was in session all five weekdays and a 1 indicating five vacation days. Next, the representative state schedule was produced by averaging the three weekly district schedules. Region schedules are obtained by a population-weighted average of the state schedules. Similarly, the national schedule is generated by a population-weighted average of the regions.

For the 2016-2017 season we determine start and end times as well as spring and fall breaks from the previous years' schedules. Thanksgiving and winter vacation timing was taken from the calendar where the winter break is assumed to be the last two calendar weeks of the year. Based on the proportion of schools closed and number of days closed, $p(t)$ is assigned a value in the range $[0, 1]$. For example in week t_i, if all schools are closed for the entire week then we define the proportion of open schools $p(t_i) = 1$. However, if all schools have Monday and Tuesday off (missing 2 of 5 days), then $p(t_i) = 0.4$. Similarly, if 3 of 10 schools have spring break (entire week off), but the other 7 schools have a full week of class then $p(t_i) = 0.3$. If all schools have a full week of class then $p(t_i) = 0$.

Basic model

The DICE package has been designed to implement meta-population epidemic modeling on an arbitrary spatial scale with or without coupling between the regions. Our model for coupling between spatial regions follows ref [12]: We assume a system of coupled S-I-R equations (susceptible-infectious-recovered) for each spatial region. In this scenario, the rate at which a susceptible person in region j becomes infectious (that is

transitions to the I compartment in region j) depends on: (1) the risk of infection from those in the same region j, (2) the risk of infection from infected people from region i who traveled to region j, and (3) the risk of infection encountered when traveling from region j to region i. To account for the three mechanisms of transmission, ref [12] defined the force of infection, or the average rate that susceptible individuals in region i become infected per time step as:

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

where D is the total number of regions. In our-case, unlike reference [12], the transmissibility is not the same for all regions and it is allowed to depend on time: $\beta_j(t)$.

Given this force of infection we can write the coupled S-I-R equations for each region as:

$$\frac{dS_j}{dt} = -\lambda_j(t) S_j, \tag{2}$$

$$\frac{dI_j}{dt} = \lambda_j(t) S_j - \frac{I_j}{T_g}, \tag{3}$$

$$\frac{dR_j}{dt} = \frac{I_j}{T_g}. \tag{4}$$

Eqs. (2-4) are the coupled version of the familiar S-I-R equations, where T_g is the recovery rate (assumed to be 2-3 days in the case of influenza). The mobility matrix, m_{ij} of Eq. 1 describes the mixing between regions. Thus, element i, j is the probability for an individual from region i, given that the individual made a contact, that that contact was with an individual from region j. As shown below, the sum over each row in the mobility matrix is one and in the limit of no mobility between regions the mobility matrix m_{ij} is the identity matrix so that $\lambda_i(t) = \beta_i(t) \frac{I_i}{N_i}$ and we recover the familiar (uncoupled) S-I-R equations:

$$\frac{dS_j}{dt} = -\beta_j(t) \frac{S_j I_j}{N_j}, \tag{5}$$

$$\frac{dI_j}{dt} = \beta_j(t) \frac{S_j I_j}{N_j} - \frac{I_j}{T_g}, \tag{6}$$

$$\frac{dR_j}{dt} = \frac{I_j}{T_g}. \tag{7}$$

The level of interaction between spatial regions is determined by the mobility matrix and its interaction kernel , $\kappa(r_{ik})$:

$$m_{ij} = N_j \kappa(r_{ij}) \frac{1}{\sum_k N_k \kappa(r_{ik})} \tag{8}$$

This kernel is expected to depend on the geographic distance between the regions (r_{ij}), and following Mills and Riley [12] we use a variation of the off-set power function for it:

$$\kappa(r_{ij}) = \frac{1}{1 + (r_{ij}/s_d)^{\gamma}} \tag{9}$$

where s_d is a saturation distance in km and the power γ determines the amount of mixing between the regions: as γ decreases there is more mixing while as γ increases, mixing is reduced. In the limit that $\gamma \to \infty$there is no mixing between regions and we recover the uncoupled SIR Eqs. (5-7). The DICE package is designed to allow the estimation of these two parameters (γ and s_d), but they can also be set to fixed values.

The S-I-R equations model the total population, but the data are the number of weekly observed cases or incidence rate for each spatial region (I_j^R). The weekly incidence rate is calculated from the continuous S-I-R model by discretizing the rate-of-infection term$\lambda_j(t)S_j$ (or $\beta_j(t)\frac{S_j I_j}{N_j}$ in the uncoupled case):

$$I_j^R(t_i) = B_j + p_j^C \int_{t_i - 1 - \Delta_t}^{t_i - \Delta_t} \lambda_j(t) S_j(t) dt \ , \tag{10}$$

scaling by percent clinical p_j^C, and adding a baseline B_j. p_j^C is the proportion of infectious individuals that present themselves to a clinic with ILI symptoms and B_j is a constant that estimates non-S-I-R or false-ILI cases. The integral runs over one week determining the number of model cases for week t_i. Δ_t approximates the time delay from when an individual becomes infectious to when they visit a sentinel provider for ILI symptoms and is to 0.5 weeks based on prior calibration [13, 14]. Eq. 10 describes how DICE relates its internal, continuous S-I-R model to the discrete ILI data. In the next section we describe the procedure used for fitting this property to an ILI profile.

To allow for different models for the force of infection/contact rate, we write this term in the most general way as a product of a basic force of infection, R_j^0, multiplied by three time dependent terms:

$$\beta_j^e(t) = \frac{R_j^0}{T_g} \cdot F_1(t) \cdot F_2(t) \cdot F_3(t) \tag{11}$$

The first time dependent term, $F_1(t)$, allows for a dependence of the transmission rate on specific-humidity, the second ($F_2(t)$) on the school vacation schedule, and the third ($F_3(t)$) allows the User to model an arbitrary behavior modification that can drive the transmission rate up or down for a limited period of time. For the purpose of the CDC challenge we only considered models involving either $F_1(t)$, $F_2(t)$, both, or none (i.e., the contact rate does not depend on time), and the functional form of these terms is discussed in sections S1 Text and S2 Text of the Supporting Information.

