Beyond Hydrology: Shaping Sustainable Future with Intuitive Water Risk Metrics

Data Visualization Final Project

AUTHOR
Arianna Tessari - Student ID: 19322

Outline

- Introduction
- Data and Exploratory Analysis: Data Cleaning and Preprocessing, Descriptive Statistics
- Global Overview: Overall Water Risk
- Global Overview: Overall Water Risk Across Economic Regions
- Sector-wise Comparison of Risk Indicators
- Risk Categories: Their Role in Determining Water Risk
- Understanding the Cost of a Sustainable Water Future
  - Data and Exploratory Analysis: Data Cleaning and Preprocessing, Descriptive Statistics
  - Global Overview: Total Expected Costs in 2030
  - Projected 2030 Average Costs Across Economic Regions and Income Levels
  - Correlation between GDP per Capita, Average Risk Score, and Projected Total Costs
- Conclusions
- References

Introduction

As it promotes both environmental health and economic progress, water is a vital incentive for the advancement of human societies. Important industries like manufacturing, energy production, and agriculture depend on it. Still, there is a significant barrier because many policymakers find it difficult to understand the complexities of hydrological data. Aim of this project is precisely to explore and understand, with the support of visualization techniques, the risks and implications associated with access to water across different regions of the world.

This project relies on two main datasets, both retrieved from the WORLD RESOURCES INSTITUTE database.

The first one, "Aqueduct 4.0", was specifically designed to translate complex hydrological data into intuitive indicators of water-related risks. It represents a long-term, chronic trend from 1979 to 2019 and its strength lies in its ability to provide intuitive understanding and actionable information for informed decision making.

On the other hand, the second dataset, utilized mainly in the latter part of the analysis, computes the costs associated with the measures required to close the gap between current conditions and desired conditions for sustainable water management. The focus there shifts to future estimates for the year 2030.

Data and Exploratory Analysis

Data Cleaning and Preprocessing

```
data <- read_csv("data/Aqueduct40_baseline_annual_y2023m07d05.csv",
                 show_col_types = FALSE)
```

Here there is an overview on how the dataset is initially presented:

```
head(data)
```

```
# A tibble: 6 × 231
  string_id     aq30_id pfaf_id gid_1  aqid gid_0 name_0 name_1 area_km2 bws_raw
  <chr>           <dbl>   <dbl> <chr> <dbl> <chr> <chr>  <chr>     <dbl>   <dbl>
1 111011-EGY.1…       0  111011 EGY.…  3365 EGY   Egypt  Al Qa…    4.22     9999
2 111011-EGY.1…       1  111011 EGY.…  3365 EGY   Egypt  As Su… 1846.       9999
3 111011-EGY.1…       2  111011 EGY.… -9999 EGY   Egypt  As Su…   30.5      9999
4 111011-None-…       3  111011 -9999  3365 <NA>  <NA>   <NA>      0.743    9999
5 111011-None-…       4  111011 -9999 -9999 <NA>  <NA>   <NA>     13.4      9999
6 111012-EGY.1…       5  111012 EGY.…  3365 EGY   Egypt  Al Qa…  258.          1
# ℹ 221 more variables: bws_score <dbl>, bws_cat <dbl>, bws_label <chr>,
#   bwd_raw <dbl>, bwd_score <dbl>, bwd_cat <dbl>, bwd_label <chr>,
#   iav_raw <dbl>, iav_score <dbl>, iav_cat <dbl>, iav_label <chr>,
#   sev_raw <dbl>, sev_score <dbl>, sev_cat <dbl>, sev_label <chr>,
#   gtd_raw <dbl>, gtd_score <dbl>, gtd_cat <dbl>, gtd_label <chr>,
#   rfr_raw <dbl>, rfr_score <dbl>, rfr_cat <dbl>, rfr_label <chr>,
#   cfr_raw <dbl>, cfr_score <dbl>, cfr_cat <dbl>, cfr_label <chr>, …
```

The first steps involved understanding and cleaning the data and preparing it for subsequent visualizations. Four risk indicators were selected, including quantity, quality and reputational concerns. They all group together different indicators within the same category.

Here follows a summary of the indicators chosen:

- **Physical Risk Quantity**

  Physical Risk Quantity assesses the risk associated with insufficient or excessive water by consolidating various indicators within the Physical Risk Quantity category. Elevated values signify greater risks related to water quantity.

- **Physical Risk Quality**

  Physical Risk Quality assesses the risk linked to water unsuitable for use by integrating specific indicators within the Physical Risk Quality category. Elevated values indicate increased risks related to water quality.

- **Regulatory and Reputational Risk**

  Regulatory and Reputational Risk assesses the uncertainty associated with regulatory changes and potential conflicts with the public regarding water issues. Elevated values signify greater regulatory and reputational risks in the context of water management.

- **Overall Water Risk**

  The combined water risk scores from the three categorized groups can be employed to derive the overall water risk score. The summation of weights is utilized to calculate the relative contribution of each group.

  Overall Water Risk assesses all water-related risks by consolidating selected indicators from the Physical Risk Quantity, Physical Risk Quality, and Regulatory and Reputational Risk categories. Increased values indicate a higher level of water risk.

Numerically they are translated into two quantitative variables: raw score and score. The score is simply the raw value translated to a scale of 0 to 5, with higher scores indicating higher water risk.

The analysis covers nine key sectors across 173 countries. To make visualization easier, a merge was made with the “countries” dataset from the rnaturalearth R package. This not only provided geometries, but also categorized countries into seven economic regions and five income levels.

To achieve the primary goal of a tidy dataset, the following pipeline was implemented. This data transformation process involves several key steps to ensure that the dataset is cleaned, structured, and well prepared for subsequent analysis and visualization.

1. **Data Selection:** Initially, specific columns are selected from the original dataset, including identifiers (`gid_0`, `name_0`, `name_1`) and a range of numeric columns containing the variables of interest.

2. **Missing Values:** Missing data is handled by replacing specific placeholders with `NA` values. Numeric columns with a value of `-9999` and character columns containing `"No data"` are transformed accordingly.

```
GroupedWaterRisk <- data %>%

  select(gid_0,
         name_0,
         name_1,
         (62:ncol(.))) %>%

  mutate(across(where(is.numeric), ~ifelse(. == -9999, NA, .))) %>%
  mutate(across(where(is.character), ~ifelse(. == "No data", NA, .)))
```

3. **Missing Value Removal:** Rows containing any missing values are removed from the dataset using the `drop_na` function, ensuring data integrity.
4. **Column Renaming and Elimination:** Columns are renamed to improve clarity (`gid_0` becomes `ISO`, `name_0` becomes `country`), and columns not pertinent to the analysis, such as those containing information about categories or weight fractions, are eliminated.
5. **Grouping and Summarization:** The data is grouped by geographical identifiers (`ISO` and `country`), and summary statistics, specifically means, are calculated for numeric columns within each group.
6. **Reshaping Data:** The structure of the data is transformed from wide to long format using the `pivot_longer` function, facilitating further analysis.
7. **Column Splitting:** The `name` column is split into multiple columns (`w`, `awr`, `sector`, `group`, `type`) based on underscores, enhancing the granularity of the dataset.
8. **Reshaping Data (again):** Following the split, the data is reshaped from long to wide format based on the `type` column, potentially improving readability and analysis flexibility.
9. **Labeling:** A new column named `label` is introduced, categorizing the `score` column into distinct bins representing different levels of risk.
10. **Column Selection and Transformation:** Finally, unnecessary columns (`w` and `awr`) are removed, and categorical values in the `sector` and `group` columns are transformed into more descriptive labels using `case_when` statements.

```
GroupedWaterRisk <- GroupedWaterRisk %>%

  drop_na() %>%

  rename(ISO = gid_0, country = name_0) %>%

  # eliminate columns not used
  select(-contains(c("cat", "weight_fraction"))) %>%

  group_by(ISO, country) %>%

  summarise(across(where(is.numeric), mean)) %>%

  pivot_longer(cols = contains(c("raw","score"))) %>%

  separate(name, into = c("w", "awr", "sector", "group", "type"), sep = "_") %>%
  # w = weighted
  # awr = aggregated water risk

  pivot_wider(names_from = type, values_from = value) %>%

  mutate(label = cut(score, breaks = c(0, 1, 2, 3, 4, 5),
                     labels = c("Low", "Low-medium", "Medium-high", "High", "Extremely high"))) %>%

  select(-c("w", "awr")) %>%

  mutate(sector = case_when(
    sector == "def" ~ "Default",
    sector == "agr" ~ "Agriculture",
    sector == "che" ~ "Chemicals",
    sector == "con" ~ "Construction Materials",
```

```
              sector == "elp" ~ "Electric Power",
              sector == "fnb" ~ "Food & Beverage",
              sector == "min" ~ "Mining",
              sector == "ong" ~ "Oil & Gas",
              sector == "smc" ~ "Semiconductor",
              sector == "tex" ~ "Textile")) %>%

            mutate(group = case_when(
              group == "qan" ~ "Physical risk quantity",
              group == "qal" ~ "Physical risk quality",
              group == "rrr" ~ "Regulatory and reputational risk",
              group == "tot" ~ "Total, Overall water risk"))

          head(GroupedWaterRisk)
```

```
# A tibble: 6 × 7
# Groups:   ISO [1]
  ISO   country     sector      group                          raw score label
  <chr> <chr>       <chr>       <chr>                        <dbl> <dbl> <fct>
1 AFG   Afghanistan Default     Physical risk quantity        2.51  3.59 High
2 AFG   Afghanistan Default     Physical risk quality         4.10  4.23 Extr…
3 AFG   Afghanistan Default     Regulatory and reputational r… 4.65  4.68 Extr…
4 AFG   Afghanistan Default     Total, Overall water risk     3.21  4.12 Extr…
5 AFG   Afghanistan Agriculture Physical risk quantity        2.62  3.82 High
6 AFG   Afghanistan Agriculture Physical risk quality         2.85  2.87 Medi…
```

11. **Geometrical Data Retrieval**: Geometrical data for countries is obtained using the `rnaturalearth::ne_countries` function, specifying a large scale and returning the result as an sf object. Specific columns are selected from the retrieved geometrical data, including country identifiers, economy, income group, and geometry, with the ISO code column renamed to "ISO" for clarity and for matching the main dataset. Additionally, the income group column is renamed to "income" for improved readability.

```
          # obtain the geometries
          countries <- rnaturalearth::ne_countries(
            scale = "large",
            returnclass = "sf") %>%
            select(geounit, gu_a3, economy, income_grp, geometry) %>%
            rename(ISO = gu_a3, income = income_grp)
```

12. **Geometry Addition**: The geometrical data retrieved for countries is merged with the `GroupedWaterRisk` dataset using the ISO code as the common identifier. At the end, it is also possible to have a look at the tidy dataset that has been obtained, now comprehensive of geometrical data.

```
          # add the geometry
          GroupedWaterRisk <- merge(countries[,-1], GroupedWaterRisk, by="ISO")

          head(GroupedWaterRisk)
```

```
Simple feature collection with 6 features and 9 fields
Geometry type: MULTIPOLYGON
Dimension:     XY
Bounding box:  xmin: 60.48678 ymin: 29.38661 xmax: 74.89231 ymax: 38.47367
Geodetic CRS:  WGS 84
  ISO                   economy        income     country
1 AFG 7. Least developed region 5. Low income Afghanistan
2 AFG 7. Least developed region 5. Low income Afghanistan
3 AFG 7. Least developed region 5. Low income Afghanistan
4 AFG 7. Least developed region 5. Low income Afghanistan
5 AFG 7. Least developed region 5. Low income Afghanistan
6 AFG 7. Least developed region 5. Low income Afghanistan
                  sector                    group      raw    score
1              Oil & Gas     Physical risk quality 5.000000 5.000000
2          Electric Power     Physical risk quality 2.849489 2.874574
3         Food & Beverage    Physical risk quantity 2.576878 3.621153
4         Food & Beverage     Physical risk quality 3.655931 3.807325
5 Construction Materials Total, Overall water risk 3.074014 4.050848
```

```
6         Electric Power Total, Overall water risk 2.353315 3.212093
           label                       geometry
1 Extremely high MULTIPOLYGON (((74.54235 37...
2    Medium-high MULTIPOLYGON (((74.54235 37...
3           High MULTIPOLYGON (((74.54235 37...
4           High MULTIPOLYGON (((74.54235 37...
5 Extremely high MULTIPOLYGON (((74.54235 37...
6           High MULTIPOLYGON (((74.54235 37...
```

As a final step, a new dataset is created, containing the countries that are present in the geometrical data but missing in the GroupedWaterRisk dataset, identified by the ISO codes. This will be useful later on for visualization purposes.

```
# create a dataset with the countries missing in GroupedWaterRisk
missing_data <- countries[!(countries$ISO %in% GroupedWaterRisk$ISO), ]
```

Descriptive Statistics

In this section, a brief overview of descriptive statistics derived from the cleaned dataset is provided. Descriptive statistics can offer valuable insights into the central tendencies, distributions, and variability of the data, enabling a comprehensive understanding of the dataset's characteristics.

