# Probabilistic Forecasting of Seasonal Influenza Peaks By Comparing Extreme Value Theory and SIR

**Abstract**

Predicting the magnitude of seasonal influenza peaks is important for public health resource allocation. I compare two approaches: Extreme Value Theory (EVT) with the Generalized Extreme Value (GEV) distribution and the mechanistic Susceptible-Infected-Recovered (SIR) compartmental model. I use 14 seasons (2010–2024) of CDC FluView data [**?**] across 10 US HHS regions and generate probabilistic peak forecasts evaluated with proper scoring rules [**?**]. Depending on how I evaluate performance, SIR may perform better on test data (CRPS: 0.996 vs 1.795), or GEV may perform better (Log-Score: 2.622 vs 7.682).

## 1 Introduction

Seasonal influenza peak intensity varies dramatically between seasons (2–13% ILI in 2010–2024). I compare statistical (Extreme Value Theory) vs mechanistic (SIR) approaches for probabilistic forecasting using CDC FluView data: 14 seasons (2010–2024) across 10 HHS regions, which gives 4,620 weekly observations and 140 seasonal peaks (Figure **??**). Flu seasons span week 40 through week 20 of the following year. The GitHub link can be found at [**?**].

## 2 Model Description

### 2.1 Generalized Extreme Value (GEV) Model

Extreme Value Theory provides a rigorous mathematical framework for modeling the behavior of maxima [**?**]. The Generalized Extreme Value distribution governs the asymptotic distribution of block maxima (in our case, seasonal peaks).

The GEV distribution governs block maxima with CDF: $F(x) = \exp\{-[1 + \xi(x - \mu)/\sigma]^{-1/\xi}\}$ where $\mu$ is location, $\sigma > 0$ is scale, and $\xi$ is the shape parameter controlling tail behavior ($\xi > 0$: heavy tail, $\xi = 0$: exponential, $\xi < 0$: bounded). The $N$-year return level $x_N = \mu - (\sigma/\xi)[1 - (1 - 1/N)^{-\xi}]$ gives the value exceeded once per $N$ years on average. I fit parameters via maximum likelihood (scipy.stats.genextreme). For predictions, I draw 200 samples from the fitted GEV to quantify uncertainty. Figure **??** shows GEV diagnostics.

### 2.2 Susceptible-Infected-Recovered (SIR) Model

The compartmental SIR model [**?**] divides the population into three groups with dynamics:

$$\frac{dS}{dt} = -\beta\frac{SI}{N}, \quad \frac{dI}{dt} = \beta\frac{SI}{N} - \gamma I, \quad \frac{dR}{dt} = \gamma I \tag{1}$$