Fitting the model

The DICE fitting procedure determines the joint posterior distribution for the model parameters using a Metropolis-Hastings Markov Chain Monte Carlo (MCMC) procedure. We describe the procedure starting with the simpler uncoupled case. In the uncoupled scenario infection can only happen within each HHS region, since there is no interaction between different spatial regions. The uncoupled regions are run sequentially and posterior distributions for the model parameters and forecasts are obtained. For each region, we simulate three MCMC chains each with 10^7 steps and a burn time of 2×10^6 steps. The smallest effective sample size that we report for any parameter is greater than 100. After sampling from the individual posterior densities of each region, we calculate our national forecast as the weighted sum of the regional profiles with the weights given by the relative populations of the regions. It is also possible to fit the national % ILI profile directly, without using any information from the ten HHS regions.

In the coupled scenario, the MCMC procedure uses Eqs. (2-4) along with Eq. (10) to simultaneously generate candidate profiles for the coupled ten HHS regions. The log-likelihood of the ten regional profiles is calculated and combined with the proper relative weights to generate a national log-likelihood which is minimized. It is important to note that in the coupled scenario we *only* optimize the national log-likelihood, and not the individual region-level likelihoods. We also tried fitting the coupled model to the regional log-likelihoods, however the results of the fits were not as accurate as the ones obtained when the national likelihood is optimized (see discussion).

Specific Model variants

We ran the model using four variants for the force of infection: (i) The force of infection depends on specific humidity only (H), (ii) school vacation only (V), (iii) both (HV) or, (iv) none (F).

Coupling HHS regions

We ran the model with and without connectivity between the different regions.

Informative priors

In the previous section we described a traditional MCMC procedure which uses an uninformed prior (UP), that uses a log uniform distribution for the parameters. Early in the flu season, before the ILI curve takes off, this fitting can result in peak intensities that are significantly larger/lower than expected (based on historic values) and/or peak weeks that are inconsistent with past values. One way to constrain the predictions, which has been used by others [15, 16] is to use an informed prior (IP). We have used each of the models supported by DICE to fit all previous seasons (starting from 2004) at both the national and regional levels. Using the history of the MCMC chain we then built a posterior distribution for each parameter and fitted it to a 1-D Gaussian. (This assumes that the model parameters are independent.) By repeating this procedure for each season and each model, we create a database of prior knowledge which can be used to inform the MCMC procedure for the current season. Specifically, at each week during the CDC challenge, and for each of the ten HHS regions and the national, we find the past season that is most similar to the current data (based on the value of the Pearson correlation) and use the posterior distributions of that season as an informed prior for the current season. Each region/national has its own informed prior which is allowed to change from one epidemic week to the other. To allow for an informed prior that is less restrictive, we also use a heated informed prior, where the Gaussian temperature is increased by an order of magnitude. In the Results section we refer to the fitting procedures that use a prior as IP and HIP for informed prior and heated informed prior, respectively. Informed priors were used *only* with the uncoupled SIR Eqs. In a future study we plan to explore how they would extend to the coupled MCMC procedure.

Using discounted historical data for not-yet-seen future time points

In addition to informative priors, we also used data augmentation to make maximum use of prior data within a mechanistic framework. For each week during the challenge, we augment the available ILI data using average historic data or using the season that is most similar to the current season. We shifted from the historic data to the most similar data at EW 6 when it was clear that the current season is very different from the historic average. The augmented data was also y-shifted so that it matched the last data point for the current season and it is given a lower weight in the MCMC procedure, determined by the value of the Pearson correlation between the available data and the data used for augmentation. The augmented data procedure was used for both the coupled and uncoupled fits and also using a heated augmented procedure (where the log-likelihood is again heated by a factor of ten). In what follows, we refer to the fitting procedures that use data augmentation and heated data augmentation as DA and HDA, respectively.

From model predictions to forecasts

During each of the CDC weeks DICE was used to fit both the regional and the national curves using the combinations of coupling, priors and models described in the previous subsections. The uncoupled procedure (and direct fitting of the National %ILI profile) were used throughout the season with all five priors and with the four models for the force of infection leading to: $5 \times 4 = 20$ forecasts. For the coupled procedure we fitted with the following combinations of priors and data augmentation: uninformed prior with no data augmentation, uninformed prior with data augmentation and uninformed prior with heated data augmentation (UP, DA and HDA, respectively), but with all four models for the force of infection, leading to: $3 \times 4 = 12$ forecasts.

This total of 32 model-runs were used to make predictions of incidence at both the national and regional levels. The national curve was also fitted directly (without any regional information) using all the models and priors, but these direct results were only used at the end of the season when estimating the performance of each of our procedures.

Early in the season we were experimenting with the coupled procedure and we began

to use it as described in the manuscript with the DA and HDA priors on EW 50 and with the UP prior only on EW 9. Hence, some of the coupled results reported in this section were not available in real-time and were generated at the end of the season (but using only the %ILI data that was available in real-time at each week.)

Each week a single forecast was selected from these results for each of the ten HHS regions and the national. At the regional level we selected a single forecast from one of the (32) uncoupled or coupled procedures enumerated in the previous paragraph. We first reviewed the results for each region and made a selection which took into account: (1) the historic profile for the region, (2) the quality of the fit (both mean and width), (3) the extent of data adjustments in the past weeks, (4) the lab strain % positive data, and when relevant (5) the impact of upcoming school vacations (particularly the winter break). For the national level we selected either one of the 12 coupled results or the aggregated result obtained as the weighted average of our ten regional selections (with the weights given by the relative population of each region).

Results

Models selected for forecasts

We selected different FOI variants during different weeks. At the regional level, although we selected the most flexible humidity and school vacation assumptions (HV) more often (47.9% 134/280) than the alternatives (Fig 2), we did select humidity-only (H, 47/280 16.8%), school-vacations-only (V, 62/280 22.1%) and fixed transmissibility (F, 37/280 13.2%) models on a number of occasions. For the national model, we only used the aggregated forecast of the regional models on 9/28 (32.1%) occasions.

