When Data Disappear: Public Health Pays As Policy Strays

tom mcandrew
Assistant Professor
Dept. of Biostatistics and Health Data Science
College of Health
Lehigh University

mcandrew@lehigh.edu
[Website=https://compuncertlab.org/](https://compuncertlab.org/)
[Github=https://github.com/computationalUncertaintyLab](https://github.com/computationalUncertaintyLab)

(Brief) Timeline

National Strategy For Pandemic Influenza Implementation Plan (2005, Bush)

a comprehensive approach to addressing the threat of pandemic influenza. Our Strategy outlines how we are preparing for, and how we will detect and respond to, a potential pandemic.

2009 H1N1, Swine Flu

CDC builds Emergency Operation Center

EO: 2013 Open Govt Initiative

EO making open and machine readable data the default

Project Open Data

Resources.data.gov is an online repository of policies, tools, case studies, and other resources to support data governance, management, exchange, and use throughout the federal government

The Trump administration has ordered hospitals to bypass the Centers for Disease Control and Prevention and send all Covid-19 patient information to a central database in Washington

EO: Advancing Equity and Racial Justice Through the Federal Government

Equitable Data Working Group to coordinate with agencies to expand their collection and use of demographic data and other equity data

EO: DEFENDING WOMEN FROM GENDER IDEOLOGY Extremism

Got agencies take down data, websites, information to assess whether they comply with EO

EO: Federal Funding Freeze

Makes it more difficult to collect data

Can disruptions to public health data cause issues with modeling (and so decision making)?

Lets collect data and compare two model forecasts: one with tons of data and one with the minimum needed

mcandrew@lehigh.edu

Data sources (https://github.com/computationalUncertaintyLab/importance_of_data)

Define two models: a data-poor model (Yellow) and a data-rich model (Blue)

D1. National Hospital Safety Network - extracted Flu Hospitalizations

1. get_target_data.R from cdcepi/FluSight-forecast-hub

D2. National Hospital Safety Network - Percent of facilities reporting

2. importance_of_data/download_percent_reported_hosps.py (totalconffluhosppatsperc)

D3. Outpatient Illness and Viral Surveillance dataset (ILINET) - Public health lab data

3. importance_of_data/download_epidata.py which uses Epidata

D4. Outpatient Illness and Viral Surveillance dataset (ILINET) - Clinical lab data

4. importance_of_data/download_lab_percentage_data.R

D5. Morbidity and Mortality Weekly Report - Interim vaccine effectiveness

5. importance_of_data/data_sets/VE_mmwr.csv

D6. National Oceanic and Atmospheric Administration - Temperature and pressure

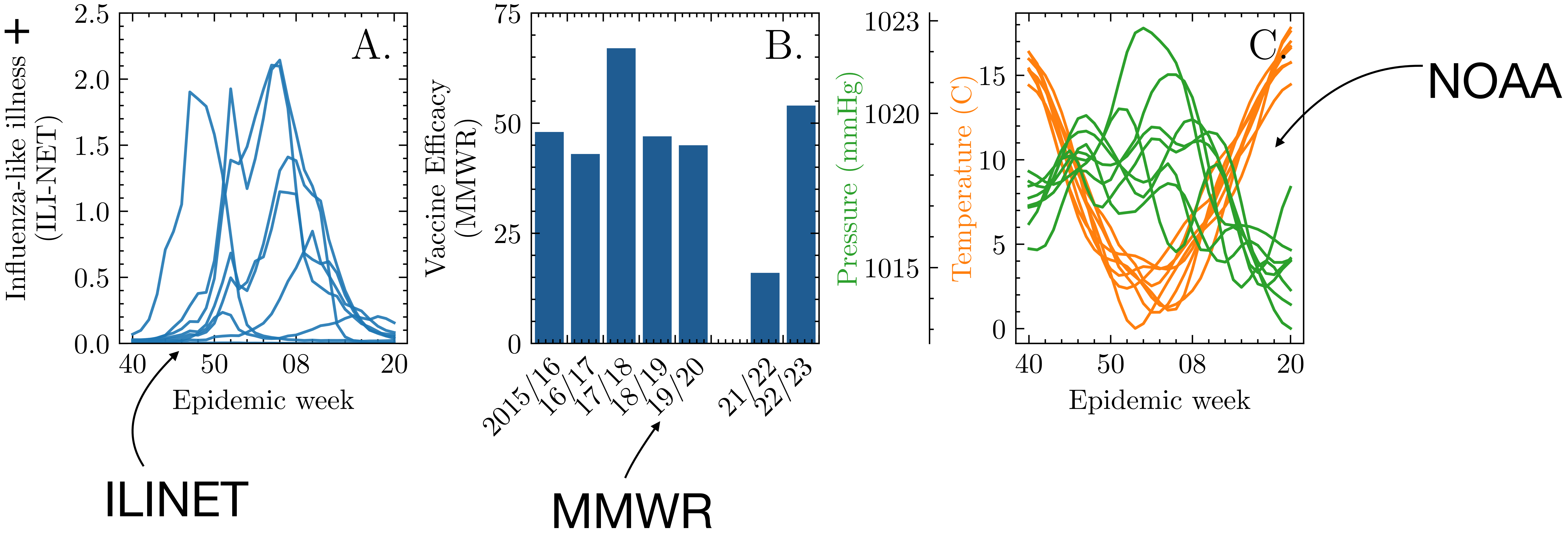
6. importance_of_data/download_weather_data.py