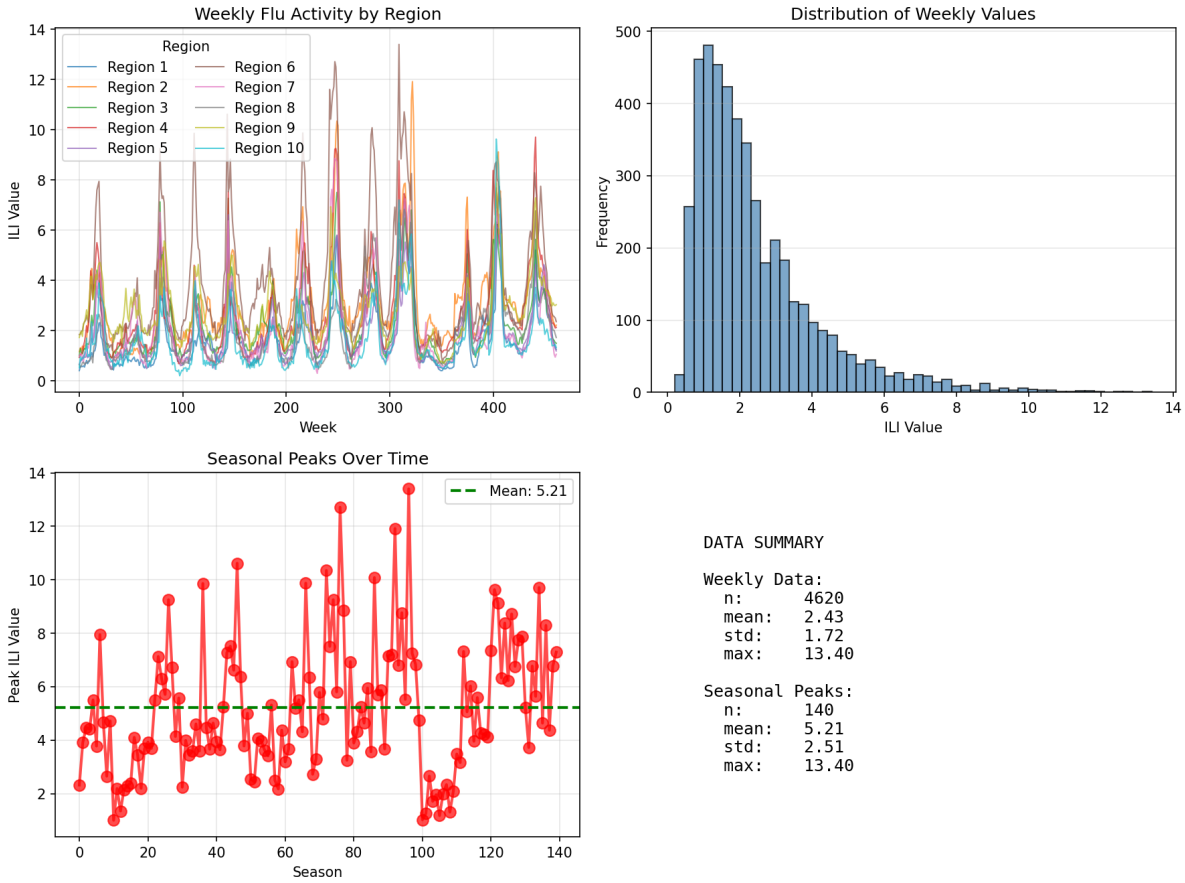


Figure 1: CDC FluView data overview: time series of %ILI across 10 HHS regions (2010–2024) and distribution of seasonal peaks. The 2017-18 season shows the historical maximum of 13.4% ILI.

where $S, I, R$ are susceptible, infected, and recovered counts; $N$ is total population; $\beta$ is transmission rate; $\gamma$ is recovery rate; and $R_0 = \beta/\gamma$ is the basic reproduction number. Peak occurs when $dI/dt = 0$, i.e., $S = N/R_0$.

I fit $(\beta, \gamma)$ to the first 20 weeks of each test season by minimizing $\sum(I_{\text{obs}} - I_{\text{model}})^2$ via L-BFGS-B, then integrate ODEs to find predicted peak (Figure **??**). For uncertainty quantification, I bootstrap resample the 20-week data 200 times, refit each sample, and predict peaks to obtain a distribution.

# 3 Analysis and Results

## 3.1 Experimental Design

I use temporal train/test split: 120 training region-seasons (2010–2021) and 20 test region-seasons (2 final seasons $\times$ 10 regions). SIR serves as baseline because it's the standard mechanistic epidemic model, and the goal is to see if GEV performs better in predicting seasonal peak magnitude.

I evaluate with proper scoring rules: First, CRPS (Continuous Ranked Probability Score, $\text{CRPS} = \mathbb{E}_F|X - y| - \frac{1}{2}\mathbb{E}_F|X - X'|$, lower is better) generalizes MAE to probabilistic predictions; Next, Log Score ($-\log f(y)$ via KDE, lower is better) heavily penalizes wrong confident predictions. I present more detailed interpretations of these metrics in context when discussing results.

I also report MAE and RMSE using distribution medians.

## 3.2 Model Fitting Results

GEV fitted to 120 training peaks: $\mu = 3.76$, $\sigma = 1.90$, $\xi = 0.024$ with KS $p = 0.788$ (excellent fit). The slightly positive $\xi$ indicates heavy tail (Fréchet domain), but different train-test splits reveal that this parameter does not stray meaningfully from 0. Return levels: 10-year = 8.2% ILI, 100-year = 13.0% ILI (nearly matches observed 2017–18 maximum of 13.4%). Given the data (2010–2024), this is a reasonable prediction. However, it's worth noting that the 2009–2010 flu season, which is not included in this data, was also one of the worst in recent history, and would certainly increase the ILI of the 100-year return level.

