



Modeling the Projected Effect of Climate Change on the Prevalence of Food and Waterborne Illness Outbreaks

Nicolas Beltran, Sweta Karlekar

Department of Computer Science, Columbia University



Introduction

Climate change has increased the volatility of various weather phenomena, leading increased frequency of events like droughts and flooding. Moreover, changes in weather patterns have been strongly hypothesized to have a negative impact on food safety in the form of increases in foodborne, waterborne, and enteric diseases. Previous studies have shown that there is a significant association between climatic factors like drought, flooding, temperature, humidity, etc. on the prevalence of salmonellosis, vibriosis, and Escherichia coli infection [3] [6] [13] [14] [15]. Existing approaches assess the correlation between these factors and hypothesize causal mechanisms, but fail to quantify the causal impact and understand non-linear and/or hierarchical effects [12].

This work uses existing physics-based climate models alongside probabilistic machine learning methods to estimate the prevalence of food and waterborne illness outbreaks given different climate change scenarios under a causal inference framework.

We compare various probabilistic models and show projected outbreaks from 2030 to 2070, as well as state-specific outliers. While our findings display high variance, they contradict our initial assumptions based on related biological and ecological literature that the severity of climate change would cause increases in the number of outbreaks. We discuss several possible reasons for this contradiction, as well as areas for future work.

Datasets

Weather Dataset: We use the Climate at a Glance dataset developed by the National Oceanic and Atmospheric Administration (NOAA) [10]. This dataset provides monthly temperature and precipitation data, as well as drought severity indexes, z-indexes, cooling degree days, and heating degree days from 1865 to 2023 for each state in the contiguous US.

Outbreak Dataset: We use the National Outbreak Reporting System (NORS) developed by the Center for Disease Control (CDC) [5] to access the total number of outbreaks in a month from each contiguous state in the US, ranging from 2009 to 2019. NORS includes both foodborne and waterborne outbreaks, as well as outbreaks spread through animal, environmental, person-to-person, and indeterminate vectors. We use the years 2009 to 2018 to train our models and we hold out the year 2019 as our test set.

Structural Causal Model

We posit the following causal model for food-borne outbreaks: We assume there exists a set of states represented by $\mathcal{S}$, a set of time-points $\mathcal{T} = \mathcal{T}_p \cup \mathcal{T}_f$ where $\mathcal{T}_p$ and $\mathcal{T}_f$ represent past and future time-points, vectors of time and state-specific climate variables $C_{s,t}$, unobserved time confounders M_t , unobserved state confounders H_s , per time point and state outbreaks $O_{s,t}$, and a climate scenario variable I , which in accordance to the IPCC can be any of the established Shared Socioeconomic Pathways (SSP) [9]. These variables are formally related via the following SCM depicted in figure (1).

1. Endogenous variables $V = \{C_{s,t}, O_{s,t}, I\}$
2. Exogenous variables $U = \{H_s, \epsilon_{s,t}^c, \epsilon_{s,t}^o, \epsilon^I, M_t\}$
3. Structural equations:

$$\begin{aligned} I &\leftarrow \epsilon^I \\ C_{s,t} &\leftarrow 1_{\{t \in \mathcal{T}_p\}} f_C^p(M_t, H_s, \epsilon_{s,t}^c) + 1_{\{t \in \mathcal{T}_f\}} f_C^f(M_t, H_s, I, \epsilon_{s,t}^c) \\ O_{s,t} &\leftarrow f_O(M_t, H_s, \epsilon_{s,t}^o) \end{aligned}$$

