# Modelling Outputs

In this section, we create 3 classes of models – (1) Univariate time series ARIMA (2) Generalised Linear Models GLM and, (3) Random Forest

## Generalised Linear Model (GLM)

We begin by creating a full model i.e. a model contain individual variables and their interaction effects. Then we eliminate variables one by one to find the most optimal model (with lowest AIC).

In doing so, we consider the following underlying distributions for the output variable ‘flood’ – (a) Normal (b) Poisson (c) Quasi-Poisson and, (d) negative binomial.

**Comparison of Best GLM Models by Family**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Family** | **Best Model Terms** | **Residual Deviance** | **AIC** | **Key Comments** |
| **Gaussian** | 5 main effects + 4 interactions (e.g., Electricity × Drainage) | 42,282 | 278.78 | Gaussian assumes constant error around predictions – which may not be valid for flood count |
| **Poisson** | 5 main effects + 10 interactions (e.g., Electricity × Urban Roads, GDP × Roads) | 167.64 | 335.24 | Many variables are significant, but errors are too large, thus, Poisson is not flexible enough. |
| **Quasi-Poisson** | *Intercept-only model* | 5519.9 | NA | Fails to model anything; confirms overdispersion |
| **Neg. Binomial** | Electricity, GDP, Electricity × GDP | **32.48** | **269.24** | fits flood counts well, handles over-variation, and has a simple, meaningful structure |

**Why choose Negative Binomial family?**

* Flood is a **count variable** (discrete, non-negative)
* Poisson model had **massive overdispersion** → invalid standard errors
* Quasi-Poisson confirms overdispersion but doesn’t allow AIC-based comparison
* Gaussian assumes symmetric residuals, which may not hold in your flood distribution
* Negative Binomial:
  + **Handles overdispersion explicitly** via dispersion parameter θ = 1.66
  + Provides **valid likelihood-based comparison (AIC)**
  + Yields **interpretable coefficients and interactions**

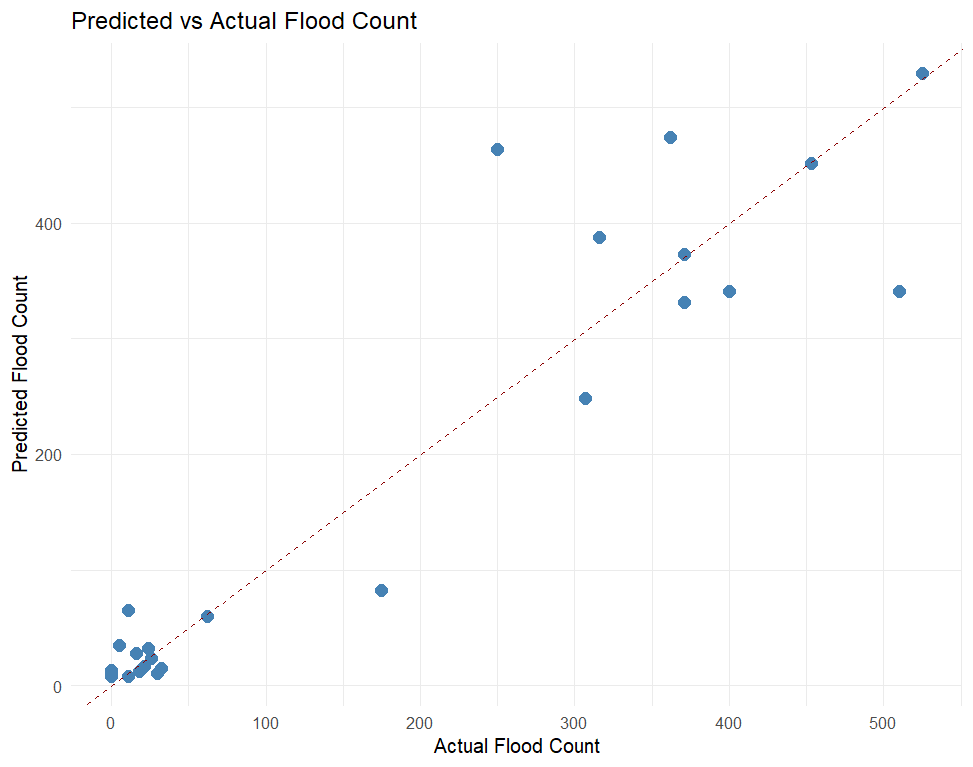
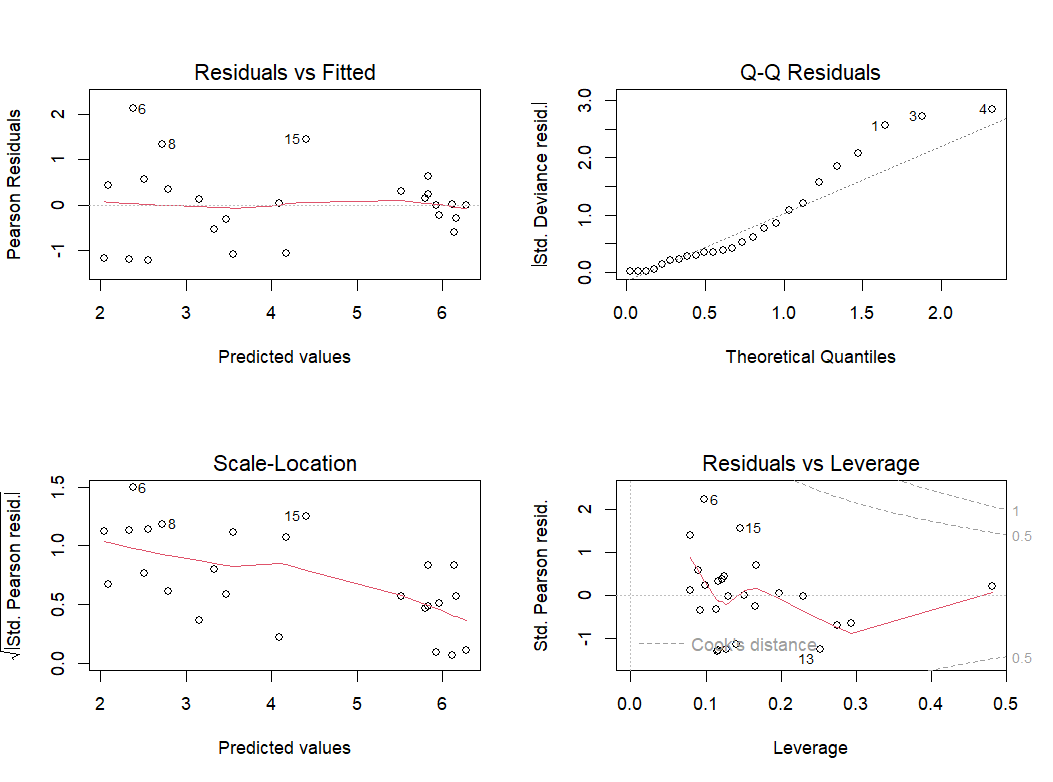
**In Context of Your Research**