For assumptions about the inferential prior, before the epidemic peaked (EW 06 for the nation) we rarely chose an uninformed prior (UP). Most often we chose the heated data augmentation option (HDA). Early in the season (EW 43-49), when our options did not include the coupled procedure, we chose the informed prior (IP) or heated informed prior (HIP) most often. Once the season had peaked the uninformed prior was selected often both for the national profile and individual regions, we continued to select the data augmentation (DA) and HDA options for the nation well after the season has peaked (EWs 12-14, 16 and 18).

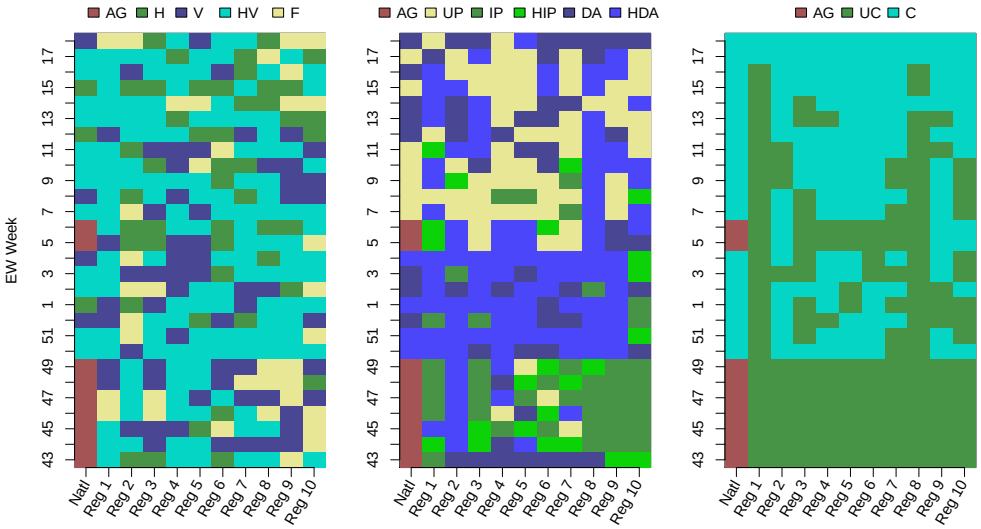


Fig 2. Forecast model choices for force of infection, use of priors, coupling, and data augmentation; for national and regional estimates by week Left panel: the selected model for the force of infection: H - specific humidity only, V - school vacation only, HV - both specific humidity and school vacation and F - fixed value for the force of infection, AG - national aggregated from the regional choices (weighted average of the direct uncoupled fits to each of the ten HHS regions). Middle panel: selected model for the prior distribution. UP - uninformed prior, IP - informed prior, HIP - heated informed prior, DA - data augmentation, and HDA - heated data augmentation. Right panel: selected spatial coupling model: UC - uncoupled, C - coupled. For the national AG indicates aggregate where the national forecast .

For assumptions about coupling, once the coupled procedure was available, it was often selected for both the national profile and most regions (2, right panel), with the exception of regions 1 and 8. We found that the coupled procedure used regions 8 (and to a lesser extent 1) as a way to reduce the error to the national fit, at the cost of producing poor fits to these regions, hence their coupled results were rarely selected for submission. The aggregate option for the national selection was only selected at EWs 5 and 6, the weeks prior to the peak and the peak week itself. For these two weeks our errors for both season targets and $1-4$ week forecasts were large (see Figure 4 below).

Accuracy of forecasts

At both the national and regional levels, the accuracy of the weekly %-ILI forecasts decreased as the lead time increased. The %ILI 1-4 week forecast and observed data for the national data and the three largest HHS regions (Figure 1): 4, 5, and 9 is shown in Figure 3. Early in the season, up to and including EW01, the national curve is nearly identical to the historic national curve, whereas our mechanistic forecasts consistently

