Sandia National Laboratories

Disease forecasting using open-source indicators

S. Lefantzi[1], J. Ray[1], G. Lambert[2], P. Finley[2] and H. Smith[2]

[1]Sandia National Laboratories, Livermore, CA, and [2]Sandia National Laboratories, Albuquerque, NM

OBJECTIVE

To explore the use of open-source indicators (OSI) for epidemiological nowcasting and forecasting

- OSI datastreams: weather datasets, Google Flu Trends (GFT), Google Dengue Trends (GDT) and data collected by Healthmap.org (HM)
- Nowcasting refers to using OSI to compensate for the 1-4 week lag needed to collect and disseminate information via public health reports

Demonstration problems

- Nowcast swine flu in US and France using Healthmap data
- Forecast dengue in India, using GDT and meteorology
- Predict influenza activity in the San Francisco Bay Area (SFBA) using data assimilation (GFT and meteorology)

BACKGROUND

Epidemiological activity is often reflected in our online behavior

- Outbreaks lead to media reports, which are often posted online
    - These reports are automatically collected and classified by web-crawlers and organizations e.g., Healthmap.org
- Outbreaks lead to spikes in online searches about the disease or the outbreak in question
    - GFT monitors influenza-related searches and has been shown to nowcast CDC's FluNet data (Ginsberg et al, 2008); Google Dengue Trends (GDT) is predictive in some countries (Chan et al, 2011)
    - ED data, along with GFT and meteorology, have been used to forecast hospital ED loads in Baltimore (Dugas et al, 2013)
- Data assimilation (DA) has been used to forecast disease activity
    - Data assimilation refers to using OSI to calibrate a "mechanistic" disease model e.g., SEIR models, to observations and using it to forecast future outbreak evolution
    - Absolute humidity is known to affect viability of influenza virus and is a leading indicator of flu outbreaks (Shaman et al, 2010)
    - GFT and humidity data were assimilated to calibrate a SIR model of flu and provide accurate forecasts of peaks of seasonal influenza outbreaks, (Shaman and Karspeck, 2012; Shaman et al, 2013)

Research questions

- Can Healthmap.org data be used to nowcast influenza activity?
    - Its quality is lower than GFT, but is available even where Google's and Internet's penetration is low
- Can the seasonal nature of Indian dengue outbreaks (after the monsoons) be predicted, guided by precipitation levels?
- Can GFT be used to provide regional forecasts i.e., outside the municipalities tracked by it?

PREDICTIONS USING ARX MODELS

Exploiting HM's correlation with syndromic surveillance data using ARX models

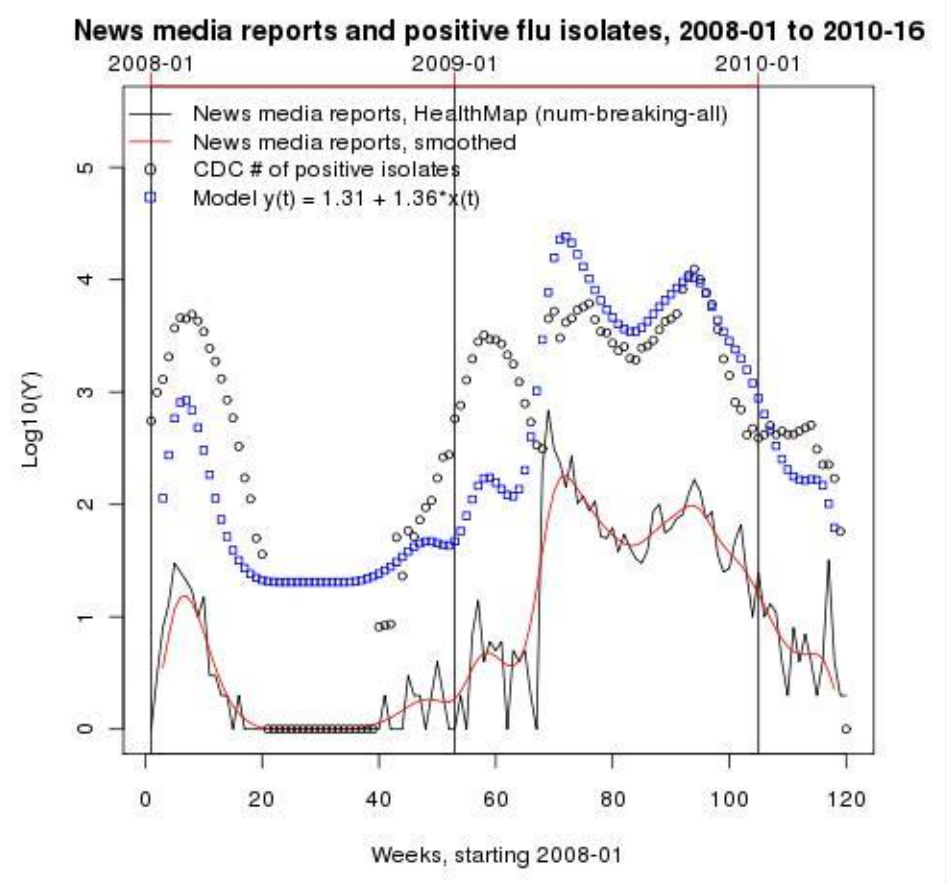
- CDC's FluNet reports # of ILI cases per 100,000 physician visits; HM can supply the weekly totals of flu-related media articles
- Propose an ARX (Auto-Regressive with eXogenous inputs) model

