Forecasting seasonal outbreaks of influenza

Jeffrey Shaman[a,1] and Alicia Karspeck[b]

[a]Department of Environmental Health Sciences, Mailman School of Public Health, Columbia University, New York, NY 10032; and [b]Climate and Global Dynamics Division, National Center for Atmospheric Research, Boulder, CO 80305

Edited* by Aaron A. King, University of Michigan, Ann Arbor, MI 48109, and approved October 25, 2012 (received for review May 23, 2012)

Influenza recurs seasonally in temperate regions of the world; however, our ability to predict the timing, duration, and magnitude of local seasonal outbreaks of influenza remains limited. Here we develop a framework for initializing real-time forecasts of seasonal influenza outbreaks, using a data assimilation technique commonly applied in numerical weather prediction. The availability of real-time, web-based estimates of local influenza infection rates makes this type of quantitative forecasting possible. Retrospective ensemble forecasts are generated on a weekly basis following assimilation of these web-based estimates for the 2003–2008 influenza seasons in New York City. The findings indicate that real-time skillful predictions of peak timing can be made more than 7 wk in advance of the actual peak. In addition, confidence in those predictions can be inferred from the spread of the forecast ensemble. This work represents an initial step in the development of a statistically rigorous system for real-time forecast of seasonal influenza.

Kalman filter | absolute humidity

Worldwide, influenza produces 3–5 million severe illnesses annually and kills an estimated 250,000–500,000 people (1). In temperate regions, influenza characteristically recurs during winter when absolute humidity levels are low (2, 3), but at present our ability to predict important details of these seasonal influenza outbreaks is limited. Indeed, much public health benefit could be gleaned from early, skillful prediction of the onset, peak, duration, and magnitude of local influenza outbreaks.

Mathematical models of infectious disease transmission have been in use for over a century (4). These models have been developed to study the dynamic properties of disease transmission (5–7), determine the biological characteristics of specific pathogens (8, 9), and analyze historical transmission behavior during documented outbreak events (10).

More recently, infectious disease model simulations have been performed retrospectively in conjunction with statistical filtering methods to provide maximum-likelihood parameter estimation (11, 12) and improved epidemic simulation through time and physical space (13–16). Filtering techniques iteratively update, or adjust, model simulation estimates of the dynamic state, e.g., population infection rates, using real-world observations of that state, as the model is integrated through time. Because the state is only intermittently, or partially, observed—i.e., infections may be observed only for some locations and times, and some state variables, such as population susceptibility rates, may not be observed at all—and because these partial observations themselves contain error, the filter endeavors to balance the relative information contained in the observations and the model simulation. At the same time, the filtering process can also be used to estimate epidemiologically significant parameters within a model.

These same filtering techniques, by constraining the model state and parameters, can potentially be used to enhance the ability of a model to skillfully forecast disease transmission. Heretofore, such attempts at prediction have been applied in a very limited fashion without thorough evaluation of the skill, or uncertainty, of the forecast (17–19).

In theory, infectious disease predictions could be developed, validated, and produced in a fashion similar to that of the forecasts generated for numerical weather prediction (NWP) (20). Much like weather dynamics, infectious disease dynamics are nonlinear and intrinsically chaotic. These characteristics make the evolution of these purely deterministic systems highly sensitive to their current conditions, such that very small differences in the current state can rapidly amplify through time to divergent future outcomes (21, 22). Similarly, for mathematical models representing these nonlinear systems, errors in the estimate of the initial dynamic state of the system grow as these models are integrated into the future. This error growth degrades the accuracy of a forecast the farther into the future it is extended and imposes a formal limit on predictability (23). Such initial error growth is additionally exacerbated by the use of imperfect, simplified model representations of the true system. Together, the nonlinearity of the modeled system and the shortcomings of the model representation lead to a degradation of prediction quality the farther forward in time one forecasts.

To counter this error growth, NWP uses data assimilation techniques, such as filtering, to repeatedly reinitialize dynamic weather models, using the latest available observations and knowledge of the error associated with those observations. In NWP this reinitialization or assimilation is performed frequently (e.g., every 6 h) and the newly initialized model is then integrated forward in time (e.g., 7 d) to create a new set of weather forecasts. Metrics for evaluating the real-time skill of these model–data assimilation system forecasts have been developed to validate these predictions.

To realize such a model–data assimilation prediction system for infectious disease, real-time or near real-time observations of population-level disease status must be available. For influenza, near real-time online-search query estimates of influenza infection rates have recently been developed and validated historically against US Centers for Disease Control and Prevention (CDC) official estimates of influenza-like illness (ILI) (24). At the time of this writing, these Google Flu Trends (GFT) data were available for 28 countries, as well as by district, province, state, or municipality within these countries. These near real-time population-level data make application of model–data assimilation prediction practicable.

Here we apply a data assimilation method called the ensemble adjustment Kalman filter (EAKF) (20) to entrain weekly GFT estimates of ILI (24, 25) into a simple humidity-forced susceptible–infectious–recovered–susceptible (SIRS) mathematical model of influenza (3). The EAKF is a recursive filtering technique that combines observations with a temporally evolving ensemble of model simulations to generate a posterior estimate of the model state (20). This process nudges the ensemble mean toward the observations and simultaneously contracts the ensemble variance, thus constraining the model state and parameters.