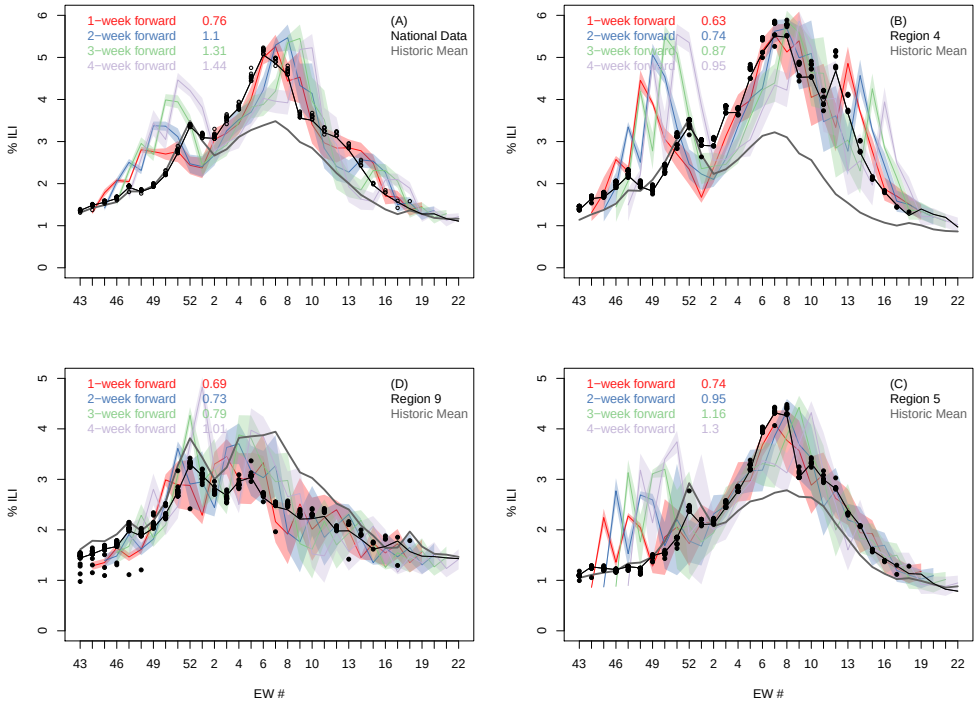


Fig 3. Comparison of submitted forecast and final CDC-reported % of clinic visits that were for ILI for the continental USA and three selected regions. Final season CDC reported (black line), reported during the season (black circles) and predicted %ILI (colored bands) as a function of epidemic week at the national (**A**) and three of the ten regional levels (**B** to **D**). In each panel, the four colored shaded bands denote our one (red), two (blue), three (green) and four (light purple) week-ahead predictions made at each week during the challenge. The width of the colored bands represent our 5-95% prediction intervals as submitted for the forecast. The gray line denotes the historic average. For similar plots for the other seven HHS regions see S1 Fig.

overestimated incidence. After EW01, our national predictions improve significantly, while the historic curve no longer follows the 2016-2017 national profile. Similarly, for the three largest HHS regions, the historic curve is similar to the 2016-2017 profile until EW 01, at which point they start to deviate and our forecasts become more accurate.

Averaged over the entire season, our selected national forecast does better than the historic NULL model only for the 1-week prediction window (Figure 4). However, for the largest HHS region (Region 4) we perform better for all four prediction horizons, for region 9 (bottom left panel) for the first three, and for region 5 (bottom right panel) for the first two.

Although our forecasts gave potentially useful information over and above the null model for the timing of the peak week (Figure 4) and for the amplitude of peak

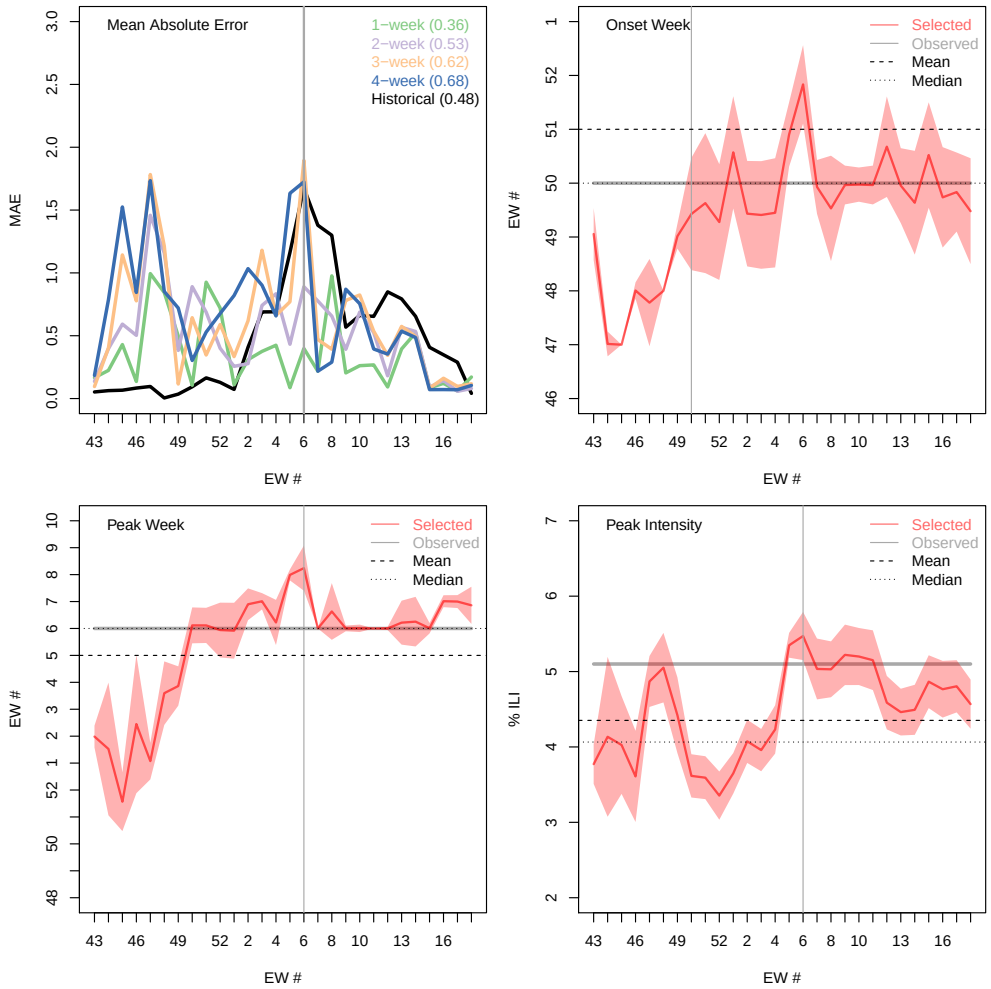


Fig 4. A Weekly Mean absolute error for the n-week ahead forecast (1 week, green; 2 weeks, purple; 3 weeks, orange; and 4 weeks, blue) and the historic NULL model (black). The average mean absolute error of each forecast horizon, and of the historic prediction, is shown in the legend. Forecast and actual onset week (**B**, peak week (**c** and peak intensity (**D** during each week of the CDC challenge: selected forecast (red line); 95% prediction interval (red shading); observed value (gray horizontal line); historic mean (horizontal dashed line); and historic median (horizontal dotted line). The vertical gray line shows the observed peak week (A, C and D) and the observed onset time (B).

intensity, the peak week of EW 6 was the same as the historical mean. Between EW50 (eight weeks before the season peaks) and EW 4 (two weeks before the season peak) our forecast correctly predicted to within ±1 week of the observed peak week (EW 6). One week before the season peaks, and at the peak week (EWs 5 and 6), our model forecast has an error of two weeks.

