DATA 6500 FINAL PROJECT REPORT

Analyzing Flood Distribution and Intensity in Canada: Integrating Geospatial and Hydrometric Data for Effective Disaster Preparedness

Somaye Ahangar, Poojan Umeshbhai Vadaliya

2024-07-16

Introduction

Canada's geography and climate make it particularly vulnerable to natural disasters, with floods being the most frequent and costly. Factors such as heavy rainfall, storm surges, and inadequate drainage systems contribute to these floods, posing significant threats to communities, property, and the environment across the nation. This study investigates flood distribution and intensity in Canada by analyzing the Historical Flood Events (HFE) dataset and Hydrometric data from the Water Survey of Canada, and National Hydronetwork Dataset.

Using spatial techniques like Gaussian Kernel Density Estimation (KDE), the research focuses on identifying flood patterns and contributing factors, particularly in the five most affected provinces. The study employs both non-parametric and parametric methods to estimate flood intensity. The non-parametric approach uses point pattern analysis and KDE to visualize flood hotspots, offering insights into spatial distribution patterns. The parametric approach explores the impact of hydrometric factors, specifically discharge values, on flood occurrences, with a detailed analysis of Quebec. By incorporating discharge values and spatial coordinates, the study provides a comprehensive understanding of flood intensity patterns and contributes valuable insights for disaster preparedness and risk mitigation in Canada. We then proceeded to predict peak discharge levels for the years 2023 to 2026, as the current hydrometric dataset only includes data up to 2022. Using INLA for forecasting, we estimated the potential timing of floods by analyzing the projected rise in peak discharge levels over the coming years.

Dataset

Flood Historical dataset[1]

Dataset for all the flood points have been obtained from the Historical Flood events(HFE) dataset where it contains the geospatial flood information including the flood location, provincial data, cause of flood and also time of the event. Floods affecting multiple locations are represented by multi-points which highlight the multiple locations of flood points affected from a single flood event.The raw dataset do contain the data from 1730s but for our study we have norrowed it down to 1970s to 2000s.

Hydrometric dataset [2]:

The Water Survey of Canada (WSC), in collaboration with provinces, territories, and other agencies through the National Hydrometric Program, collects, interprets, and disseminates standardized water resource data. This historical hydrometric data includes daily mean, monthly mean, annual maximum and minimum daily

mean, and instantaneous peak water level and discharge information from over 2700 active and 5080 discontinued monitoring stations across Canada. In this dataset there is a feature called "peak value" refers to Peak discharge term which represent the point where there is the largest amount of water in the river. The instantaneous peak water level is often used in flood forecasting, designing flood defenses, and managing water resources, so we plan to use this information from the hydrometric dataset as a potential factor in causing the flood.

Data for the peak value of the considered region of the Canada have been extracted as per the province where multiple files of the Peak Value information of the stations located over each province were merged into a single file to get the complete data for the considered duration of years(i.e.: 1970 to 2024)

NHN Waterbed Dataset [3]:

The National Hydro Network (NHN) dataset is a comprehensive representation of Canada's hydrographic features, including rivers, lakes, streams, and other water bodies. Developed and maintained by Natural Resources Canada (NRCan), the NHN dataset is designed to provide detailed, consistent, and accurate hydrographic information that supports various applications, including environmental monitoring, water resource management, flood risk assessment, and geographic information systems (GIS) analysis.

The NHN dataset is a crucial tool for understanding and managing Canada's water resources. Its extensive coverage, detailed attributes, and standardized format make it essential for hydrological studies, environmental assessments, and various GIS applications. In our project, we utilized the NHN geodatabase in ArcGIS to analyze and process relevant data. Initially, we extracted regions within Quebec based on flood locations, identifying 89 waterbody across the province. To refine our focus, we included only waterbeds with more than 20 flood locations, narrowing it down to the 23 most flood-prone waterbody in Quebec.

Waterbody : A waterbody here typically refers to the bottom or foundation of a waterbody, In datasets like the NHN, boundaries of waterbeds might refer to the extent or area of these foundational surfaces, providing a basis for understanding how water flows and interacts with the surrounding environment.

DATA ANALYSIS

YEARLY DISTRIBUTION OF THE FLOOD DATA

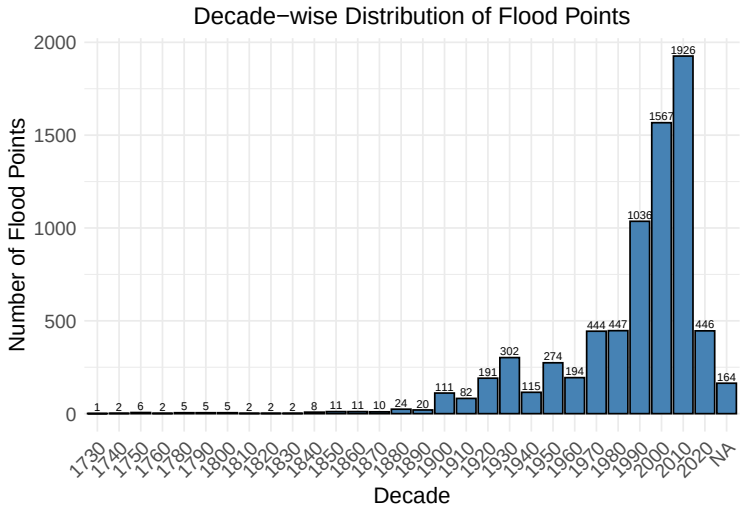


Figure 1: The plot shows the yearly distibution of the flood data

Till 1900s, the number of flood points per decade is quite low, usually below 20.A noticeable increase starts in the early 1900s, with a significant spike in flood points from the 1990s onwards. The highest number of flood points recorded in a decade is 1926 in the 2000s, followed by 1567 in the 2010s.There is a slight decline in the 2020s, with 446 flood points recorded.This increasing trend in flood points could be due to various factors, including climate change, urban development, and improved reporting mechanisms. Due to the inconsistency and the trend observed in the data for the flood points we have decided to considered to analyse for floods happened in 1970s and later as it is cant be the case that there are less than 10 floods occured before 1870s.