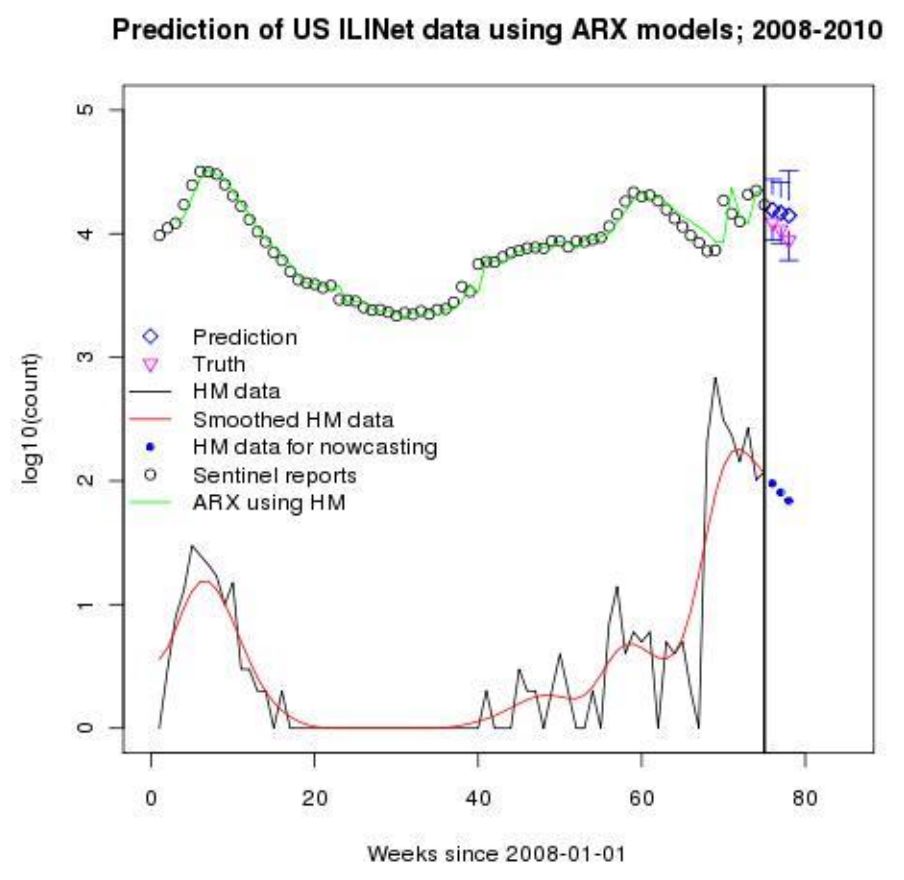
$$y_i = \sum_{j=1}^{N} \alpha_j y_{i-j} + \sum_{k=1}^{M} \beta_k x_{i-k} + \varepsilon$$