Forecasts based on the mechanistic model performed better than the historic null model for the peak intensity (Figure 4 A/B/C/D). Two weeks before the peak week

(and three weeks early in the season) we start predicting the correct peak intensity of 5.1% (to within ±0.5%). The mean and median historic values are significantly lower (4.4% and 4.1% respectively) and outside the ±0.5% range. Our forecast performance for intensity appears to drop off at the end of the season. However, this is an artifact of the forecasting work flow. Once the peak had clearly passed, the final model was selected for reasons other than the peak intensity and the observed peak intensity was submitted.

Selected forecasts based on the mechanistic model did not accurately predict onset. Both the mean of onset (EW 51) and median (EW 50) historic values are within a week of the observed 2016/17 onset week (EW 50). However, our model was unable to properly predict the onset until it happens. As with peak values, once onset has been observed in the data, we use the observed value in our formal submission, which is not reflected in onset values from the chosen model.

Retrospective analyses of model forecasting ability

Once the challenge was over, we examined retrospectively the performance of all mechanistic model variants over the course of all seasons in the historical database and separately for the 2016-17 season. To assess the quality of all the near-term forecasts ($1-4$ weeks) from the different models and assumptions about priors, we show in Figure 5 their weekly CDC score (see Methods), for the 4 different forecast lead times (1-, 2-, 3- and 4- weeks ahead), for the prediction of the National %ILI intensity. Coupled models were more accurate than non-coupled models for the 2016-17 season and for historical seasons for all 4 lead times.

The performance of mechanistic models was comparable to that of the historical average null model at the beginning and end of the season. However, in the middle of the season when there is greater variation in the historical data, the performance of the best mechanistic model variants was substantially better than that of the historical average model.

For predicting ILI incidence for the 2016-17 season, which followed similar trend to the historical average, coupled models that used data augmentation were more accurate than coupled models that did not use data augmentation. However, on average for historical seasons, coupled models that did not use augmented data were more accurate than those that did. Also, on average for historical seasons, coupled models that included humidity were more accurate than those that did not (see dark banding in

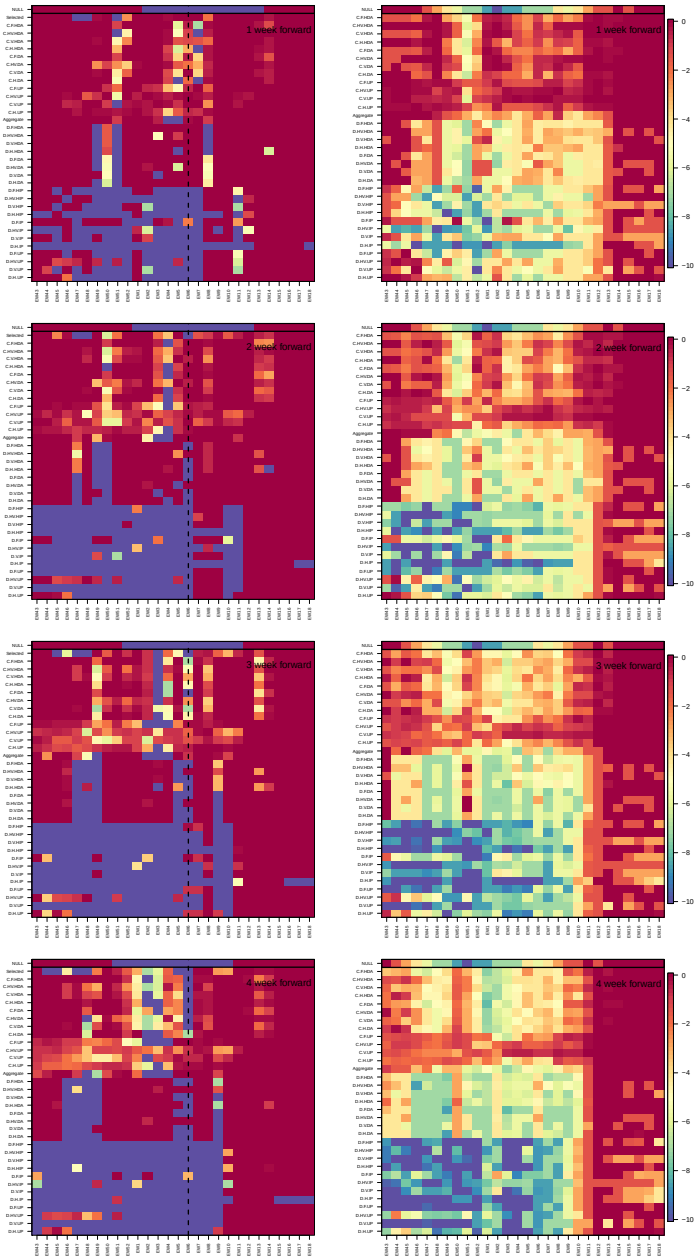


Fig 5. Weekly CDC score for the 1 to 4 week forward national %ILI forecast. Left column: 2016-2017 season. Right column: averages over seven flu seasons, 2010-2016. In the method/model labels (y-axis): D- direct, C-coupled, and aggregate, H, V, HV and F denote the four models for the force of infection: Humidity only, vacation only, both and fixed. The prior models are: uniform prior (UP), informed prior (IP), heated informed prior (HIP), data augmentation (DA), and heated data augmentation (HDA). Selected (only available for the 2016-2017 season), is what we submitted each week and the NULL model is the historic average. The vertical dashed black line, in the left panels, denotes the national season peak week for the 2016-2017 season.

upper portion of charts on the right hand side of Figure 5).

We examined the performance of the different model variants for individual regions for the near-term forecasting of %-ILI (Figures S2 Fig and S3 Fig). Again, the coupled models with uninformative prior outperformed other model variants. Although for some regions, the improvement in forecast score for the uninformative prior variants over other coupled variants was less pronounced (Regions 1 and 7), these models never appeared to be inferior to the other variants.