D7. Census- Number of individuals living in the United States

7. importance_of_data/data_sets/locations.csv

mcandrew@lehigh.edu

Data sources



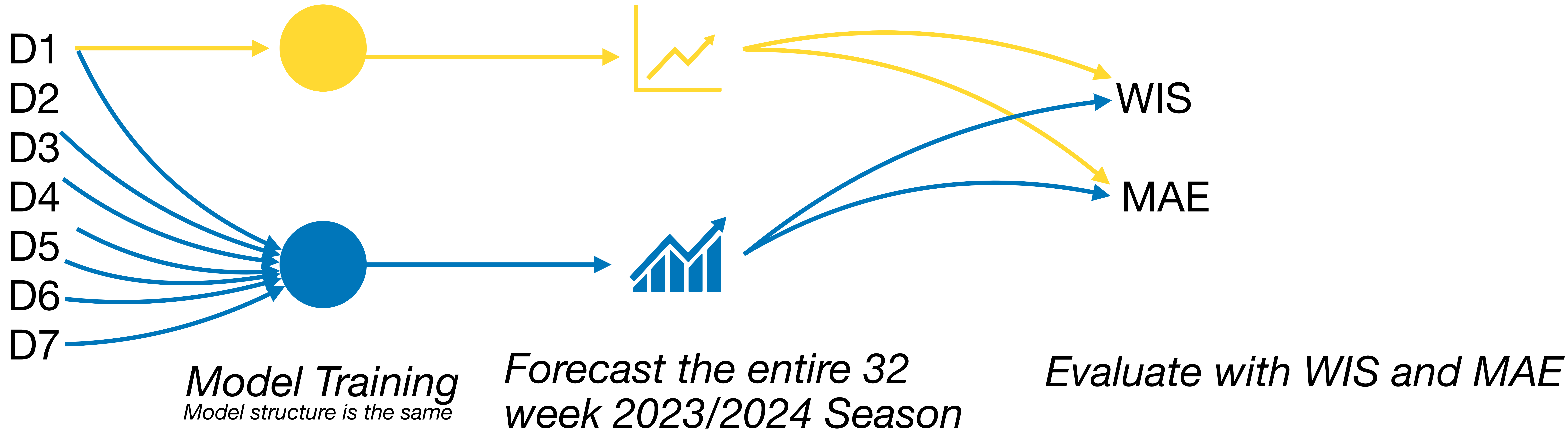
Can i incorporate the above data signals into a model to predict US national incident hospitalizations?

Does having this additional information in the model improve forecasts of flu hospitalizations?

Overarching goal: To demonstrate that these data sources are valuable and so should be insulated from executive action