Results

Before using the GFT estimates we first validated the functionality of the combined SIRS-EAKF framework, using a synthetic,

Author contributions: J.S. and A.K. designed research, performed research, analyzed data, and wrote the paper.

The authors declare no conflict of interest.

This article is a PNAS Direct Submission. A.A.K. is a guest editor invited by the Editorial Board.

[1]To whom correspondence should be addressed. E-mail: jls106@columbia.edu.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1208772109/-/DCSupplemental.

Downloaded from https://www.pnas.org by 183.230.183.44 on April 24, 2023 from IP address 183.230.183.44.

model-generated record. This preliminary validation of the combined SIRS-EAKF framework with a wholly known synthetic record was performed to determine whether the EAKF appropriately constrains the SIRS model state and parameters.

To conduct the validation, a single run of the humidity-driven SIRS model was initiated with prescribed parameters and initial conditions and integrated from October 1, 1972 into the spring of 1973, using daily absolute humidity (AH) conditions for New York state. This simulation produced a January outbreak of influenza (Fig. 1*A*), which we defined as the "true" outbreak. Prescribed observational error was then added to this truth to create a synthetic, error-laden observational record of influenza infection to be assimilated in the combined SIRS-EAKF framework.

For this first assimilation, an ensemble of 100 simulations was initiated, each with randomly chosen parameters and initial conditions, and integrated through time using the same daily AH record for New York state used to generate the synthetic truth. The synthetic observations were then iteratively assimilated using the EAKF as the model moved forward in time. Ensemble posterior mean estimates of the true state and parameters were well constrained by the combined SIRS-EAKF framework (Fig. 1*A*).

The observed state variable, I, the number of infected persons, is well captured in the mean by the combined SIRS-EAKF framework. The nonobserved state variable, S, the number of susceptible persons, is also well captured, such that before the cessation of the outbreak, there is little bias between the ensemble mean and the truth. Through time the variance of the 100-member ensemble decreases continuously for S (Fig. 1*B*). For I, error growth and ensemble spread occur before the outbreak as individual ensemble members engender premature outbreaks, only to be constrained by the observations. During the outbreak, I undergoes more intense error growth as outbreaks are then supported by the observations.

Epidemiologically significant parameters within the SIRS model, including the mean infection period, D, and $R_{0\max}$, which sets an upper bound on the transmissibility of the virus, are also constrained by the EAKF. By the end of the true outbreak, the ensemble mean estimates of these parameters are close to the true values used to generate the synthetic observations.

For comparison with the EAKF approach, we alternatively tested a particle filter method, using the same SIRS model and synthetic observations (*SI Methods*). Unlike the EAKF, particle filters make no assumptions on the linearity underlying the model or the mapping from state to observational space. However, particle filter methods are prone to particle degeneracy, which can lead to model divergence and poor state estimation (26, 27). The particle filter method tested performed reasonably well; however, the EAKF method provided better and more computationally efficient estimates of the unobserved state, S, and parameters (Fig. S1). This better performance may stem from the fact that the EAKF approach maintains greater spread during the preoutbreak period, which may enable better state and parameter estimation once the outbreak commences.

We have also elected to present the EAKF as the method of data assimilation because the EAKF can be readily applied to higher-dimensional systems, unlike particle filter methods (28). Whereas the current system is only dimension 6 (two state variables and four parameters), we anticipate use of higher-dimensional influenza model systems that include multiple strains, age stratification, and geographic structure in the future.

Additional simulations were performed to gauge the sensitivity of the combined SIRS-EAKF framework to changes of the ensemble size, the random sampling of initial ensemble member states, the frequency of observations (e.g., every second day, weekly, etc.), and the prescribed error variance used to generate the synthetic observations from the truth (Figs. S2–S4). These sensitivity tests indicated that the SIRS-EAKF framework provides a robust estimate of the truth provided the ensemble is of sufficient size (100 or more members), the observations are available with sufficient frequency (at least every 10 d), and the observational error variance is not too large.

We next applied the combined SIRS-EAKF framework to perform retrospective simulation of influenza infection in New York City during the 2004–2005 and 2007–2008 seasons, using GFT weekly estimates of ILI (Figs. S5–S7). A small level of multiplicative inflation was included in these simulations to avoid filter divergence (29, 30) and better replicate the GFT observations (*SI Methods*). This need for inflation, which is commonly used during data assimilation, in part stems from the fact that the GFT observations reflect real-world transmission dynamics, which include structures and processes, such as spatial heterogeneity and preferential mixing, that are not represented in the SIRS model and contribute to model error.

