Final Project - Wood Thrush Response

# Climate Change and Avian Response: Assessing the Effects of CO₂ and Temperature on Wood Thrush Populations

## Abstract -

As climate change continues to drive research goals, different areas of impact arise surrounding not only humans but environmental and wildlife harm. Carbon dioxide is one of the main contributors to climate change, directly impacting annual temperatures and disturbing typical weather patterns. Wildlife, specifically avian species, are queue dependent on temperatures for specific phenological stages like migration (Lang et al. (2024)). In recent decades, there has been an observed decline to several avian populations, leading both avian and climate scientists to answer if climate change has disturbed the migration periods, which can directly impact the effectiveness of their breeding cycles. By analyzing annual temperatures and carbon emissions, this study aims to identify a correlation between changing variables and observed migration dates.

## Part 1. Introduction -

**Problem Overview**: In recent decades, global climate change, driven by increased carbon dioxide emissions, has contributed to rising temperatures and disrupted ecological systems (NOAA National Centers for Environmental Information (2024)). Many migratory bird species, such as the subject of this study, the Wood Thrush (Hylocichla mustelina), are particularly vulnerable to these changes due to their reliance on specific environmental cues for migration, breeding, and survival. The Wood Thrush is included on the Yellow Watch List for birds most at risk of extinction, and it is one of the most prominent examples of declining forest songbirds in North America (Ornithology (2024)). Populations have declined by 1.3% per year, resulting in a cumulative decline of approximately 50% between 1966 and 2019 (Ornithology (2024)). Declines in Wood Thrush populations have been observed across North America, yet the extent to which atmospheric CO2 levels and associated temperature changes contribute to these declines remains unclear.

knitr::include\_graphics("~/Github/ESS-330-Final-Project-REDO/images/Wood\_Thrush\_Photo.jpg")

|  |
| --- |
| Figure 1: Wood Thrush - Lance Leonhardt (Audubon Vermont) |

Source: [Article Notebook](https://andiehal.github.io/ESS-330-Final-Project-REDO/index.qmd.html)

## Part 2. Materials and Methods -

### I. Research Goal and Data Sources

**This research will investigate the effects of atmospheric CO₂ concentrations and temperature changes on the population dynamics of the Wood Thrush (Hylocichla mustelina)**, a migratory songbird of conservation concern. The study integrates climate and ecological data from multiple large-scale sources: temperature records from the National Weather Service, (NOAA National Centers for Environmental Information (2024)) bird abundance and distribution data from Cornell’s eBird Dataset (Ornithology (2024)), and atmospheric CO₂ data from the Global Carbon Budget (Friedlingstein et al. (2023)). Using these data sources, we can create a model to help determine the effect of these variables on the arrival and departure dates of these species along the mid-northeastern coast of the United States from 1966 to 2019.

### II. Methods Overview

1. **Data acquisition**
   * Climate Data: National Weather Service
   * CO₂ Data: Global Carbon Budget
   * Bird Observation Data: Cornell Lab’s eBird dataset
2. **Data Cleaning and Processing**
   * **Climate Data -** Imported 7 different datasets from the states above that recorded annual temperatures for each month between 1966 and 2019 and combined into one large dataset, “All\_Weather\_Data.csv”
     + **Difficulties**: The data was extremely time-consuming to clean and process into a large dataset due to the website’s pdf format for data exports. Using Excel’s data from the file helped to import the data and allowed for processing and cleaning of the data before being inserted into the rProject data directory.
   * **CO2 Data -** Downloaded from the Global Carbon Budget and then uploaded to Rstudio. No additional cleaning was needed.
     + **Difficulties**: This data set only includes annual emission while other data sets include monthly data, but was the most accessible to our report.
   * **Wood Thrush Data -** Filtered Wood Thrush eBird data to between the years of 1966 and 2019, along with filtering the location of the bird to these specific states: Pennsylvania, West Virginia, Virginia, Kentucky, Ohio, Indiana, and Tennessee. “Wood\_Thrush\_Raw.csv”
     + **Difficulties**: The dataset is about 131.5 MB and cannot be uploaded to the git repository, meaning data sharing had to be done via email rather than sharing the actual repo for us to pull and push.
3. **Statistical Analysis**

