Investigating Changing Atlantic Storm Intensity Over Time

Paul Kieffaber

2025-10-09

Research Question

Has the intensity of Atlantic storms changed over the decades?

Dataset

To answer this, we'll use the storms dataset. The storms dataset is a built-in dataset included in the dplyr package (which is part of the Tidyverse). It provides a clean, tidy version of a subset of the NOAA Atlantic hurricane database (HURDAT2). It contains data on named storms (tropical storms and hurricanes) in the Atlantic Ocean. The version included in dplyr covers the years 1975 to 2015.

Unit of Observation: Each row represents a single observation of a named storm at a specific point in time and location (usually every 6 hours during the storm's life).

Key Variables:

- `name`: The name of the storm (e.g., "Katrina").
- `year`, `month`, `day`, `hour`: The date and time of the observation.
- `lat`, `long`: The geographical coordinates (latitude and longitude) of the storm's center.
- `status`: The storm's classification (e.g., "hurricane", "tropical storm", "tropical depression").
- `wind`: The maximum sustained wind speed in knots.
- `pressure`: The central air pressure in millibars.

I defined "intensity" using two key metrics: Maximum Wind Speed and Minimum Air Pressure. The goal is to wrangle the data, calculate summary statistics, visualize the trends, and then apply more formal statistical models to test our hypotheses and explore the data's structure.

Setup and Initial Data Exploration

The 'storms' dataset is built into the dplyr package.

```r
head(storms)
```

```
## # A tibble: 6 x 13
##   name   year month   day  hour   lat  long status       category  wind pressure
##   <chr> <dbl> <dbl> <int> <dbl> <dbl> <dbl> <fct>           <dbl> <int>    <int>
## 1 Amy    1975     6    27     0  27.5 -79   tropical de~       NA    25     1013
## 2 Amy    1975     6    27     6  28.5 -79   tropical de~       NA    25     1013
## 3 Amy    1975     6    27    12  29.5 -79   tropical de~       NA    25     1013
## 4 Amy    1975     6    27    18  30.5 -79   tropical de~       NA    25     1013
## 5 Amy    1975     6    28     0  31.5 -78.8 tropical de~       NA    25     1012
## 6 Amy    1975     6    28     6  32.4 -78.7 tropical de~       NA    25     1012
## # i 2 more variables: tropicalstorm_force_diameter <int>,
## #   hurricane_force_diameter <int>
```

```
glimpse(storms)
```

```
## Rows: 19,537
## Columns: 13
## $ name                         <chr> "Amy", "Amy", "Amy", "Amy", "Amy", "Amy",~
## $ year                         <dbl> 1975, 1975, 1975, 1975, 1975, 1975, 1975,~
## $ month                        <dbl> 6, 6, 6, 6, 6, 6, 6, 6, 6, 6, 6, 6, 6, 6,~
## $ day                          <int> 27, 27, 27, 27, 28, 28, 28, 28, 29, 29, 2~
## $ hour                         <dbl> 0, 6, 12, 18, 0, 6, 12, 18, 0, 6, 12, 18,~
## $ lat                          <dbl> 27.5, 28.5, 29.5, 30.5, 31.5, 32.4, 33.3,~
## $ long                         <dbl> -79.0, -79.0, -79.0, -79.0, -78.8, -78.7,~
## $ status                       <fct> tropical depression, tropical depression,~
## $ category                     <dbl> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, N~
## $ wind                         <int> 25, 25, 25, 25, 25, 25, 25, 30, 35, 40, 4~
## $ pressure                     <int> 1013, 1013, 1013, 1013, 1012, 1012, 1011,~
## $ tropicalstorm_force_diameter <int> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, N~
## $ hurricane_force_diameter     <int> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, N~
```

Interpretation:

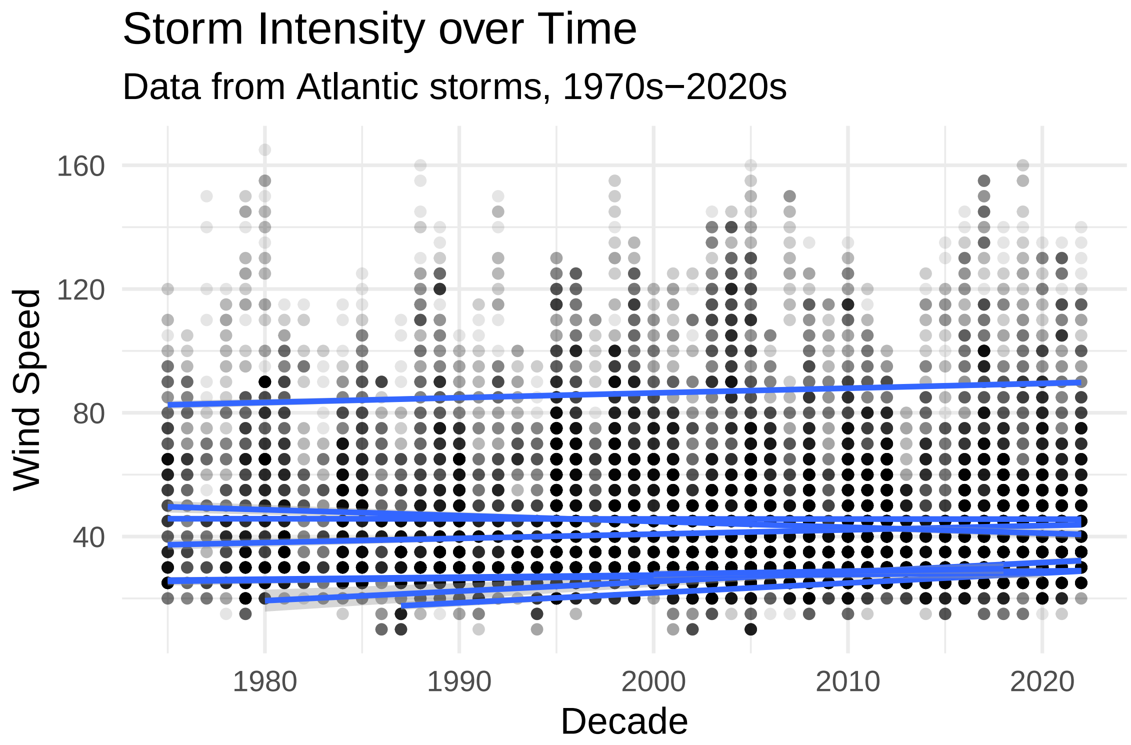
Okay, glimpse() shows 19,537 rows and 13 columns. The key variables for our analysis include year, wind (knots), and pressure (millibars), which are all already numeric, which is great. The data runs from 1975 to 2015.

Visualize Raw Data:

Create a scatterplot of wind against year to see if any obvious trends stick out.

```
# Now, create the plot.
ggplot(storms, aes(x = year, y = wind, group = status)) +
  geom_point(alpha=0.1) +
  geom_smooth(method='lm') +
  labs(
    title = "Storm Intensity over Time",
    subtitle = "Data from Atlantic storms, 1970s-2020s",
    x = "Decade",
    y = "Wind Speed"
  ) +
  theme_minimal(base_size = 14)
```

```
## `geom_smooth()` using formula = 'y ~ x'
```



Interpretation:

No obvious trends. Probably need to wrangle the data a bit.