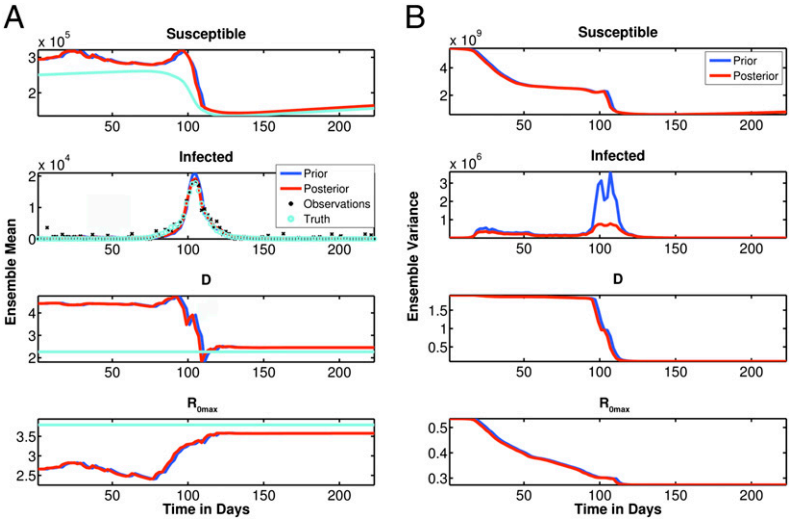

Fig. 1. Results for SIRS-EAKF scheme using model-generated truth plus prescribed observation error variance. (*A*) Time series of prior and posterior ensemble mean and truth for the susceptible (*S*) and infected (*I*) states and the model parameters *D* (mean infectious period) and $R_{0\max}$ (a transmissibility parameter). Also shown for the infected state are the synthetic observations (truth plus observation error). (*B*) Time series of prior and posterior ensemble variance for *S*, *I*, *D*, and $R_{0\max}$.

Downloaded from https://www.pnas.org by 183.230.183.44 on April 24, 2023 from IP address 183.230.183.44.

Two hundred-member ensembles of the SIRS-EAKF framework were run with observed daily AH for New York City as forcing. Each ensemble member was initialized with a unique sample of state variables and SIRS parameters in September before the wintertime influenza season. The ensemble was then integrated through time and GFT ILI observations were assimilated using the EAKF to generate a posterior estimate of infection rates. This process was repeated 250 times, each time with a different 200-member ensemble of initial parameter and state variable combinations. With a small amount of inflation, the average ensemble mean posterior estimates match the GFT weekly estimates of ILI well (Fig. 2). The complicated outbreak structure during the 2004–2005 influenza season, which includes three individual peaks, is represented by the SIRS-EAKF framework. The model–data assimilation system also captures the long rise and single peak of infection during 2007–2008.

A retrospective forecast for the 2007–2008 influenza season in New York City was next generated. As in preceding simulations, after the assimilation of a weekly GFT observation, the model posterior was integrated forward 7 d to the next GFT observation; however, in addition, the same model posterior was also integrated forward 300 d without any further EAKF constraint, although with perfect knowledge of the daily AH conditions. These second longer runs constitute retrospective forecasts; i.e., after assimilation of a GFT ILI estimate in near-real time and consequent adjustment of SIRS ensemble member parameters and state variables by the EAKF process, the model is integrated into the future, using those new parameters and the current state, to generate anticipated possible outbreak outcomes.

The SIRS-EAKF system can be evaluated for its predictive skill for any number of outcomes, such as outbreak magnitude, duration, or onset. Here we choose to focus on one metric: the timing of the peak of the outbreak—i.e., the week during the influenza season in which the highest GFT ILI estimate is recorded (e.g., week 25, or the week ending February 17 for the 2007–2008 season). The retrospective forecasts indicate that model predictions of influenza outbreak peak converge toward the observed peak in the weeks before the actual event (Fig. 3 *A* and *B*). In fact, the ensemble mode predicted peak is within 1 wk of being correct 5 wk in advance of the observed peak. In addition, the spread of peak predictions among the 200-member ensemble decreases as the forecast nears the observed peak. This decrease in ensemble spread can be seen in weekly histograms of predicted peak timing, as well as in the ensemble variance of these same predictions.

The percentage of ensemble members predicting the peak timing within ±1 wk increases dramatically beginning in week 18 as model spread decreases (Fig. 3 *C* and *D*). This relationship indicates that, for a real-time prediction performed in the absence of knowledge of when the actual observed peak will occur, a measure of confidence in the prediction can be made on the basis of the ensemble spread. That is, decreased ensemble spread will coincide with increased confidence in the prediction.

