Flu Forecasting using Climate Data and Bayesian Hierarchical Modeling

Maia Richards-Dinger[1] and Dr. Stephen Kissler[2]

[1]CU Boulder - Applied Math Department, [2]CU Computer Science Department

Motivation and Background

Importance of Forecasting

- Influenza places a significant disease burden on temperate countries in the Northern Hemisphere every winter
- Accurate forecasts can help hospitals prepare for spikes and decide how to ration medication, and can impact how individuals make decisions regarding vaccination, mask wearing, travel, etc.

Original motivating questions

- How do scoring rules (explicit or implicit) potentially impact infectious disease forecasting competition submissions?
  - Frongillo [3] studied how improper scoring rules can incentivize forecasters to report predictions that do not align with their true beliefs - could this be affecting the submissions to disease forecasting competitions?
- CDC organizes the FluSight competition and uses the multibin logarithmic scoring rule, which is not proper [1]
- What is the optimal way to aggregate ensemble forecasts?
  - CDC currently seems to just take the median

Currect Project

- Goal: Forecast flu incidence at the state level using climate data
- Cold and dry conditions have been shown to correlate with increased flu (and other respiratory infection) activity [2]
- Can incorporating climate data improve flu forecasts?

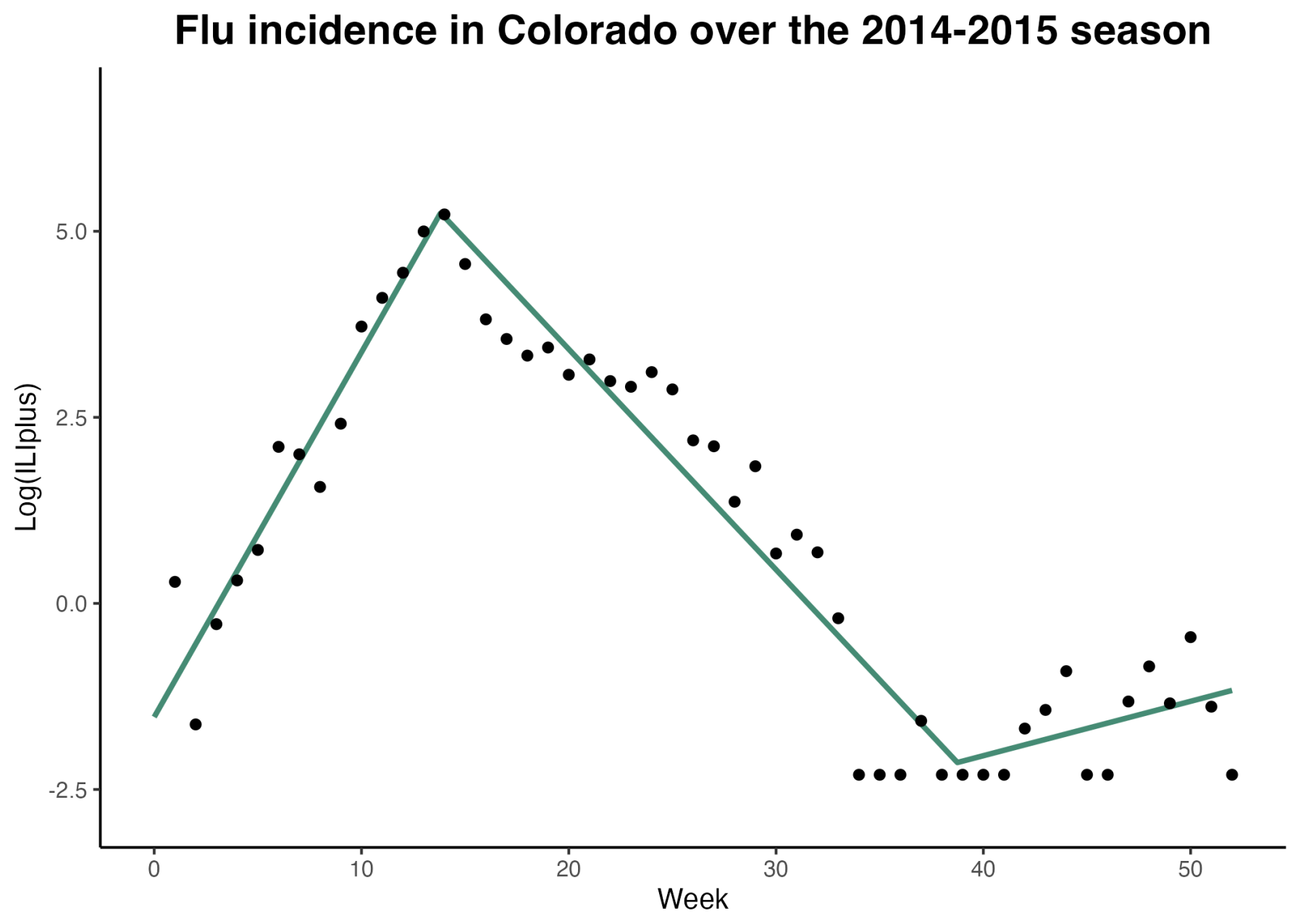
Methods

Step 1: Calculate ILI plus