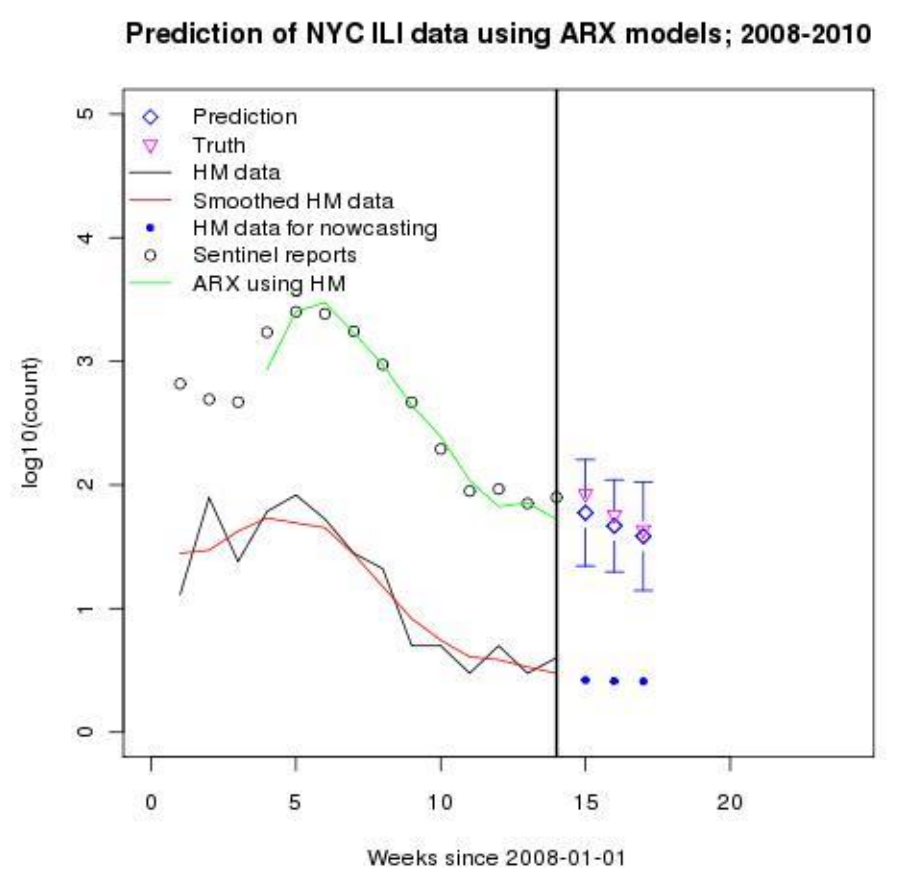
where y_i is the log-transformed FluNet report and x_i is the log-transformed HM data for week *i*

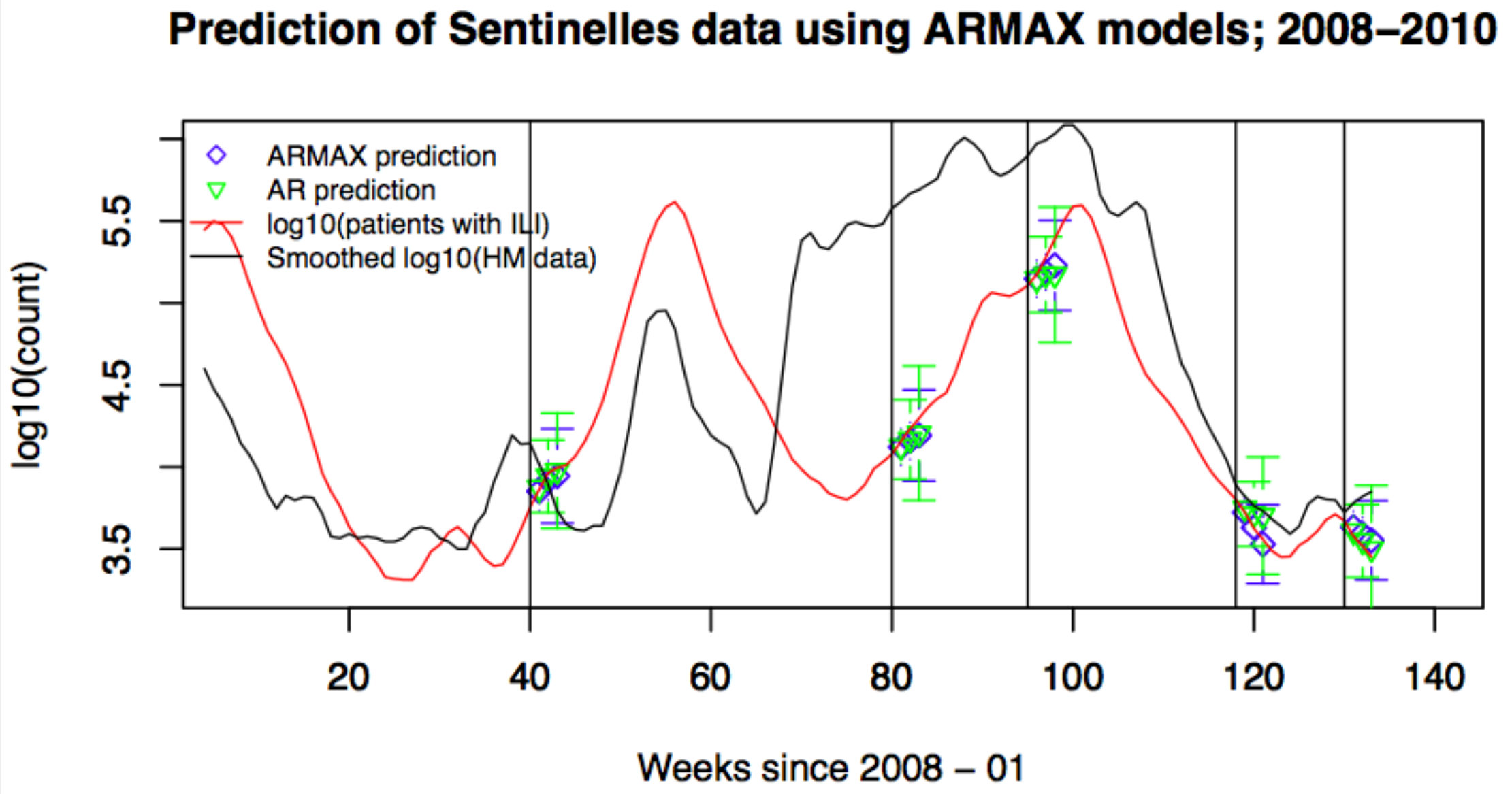
- α_i, β_i, M & N obtained by fitting to historical data
- Exploits the smooth nature of syndromic surveillance time-series and its cross-correlation with HM data

