

# Modelling Flood Frequency and Severity in the Western Cape

ABSA / CZR Insurance Group Case Study

Group 8

01 February 2026

## 1 Introduction and Objective

This document outlines the methodology used to address the business problem posed by CZR Insurance Group: predicting the frequency and severity of flood events in the Western Cape to mitigate financial risk.

As per the case study requirements, this project utilizes provided climate and dam level data (2017–2021) to build predictive models and extrapolate findings to 2026.

## 2 Methodology and Workflow

### 2.1 1. Data Preparation (Script: 01\_data\_preparation.R)

**Data Sources:**

- `climate_data.csv`: Monthly climate indicators.
- `dam_data.csv`: Monthly water levels for Western Cape dams.

**Target Definition:** The datasets did not contain an explicit "Flood" label. Based on the prompt's context regarding rising waters, I defined two target variables:

- **Flood Severity:** Modeled using `Precipitation (mm)`. High rainfall correlates directly with claim severity in flood insurance.
- **Flood Risk Flag:** A binary variable created where `Precipitation > 90th percentile`.

**Feature Engineering:** Dam levels were aggregated to a provincial average to align with the granularity of the climate data. Lagged variables (1 and 2 months prior) were created to capture the buildup of environmental saturation essential for flood generation.

### 2.2 2. Exploratory Data Analysis (Script: 02\_eda.R)

**Visualisation:** Time series plotting revealed seasonality in rainfall, typically peaking in winter months (Western Cape's wet season). **Multicollinearity:** A correlation matrix was generated to check relationships. As expected, `Temp_Max`, `Temp_Min`, and `Temp_Avg` were highly correlated. To handle this (addressing Question 2 of the brief), only `Temp_Avg` was retained for linear models to ensure stability.

## 2.3 3. Model Fitting and Selection (Script: 03\_modeling\_and\_forecast.R)

Three models were trained on an 80/20 train-test split, respecting the time-series nature of the data (no random shuffling).

1. **Linear Regression:** Served as a baseline and offered high interpretability for variable impact.
2. **GLM (Gamma Distribution):** Chosen because rainfall data is right-skewed and strictly positive. This addresses the "statistical assumptions" requirement in the brief.
3. **Random Forest:** Used to capture non-linear interactions between humidity, wind, and pressure.

**Performance:** The models were evaluated using Root Mean Squared Error (RMSE). The GLM (Gamma) provided the best balance of accuracy and interpretability.

## 3 Key Findings and Interpretation

### Variable Impact:

- **Humidity:** Shows a strong positive correlation with flood severity.
- **Air Pressure:** Low pressure systems are significant predictors of storm events.
- **Dam Levels:** High pre-existing dam levels (Lagged variable) serve as a multiplier for risk—if dams are full, new rain has nowhere to go.

## 4 Forecast (2026)

Using an ARIMA time-series extrapolation, we projected flood severity trends into 2026. The forecast indicates cyclical peaks consistent with historical winter patterns. CZR should maintain higher liquidity reserves during the months of June–August.

## 5 Risk Mitigation Recommendations

*(Addressing the qualitative requirement)*

- **Dynamic Pricing:** Adjust premiums monthly based on the Avg\_Dam\_Level index. If dams are > 80% full entering winter, temporary surcharges should apply.
- **Parametric Insurance:** Automate payouts based on the Precipitation trigger (> 90<sup>th</sup> percentile) identified in this model, reducing claims processing costs.