- CDC collects data on percent of healthcare visits due to influenza-like illness (ILI) by state and publishes it weekly
- Report percent of flu tests that are positive by state each week
- As in Goldstein et al. [4], we use the product of percent visits due to ILI and percent positive flu tests as a proxy for flu incidence, which we call ILI plus

Step 2: Fit a piecewise linear functions

- Since we expect flu incidence to grow and decay exponentially, we expect the log of ILI plus to grow and decay linearly
- For each flu season and state, fit a piecewise linear function with two breaks to the log of ILI plus data
- Record up slope, down slope, peak week, and peak value



Methods (cont.)

Step 3: Predict parameters using climate and population data

- Calculate the population density (x_1), percent of people under 18 (x_2), latitude (x_3), mean maximum temperature (x_4), mean maximum relative humidity (x_5), mean minimum relative humidity (x_6), and mean average absolute humidity (x_7) for each state and flu season, using census population data and gridMET climate data
- Use Monte Carlo Markov Chain methods in Stan to fit the following statistical model

$$y_k \sim \alpha_k + \sum_{i=1}^{7} \beta_{ik} x_i + \text{Normal}(0, \sigma_k)$$

for $k = 1, 2, 3, 4$, where $y_1 =$ up slope, $y_2 =$ down slope, $y_3 =$ peak week, and $y_4 =$ peak value

- Use α_k and β_{ik} to predict the piecewise linear fit for each flu season and state based on the climate and population variables for that season/state and compare to the actual fit

Results (so far)