Province wise flood occurance in each decade

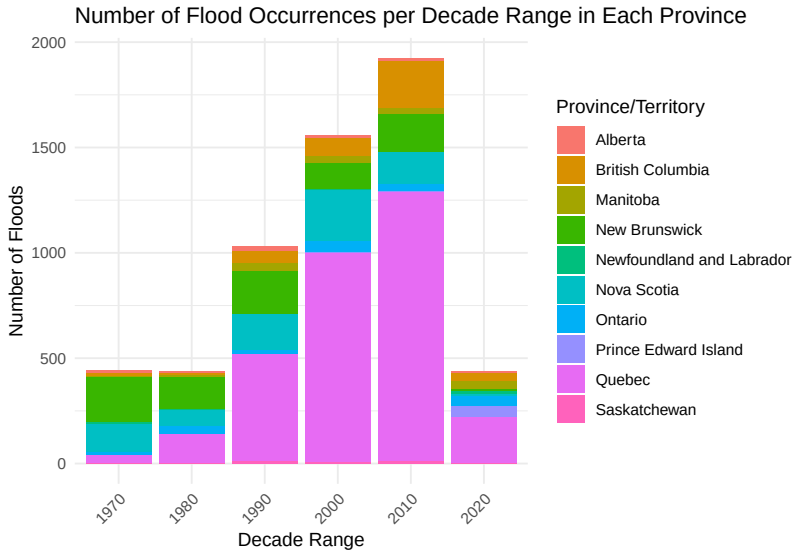


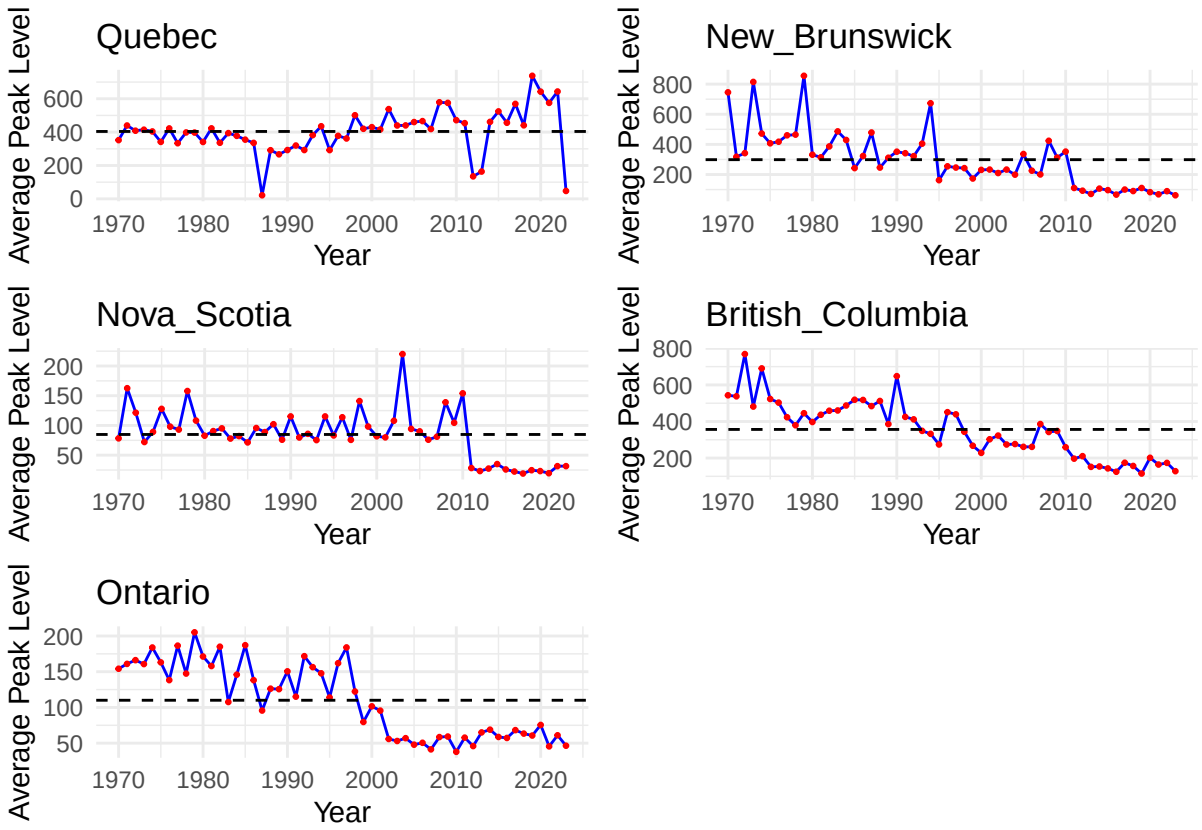
Figure 2: The plot describes province wise flood occurances in each decade

Observing the plot above following points are considered for the selection of the provinces for further data usage :

- **Quebec:** Consistently had the highest number of floods across all decades, with a significant spike in the 2000s and 2010s.
- **New Brunswick:** Showed substantial flood activity, especially in the 2000s and 2010s, indicating a high flood risk.
- **Nova Scotia and British Columbia:** Also exhibit notable flood occurrences, particularly in recent decades.
- **Ontario:** Displays moderate flood activity with a peak in the 2000s.

Considering the above all factors contributing to the number of flood activities especially after 2000s led us to narrowing the consideration of the province to just 5 province.

Annual Peak Level distribution for each province



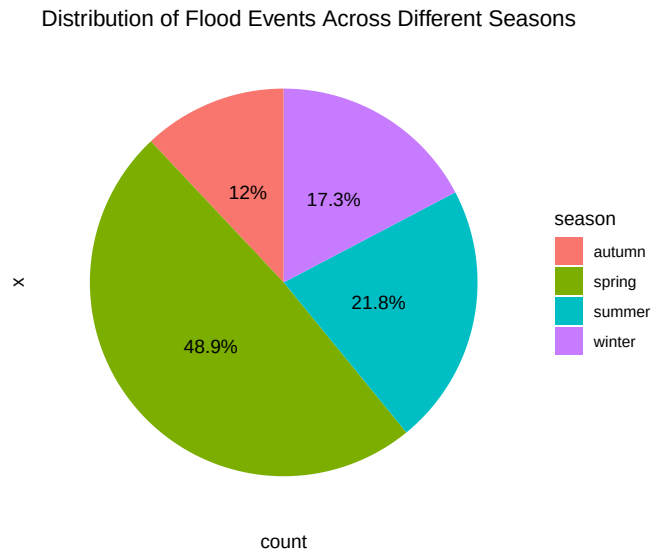
Quebec stands out as the most flood-prone province with increasing severity, highlighting an urgent need for improved flood mitigation strategies.British Columbia, New Brunswick, Nova Scotia, and Ontario showed effective flood management, evidenced by declining peak levels despite frequent floods.The comparison between frequency and severity suggests that provinces with high flood frequencies but declining peak levels have likely invested in effective flood management and mitigation measures.

Distribution of Floods across seasons

This pie chart visually represents the proportion of floods occurring in each season. This distribution highlights that floods are most prevalent in the spring and summer seasons, with a decrease during the winter and autumn.

And as observing the trend for the most occuring flood time period after we have found that there are mainly two time periods which are:

- **April to June (Spring Flooding):** This period is particularly prone to flooding due to the combination of snowmelt and spring rains. As temperatures rise in April, the snow begins to melt, contributing to increased water levels in rivers and streams. When this is combined with spring rainfall, the risk of flooding is heightened.
- **October to November (Fall Flooding due to Autumn Rains):** Heavy rains during the fall can saturate the ground and raise water levels in rivers, leading to floods. This period can also see flooding due to the fact that the ground is less able to absorb water as the soil becomes cooler and potentially more compacted.



Types of cause affecting the flood

The following pie chart visually represents the proportion of the distribution of flood causes. It shows freshet (a significant rise in water level) and heavy rain are the most significant reason for the flood to occur. Upon observing the freshet as the primary reason for affect for the flood which lead to decision of selecting the hydrometric data which contains the peak discharge of the waterbodies around Canada.

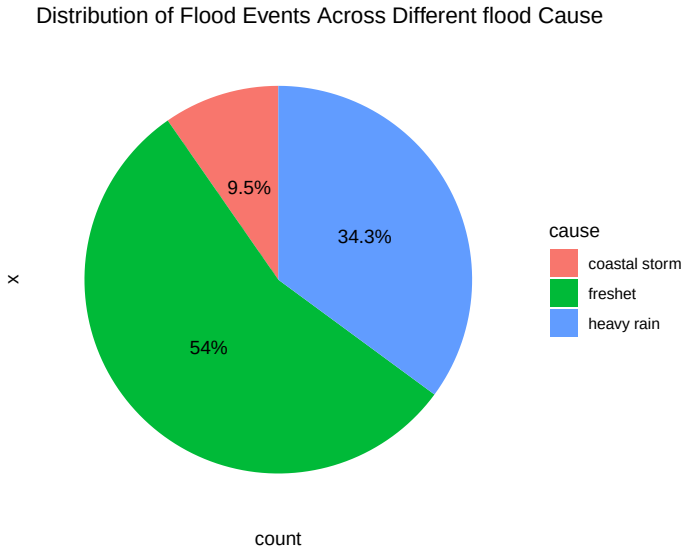
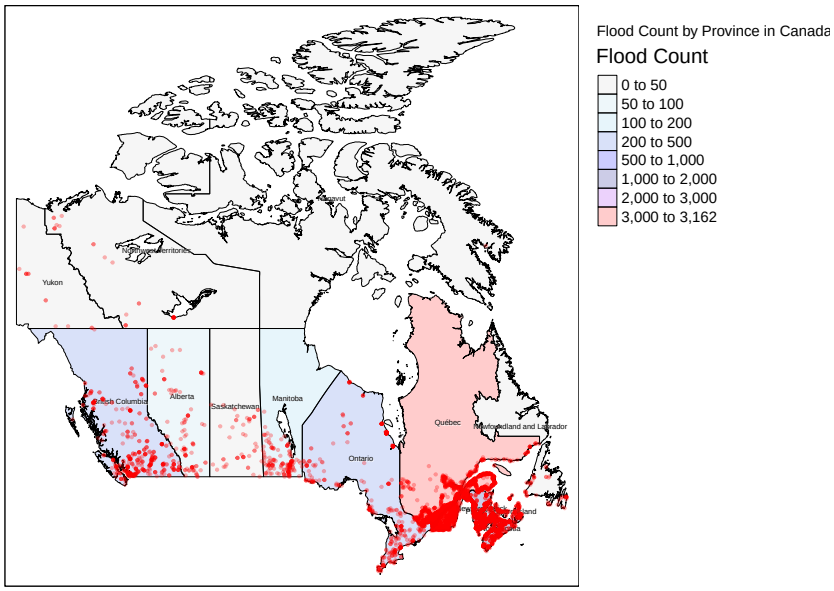