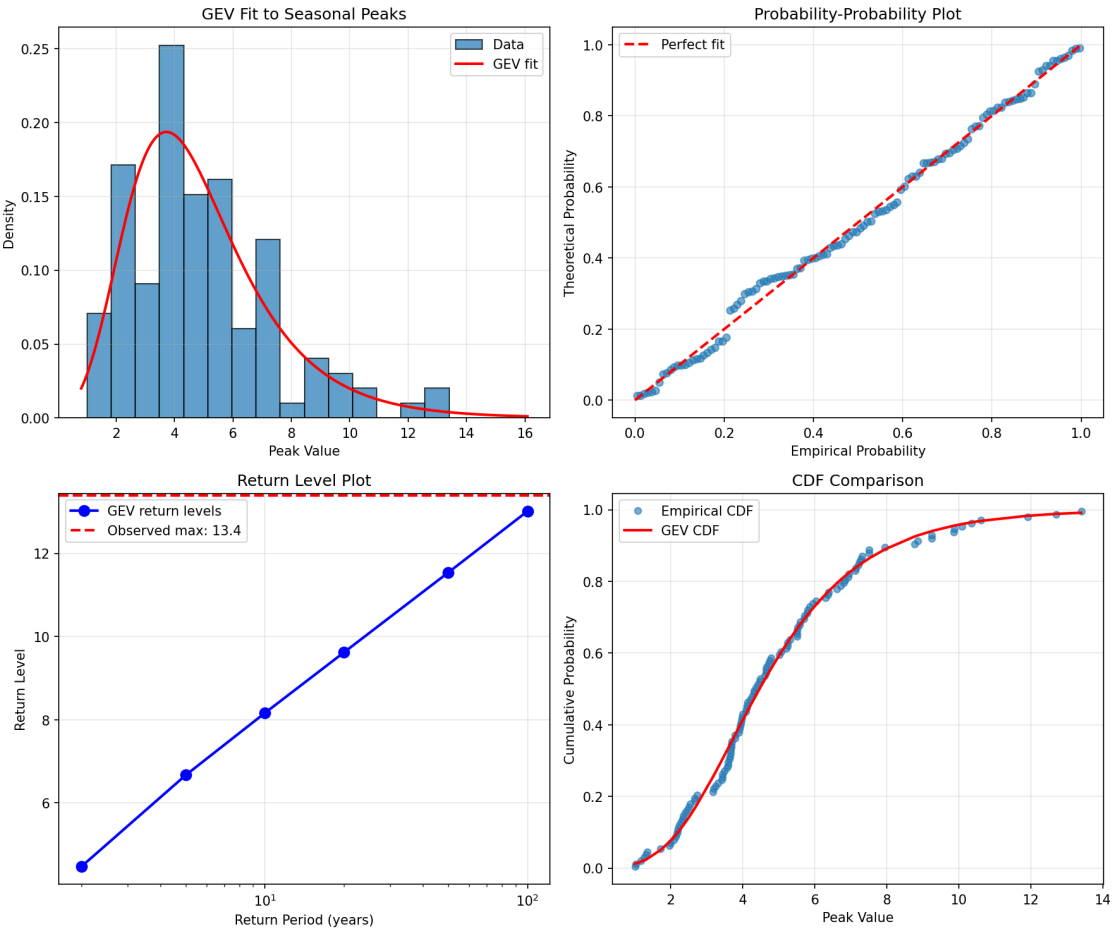


Figure 2: GEV model diagnostics: fitted PDF and CDF to 120 training peaks, and return level plot showing 10-year (8.2%) and 100-year (13.0%) values.

SIR example (Region 6, year 13, 20-week fit): $\beta = 0.50$, $\gamma = 0.38$, $R_0 = 1.30$, $R^2 = 0.45$. Modest $R^2$ reflects model simplicity but $R_0 > 1$ correctly identifies epidemic spread. Keep in mind that this season corresponds to the last season in Figure **??**, which is a relatively more peaked distribution than others like season 2 or 11. These flatter seasons cause the model to identify $R_0 < 1$, so the epidemic does not spread.

