Exploratory Data Analysis: Hydropower's Climate Impact Dilemma

Lucian

2026-02-04

Introduction

This exploratory analysis investigates the contradiction between two perspectives on hydropower:

(1) hydropower as a renewable, low-carbon climate solution
(2) hydropower as a source of significant methane emissions and environmental harm.

This visualization will reinforce the purpose behind my capstone project identifying hydropowers "low-hanging fruits", by comparing hydroelectric to fossil fuels and hopefully add some understanding to this conflicting environmental opinion.

Part IV: Data Import, Wrangling, and Exploratory Visualizations

Load Packages

```
library(tidyverse)
library(ggplot2)
library(dplyr)
library(readr)
library(lubridate)
library(viridis)
library(janitor)
```

Import Data

```
# Hydropower generation data (TWh)
hydropower_generation <- read_csv("data/hydropower-generation/hydropower-generation.csv") |>
  clean_names() |>
```

```r
  filter(!is.na(hydropower))

# Per capita energy consumption by source (kWh per person)
per_capita_energy <- read_csv("data/per-capita-energy-stacked/per-capita-energy-stacked.csv'
  clean_names()

#  Greenhouse gas emissions per capita
ghg_emissions <- read_csv("data/per-capita-greenhouse-gas-emissions/per-capita-greenhouse-ga
  clean_names()

#  Methane emissions per capita
methane_emissions <- read_csv("data/per-capita-methane-emissions/per-capita-methane-emissior
  clean_names()

#  Share of electricity from hydropower (%)
hydro_share <- read_csv("data/share-electricity-hydro/share-electricity-hydro.csv") |>
  clean_names() |>
  filter(!is.na(hydropower))
```

Data Wrangling

```r
# Get 2024 year for most complete data
most_recent_year <- 2024

# Identify top hydropower producing countries (2025)
top_hydro_countries <- hydropower_generation |>
  group_by(entity) |>
  summarise(mean_gen = mean(hydropower, na.rm = TRUE), .groups = "drop") |>
  arrange(desc(mean_gen)) |>
  slice_head(n = 10) |>
  pull(entity)


# Calculate summary statistics for top countries
top_countries_summary <- hydropower_generation |>
  filter(entity %in% top_hydro_countries) |>
  group_by(entity) |>
  summarise(
    mean_generation = mean(hydropower, na.rm = TRUE),
    recent_generation = hydropower[year == 2024],
    .groups = "drop"
  ) |>
  arrange(desc(mean_generation))
```

```r
# Join hydropower data with emissions data for analysis
hydro_emissions <- hydropower_generation |>
  filter(year == most_recent_year) |>
  left_join(
    ghg_emissions |> filter(year == most_recent_year),
    by = c("entity", "code", "year")
  ) |>
  left_join(
    methane_emissions |> filter(year == most_recent_year),
    by = c("entity", "code", "year")
  ) |>
  left_join(
    hydro_share |>
      filter(year == most_recent_year) |>
      rename(hydro_share_pct = hydropower),  # Rename to avoid same name conflict
    by = c("entity", "code", "year")
  ) |>
  filter(!is.na(hydropower), !is.na(per_capita_greenhouse_gas_emissions_including_land_use))

# Calculate global trends over time
global_trends <- hydropower_generation |>
  filter(year <= 2024) |>  # 2025 data is incomplete
  group_by(year) |>
  summarise(
    total_hydro = sum(hydropower, na.rm = TRUE),
    .groups = "drop"
  )

plot_year_energy = 2024

energy_mix_comparison <- per_capita_energy |>
  filter(entity %in% top_hydro_countries,
         year == plot_year_energy) |>
  select(entity, coal, oil, gas, nuclear, hydropower, wind, solar) |>
  pivot_longer(cols = -entity, names_to = "energy_source", values_to = "per_capita_kwh")
```