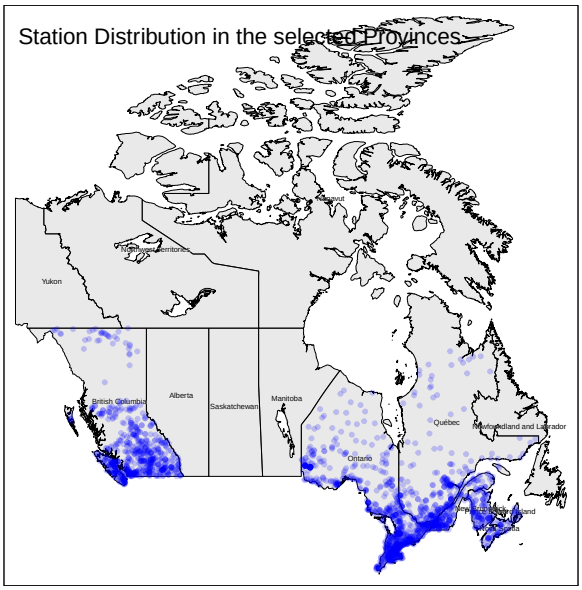
Figure 3: Reason for floods to happen.

Flood categorized by each province



Here in the above graph an overall picture of the flood count distribution can be seen for the whole Canada.As per the graph it can be observed that the mostly the flood observation noted are along the southern boundary of the Canada. And the region around southern Quebec, New Brunswick and Nova Scotia can be observed as being the hotspot of the overall region of Canada. As per the observation it can the presence of those region in along the coastal area is one of the primary reason of more floods to happen in that area.

Station distibution of peak value for overall canada for the selected 5 province



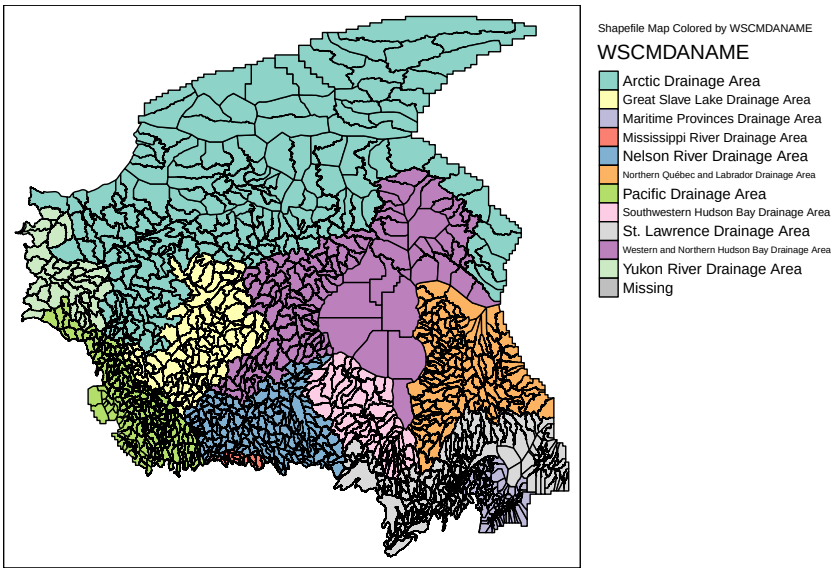
After considering the observation of the top 5 provinces with most number of floods the peak discharge value was been gathered from the hydrometric dataset for each province (i.e. : Quebec, Ontario, British Columbia,

New Brunswick and Nova Scotia). And the overall spatial distribution of the station of the hydrometric data to can be observed as the blue points on in the presented plot which shows that the hydrometric stations are spreaded across each province to record the peak discharge values.

Overview of National Hydro Network Dataset

Also considering the boundaries for especially hydrometric data analysis waterbeds are generally considered for which act as a container of multiple waterbodies as a single region. For this purpose we have considered partitioning the analysis of peak discharge values based on the waterbeds.

Spatial join was performed between the waterbeds polygons and the flood point coordinates which to find find out the flood points lying in each waterbed. Same was done for the hydrometric station which assisted us to fetch the flood points and hydrometric stations lying in the same waterbed making it relevant to perform a comprehensive analysis for a specific region.



From the Plot displaying the National HydroNetwork Dataset the wholde data is divided into 11 Drainage Area regions which are further divided into individual waterbeds. As there were no specific mentioning of the province so as per performing a boundary analysis it was found that two Drainage Area named **"St. Lawrence Drainage Area"** and **"Maritime Provinces Drainage Area"** drainage area felt under our primary analysis region of Quebec where the majority of the flood points were located which led us ton subset the complete national hydronetwork dataset to one subregion lying inside Quebec which would then look like the plot below where the red dots are the flood points around in the region of Quebec.

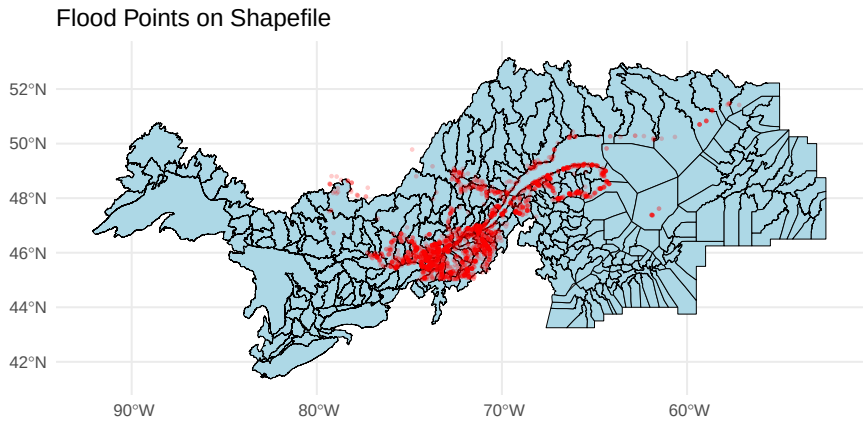


Figure 4: Visual representation of the map after selecting the drainage areas located in quebec.

Objective 1 : Using Non-Parametric Intensity Estimation

For doing the first step of our proposed methodology, we consider Point pattern analysis to examines the arrangement of points in space to understand the processes that generate these patterns.The analysis aims to describe and quantify the distribution of these points. We applied point pattern analysis to flood locations in Canada since 1970 use Gaussian Kernel Density Estimation (KDE) to create a continuous density surface, estimating the spatial density of flood points. This technique involves summing Gaussian (normal) distributions around each point to visualize and analyze the spatial distribution and clustering of points. We selected top five provinces in Canada based on Exploratory Data Analysis (EDA) results that we did. The main focus of a non-parametric intensity plot is to provide a visual representation of where floods are most densely concentrated within each province, helping to identify hotspots and understand spatial distribution patterns.