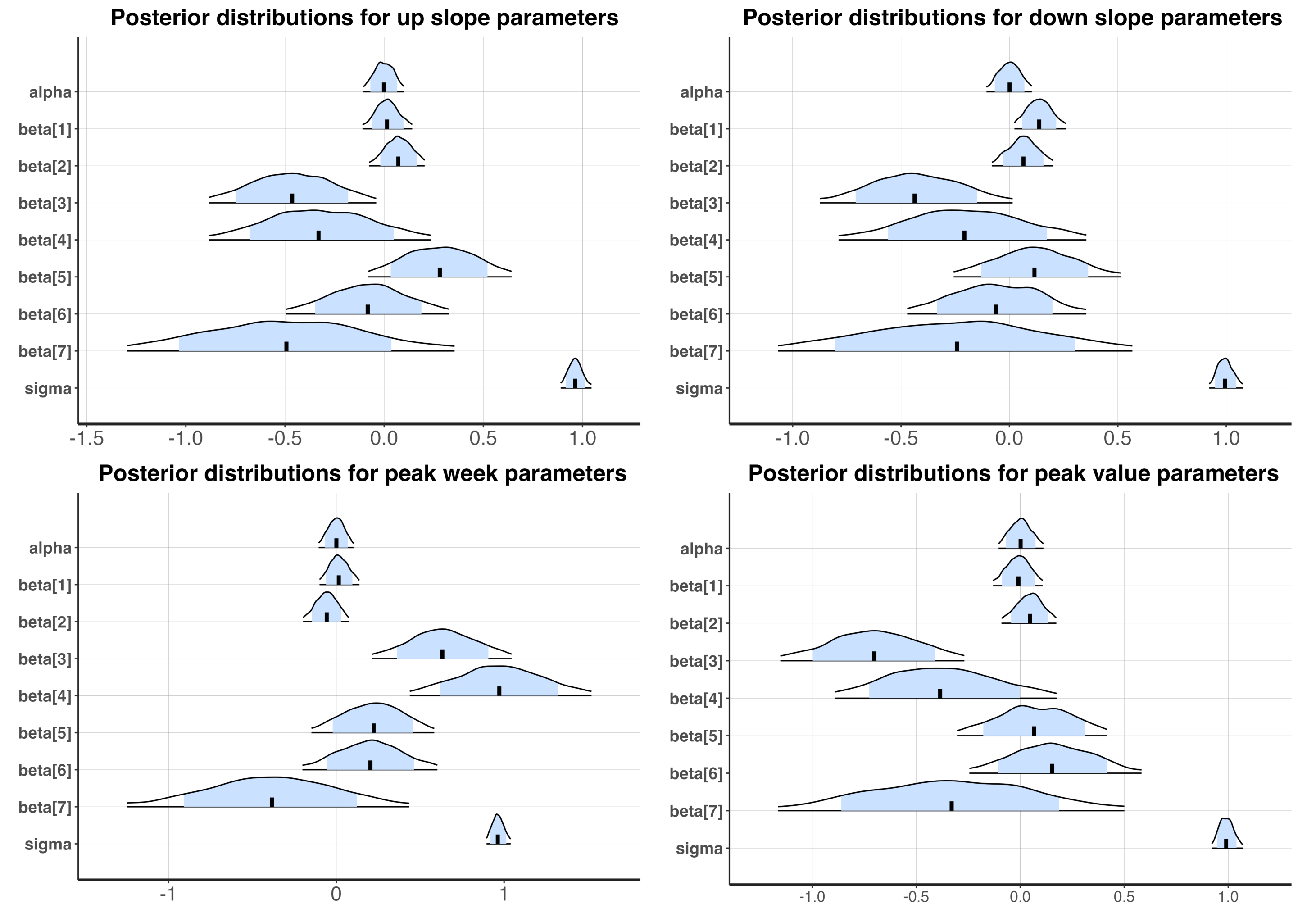


Figure: Posterior distributions for $\alpha_k, \beta_{ik}, \sigma_k$ parameters for up slope, down slope, peak week, and peak value. The blue shaded regions are 80% credible intervals and the total lines plotted are 95% credible intervals.

- Overall β_3, which corresponds to latitude, is the most significant
- The further north a state is, the less steep their up slope tends to be, the steeper their down slope tends to be, the later their peak week tends to be, and the smaller their peak value tends to be
- The more dense a state is, the less steep their down slope tends to be
- The hotter a state is in a given year, the later their peak week tends to be

Results so far (cont.)