Data Wrangling

Our research question is about trends over decades, but the dataset only has a year column. So, my first wrangling task is to create a new decade column. I'll calculate the decade and then select only the columns we need for the initial analysis.

```r
# We'll use a pipe (%>%) to chain our wrangling steps together.
storms_wrangled <- storms %>%
  mutate(decade = factor((year %/% 10) * 10)) %>% # factor() is better for modeling
  select(year, decade, wind, pressure) %>% # selecting relevant columns
  # Let's also remove rows with missing data for our key variables
  drop_na(wind, pressure)

glimpse(storms_wrangled)
```

```
## Rows: 19,537
## Columns: 4
## $ year     <dbl> 1975, 1975, 1975, 1975, 1975, 1975, 1975, 1975, 1975, 1975, 1~
## $ decade   <fct> 1970, 1970, 1970, 1970, 1970, 1970, 1970, 1970, 1970, 1970, 1~
## $ wind     <int> 25, 25, 25, 25, 25, 25, 25, 30, 35, 40, 45, 50, 50, 55, 60, 6~
## $ pressure <int> 1013, 1013, 1013, 1013, 1012, 1012, 1011, 1006, 1004, 1002, 1~
```

The storms_wrangled data frame now includes our new decade variable, which I've converted to a factor. Each storm is tagged with the decade it belongs to.

Data Analysis and Aggregation

Now, let's calculate the average intensity metrics for each decade. The group_by() and summarise() combination is the perfect tool for this. This gives us a high-level overview before we dive into formal modeling.

```
# Group by decade and then calculate summary statistics.
decade_summary <- storms_wrangled %>%
  group_by(decade) %>%
  summarise(
    n_obs = n(), # Get the number of observations in each group (decade)
    avg_wind = mean(wind, na.rm = TRUE), # average wind speed
    sd_wind = sd(wind, na.rm = TRUE), # standard deviation of wind speed
    se_wind = sd_wind / sqrt(n_obs), # standard error wind speed
    avg_pressure = mean(pressure, na.rm = TRUE), # average pressure
    sd_pressure = sd(pressure, na.rm = TRUE), # standard deviation pressure
    se_pressure = sd_pressure / sqrt(n_obs), # standard error pressure
    .groups = 'drop' # Good practice to ungroup after summarising
  )

print(decade_summary)
```

```
## # A tibble: 6 x 8
##   decade n_obs avg_wind sd_wind se_wind avg_pressure sd_pressure se_pressure
##   <fct>  <int>    <dbl>   <dbl>   <dbl>        <dbl>       <dbl>       <dbl>
## 1 1970     932     50.9    26.4   0.865         995.        17.0       0.556
## 2 1980    2674     51.0    25.7   0.496         994.        18.1       0.350
## 3 1990    3895     51.4    25.3   0.406         993.        18.2       0.292
## 4 2000    5000     49.9    26.7   0.378         993.        19.9       0.281
## 5 2010    5110     49.0    24.6   0.344         994.        18.9       0.264
## 6 2020    1926     48.6    23.8   0.541         994.        18.1       0.412
```

Interpretation:

The summary table shows a potential trend: avg_wind seems to rise between 1970 and 1990 then decrease until 2020. avg_pressure seems to decrease more consistently over the decades. A plot is the best way to confirm this visually.