4. We assume the exogenous noise factorizes as

$$P(H_s)P(M_T)P(\epsilon^I) \prod_{s \in \mathcal{S}, t \in \mathcal{T}} P(\epsilon_{s,t}^c)P(\epsilon_{s,t}^o)$$

and

$$P(M_t = M_t \bmod 12) = 1$$

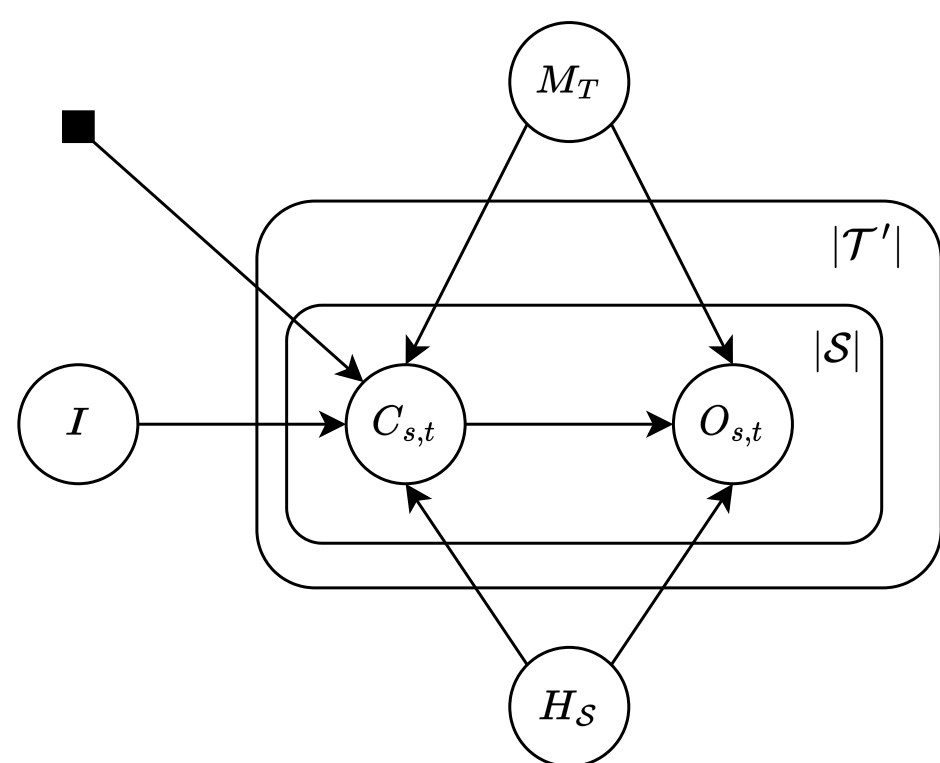


Figure 1. Visual Representation of SCM: $\mathcal{T}' = \mathcal{T}_p, \mathcal{T}_f$ and black square denotes the difference in mechanism depending on the setting

Goal - Estimand

Following [1], our goal is to estimate

$$P(O_{s,t} | \text{do}(I = i, H_s = h_s, M_t = m_t)) \quad t \in \mathcal{T}_f$$

where i is a climate change scenario, and h_s and m_t are the realized values of the random variables.

By simple do calculus we have,

$$\begin{aligned} &= \int P(O_{s,t}, C_{s,t} | \text{do}(i, h_s, m_t)) dC_{s,t} \\ &= \int P(O_{s,t} | C_{s,t}, \text{do}(i, h_s, m_t)) P(C_{s,t} | \text{do}(i, h_s, m_t)) dC_{s,t} \\ &= \int \underbrace{P(O_{s,t} | C_{s,t}, h_s, m_t)}_{\text{compute w/ data}} \underbrace{P(C_{s,t} | \text{do}(i, h_s, m_t))}_{\text{compute w/ climate model}} dC_{s,t} \end{aligned}$$

We emphasize that although H_s and M_t are unobserved, the quantities above can be estimated because we have multiple observations for each latent pair h_s, m_t . See [1] for further details.

Strategy

In line with the estimand we adopt the following strategy:

1. Use a physics-based climate models to compute

$$P(C_{s,t}^c | \text{do}(i, h_s, m_t))$$

where $C_{s,t}^c$ might be a different but related set of climate variables.

2. Use a probabilistic "calibration model" to link $C_{s,t}^c$ with $C_{s,t}$. This is justified under the reasonable assumptions that there is no backdoor from $C_{s,t}^c, H_s, M_t, I$ to $C_{s,t}$ and that $C_{s,t} \perp M_t, I | C_{s,t}^c, H_s$ in which case $P(C_{s,t} | \text{do}(i, h_s, m_t))$ is equal to

$$\int P(C_{s,t} | C_{s,t}^c, h_s) P(C_{s,t}^c | \text{do}(i, h_s, m_t)) dC_{s,t}^c$$

We use this expression by estimating $P(C_{s,t} | C_{s,t}^c, h_s)$ from data.