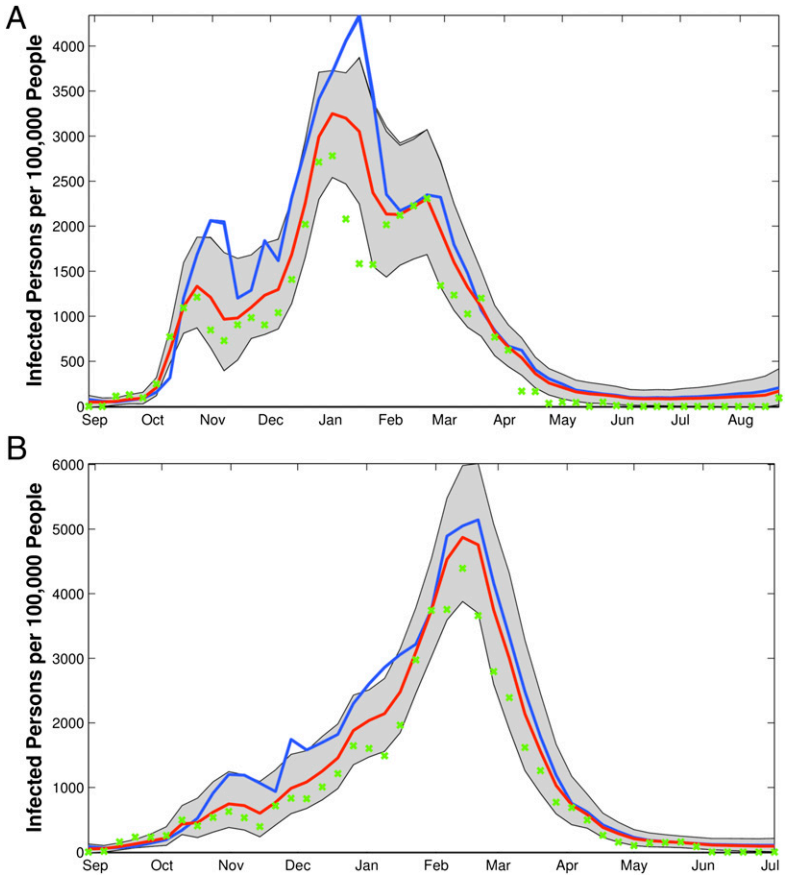

Fig. 2. Results for the SIRS-EAKF scheme using weekly GFT estimates of ILI for New York City during the 2004–2005 and 2007–2008 influenza seasons. Two hundred fifty 200-member ensembles were run for each season, each with a different suite of initial conditions. Multiplicative inflation of $\lambda = 1.02$ is applied. SIRS model simulations were forced with observed daily specific humidity conditions for New York City. (*A*) The 2004–2005 influenza season weekly GFT ILI estimates (green) and SIRS-EAKF ensemble mean prior (blue) and ensemble mean posterior (red) of infected (*I*) averaged for all 250 ensembles. Also shown are the 10th and 90th percentiles of the ensemble posterior *I* (black lines) averaged for all 250 ensembles. The range between the average 10th and 90th percentiles is shaded gray. (*B*) Same as *A* but for the 2007–2008 influenza season.

Downloaded from https://www.pnas.org by 183.230.183.44 on April 24, 2023 from IP address 183.230.183.44.

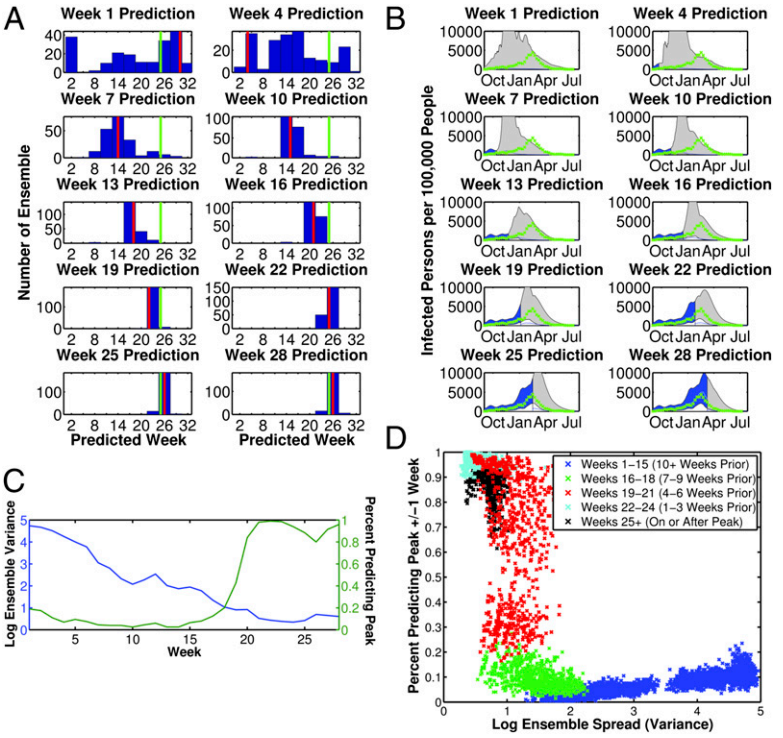

Fig. 3. Results for SIRS-EAKF retrospective forecasting using GFT ILI estimates and observed AH conditions for the 2007–2008 season in New York City. Retrospective assimilation began September 2, 2007, and the first forecast occurred after assimilation of the week 1 GFT ILI estimate (representing September 2–8, 2007). Inflation was set at $\lambda = 1.02$. (*A–C*) Results for a single 200-member SIRS-EAKF run with prediction for 300 d following each assimilation, using the new posterior state and parameter values. (*A*) Histogram of ensemble forecast peak timing for predictions initiated at the end of weeks 1, 4, 7, 10, 13, 16, 19, 22, 25, and 28 (blue). Also shown are the observed peak (green, week 25) and the ensemble mode (red). Note that the peak for any given ensemble member may occur before the forecast week. (*B*) Time series of the ensemble mean posterior and spread (blue), the mean prediction and spread (gray), and the observations (green) for predictions initiated at the end of weeks 1, 4, 7, 10, 13, 16, 19, 22, 25, and 28. (*C*) Time series of the ensemble predicted peak variance log transformed (blue) and percentage of the 200-member ensemble predicting the observed peak (week 25) within ±1 wk (green). Note that for the forecast made the week after the week 25 peak, the percentage of ensemble members accurately predicting the peak drops as some ensemble members erroneously forecast still more infections in the future; however, after 2 wk of continued decline in GFT observations the forecasts "recognize" the abatement and no longer forecast any future resurgence of infection. (*D*) Percentage of the 200-member ensemble predicting the peak within ±1 wk of the actual peak (week 25) for each weekly prediction (weeks 1–28) for each of 250 SIRS-EAKF assimilation runs plotted as a function of the ensemble predicted peak variance log transformed. Each of the two hundred fifty 200-member SIRS-EAKF assimilation runs was initiated with a different suite of initial state and parameter combinations. Color coding denotes the week the prediction was made relative to the actual week 25 peak: 10 wk or more prior (blue), 7–9 wk prior (green), 4–6 wk prior (red), 1–3 wk prior (cyan), and on or after the peak (black).