The numerical results show that water risk scores vary significantly between countries and economic sectors. The average score is about 2.68, with a variance of 1.47, indicating a discrete variability in the data.

The boxplots clearly show the distribution of scores across economies, highlighting differences and potential outliers. It can be seen that the level of risk tends to be distributed over lower values for more developed regions. The quasirandom plot on the right visualizes the risk scores in a non-overlapping manner, making it easier to identify the density of the data points. Here it is observable that for some regions there is a high density of data, probably caused by the higher number of countries belonging to a certain region.

```
# number of countries
n_distinct(GroupedWaterRisk$ISO)
```

```
[1] 173
```

```
GroupedWaterRisk %>%
  st_drop_geometry() %>%
  select_if(is.numeric) %>%
  summary()
```

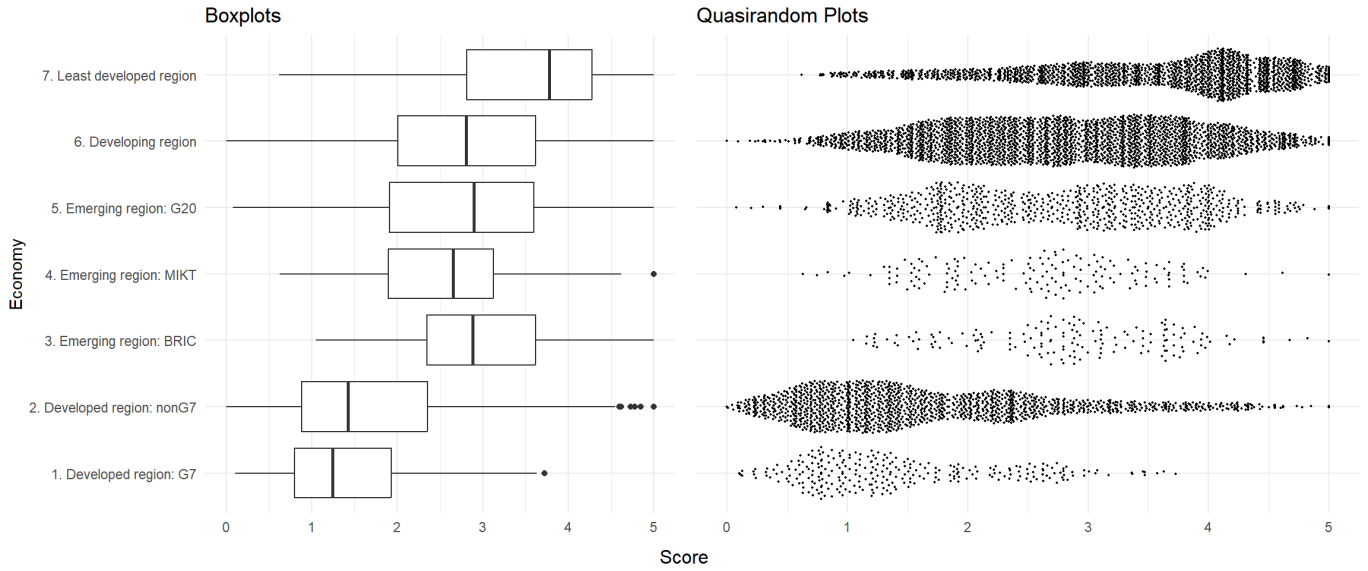
```
      raw            score      
 Min.   :0.000   Min.   :0.000  
 1st Qu.:1.280   1st Qu.:1.701  
 Median :2.039   Median :2.701  
 Mean   :2.210   Mean   :2.680  
 3rd Qu.:2.976   3rd Qu.:3.698  
 Max.   :5.000   Max.   :5.000  
```

```
GroupedWaterRisk %>%
  st_drop_geometry() %>%
  select_if(is.numeric) %>%
  cbind() %>%
  apply(MARGIN = 2, var) %>%      # MARGIN = 2 - function applied on the columns
  round(digits = 3)
```

```
  raw score 
1.457 1.470 
```

```
plot1 <- ggplot(data = GroupedWaterRisk, mapping = aes(x = economy, y = score)) +
  geom_boxplot() +
  coord_flip() +
  labs(title="Boxplots", x="Economy") +
  theme_minimal() +
  theme(axis.title.x = element_blank())
```

```
plot2 <- ggplot(data = GroupedWaterRisk, mapping = aes(x = economy, y = score)) +
  geom_quasirandom(size=0.4) +
  coord_flip() +
  labs(title="Quasirandom Plots") +
  theme_minimal() +
  theme(axis.title.y = element_blank(),
        axis.text.y = element_blank(),
        axis.ticks.y = element_blank(),
        axis.title.x = element_blank())

grid.arrange(plot1, plot2, ncol = 2, bottom="Score")
```



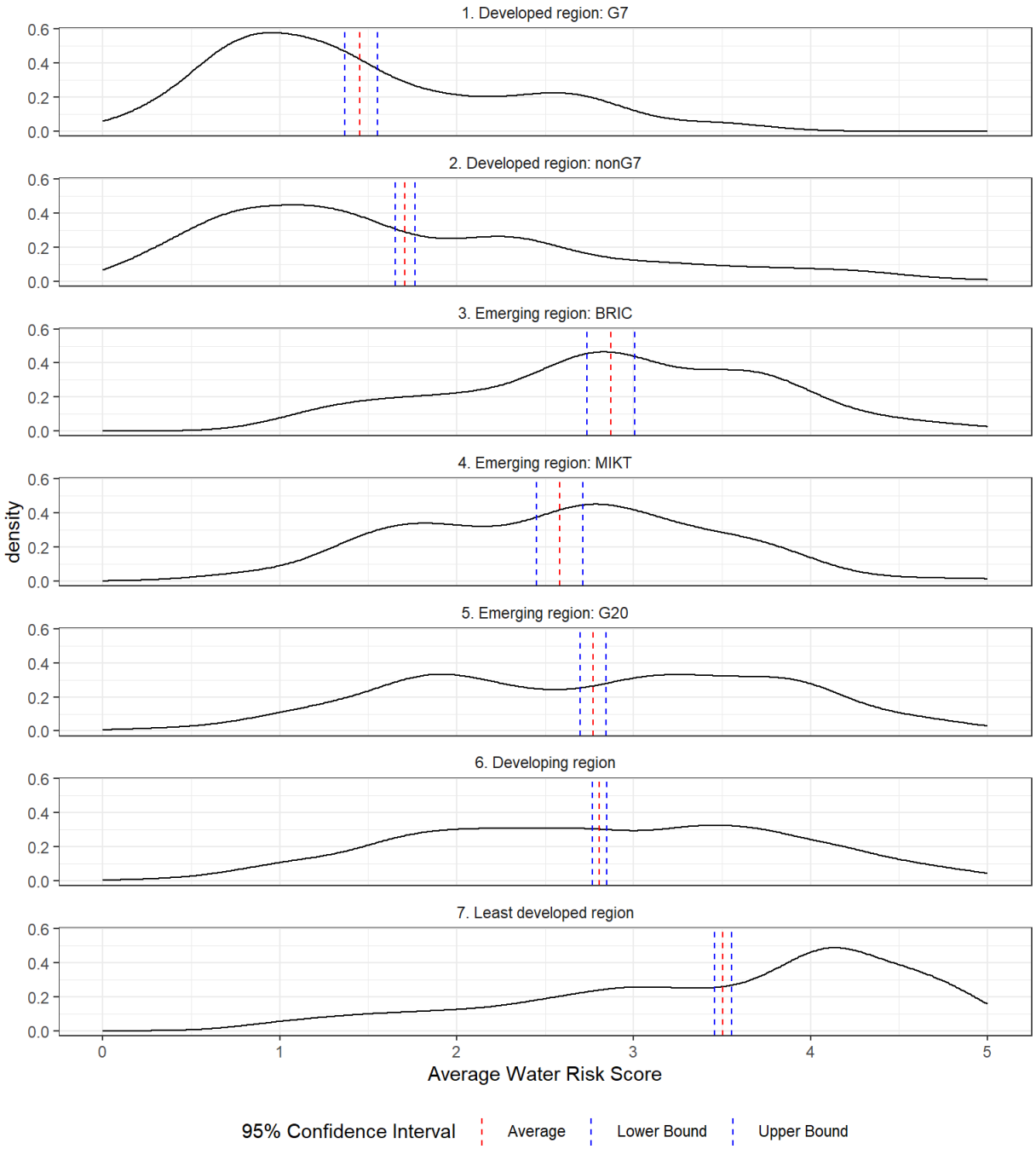
The density distribution of risk scores is then visualized, with the dashed lines representing the mean and the limits of the 95% confidence interval. This final display provides a better understanding of the distribution of risk scores and the variability of the data, as well as highlighting the differences between economies in terms of water risk. It confirms that developed regions tend to have lower risk scores and less variability, while developing and less developed regions have higher water risk scores and greater variability. The confidence intervals provide an indication of the precision of the average estimates for each region, with less developed regions generally showing narrower ranges.

```
# 95% Confidence Interval

conf_int_score95 <- GroupedWaterRisk %>%
  st_drop_geometry() %>%
  group_by(economy) %>%
  summarise(mean_cl_boot(score, conf.int=.95))

conf_int_score95
```

```
# A tibble: 7 × 4
  economy                      y  ymin  ymax
  <chr>                    <dbl> <dbl> <dbl>
1 1. Developed region: G7    1.45  1.37  1.55
2 2. Developed region: nonG7 1.71  1.65  1.77
3 3. Emerging region: BRIC   2.87  2.74  3.01
4 4. Emerging region: MIKT   2.58  2.45  2.71
5 5. Emerging region: G20    2.77  2.70  2.85
6 6. Developing region       2.81  2.77  2.85
7 7. Least developed region  3.50  3.46  3.55
```

```
ggplot(GroupedWaterRisk, aes(x = score)) +

  geom_density() +
  facet_wrap(vars(economy), ncol = 1) +

  geom_vline(data = conf_int_score95,
             mapping = aes(xintercept = y, color = "Average"),
```

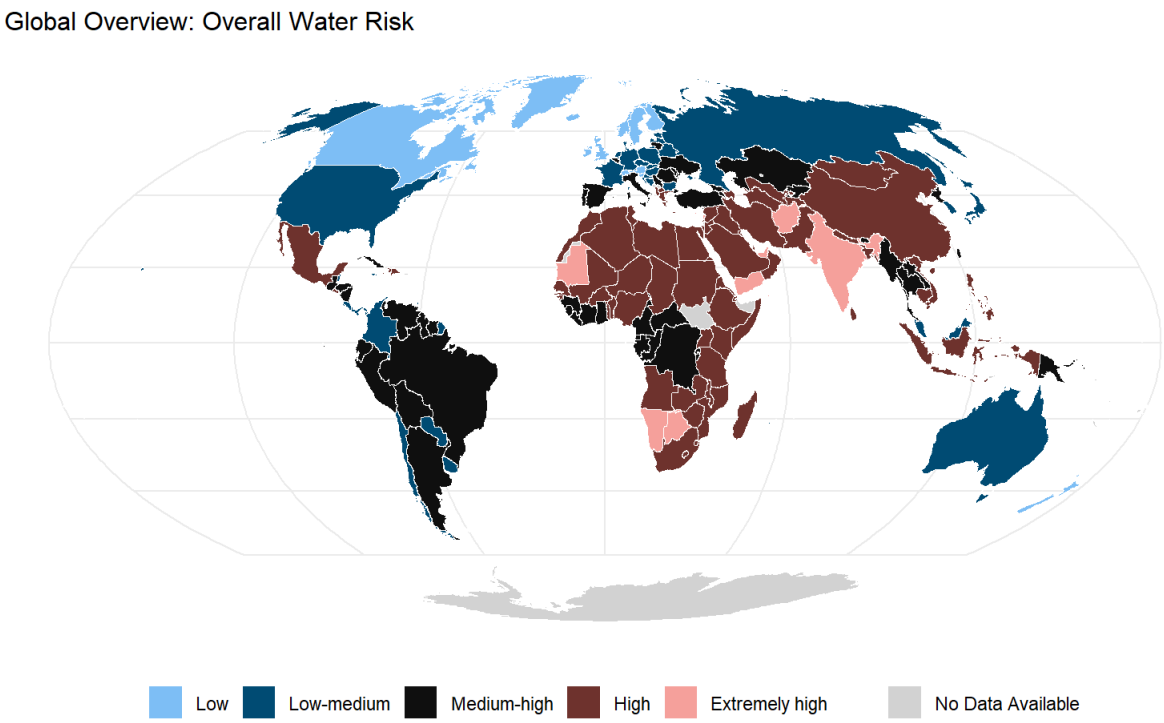
```
            linetype = "dashed") +

  geom_vline(data = conf_int_score95,
             mapping = aes(xintercept = ymin, color = "Lower Bound"),
             linetype = "dashed") +

  geom_vline(data = conf_int_score95,
             mapping = aes(xintercept = ymax, color = "Upper Bound"),
             linetype = "dashed") +

  scale_color_manual(values = c("Average" = "red",
                                "Lower Bound" = "blue",
                                "Upper Bound" = "blue"),
                     name = "95% Confidence Interval") +

  labs(x = "Average Water Risk Score") +
  theme_bw() +
  theme(legend.position = "bottom",
        strip.background = element_rect(fill="transparent", color="transparent"))
```