mcandrew@lehigh.edu

Experiment

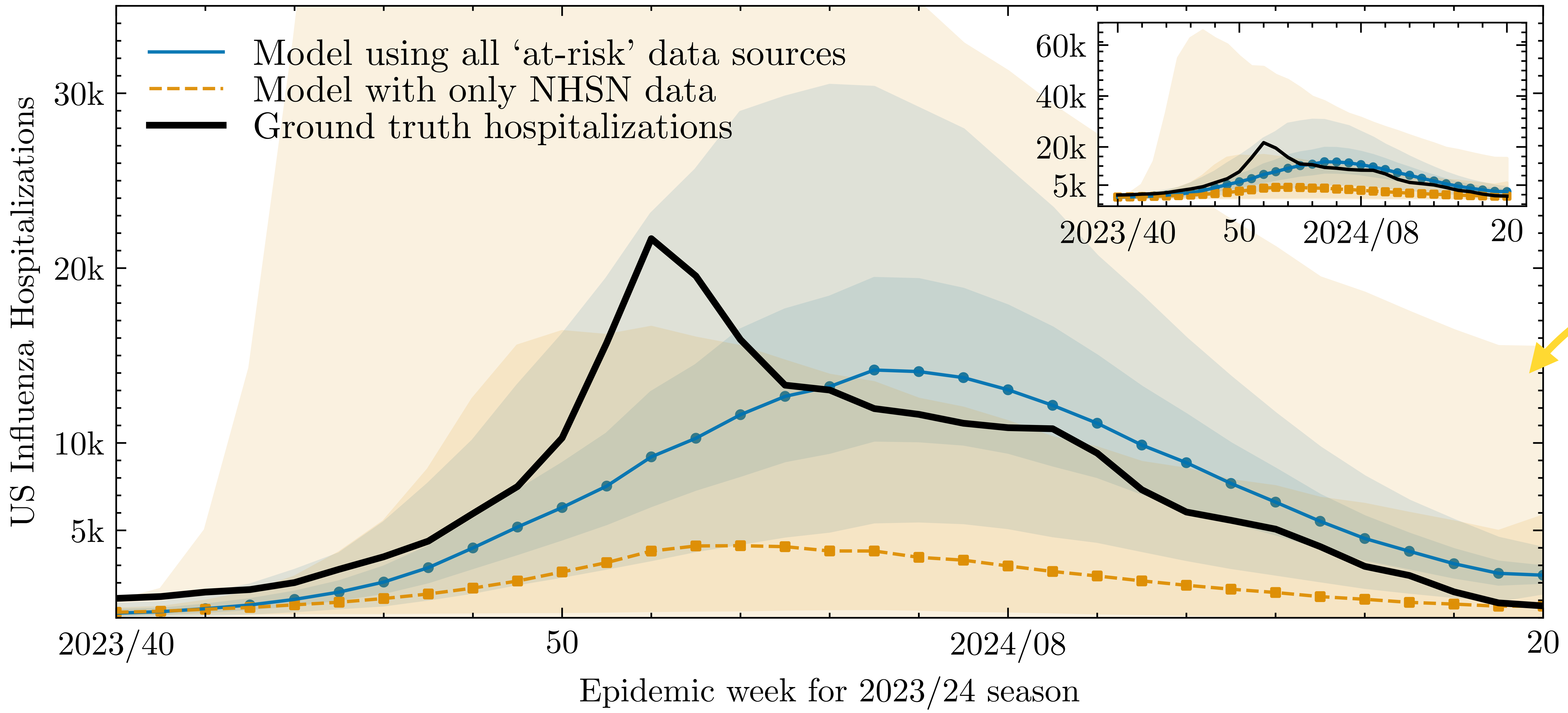


If the Blue model, trained on all data, produces better forecasts than **we should observe** smaller WIS scores and smaller MAE scores over the course of the season.

Model will be compartmental. Model structure same between Yellow and Blue models.

mcandrew@lehigh.edu

Results First, Details Later



Yellow model only has 2021/22 and 2022/23 seasons

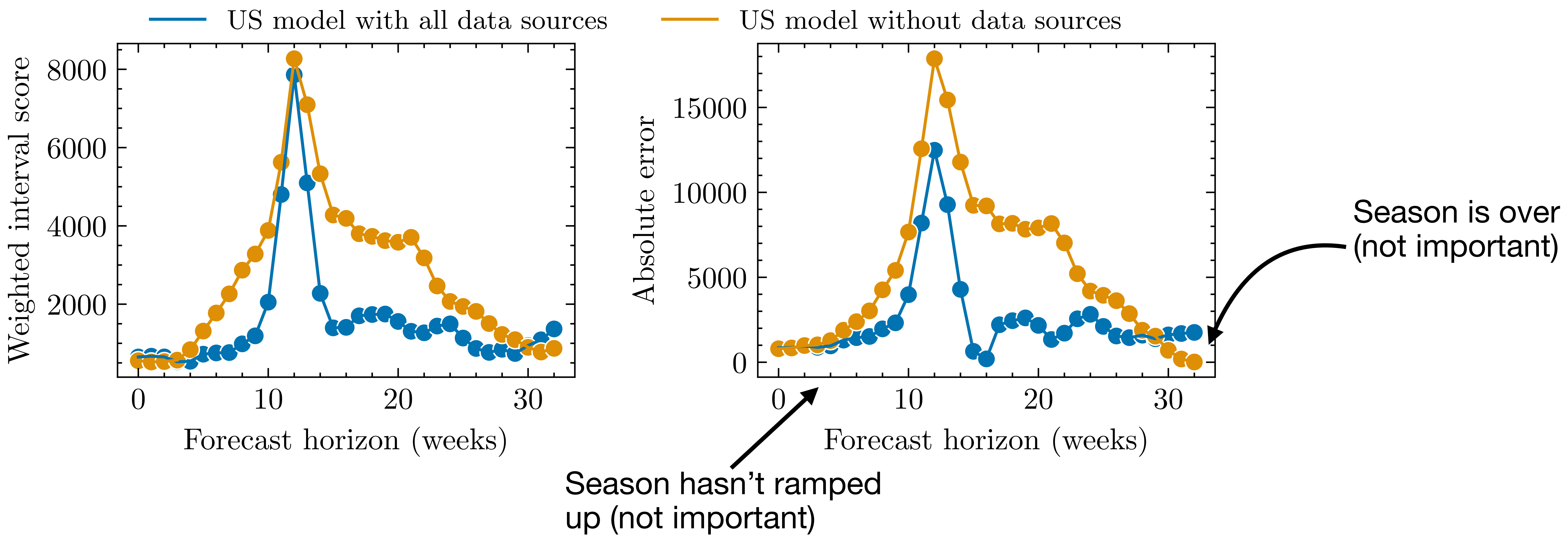
Blue model appear to be a more informative forecast