To further codify this finding and verify its generality across a larger set of seasonal outbreaks, we next ran repeated retrospective ensemble forecasts for the 2003–2004 through 2008–2009 influenza seasons in New York City. These forecasts were run in 200-member ensembles, using two different approaches: (*i*) with perfect knowledge of future daily AH conditions for only the first 5 d of the influenza forecast—these conditions would be locally available from numerical weather forecasts—and daily climatological AH conditions for the remaining 295 d of each forecast; and (*ii*) with perfect knowledge of future daily AH conditions for the first 5 d of the forecast followed by 295 d of AH climatology, but with only the model infected state constrained by the EAKF, i.e., the posterior parameters and other state variable are reset to an initial distribution before commencing each forecast.

The first approach is how a real-time forecast could be run for this SIRS-EAKF framework: with full use of the EAKF constraint and only limited knowledge of future AH conditions. The difference between the outcomes of the first and second approaches indicates the forecast improvement due to full EAKF constraint of all variables and parameters.

The same relationship between the accuracy of the ensemble mode predicted peak and the spread of the predictions within the ensemble holds for these more numerous retrospective forecasts (Fig. S8). A clearer picture of this predictive skill emerges when the forecasts are grouped by how many weeks into the future the ensemble mode peak is predicted (*SI Methods*). The two different forecast approaches have comparable predictive skill when the ensemble peak mode is predicted to be 1–3 wk in the future (Fig. 4 and Fig. S9); however, for ensemble peak modes forecast greater than 3 wk in the future, only the first forecast form, which makes full use of the EAKF state variable and parameter constraint, has demonstrated skill (Fig. 4). The second forecast form, which lacks full EAKF constraint, has no forecast skill for mode peak predictions greater than 3 wk; indeed, without full EAKF constraint mode peak predictions rarely occur more than 3 wk in the future.

Results from the first forecast form, which uses climatological AH conditions and full EAKF constraint of model variables and parameters, can now be used to explore how a real-time influenza forecast would be interpreted (Fig. 4, red lines). On the basis of the ensemble spread and how many weeks in the future the outbreak peak is predicted, confidence in a given prediction, or, conversely, the forecast uncertainty, can be assigned on the basis of the probabilities shown in Fig. 4. For instance, an ensemble mode peak predicted for 7 wk in the future with a log-transformed ensemble variance of 2.5 wk squared is expected to be accurate within ±1 wk about 40% of the time.

Discussion

The results presented here demonstrate that weekly, local predictions of influenza risk, with estimates of forecast certainty, can be made in real time. These forecasts can be performed using GFT ILI data at municipal, state, and country levels and used to help inform public health decisions including vaccine allocation and antiviral drug distribution.

In the future, as more years of GFT ILI estimate data and more locations are entrained into this analysis, a more robust estimate of the relationship between predictions of peak timing and ensemble variance will be developed. This will further codify the relationships presented in Figs. 3 and 4 and improve prediction skill. Forecasts can also be developed for other outcome metrics, such as peak magnitude and outbreak duration. Predictions of large local maxima, rather than the one seasonal peak we have used here, might also prove useful; such a metric may improve prediction skill, as some years (e.g., 2005 in New York City) have no distinct, substantive peak. Additional information, such as influenza strain or school calendar, might also be used to further optimize predictions.

Information on influenza strain may be particularly important. The three peaks during 2004–2005 (Fig. 2*A*) stem principally from differently timed outbreaks of three distinct influenza subtypes in

Downloaded from https://www.pnas.org by 183.230.183.44 on April 24, 2023 from IP address 183.230.183.44.

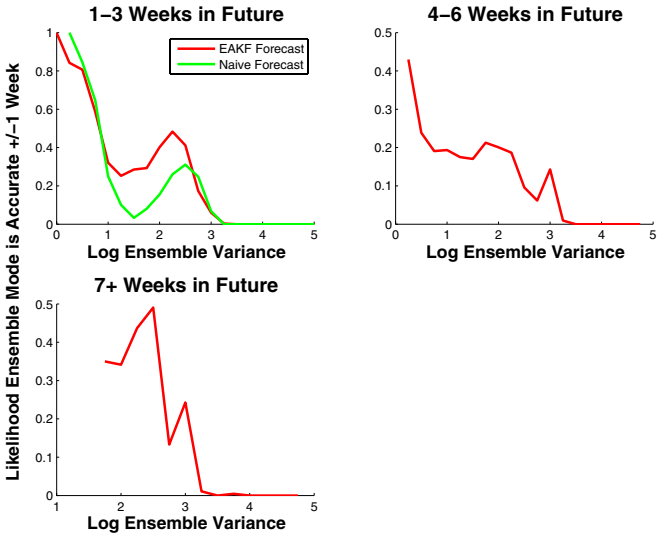

Fig. 4. Results for SIRS-EAKF retrospective forecasting using GFT ILI estimates for the 2003–2004 through 2008–2009 seasons in New York City. Retrospective assimilations began in either August or October. Two hundred 200-member SIRS-EAKF assimilations were performed for each season and each of these start months, excepting 2003–2004 for which only October was run as August 2003 GFT ILI estimates are unavailable. Inflation was set at $\lambda = 1.02$. Forecasts were run for 300 d following each assimilation, using the new posterior state and parameter values. The forecasts are grouped using 0.25-wk squared bins of ensemble predicted peak variance log transformed. The proportions of ensemble mode predicted peak within each bin that are within ±1 wk of the observed GFT ILI estimate for that season are plotted as a function of the ensemble predicted peak variance log transformed. The different subplots group the predictions by lead time; i.e., "4–6 wk in the future" indicates the forecast mode, whether correct or not, was predicted to be 4–6 wk in the future. The different color lines are for (*i*) the first forecast form, 5 d of true AH conditions followed by daily 1979–2002 AH climatology and full EAKF constraint (red); and (*ii*) the second forecast form, AH as in the first form but no EAKF constraint of model parameters and the susceptible variable (green).