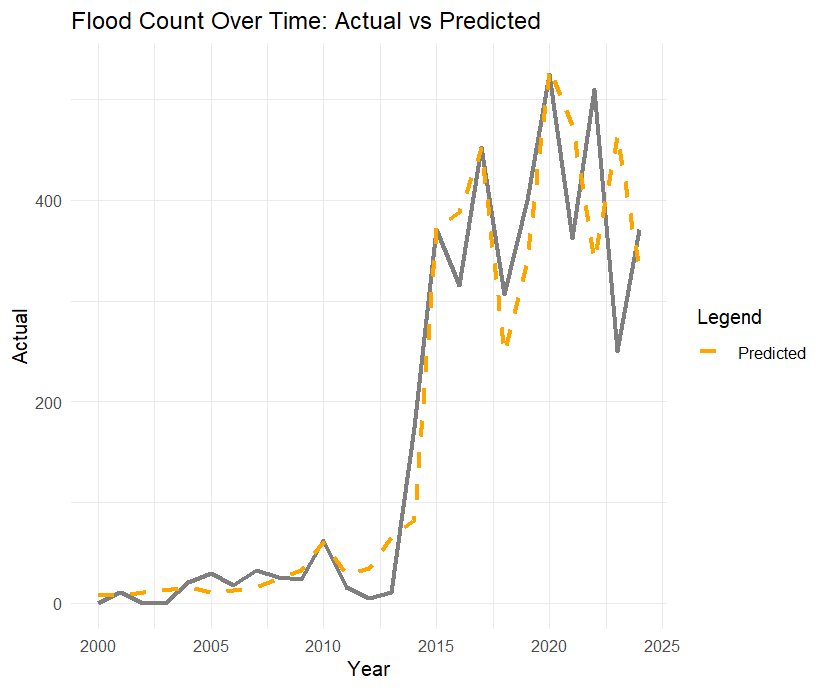
* Flood risk is driven by urban pressure (Electricity) but may be modulated by economic resources (GDP)
* Negative Binomial allows modelling non-linear count behavior and interaction effects in a city growing at varying paces
* Model reflects a nuanced view: urban growth → more floods, but GDP may offset this pressure through better planning, infrastructure, or resilience investment

|  |  |  |
| --- | --- | --- |
| **Term** | **Estimate** | **Interpretation** |
| **Intercept** (log-scale) | 2.0445 | Baseline log-floods at zero predictors (irrelevant by itself due to normalization) |
| **Electricity Consumption** | **+14.15** | Strong positive impact on flood count when GDP is low |
| **State GDP** | −1.74 | Alone, weak and non-significant (p = 0.52) |
| **Electricity × GDP (Interaction)** | **−8.65** | Significant: **GDP moderates the flood impact of electricity growth** |

Equation:

* When Electricity Consumption increases, flood count increases — representing urban sprawl, increased population density, and more impervious surfaces.
* But when State GDP is higher, the effect of electricity on flood count reduces.
* Suggests that wealthier states may invest in better infrastructure (e.g., drainage, zoning) or have more adaptive capacity.
* In low-GDP scenarios, uncontrolled growth leads to high vulnerability.



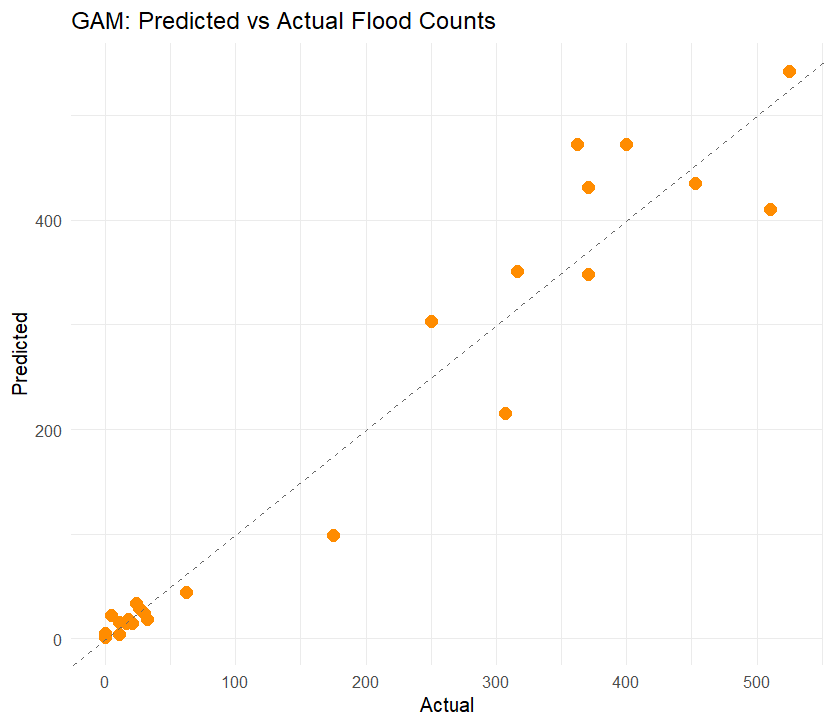
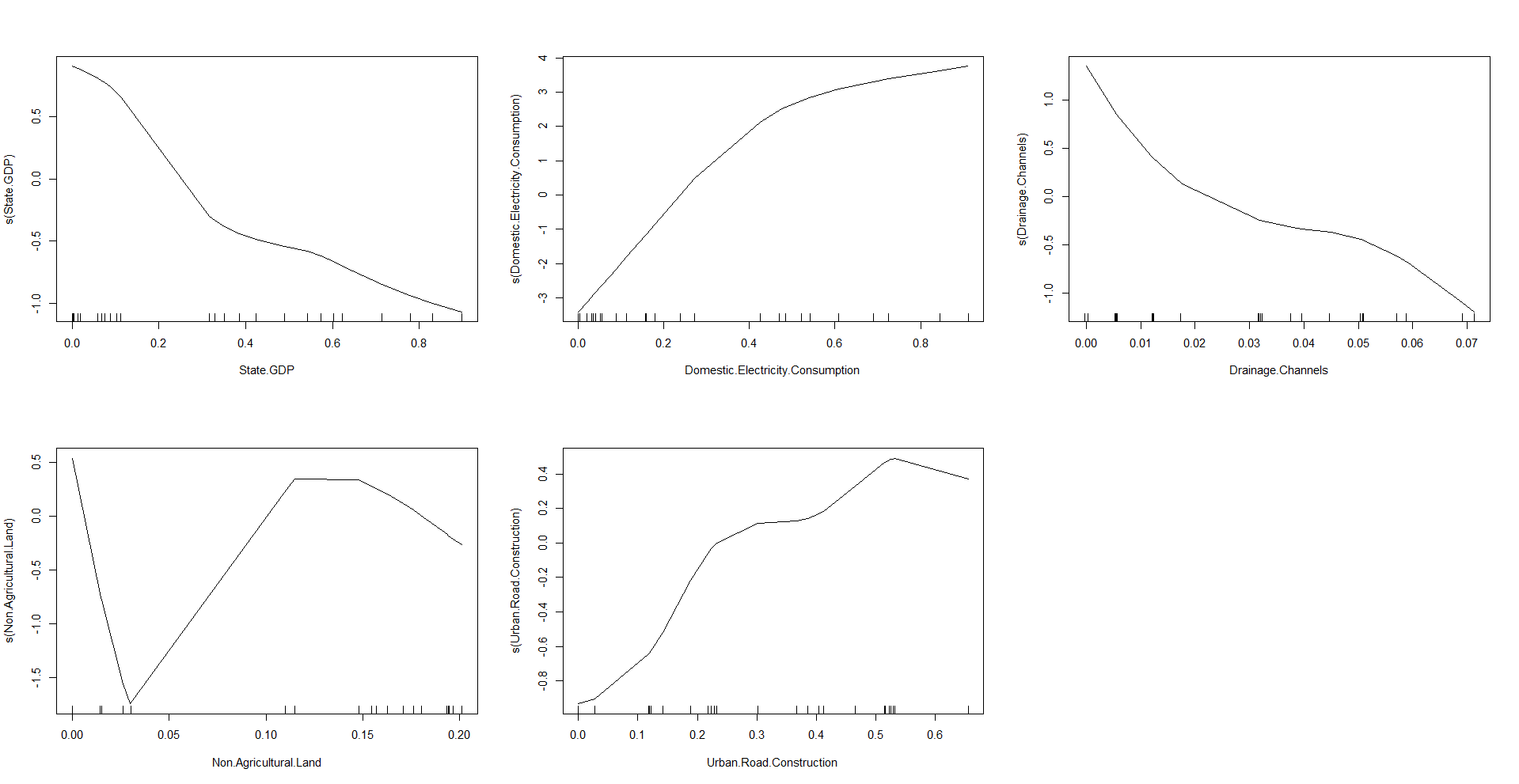
1. **Predicted vs Actual Flood count (Scatter plot)** - Most of the points lie close to the dashed diagonal line. However, slight over-prediction is visible for mid-range values (~200–400) and under-prediction for a couple of high-flood years. The model captures the overall trend well with low bias.
2. **Residual Diagnostic Plots**
   1. **Residual vs fitted** – ideally, residuals should be randomly scattered around 0. Here, there’s some structured (curvature), suggesting non-linearity not fully captured even after including interaction terms. This hints towards a potential benefit from a GAM model.
   2. **Q-Q Plot** - Deviations from the 45° line suggest that residuals aren’t perfectly normal, which is acceptable for count data.
   3. **Scale-Location** - Slight trend observed, with heteroscedasticity.
   4. **Residuals vs Leverage** - Points with high leverage and large residuals are influential observations. No strong outlier beyond Cook’s distance boundary, so no single year is distorting the model, which is good.
3. **Predicted vs Actual (over time)**
   1. The model tracks the general trajectory of actual floods very well (especially from 2010 onwards), there is some lag & smoothing of sharp spikes (like 2017 and 2022) indicating that the model generalizes well.
   2. The sudden increase in 2014 is well captured.
   3. The model’s underfit in earlier years might indicate that natural causes dominated earlier, while urban effects have become prominent more recently.