* **Preliminary Analysis - Linear Regression Model**
  + To get a basic understand of the correlation relationship between temperature, CO2, and dates of observation, we created a linear model to visualize and create statistical information.
* **Main Analysis - Machine Learning Prediction Models**
  + Using the multiple datasets to account for temperature, CO2, and observations, we created multiple machine learning models: Linear Regression, Random Forest, and XGBoost. To identify the model of best fit, we created a workflow that would allow us to collect the metrics and determine the model we should choose based on the R-squared and RMSE values.
  + Using two separate recipes to account for both the integration of temperature on observation dates, along with CO2 emissions, to ensure that they would not interfere with each other, and give us much more data to analyze. After prepping and baking the data, we created workflows to run the models. In total, we tested about 6 models, three for each recipe, and deciphered that Random Forest gives us the best predictions for both predictor variables.
  + Once models were defined with both models and fitted to the resamples of the CV fold, we ran the models to determine their statistical significance and to see if the outcome of the model would give us data that would either prove or disprove our hypothesis. We then, with these models we create visualizations to represent and understand the data much better.

1. **Visualization**
   * Creating plots for the relationship between observation date and CO2, along with a separate relationship between observation date and temperature. In total, there will be four main graphs to understand their impact on both Spring Arrival dates and Fall Departure dates.

## Part 3. Preliminary Analysis - Exploring the Data

# Library Codes used  
library(tidyverse)

── Attaching core tidyverse packages ──────────────────────── tidyverse 2.0.0 ──  
✔ dplyr 1.1.4 ✔ readr 2.1.5  
✔ forcats 1.0.0 ✔ stringr 1.5.1  
✔ ggplot2 3.5.2 ✔ tibble 3.2.1  
✔ lubridate 1.9.4 ✔ tidyr 1.3.1  
✔ purrr 1.0.4   
── Conflicts ────────────────────────────────────────── tidyverse\_conflicts() ──  
✖ dplyr::filter() masks stats::filter()  
✖ dplyr::lag() masks stats::lag()  
ℹ Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors

library(tidymodels)

── Attaching packages ────────────────────────────────────── tidymodels 1.3.0 ──  
✔ broom 1.0.8 ✔ rsample 1.3.0  
✔ dials 1.4.0 ✔ tune 1.3.0  
✔ infer 1.0.8 ✔ workflows 1.2.0  
✔ modeldata 1.4.0 ✔ workflowsets 1.1.0  
✔ parsnip 1.3.1 ✔ yardstick 1.3.2  
✔ recipes 1.3.0   
── Conflicts ───────────────────────────────────────── tidymodels\_conflicts() ──  
✖ scales::discard() masks purrr::discard()  
✖ dplyr::filter() masks stats::filter()  
✖ recipes::fixed() masks stringr::fixed()  
✖ dplyr::lag() masks stats::lag()  
✖ yardstick::spec() masks readr::spec()  
✖ recipes::step() masks stats::step()

library(here)

here() starts at /Users/andiehall/Github/ESS-330-Final-Project-REDO

library(readr)  
library(lubridate)  
library(zoo)

Attaching package: 'zoo'  
  
The following objects are masked from 'package:base':  
  
 as.Date, as.Date.numeric

Source: [Article Notebook](https://andiehal.github.io/ESS-330-Final-Project-REDO/index.qmd.html)

### I. Reading in Data Bases

# CO2 Emissions  
Annua\_CO2\_Emissions <- read\_csv("data/annual-co2-emissions-per-country.csv")

Rows: 54 Columns: 4  
── Column specification ────────────────────────────────────────────────────────  
Delimiter: ","  
chr (1): Entity  
dbl (2): Year, Annual CO₂ emissions  
lgl (1): Code  
  
ℹ Use `spec()` to retrieve the full column specification for this data.  
ℹ Specify the column types or set `show\_col\_types = FALSE` to quiet this message.

# Temperature Data  
AllWeatherData <- read\_csv("data/AllWeatherData.csv")

Rows: 378 Columns: 15  
── Column specification ────────────────────────────────────────────────────────  
Delimiter: ","  
chr (1): STATE  
dbl (14): Year, Jan, Feb, Mar, Apr, May, Jun, Jul, Aug, Sep, Oct, Nov, Dec, ...  
  
ℹ Use `spec()` to retrieve the full column specification for this data.  
ℹ Specify the column types or set `show\_col\_types = FALSE` to quiet this message.