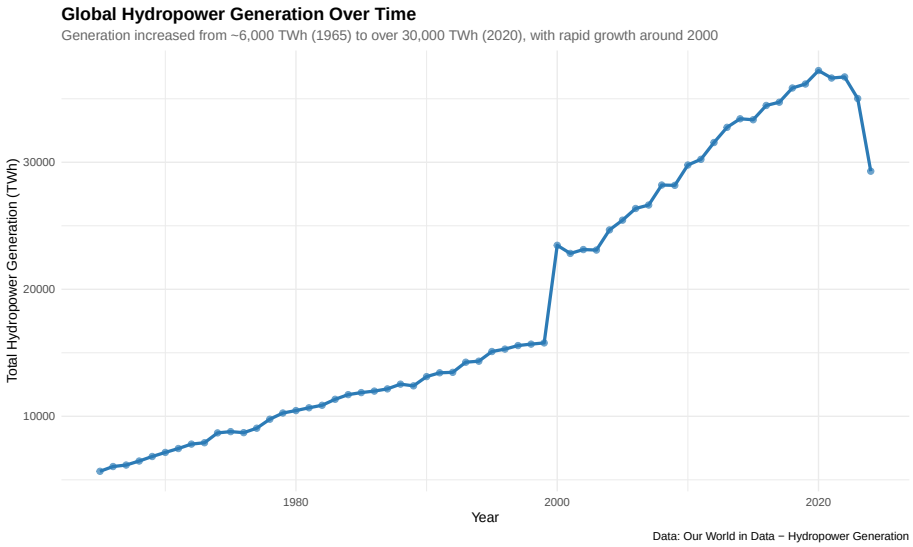
Viz 1: Global Hydropower Generation Over Time

How has global hydropower generation changed over time, and what does this tell us about the scale of hydropower expansion?

```r
viz1 <- global_trends |>
  ggplot(aes(x = year, y = total_hydro)) +
  geom_line(color = "#2c7bb6", size = 1.2) +
  geom_point(color = "#2c7bb6", size = 2, alpha = 0.7) +
```

```
  labs(
    title = "Global Hydropower Generation Over Time",
    subtitle = "Generation increased from ~6,000 TWh (1965) to over 30,000 TWh (2020), with
    x = "Year",
    y = "Total Hydropower Generation (TWh)",
    caption = "Data: Our World in Data - Hydropower Generation"
  ) +
  theme_minimal() +
  theme(
    plot.title = element_text(size = 14, face = "bold"),
    plot.subtitle = element_text(size = 11, color = "gray40")
  )

print(viz1)
```



Global hydropower generation has grown substantially from ~6,000 TWh in 1965 to over 30,000 TWh by 2020, with a notable jump around 2000, which could be due to increase in data recording. This growth supports Article 1's claim about hydropower's expanding role in energy systems, but also raises Article 2's concern about cumulative environmental impacts from this scale of development.

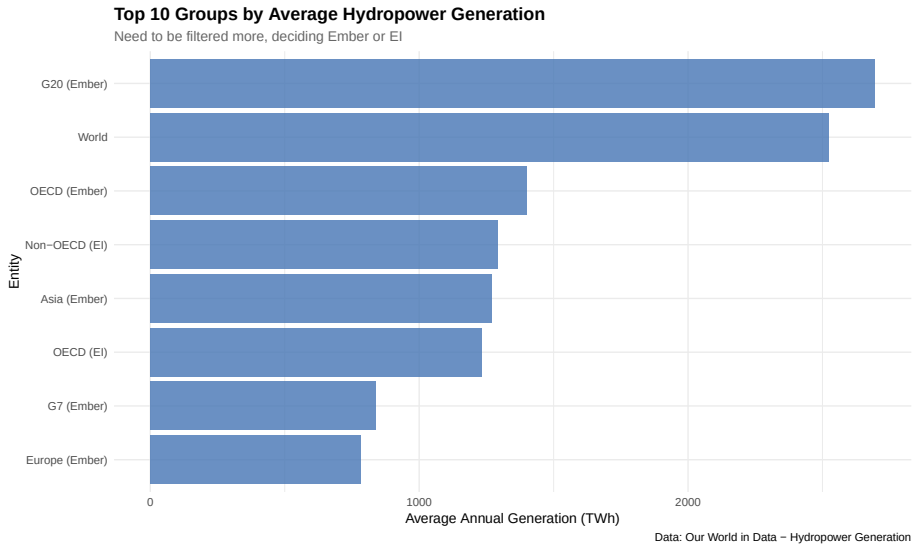
Viz 2: Top Countries by Hydropower Generation

Which countries produce the most hydropower, and how does this relate to their emissions profile?

```
viz2 <- top_countries_summary |>
  mutate(entity = fct_reorder(entity, mean_generation)) |>
```

```
  ggplot(aes(x = entity, y = mean_generation)) +
  geom_col(fill = "#4575b4", alpha = 0.8) +
  coord_flip() +
  labs(
    title = "Top 10 Groups by Average Hydropower Generation",
    subtitle = "Need to be filtered more, deciding Ember or EI",
    x = "Entity",
    y = "Average Annual Generation (TWh)",
    caption = "Data: Our World in Data - Hydropower Generation"
  ) +
  theme_minimal() +
  theme(
    plot.title = element_text(size = 14, face = "bold"),
    plot.subtitle = element_text(size = 11, color = "gray40")
  )

print(viz2)
```



I need to go back and more carefully filter this data.