- ARX can be successfully used at the country-level and the city-level
    - Tested for US & NYC during the swine flu outbreak
- ARX successfully applied to Sentinelles data (France) during 2009 swine flu outbreak
    - France escaped swine flu, but not the media storm. ARX models detected the lack of correlation between Sentinelles and HM data and ignored HM (see figure below)


News media reports and positive flu isolates, 2008-01 to 2010-16

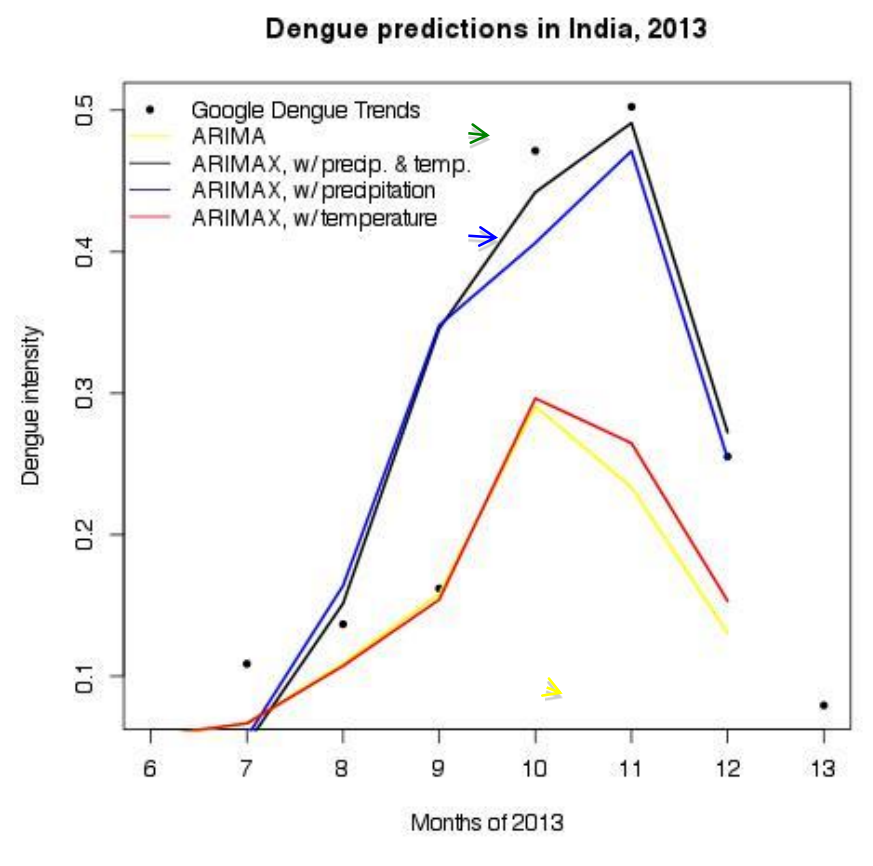

Prediction of US ILINet data using ARX models; 2008-2010


Prediction of NYC ILI data using ARX models; 2008-2010


Prediction of Sentinelles data using ARMAX models; 2008–2010

Prediction of Indian dengue outbreaks using ARX models

- Indian post-monsoon dengue outbreaks are correlated with precipitation (with a lag to establish the mosquito population)
- Dengue activity can be tracked using GDT
- Arrival of the monsoons & rainfall is routinely predicted accurately
- ARX models (rainfall as the exogenous variable) can predict GDT levels for an entire season