Here is the function that implemented for this purpose. The function analyzes and visualizes the none -parametric intensity plot flood distribution within a specified Canadian province. It filters flood data by province, extracts temporal features, and converts location data into spatial coordinates using a specified CRS and performs kernel density estimation, optionally using the `bw.CvL` method or default(Diggle) for bandwidth selection.The input use_cvl_bandwidth as TRUE including as `bw.CvL` bandwidth selection. Using the most appropriate CRS for each province ensures better spatial accuracy over the region. Also using different bandwidths for each province ensures that kernel density estimation accurately reflects the unique spatial distribution pattern and clustering of events in each region. The resulting density raster is re projected and visualized, overlaying flood locations.

```
# Define the function
# Input : province_name, input_crs, shape_file, use_cvl_bandwidth
process_flood_data = function(province_name, input_crs, shape_file,  use_cvl_bandwidth) {

  # Read the data from flood Historical Locations dataset
  flood_data = read.csv("finalfloodHistoricalLocations-.csv") |>
    filter(province_territory_description == province_name) |>
   mutate(
    # Handle different date formats
```

```
  start_date_clean = case_when(
    str_detect(start_date, "^\\d{1,2}/\\d{1,2}/\\d{4}$") ~ as.Date(start_date, format = "%m/%d/%Y"),
    str_detect(start_date, "^\\d{4}-\\d{2}-\\d{2}$") ~ as.Date(start_date, format = "%Y-%m-%d"),
    str_detect(start_date, "^\\d{4}$") ~ as.Date(paste0(start_date, "-01-01"), format = "%Y-%m-%d"),
    TRUE ~ NA_Date_
  ),
  year_l = year(start_date_clean),
  month_l = month(start_date_clean),
  decade = floor(year(start_date_clean) / 10) * 10
) |>
  filter(year_l>=1970)|>
  mutate(
    split_location = str_split(location, " "),
    x = sapply(split_location, `[`, 1),
    y = sapply(split_location, `[`, 2)
  ) |>
  dplyr::select(start_date, year_l, month_l, decade, locality, season, x, y)

# Convert to sf object
flood_data_sf = st_as_sf(flood_data, coords = c("x", "y"), crs = 4326)
flood_data_sf = st_transform(flood_data_sf, crs = input_crs)

# Read and transform the shapefile to province locally crs

province_shape = st_read(shape_file, quiet = TRUE)
province_shape_flat <- st_transform(province_shape, crs = input_crs)
province_owin = as.owin(province_shape_flat)

# Convert to ppp object
proj_flood = flood_data_sf |>
  st_geometry() |>
  st_transform(crs = input_crs)
flood_ppp = as.ppp(proj_flood)
Window(flood_ppp) = province_owin

# calculaying Density estimation

# bandwidth selection
if (use_cvl_bandwidth) {
  dens = density.ppp(flood_ppp, sigma = bw.CvL(flood_ppp)) |> rast()
} else {
  dens = density.ppp(flood_ppp) |> rast()
}
crs(dens) = paste0('EPSG:', input_crs)

# Convert CRS to a standard one
intensity_rast = project(dens, paste0("EPSG:", input_crs))
# Set plot title
plot_title = paste0("Non-Parametric Intensity Estimation", province_name, " Since 1970")
# Plotting
map = tm_shape(intensity_rast) +
  tm_raster(labels = c("very low", "low", "med-low", "medium", "med-high", "high", "very high"),
            palette = "-Spectral",
```

```
              title = plot_title) +
    tm_shape(flood_data_sf) +
    tm_dots() +
    tm_layout(legend.outside = TRUE,
              main.title.position = "center" ,
              main.title.size = 5) +
    tm_scale_bar()

  # Print the tmap object to show the map
  print(map)

  # Return intensity_rast and discharge_data (non_na_discharge_rows)
  return(list(intensity_rast = intensity_rast, flood_data_sf = flood_data_sf, province_owin= province_ou

}
```

Results of Objective 1 : The Non-Parametric Intensity Estimation for top 5 Provinces

Ontario

After defining the function , the function is applied to estimate and visualize the none-parametric intensity plot, for each province. The result for Ontario shows that except for some regions,flood locations spread over province but the hot spots are mostly occurred in southern Ontario which has more related to having Great Lakes, numerous rivers making it a region with abundant water resources.

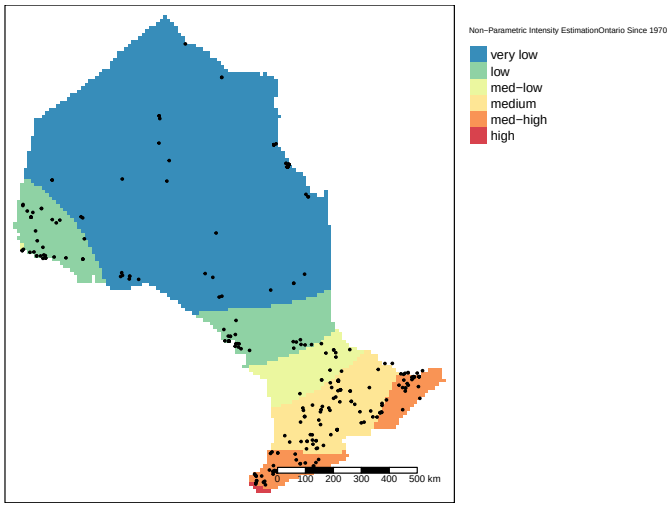


Figure 5: Non-Parametric Intensity Estimation of Ontario

British Columbia

The result for British Columbia shows that the Flood hotspots in the southern part of province happens which are largely due to the region's unique geography, with its combination of mountainous terrain, major river systems and low-Lying areas that are more prone to flooding because they are lower in elevation compared to the surrounding mountainous regions.

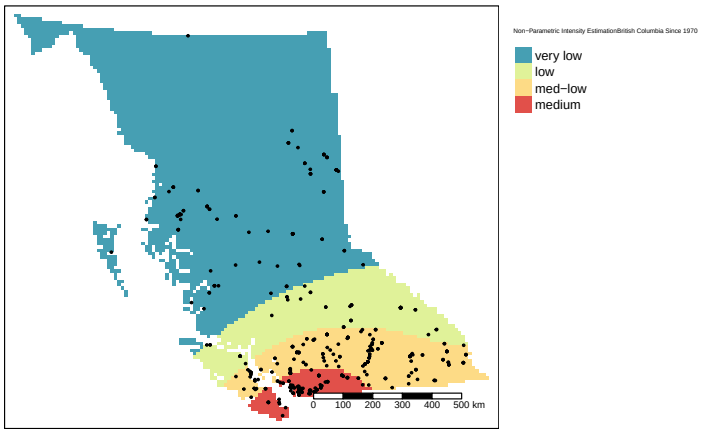


Figure 6: Non-Parametric Intensity Estimation of British Columbia

New Brunswick

The result for New Brunswick shows that the Flood hotspots mostly happens in the middle of the region, the area is significantly impacted by river flooding, particularly from the Saint John River and possibly other central river systems like the Miramichi.The central part of New Brunswick includes several low-lying areas as well . This could be due to a combination of natural geography and hydrology.

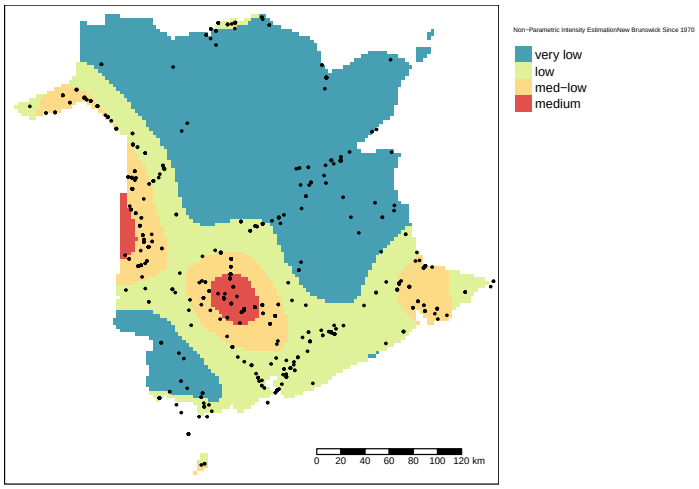


Figure 7: Non-Parametric Intensity Estimation of New Brunswick

Nova Scotia

The result for Nova Scotia shows that the Flood hotspots mostly happens in the middle of the region. The central part of the province, is a low-lying region that is naturally prone to flooding. The central lowlands, particularly the Annapolis Valley and areas along the Shubenacadie River, are likely the main contributors to these hotspots.

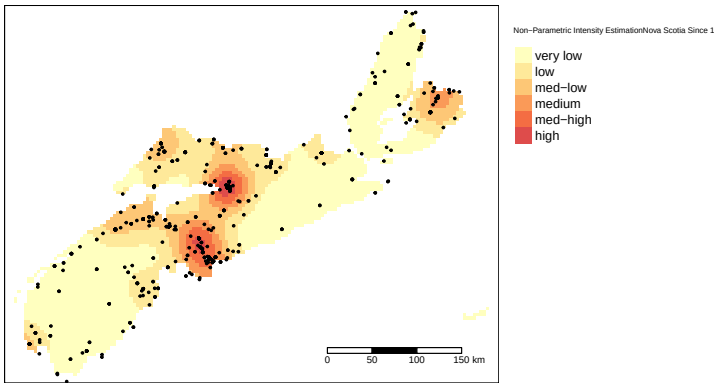


Figure 8: Non-Parametric Intensity Estimation of Nova Scotia

Quebec

The result for Quebec shows that the southern part of the province includes hotspot which have major bodies of water. The Saint Lawrence River. Additionally, rivers like the Ottawa River, which flows into the Saint Lawrence River, can also contribute to flooding in southern Quebec.

