Extreme Value Theory Analysis of CDC Influenza Data

1 Introduction

I applied Extreme Value Theory (EVT) to model seasonal influenza peaks using CDC FluView ILINet data with 14 seasons (2010–2024) across 10 HHS regions. I compared the Generalized Extreme Value (GEV) distribution against a baseline SIR model for predicting peak influenza-like illness (ILI) activity. I use these data instead of more robust lab tests as it is broader and contains many more cases, whereas lab-confirmed cases are scarce.

2 Methods

Data: CDC FluView data with 4,620 weekly observations across 10 HHS regions yielded 140 seasonal peaks (region-season pairs). I used an 80/20 train-test split (112 training, 28 test peaks).

GEV Model: For block maxima, the GEV distribution is $F(x) = \exp\{-[1 + \xi(x - \mu)/\sigma]^{-1/\xi}\}, \quad 1+\xi(x-\mu)/\sigma > 0$ where μ is location, $\sigma > 0$ is scale, and ξ determines tail behavior ($\xi > 0$: heavy, $\xi = 0$: exponential (Gumbel, as discussed in class), $\xi < 0$: bounded). I fitted via maximum likelihood.

GPD Model: For threshold exceedances, the Generalized Pareto Distribution is $F(x) = 1-(1+\xi x/\sigma)^{-1/\xi}$. I used the 90th percentile ($u = 4.64\%$ ILI) as threshold.

SIR Model: As baseline, the SIR model predicts dynamics via $dS/dt = -\beta SI/N$, $dI/dt = \beta SI/N - \gamma I$, $dR/dt = \gamma I$ where $R_0 = \beta/\gamma$. For each test season, I fitted SIR and extracted predicted peaks.

3 Results

Model Performance on 28 Test Region-Seasons

Model	MAE	RMSE	MAPE (%)
GEV	**2.19**	**2.72**	**29.4**
SIR	6.28	7.09	95.3

The fitted GEV parameters ($\mu = 3.70$, $\sigma = 1.92$, $\xi = 0.045$) showed excellent fit (KS $p = 0.888$). ξ is positive, but when changing my train-test split, it would sometimes be negative, and was generally very close to 0. The 100-year return level in this case is 13.5% ILI. The GPD model ($\xi = -0.118$, $\sigma = 1.99$, KS $p = 0.919$) showed $P(\text{exceed } 9.3\%) = 0.0065$. The observed training maximum (13.4% ILI, 2017–18) aligns with GEV's 100-year estimate.

GEV Return Level Estimates

Return Period (years)	2	5	10	20	100
Return Level (% ILI)	4.5	6.8	8.2	9.7	13.5

4 Discussion

GEV substantially outperformed SIR (MAE: 2.19 vs. 6.28). This is expected because GEV directly models the statistical distribution of observed extremes, while SIR attempts mechanistic modeling with simplifying assumptions (homogeneous mixing, no waning immunity, no antigenic drift). The heavy-tailed GEV ($\xi > 0$) captures unusually severe seasons like 2017–18, which lighter-tailed distributions would underestimate. EVT provides robust probabilistic forecasts: the 10-year return level of 8.2% ILI would help planners allocate resources for expected severe seasons. Extreme Value Theory has better results when characterizing influenza peak distributions across multiple regions and seasons. It gives better predictive accuracy and interpretable return levels that would be valuable for public health surveillance and resource planning.

I now plan to expand my cross-validation/evaluation and make a more extensive package for my code.