Global Overview: Overall Water Risk

The following step involved obtaining an glimpse of the big picture. Consequently, a map displaying the overall water risk indicator was created, highlighting how regions with extremely high water risk are concentrated in Africa and South Asia. In contrast, areas with low to medium-low water risk, indicated in light blue and dark blue respectively, are mainly located in North America, Europe and Oceania. Countries with medium to high water risk, marked in black, are mainly located in Latin America and parts of Asia.

```
ggplot() +

  geom_sf(data = filter(GroupedWaterRisk,
                        sector == "Default" & group == "Total, Overall water risk"),
          mapping = aes(fill = label),
          col = "white") +

  scale_fill_discrete_diverging(palette = "Berlin", name = "") +

  new_scale_fill() +

  geom_sf(data = missing_data,
          mapping = aes(
            fill = "No Data Available"),
          col = "white") +

  scale_fill_manual(values = c("No Data Available" = "lightgray"), name = "") +

  labs(title = "Global Overview: Overall Water Risk") +
  theme_bw() +
  theme(panel.border = element_blank(),
        legend.position = "bottom",
        title = element_text(size=rel(0.9), angle=0)) +

  coord_sf(crs="+proj=moll")
```



Sorting the countries according to their overall water risk score, the results show that the country facing the highest danger is Yemen, while Iceland has the lowest risk. Italy and Germany have been highlighted for comparison, allowing a better visualization of their relative position to other countries in the water risk distribution. Germany, in particular, ranks among the range of lowest water risk countries, possibly attributable to its favourable geological and climatic conformation.

```r
country_ordered <- GroupedWaterRisk %>%
  st_drop_geometry() %>%
  filter(sector == "Default" & group == "Total, Overall water risk") %>%
  arrange(desc(score)) %>%
  # useful for subsequent plot (it allows to visualize data in descending order)
  mutate(ISO = factor(ISO, levels = unique(ISO)))

head(country_ordered)
```

```
  ISO                   economy                  income    country  sector
1 YEM 7. Least developed region  4. Lower middle income      Yemen Default
2 BWA      6. Developing region  3. Upper middle income   Botswana Default
3 NAM      6. Developing region  3. Upper middle income    Namibia Default
4 BHR      6. Developing region 2. High income: nonOECD    Bahrain Default
5 LBN      6. Developing region  3. Upper middle income    Lebanon Default
6 MRT 7. Least developed region           5. Low income Mauritania Default
                     group      raw    score          label
1 Total, Overall water risk 3.347963 4.287631 Extremely high
2 Total, Overall water risk 3.365076 4.212266 Extremely high
3 Total, Overall water risk 3.419828 4.164558 Extremely high
4 Total, Overall water risk 3.049316 4.159188 Extremely high
5 Total, Overall water risk 3.019673 4.146411 Extremely high
6 Total, Overall water risk 3.306087 4.124800 Extremely high
```

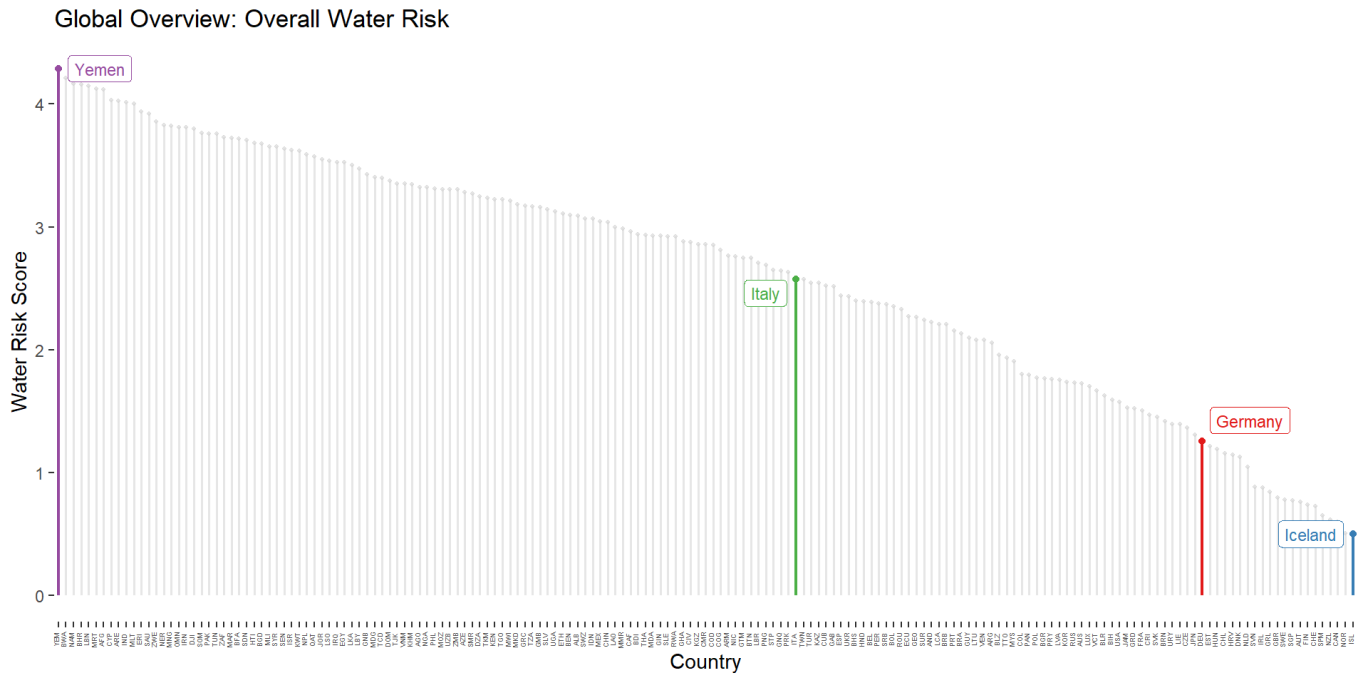
```r
tail(country_ordered)
```

```
    ISO                    economy                 income
168 CHE 2. Developed region: nonG7   1. High income: OECD
169 SPM 2. Developed region: nonG7 3. Upper middle income
170 NZL 2. Developed region: nonG7   1. High income: OECD
171 CAN    1. Developed region: G7   1. High income: OECD
172 NOR 2. Developed region: nonG7   1. High income: OECD
173 ISL 2. Developed region: nonG7   1. High income: OECD
                      country  sector                    group       raw
168               Switzerland Default Total, Overall water risk 0.7279870
169 Saint Pierre and Miquelon Default Total, Overall water risk 0.6612553
170               New Zealand Default Total, Overall water risk 0.6234959
171                    Canada Default Total, Overall water risk 0.5507107
172                    Norway Default Total, Overall water risk 0.5094146
173                   Iceland Default Total, Overall water risk 0.5072003
        score label
168 0.7274644   Low
169 0.6547082   Low
170 0.6181655   Low
171 0.6019038   Low
172 0.5070363   Low
173 0.5021785   Low
```

```r
# countries to highlight in the plot

highlight <- c(
  country_ordered$country[1],                      # highest score
  country_ordered$country[nrow(country_ordered)],  # lowest score
  "Italy",
  "Germany"
)
```

```r
ggplot(data = country_ordered,
       mapping = aes(x = ISO,
                     y = score)) +

  geom_point(col="lightgrey", size=0.8, alpha=0.5) +

  geom_segment(mapping = aes(xend = ISO,
                             y = 0, yend = score),
```

```
                col="lightgrey", linewidth=0.6, alpha=0.5) +

  geom_point(data = function(d) {filter(d, country %in% highlight)},
             aes(color=country)) +

  geom_segment(data = function(d) {filter(d, country %in% highlight)},
               aes(xend = ISO, y = 0, yend = score,
                   color=country), linewidth=0.8) +

  geom_label_repel(data = function(d) {filter(d, country %in% highlight)},
                   aes(label=country, color=country),
                   size = 3) +

  scale_color_brewer(palette="Set1", type="qual") +

  labs(title = "Global Overview: Overall Water Risk",
       x = "Country",
       y = "Water Risk Score") +
  theme_classic() +
  theme(legend.position = "none",
        axis.line = element_blank(),
        axis.text.x = element_text(angle=90, size=3.5, vjust=0))
```



Global Overview: Overall Water Risk Across Economic Regions

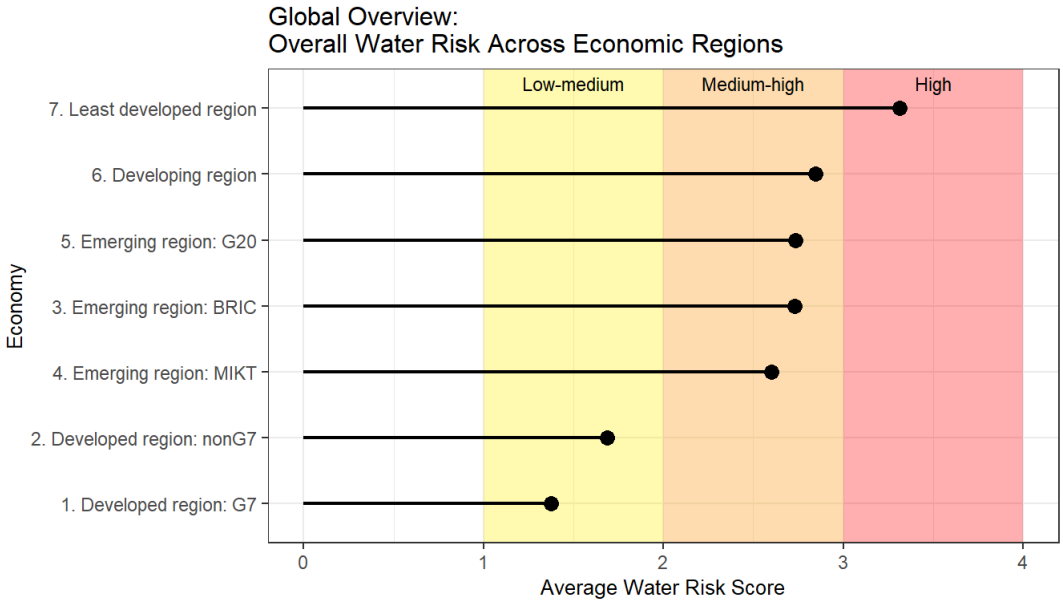
Further insight was gained through a bar chart plotting the same data by economic region. This chart enabled a comparison of the average water risk scores across different economies, revealing once again the least developed region as the most vulnerable in terms of water risk.

```
risk_economy <- GroupedWaterRisk %>%

  # drop the geometry to facilitate and improve the speed of calculation
  st_drop_geometry() %>%

  filter(sector == "Default" & group == "Total, Overall water risk") %>%
  group_by(economy) %>%
  summarise(avg_score = mean(score)) %>%
  mutate(economy = fct_reorder(economy,
                               avg_score))
```

```
risk_economy
```

```
# A tibble: 7 × 2
  economy                    avg_score
  <fct>                          <dbl>
1 1. Developed region: G7          1.37
2 2. Developed region: nonG7       1.69
3 3. Emerging region: BRIC         2.73
4 4. Emerging region: MIKT         2.60
5 5. Emerging region: G20          2.74
6 6. Developing region             2.84
7 7. Least developed region        3.31
```

```
ggplot(data = risk_economy) +

  geom_rect(aes(xmin = -Inf,
                xmax = Inf,
                ymin = 1,
                ymax = 2),
            fill="#fff200",
            alpha=0.05) +

  geom_rect(aes(xmin = -Inf,
                xmax = Inf,
                ymin = 2,
                ymax = 3),
            fill="#ff9900",
            alpha=0.05) +

  geom_rect(aes(xmin = -Inf,
                xmax = Inf,
                ymin = 3,
                ymax = 4),
            fill="#ff0000",
            alpha=0.05) +

  annotate("text", y = 1.5, x = Inf, label = "Low-medium",
           size=3, vjust = 1.5) +
  annotate("text", y = 2.5, x = Inf, label = "Medium-high",
           size=3, vjust = 1.5) +
  annotate("text", y = 3.5, x = Inf, label = "High",
           size=3, vjust = 1.5) +

  geom_point(mapping = aes(x = economy,
                           y = avg_score),
             col="black", size=3) +
  geom_segment(mapping = aes(x = economy, xend = economy,
                             y = 0, yend = avg_score),
               col="black", linewidth=0.8) +

  coord_flip() +
  labs(title = "Global Overview:\nOverall Water Risk Across Economic Regions",
       x = "Economy",
       y = "Average Water Risk Score") +
    theme_bw() +
  theme(title = element_text(size=rel(0.9), angle=0))
```



Sector-wise Comparison of Risk Indicators

To elaborate further, it is worth examining some comparisons to understand the variations in risk under different circumstances.