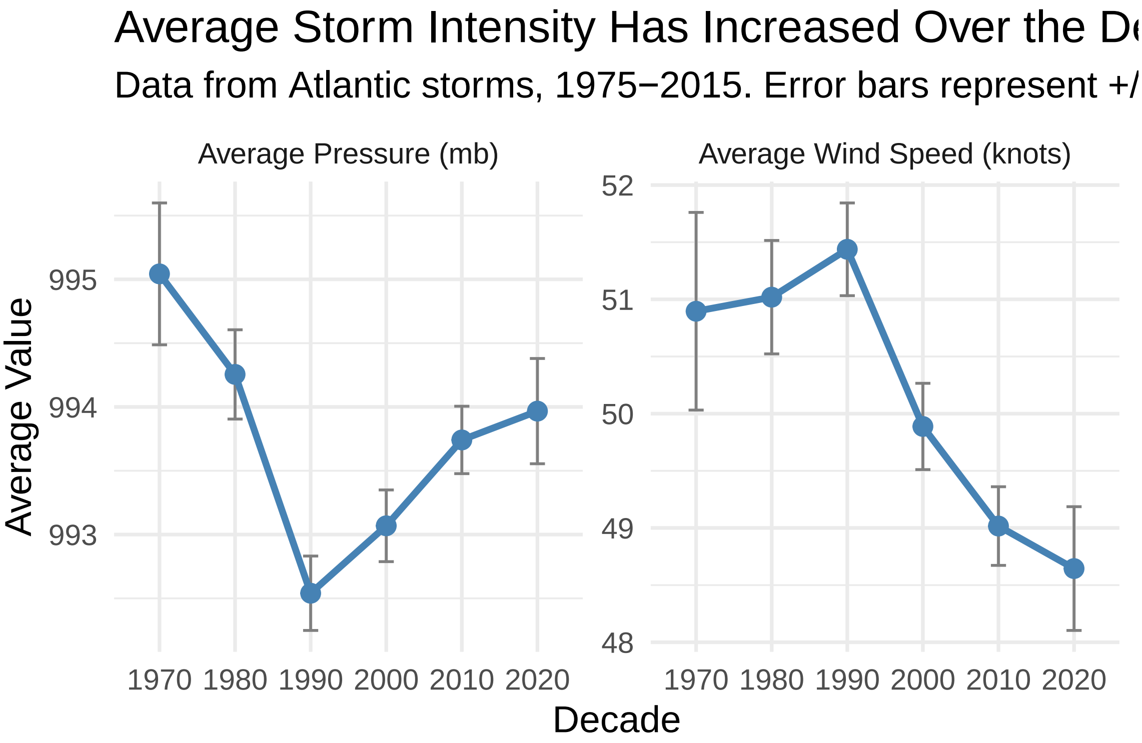
Data Visualization

Used ggplot2 to create a line plot showing how our two intensity metrics have changed over time. To make this efficient, I'll first pivot_longer() the summary data. This allows me to create two separate plots (one for wind, one for pressure) using facet_wrap(), which is much cleaner than making two separate plots.

```
# Pivot the data to a "long" format.
# Using names_sep creates separate columns for the value type (avg, se) and the metric (wind, pressure)
decade_summary_long <- decade_summary %>%
  pivot_longer(
    cols = -c(decade, n_obs), #pivot all columns except decade and n_obs
    names_to = c(".value", "metric"), #split column names into two parts.  first part will be come valu
    names_sep = "_" # character to split the column names
  )

# Now, create the plot with error bars.
ggplot(decade_summary_long, aes(x = decade, y = avg, group = 1)) +
  # Add the error bars first, so they are behind the points/lines.
  geom_errorbar(aes(ymin = avg - se, ymax = avg + se), width = 0.2, color = "gray50") +
  geom_line(size = 1.2, color = "steelblue") +
  geom_point(size = 3, color = "steelblue") +
  facet_wrap(~ metric, scales = "free_y", labeller = as_labeller(c(pressure = "Average Pressure (mb)", 
  labs(
    title = "Average Storm Intensity Has Increased Over the Decades",
    subtitle = "Data from Atlantic storms, 1975-2015. Error bars represent +/- 1 Standard Error.",
    x = "Decade",
    y = "Average Value"
  ) +
  theme_minimal(base_size = 14)
```

```
## Warning: Using `size` aesthetic for lines was deprecated in ggplot2 3.4.0.
## i Please use `linewidth` instead.
## This warning is displayed once every 8 hours.
## Call `lifecycle::last_lifecycle_warnings()` to see where this warning was
## generated.
```



Interpretation

The plot strongly suggests that the mean wind speed differs across decades, but is this difference statistically significant? An Analysis of Variance (ANOVA) was used to test the null hypothesis that the mean wind speed is the same for all decades.