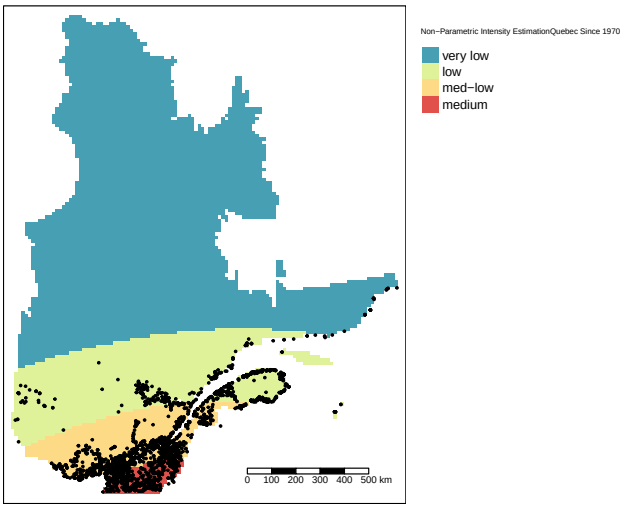
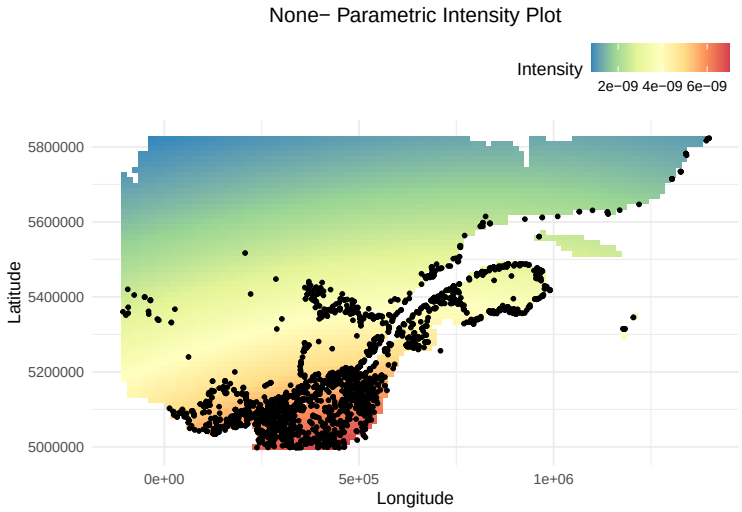


Figure 9: Non-Parametric Intensity Estimation of Quebec

Objective 2: Using Parametric Intensity Estimation

Flooding can result from heavy rainfall, rapid snow melt, ice jams, and changes in land use. Hydro metric data is crucial in this context as it provides real-time measurements of water levels and flow rates in rivers and lakes. So for exploring the effect of hydro metric factors on flood occurrences, in this section we did the second step of our proposed methodology that aimed to consider historical hydro metric data, including discharge value ,to perform Parametric Intensity Estimation. As the EDA shows , Quebec is the first provinces among the top 5 .So we keep our further analysis for this province. For this, first let's have a look at a close-up view of none- Parametric Intensity plot in areas likely to flood.This plot can help us to make better comparison to parametric intensity plot in next steps as they are both in a similar mapping scale.



Parametric Intensity Plot Using Discharge Value and Coordinate for Quebec

For exploring the effect of discharge values in flood location, we load the monthly discharge values dataset and according to our proposed study, we more specifically focus on discharge value for Quebec. We have data over years for mostly each month. So we need to consider the distinct stations for Quebec Province and based on the distinct station that extracted, next step would be add discharge value for each flood location. To do this, we first find the nearest station for each flood location and combine flood locations with the nearest stations and then join the relevant monthly discharge data based on the month and year that flood happen. The output dataset would be flood locations that happened over Quebec combined with the relevant discharge value, considering non na data for discharge values. This result would be used for parametric intensity plot that includes both the flood location and corresponding discharge value. For intensity estimation, as each flood point has a discharge value associated with it, the discharge value would be considered a mark, so it would be treated as charactritics of each point. Also at the next step, including the x-coordinte as predictors to see how the intensity would vary with the discharge level water and spatial coordinate.

In both senarios, the intensity estimation involves fitting a log-linear model (point process model) where the intensity function is predicted based on interpolated discharge level water values across whole region. The model uses these interpolated values to estimate the spatial distribution of flood intensities. By interpolating the discharge data, we create a continuous surface of discharge value values that allows for a more detailed and accurate modeling of flood intensity patterns. Here the akima::interp function is used to perform interpolation, generating a grid of discharge value based on the coordinates and known discharge values.

```r
#Merging the flood Location and Hydrometric Dataset for Parametric Intensity Estimation
# Reading the dataset and filtering Qc
monthly_hydro_Data = read.csv("monthly_peak_values.csv") |>
  filter(PROV_TERR_STATE_LOC=="QC")|>
  mutate(date_h = as.Date(paste0(DATE, "-01"), format = "%Y-%m-%d")) |>
  mutate(year_h = year(date_h)) |>
  mutate(month_h = month(date_h)) |>
  mutate(decade_h = floor(year_h / 10) * 10)|>
         dplyr::select(x,y, year_h, month_h, decade_h, STATION_NAME, PROV_TERR_STATE_LOC, STATION_NUMBE

#Distinct stations - Monthly_peak_values

monthly_hydro_Data_dist = read.csv("monthly_peak_values.csv") |>
  filter(PROV_TERR_STATE_LOC=="QC")|>
  dplyr::select(STATION_NAME, STATION_NUMBER, x, y) |>
  distinct(STATION_NUMBER, .keep_all = TRUE) #

# Add peak Value for each flood location information (finding nearest station)

# Set Crs and Define sf objects
floodData_Qc_sf = st_as_sf(flood_data_sf, coords = c("x", "y"), crs = 4326) # Assuming WGS84
monthly_hydro_Data_dist_sf = st_as_sf(monthly_hydro_Data_dist, coords = c("x", "y"), crs = 4326)

# Define sf objects

floodData_Qc_sf = st_transform(flood_data_sf, crs = 2950)
monthly_hydro_Data_dist_sf = st_transform(monthly_hydro_Data_dist_sf, crs = 2950)

# Ensure both datasets are in the same CRS
floodData_Qc_sf = st_transform(floodData_Qc_sf, crs = st_crs(monthly_hydro_Data_dist_sf))

# Find the index of the nearest station for each flood location
nearest_station_indices = st_nearest_feature(floodData_Qc_sf, monthly_hydro_Data_dist_sf)

# Extract the nearest station info for each flood location
nearest_stations = monthly_hydro_Data_dist_sf[nearest_station_indices, ]

# Combine flood locations with their nearest stations
merged_sf = bind_cols(floodData_Qc_sf, st_drop_geometry(nearest_stations)) |>
  dplyr::select(year_l, month_l, locality, STATION_NAME, STATION_NUMBER,)

# Convert to sf object if needed
merged_sf = st_as_sf(merged_sf)

# Filter and join the relevant monthly discharge data
flood_with_discharge = merged_sf |>
  left_join(monthly_hydro_Data, by = c("year_l" = "year_h", "month_l" = "month_h", "STATION_NUMBER" = "
  dplyr::select(year_l, month_l, locality, STATION_NAME.x, STATION_NUMBER, MONTHLY_MEAN_DISCHARGE, geom

non_na_discharge_rows = flood_with_discharge |>
  filter(!is.na(MONTHLY_MEAN_DISCHARGE))

# Convert the sf object to a data frame
```