3. Use a probabilistic model to compute

$$P(O_{s,t} | C_{s,t}, h_s, m_t)$$

Although h_s and m_t are not observed, we can train a model to predict $P_{\text{state, month}}(O_{s,t} | C_{s,t})$ which is computable from the data. From the the SCM then

$$P_{\text{state, month}}^{\text{obs}}(O_{s,t} | C_{s,t}) = P(O_{s,t} | C_{s,t}, h_s, m_t)$$

Part 1: Climate Model

We use the Canadian Earth System Model [16] to obtain point estimates of the distribution $P(C_{s,t} | \text{do}(i, h_s, m_t))$. We define the interventions I to be 4 different socioeconomic pathways specifying increasingly worse scenarios $\{\text{SSP}_1, \text{SSP}_2, \text{SSP}_3, \text{SSP}_5\}$. For each time-point we spatially average the predictions by state.

Part 2: Outbreak Models

We used 4 model classes to estimate the distribution $P(O_{s,t} | C_{s,t}, h_s, m_t)$. All hyper-parameters were chosen by maximizing log-likelihood on a held-out validation set. Throughout, s and m represent one hot encoded variables corresponding to the state and month.

Neural Poisson Regression: A Poisson regression parameterized by a neural network.

$$\begin{aligned} \theta &\sim \mathcal{N}(0, \sigma^2 I_p) \\ O_{s,t} &\sim \text{Poisson}(\exp(f_\theta(C_{s,t}, s, m))) \end{aligned}$$

Where f is a neural network. We refer to the one-layer version as the **Poisson** model as it equivalent to a traditional linear model with poisson likelihood.

Spatial Gaussian Process: We propose a model that explicitly models the latent confounders H_s and M_t using latent processes. In particular, if $C_{s,t} \in \mathbb{R}^d$, we assume that $H_s \in \mathbb{R}^d$, $M_t \in \mathbb{R}$ and that the vector $H^j = (H_{s_1}^j, \dots, H_{s_{|S|}}^j)$ is jointly distributed as a normal distribution with covariance

$$K(\lambda)_{s,s'} = \text{Cov}(H_s^j, H_{s'}^j) = \exp\left(-\frac{\|x_s - x_{s'}\|^2}{2\lambda^2}\right)$$

where the super-script ^{j} means the j -th coordinate of the vector and x_s represents the average longitude and latitude of each state. Formally, the sampling process for the model is the following (we omit the bias term for clarity):

$$\begin{aligned} \lambda &\sim \text{Gamma}(\alpha', \beta') \\ M_k &\sim \mathcal{N}(0, \tau^2) \quad \text{for } k = 1, 2, \dots, 12 \\ H^j &\sim \mathcal{N}(0, K(\lambda)) \quad \text{for } j \in [d] \\ O_{s,t} &\sim \text{Poisson}(\exp(H_s^\top C_{s,t} + M_t \bmod 12)) \end{aligned}$$

Conditional Normalizing Flow: We use neural spline conditional normalizing flow as described in [4]. Formally, our sampling process is described as

$$\begin{aligned} \theta &\sim \mathcal{N}(0, \sigma^2 I_p) \\ U_{s,t} &\sim \mathcal{N}(0, 1) \\ O'_{s,t} &= f_\theta(U_{s,t}; C_{s,t}, s, m) \\ O_{s,t} &= \text{Nearest-int}(\text{ReLu}(O'_{s,t})) \end{aligned}$$

where $f_\theta(\cdot; C_{s,t}, s, m)$ is an invertible neural network as specified by [4]. The last line of the process is so that we can convert continuous values to discrete ones.

Inference

We used SVI to train all of the models except the Normalizing Flow for which we used a special procedure detailed below. For the other models the variational family was chosen as described by ADVI [8] except when using a neural network in which case used a delta distribution. Optimization was done always using Adam.