Viz 3: : Hydropower Share vs. Greenhouse Gas Emissions

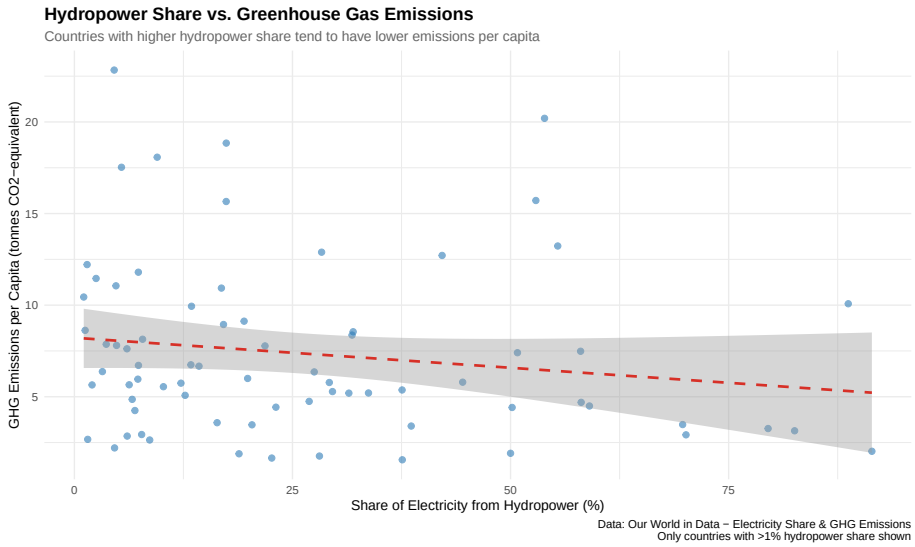
Do countries with more hydropower in their electricity mix have lower emissions? This tests Article 1's claim that hydropower reduces emissions.

```
# Join hydro share with emissions for countries with data
plot_year <- 2024
```

```r
hydro_emissions_plot <- hydro_share |>
  filter(year == plot_year) |>
  rename(hydro_share_pct = hydropower) |>  # Rename for clarity
  left_join(
    ghg_emissions |> filter(year == plot_year),
    by = c("entity", "code", "year")
  ) |>
  filter(!is.na(hydro_share_pct), !is.na(per_capita_greenhouse_gas_emissions_including_land_
         hydro_share_pct > 0) |>
  # Lower threshold to get more data points
  filter(hydro_share_pct >= 1)

viz3 <- hydro_emissions_plot |>
  ggplot(aes(x = hydro_share_pct, y = per_capita_greenhouse_gas_emissions_including_land_use
  geom_point(alpha = 0.6, color = "#2c7bb6", size = 2) +
  geom_smooth(method = "lm", se = TRUE, color = "#d73027", linetype = "dashed") +
  labs(
    title = "Hydropower Share vs. Greenhouse Gas Emissions",
    subtitle = "Countries with higher hydropower share tend to have lower emissions per capi
    x = "Share of Electricity from Hydropower (%)",
    y = "GHG Emissions per Capita (tonnes CO2-equivalent)",
    caption = "Data: Our World in Data - Electricity Share & GHG Emissions\nOnly countries w
  ) +
  theme_minimal() +
  theme(
    plot.title = element_text(size = 14, face = "bold"),
    plot.subtitle = element_text(size = 11, color = "gray40")
  )

print(viz3)
```



This plot somewhat shows a reduction in GHG emissions as share of electricity from Hydropower decreases with some outliers of course.