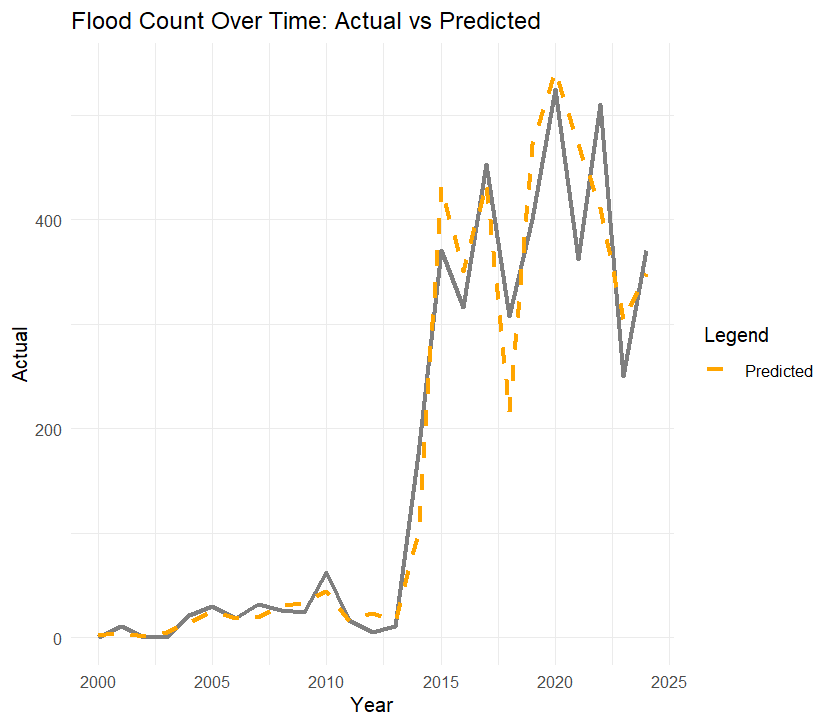
## Generalized Additive Models (GAMs)

We consider three families – Normal, Poisson and Negative Binomial and find the best models within each family. Results :

|  |  |  |  |
| --- | --- | --- | --- |
| **Metric** | **Gaussian GAM** | **Poisson GAM** | **Negative Binomial GAM** |
| **AIC** | 288.10 | 380.77 | **227.33** |
| **Residual Deviance** | 25468.76 | 201.18 | **11.39** |
| **Null Deviance** | 864573.4 | 5519.93 | **68.67** |
| **Dispersion Parameter** | 6366.85 | 1.00 | **1.5715** |
| **Best Predictor (Linear Term)** | GDP (p < 0.001) | GDP (p < 0.01) | GDP (p < 0.01) |
| **Non-linear Effects (Smooth Terms)** | Not significant | All smooths significant | None are significant |

1. **Negative Binomial GAM**
   1. Lowest AIC
   2. Dispersion parameter of 1.57 confirms over-dispersion which NB family is meant to handle.
   3. State GDP is the only statistically significant linear term, again reinforcing its dominant influence.
   4. None of the smooth terms are statistically significant indicating that most relationships might be linear, or that the dataset is too small to detect meaningful curvature.
   5. So far, we can conclude that rapid economic expansion is a strong proxy for urbanization-driven flood risk.
2. **Poisson GAM** 
   1. Overestimates deviance residuals, suggesting poor fit.
   2. All smooth terms (nonparametric) show strong significance — this contradicts NB’s results and may reflect Poisson’s sensitivity to distributional assumptions.
   3. AIC = 380.77, which is considerably worse than both Gaussian and NB.
3. **Gaussian GAM**
   1. Performs better than Poisson, but worse than NB.
   2. GDP and Electricity show significance among linear terms.
   3. Smoothing terms are not significant (p > 0.1), similar to NB.
   4. Suggests most predictors can be modelled with linearity, possibly better suited for GLM with interactions if interpretability is key.