circulation that season. The SIRS model, as run here with only one strain, is not equipped to produce separate parameter and state estimates for each subtype, which would be a truer representation of actual seasonal dynamics. A model that accounts for and assimilates infection data by strain type would likely improve prediction skill and uncertainty estimates for each strain and the season as a whole; however, at present, such partitioned data are not available in real time. Alternatively, a segregated analysis of predictions from seasons with one predominant strain, vs. those with more than one strain, might improve prediction and uncertainty estimates for those single-strain seasons, simply by aggregating those "cleaner" years.

In weather and climate, forecasts made with a suite of models are generally more robust than forecasts made with any single model (31, 32). This finding motivates the development and use of additional model–data assimilation influenza forecast frameworks to be used in conjunction with the approach presented here. Other models might be spatially distributed, discrete, or age stratified or include stochasticity. Alternate assimilation approaches should also be tested further, including additional particle filtering methods (18, 26, 27) or alternate ensemble Kalman filtering forms (33, 34). Indeed, no one model or assimilation method need be used to the exclusion of all others, and it is not our intention to advocate exclusive use of either the SIRS model or the EAKF.

The general model–data assimilation framework presented here is flexible and adaptive and could be used to develop predictions for other seasonally recurring respiratory diseases, such as respiratory syncytial virus and rhinovirus, should real-time estimates of these pathogens become available. The forecasts developed here indicate that we will soon reach an era when reliable forecasts of some infectious pathogens are as commonplace as weather predictions.

Methods

Description of the SIRS Model. The model used for this study is a perfectly mixed, absolute humidity-driven SIRS construct. The form is similar to that of the model described in Shaman et al. (3). The SIRS model equations are

$$\frac{dS}{dt} = \frac{N-S-I}{L} - \frac{\beta(t)IS}{N} - \alpha \quad [1]$$

$$\frac{dI}{dt} = \frac{\beta(t)IS}{N} - \frac{I}{D} + \alpha, \quad [2]$$

where S is the number of susceptible people in the population, t is time in years, N is the population size, I is the number of infectious people, $N - S - I$ is the number of resistant individuals, $\beta(t)$ is the contact rate at time t, L is the average duration of immunity, D is the mean infectious period, and α is the rate of travel-related import of influenza virus into the model domain. The basic reproductive number, which is the number of secondary infections the average infectious person would produce in a fully susceptible population, at time t is related to the contact rate through the expression $R_0(t) = \beta(t)D$.

AH modulates transmission rates within this model by altering $R_0(t)$ through an exponential relationship similar to how AH has been shown to affect both influenza virus survival and transmission in laboratory experiments (2),

$$R_0(t) = \exp(a \times q(t) + b) + R_{0\min}, \quad [3]$$

where $a = -180$, $b = \log(R_{0\max} - R_{0\min})$, $R_{0\max}$ is the maximum daily basic reproductive number, $R_{0\min}$ is the minimum daily basic reproductive number, and $q(t)$ is the time-varying specific humidity, a measure of AH. The value of a is estimated from the laboratory regression of influenza virus survival upon AH (2).

This SIRS model, as run previously, reproduces both the observed 1972–2002 seasonal cycle of excess pneumonia and influenza mortality within the United States and the association between negative AH anomalies and the onset of local influenza outbreaks (3). It is a strong starting point for the proposed project because it has been validated against datasets of historical influenza outbreaks, and its inclusion of environmental forcing (i.e., AH conditions) may help the model better simulate outbreak dynamics.

For this study, the model was run continuously and without stochasticity. That is, fractional persons were generated within model state classes, and transitions between model states were calculated deterministically directly from Eqs. **1** and **2**. The complete temporal evolution of the SIRS model is fully described by the model equations (Eqs. **1**–**3**), initial conditions, and AH boundary forcing. Simulations were performed with an influenza importation rate, α, of 0.1 infections per day (1 infection every 10 d). Only one influenza strain was simulated per season, and partial and cross-immunities were not represented.

Further description of the model initialization and forcing for individual SIRS-EAKF experiments is provided below.

Description of the EAKF. Generally, sequential ensemble filtering can be viewed as the problem of estimating the probability of the system state at a given time (Z_t) conditional on observations (y_t) taken up until and including time t. In the case of the SIRS model described above, the state is composed of the number of individuals in a population that are susceptible and infectious at a given time and the set of independent model parameters; that is, $Z_t = (S_t, I_t, R_{0\max}, R_{0\min}, L, D)$ and y_t is a time series representing observations (e.g., the Google Flu Trends ILI estimates).