Blue model could be used for planning. Yellow model it too vague.

Yellow model suspects too mild a season. Blue model scale is closer to truth

mcandrew@lehigh.edu

Results, WIS and MAE



We observe smaller WIS and smaller MAE scores for the Blue model when compared to Yellow. Blue and Yellow models have similar performance for forecasts at the very beginning and very end of the season. These two time points are the least important.

<u>This lends support to demonstrating that MORE DATA MEANS BETTER MODELING.</u>
(Documenting the obvious)

mcandrew@lehigh.edu

Implications / Discussion

Potential paths forward

1. **Public health data as public good.**
    I. Use of health data does not exclude or reduce availability to others.
    II. 2013 EO by Obama supports above by making data freely available in machine readable format.
    III. Congressional act to officially designate public health data as a public good.

2. **Public health data at sub-national levels.**
    I. Academia, industry, state government collaborations.
    II. Major burden is coordination.

3. **Proxy data for forecasting.**
    I. Prepare readily available alternatives that correlate with public health data.
    II. Human mobility, social media data, OTC medication data.
    III. Delphi is likely the most concerted effort on this front.

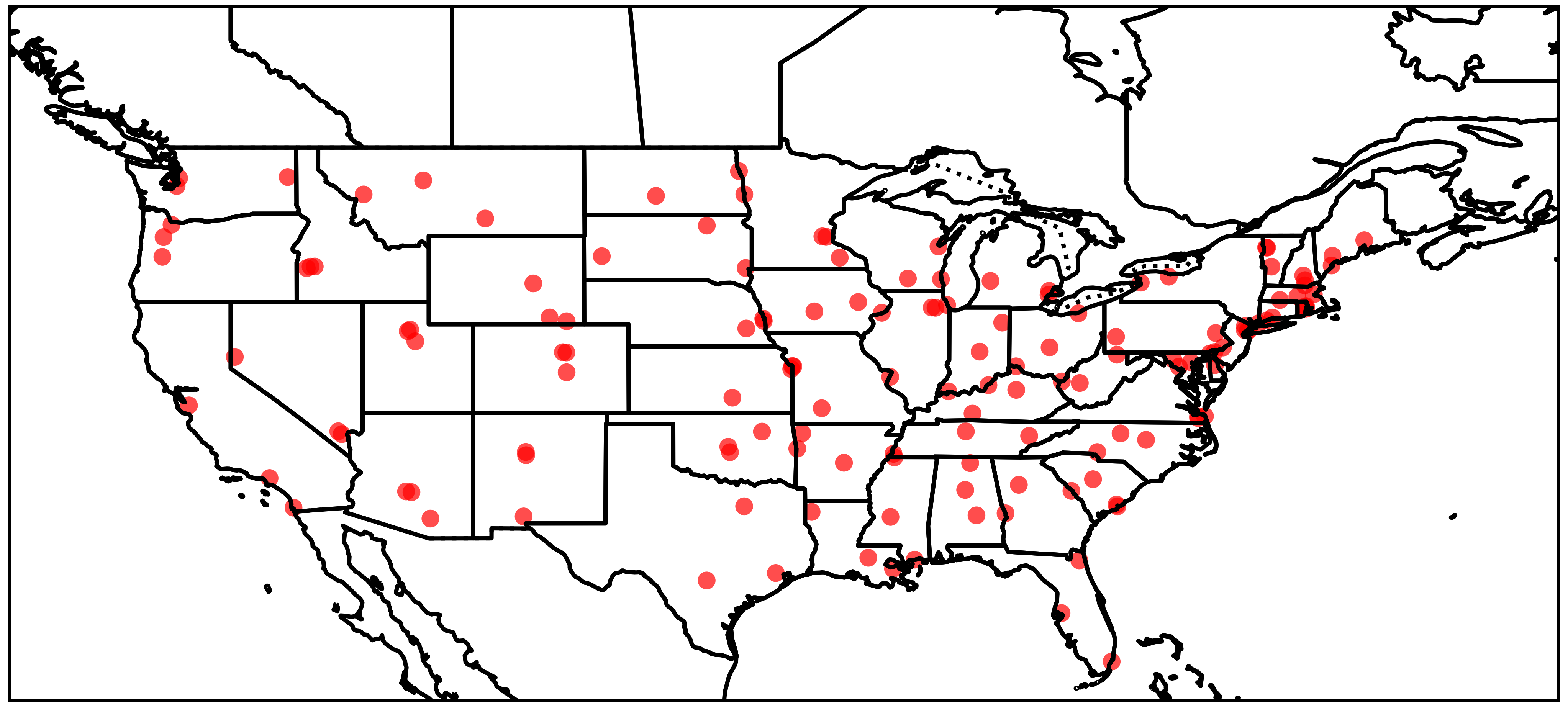
Main message about Public health data and why we should care at all:
Public health data serve as a common ground to discuss how to change policy to better serve our citizens