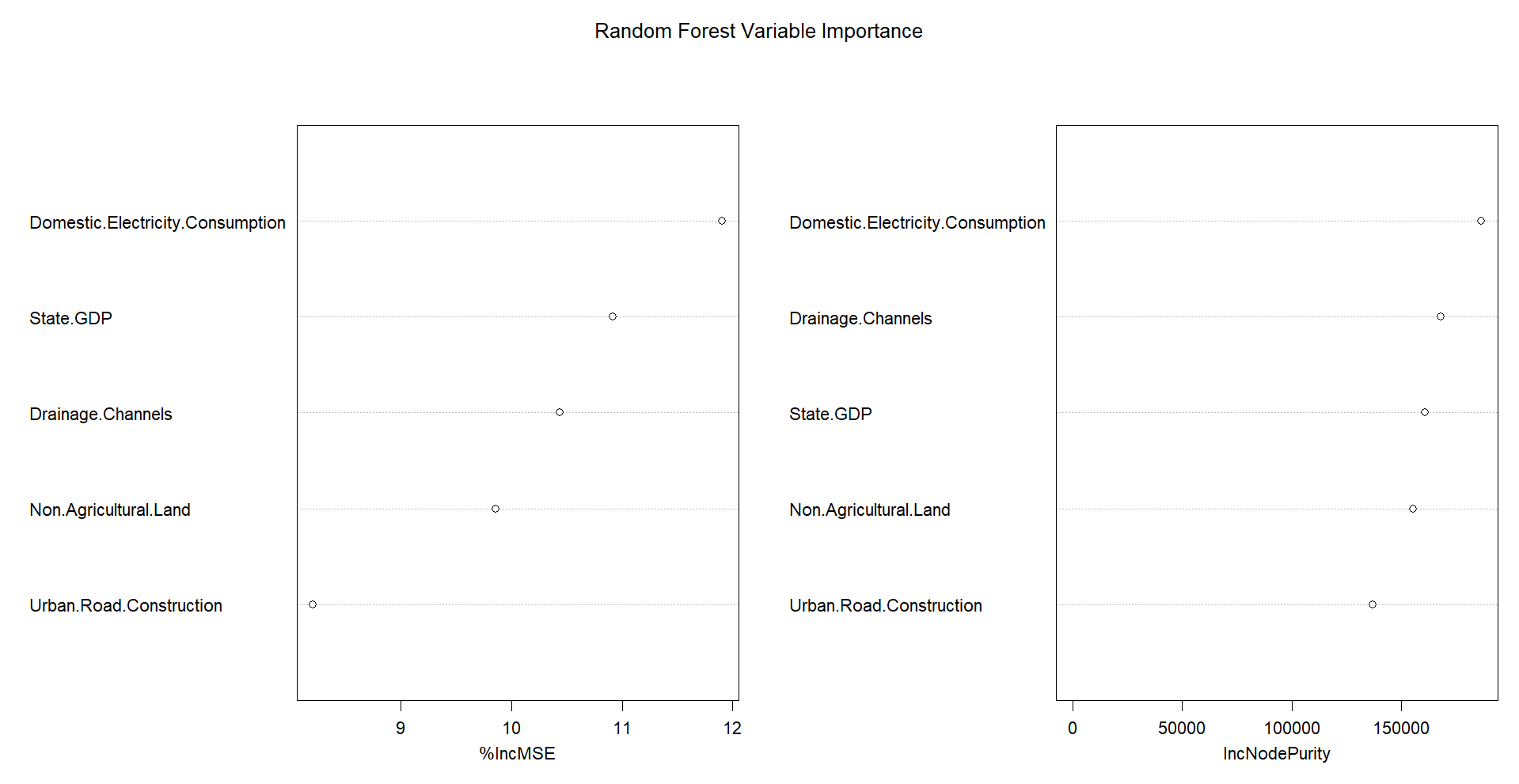


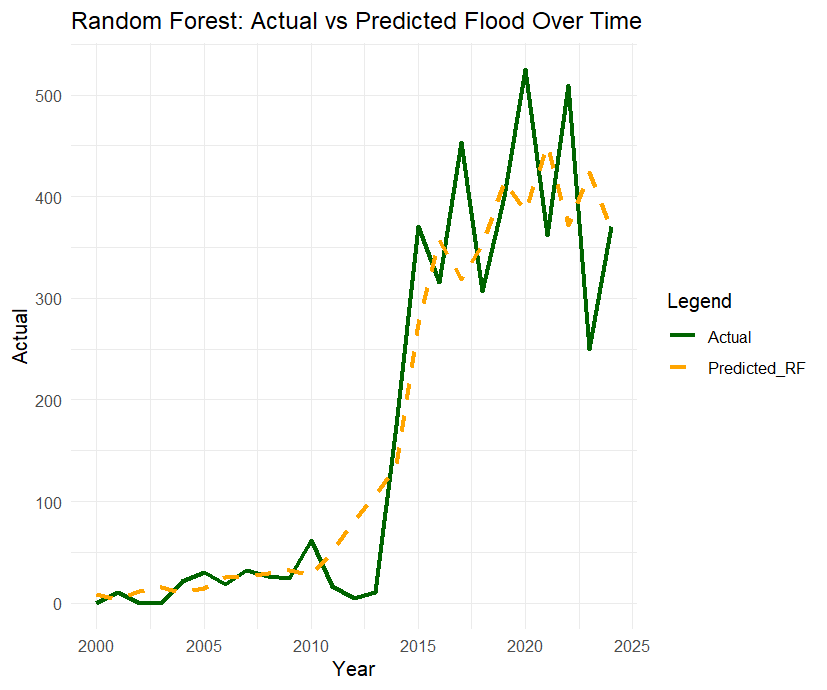
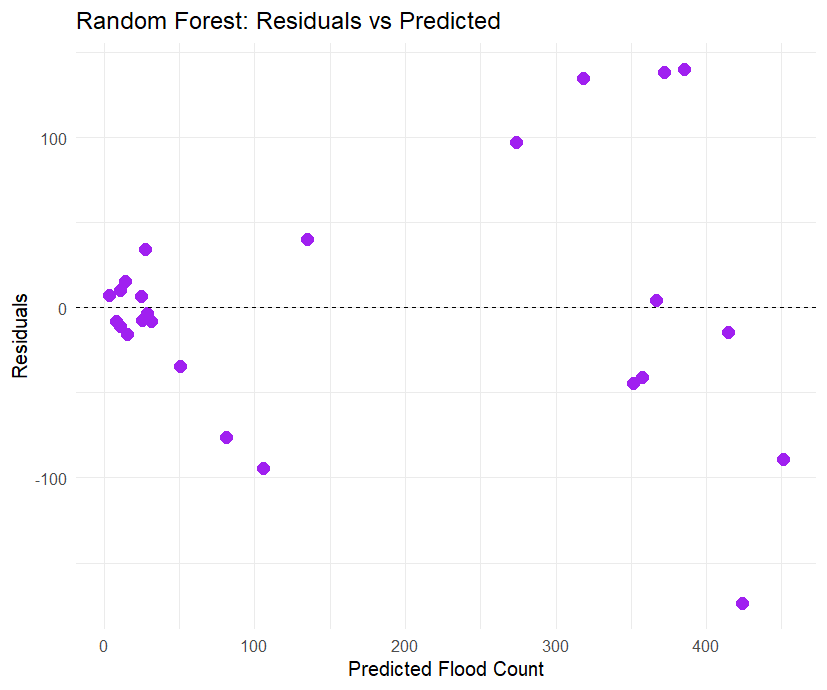
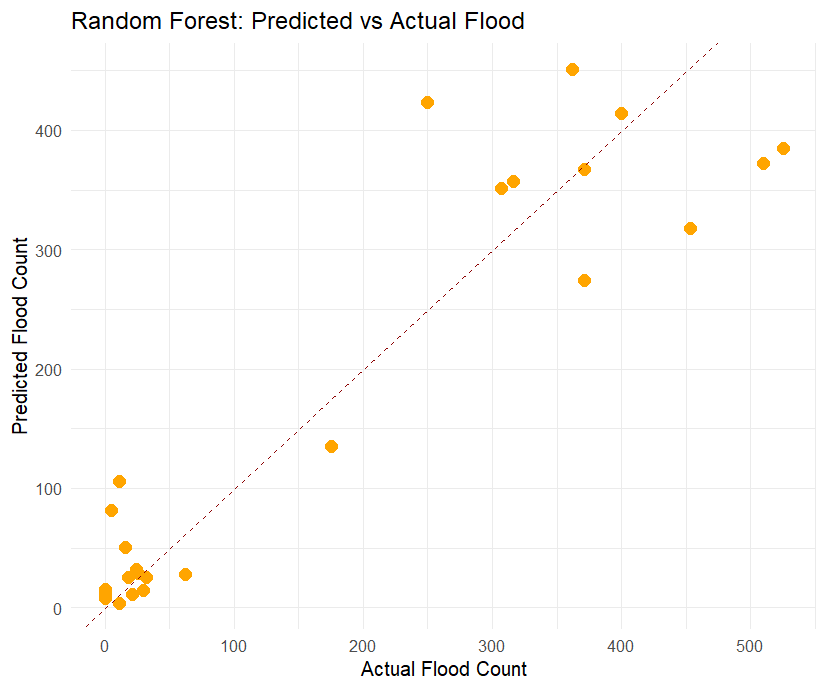


1. **Smooth Term Plot**
   1. **State GDP:** Decreasing effect → Higher GDP = Lower flood count? Possibly an artefact due to smoothing on normalized scale.
   2. **Electricity Consumption:** Clear upward curve → More consumption = More floods.
   3. **Drainage Channels:** Declining curve → Better drainage may reduce floods.
   4. **Non-Agricultural Land:** U-shaped → Low/very high proportions not helpful.
   5. **Urban Road Construction:** Increasing till a peak → Moderate infrastructure worsens flooding, but levels off.
2. **Predicted vs Actual Plot (scatter)**
   1. Most points align fairly well along the 45° reference line (dashed), suggesting a good fit.
   2. However, slight deviations exist in the mid-to-high range (e.g., some underestimation of mid-range actual flood counts around 200–300).