Bayes' rule provides a target for the update of the system state given an observation:

$$p(Z_t|y_t, y_{t-1}, \ldots) \propto p(y_t|Z_t)p(Z_t|y_{t-1}, \ldots). \quad [4]$$

Here the first term on the right-hand side is the likelihood of observing the data given the state and the second term is the prior distribution of the system state. The updated distribution (the left-hand side of Eq. **4**) is called the posterior. Generally, Kalman filters fall into a class of filters that assumes normality of both the likelihood and prior distributions during an update. This assumption allows parameterization of these distributions in terms of the first two

Downloaded from https://www.pnas.org by 183.230.183.44 on April 24, 2023 from IP address 183.230.183.44.

moments only (mean and covariance). However, in ensemble filtering, one need never form the covariance of the prior or posterior; instead we have a finite ensemble of states that are samples from these distributions. The model mean and covariance are computed directly from the ensemble.

Knowledge of these prior moments, as well as observations and their error, allows for the computation of the mean and covariance of the posterior. The method of transforming the ensemble members of the prior into ensemble members of the posterior is what distinguishes different types of ensemble filters. The EAKF [see Anderson (20) for algorithmic details] adjusts the prior ensemble members such that their new moments match the target moments of the posterior predicted by Bayes' theorem. The method is sequential in that given samples from the posterior, the SIRS model can be used to integrate each ensemble member forward in time to the point at which new observations become available. The update is then repeated.

In models that consist of multiple prognostic variables—in this instance, the state variables S and I, as well as the model parameters—covariant relationships between variables arise naturally from the dynamics of the system. Mathematically rigorous methods for assimilating real-world data into numerical models rely on knowledge of these intervariable relationships. This information allows "balanced" adjustments to the entire model solution, even when some of the prognostic variables, such as S, and the parameters are not directly observable quantities. Ensemble filters store all of the information about the state variable and model parameter interrelationships in the form of multiple model solutions (the ensemble), all of which are possible realizations given past measurements.

In Kalman filtering these intervariable relationships are assumed linear (i.e., jointly distributed by a multivariate Gaussian). The EAKF inherits this assumption, but the adjustment operates only on the first two moments of the prior distribution, leaving the higher-order moments unchanged. In the case of a system in which intervariable relationships are nearly linear (such as the SIRS model), the ensemble filter can have high utility even when it is not strictly optimal. Ensemble Kalman filtering (like particle filtering) also has the attractive quality that the algorithm can be implemented independently of the dynamic model. In contrast to variational methods, which require adjoints or parametric covariance forms, ensemble methods use intervariable relationships that are statistically diagnosed from the ensemble of model solutions.

Generation of Synthetic Truth and Observations. A single combination of initial conditions and SIRS model parameters was used to generate a synthetic "truth," which was then used to validate the model–data assimilation scheme. The truth run had parameters $L = 3.86$ y; $D = 2.27$ d; $R_{0\max} = 3.79$; and $R_{0\min} = 0.97$. This combination of parameters was chosen because it produced a good representation of 1972–2002 excess observed pneumonia and influenza mortality in New York State.

To generate the synthetic truth, the SIRS model was run with the chosen parameter combination and forced with New York State daily absolute humidity conditions from October 1, 1972 until May 15, 1973. A total model population of 500,000 was used, and initial susceptibility was set at 250,000 persons with 1 person initially infected. Only one strain of influenza was simulated. This simulation produced a single outbreak that peaked during January 1973. A time series of the number of infected people, I, was formed by sampling the simulation every 2 d. Synthetic observations were then generated by adding to this synthetic truth a normally distributed random observational error with mean 0 and SD, σ_0, where the SD is proportional to the percentage of the infected population (i.e., $\sigma_0 = 1,000 \times I/N$). The resulting time series was then used for assimilation in the combined SIRS-EAKF framework.

Initialization of the SIRS-EAKF System. The SIRS-EAKF system is initialized with an ensemble of state vectors, Z_0. The values of these state vectors, which include the SIRS variables S and I and the parameters L, D, $R_{0\max}$, and $R_{0\min}$, are drawn from a broad distribution of possible variable/parameter combinations. To generate this distribution, 100,000 simulations of the SIRS model forced with New York State AH were integrated from 1972 to 2002 (31 y). Each of these 100,000 integrations was performed using a unique set of parameters. The parameter ranges for this initial random selection were $2 \leq L \leq 10$, $2 \leq D \leq 7$, $1.3 \leq R_{0\max} \leq 4$, and $0.8 \leq R_{0\min} \leq 1.3$, as in ref. 3, and combinations were selected using a Latin hypercube sampling strategy. The state vectors for a given SIRS-EAKF ensemble were then drawn randomly from the collection of all possible October 1 combinations.

For this work, each model initialization and seasonal forecast is approached naively with no information gleaned from the preceding season. In the future, the initial distribution of parameter values and S might be informed by prior season outbreaks or knowledge of circulating strains, which might allow earlier constraint and prediction leads.

ACKNOWLEDGMENTS. Funding was provided by US National Institutes of Health (NIH) Grant GM100467 (to J.S. and A.K.) and the NIH Models of Infectious Disease Agent Study program through Cooperative Agreement 1U54GM088558 (to J.S.), as well as by National Institute on Environmental Health Sciences Center Grant ES009089 (to J.S.) and the Research and Policy for Infectious Disease Dynamics (RAPIDD) program of the Science and Technology Directorate, US Department of Homeland Security (to J.S.).