In a similar way, we examined the forecast accuracy of different mechanistic model variants in forecasting season-level targets: onset, peak time and peak intensity for the 2016-17 season and on average across all seasons (6). Again, for all three targets during 2016-17, the coupled uninformative model variant was at least as good as other coupled options and better than the non-coupled variants. We note that in the latter part of the season, after the single observed onset and peak had passed, results from a single season do not contain much information about model performance. However, the performance of the coupled uninformative prior model was on average better than other model variants across the historical data and different epidemic weeks for all three targets, other than one exception. From EW1 onwards for peak intensity, uncoupled heated augmented prior variants performed better than did coupled uninformative prior model variants.

Discussion

In this study, we have described our participation in a prospective forecasting challenge. Although we drew on results from a large set of mechanistic models, our single forecast for each metric was made after choosing between available model results for that metric in that week and was therefore somewhat subjective. We performed poorly at the start of the competition when our mechanistic models consistently over-estimated incidence. However, during the middle phase of the season, our models produced less biased estimates and consistently outperformed non-mechanistic models based on the average of historical data. A systematic testing of model variants using historical data suggests that spatially coupled models are systematically better than historical null models during the middle of the season and are not significantly worse even at the start of the

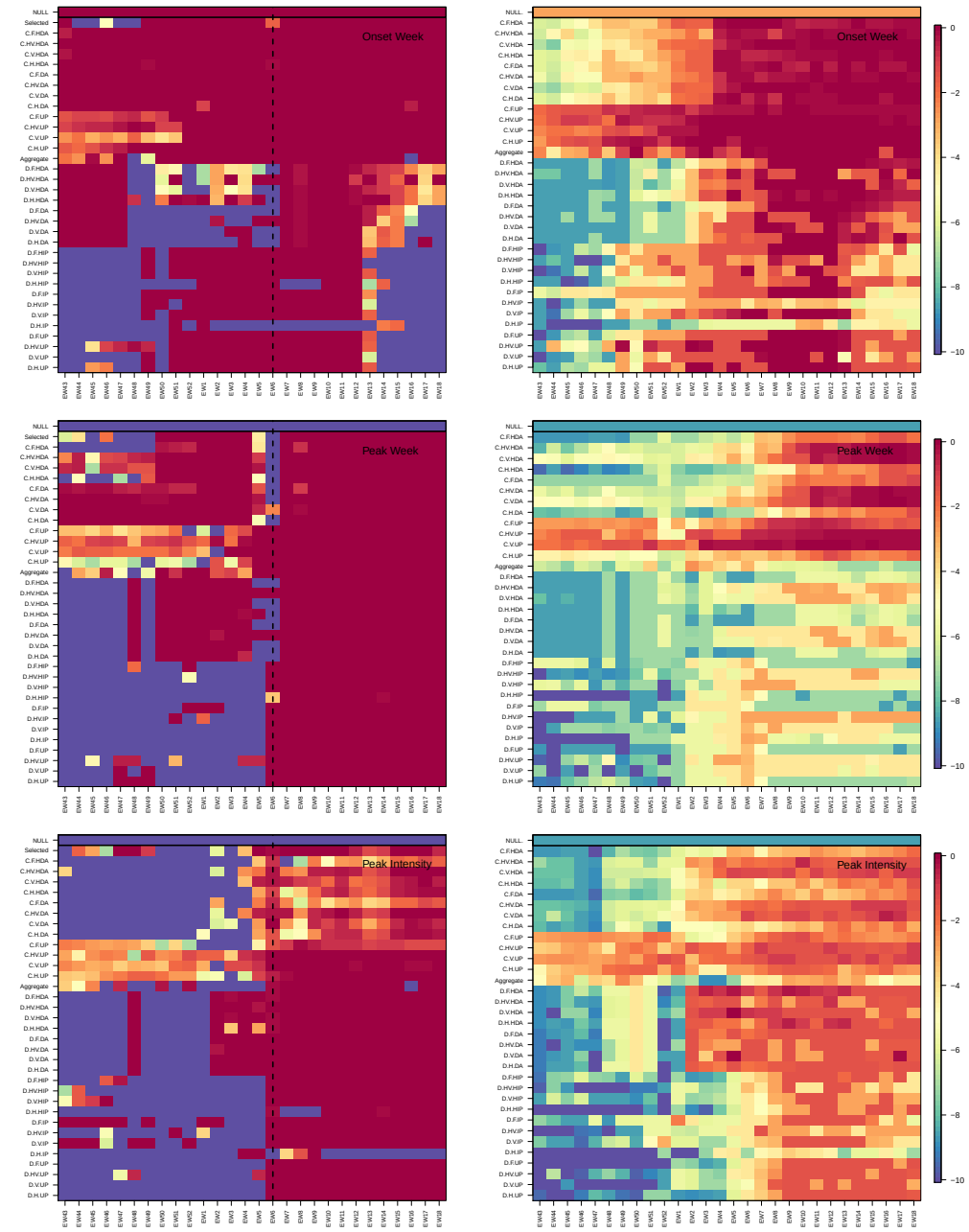


Fig 6. As in Figure 5 but for the seasonal targets: onset week (top), peak week (middle) and peak intensity (bottom). Left column: 2016-2017 season, right column: averaged over seven seasons, 2010-2016. For the 2016-2017 season, the coupled procedure with data augmentation correctly predicts onset week and peak week from the start of the CDC challenge. Peak intensity is predicted best with a coupled uninformed prior. When averaged over seven seasons (right column) the coupled uninformed prior does best for all three season targets.

season.

This study is slightly different from many prior studies of influenza forecasting [17]

in that it describes and assesses a subjective choice between multiple mechanistic models as the basis of a prospective forecast, rather than describing the performance of a single model or single ensemble of models used for an entirely objective forecast. Although this could be viewed as a limitation of our work, because individual subjective decisions cannot be reproduced, we suggest that the explicit description of a partially subjective process is a strength. In weather forecasting, there is a long history of evaluating the accuracy of entirely objective forecasts versus partially subjective forecasts [18, 19]. Broadly, for each different forecast target and each forecast lead-time, there has been a gradual progression over time such that objective forecasts are becoming more accurate than subjective forecasts. We note also that although we describe the subjective process as it was conducted, we also provide a thorough retrospective assessment of the predictive performance of each model variant.