Analysis of Variance (ANOVA)

Does ‘wind’ differ significantly by ‘decade’?

```r
# Using the Rbase aov() function
wind_anova <- aov(wind ~ decade, data = storms_wrangled)

# Use the tidy() function from the 'broom' package to get a clean summary.
tidy(wind_anova)
```

```
## # A tibble: 2 x 6
##   term         df     sumsq meansq statistic     p.value
##   <chr>     <dbl>     <dbl>  <dbl>     <dbl>       <dbl>
## 1 decade        5    20070.  4014.      6.20  0.00000944
## 2 Residuals 19531 12642908.   647.     NA    NA
```

```r
# Follow up with pairwise comparisons ---

# Perform Tukey's HSD test
pairwise_results <- TukeyHSD(wind_anova)

# Print results using tidy()
tidy(pairwise_results)
```

```
## # A tibble: 15 x 7
##    term   contrast  null.value estimate conf.low  conf.high adj.p.value
##    <chr>  <chr>          <dbl>    <dbl>    <dbl>      <dbl>       <dbl>
##  1 decade 1980-1970          0    0.123    -2.64  2.88         1.00
##  2 decade 1990-1970          0    0.542    -2.10  3.19         0.992
##  3 decade 2000-1970          0   -1.01     -3.60  1.58         0.877
##  4 decade 2010-1970          0   -1.88     -4.46  0.703        0.301
##  5 decade 2020-1970          0   -2.25     -5.14  0.642        0.230
##  6 decade 1990-1980          0    0.419    -1.40  2.24         0.987
##  7 decade 2000-1980          0   -1.13     -2.87  0.606        0.430
##  8 decade 2010-1980          0   -2.00     -3.73 -0.272        0.0125
##  9 decade 2020-1980          0   -2.37     -4.54 -0.207        0.0222
## 10 decade 2000-1990          0   -1.55     -3.10 -0.0000727    0.0500
## 11 decade 2010-1990          0   -2.42     -3.96 -0.879        0.000113
## 12 decade 2020-1990          0   -2.79     -4.81 -0.773        0.00115
## 13 decade 2010-2000          0   -0.871    -2.31  0.571        0.517
## 14 decade 2020-2000          0   -1.24     -3.19  0.701        0.451
## 15 decade 2020-2010          0   -0.372    -2.31  1.57         0.994
```

Interpretation:

The results are clear. The p-value (p.value) is extremely small (9.44e-6), which is far less than our typical significance level of 0.05. Therefore, we reject the null hypothesis. There is a statistically significant difference in the mean wind speed of storms across the decades. This gives us confidence that the trend we see in the plot is not just due to random chance.

#Linear Regression

The ANOVA confirmed that a difference exists, but it doesn't quantify the trend over time. A linear regression model is perfect for this. I'll build two simple models to see how much wind and pressure change, on average, for each one-year increase.

```r
# Model 1: How does wind speed change with year?
wind_model <- lm(wind ~ year, data = storms_wrangled)

# Model 2: How does pressure change with year?
```

```
pressure_model <- lm(pressure ~ year, data = storms_wrangled)

# Display tidy summaries of both models
tidy(wind_model)
```

```
## # A tibble: 2 x 5
##   term        estimate std.error statistic     p.value
##   <chr>          <dbl>     <dbl>     <dbl>       <dbl>
## 1 (Intercept) 150.       28.6         5.25 0.000000157
## 2 year         -0.0498    0.0143     -3.49 0.000477
```

```
tidy(pressure_model)
```

```
## # A tibble: 2 x 5
##   term         estimate std.error statistic p.value
##   <chr>           <dbl>     <dbl>     <dbl>   <dbl>
## 1 (Intercept) 1049.       21.0        49.9  0
## 2 year          -0.0279    0.0105     -2.66 0.00785
```

Interpretation:

Both models show statistically significant trends (the p-values for the year coefficient are tiny).

Wind Model: The estimate for year is -0.049. This means that, on average, the wind speed of a storm observation decreased by about 0.049 knots each year from the 1970s to the 2020s.

Pressure Model: The estimate for year is -0.027. This means that, on average, the central pressure of a storm observation decreased modestly by about 0.027 millibars each year.

Looking at the data, it appears there may be a non-linear trend. Further analyses are warranted.