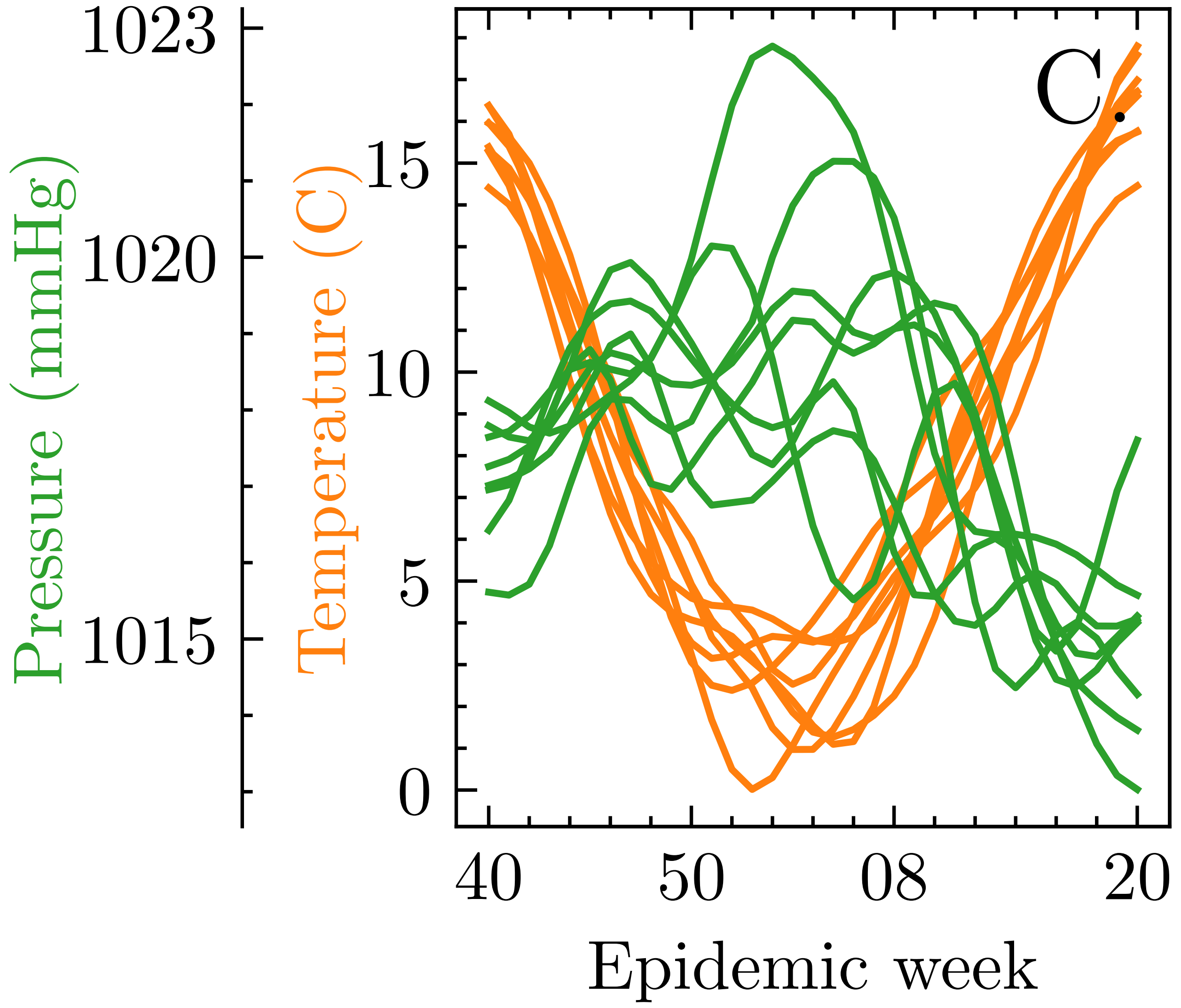
Ok, time for modeling details->

mcandrew@lehigh.edu

The model - Computing temperature and pressure for the US

NOAA data measured at 3 largest cities per state





Find Lat / Long for three largest cities in each state.