# Wood Thrush Data  
Wood\_Thrush <- read\_csv("data/Wood\_Thrush.csv")

New names:  
Rows: 301512 Columns: 27  
── Column specification  
──────────────────────────────────────────────────────── Delimiter: "," chr  
(6): COUNTRY, COUNTRY CODE, STATE, COUNTY, OBSERVATION COUNT, OBSERVATI... dbl  
(2): LATITUDE, LONGITUDE lgl (19): ...6, ...10, ...11, ...12, ...13, ...14,  
...15, ...16, ...17, ...1...  
ℹ Use `spec()` to retrieve the full column specification for this data. ℹ  
Specify the column types or set `show\_col\_types = FALSE` to quiet this message.  
• `` -> `...6`  
• `` -> `...10`  
• `` -> `...11`  
• `` -> `...12`  
• `` -> `...13`  
• `` -> `...14`  
• `` -> `...15`  
• `` -> `...16`  
• `` -> `...17`  
• `` -> `...18`  
• `` -> `...19`  
• `` -> `...20`  
• `` -> `...21`  
• `` -> `...22`  
• `` -> `...23`  
• `` -> `...24`  
• `` -> `...25`  
• `` -> `...26`  
• `` -> `...27`

Source: [Article Notebook](https://andiehal.github.io/ESS-330-Final-Project-REDO/index.qmd.html)

### II. Prepping/Cleaning Data Sets

# Formatting Wood Thrush  
Wood\_Thrush <- Wood\_Thrush |>  
 select(COUNTRY, `COUNTRY CODE`, STATE, `OBSERVATION COUNT`, `OBSERVATION DATE`, LATITUDE, LONGITUDE) |>  
 mutate(`OBSERVATION DATE` = as.Date(`OBSERVATION DATE`, format = "%m/%d/%Y"),  
 Year = year(`OBSERVATION DATE`))  
  
# Formatting Weather  
Annual\_Weather <- AllWeatherData |>  
 select(Year, Annual, STATE)  
  
# Formatting CO2  
Annua\_CO2\_Emissions <- Annua\_CO2\_Emissions |>  
 select(Year, `Annual CO₂ emissions`)  
  
# Joined Data  
Main\_Data <- Wood\_Thrush |>  
 inner\_join(Annual\_Weather, by = c("Year", "STATE")) |>  
 inner\_join(Annua\_CO2\_Emissions, by = "Year") |>  
 mutate(DayOfYear = yday(`OBSERVATION DATE`)) |>  
 drop\_na(Annual, `Annual CO₂ emissions`, DayOfYear)  
  
# Creating Springtime Dataset  
Spring\_Arrival <- Main\_Data |>  
 group\_by(year(`OBSERVATION DATE`)) |>  
 slice\_min(`OBSERVATION DATE`, with\_ties = FALSE)  
  
# Creating Falltime Dataset  
Fall\_Departure <- Main\_Data |>  
 group\_by(year(`OBSERVATION DATE`)) |>  
 slice\_max(`OBSERVATION DATE`, with\_ties = FALSE)

Source: [Article Notebook](https://andiehal.github.io/ESS-330-Final-Project-REDO/index.qmd.html)

### III. Creating Linear Model for Observation Dates

#### Temperature Models -

# Temperature Model - Spring  
Temp\_S\_Main\_Model <- lm(formula = DayOfYear ~ Annual, data = Spring\_Arrival)  
  
summary(Temp\_S\_Main\_Model)

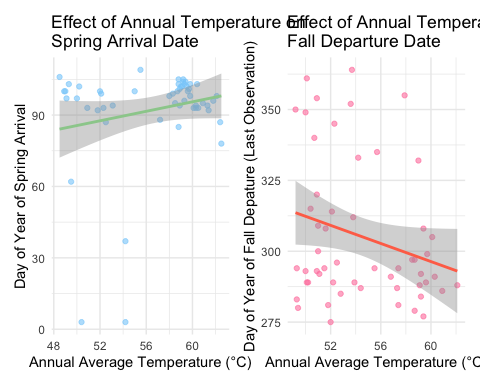
Call:  
lm(formula = DayOfYear ~ Annual, data = Spring\_Arrival)  
  
Residuals:  
 Min 1Q Median 3Q Max   
-86.777 -1.755 5.228 9.191 21.902   
  
Coefficients:  
 Estimate Std. Error t value Pr(>|t|)  
(Intercept) 35.7705 36.6956 0.975 0.334  
Annual 0.9964 0.6456 1.543 0.129  
  