   3. Predictions for low flood counts are very tight and accurate (bottom left cluster).
   4. Conclusion: The model performs very well overall, especially in the tails. A few central outliers are tolerable and expected
3. **Predicted vs Actual Plot (Over time)**
   1. The orange dashed line (predictions) closely tracks the grey line (actuals) — showing strong temporal alignment.
   2. Slight lag in years like 2017–2018 and overshooting in early 2020s, but the broader shape is maintained.
   3. Sharp jump post-2013 in both actual and predicted flood counts is well captured, which is crucial for policy inference.

## Random Forest (RF)

|  |  |
| --- | --- |
| **Model Structure** | **Metric** |
| % of variance explained | 85.2% (Good Fit) |
| Mean Squared Residuals | 5116.9 |
| Number of trees | 500 |
| Max Features per split | 1 |



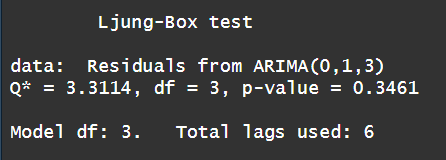
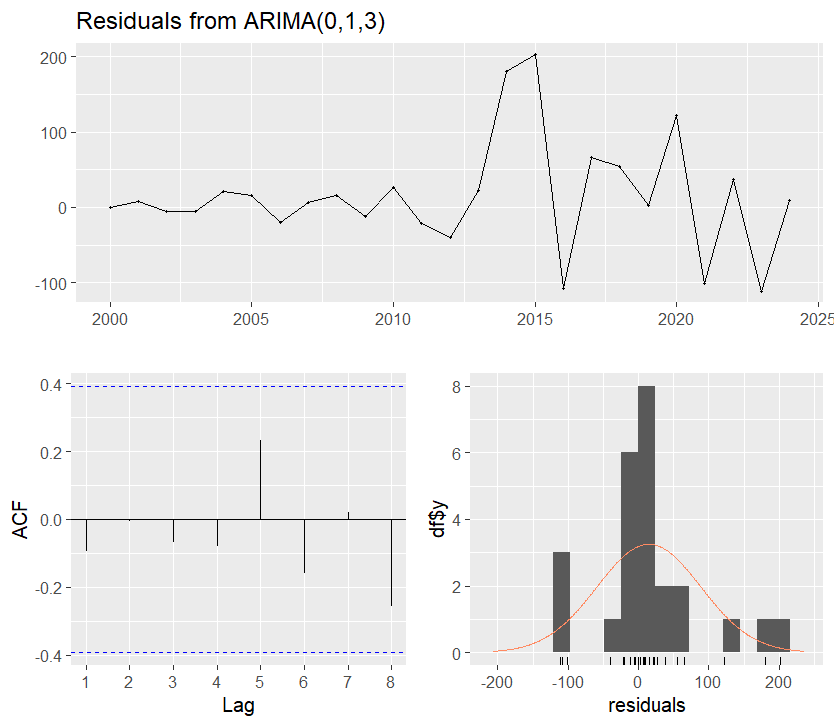