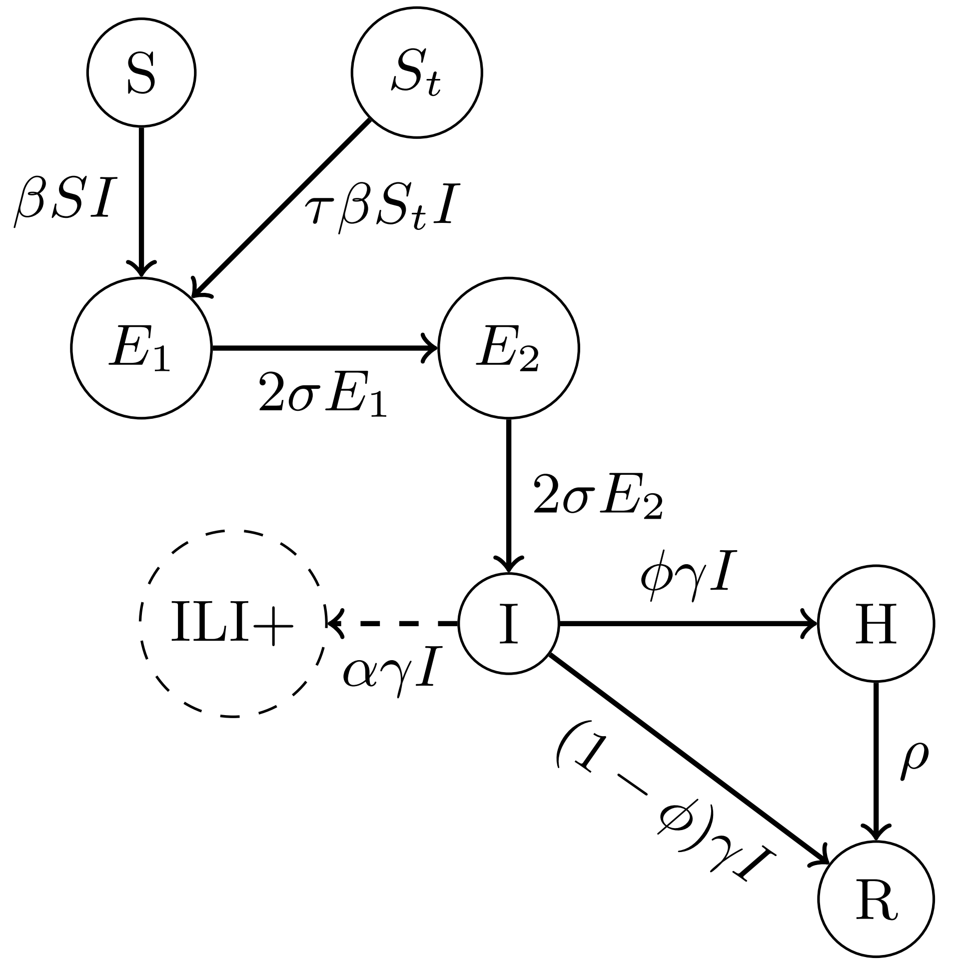
Collect temperature and pressure data from nearest airport or station

Temperature for US at time *t* is average over all cities

Limitations: Woefully misses stochasticity in the country

mcandrew@lehigh.edu

The model - Compartmental structure



Unvaccinated

$$
\begin{aligned}
\dot{S} &= -\beta(t)SI \\
\dot{S}_t &= -\tau\beta(t)S_t I \\
\dot{E}_1 &= \beta I(S+\tau S_t) && -2\sigma E_1 \\
\dot{E}_2 &= 2\sigma E_1 && -2\sigma E_2 \\
\dot{I} &= 2\sigma E_2 && -\gamma I \\
\dot{H} &= \phi\gamma I && -\rho H \\
\dot{R} &= (1-\phi)\gamma I && +\rho H
\end{aligned}
$$

Vaccinated reduces transmission

Linear Chain trickery

FIG. (3) (Left) A flow diagram that presents how individuals move through disease states in this dynamical system (right). Disease states are represented as circles. Rates are placed on arrows to describe the rate at which individual move from one state to another. The ILI state is placed in a dashed circle because this state is only an observed state that is used to inform prevalent infections (I).