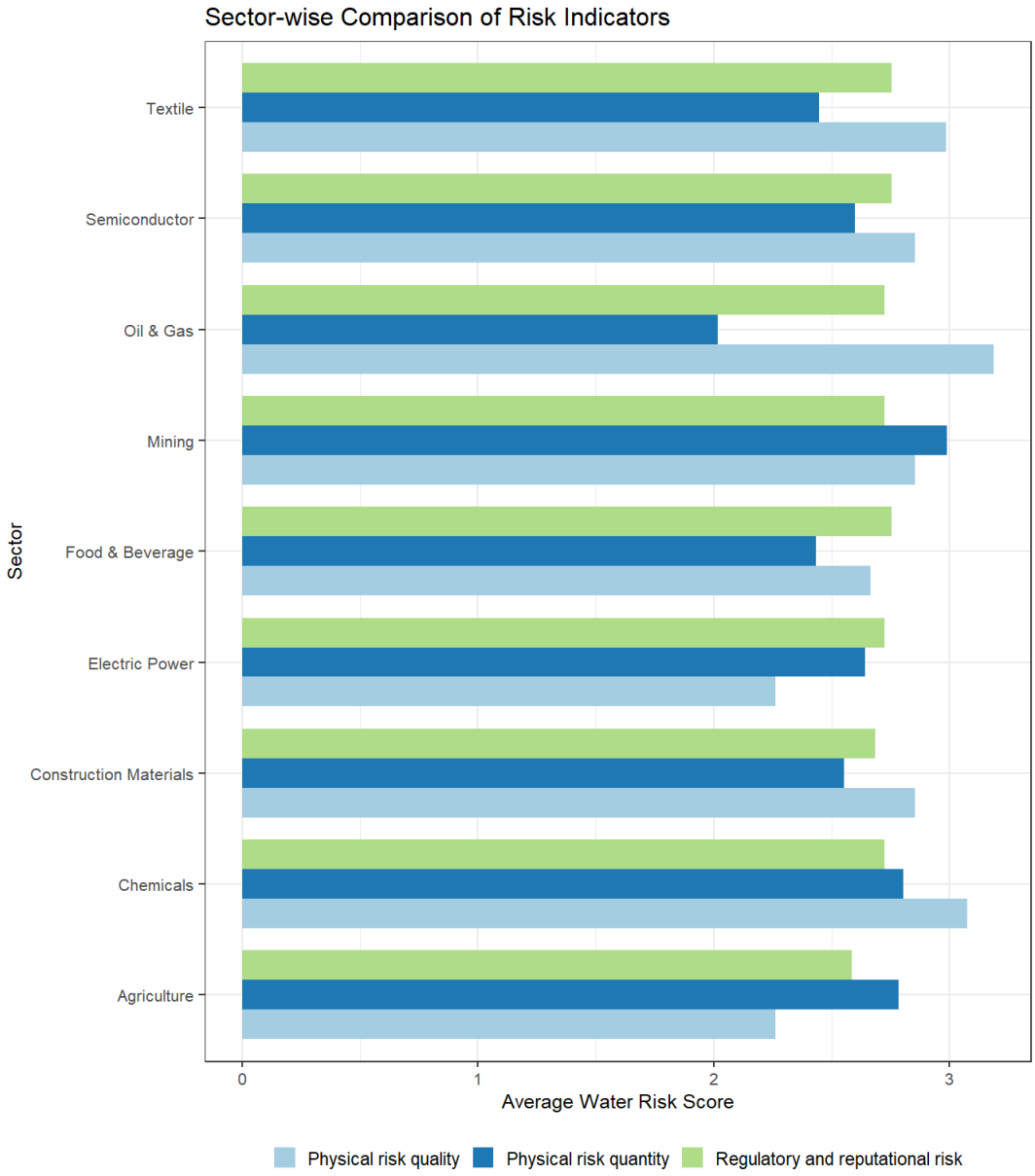
The plots that follow allow to see whether the three risk indicators have different impacts across economic sectors. The physical risk quality appears to be predominant, but the regulatory and reputational side also plays a significant role across industries. For the purpose of achieving an even clearer representation, the graph was further divided into three parts, one for each indicator. This made it possible to see, for each risk indicator, which sector is most strongly impacted by it.

```
avg_sector_group <- GroupedWaterRisk %>%
  st_drop_geometry() %>%
  filter(sector != "Default" & group != "Total, Overall water risk") %>%
  group_by(sector, group) %>%
  summarise(avg_score = mean(score)) %>%
  mutate(sector = fct_reorder(sector, avg_score))

head(avg_sector_group)
```

```
# A tibble: 6 × 3
# Groups:   sector [2]
  sector      group                             avg_score
  <fct>       <chr>                                 <dbl>
1 Agriculture Physical risk quality                  2.26
2 Agriculture Physical risk quantity                 2.79
3 Agriculture Regulatory and reputational risk       2.59
4 Chemicals   Physical risk quality                  3.08
5 Chemicals   Physical risk quantity                 2.80
6 Chemicals   Regulatory and reputational risk       2.73
```

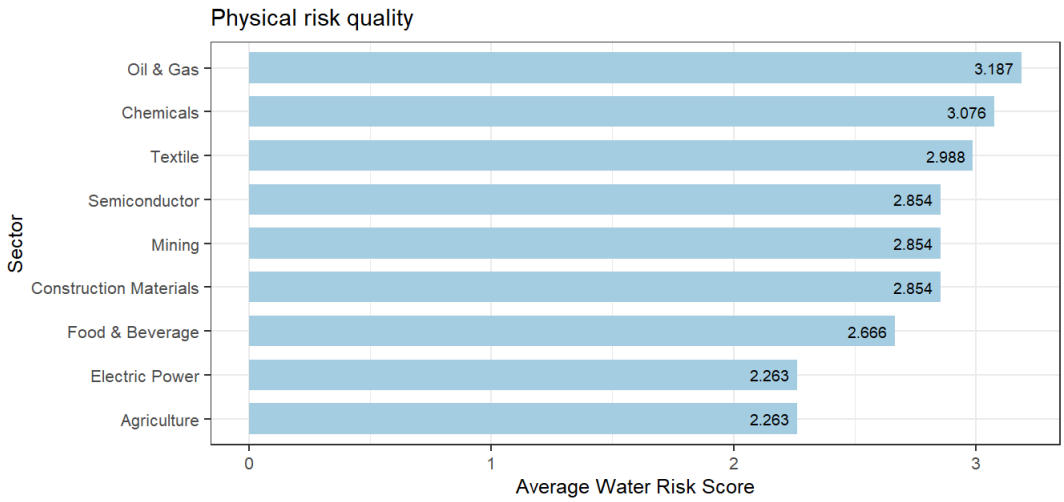
```
avg_sector_group %>%

  ggplot(mapping = aes(x = sector, y = avg_score, fill = group)) +

  geom_col(position = "dodge", width = 0.8) +

  scale_fill_brewer(palette="Paired",
                    type="qual") +

  coord_flip() +
  labs(title = "Sector-wise Comparison of Risk Indicators",
       x = "Sector",
       y = "Average Water Risk Score") +
  theme_bw() +
```

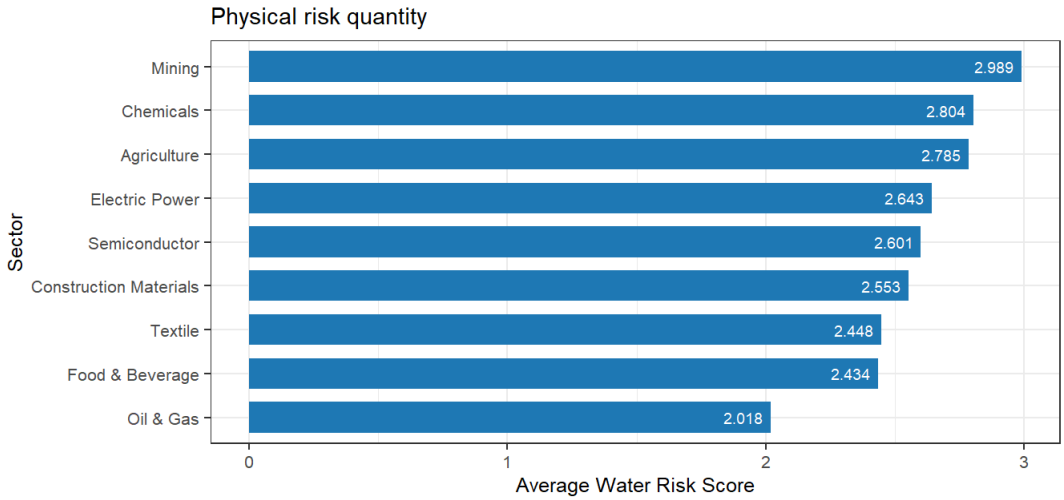
```
  theme(legend.position = "bottom",
        legend.key.size = unit(0.4, "cm"),
        legend.title = element_blank(),
        title = element_text(size=10),
        axis.text = element_text(size=8),
        axis.title = element_text(size=9.5),
        legend.text = element_text(size=8.8))
```

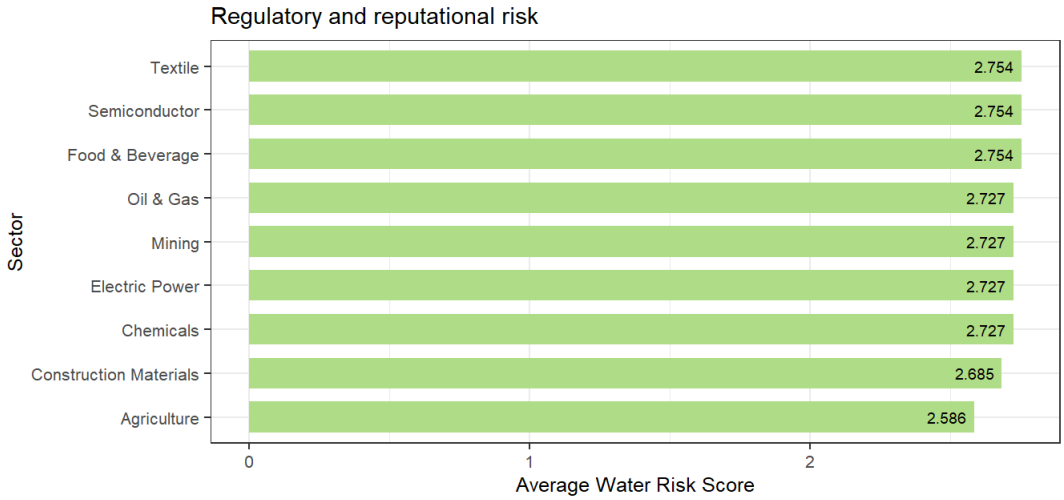



```
plot1 <- avg_sector_group %>%
  filter(group == "Physical risk quality") %>%

  ggplot(mapping = aes(x = fct_reorder(sector, avg_score), y = avg_score)) +

  geom_col(fill="#a6cee3", width = 0.7) +
  geom_text(
    aes(label = sprintf("%.3f", avg_score)),
    col = "black", size = 2.5, hjust = 1.2) +

  coord_flip() +
  labs(title = "Physical risk quality",
       x = "Sector",
       y = "Average Water Risk Score") +
  theme_bw() +
  theme(legend.position = "none",
        title = element_text(size=9),
```

```
        axis.text = element_text(size=8),
        axis.title = element_text(size=9.5))

plot2 <- avg_sector_group %>%
  filter(group == "Physical risk quantity") %>%

  ggplot(mapping = aes(x = fct_reorder(sector, avg_score), y = avg_score)) +

  geom_col(fill="#1f78b4", width = 0.7) +
  geom_text(
    aes(label = sprintf("%.3f", avg_score)),
    col = "white", size = 2.5, hjust = 1.2) +

  coord_flip() +
  labs(title = "Physical risk quantity",
       x = "Sector",
       y = "Average Water Risk Score") +
  theme_bw() +
  theme(legend.position = "none",
        title = element_text(size=9),
        axis.text = element_text(size=8),
        axis.title = element_text(size=9.5))

plot3 <- avg_sector_group %>%
  filter(group == "Regulatory and reputational risk") %>%

  ggplot(mapping = aes(x = fct_reorder(sector, avg_score), y = avg_score)) +

  geom_col(fill="#b2df8a", width = 0.7) +
  geom_text(
    aes(label = sprintf("%.3f", avg_score)),
    col = "black", size = 2.5, hjust = 1.2) +

  coord_flip() +
  labs(title = "Regulatory and reputational risk",
       x = "Sector",
       y = "Average Water Risk Score") +
  theme_bw() +
  theme(legend.position = "none",
        title = element_text(size=9),
        axis.text = element_text(size=8),
        axis.title = element_text(size=9.5))

grid.arrange(plot1, plot2, plot3, ncol = 1)
```