1. **Variable Importance Plot**
2. Top predictors are Domestic Electricity consumption, state GDP & Drainage Channels, while Urban roads are relatively less important.
3. This largely aligns with our GLM/LASSO results and supports their inclusion in final model selection.
4. **Predicted vs Actual (Scatter plot)**
5. The points largely follow the diagonal line, especially for moderate-to-high flood counts.
6. Slight underestimation for extremely high values and overestimation for a few lower ones.
7. This suggests the model generalizes well but may struggle slightly with extremes — expected given the relatively small dataset.
8. **Residual vs Predicted Plot**
9. Residuals are somewhat scattered, but there is no strong pattern — a good sign.
10. A few residuals exceed ±100, indicating occasional poor fits, likely for abrupt spikes in flood count.
11. The spread seems wider for mid-range predictions (~200–400), hinting that the model occasionally struggles with the volatile middle regime of flood count growth.
12. **Predicted vs Actual (Over time)**
13. The dashed predicted line closely follows the actual flood line.
14. Peaks and troughs are generally well captured, especially after 2014, when flood counts surged.
15. There’s slight smoothing — expected from ensemble models like RF — but no major lag or drift.
16. This is one of the strongest pieces of evidence that the model is learning structural temporal patterns, despite not being explicitly a time series model.

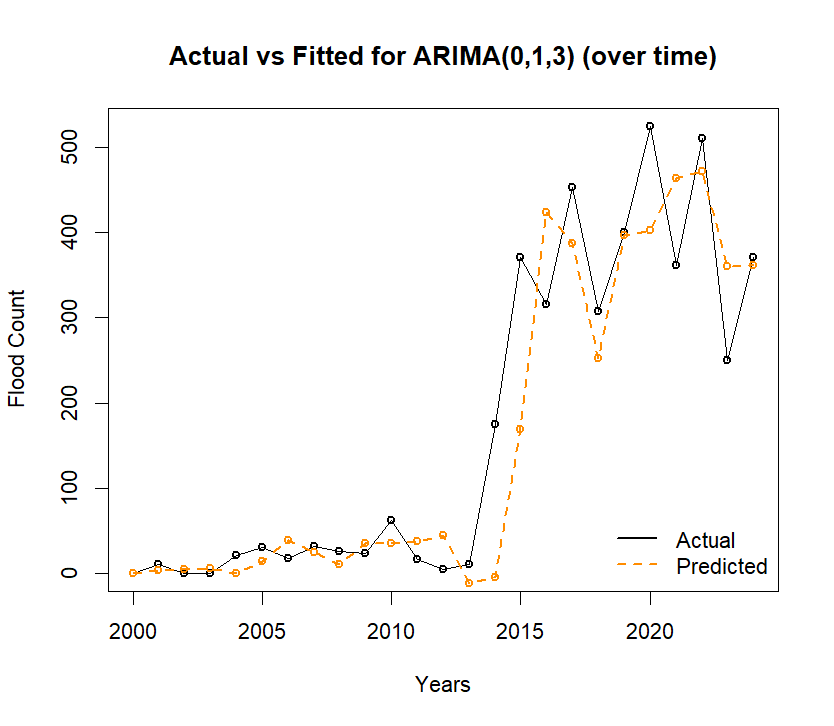
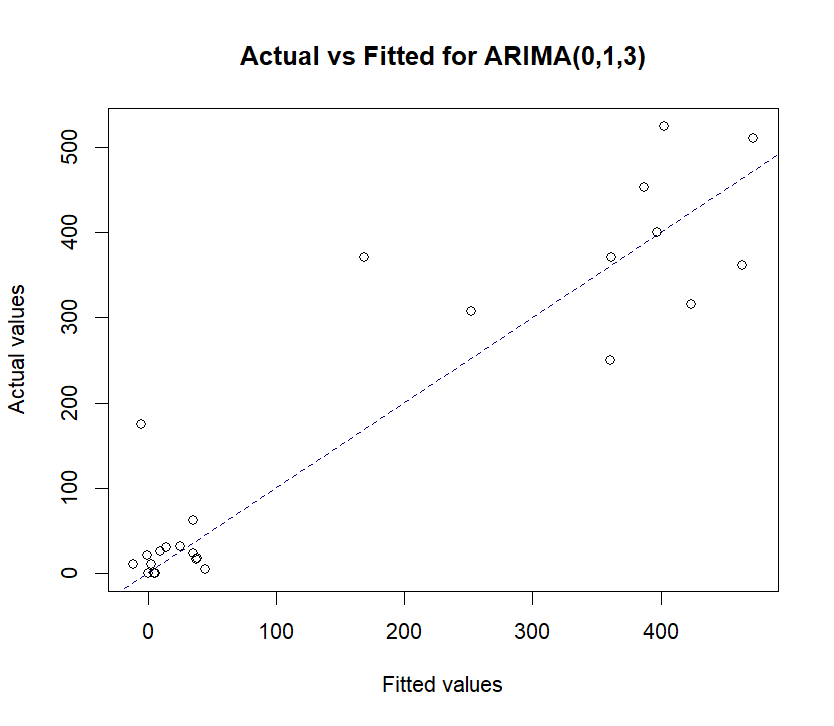
## Time Series Modelling

### ARIMA Model

Equation of the best univariate ARIMA Model (0,1,3):



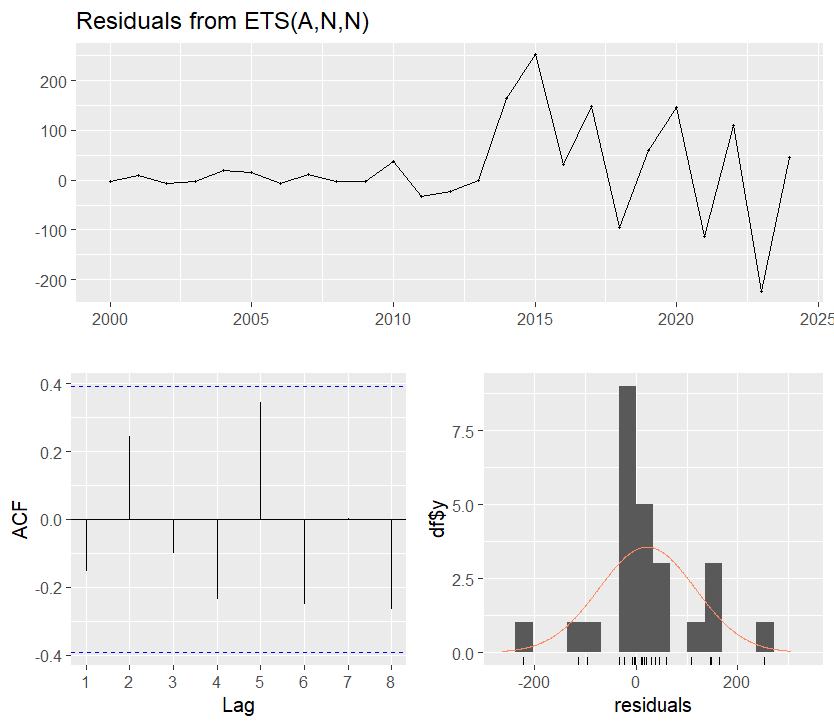
* Normality: Reasonably bell-shaped
* Autocorrelation: Ljung-Box p = 0.35 → no significant autocorrelation.