Helper states for incidence

$$\dot{cI} = 2\sigma E_2;\ \dot{cILI+} = \alpha\gamma I;\ \dot{cH} = \phi\gamma I$$

Assumptions

Dont expect removed individuals to return to *S* within one season

All individuals have the same protection, tau

Homogeneous mixing and a closed system

Two items to address: Init Conds and Fixing params

mcandrew@lehigh.edu

The model - Initial conditions

Goal is to assign, as many as possible, of the states to zero

Q: Why? A: Non-identifiability

Set six states equal to zero

$$E_1 = E_1 = R = H = cH = cILI+ = 0$$

Assume a certain proportion, v, are vaccinated to determine Susceptible and Susceptible but treated

$$S_{\text{ttl}} = S + S_t$$

$$S_t = vS_{\text{ttl}};\ S = (1-v)S_{\text{ttl}}.$$

This is an informative estimate (more on this soon)

$$\text{logit}\left(S_{\text{ttl}}\right) \sim \mathcal{N}\left(\text{logit}\left(\widehat{S_{\text{ttl}}}\right), \sigma_S\right).$$

$$I = (1 - S_{\text{ttl}})$$

Final initial conditions

$$(S, S_t, 0, 0, I, 0, 0, I, 0, 0)$$

We focus on initial conditions for S and I

mcandrew@lehigh.edu

The model - Fixing parameters based on Literature (reducing param space)

Goal is to assign, as many as possible, of the parameters to fixed values

Q: Why? A: Non-identifiability (again)

Parameter	Fixed value
$1/\gamma$	3/7 week
$1/\sigma$	2/7 of a week
$1/\rho$	5/7 of a week

Values are taken from lab studies on the infectious and latent periods, and on the typical hospital stay due to influenza.

Be careful to make sure that your parameter values are on the same scale as your model (in this case a week)

mcandrew@lehigh.edu

The model - Priors

$$\sigma_\beta \sim \text{Half-Cauchy}(10);\ \log(\beta_0) \sim \mathcal{N}(\beta_{0,\text{mle}}, \sigma_\beta)$$

$$\sigma_{\beta s} \sim \text{Half-Cauchy}(1);\ \log(\beta_{0,s}) \sim \mathcal{N}(\log(\beta_0), \sigma_{\beta s})$$

$$\sigma_\phi \sim \text{Half-Cauchy}(10);\ \text{logit}(\phi) \sim \mathcal{N}(\phi_\text{mle}, \sigma_\phi)$$

$$\sigma_\alpha \sim \text{Half-Cauchy}(10);\ \log(\alpha) \sim \mathcal{N}(\alpha_\text{mle}, \sigma_\alpha)$$

$$\text{logit}(S_\text{ttl}) \sim \mathcal{N}(S_{\text{ttl,mle}}, 10)$$

$$S_0 = (1-\nu)S_\text{ttl};\ S_t = \nu S_\text{ttl}$$

$$I_0 = (1 - S_\text{ttl})$$

Transmission rate intercept are linked

Data

$$\tau_s = \text{VE}_{\text{MMWR},s}$$

$$\log(\beta_{s,t}) = \log(\beta_{0,s}) + b_1 \text{temp}_{s,t} + b_2 \text{pres}_{s,t}$$

Model fit is stochastic variational inference via Numpyro

Normal “Guide”, 20k iterations, make use of Jax to speed up computation.

mcandrew@lehigh.edu

The model - Setting some informative priors via fit

Fit a simpler, pooled model across all seasons

Estimate the following parameters:

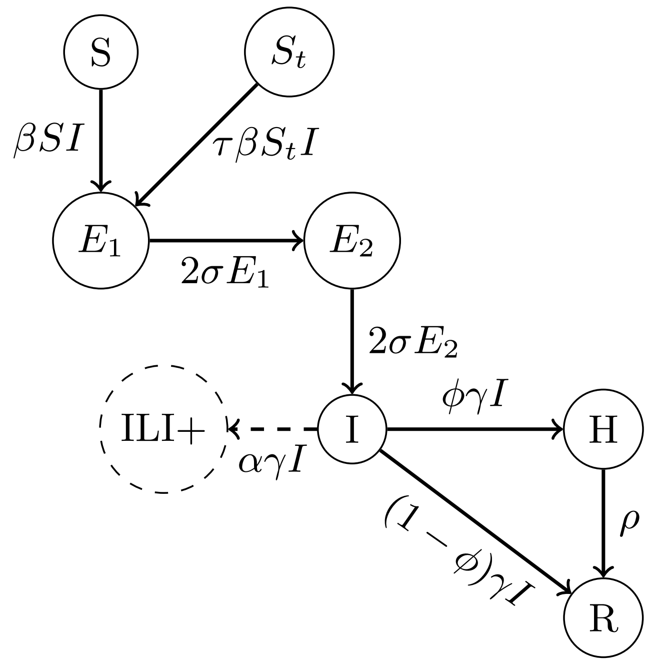
1. Total proportion of susceptible
2. Proportion of ILI cases reported
3. Proportion of hospitalizations reported
4. Transmission rate (just the intercept)