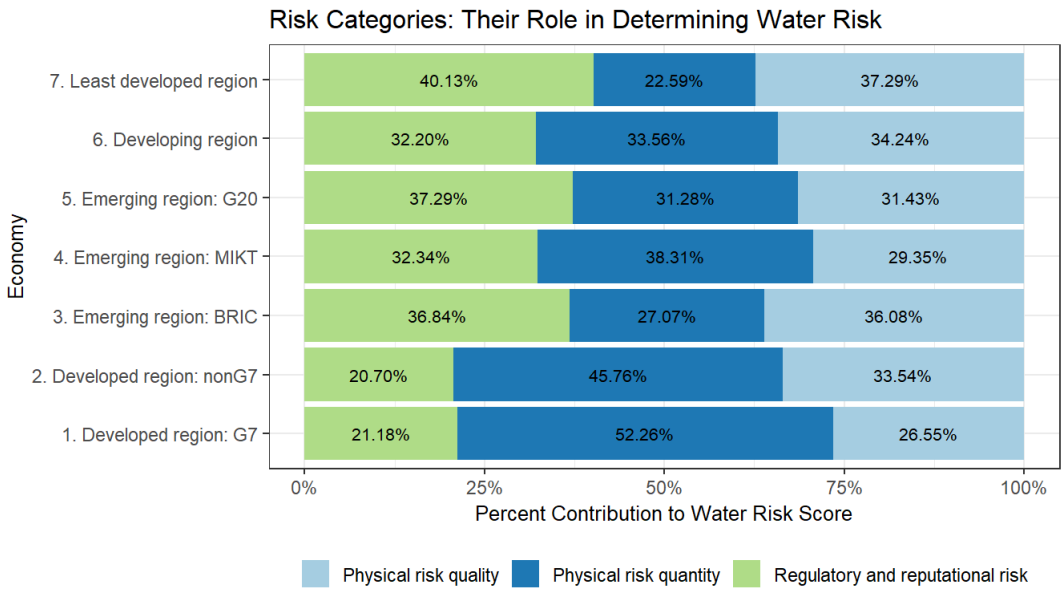


Risk Categories: Their Role in Determining Water Risk

Regarding the impact of different indicators in determining water risk, evidence shows that regulatory and reputational aspects play an equally important, if not more dominant role than physical water quantity and quality. This is particularly noticeable in less developed countries or emerging economies, with a contribution near to 40%.

```
GroupedWaterRisk %>%
  st_drop_geometry() %>%
  filter(group != "Total, Overall water risk") %>%
  group_by(economy, group) %>%
  summarise(avg_score = mean(score)) %>%
  mutate(percent = avg_score / sum(avg_score)) %>%
```

```
ggplot(mapping = aes(x = economy,
                     y = avg_score,
                     fill = group)) +

geom_col(position = "fill") +

stat_summary(geom = "text",
             aes(label = percent(percent, accuracy = 0.01)),
             fun = "mean", vjust = 0.5, size = 2.8,
             position = position_fill(vjust = 0.5),
             color = "black") +

scale_fill_brewer(palette="Paired",
                  type="qual") +

scale_y_continuous(labels = scales::percent) +

coord_flip() +
labs(title = "Risk Categories: Their Role in Determining Water Risk",
     x = "Economy",
     y = "Percent Contribution to Water Risk Score") +
theme_bw() +
theme(legend.position = "bottom",
      legend.key.size = unit(0.5, "cm"),
      legend.title = element_blank(),
      title = element_text(size=rel(0.9), angle=0))
```



Understanding the Cost of a Sustainable Water Future

The final step was to see what the future might hold, and this is where the second dataset comes into play. At this stage, the costs required to achieve a sustainable future of water resources in different countries, regions and income groups by 2030 are therefore explored.

Data and Exploratory Analysis

Data Cleaning and Preprocessing

Following on from the initial portion of the analysis, the first step is to clean and prepare the data. This process includes loading the dataset and renaming the columns for clarity. It can be seen that there are no missing values in this case, so no further action is needed in this direction.

```r
Achieving_Abundance <- read_excel(
  "data/Achieving_Abundance_Countries.xlsx",
  sheet = "Countries", skip = 2,
  col_names = c(
    "ISO",
    "country",
    "Access_Drinking_Water",
    "Access_Sanitation",
    "Industrial_Pollution",
    "Agricultural_Pollution",
    "Water_Scarcity",
    "Water_Management",
    "Total",
    "PPP",
    "GDP2030",
    "Population2030",
    "WaterDemand2030"
  ))
```

```r
str(Achieving_Abundance)
```

```
tibble [162 × 13] (S3: tbl_df/tbl/data.frame)
 $ ISO                  : chr [1:162] "AFG" "ALB" "DZA" "AND" ...
 $ country              : chr [1:162] "Afghanistan" "Albania" "Algeria" "Andorra" ...
 $ Access_Drinking_Water : num [1:162] 4.27e+08 2.46e+07 8.91e+08 9.64e+05 1.91e+09 ...
 $ Access_Sanitation     : num [1:162] 3.70e+08 1.97e+07 1.25e+09 9.90e+05 1.16e+09 ...
 $ Industrial_Pollution  : num [1:162] 4.18e+06 2.26e+07 7.86e+08 1.27e+04 9.88e+08 ...
 $ Agricultural_Pollution: num [1:162] 3.80e+07 3.33e+07 1.43e+08 5.31e+05 6.13e+08 ...
 $ Water_Scarcity        : num [1:162] 1.29e+09 2.35e+08 1.10e+09 9.37e+04 3.14e+08 ...
 $ Water_Management      : num [1:162] 4.26e+08 6.70e+07 8.33e+08 5.18e+05 9.97e+08 ...
 $ Total                 : num [1:162] 2.56e+09 4.02e+08 5.00e+09 3.11e+06 5.98e+09 ...
 $ PPP                   : num [1:162] 0.296 0.349 0.285 1 0.555 ...
 $ GDP2030               : num [1:162] 1.07e+11 4.07e+10 5.46e+11 3.97e+09 2.18e+11 ...
 $ Population2030        : num [1:162] 45487093 3075379 41844172 105440 28072257 ...
 $ WaterDemand2030       : num [1:162] 2.70e+10 2.18e+09 1.42e+10 3.88e+05 4.07e+09 ...
```

```r
sum(is.na(Achieving_Abundance))
```

```
[1] 0
```

After the above steps have been completed, geometry data have been added.

```r
# add the geometry to the main dataset
Achieving_Abundance <- merge(countries[,-1], Achieving_Abundance, by="ISO")
```

The calculation of the estimated GDP per capita for 2030 (`GDPperCap2030`) is subsequently calculated by dividing the total projected GDP for 2030 (`GDP2030`) by the estimated population for 2030 (`Population2030`). This value is important for understanding the economic capacity of each country to meet the costs required to achieve water sustainability.

Next, data were filtered to include only those countries in the GroupedWaterRisk dataset, using the ISO country codes. This filtering is essential to ensure that the analysis only focuses on countries for which comprehensive water risk data are available, as well as to ensure consistency with the countries considered in the first part of the project.

```r
Achieving_Abundance <- Achieving_Abundance %>%

  # Calculate the estimated GDP per Capita in 2030
  mutate(GDPperCap2030 = GDP2030 / Population2030) %>%

  # filter to select just the countries present in the GroupedWaterRisk dataset
  filter(ISO %in% GroupedWaterRisk$ISO)
```

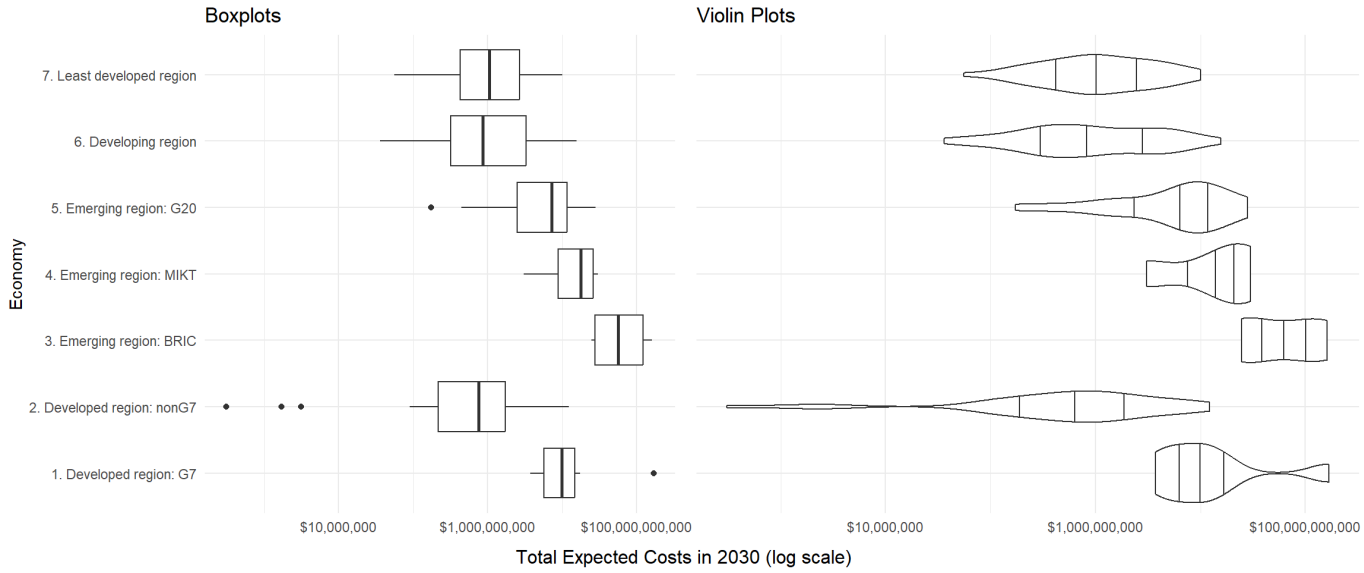
Descriptive Statistics

Results of the descriptive statistics reveal considerable variability in key variables such as access to drinking water, sanitation, industrial and agricultural pollution, water scarcity, water management, total and per capita GDP, population and projected water demand in 2030. The plots, which combine boxplots and violin plots, show the distributions of expected total costs and GDP per capita in 2030 for different economic categories. Less developed regions present low expected costs and GDP per capita with narrow distributions, while developed regions (G7) show very high costs and GDP per capita with more concentrated distributions. Emerging regions (BRIC, MIKT, G20) exhibit significant variability in both costs and GDP per capita, indicating significant internal differences.