1. WHO (2009) *Influenza (Seasonal), Fact Sheet Number 211*. Available at http://www.who.int/mediacentre/factsheets/fs211/en/index.html. Accessed April, 2012.
2. Shaman J, Kohn MA (2009) Absolute humidity modulates influenza survival, transmission and seasonality. *Proc Natl Acad Sci USA* 106:3243–3248.
3. Shaman J, Pitzer VE, Viboud C, Grenfell BT, Lipsitch M (2010) Absolute humidity and the seasonal onset of influenza in the continental US. *PLoS Biol,* e1000316. doi:10.1371/journal.pbio.1000316.
4. Ross R (1911) Some quantitative studies in epidemiology. *Nature* 87:466–467.
5. Macdonald G (1952) The analysis of the sporozoite rate. *Trop Dis Bull* 49:569–586.
6. Anderson RM, May RM (1979) Population biology of infectious diseases: Part I. *Nature* 280:361–367.
7. May RM, Anderson RM (1979) Population biology of infectious diseases: Part II. *Nature* 280:455–461.
8. Macdonald G (1952) The analysis of equilibrium in malaria. *Trop Dis Bull* 49:813–1129.
9. Keeling MJ, Grenfell BT (1997) Disease extinction and community size: Modeling the persistence of measles. *Science* 275:65–67.
10. Mills CE, Robins JM, Lipsitch M (2004) Transmissibility of 1918 pandemic influenza. *Nature* 432:904–906.
11. Ionides EL, Bretó C, King AA (2006) Inference for nonlinear dynamical systems. *Proc Natl Acad Sci USA* 103:18438–18443.
12. He D, Ionides EL, King AA (2010) Plug-and-play inference for disease dynamics: Measles in large and small populations as a case study. *J R Soc Interface* 7:271–283.
13. King AA, Ionides EL, Pascual M, Bouma MJ (2008) Inapparent infections and cholera dynamics. *Nature* 454:877–881.
14. Bretó C, He D, Ionides EL, King AA (2009) Time series analysis via mechanistic models. *Ann Appl Stat* 3:319–348.
15. Krishnamurthy A, Cobb L, Mandel J, Beezley J (2010) Bayesian tracking of emerging epidemics using ensemble optimal statistical interpolation (EnOSI). *Sect Stat Epidem* arXiv:1009.4959.
16. Mandel J, Beezley JD, Cobb L, Krishnamurthy A (2010) Data driven computing by the morphing fast Fourier transform ensemble Kalman filter in epidemic spread simulations. *Proc Comp Sci* 1:1215–1223.
17. Rhodes CJ, Hollingsworth TD (2009) Variational data assimilation with epidemic models. *J Theor Biol* 258:591–602.
18. Dukic VM, Lopes HF, Polson N (2010) Tracking flu epidemics using Google Flu Trends data and a state-space SEIR model. *J Am Stat Assoc*, 10.1080/01621459.2012.713876.
19. Ong JBS, et al. (2010) Real-time epidemic monitoring and forecasting of H1N1-2009 using influenza-like illness from general practice and family doctor clinics in Singapore. *PLoS ONE*, e10036. doi:10.1371/journal.pone.0010036.
20. Anderson JL (2001) An ensemble adjustment Kalman filter for data assimilation. *Mon Weather Rev* 129:2884–2093.
21. Lorenz EN (1963) Deterministic nonperiodic flow. *J Atmos Sci* 20:130–141.
22. Strogatz SH (1998) *Nonlinear Dynamics and Chaos* (Addison-Wesley, Reading, MA).
23. Lighthill J (1986) The recently recognized failure of predictability in Newtonian dynamics. *Proc R Soc Lond A Math Phys Sci* 407:35–50.
24. Ginsberg J, et al. (2009) Influenza epidemics using search engine query data. *Nature* 457:1012–1014.
25. Google Flu Trends (2012) Available at http://www.google.org/flutrends. Accessed December, 2011.
26. Arulampalam MS, Maskell S, Gordon N, Clapp T (2002) A tutorial on particle filters for online nonlinear/non-Gaussian Bayesian tracking. *IEEE Trans Signal Process* 50:174–188.
27. van Leeuwen PJ (2009) Particle filtering in geophysical systems. *Mon Weather Rev* 137:4089–4114.
28. Snyder C, Bengtsson T, Bickel P, Anderson J (2008) Obstacles to high-dimensional particle filtering. *Mon Weather Rev* 136:4629–4640.
29. Anderson JL, Anderson SL (1999) A Monte Carlo implementation of the nonlinear filtering problem to produce ensemble assimilations and forecasts. *Mon Weather Rev* 127:2741–2758.
30. Whitaker JS, Hamill TM (2002) Ensemble data assimilation without perturbed observations. *Mon Weather Rev* 130:1913–1924.
31. Krishnamurti TN, et al. (1999) Improved weather and seasonal climate forecasts from multi-model superensemble. *Science* 285:1548–1550.
32. Palmer TN, et al. (2005) Representing model uncertainty in weather and climate prediction. *Annu Rev Earth Planet Sci* 33:163–193.
33. Anderson JL (2010) A non-Gaussian ensemble filter update for data assimilation. *Mon Weather Rev* 138:4186–4198.
34. Lei J, Bickel P, Snyder C (2010) Comparison of ensemble Kalman filters under non-Gaussianity. *Mon Weather Rev* 138:1293–1306.

Downloaded from https://www.pnas.org by 183.230.183.44 on April 24, 2023 from IP address 183.230.183.44.