Normalizing Flow: To handle the discrepancy between continuous and count data, we used MAP on a noised version of the standard normalizing flow objective. In particular, we maximized an estimate of the objective

$$\mathbb{E}_{U_{s,t} \sim P(U)} \left[\sum_{s \in \mathcal{S}, t \in \mathcal{T}_p} \ell(x_{s,t}, O_{s,t}, U_{s,t}, \theta) + p(\theta) \right]$$

where we have used the following shorthands.

$$\begin{aligned} x_{s,t} &= (C_{s,t}, s, m_t) \\ \ell(\tilde{x}, \tilde{o}, \tilde{u}, \theta) &= \log p_{\mathcal{N}}(f_\theta^{-1}(\tilde{o} + \tilde{u}; \tilde{x})) + |\det J_{f^{-1}(\cdot; \tilde{x})}(\tilde{o} + \tilde{u})| \\ P(U) &= \text{Uniform}[-2, 2] \end{aligned}$$

where J is the Jacobian and $P_{\mathcal{N}}$ denotes the p.d.f of a normal distribution. Despite how complicated it looks, the objective is simply the MAP estimate of a standard conditional normalizing flow but using noise terms U to correct for the discreteness of the data.

Part 3: Calibration Model

For our calibration model we also used NFs but opted for autoregressive conditional flows[7] as they produced higher out-of-sample log-likelihoods. We also attempted to use neural normal models for our likelihood but discarded them due to lower performance. We trained the model on the years 1997 to 2013 and tested on the year 2014. We used this range as it constitutes the historical predictions for the climate model. The average R^2 for predicted variables in 2014 was 0.87.

Evaluation

To evaluate the fitness of each outbreak model, we used R^2 , posterior predictive checks (PPCs), and calibration plots. We don't use predictive log-likelihood because the Conditional Normalizing Flow doesn't have a straightforward log-likelihood.

Model	R^2
Poisson	0.508
Neural Poisson	0.653
Spatial GP	0.653
Normalizing Flow	0.656