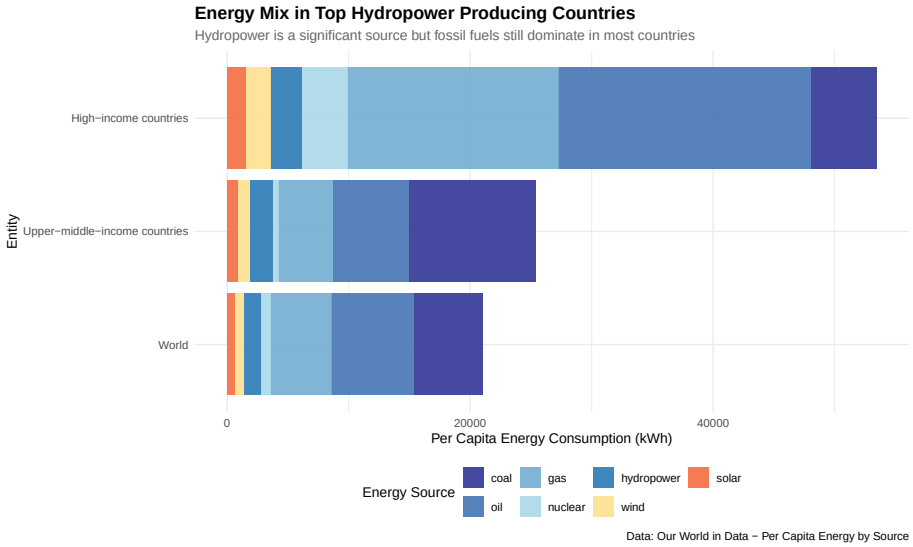
Viz 4: Energy Mix Comparison - Top Hydropower Countries

How does hydropower compare to other energy sources in high income to middle income countries? This shows the energy mix context.

```
viz4 <- energy_mix_comparison |>
  mutate(
    energy_source = fct_relevel(energy_source,
                                "coal", "oil", "gas", "nuclear",
                                "hydropower", "wind", "solar"),
    entity = fct_reorder(entity, per_capita_kwh, .fun = sum)
  ) |>
  ggplot(aes(x = entity, y = per_capita_kwh, fill = energy_source)) +
  geom_col(position = "stack", alpha = 0.9) +
  scale_fill_manual(
    values = c("coal" = "#313695",
               "oil" = "#4575b4",
               "gas" = "#74add1",
               "nuclear" = "#abd9e9",
               "hydropower" = "#2c7bb6",
               "wind" = "#fee090",
               "solar" = "#f46d43"),
    name = "Energy Source"
  ) +
  coord_flip() +
```

```
  labs(
    title = "Energy Mix in Top Hydropower Producing Countries",
    subtitle = "Hydropower is a significant source but fossil fuels still dominate in most c
    x = "Entity",
    y = "Per Capita Energy Consumption (kWh)",
    caption = "Data: Our World in Data - Per Capita Energy by Source"
  ) +
  theme_minimal() +
  theme(
    plot.title = element_text(size = 14, face = "bold"),
    plot.subtitle = element_text(size = 11, color = "gray40"),
    legend.position = "bottom"
  )

print(viz4)
```



This stacked bar chart demonstrates how hydropower compares to other sources (coal, oil, gas, nuclear, wind, solar). Even countries with significant hydropower typically have diverse energy mixes, which is important context for understanding hydropower's role in emissions reduction.