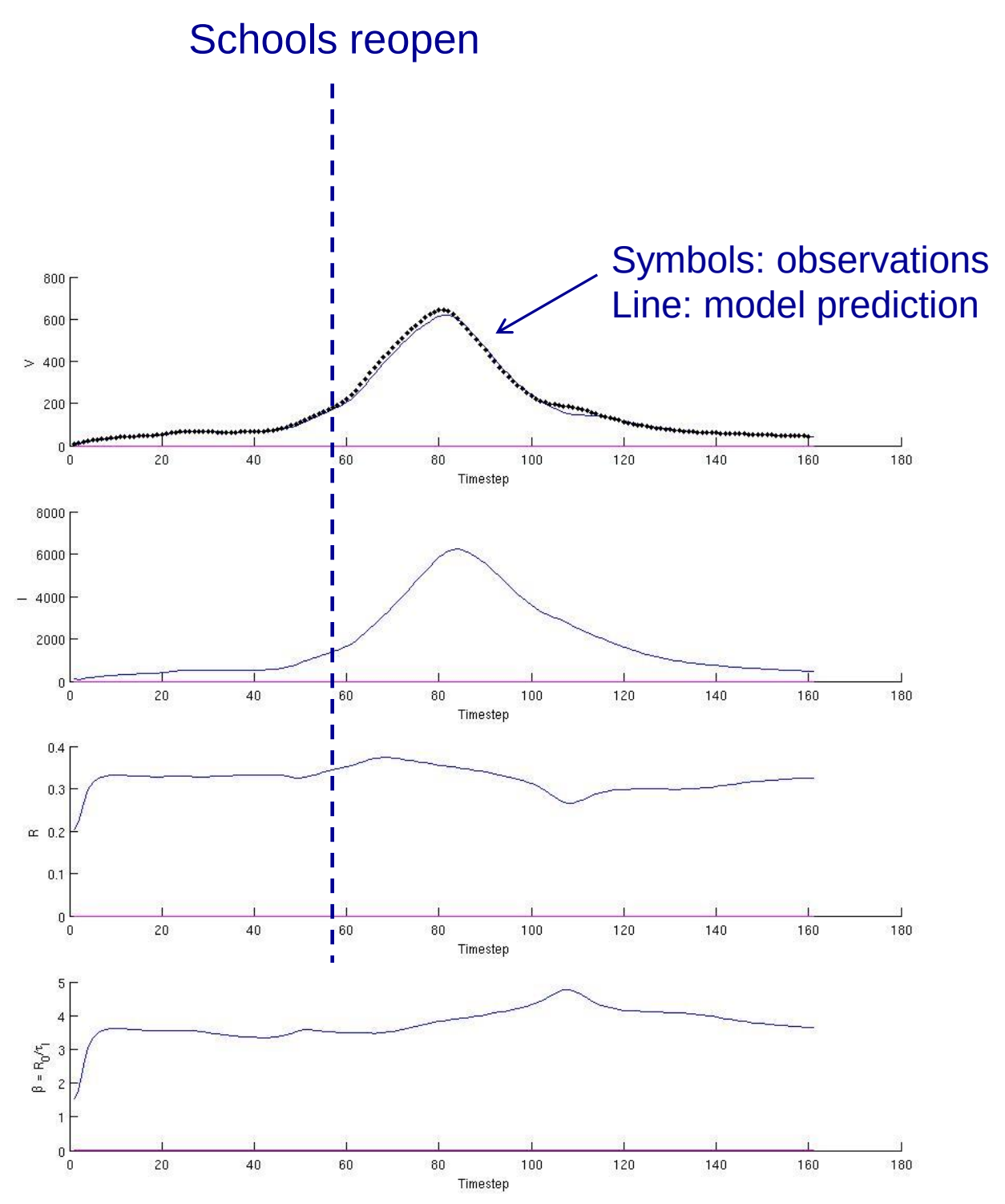

Dengue predictions in India, 2013

ARX model trained on data from 2008 to 2012, and used to predict the 2013 dengue season