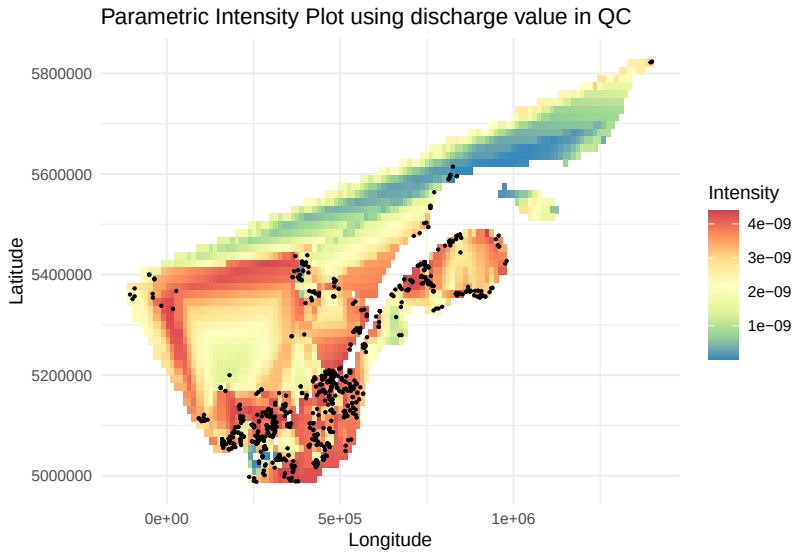
```r
non_na_discharge_df = st_drop_geometry(non_na_discharge_rows)

# Extract coordinates and marks
coords = st_coordinates(non_na_discharge_rows)
marks = non_na_discharge_df$MONTHLY_MEAN_DISCHARGE

# Ensure there are no NA values
valid_indices = !is.na(marks)
coords = coords[valid_indices, ]
marks = marks[valid_indices]

# Create a data frame from the coordinates and marks
df = data.frame(x = coords[,1], y = coords[,2], marks = marks)

# Aggregate marks by coordinates using base R
df_aggregated = aggregate(marks ~ x + y, data = df, FUN = mean)

# Extract the cleaned coordinates and marks
coords_aggregated = as.matrix(df_aggregated[, c("x", "y")])
marks_aggregated = df_aggregated$marks

# Create a ppp object for the point pattern
discharge_ppp = ppp(x = df$x, y = df$y, window = province_owin)

# Interpolate the data using akima::interp
interpolated_marks = interp(x = coords_aggregated[,1], y = coords_aggregated[,2], z = marks_aggregated,
                            xo = seq(from = min(coords_aggregated[,1]), to = max(coords_aggregated[,1]
                            yo = seq(from = min(coords_aggregated[,2]), to = max(coords_aggregated[,2]

# Convert the interpolated grid to an 'im' object
raster_im = as.im(interpolated_marks, W = province_owin)

# Fit the parametric model (log-linear model) using the covariate
parametric_model = ppm(discharge_ppp ~ raster_im)

#Plot Parametric intensity, including discharge value for QC
intensity_rast = raster::raster(predict(parametric_model))

# Convert raster to data frame
intensity_df = as.data.frame(rasterToPoints(intensity_rast))
names(intensity_df) = c("x", "y", "intensity")
```

Results of Objective 2: Parametric Intensity Estimation

Parametric Intensity Estimation Using Discharge Value in Quebec

As result shows, this plot incorporates a parametric model that uses the discharge level water value in Quebec as a variable to create the intensity plot. It indicates that how the discharge value correlates with spatial density. The areas with warmer colors (red and orange) indicate higher intensities, suggesting that the discharge value value has a significant impact in these regions.It also confirms that the most flood location are in the higher intensity which shows that the discharge value, has significant in causing the flood in most regions. It suggests regions where the discharge value is more significant, indicating a stronger influence of

discharge water on the density of data points.




```r
# Convert the sf object to a data frame
non_na_discharge_df = st_drop_geometry(non_na_discharge_rows)

# Extract coordinates and marks
coords = st_coordinates(non_na_discharge_rows)
marks = non_na_discharge_df$MONTHLY_MEAN_DISCHARGE

# Ensure there are no NA values
valid_indices <- !is.na(marks)
coords = coords[valid_indices, ]
marks = marks[valid_indices]

# Create a data frame from the coordinates and marks
df = data.frame(x = coords[,1], y = coords[,2], marks = marks)

# Aggregate marks by coordinates using base R
df_aggregated = aggregate(marks ~ x + y, data = df, FUN = mean)

# Extract the cleaned coordinates and marks
coords_aggregated = as.matrix(df_aggregated[, c("x", "y")])
marks_aggregated = df_aggregated$marks

# Interpolate the data using akima::interp
interpolated_marks = interp(x = coords_aggregated[,1], y = coords_aggregated[,2], z = marks_aggregated,
                            xo = seq(from = min(coords_aggregated[,1]), to = max(coords_aggregated[,1]
                            yo = seq(from = min(coords_aggregated[,2]), to = max(coords_aggregated[,2]

# Convert the interpolated grid to an 'im' object
raster_im_x = as.im(interpolated_marks, W = province_owin)

# Create a ppp object for the point pattern
discharge_ppp = ppp(x = df$x, y = df$y, window = province_owin)
```

```
# Fit the parametric model (log-linear model) using the raster image and the x coordinate
parametric_model = ppm(discharge_ppp ~ raster_im_x + x)

# Get predicted intensity values
predicted_intensity = predict(parametric_model)

# Convert the predicted intensity to a data frame for ggplot
predicted_df = as.data.frame(as.im(predicted_intensity))
```

Parametric Intensity Estimation Using discharge value and X Coordinate in Quebec

As result shows, this plot incorporates a parametric model that uses the discharge level water value and also and X Coordinate in Quebec as a variable. It indicates how the combination of these two variables influences the spatial distribution. It makes a more refined and potentially more accurate representation of intensity distribution. This means that certain areas might show different intensities not just because of the discharge water values but also due to their location in the X direction. For instance, in the this plot, an area that has a moderate discharge water value might still show high intensity if it is in a specific X coordinate region that the model finds significant.

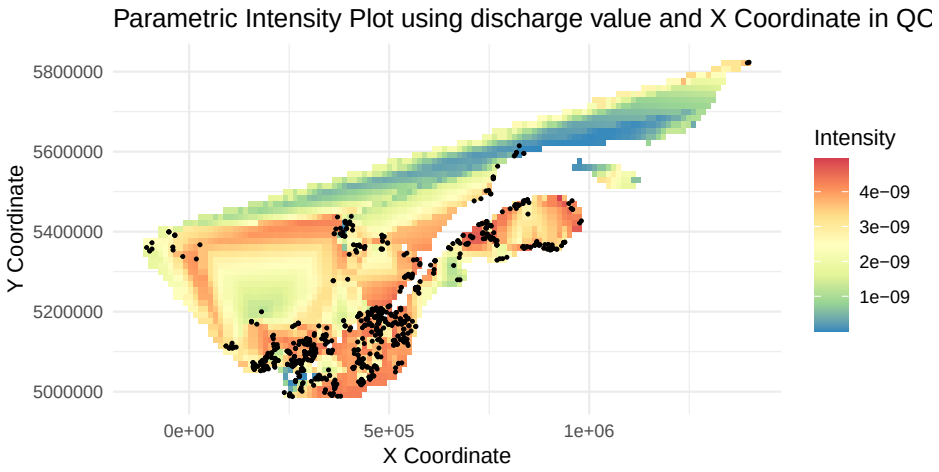


Figure 10: Parametric Intensity Plot using discharge value and X Coordinate in QC

Objective 3 : Forecasting hydrometric data and Predicting Flood Possibilities

When we initially proposed using Kriging for prediction, we discovered that while Kriging is a powerful tool for many types of spatial prediction, it may have limitations in predicting flood points based on hydrometric data. The complexity of the factors that cause floods suggests that a deeper understanding of the physical processes driving these events might be more suitable for accurate predictions. Due to the spatial autocorrelation of flood data points and the intricacies of hydrometric datasets, identifying potential flood locations with Kriging can be challenging, making it less ideal for this task.

Therefore, we opted to use the Integrated Nested Laplace Approximation (INLA) for forecasting flood events in water bodies. This method offers a more robust approach, particularly when considering the spatial and temporal complexities involved in flood prediction.

Flood Distribution Across Each Waterbody

Based on the overview of the NHN dataset, two drainage areas located in the province of Quebec were selected for further analysis. For the analysis of peak distribution across individual waterbodies, only those with more than 20 flood points were considered.