By reflecting on our choices and their performance, we can evaluate the importance of a number of different model features. Our coupled model variants performed much better than uncoupled variants consistently across the 2016-17 season, for different targets and when evaluated using the historical data. This prospective study supports recent retrospective results suggesting that influenza forecasts can be more accurate if they explicitly represent spatial structure [20, 21]. Given that the model structure we used to represent space was relatively coarse [12], further work is warranted to test how forecast accuracy at larger spatial scales can be improved by models that include iteratively finer spatial resolution.

In submitting forecasts based on uninformed mechanistic priors using an uncoupled model at the start of the season, we failed to learn lessons that have been present in the influenza forecasting literature for some time [17]. Historical variance is low during the start of the season and the growth pattern is not exponential. Therefore, it would be reasonable to forecast early exponential growth only in the most exceptional of circumstance, such as during the early days of a pandemic. Model solutions that are anchored to the historical average in some way, such by the use of augmented data for not-yet-seen time points, are likely to perform better.

Models that included humidity forcing performed better on average in our analysis of all historical data than equivalent models that did not include those terms, especially for the forecasting of ILI 1- to 4-weeks ahead [22]. However, we did not see similar

support for the inclusion of school vacation terms improving accuracy, which has been suggested in a retrospective forecasting study at smaller spatial scales (by this group) [23]. The lack of support for school vacations in the present study could indicate that the prior work was under-powered or that sampling and then averaging of school vacation data across large spatial scales degrades its contribution to forecasts.

We found the experience of participating in a prospective forecasting challenge to be different to that of a retrospective modeling study. The feedback in model accuracy was much faster and the need for statistically robust measures of model likelihood or parsimony less obvious. We encourage the use of forecasting challenges for other infectious disease systems as a focus for better understanding of underlying dynamics in addition to any actionable information arising from the forecasts themselves.

Acknowledgments

Disclaimer: The findings and conclusions in this report are those of the author(s) and do not necessarily represent the views of the Department of Health and Human Services or its components, the US Department of Defense, local country Ministries of Health, Agriculture, or Defense, or other contributing network partner organizations. Mention of any commercial product does not imply DoD endorsement or recommendation for or against the use of any such product. No infringement on the rights of the holders of the registered trademarks is intended. No funding bodies had any role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Supporting information

S1 Text Details of the parametric dependence of the force of infection on specific humidity

Eq. 11 allows the transmission rate to depend on time using three terms. The first time dependent term, $F_1(t)$, allows for a dependence of the transmission rate on specific-humidity. In temperate regions specific humidity has a seasonal oscillation with a minimum in the winter and a maximum in the summer. We follow Shaman et al. [24] and relate the local SH, $q_j(t)$, to the reproduction number as:

$$\beta_j^e(t) = \frac{R_j^0}{T_g} \cdot F_1(t) = R_0 \times [1 + \Delta_R \cdot \mathrm{e}^{-a \cdot q_j(t)}] \tag{12}$$

In the above equation, and unlike the work published by others, the values of the parameters a and Δ_R are fitted. The effect of specific humidity can be combined with that of school vacation which is discussed in the following sub-section.

S2 Text Details of the parametric dependence of the force of infection on school vacation schedule

The second term in Eq. 11 allows the transmission rate to depend on the weekly school vacation schedule ($p_j(t)$) and we implement is as:

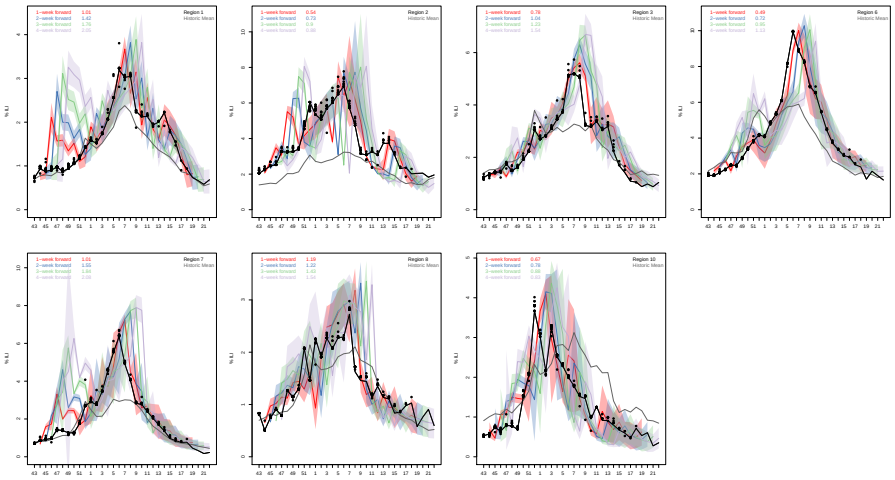
$$\beta_j^e(t) = \frac{R_j^0}{T_g} \cdot F_2(t) = R_0 \times [1 - \alpha \cdot p_j(t)] \tag{13}$$

DICE fits the effect of school closure by optimizing the parameter α, which is in the range $0 - 1$. Larger values of α indicate a more significant lowering of R_e as a result of planned school vacations. Conversely small values of α indicate that the school vacation schedule is not an important factor in determining the ILI profile. The effect of school vacation can be combined with that of specific humidity, i.e. $\beta_j^e(t) = \frac{R_j^0}{T_g} \cdot F_1(t) \cdot F_2(t)$.

S1 Fig. Predicted %ILI as a function of epidemic week for seven of the ten HHS regions.

Final season CDC reported (black line), reported during the season (black circles) and predicted %ILI (colored bands) as a function of epidemic week for seven of the ten HHS regions not shown in the main text panel. In each panel the four colored shaded

bands denote our n-week forward prediction ($n = 1, 2, 3, 4$), and the gray line denotes the historic average. The average relative error (measured with respect to the error of the historic NULL model) is indicated for each of the four prediction horizons in the legend. See main text and 3 for more details.