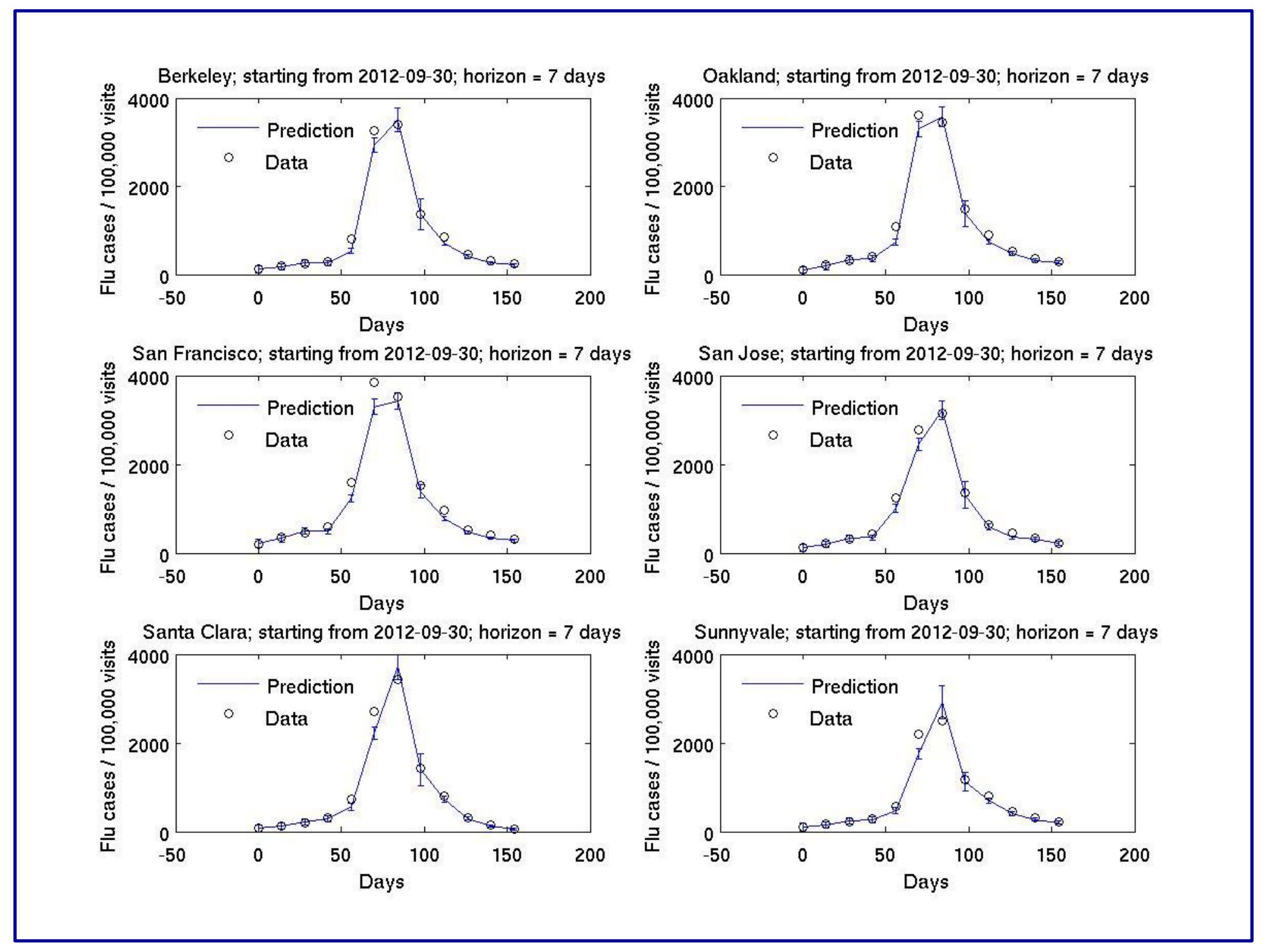
FLU FORECASTING USING DA

Data assimilation to calibrate flu models

- Epidemiological forecasting is best done with a properly calibrated model
    - The challenge lies in seeding the model with the correct infectious cohort, spread rate, and infectious period, which change every year
- GFT, corrected using CDC's lab reports on ILI cases testing positive for flu, provides observational data to calibrate a flu model
- We use a SIR model for flu
- We use Ensemble Transform Kalman Filters (ETKF) to calibrate the SIR model
    - The calibration produces a calibrated ensemble of 200 flu models
    - Provides probabilistic forecasts (best forecast and a standard deviation)