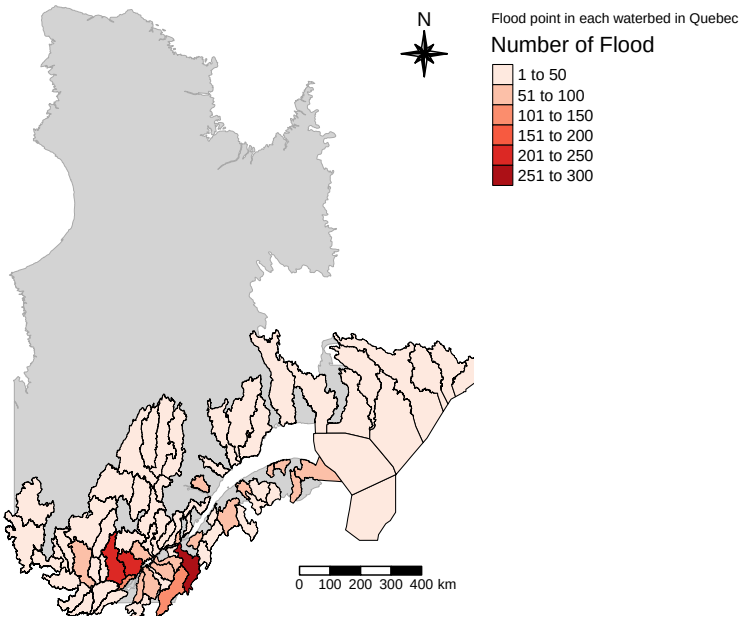


Figure 11: Flood point count in each waterbed in Quebec

Here is the over all distribution of the number of flood point intensity for in each waterbody indicating that the waterbody polygon with lighter color having less number of flood point which gradually increases as the color gets darker in the plot of flood distribution.

Preparation of the data

For the preparation of the data the peak values and the flood points were first spatially joined with the waterbody dataset to get an idea of the which flood point and the station lies within the same waterbody.

Calculating of the threshold for the peak discharge value for each waterbed for flood to happen

After preprocessing the peak value dataset the hydrometric dataset was joined with the flood points where which had the same waterbody , year and month of the flood to happen. Based on the joined data we filtered the peak discharge values only for at the time when flood had happened for the waterbodies.

```
# Remove the sf class and convert to a regular data frame
flood_points_df <- st_drop_geometry(flood_points_within_shapes)
peak_values_df <- st_drop_geometry(peak_values_within_shapes)


# Perform an inner join on WSCMDANAME, year, and month
joined_data <- peak_values_df |>
  right_join(flood_points_df, by = c("WSCSSDANAM" = "WSCSSDANAM", "year_p"="year_f", "month_p"="month_f
             relationship = "many-to-many")


cat("\n Peak values with no discharge and mean level:",sum(is.na(joined_data$MONTHLY_MEAN_LEVEL) & is.na
```

```
##
##  Peak values with no discharge and mean level: 453
```

```
#Only keeping the rows which have either of monthly mean level or monthly mean discharge value or both.
joined_data <- joined_data %>%
  filter(!is.na(MONTHLY_MEAN_LEVEL) | !is.na(MONTHLY_MEAN_DISCHARGE))

# Considering the maximum value from the MONTHLY_MEAN_LEVEL and MONTHLY_MEAN_DISCHARGE and keeping a si
peak_data <- joined_data %>%
             mutate(peak_discharge = pmax(MONTHLY_MEAN_LEVEL, MONTHLY_MEAN_DISCHARGE, na.rm = TRUE))

#Filtering out unwanted columns
peak_data <- peak_data |> dplyr::select(WSCSSDANAM,year_p,month_p,STATION_NUMBER,peak_discharge,WSCMDANA

# Sorting the peak discharge values based on the year month and Waterbed
peak_data_sorted <- peak_data |> arrange(WSCSSDANAM,year_p,month_p)

# If the dataset has duplicate stations and you want only distinct ones
station_points_sf <- peak_values_within %>%
  distinct(STATION_NUMBER, .keep_all = TRUE)
```

After retrieving the peak discharge values for each waterbody the discharge data was aggregated by individual waterbody and the upper quantile which would be above 80% was considered as the peak threshold value which was the scaled using Min max scaling as it also preserves the linear relationship between the data to match with the scale of forecasted data.

```
# Min-Max Scaling
min_max_scale <- function(x) {
  return ((x - min(x, na.rm = TRUE)) / (max(x, na.rm = TRUE) - min(x, na.rm = TRUE)))
}

# Getting the quantile data for each waterbed to find the threshold of the waterbed
quantile_data <- peak_data %>%
  group_by(WSCSSDANAM) %>%
```

```
  summarise(
    lower_quartile = quantile(peak_discharge, 0.25, na.rm = TRUE),
    upper_quartile = quantile(peak_discharge, 0.80, na.rm = TRUE),
    mean_discharge = mean(peak_discharge, na.rm = TRUE)
  )

#Scaling the upper quartile for the threshold consideration
quantile_data$scaled_threshold <- min_max_scale(quantile_data$upper_quartile)
```

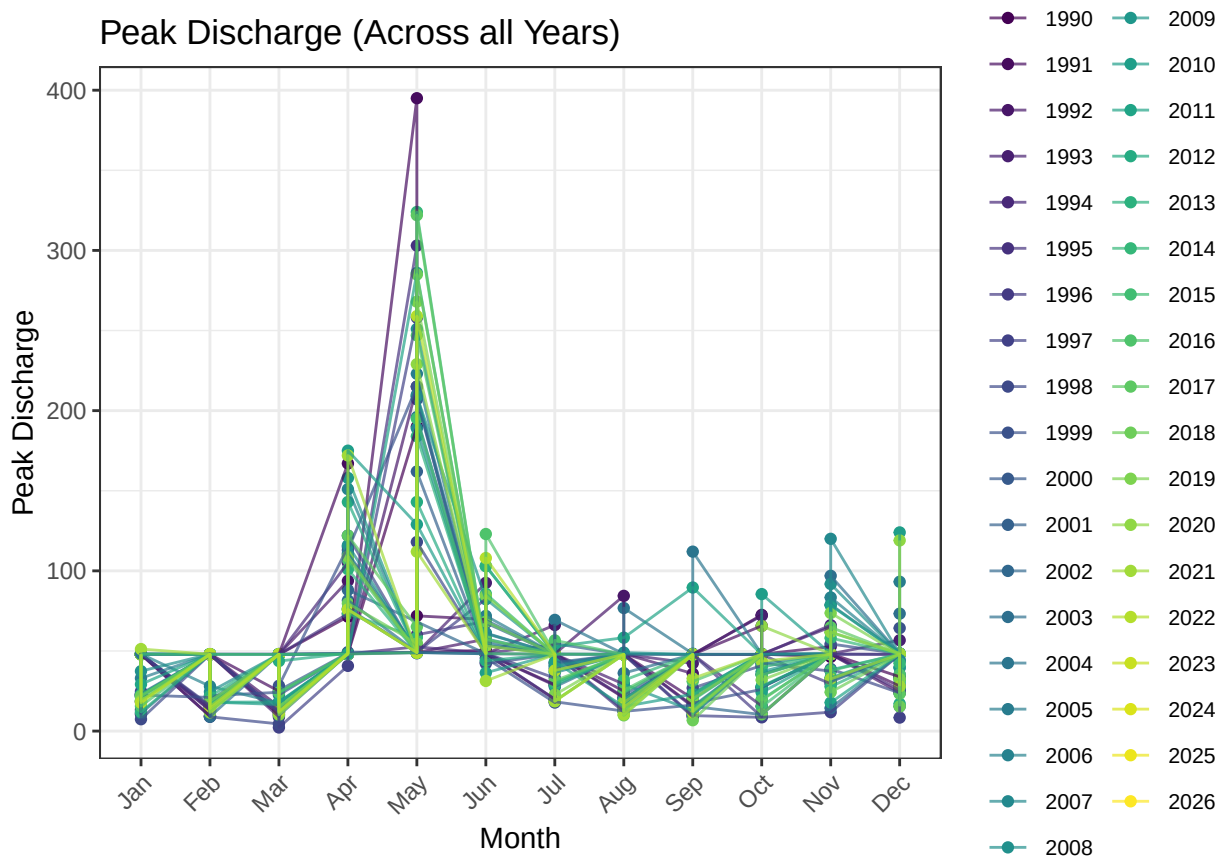
Forecasting the hydrometric data for Matapédia waterbody

Considering the availability and the completeness of the data "Matepedia" was choosen as the waterbody for forecasting as it had all values for each year and each months peak discharge value present which made it an ideal candidate to choose for a better reasonable analysis and forecasting of the following years peak discharge.