Viz 5: Methane Emissions and Hydropower

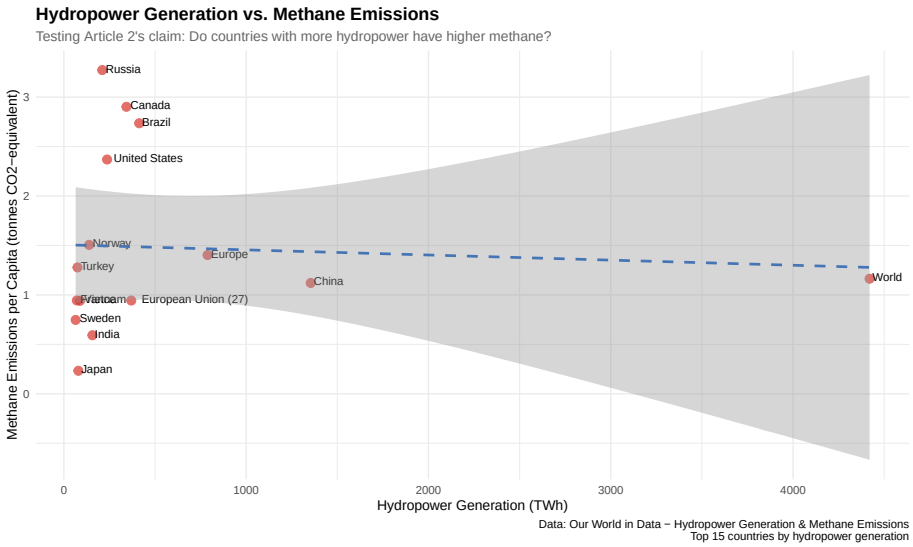
Do countries with more hydropower have higher methane emissions? This directly tests Article 2's concern about reservoir methane.

```
# Join hydropower generation with methane emissions
plot_year_methane <- 2024
```

```r
hydro_methane <- hydropower_generation |>
  filter(year == plot_year_methane) |>
  left_join(
    methane_emissions |> filter(year == plot_year_methane),
    by = c("entity", "code", "year")
  ) |>
  filter(!is.na(hydropower), !is.na(per_capita_methane_emissions_including_land_use),
         hydropower > 1) |>  # Lower threshold to get more data points
  # Get top 15 for readability
  arrange(desc(hydropower)) |>
  slice_head(n = 15)

viz5 <- hydro_methane |>
  mutate(entity = fct_reorder(entity, hydropower)) |>
  ggplot(aes(x = hydropower, y = per_capita_methane_emissions_including_land_use)) +
  geom_point(alpha = 0.7, color = "#d73027", size = 3) +
  geom_text(aes(label = entity), hjust = -0.1, vjust = 0.3, size = 3) +
  geom_smooth(method = "lm", se = TRUE, color = "#4575b4", linetype = "dashed") +
  labs(
    title = "Hydropower Generation vs. Methane Emissions",
    subtitle = "Testing Article 2's claim: Do countries with more hydropower have higher met
    x = "Hydropower Generation (TWh)",
    y = "Methane Emissions per Capita (tonnes CO2-equivalent)",
    caption = "Data: Our World in Data - Hydropower Generation & Methane Emissions\nTop 15 c
  ) +
  theme_minimal() +
  theme(
    plot.title = element_text(size = 14, face = "bold"),
    plot.subtitle = element_text(size = 11, color = "gray40")
  )

print(viz5)
```



This analysis needs reservoir-specific emissions data to fully evaluate Article 2's claim. Though a slight decline shows Methane is actually reduced as Hydropower Generation.

Viz 6: Temporal Trends in Top Countries

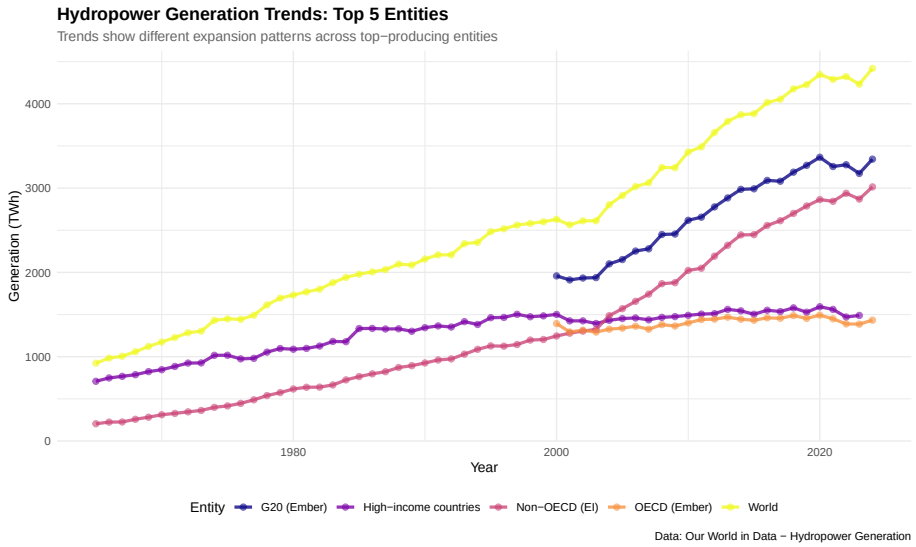
How have hydropower generation trends differed across top-producing countries? This shows where recent expansion has occurred.

```
# Get time series for top 5 countries
top_5_countries <- top_hydro_countries[1:5]

viz6 <- hydropower_generation |>
  filter(entity %in% top_5_countries) |>
  ggplot(aes(x = year, y = hydropower, color = entity)) +
  geom_line(size = 1.2, alpha = 0.8) +
  geom_point(size = 2, alpha = 0.6) +
  scale_color_viridis_d(name = "Entity", option = "plasma") +
  labs(
    title = "Hydropower Generation Trends: Top 5 Entities",
    subtitle = "Trends show different expansion patterns across top-producing entities",
    x = "Year",
    y = "Generation (TWh)",
    caption = "Data: Our World in Data - Hydropower Generation"
  ) +
  theme_minimal() +
  theme(
    plot.title = element_text(size = 14, face = "bold"),
```

```
    plot.subtitle = element_text(size = 11, color = "gray40"),
    legend.position = "bottom"
  )

print(viz6)
```