Data assimilation starting 11/4/2013



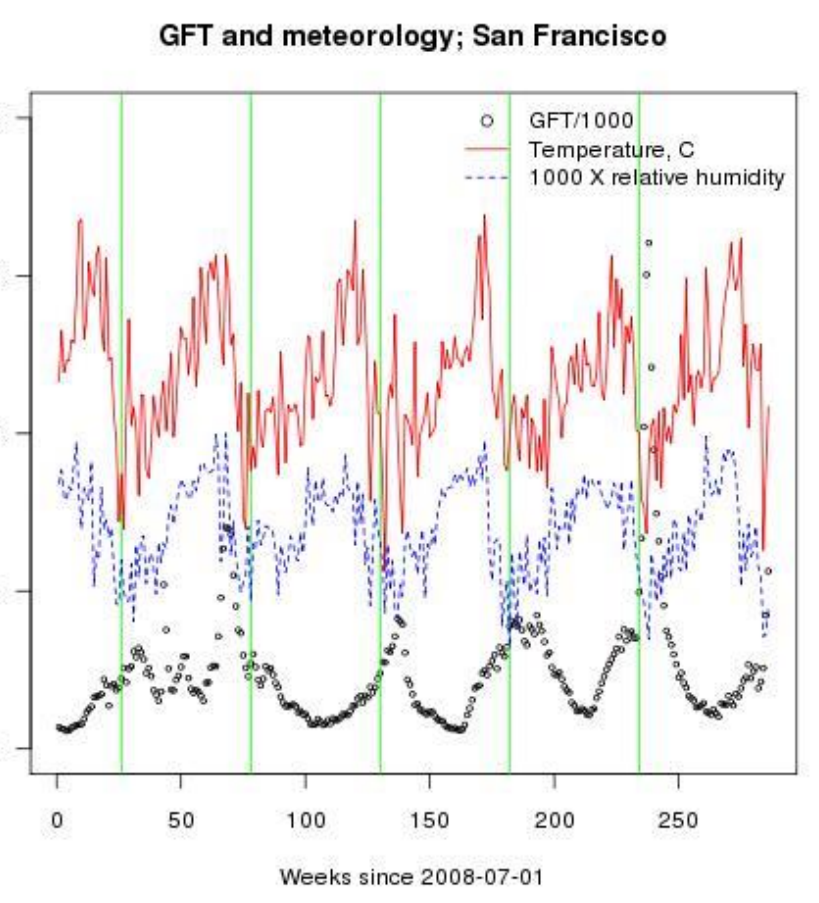
Tested on 6 San Francisco Bay Area (SFBA) cities tracked by GFT and compared against data

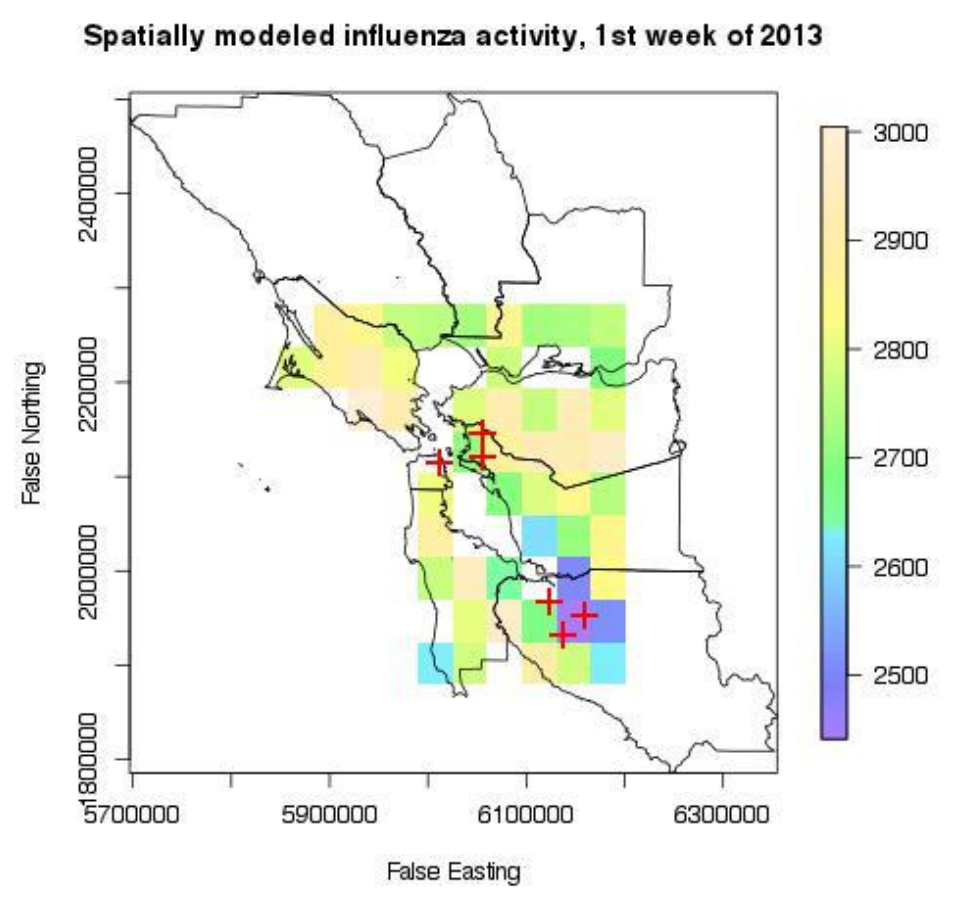
- Week-ahead forecasts usually capture the observational data
- *Predictions come with +/- 3σ bounds*