Table 1. Comparison of R^2 Scores for Different Models

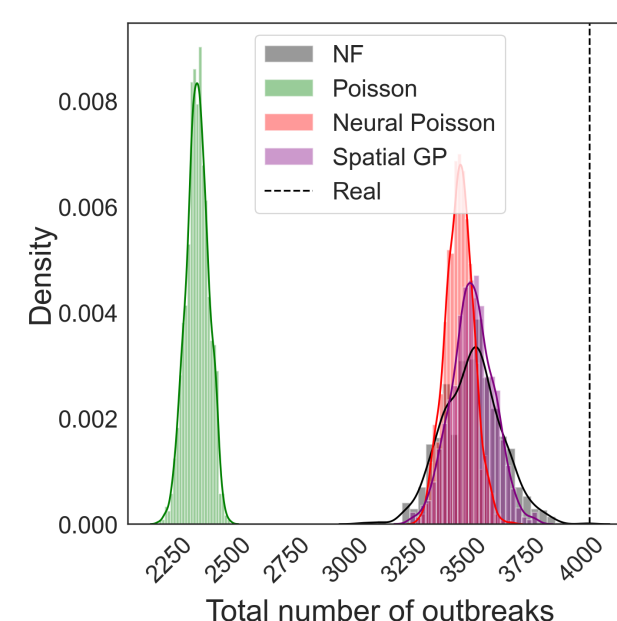


Figure 2. PPC for total number of outbreaks

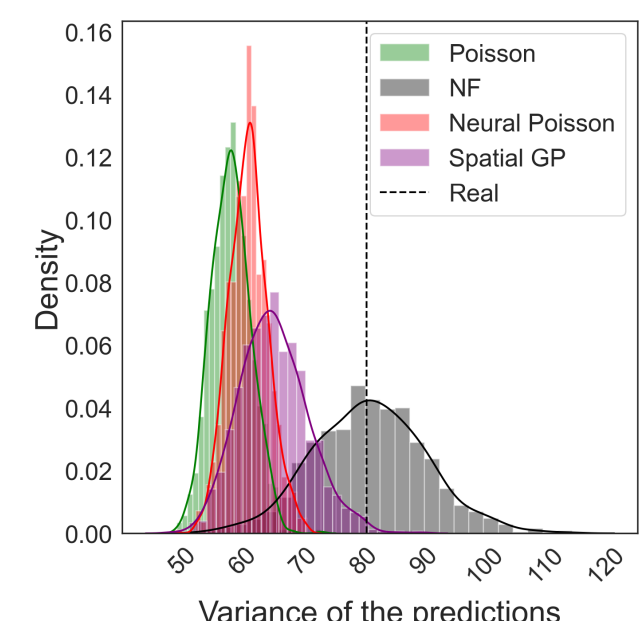


Figure 3. PPC for variance of outbreaks

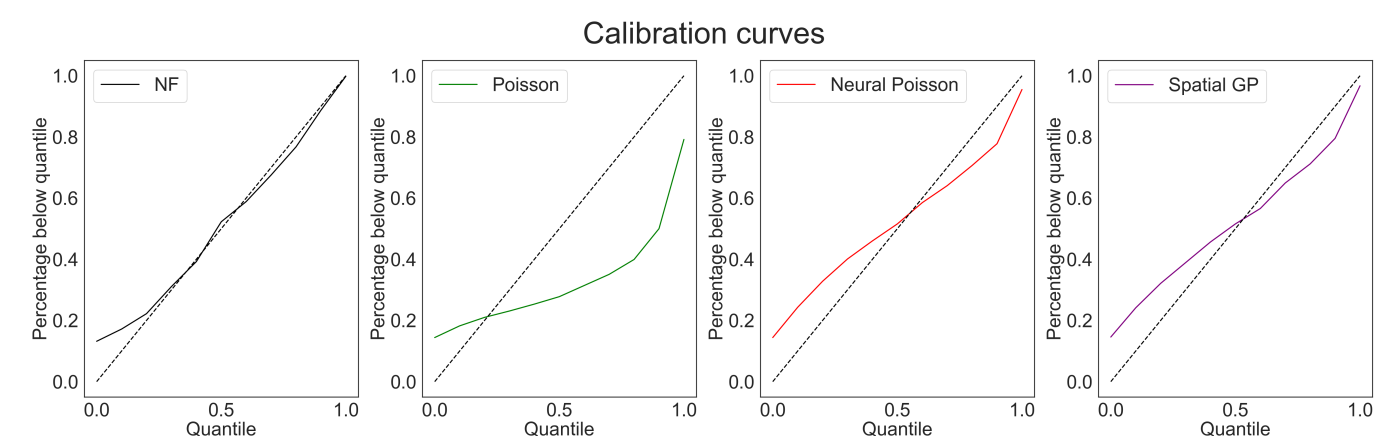


Figure 4. Calibration curves for outbreak models

Results

The best performing model, the normalizing flow, was used in our end-to-end sampling procedure for future outbreaks given each of the four climate scenarios I .