* Close fit to 45° line, though slight deviation at high flood counts.
* Captures rising trend post-2013 quite well, despite small underestimation in peaks

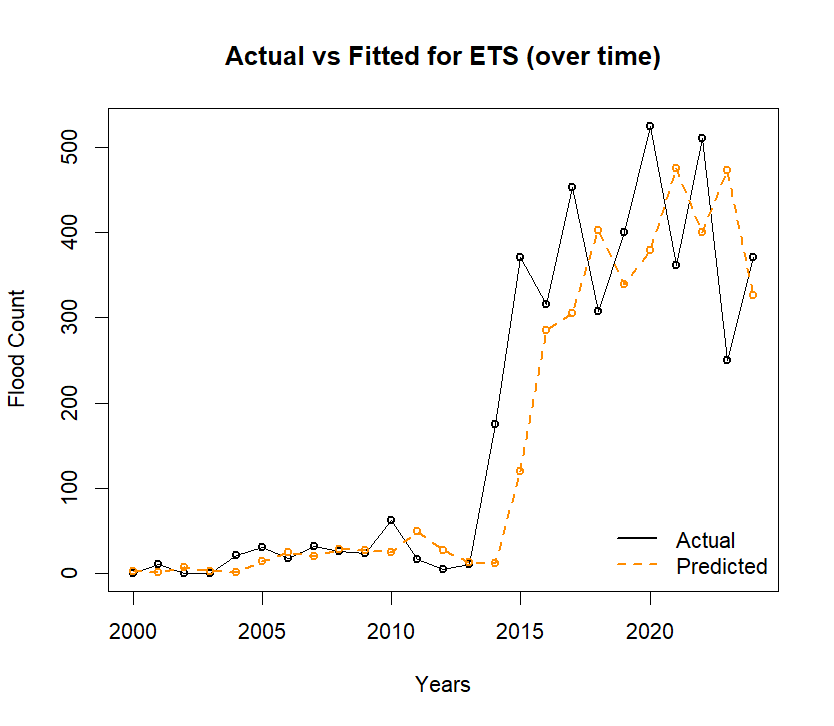
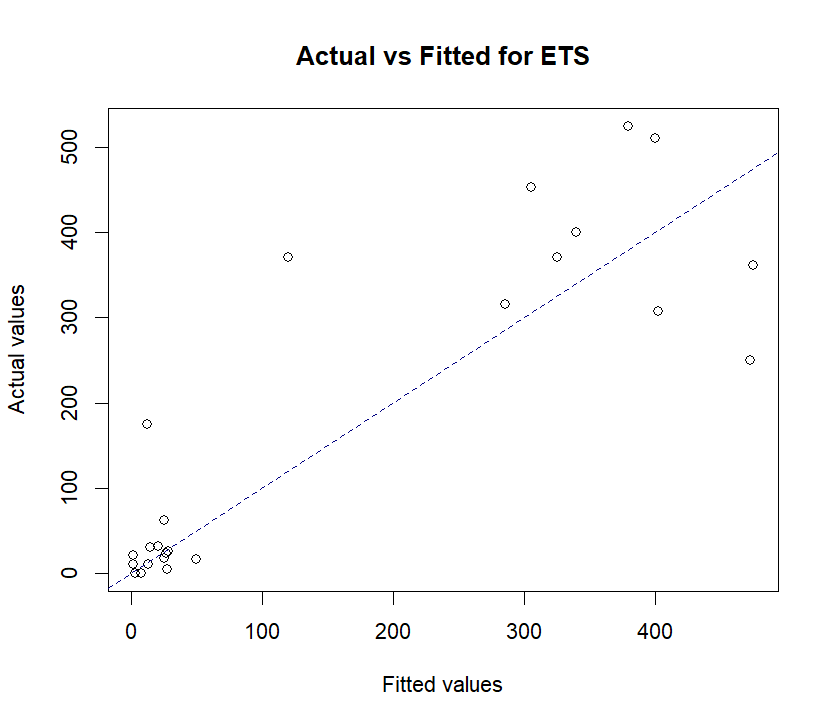
### ETS

The smoothing parameter alpha = 0.6599 (model gives more weight to recent observations which is expected for non-stationary processes)

A black background with white text

AI-generated content may be incorrect.

* Residuals: More erratic.
* Ljung-Box p = 0.13 → acceptable but weaker autocorrelation structure.

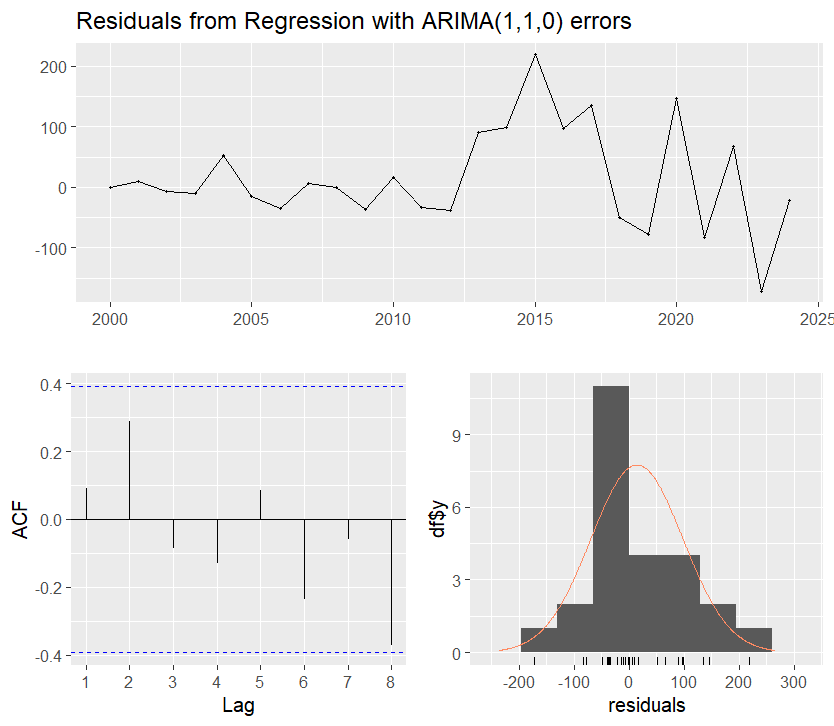
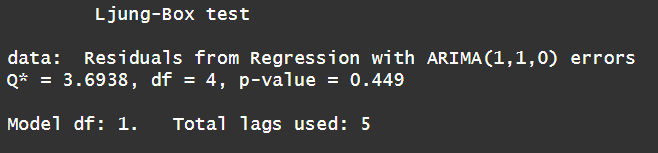


* Slightly more scatter around 45° line.
* Tracks broad trend but lags behind ARIMA in capturing year-to-year variability.