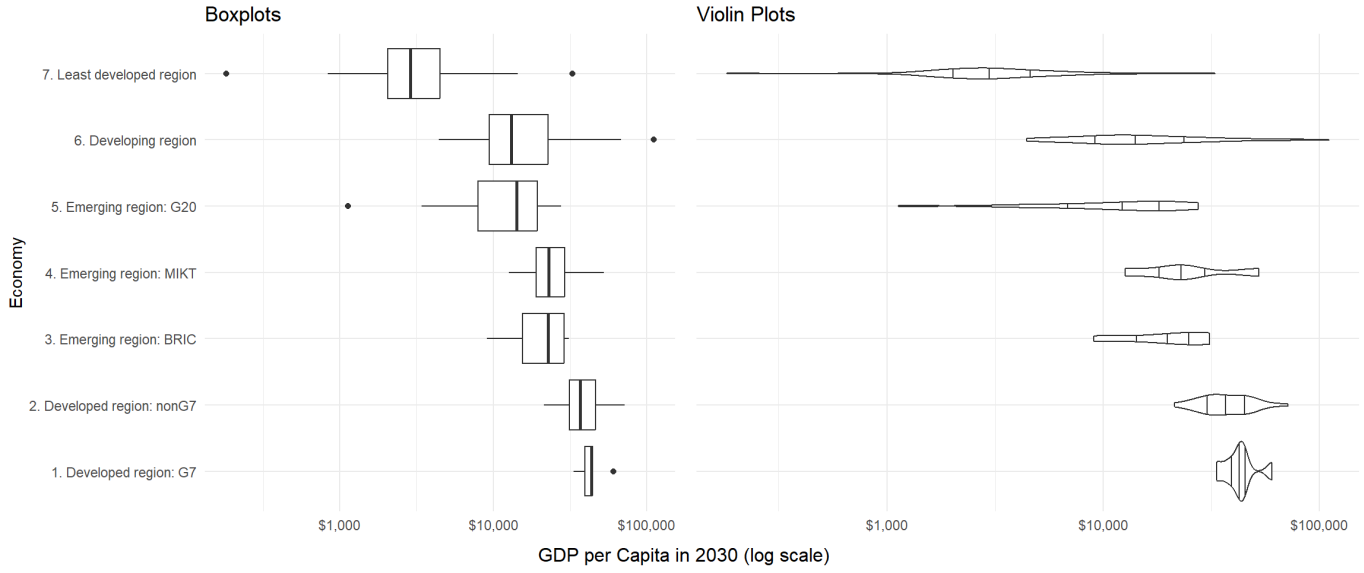
```
Achieving_Abundance %>%
  select_if(is.numeric) %>%
  summary()
```

```
Access_Drinking_Water Access_Sanitation   Industrial_Pollution
Min.   :0.000e+00     Min.   :0.000e+00   Min.   :0.000e+00
1st Qu.:5.977e+07     1st Qu.:6.599e+07   1st Qu.:1.242e+07
Median :1.825e+08     Median :1.947e+08   Median :4.925e+07
Mean   :7.053e+08     Mean   :9.351e+08   Mean   :5.431e+08
3rd Qu.:6.787e+08     3rd Qu.:6.597e+08   3rd Qu.:3.712e+08
Max.   :1.263e+10     Max.   :2.753e+10   Max.   :1.456e+10
Agricultural_Pollution Water_Scarcity      Water_Management
Min.   :0.000e+00      Min.   :0.000e+00   Min.   :5.119e+04
1st Qu.:1.588e+07      1st Qu.:2.670e+07   1st Qu.:6.948e+07
Median :6.082e+07      Median :2.342e+08   Median :2.105e+08
Mean   :4.132e+08      Mean   :2.766e+09   Mean   :1.073e+09
3rd Qu.:3.097e+08      3rd Qu.:1.368e+09   3rd Qu.:8.301e+08
Max.   :7.735e+09      Max.   :1.139e+11   Max.   :2.806e+10
    Total                PPP             GDP2030          Population2030
Min.   :3.071e+05   Min.   :0.2620   Min.   :1.000e+09   Min.   :2.995e+04
1st Qu.:4.169e+08   1st Qu.:0.3628   1st Qu.:4.566e+10   1st Qu.:4.268e+06
Median :1.263e+09   Median :0.4424   Median :1.332e+11   Median :1.110e+07
Mean   :6.435e+09   Mean   :0.5312   Mean   :9.276e+11   Mean   :4.899e+07
3rd Qu.:4.981e+09   3rd Qu.:0.5980   3rd Qu.:4.892e+11   3rd Qu.:3.826e+07
Max.   :1.684e+11   Max.   :1.2886   Max.   :3.805e+13   Max.   :1.449e+09
WaterDemand2030     GDPperCap2030              geometry
Min.   :3.880e+05   Min.   :   182.1   MULTIPOLYGON :161
1st Qu.:1.024e+09   1st Qu.:  5786.4   epsg:4326    :  0
Median :3.939e+09   Median : 14483.8   +proj=long...:  0
Mean   :2.890e+10   Mean   : 20145.3
3rd Qu.:2.167e+10   3rd Qu.: 30984.7
Max.   :9.589e+11   Max.   :110725.3
```

```
plot1 <- ggplot(data = Achieving_Abundance, mapping = aes(x = economy,
                                                          y = Total)) +
  geom_boxplot() +
  scale_y_log10(labels = scales::dollar) +
  coord_flip() +
  labs(title="Boxplots", x="Economy") +
  theme_minimal() +
  theme(axis.title.x = element_blank())

plot2 <- ggplot(data = Achieving_Abundance, mapping = aes(x = economy,
                                                          y = Total)) +
  geom_violin(draw_quantiles = c(.25,.5,.75)) +
  scale_y_log10(labels = scales::dollar) +
  coord_flip() +
  labs(title="Violin Plots") +
  theme_minimal() +
  theme(axis.title.y = element_blank(),
        axis.text.y = element_blank(),
        axis.ticks.y = element_blank(),
        axis.title.x = element_blank())

grid.arrange(plot1, plot2, ncol = 2,
             bottom = "Total Expected Costs in 2030 (log scale)")
```




```
plot1 <- ggplot(data = Achieving_Abundance, mapping = aes(x = economy,
                                                          y = GDPperCap2030)) +
  geom_boxplot() +
  scale_y_log10(labels = scales::dollar) +
  coord_flip() +
  labs(title="Boxplots", x="Economy") +
  theme_minimal() +
  theme(axis.title.x = element_blank())

plot2 <- ggplot(data = Achieving_Abundance, mapping = aes(x = economy,
                                                          y = GDPperCap2030)) +
  geom_violin(draw_quantiles = c(.25,.5,.75)) +
  scale_y_log10(labels = scales::dollar) +
  coord_flip() +
  labs(title="Violin Plots") +
  theme_minimal() +
  theme(axis.title.y = element_blank(),
        axis.text.y = element_blank(),
        axis.ticks.y = element_blank(),
        axis.title.x = element_blank())

grid.arrange(plot1, plot2, ncol = 2,
             bottom = "GDP per Capita in 2030 (log scale)")
```



```
# 95% Confidence Interval

conf_int_Total95 <- Achieving_Abundance %>%
  st_drop_geometry() %>%
  summarise(mean_cl_boot(Total, conf.int=.95))

conf_int_GDPperCap95 <- Achieving_Abundance %>%
  st_drop_geometry() %>%
  summarise(mean_cl_boot(GDPperCap2030, conf.int=.95))

conf_int_Total95
```

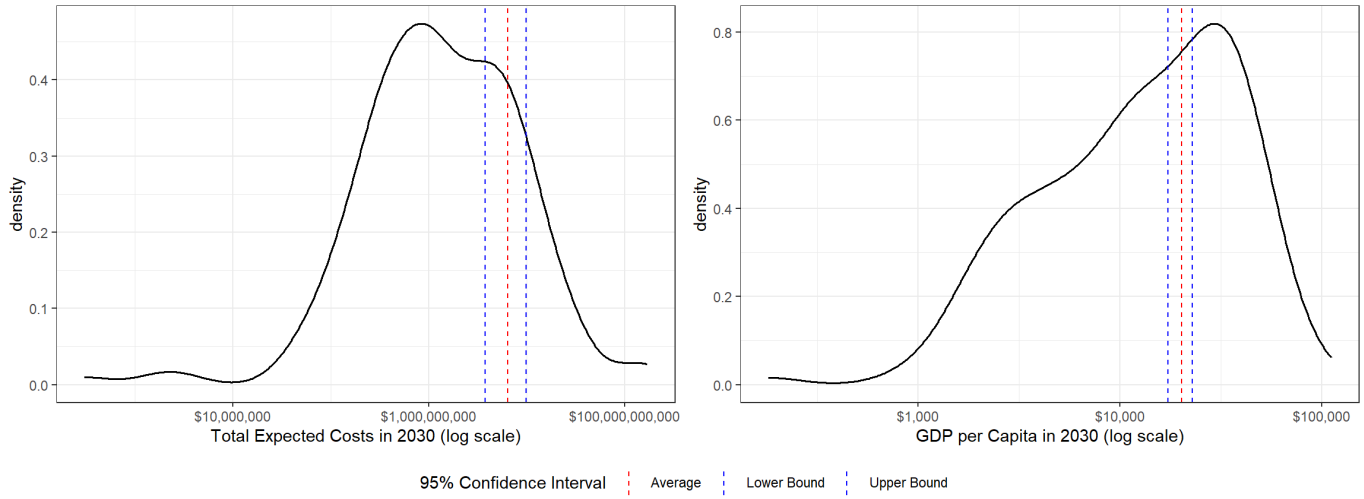
```
          y       ymin       ymax
1 6435394975 3773059140 9893335182
```

```
conf_int_GDPperCap95
```

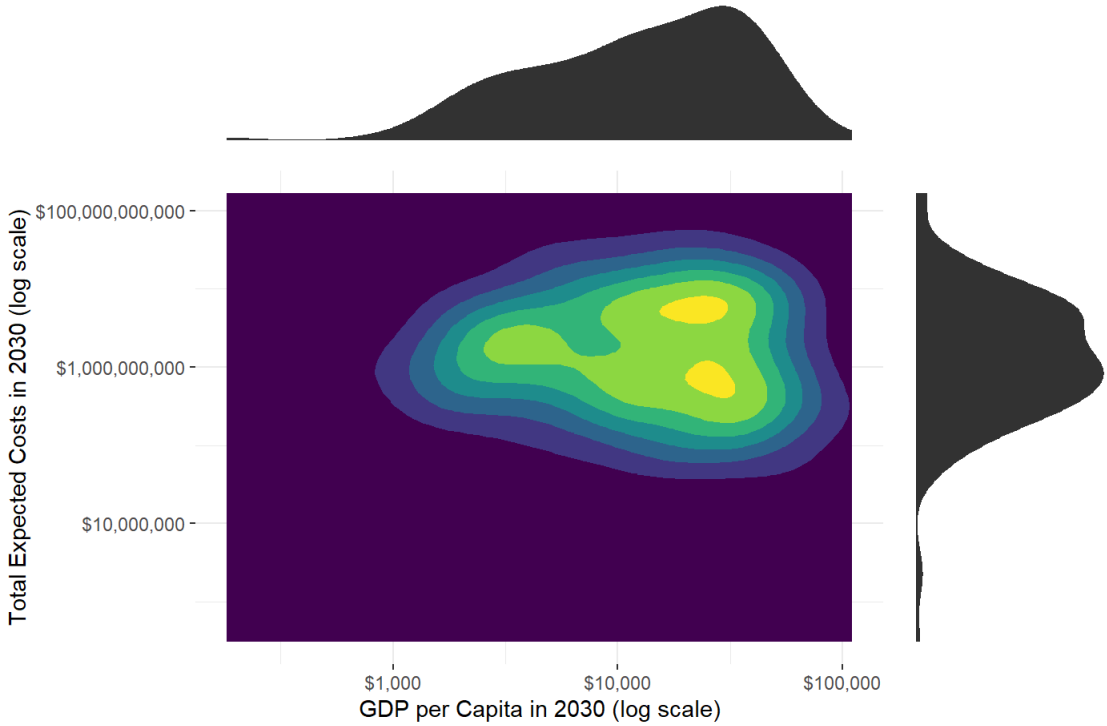
```
        y     ymin     ymax
1 20145.34 17255.56 22768.81
```

```
# Function to extract the legend from a graph (avoids repetition)
extract_legend <- function(plot) {
  g <- ggplotGrob(plot)
  legend <- g$grobs[[which(sapply(g$grobs, function(x) x$name) == "guide-box")]]
  return(legend)
}

plot1 <- ggplot(data = Achieving_Abundance,
                mapping = aes(x = Total)) +
  geom_density(linewidth=0.6) +
  geom_vline(data = conf_int_Total95,
             mapping = aes(xintercept = y, color = "Average"),
             linetype = "dashed") +
  geom_vline(data = conf_int_Total95,
             mapping = aes(xintercept = ymin, color = "Lower Bound"),
             linetype = "dashed") +
  geom_vline(data = conf_int_Total95,
             mapping = aes(xintercept = ymax, color = "Upper Bound"),
             linetype = "dashed") +
  scale_color_manual(values = c("Average" = "red",
                                "Lower Bound" = "blue",
                                "Upper Bound" = "blue"),
                     name = "95% Confidence Interval") +
  scale_x_log10(labels = scales::dollar) +
  labs(x = "Total Expected Costs in 2030 (log scale)") +
  theme_bw() +
  theme(legend.position = "none")

plot2 <- ggplot(data = Achieving_Abundance,
                mapping = aes(x = GDPperCap2030)) +
  geom_density(linewidth=0.6) +
  geom_vline(data = conf_int_GDPperCap95,
             mapping = aes(xintercept = y, color = "Average"),
             linetype = "dashed") +
  geom_vline(data = conf_int_GDPperCap95,
             mapping = aes(xintercept = ymin, color = "Lower Bound"),
             linetype = "dashed") +
  geom_vline(data = conf_int_GDPperCap95,
             mapping = aes(xintercept = ymax, color = "Upper Bound"),
             linetype = "dashed") +
  scale_color_manual(values = c("Average" = "red",
                                "Lower Bound" = "blue",
                                "Upper Bound" = "blue"),
                     name = "95% Confidence Interval") +
  scale_x_log10(labels = scales::dollar) +
  labs(x = "GDP per Capita in 2030 (log scale)") +
```

```
  theme_bw() +
  theme(legend.position = "none")

legend <- extract_legend(plot1 + theme(legend.position = "bottom"))

grid.arrange(plot1, plot2, ncol = 2, bottom=legend)
```




```
p_main <- ggplot(Achieving_Abundance, aes(x = GDPperCap2030,
                                          y = Total)) +
  stat_density2d_filled() +
  scale_x_log10(labels = scales::dollar) +
  scale_y_log10(labels = scales::dollar) +
  scale_fill_viridis(discrete=T, option="D") +
  labs(x = "GDP per Capita in 2030 (log scale)",
       y = "Total Expected Costs in 2030 (log scale)") +
  theme_bw() +
  theme(panel.border = element_blank(),
        legend.position="none")

p_GDPperCap2030 <- ggplot(Achieving_Abundance, aes(x = GDPperCap2030)) +
  stat_density() +
  scale_x_log10(labels = scales::dollar) +
  theme_void()

p_Total <- ggplot(Achieving_Abundance, aes(x = Total)) +
  stat_density() +
  scale_x_log10(labels = scales::dollar) +
  coord_flip() +
  theme_void()

wrap_plots(p_GDPperCap2030, plot_spacer(),
           p_main, p_Total,
           nrow = 2,
           widths = c(1, 0.3),
           heights = c(0.3, 1)
)
```