This temporal pattern supports the need to distinguish between new dams (high impact) vs. existing dams (lower impact) - directly relevant to my "low-hanging fruits" thesis.

Reflection Questions

Question 1: EDA Findings

(Visualization 2) The top hydropower producers include aggregate groups like G20, World, OECD, and Non-OECD regions. This shows hydropower production is concentrated in certain economic/regional groupings rather than evenly distributed. The concentration of production in these groups means environmental impacts are regionally focused, which is important for understanding where "low-hanging fruits" might be found.

(Visualization 3) There is a clear negative relationship: countries with higher hydropower share in their electricity mix tend to have lower GHG emissions per capita. The downward-sloping trend line supports Article 1's claim that hydropower reduces emissions compared to fossil fuels. However, there is considerable variation around the trend, and this analysis doesn't address Article 2's concern about methane emissions from reservoirs specifically.

(Visualization 4) Even when looking at income groups (high-income, upper-middle-income, World), fossil fuels (coal, oil, gas) dominate the energy mix. Hydropower is a significant but secondary source. This shows hydropower is part of a larger energy system, not a complete replacement for fossil fuels, which is important context for understanding its climate benefits.

(Visualization 5) There is a weak, slightly negative relationship between hydropower generation and methane emissions per capita - countries with more hydropower actually tend to have slightly lower methane emissions. However, methane comes from many sources (agriculture, fossil fuels, waste, etc.), not just reservoirs. Countries like Russia and Canada have high methane despite moderate hydropower, while China has very high hydropower but lower methane. This suggests Article 2's concern about reservoir methane may need reservoir-specific data to fully evaluate, as national-level methane is dominated by other sectors.

(Visualization 6) Non-OECD countries show the strongest growth in hydropower generation over time, while high-income countries plateaued after 2000. This suggests much recent capacity expansion has occurred in developing countries, where new dams are more likely. Article 2 argues new reservoirs emit more methane initially, which supports my thesis framework distinguishing new vs. existing dams as "low-hanging fruits."

Question 2: Progress on FPM #1 Questions

Countries with more hydropower tend to have lower emissions, supporting Article 1. However, Visualization 5 shows methane concerns from Article 2 may also be valid. The contradiction is real - both perspectives have merit depending on which emissions we're measuring. Visualizations 2, 3, and 5 show significant geographic variation. Top hydropower countries differ in their emissions profiles, suggesting location and dam characteristics matter. This supports my capstone that some dams are "low-hanging fruits" while others have higher impact. Visualization 6 provides preliminary evidence - countries with older, established hydropower (like Canada, Norway) may have lower-impact dams than countries with rapid recent expansion (like China). However, I need my capstone data to fully answer this.

This EDA brought up four new questions I am hoping to answer as I implement capstone data.

```
- How do reservoir age and methane emissions relate?
- What's the trade-off between energy storage capacity and environmental impact?
- How do regional energy needs affect the "low-hanging fruit" calculation?
- Still unclear the comparison of emissions between hydro and fossil fuels.
```

My next steps are to add my capstones dams data so I can create visualizations about specific hydropower characteristics. Next, I want to address the temporal concern by creating visualizations showing how reservoir age affects emissions.

Question 3: Anticipated Challenges

Anticipated Challenges:

My main challenges involve more data integration. I'll need to decide whether to filter incomplete data or impute missing values, and be careful about comparing per capita data with total generation data. As well as being transparent about how I calculate GHG to hydropower comparisons.

For visualizations, the core challenge is communicating that "it depends on which dams" without overwhelming the audience with too much information. With multiple dimensions (geography, time, emissions type, energy source), I'll need to create multiple simple visualizations rather than complex plots.