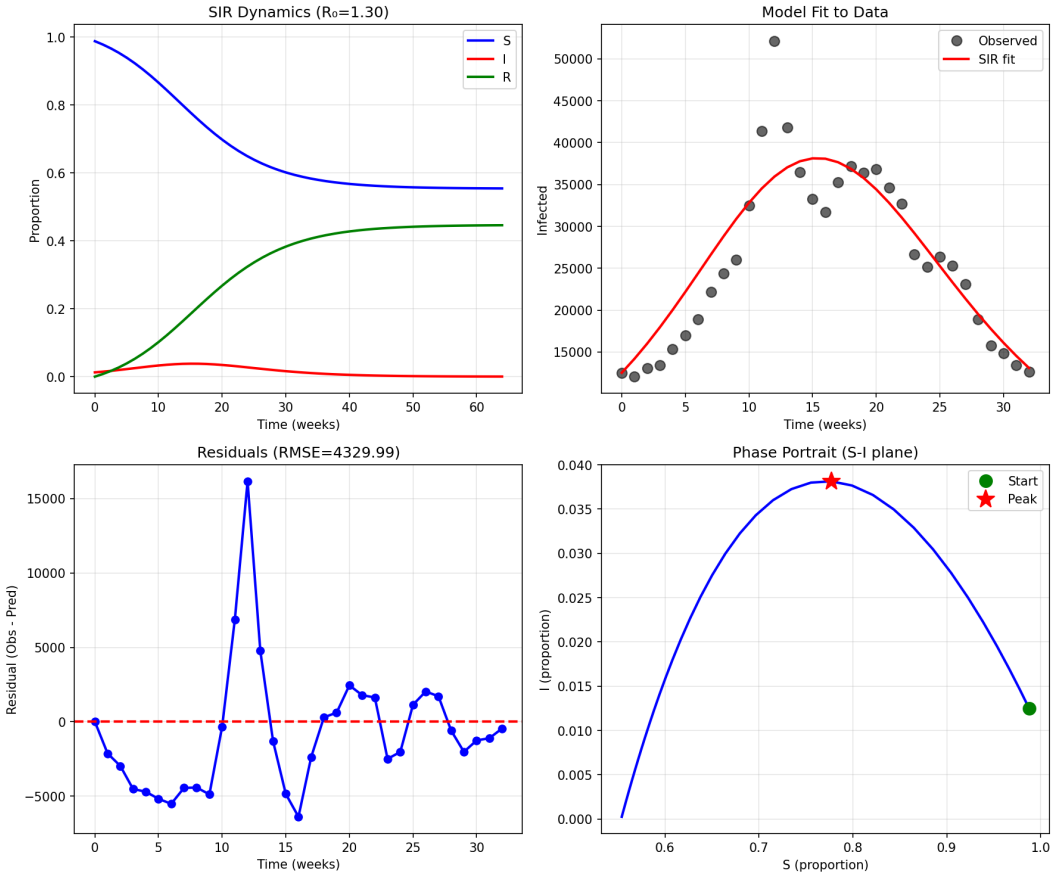


Figure 3: SIR model fit to early-season data (first 20 weeks) with predicted peak from full trajectory. Bootstrap uncertainty bands shown around the mean prediction.

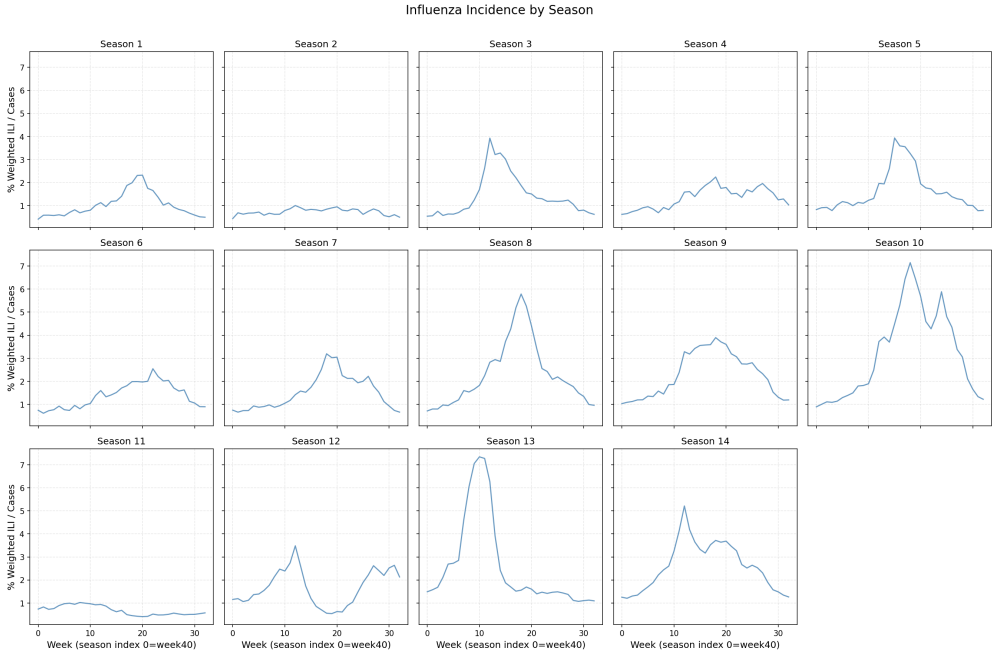


Figure 4: Influenza Incidence by Season for 2010-2024 in Region 6, with each season starting from week 40 and ending in week 20 of the next year.