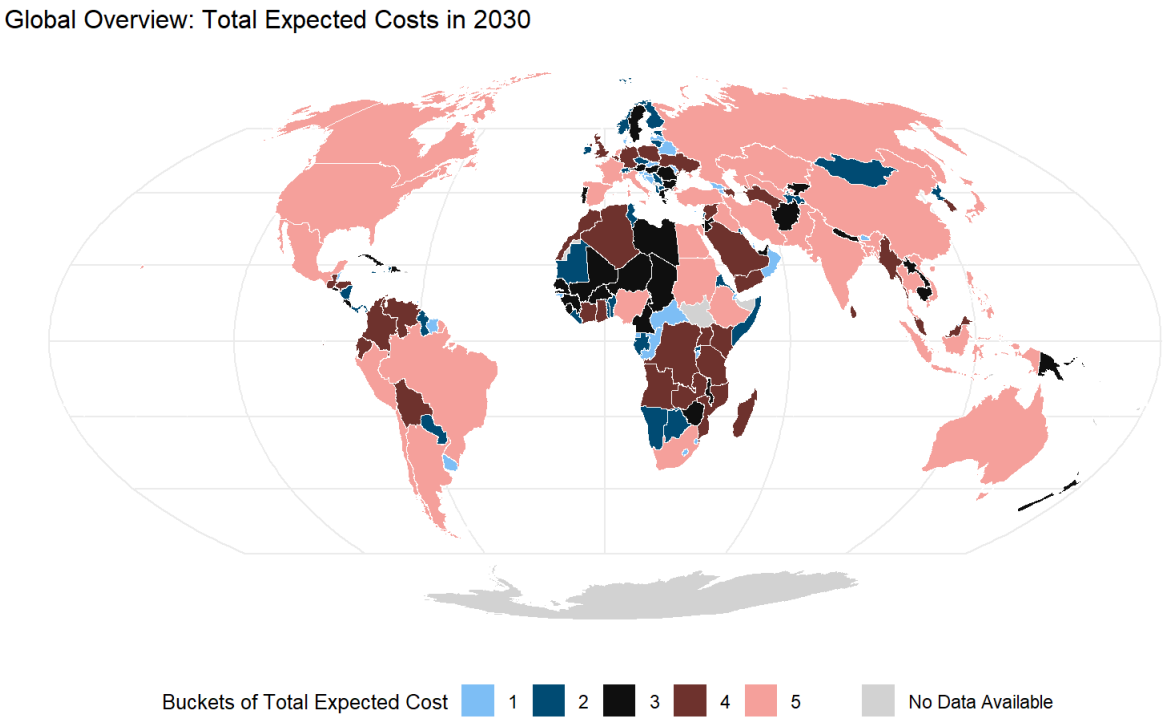
The single density plots show asymmetric distributions of expected total costs and GDP per capita in 2030, with higher densities in specific ranges. The graph on the left has a main peak around $1 billion, while the second has a peak around $50,000. Both include a 95% confidence interval, shown by blue dotted lines for the lower and upper bounds and a red dashed line for the mean. Furthermore, the relationship between total costs and GDP per capita is shown in the bivariate density graph, with coloured contours representing different density levels and marginal distributions displayed on the sides. This graph suggests a correlation between the two parameters, with areas of higher density indicating areas of forecast concentration. In conclusion, the data and graphs highlight economic disparities between regions, suggesting that strategies for a sustainable water future must be tailored to the specific needs and capabilities of each area.

Global Overview: Total Expected Costs in 2030

Once again, it may be of interest to obtain an initial overall picture of the estimated situation. The following map was constructed using the same colour palette as the first Overall Water Risk map, in order to facilitate a direct comparison between the two. It is interesting to note that even many of those regions with a low to medium risk will have to invest heavily and incur high costs in the future to ensure sustainable water management.

```
ggplot() +

  geom_sf(data = Achieving_Abundance,
          mapping = aes(fill = factor(ntile(Total,n=5))),
          col = "white") +

  scale_fill_discrete_diverging(palette = "Berlin",
                                name = "Buckets of Total Expected Cost") +

  new_scale_fill() +

  geom_sf(data = missing_data,
          mapping = aes(
            fill = "No Data Available"),
          col = "white") +

  scale_fill_manual(values = c("No Data Available" = "lightgray"),
                    name = "") +

  labs(title = "Global Overview: Total Expected Costs in 2030") +
  theme_bw() +
  theme(panel.border = element_blank(),
```

```
        legend.position = "bottom",
        title = element_text(size=rel(0.9), angle=0)) +

  coord_sf(crs="+proj=moll")
```



Projected 2030 Average Costs Across Economic Regions and Income Levels

There will certainly be differences here too, and that is precisely what the following graphs aim to uncover.

As an initial step, the intention is to investigate how the average expected costs vary across economies. In this case, the economic regions were reduced from the original 7 to 4, mainly due to a visualization problem, as using 7 different colours would have been too much for a single graph.

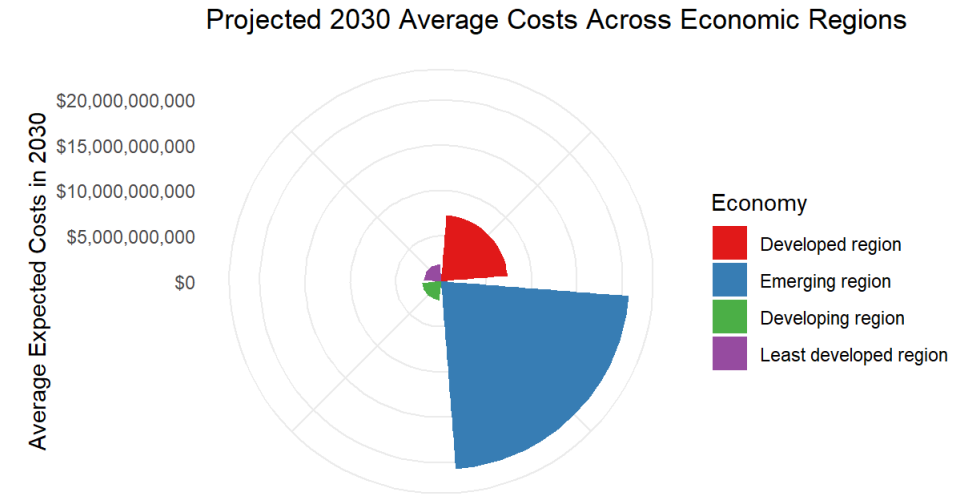
In the upper part of the figure a radar chart shows the cost magnitude for each economic region. It can be observed that the area expected to incur the highest costs is the emerging region. The map below helps to locate the various areas geographically. From the map immediately above on the global overwiew of costs actually countries belonging to the emerging region are confirmed to be in the highest cost range.

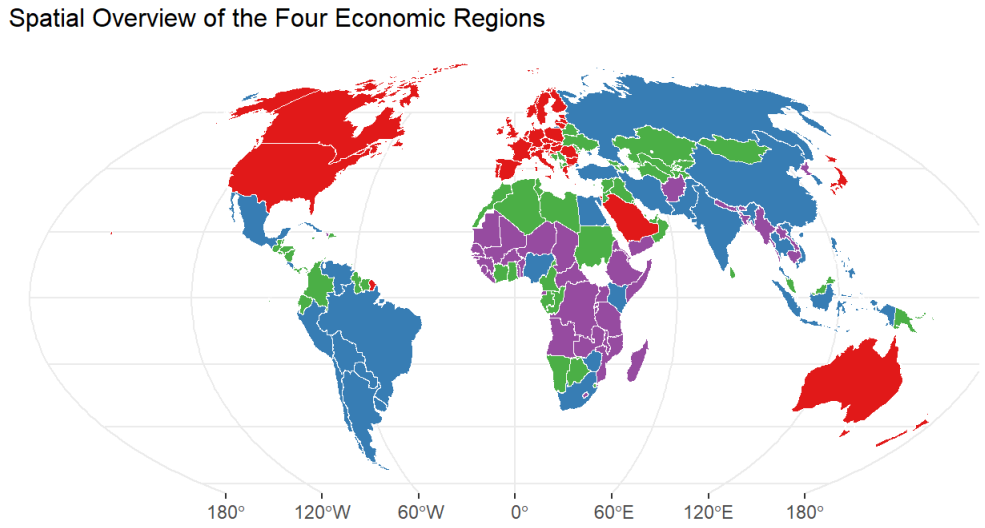
```
groups_economy <- Achieving_Abundance %>%
  mutate(economy = fct_collapse(economy,
    `Least developed region` = "7. Least developed region",
    `Developing region` = "6. Developing region",
    `Emerging region` = c("5. Emerging region: G20",
                          "4. Emerging region: MIKT",
                          "3. Emerging region: BRIC"),
    `Developed region` = c("2. Developed region: nonG7",
                           "1. Developed region: G7")
  ))
```

```
avg_cost_economy <- groups_economy %>%
  st_drop_geometry() %>%
  group_by(economy) %>%
  summarise(avg_cost = mean(Total))

avg_cost_economy
```

```
# A tibble: 4 × 2
  economy                    avg_cost
  <fct>                         <dbl>
1 Developed region        7341534199.
2 Emerging region        20770985019.
3 Developing region       2119775046.
4 Least developed region  1939907332.
```

```r
plot1 <- ggplot(data = avg_cost_economy,
                mapping = aes(x = economy, y = avg_cost, fill = economy)) +
  geom_col() +
  scale_y_continuous(labels = scales::dollar) +
  coord_polar() +
  scale_fill_brewer(palette="Set1", type="qual") +
  labs(title = "Projected 2030 Average Costs Across Economic Regions",
       fill = "Economy",
       y = "Average Expected Costs in 2030") +
  theme_minimal() +
  theme(legend.position = "right",
        axis.text.x = element_blank(),
        axis.title.x = element_blank())

plot2 <- ggplot() +
  geom_sf(data = groups_economy,
          mapping = aes(fill = economy),
          col = "white") +
  scale_fill_brewer(palette="Set1", type="qual") +
  labs(title = "Spatial Overview of the Four Economic Regions") +
  theme_bw() +
  theme(panel.border = element_blank(),
        legend.position = "none",
        title = element_text(size=rel(0.9), angle=0)) +
  coord_sf(crs="+proj=moll")

grid.arrange(plot1, plot2)
```





