Illness forecasting

A file named state_traces.csv is attached. It contains a time series representing Kinsa-derived daily estimates of the number of people ill in three different US states for a period of a little over two years. The illness count estimate is for influenza-like illness (ILI) and the values are not smoothed in either time or space. In US the ILI count typically peaks in the winter months, but the timing of the peak and its intensity can vary significantly from season to season.

	state	date	value
107	NJ	20151002	4293.414810
108	NJ	20151003	5977.121664
109	NJ	20151004	1903.517673
110	NJ	20151005	20754.601160
111	NJ	20151006	12885.652476

state_traces.csv sample

1. The first task is to develop a model (using any technique you choose) to forecast the illness signal in each of the three states in the next 30 days, using the data in state_traces.csv. Ideally your model would not treat the states in isolation, but would account for the possibility of infectious disease moving between different geographies.
    - Please report the forecasted values for the 20-day period for each of the states.
    - Please visualize the evolving ILI intensity throughout the Winter season of 2017/8, including your forecasted values.
    - Please characterize the performance of the model and suggest potential improvements.
    - Why did you select a particular model?
    - What other approaches may work?
    - How would your approach change if you had more granular data (counties, zip codes)?
    - How would your approach change if you needed to make predictions three months ahead rather than three weeks ahead?
    - Which other data sources would you include if you had more time for this problem?

   The approximate populations of the three states are: 9.0M (NJ), 19.8M (NY) and 12.8M (PA). You can disregard the population change in the given two-year time window. If you chose to build features based on epidemiological information (for example via SIR-family of models (Susceptible-Infected-Recovered)), you should keep in mind that the strain (infectiousness, recovery rate) can vary from season to season.

2. Attached is a link to a manuscript that describes one approach to real-time influenza modeling. Please read the manuscript and write a relatively brief critical review of the method used, novelty, and applicability to Kinsa.