## 3.3 Predictive Performance

Table 1: Model Comparison on 20 Test Region-Seasons

| Model | CRPS↓ | Log Score↓ | MAE↓ | RMSE↓ |
|---|---|---|---|---|
| **SIR** | **0.996** | 7.682 | **1.19** | **1.31** |
| GEV | 1.795 | **2.622** | 2.74 | 3.13 |

SIR seems to outperform GEV: 45% lower CRPS (0.996 vs 1.795) and 2.3× lower MAE (1.19 vs 2.74). Note that since SIR is fit to the first 20 weeks of the season, so it may have enough data to "see" the peaks. However, in seasons like season 6–10, the SIR model is unable or just barely able to see the peak. The SIR model's performance began to decrease significantly when it was fit to less of the initial data. Meanwhile, the EVT, which does not need any of the current season's data to be fit, performs worse in all metrics except log score. It's clear from the comparison plot (Figure **??**) why this is the case. The EVT's distribution is much wider and therefore is more likely to predict the observed peak than the narrow SIR distribution, which is closer but narrower.

This shows mechanistic modeling with uncertainty quantification can outperform statistical approaches when early-season data is available. GEV predictions cluster near the distribution center while SIR adapts to each season's trajectory, which suggests initially using EVT and refining predictions with mechanistic models could be useful for predicting peak influenza magnitude.

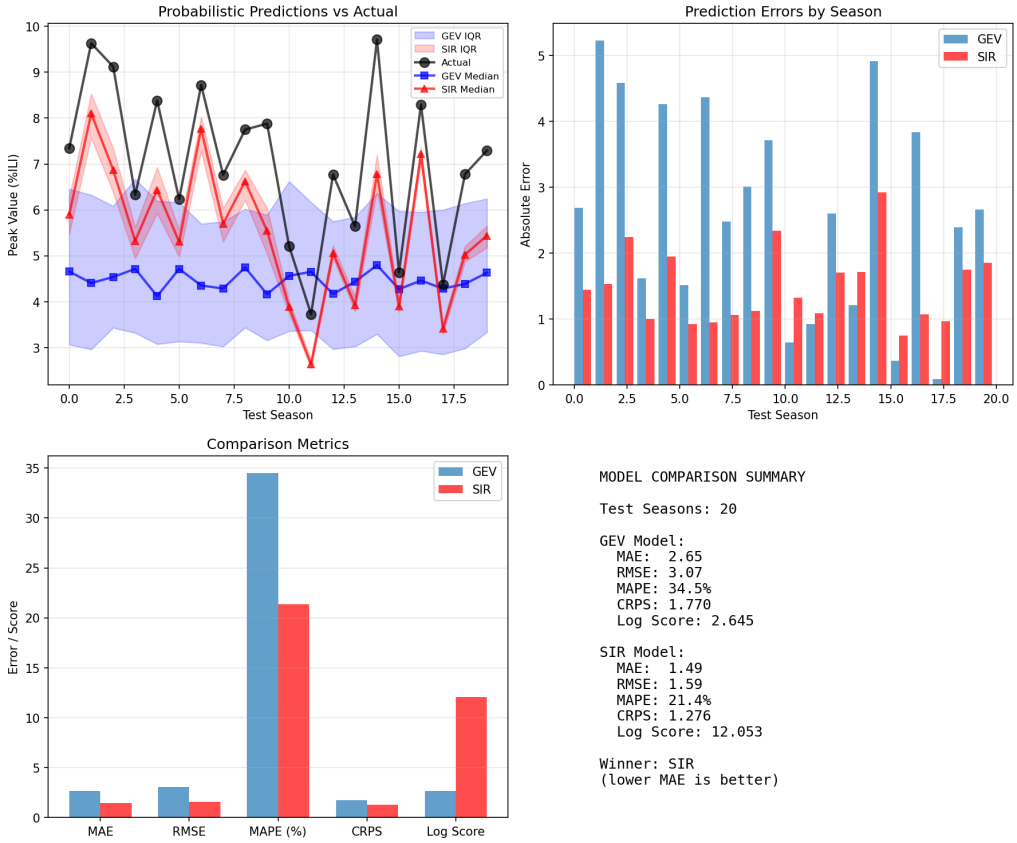


Figure 5: Probabilistic forecast comparison on 20 test region-seasons. SIR predictions (bootstrap distributions) adapt to individual season trajectories, while GEV predictions (samples from fitted distribution) remain centered around historical mean. Shaded bands show interquartile ranges.