Residual standard error: 20.59 on 52 degrees of freedom  
Multiple R-squared: 0.0438, Adjusted R-squared: 0.02541   
F-statistic: 2.382 on 1 and 52 DF, p-value: 0.1288

springp1 <- ggplot(Spring\_Arrival, aes(x = Annual, y = DayOfYear)) +  
 geom\_point(color = "#87CEFA", alpha = 0.6) +   
 geom\_smooth(method = "lm", color = "#9BCD9B", se = TRUE) +   
 labs(  
 title = "Effect of Annual Temperature on\nSpring Arrival Date",  
 x = "Annual Average Temperature (°C)",   
 y = "Day of Year of Spring Arrival"  
 ) +  
 theme\_minimal()  
  
# Temperature Model - Spring  
Temp\_F\_Main\_Model <- lm(formula = DayOfYear ~ Annual, data = Fall\_Departure)  
  
summary(Temp\_F\_Main\_Model)

Call:  
lm(formula = DayOfYear ~ Annual, data = Fall\_Departure)  
  
Residuals:  
 Min 1Q Median 3Q Max   
-34.127 -18.496 -7.249 9.039 57.588   
  
Coefficients:  
 Estimate Std. Error t value Pr(>|t|)   
(Intercept) 392.1490 47.0361 8.337 3.75e-11 \*\*\*  
Annual -1.5966 0.8621 -1.852 0.0697 .   
---  
Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
  
Residual standard error: 24.59 on 52 degrees of freedom  
Multiple R-squared: 0.06188, Adjusted R-squared: 0.04384   
F-statistic: 3.43 on 1 and 52 DF, p-value: 0.06971

fallp1 <- ggplot(Fall\_Departure, aes(x = Annual, y = DayOfYear)) +  
 geom\_point(color = "palevioletred1", alpha = 0.6) +   
 geom\_smooth(method = "lm", color = "#FF7256", se = TRUE) +   
 labs(  
 title = "Effect of Annual Temperature on\nFall Departure Date",  
 x = "Annual Average Temperature (°C)",   
 y = "Day of Year of Fall Depature (Last Observation)"  
 ) +  
 theme\_minimal()  
  
# Combine plots side by side  
  
library(patchwork)  
  
combined\_plot <- springp1 + fallp1  
  
print(combined\_plot)

`geom\_smooth()` using formula = 'y ~ x'  
`geom\_smooth()` using formula = 'y ~ x'



Source: [Article Notebook](https://andiehal.github.io/ESS-330-Final-Project-REDO/index.qmd.html)

#### CO2 Models -

library(patchwork)  
# CO2 Model - Spring  
CO2\_S\_Main\_Model <- lm(formula = DayOfYear ~ `Annual CO₂ emissions`, data = Spring\_Arrival)  
  
summary(CO2\_S\_Main\_Model)

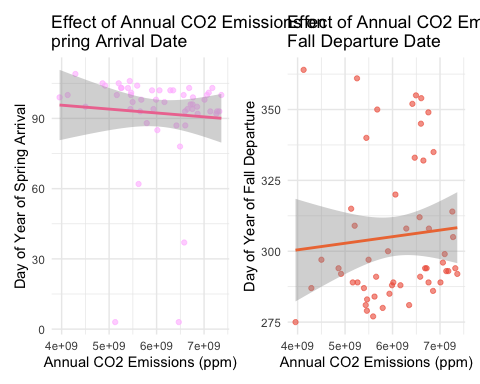
Call:  
lm(formula = DayOfYear ~ `Annual CO₂ emissions`, data = Spring\_Arrival)  
  
Residuals:  
 Min 1Q Median 3Q Max   
-90.777 1.588 4.938 9.630 13.831   
  
Coefficients:  
 Estimate Std. Error t value Pr(>|t|)   
(Intercept) 1.023e+02 2.019e+01 5.066 5.46e-06 \*\*\*  
`Annual CO₂ emissions` -1.660e-09 3.302e-09 -0.503 0.617   
---  
Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
  
Residual standard error: 21.01 on 52 degrees of freedom  
Multiple R-squared: 0.004839, Adjusted R-squared: -0.0143   
F-statistic: 0.2529 on 1 and 52 DF, p-value: 0.6172

springp2 <- ggplot(Spring\_Arrival, aes(x = `Annual CO₂ emissions`, y = DayOfYear)) +  
 geom\_point(color = "#FFBBFF", alpha = 0.6) +   
 geom\_smooth(method = "lm", color = "palevioletred2", se = TRUE) +   
 labs(  
 title = "Effect of Annual CO2 Emissions on\npring Arrival Date",  
 x = "Annual CO2 Emissions (ppm)",   
 y = "Day of Year of Spring Arrival"  
 ) +  
 theme\_minimal()  
  