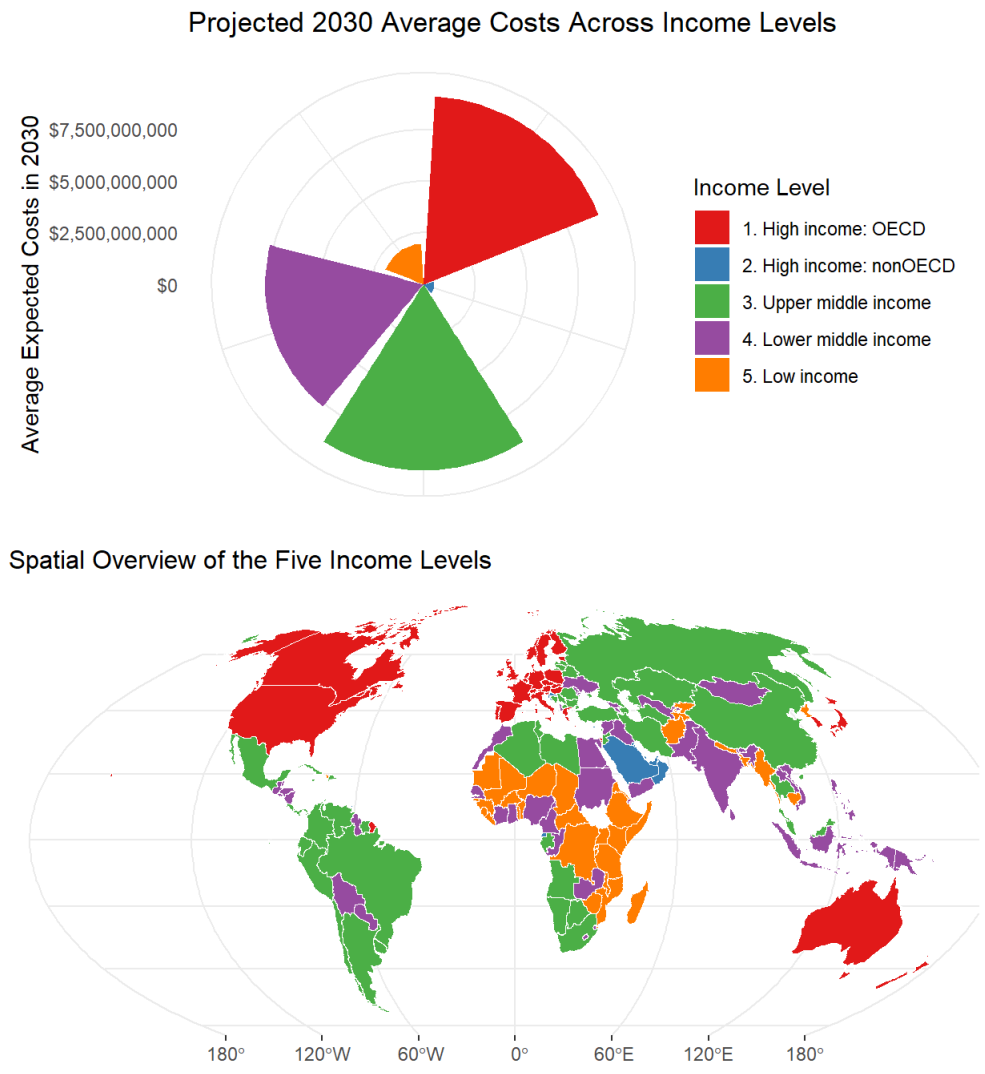
The same kind of reasoning can be made by shifting the focus to income levels. Here it is quite surprising to note that, apart from middle-income regions, a large part of the costs are expected to be borne by high-income countries, specifically those belonging to the OECD.

```
avg_cost_income <- Achieving_Abundance %>%
  st_drop_geometry() %>%
  group_by(income) %>%
  summarise(avg_cost = mean(Total))

avg_cost_income
```

```
# A tibble: 5 × 2
  income                     avg_cost
  <chr>                         <dbl>
1 1. High income: OECD    9123428632.
2 2. High income: nonOECD  549520628.
3 3. Upper middle income  9018783157.
4 4. Lower middle income  7673819407.
5 5. Low income           1970307838.
```

```
plot1 <- ggplot(data = avg_cost_income,
                mapping = aes(x = income, y = avg_cost, fill = income)) +
  geom_col() +
  scale_y_continuous(labels = scales::dollar) +
  coord_polar() +
  scale_fill_brewer(palette="Set1", type="qual") +
  labs(title = "Projected 2030 Average Costs Across Income Levels",
       fill = "Income Level",
       y = "Average Expected Costs in 2030") +
  theme_minimal() +
```

```
  theme(legend.position = "right",
        axis.text.x = element_blank(),
        axis.title.x = element_blank())

plot2 <- ggplot() +
  geom_sf(data = Achieving_Abundance,
          mapping = aes(fill = income),
          col = "white") +
  scale_fill_brewer(palette="Set1", type="qual") +
  labs(title = "Spatial Overview of the Five Income Levels") +
  theme_bw() +
  theme(panel.border = element_blank(),
        legend.position = "none",
        title = element_text(size=rel(0.9), angle=0)) +
  coord_sf(crs="+proj=moll")

grid.arrange(plot1, plot2)
```



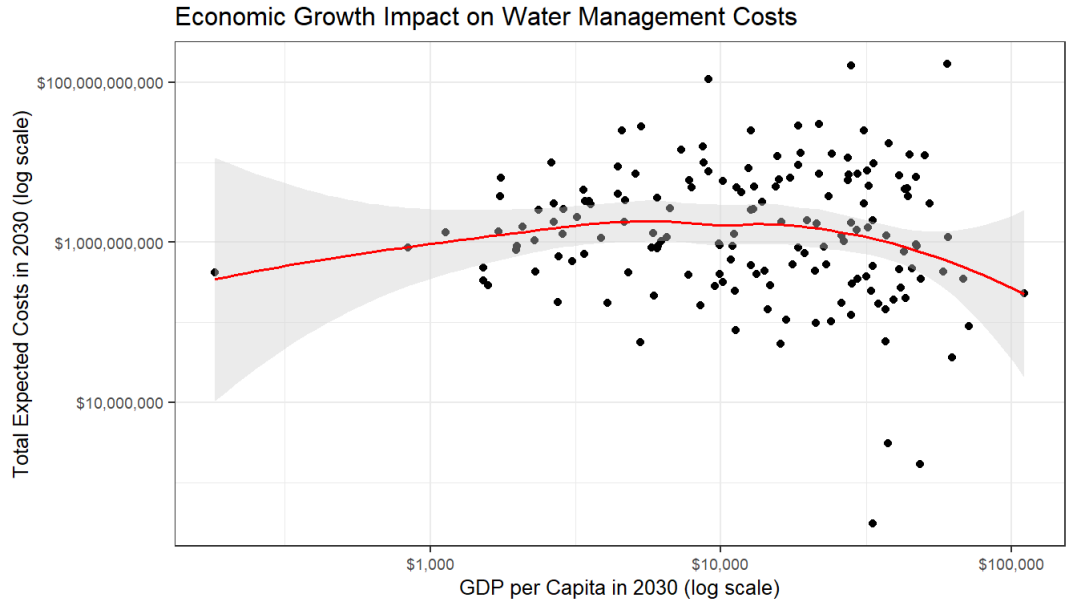
Correlation between GDP per Capita, Average Risk Score, and Projected Total Costs

The interest in discovering some patterns and correlations between the most relevant variables considered so far forms the concluding part of this project. To this end, the most useful type of graph is the scatterplot, which makes it possible to visualize the relationship between two variables, identify patterns, trends or correlations and detect any anomalies or outliers in the data.

In the first scatterplot, GDP per capita is plotted against total expected costs (both referring to forecasts for the year 2030). Analysing the graph, it is observable that, as GDP per capita increases, cost expectations tend to show a slight decrease. This suggests a potential inverse relationship between a country's economic prosperity and expected costs, implying that nations with a

higher GDP per capita may be able to manage future expenditures slightly better, perhaps due to greater economic efficiency or more effective resource management policies. However, it is also important to consider other factors that could influence this trend, such as government policies, technological innovation, and so on.

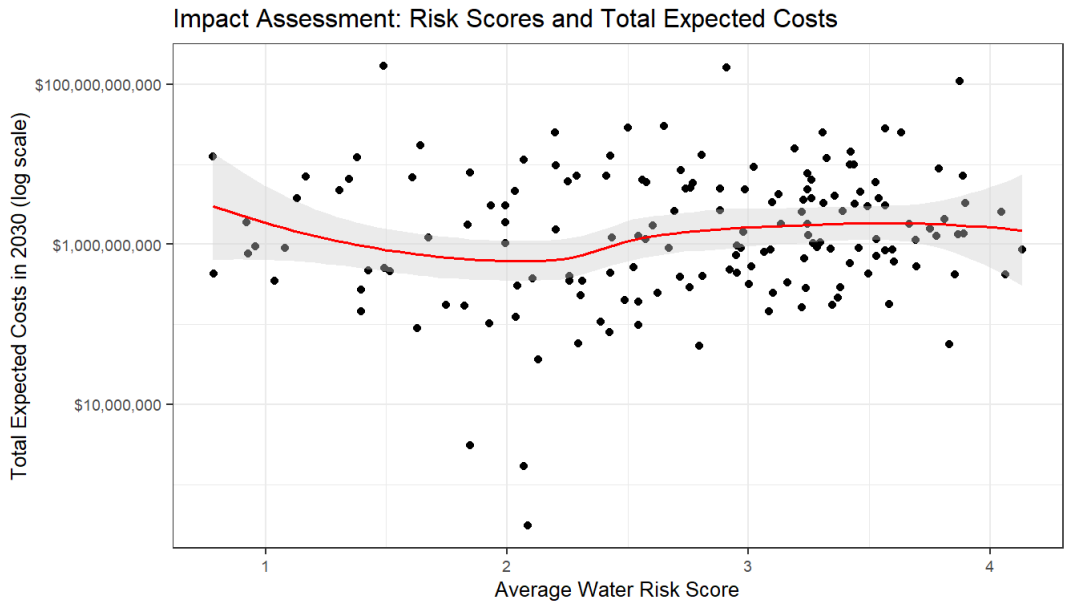
```
ggplot(data = Achieving_Abundance,
       mapping = aes(x = GDPperCap2030, y = Total)) +

  geom_point() +
  geom_smooth(col="red", fill="lightgray", linewidth=0.6) +

  scale_x_log10(labels = scales::dollar) +
  scale_y_log10(labels = scales::dollar) +

  labs(title = "Economic Growth Impact on Water Management Costs",
       x = "GDP per Capita in 2030 (log scale)",
       y = "Total Expected Costs in 2030 (log scale)") +
  theme_bw() +
  theme(axis.text = element_text(size=7),
        title = element_text(size=rel(0.9), angle=0))
```



As a closing remark, it might be relevant to examine whether the expected total costs are correlated with the value of the risk score. Surprisingly, costs seem to remain largely constant even with a low water risk score. This phenomenon may indicate that the costs of sustainable water management remain high even in countries where the risk is considered low. This may be due to a number of factors, including significant investments in infrastructure to prevent future problems, the implementation of advanced water management technologies or stringent policies requiring high standards of sustainability regardless of the current level of risk. This observation emphasizes the importance as well as the need to consider a broader range of variables when assessing water management expenditure, beyond a simple risk score.

```
GroupedWaterRisk %>%
  st_drop_geometry() %>%
  filter(ISO %in% Achieving_Abundance$ISO) %>%
  group_by(ISO) %>%
  summarise(score = mean(score)) %>%

  ggplot(mapping = aes(x = score,
                       y = Achieving_Abundance$Total)) +
  geom_point() +
  geom_smooth(col="red", fill="lightgray", linewidth=0.6) +

  scale_y_log10(labels = scales::dollar) +

  labs(title = "Impact Assessment: Risk Scores and Total Expected Costs",
       x = "Average Water Risk Score",
       y = "Total Expected Costs in 2030 (log scale)") +
```

```
        theme_bw() +
        theme(axis.text = element_text(size=7),
              title = element_text(size=rel(0.9), angle=0))
```



Conclusions

This project highlighted the crucial importance of making complex hydrological data more accessible and understandable. Through the application of visualization techniques, complex information could be translated and subsequently visualized into clear and intuitive risk indicators, facilitating understanding and informed action. In a context of increasing awareness of water-related risks, evidence suggested that the level of risk varies significantly between regions, with more developed areas generally having a lower risk than developing regions. But not only that, there are also differences between econ

However, no region is completely immune to water challenges, which underlines the importance of addressing these problems in a comprehensive and coordinated manner. Furthermore, the second phase of the analysis focused on understanding the costs associated with sustainable water management. It was found that even in regions with lower water risk, significant costs can arise, highlighting the need for a holistic and adaptive approach to addressing water-related challenges.

Ultimately, these findings can provide valuable insights into the measures needed to ensure a sustainable future for water resources globally. They underscore the urgency of targeted policies and investments that take into account regional specificities, as the world prepares to tackle the increasingly pressing challenges that climate change and global development pose to water supply and water security.

References

- Kuzma, S., M.F.P. Bierkens, S. Lakshman, T. Luo, L. Saccoccia, E. H. Sutanudjaja, and R. Van Beek. 2023. "Aqueduct 4.0: Updated decision-relevant global water risk indicators." Technical Note. Washington, DC: World Resources Institute. Available online at: doi.org/10.46830/writn.23.00061.
- Lehner, B., and G. Grill. 2013. "Global River Hydrography and Network Routing: Baseline Data and New Approaches to Study the World's Large River Systems." Hydrological Processes 27 (15): 2171–86.
- World Resources Institute. 2023. "Achieving Abundance: Understanding the Cost of a Sustainable Water Future Data." Available online at: https://www.wri.org/data/achieving-abundance-understanding-cost-sustainable-water-future-data
- World Resources Institute. 2023. "Aqueduct 4.0 Current and Future Global Maps Data." Available online at: https://www.wri.org/data/aqueduct-global-maps-40-data#download-form