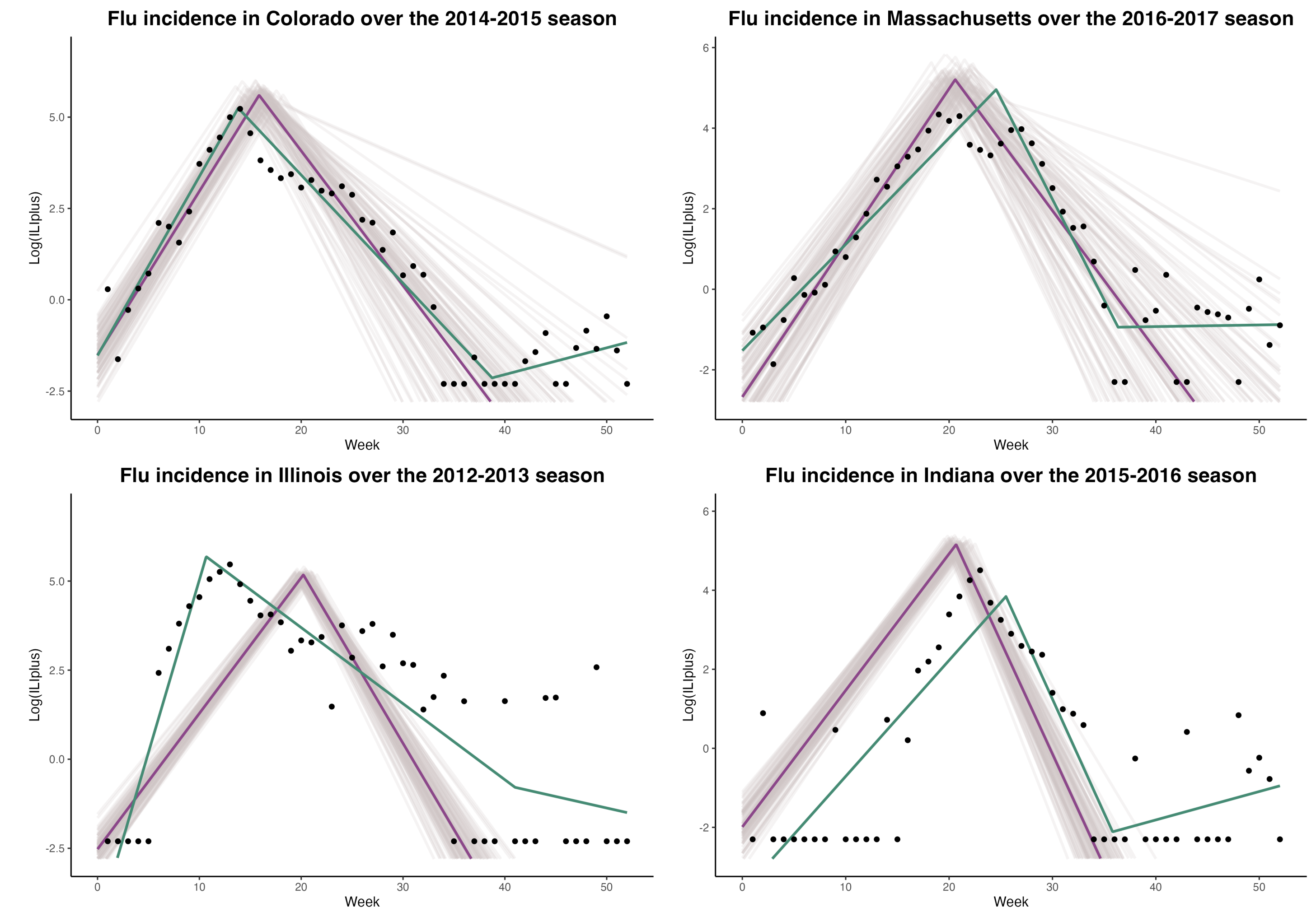


Figure: Flu incidence in four different states and years along with the piecewise linear fit (green line) and the piecewise linear fits predicted by that state/year's climate and population data (mean of posterior distribution - purple line, 100 draws from posterior distribution - light purple lines).

- Some predicted fits are quite good!
- Others are not as good
- In general, the Bayesian model parameter predictions have less variance than the true parameters (up slope, down slope, etc.)

Future Work

- Consider other climate variables (e.g. minimum temperature) and timing of climate variables (i.e. we'd expect the temperature, humidity, etc. earlier in the season to have a greater impact)
- Model evaluation and comparison
- Implement Bayesian *hierarchical* modeling
- Actually create forecasts using an ARIMA (autoregressive integrated moving average) model

References

[1] Bracher, Johannes. "On the multibin logarithmic score used in the FluSight competitions." *Proceedings of the National Academy of Sciences* 116.42 (2019): 20809-20810.

[2] du Prel, Jean-Baptist, et al. "Are meteorological parameters associated with acute respiratory tract infections?." *Clinical infectious diseases* 49.6 (2009): 861-868.

[3] Frongillo, Rafael M. *Eliciting private information from selfish agents*. University of California, Berkeley, 2013.

[4] Goldstein, Edward, et al. "Predicting the epidemic sizes of influenza A/H1N1, A/H3N2, and B: a statistical method." *PLoS medicine* 8.7 (2011): e1001051.

Mail: maia.richards-dinger@colorado.edu