Minimize the negative log likelihood for ILI cases and hospitalizations using a Genetic algorithm (Pop size = 10k)

1. Pymoo is great for this
2. GA is less likely to get stuck in a local optima

mcandrew@lehigh.edu

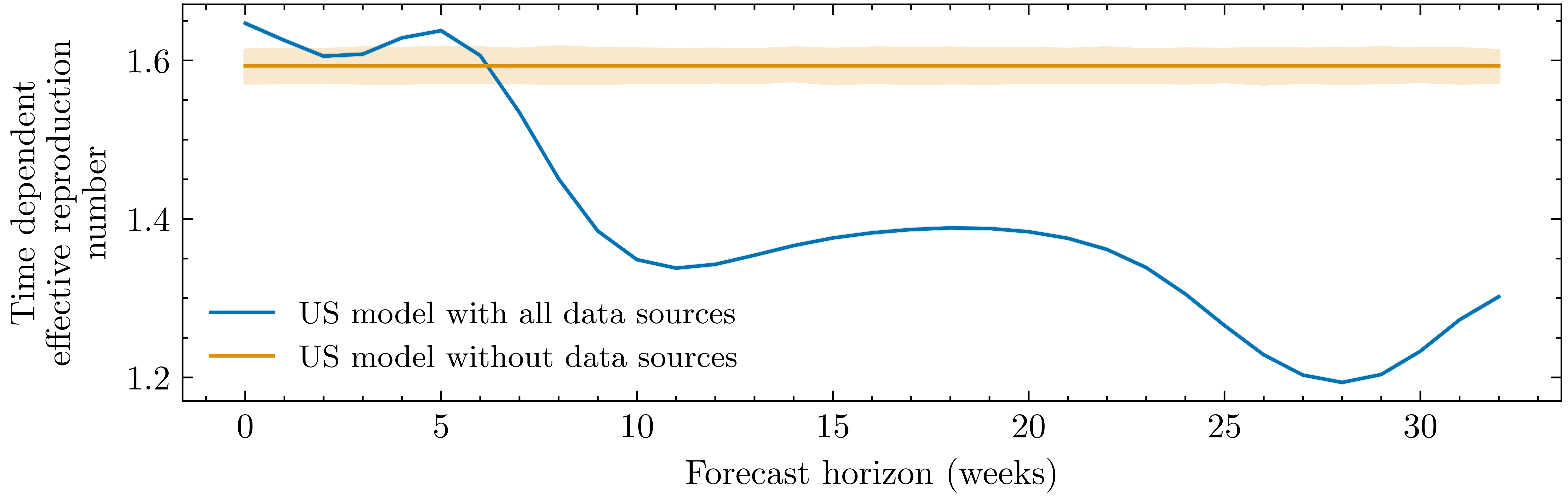
The model - Basic and Effective Repo Number



$$\begin{aligned}
\dot{S} &= -\beta(t)SI \\
\dot{S}_t &= -\tau\beta(t)S_t I \\
\dot{E}_1 &= \beta I(S+\tau S_t) && -2\sigma E_1 \\
\dot{E}_2 &= 2\sigma E_1 && -2\sigma E_2 \\
\dot{I} &= 2\sigma E_2 && -\gamma I \\
\dot{H} &= \phi\gamma I && -\rho H \\
\dot{R} &= (1-\phi)\gamma I && +\rho H
\end{aligned}$$

FIG. (3) (Left) A flow diagram that presents how individuals move through disease states in this dynamical system (right). Disease states are represented as circles. Rates are placed on arrows to describe the rate at which individual move from one state to another. The ILI state is placed in a dashed circle because this state is only an observed state that is used to inform prevalent infections (I).

$$\mathcal{R}_{eff} = S(0)\beta(t)(1+\tau)\left(\frac{1}{\gamma}\right)$$



This is likely bc of setting epi parameters from lit (i.e. from constraints)

If you don’t see estimates like the disease your studying, the model may be right and the parameters wrong.

mcandrew@lehigh.edu