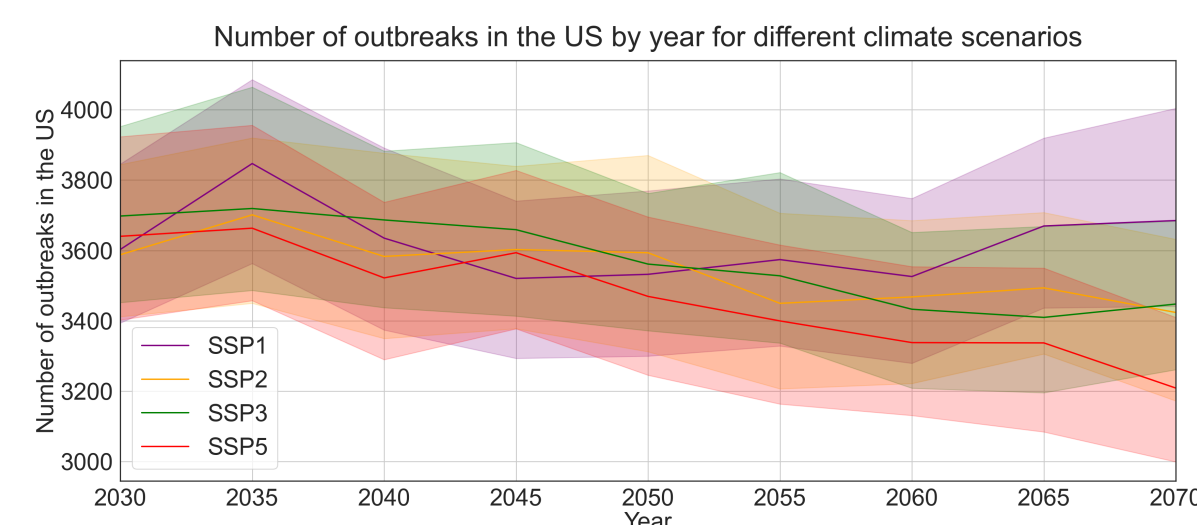


Figure 5. Predicted number of outbreaks per year per climate scenario

Predictions for future decades show high variance, with a downward trend becoming noticeably separated between best and worst climate scenarios by 2070. Although it is possible this is due to missing confounding factors, we believe this arises due to the fact that existing studies examining the climate's effect on pathogens often overlook the climate's impact on the human body. For instance, colder months weaken the immune system, leading to more disease-spread and outbreaks, indicating a negative relationship between factors like temperature and outbreak frequency. We also suspect our causal model may not be complete for the following reasons: (1) We lack data on extreme events that could trigger outbreaks, such as storms or floods, making them hard to predict accurately. (2) Rising pathogen populations don't necessarily mean more outbreaks, as advances in technology, medicine, and education may counteract increased pathogenicity. (3) Changes in governmental funding or legislation can impact outbreak reporting and confirmation.

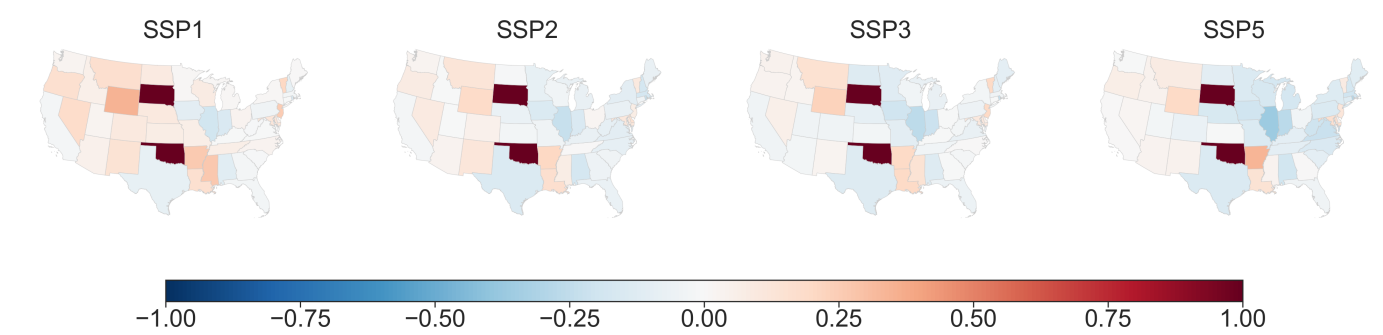


Figure 6. Percentage change in predicted number of outbreaks in 2070 per climate scenario

We assess each state's predicted percentage change in outbreaks from the training data average. Though variance across states is limited, Oklahoma and South Dakota significantly deviate, showing large positive changes in all climate scenarios. Interestingly, both Oklahoma and South Dakota have received "F" grades for low reporting by the Center for Science in the Public Interest[2]. Our model robustly adjusts for this, recovering more realistic outbreak estimates per state even with incomplete data.

Future Work

Future work may include addressing lack of data sources on "extreme" climate events and expanding the causal model to include features like yearly investment into the investigative arms of the FDA. Moreover, targeting other observable downstream effects of illness-causing bacteria and viruses like testing samples at poultry processing plants[11] (instead of number of outbreaks) could help us understand the effect of climate change independent of human immune systems.