S2 Fig. Zoom-able separate file of weekly CDC score for $1-4$ week forward regional forecast averaged over the 2010-2016 seasons.

As in the right columns of Figure 5 but for the ten HHS regions. Each columns denotes an HHS region and going down a columns we move from 1 to 2, 3 and 4 weeks forecasts. In the method/model labels (y-axis): UN- uncoupled and C-coupled. H, V, HV and F denote the four models for the force of infection: Humidity only, vacation only, both and fixed. The prior models are: uniform prior (UP), informed prior (IP), heated informed prior (HIP), data augmentation (DA), and heated data augmentation (HDA). Selected, is what we submitted each week and the NULL model is the historic average. For all ten regions, and in agreement with the results for the nation, the coupled procedure performs better than the uncoupled.

S3 Fig. Zoom-able separate file of weekly CDC score for the seasonal targets: onset week, peak week and peak intensity for the ten HHS regions averaged over the 2010-2016 seasons.

Top, middle and bottom rows are onset week, peak week and peak intensity. The NULL result is calculated using the historic mean regional profile. For all three targets, and all ten regions, the coupled method does better than the uncoupled with the details of the prior and force of infection models depending on the region.

References

1. Tsang KW, Ho PL, Ooi GC, Yee WK, Wang T, Chan-Yeung M, et al. A cluster of cases of severe acute respiratory syndrome in Hong Kong. N Engl J Med. 2003;348(20):1977–1985.

2. Centers for Disease Control and Prevention CDC. Swine influenza A (H1N1) infection in two children–Southern California, March-April 2009. MMWR Morb Mortal Wkly Rep. 2009;58(15):400–402.

3. Campos GS, Bandeira AC, Sardi SI. Zika Virus Outbreak, Bahia, Brazil. Emerg Infect Dis. 2015;21(10):1885–1886.

4. Hanratty B, Robinson M. Coping with winter bed crises. New surveillance systems might help. BMJ. 1999;319(7224):1511–1512.

5. 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.

6. Centers for Disease Control and Prevention, CDC: Flu Activity & Surveillance; accessed September 14, 2016. http://gis.cdc.gov/grasp/fluview/fluportaldashboard.html.

7. National Aeronautics and Space Administration (NASA): Land Data Assimilation Systems; accessed September 15, 2016. http://ldas.gsfc.nasa.gov/nldas/NLDAS2forcing.php.

8. Xia Y, Mitchell K, Ek M, Sheffield J, Cosgrove B, Wood E, et al. Continental-scale water and energy flux analysis and validation for the North American Land Data Assimilation System project phase 2 (NLDAS-2): 1. Intercomparison and application of model products. J Geophys Res-Atmos. 2012;117(D3).

9. Mitchell KE, Lohmann D, Houser PR, Wood EF, Schaake JC, Robock A, et al. The multi-institution North American Land Data Assimilation System (NLDAS):

Utilizing multiple GCIP products and partners in a continental distributed hydrological modeling system. J Geophys Res-Atmos. 2004;109(D7).

10. Kalnay E, Kanamitsu M, Kistler R, Collins W, Deaven D, Gandin L, et al. The NCEP/NCAR 40-year reanalysis project. B Am Meteorol Soc. 1996;77(3):437–471.

11. NCEP-NCAR Reanalysis project; accessed September 15, 2016. http://iridl.ldeo.columbia.edu/SOURCES/.NOAA/.NCEP-NCAR/.CDAS-1/.DAILY/.Diagnostic/.above_ground/.qa/index.html.

12. Mills HL, Riley S. The Spatial Resolution of Epidemic Peaks. PLoS Comput Biol. 2014;10(4):1–9.

13. Riley P, Ben-Nun M, Armenta R, Linker JA, Eick AA, Sanchez JL, et al. Early Characterization of the Severity and Transmissibility of Pandemic Influenza Using Clinical Episode Data from Multiple Populations. PLoS Comput Biol. 2013;9(5):1–15. doi:10.1371/journal.pcbi.1003064.

14. Riley P, Ben-Nun M, Linker JA, Cost AA, Sanchez JL, George D, et al. Early Characterization of the Severity and Transmissibility of Pandemic Influenza Using Clinical Episode Data from Multiple Populations. PLoS Comput Biol. 2015;11(9):1–15. doi:10.1371/journal.pcbi.1004392.

15. Moss R, Zarebski AF, Dawson P, McCaw JM. Retrospective forecasting of the 2010–2014 Melbourne influenza seasons using multiple surveillance systems. Epidemiology and Infection;145:156–169. doi:doi:10.1017/S0950268816002053.

16. Zarebski AF, Dawson P, McCaw JM, Moss R. Model Selection for Seasonal Influenza Forecasting. Infectious Disease Modelling;2:56–70. doi:https://doi.org/10.1016/j.idm.2016.12.004.

17. Chretien JP, Dylan G, Shaman J, Chitale RA, McKenzie FE. Influenza Forecasting in Human Populations: A Scoping Review. PLoS ONE;9. doi:https://doi.org/10.1371/journal.pone.0094130.

18. Murphy AH, Brown BG. A comparative evaluation of objective and subjective weather forecasts in the united states. J Forecast. 1984;3(4):369–393.

19. Murphy AH, Winkler RL. Probability Forecasting in Meteorology. J Am Stat Assoc. 1984;79(387):489–500.

20. 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;.

21. Yang W, Olson DR, Shaman J. Forecasting Influenza Outbreaks in Boroughs and Neighborhoods of New York City. PLoS Comput Biol. 2016;12(11):e1005201.

22. Axelsen JB, Yaari R, Grenfell BT, Stone L. Multiannual forecasting of seasonal influenza dynamics reveals climatic and evolutionary drivers. Proc Natl Acad Sci U S A. 2014;111(26):9538–9542.

23. Riley P, Ben-Nun M, Turtle JA, Linker J, Bacon DP, Riley S. Identifying factors that may improve mechanistic forecasting models for influenza. bioRxiv. 2017;doi:10.1101/172817.

24. 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):1–13.