Also for forecasting the empty dataframe accornding to the requirement and the description of the discharge data was prepared containing the time_id representing the number of year , year, month , and the peak discharge where the peak discharege values were was cosidered to be predicted based on the trend of the peak discharge for ther past 32 years.The empty dataframe was then combined with the populated dataframe containing the peak discharge values.

```
### Peak value forecasting for Matepedia

matepedia = filtered_peak_values_forecast_df %>%
  filter(WSCSSDANAM == "Matapédia") |> dplyr::select(WSCSSDANAM,year_p,month_p,peak_discharge)

matepedia$time_id = matepedia$year_p - 1989
month = data.frame(id = 1:12)

#Empty dataframe created here to store the forecast data.
forecasting_data = data.frame(time_id = c(rep(34,12),rep(35,12),rep(36,12),rep(37,12)),
                              year_p = c(rep(2023,12),rep(2024,12),rep(2025,12),rep(2026,12)),
                              month_p = rep(month$id,4),
                              peak_discharge = rep(NA, 12*4))

#TO check if all the months and year are present
year_check_s <- matepedia %>%
  group_by(WSCSSDANAM,year_p) %>%
  summarise(months = n_distinct(month_p))

#Joining the empty dataframe with the populated
matepedia_missing = matepedia |>
  dplyr::select(time_id,year_p,month_p,peak_discharge) |>
  rbind(forecasting_data)
```

Considering the Matepedia waterbody an trend of each year was observed for the data that was present and it was observed that averagely for majority of the years there was a rise in the peak discharge aroudn the month of may which signified towards the assumption of sprin flooding which could occur due to the melting of ice after the winter along with rainfall.



Considering the past data we considered the second timeid as month which would help to find the seasonal trend month wise for the data to forecast and for that ar1 model of inla along with seasonal model was considered to determine the trend of the peak discharge for the forecasting years. The results of the forecasting model were then scaled using min_max scaling to match the threshold scale that was determined before.

FILTERING ALL THE POSSIBLE TIMES FOR THE FLOOD TO HAPPEN

The forecasted peak values for each waterbed were compared to the scaled threshold, which was calculated to determine the likelihood of flooding. The moments when the discharge values exceeded the peak threshold were identified as potential flood occurrences in the region of Matapedia. In the forecast plot, the horizontal red line indicates the flood threshold for that specific waterbody, while the vertical green line marks the point in time after which the data was forecasted

As getting a closer look of the peak dischrage trend for the forecasted years , all the forecasted years showed the same trend which was a sharp rise at the time around May and slight rise at the end of the year which could be the possible time periods for the flood to happen.

Analysis of the limitations of the forecasted methodology :

Although the peak discharge has been forecasted in this analysis, it cannot be regarded as entirely accurate. Given the observed data, it is unlikely that the forecast will follow the same trend in the upcoming years as it has in the past. Additionally, due to the lack of sufficient data, there is no precise understanding of when or if a sudden increase in peak levels might occur. Accurately predicting such events would require continuous observation of the peak discharge trends in any given region, which would be computationally intensive.

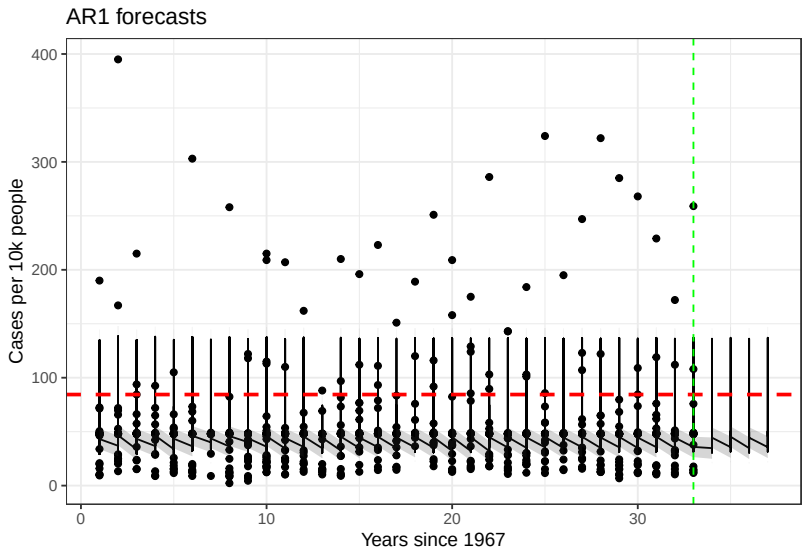


Figure 12: Forecasted data

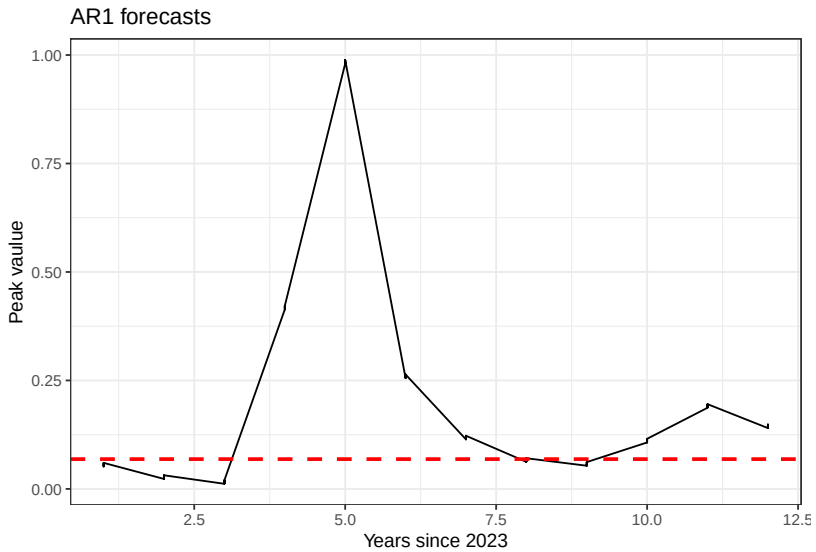


Figure 13: Forecasted data since 2023.

Conclusion and Future Works

Floods predominantly occur in British Columbia, Ontario, Quebec, New Brunswick, and Nova Scotia, with Quebec experiencing the highest frequency of flood events across Canada. These floods are primarily concentrated near coastal areas and around the Great Lakes. One of the significant factors contributing to these floods, as revealed by our exploratory data analysis (EDA), is the freshet—a marked rise in water levels due to snowmelt or heavy rainfall. To better understand this phenomenon, we included discharge values from flood monitoring stations in our analysis to assess their impact on flood occurrence.

According to our study, on-parametric intensity estimation provides a basic visualization of the spatial distribution of flood locations, identifying hotspots and cold spots across each province. However, parametric intensity estimation goes further by incorporating models that show how specific variables, such as discharge value and spatial coordinates, influence flood intensity. The parametric plots offer more detailed insights, especially when examining the spatial distribution of intensity. The comparison between parametric plots demonstrates the importance of including multiple variables, like the X coordinate, to achieve a more refined and accurate representation of intensity, revealing how spatial location and discharge values interact to shape flood intensity patterns.

For our forecasting efforts, we face several limitations. One significant challenge is the unavailability of data for selected years and months at each station, which hampers our ability to make comprehensive predictions. Additionally, as we integrate peak levels from all stations across various water bodies, it becomes clear that we cannot perform forecasting for all 23 water bodies we initially aimed to cover. Furthermore, there is a noticeable skew in peak levels across different stations, complicating the flood forecasting process. To address these challenges, we require more accurate data that maps the relationship between the NHN dataset and the monitoring stations for each waterbody. This improved mapping is crucial for enhancing the reliability of our forecasts. Our analysis indicates that different water bodies have varying discharge thresholds for triggering floods. This variation suggests that other factors, such as slope, elevation, soil moisture, and temperature—factors influenced by spatial coordinates—also play a crucial role in flood occurrence. Additionally, the capacity of a water body to manage large volumes of water is an important determinant in the likelihood of flooding.

Flood intensity estimation and forecasting can be further enhanced by considering a broader range of factors beyond hydrometric data. Key elements include meteorological data (precipitation, temperature), soil moisture and type, land use changes (urbanization, deforestation), topography, and river network characteristics. Additionally, groundwater levels, climate change projections, human activities, ice jams, and catchment area features provide valuable insights for more accurate flood predictions. These factors could be considered in future work to improve flood intensity estimation and forecasting.

Reference

[1]GeoBase. Historical flood events (HFE). https://app.geo.ca/result/en/historical-flood-events-(hfe)?id=fe83a604-aa5a-4e46-903c-685f8b0cc33c&lang=en

[2]Environment and Climate Change Canada. Water quantity data. https://climate-change.canada.ca/climate-data/#/water-quantity-data

[3]Natural Resources Canada. National hydro network (NHN). Open Canada. https://open.canada.ca/data/en/dataset/a4b190fe-e090-4e6d-881e-b87956c07977