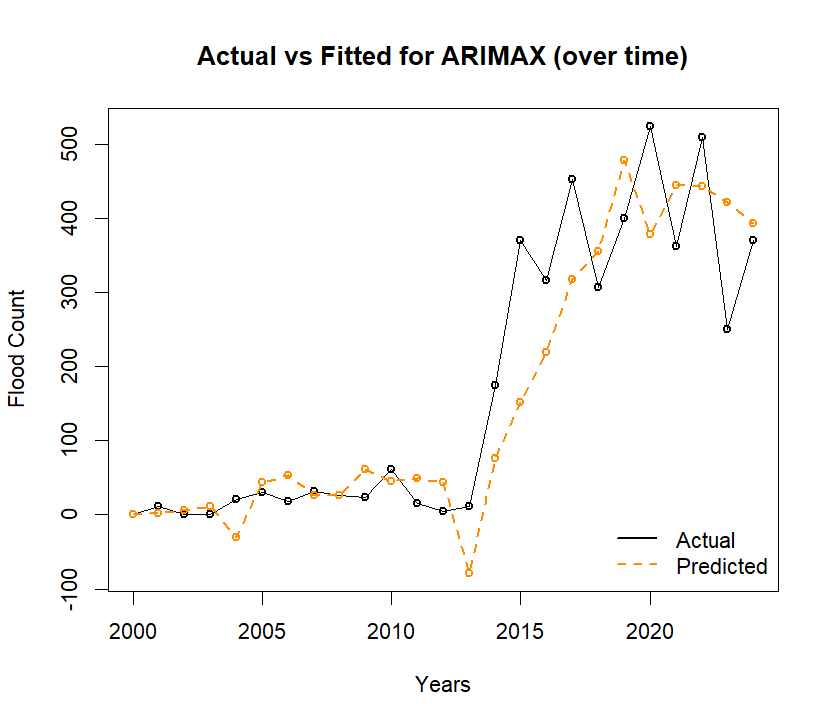
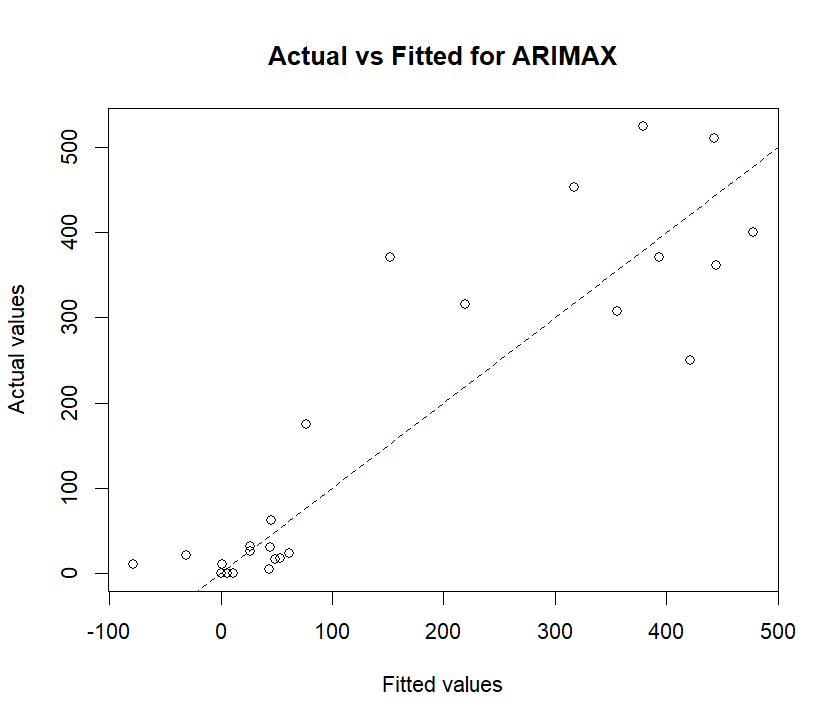
### ARIMAX Model

Equation of the best ARIMAX Model : ARIMA(1,1,0) with Urban road construction as the only regressor

* each unit increase in construction is associated with a ~3.8% rise in flood counts (holding AR structure constant)
* Other variables (GDP, electricity, drainage, land) were excluded as they did not improve model fit.

* no strong autocorrelation left
* Histogram of residuals looks reasonably symmetric.
* ACF plot shows lags within confidence bounds ⇒ good fit.
* Residuals: Ljung-Box p = 0.45 → no autocorrelation. Bell-shaped residuals, little skewness.



* Points lie mostly close to the 45° line. Slight underestimation at high flood counts (e.g., > 500), but still tightly clustered.
* The predicted values closely track the trend and turning points, especially after 2013. Model captures peaks, dips, and directional shifts — better than ETS and ARIMA-only models.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Model** | **AIC** | **RMSE** | **Normal Residuals** | **Interpretation** |
| **ARIMA(0,1,3)** | 286.75 | 73.8 | ✅ Yes | Best overall fit |
| ETS(A,N,N) | 314.34 | 95.3 | ⚠️ Noisy tail | Less robust |
| ARIMAX(1,1,0) with Urban Roads | 287.61 | 79.7 | ✅ Better shape | Interpretable, competitive |

**Final comparison (AIC & RMSE)**

|  |  |  |
| --- | --- | --- |
| **Model** | **AIC** | **RMSE** |
| NB GLM | 269.24 | 67.7 |
| **NB GAM** | **227.33** | **45.3** |
| Random Forest | [N/A](https://stats.stackexchange.com/questions/486566/akaike-information-criteria-applied-on-random-forest) | 71.5 |
| ARIMA(0,1,3) | 286.75 | 73.8 |
| ETS(A,N,N) | 314.34 | 95.3 |
| ARIMAX(1,1,0) with Urban Roads | 287.61 | 79.7 |

* GAM with Negative Binomial family is the best model in terms of AIC & RMSE.
* However, if we recall, we saw that none of the smooth terms were significant in the NB GAM and concluded that NB GLM should be sufficient. Yet, the GAM model has better performance. Why?
  + ***Non-significance does not mean non-utility*** : The smooth terms in the GAM model may not be individually significant (i.e., high *p*-values), but in combination, they help reduce residual deviance drastically and improve prediction accuracy. This happens often in smaller datasets where degrees of freedom are limited, and individual spline wiggles aren’t enough to hit statistical significance, but they still explain useful signal.
  + ***AIC rewards better fit, penalizes complexity — and still prefers GAM***

Despite the GAM having 6 effective degrees of freedom, its AIC (227) is still much lower than the simpler GLM (269), which means that - The improvement in log-likelihood (fit) more than compensates for the added complexity of the smoothing terms.

* + ***RMSE shows better generalization***

The GAM captures subtle nonlinearities or interactions that the GLM structure missed — even though it doesn’t have formally significant terms, its flexibility reduces prediction error, which is clear from the RMSE.