# 4 Discussion

## 4.1 Why SIR Outperforms GEV

SIR's superior performance stems from using early-season data (first 20 weeks) to infer current transmission dynamics, while GEV uses only historical peak distributions. This makes SIR a within-season forecast (predicting peak given early observations) vs GEV's cross-season forecast (predicting next season with no current data). Early-season dynamics are highly informative about peak magnitude, so future work finding a way to incorporate these dynamics into the EVT model would likely produce a more powerful model.

The fitted GEV ($\xi = 0.024 \approx 0$) has near-Gumbel behavior. It treats all seasons as drawn from the same distribution and predicts similar values (median $\approx$ 4–5% ILI) regardless of specific season characteristics.

Bootstrap captures SIR uncertainty and reflects both parameter uncertainty (varying $\beta, \gamma$) and structural uncertainty (model misspecification), which gives reasonable CRPS despite simplifications. However, the SIR model clearly tends to underestimate the true peak for all seasons, even though it follows the general trend. I noticed this trend even when fitting to all of the data, which suggests a correction factor may be helpful in getting the SIR predictions to match observed predictions.

## 4.2 Limitations and Future Work

The SIR model assumes homogeneous mixing, no waning immunity, no antigenic drift, no age/spatial structure, etc., and this can serve to explain the modest $R^2$. Extensions (SEIR+, age-structured, metapopulation) could improve fits.

The comparison favors SIR since GEV lacks current-season data. It's difficult to do pure pre-season forecasts with SIR, and it's not immediately obvious how to fit the GEV with seasonal data, although some form of Bayesian updating as the season progresses would likely lead to a model better than either of the ones tested here. Also, comparing time-series models (like ARIMA), machine learning models, or even ensemble methods [**?**] to these methods would be interesting. For example, the CDC's FluSight ensemble model would likely outperform either approach for influenza, but methods like these could complement or extend such models, particularly for other diseases.

For public health, SIR's MAE of 1.19% ILI enables better hospital capacity planning. Mechanistic interpretation ($R_0$, transmission rates) aids intervention policy. GEV's 100-year return levels still valuable for worst-case planning.

# 5 Summary

I compared two approaches to probabilistic forecasting of seasonal influenza peaks: Extreme Value Theory (GEV) and the mechanistic SIR model. Using 14 seasons of CDC data across 10 regions, I evaluated models on held-out test seasons using proper scoring rules (CRPS and log score). The SIR model generally outperformed GEV (CRPS: 0.996 vs 1.795), primarily because it uses early-season observations to infer current transmission dynamics, while GEV makes predictions based solely on historical peak distributions. Both models provide uncertainty quantification via sampling (GEV) or bootstrap (SIR), which allow for probabilistic forecasts that support public health decision-making. Future work should explore hybrid approaches that combine mechanistic and statistical elements, fairer pre-season forecasting comparisons, and extensions to more complex compartmental models.

# 6 Attribution of Effort

This project was completed individually, so all work, including code, analysis, and writing, was done by the author.

# References


[1] Centers for Disease Control and Prevention (2018). *2017-2018 Influenza Season Week 21 ending May 26, 2018*. FluView Report.

[2] Gneiting, T., & Raftery, A. E. (2007). Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, 102(477), 359–378.

[3] Coles, S. (2001). *An Introduction to Statistical Modeling of Extreme Values*. Springer Series in Statistics.

[4] Kermack, W. O., & McKendrick, A. G. (1927). A contribution to the mathematical theory of epidemics. *Proceedings of the Royal Society A*, 115(772), 700–721.

[5] Shaman, J., & Karspeck, A. (2012). Forecasting seasonal outbreaks of influenza. *Proceedings of the National Academy of Sciences*, 109(50), 20425–20430.

[6] Maloney, C. (2025). *Code repository for "Probabilistic Forecasting of Seasonal Influenza Peaks By Comparing Extreme Value Theory and SIR."* Available at: `https://github.com/cmaloney111/sir-evt/tree/main`