# CO2 Model - Fall  
CO2\_F\_Main\_Model <- lm(formula = DayOfYear ~ `Annual CO₂ emissions`, data = Fall\_Departure)  
  
summary(CO2\_F\_Main\_Model)

Call:  
lm(formula = DayOfYear ~ `Annual CO₂ emissions`, data = Fall\_Departure)  
  
Residuals:  
 Min 1Q Median 3Q Max   
-27.19 -17.02 -12.16 10.40 63.22   
  
Coefficients:  
 Estimate Std. Error t value Pr(>|t|)   
(Intercept) 2.912e+02 2.432e+01 11.974 <2e-16 \*\*\*  
`Annual CO₂ emissions` 2.323e-09 3.977e-09 0.584 0.562   
---  
Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
  
Residual standard error: 25.3 on 52 degrees of freedom  
Multiple R-squared: 0.006521, Adjusted R-squared: -0.01258   
F-statistic: 0.3413 on 1 and 52 DF, p-value: 0.5616

fallp2 <- ggplot(Fall\_Departure, aes(x = `Annual CO₂ emissions`, y = DayOfYear)) +  
 geom\_point(color = "tomato2", alpha = 0.6) + # Scatter points  
 geom\_smooth(method = "lm", color = "#EE7942", se = TRUE) + # Regression line with confidence interval  
 labs(  
 title = "Effect of Annual CO2 Emissions on\nFall Departure Date",  
 x = "Annual CO2 Emissions (ppm)", # Adjust units if needed  
 y = "Day of Year of Fall Departure"  
 ) +  
 theme\_minimal()  
  
combined\_plot2 <- springp2 + fallp2  
  
print(combined\_plot2)

`geom\_smooth()` using formula = 'y ~ x'  
`geom\_smooth()` using formula = 'y ~ x'



Source: [Article Notebook](https://andiehal.github.io/ESS-330-Final-Project-REDO/index.qmd.html)

## Part 4. Main Analysis - Machine Learning Model

### I. Splitting and CV folds

set.seed(123)  
WT\_split <- initial\_split(Main\_Data, prop = 0.8)  
WT\_train <- training(WT\_split)  
WT\_test <- testing(WT\_split)  
WT\_cv <- vfold\_cv(WT\_train, v = 5)

Source: [Article Notebook](https://andiehal.github.io/ESS-330-Final-Project-REDO/index.qmd.html)

### II. Recipes

# CO2 Recipe   
CO2\_rec <- recipe(formula = DayOfYear ~ `Annual CO₂ emissions`, data = WT\_train) |>  
 step\_log(all\_predictors()) |>  
 step\_naomit(all\_predictors(), all\_outcomes())  
  
# Temperature Recipe  
Temp\_rec <- recipe(formula = DayOfYear ~ Annual, data = WT\_train) |>  
 step\_log(all\_predictors(), offset = 1) |>  
 step\_naomit(all\_predictors(), all\_outcomes())

Source: [Article Notebook](https://andiehal.github.io/ESS-330-Final-Project-REDO/index.qmd.html)

### III. Define Models

lm\_model <- linear\_reg() |> set\_engine("lm") |> set\_mode("regression")  
rf\_model <- rand\_forest(trees = 100) |> set\_engine("ranger") |> set\_mode("regression")  
xg\_model <- boost\_tree() |> set\_engine("xgboost") |> set\_mode("regression")

Source: [Article Notebook](https://andiehal.github.io/ESS-330-Final-Project-REDO/index.qmd.html)

### IV. Workflows

# CO2 Workflows  
wf\_lm\_CO2 <- workflow() |>  
 add\_model(lm\_model) |>  
 add\_recipe(CO2\_rec)  
  
wf\_rf\_CO2 <- workflow() |>  
 add\_model(rf\_model) |>  
 add\_recipe(CO2\_rec)  
  
wf\_xg\_CO2 <- workflow() |>  
 add\_model(xg\_model) |>  
 add\_recipe(CO2\_rec)  
  
# Temperature Workflows  
wf\_lm\_Temp <- workflow() |>  
 add\_model(lm\_model) |>  
 add\_recipe(Temp\_rec)  
  
wf\_rf\_Temp <- workflow() |>  
 add\_model(rf\_model) |>  
 add\_recipe(Temp\_rec)  
  
wf\_xg\_Temp <- workflow() |>  
 add\_model(xg\_model) |>  
 add\_recipe(Temp\_rec)