Forecast start date: 09/30/2012

Using meteorology for spatial modeling of flu activity

- GFT bears a strong correlation with humidity & temperature in the SFBA
    - Can be modeled using a simple linear model
    - $Y^{(i)} = \alpha_0 + \alpha_1 X_1^{(i)} + \alpha_2 X_2^{(i)}$, $Y^{(i)}$, $X^{(i)}_1$ & $X^{(i)}_2$ are GFT, temperature and humidity for week *i*
- GFT data (and forecasts from our DA method) are available at 6 SFBA cities
    - They provide 6 different (α_0, $\alpha_{1,}$ α_2) triplets
    - They are spatially interpolated to points in the SFBA using Nadaraya-Watson kernel smoothing, with an exponential correlation function
    - The flu forecasts are then provided using local humidity & temperature values
    - Test conducted for 2012-2013 flu season


GFT and meteorology; San Francisco


Spatially modeled influenza activity, 1st week of 2013

CONCLUSIONS

- Preliminary results indicate that OSI can be used to nowcast and forecast epidemiological activity
- The timely nature of electronic data sources, coupled with the leading nature of meteorological data (with respect to outbreaks) make such predictions possible – and useful
- The incorporation of exogenous data e.g., meteorology, require the use of appropriate time-series methods such as ARX, ARMAX etc
- The methods are robust i.e., when the exogenous variable misleads, the method exploits the smooth temporal evolution of the outbreak for its predictive skill
- Data assimilation via ensemble methods allow the most flexible, general and potentially most rewarding means of assimilating OSI
- The nature of the OSI assimilated is limited only by the model relating it to epidemiological activity
- Tests show that Healthmap data, of lower quality (but greater availability) than GFT can be used in nowcasting flu activity
- ARX methods are robust and predictive even when Healthmap data is misleading e.g., France, 2009
- GFT and meteorology can be used to perform spatial interpolation and prediction of flu activity even in regions not monitored by GFT
- Exploits the availability of meteorological forecasts at fine spatial resolution and its correlation with flu activity

Acknowledgements

The project is funded by Sandia National Laboratories LDRD (Laboratory Directed Research and Development) program. Part of the research was supported by a grant from the US Department of Defense.

References

1. J. Ginsberg, M. H. Mohebbi, R. S. Patel, L. Brammer, M. S. Smolinski and L. Brilliant, "Detecting influenza epidemics using search engine query data", Nature 457:1012-1015, 2009.
2. E. H. Chan, V. Sahai, C. Conrad and J. S. Brownstein, "Using web search query data to monitor dengue epidemics: A new model for neglected tropical disease surveillance", PLoS Neglected Tropical Diseases 5(5):e1206, 2011.
3. A. F. Dugas, M. Jalapour, Y. Gel, S. Levin, F. Torcaso, T. Igusa and R. E. Rothman, "Influenza forecasting with Google Flu Trends", PLoS One, 8(2):e56176, 2013.
4. J. Shaman, V. E. Pitzer, C. Viboud, B. T. Grenfell and M. Lipsitch, "Absolute humidity and the seasonal onset of seasonal influenza in the continental US", PLoS Biology, 8(2):e1000316, 2010.
5. J. Shaman and A. Karspeck, "Forecasting seasonal outbreaks of influenza", PNAS USA, 109:20425-20430, 2012.
6. J. Shaman, A. Karspeck, W. Yang, J. Tamerius and M. Lipsitch, "Real-time influenza forecasting during the 2012-2013 season", Nature Communications, 4:e2837, 2013.

For additional information, please contact:

S. Lefantzi, Sandia National Laboratories, slefant@sandia.gov

Sandia is a multiprogram laboratory operated by Sandia Corporation, a Lockheed Martin Company, for the United States of Energy's National Nuclear Security Administration under contract DE-AC04-94AL85000.