Source: [Article Notebook](https://andiehal.github.io/ESS-330-Final-Project-REDO/index.qmd.html)

### V.a CO2 - Fitting Models

# Linear - CO2  
res\_lm\_CO2 <- fit\_resamples(  
 wf\_lm\_CO2,  
 resamples = WT\_cv,  
 metrics = metric\_set(rmse, rsq),  
 control = control\_resamples(save\_pred = TRUE)  
)  
  
# Random Forest - CO2  
res\_rf\_CO2 <- fit\_resamples(  
 wf\_rf\_CO2,  
 resamples = WT\_cv,  
 metrics = metric\_set(rmse, rsq),  
 control = control\_resamples(save\_pred = TRUE)  
)  
  
# XG Boosted Tree - CO2  
res\_xg\_CO2 <- fit\_resamples(  
 wf\_xg\_CO2,  
 resamples = WT\_cv,  
 metrics = metric\_set(rmse, rsq),  
 control = control\_resamples(save\_pred = TRUE)  
)

Source: [Article Notebook](https://andiehal.github.io/ESS-330-Final-Project-REDO/index.qmd.html)

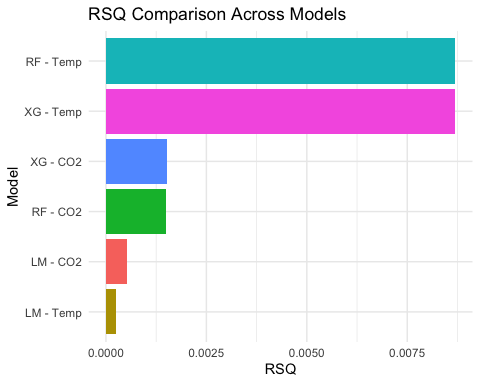
### V.b Temperature - Fitting Models

# Linear - Temperature   
res\_lm\_Temp <- fit\_resamples(  
 wf\_lm\_Temp,  
 resamples = WT\_cv,  
 metrics = metric\_set(rmse, rsq),  
 control = control\_resamples(save\_pred = TRUE)  
)  
  
# Random Forest - Temperature  
res\_rf\_Temp <- fit\_resamples(  
 wf\_rf\_Temp,  
 resamples = WT\_cv,  
 metrics = metric\_set(rmse, rsq),  
 control = control\_resamples(save\_pred = TRUE)  
)  
  
# XG Boosted Tree - Temperature  
res\_xg\_Temp <- fit\_resamples(  
 wf\_xg\_Temp,  
 resamples = WT\_cv,  
 metrics = metric\_set(rmse, rsq),  
 control = control\_resamples(save\_pred = TRUE)  
)

Source: [Article Notebook](https://andiehal.github.io/ESS-330-Final-Project-REDO/index.qmd.html)

### VI. Metrics and Comparison

all\_results <- bind\_rows(  
 collect\_metrics(res\_lm\_CO2) |> mutate(Model = "LM - CO2"),  
 collect\_metrics(res\_rf\_CO2) |> mutate(Model = "RF - CO2"),  
 collect\_metrics(res\_xg\_CO2) |> mutate(Model = "XG - CO2"),  
 collect\_metrics(res\_lm\_Temp) |> mutate(Model = "LM - Temp"),  
 collect\_metrics(res\_rf\_Temp) |> mutate(Model = "RF - Temp"),  
 collect\_metrics(res\_xg\_Temp) |> mutate(Model = "XG - Temp")  
)  
  
all\_results |>  
 filter(.metric == "rsq") |>  
 ggplot(aes(x = reorder(Model, mean), y = mean, fill = Model)) +  
 geom\_col(show.legend = FALSE) +  
 labs(  
 title = "RSQ Comparison Across Models",  
 x = "Model",  
 y = "RSQ"  
 ) +  
 coord\_flip() +  
 theme\_minimal()



Source: [Article Notebook](https://andiehal.github.io/ESS-330-Final-Project-REDO/index.qmd.html)

## Part 5. Full Analysis - Spring Arrival and Fall Departure Random Forest Predictions

### I. Fit to Full Model (Spring and Fall | CO2 and Temperature)

# Temperature Models  
rf\_temp\_spring\_fit <- fit(wf\_rf\_Temp, data = Spring\_Arrival)  
  
rf\_temp\_fall\_fit <- fit(wf\_rf\_Temp, data = Fall\_Departure)  
  
# CO2 Models  
rf\_CO2\_spring\_fit <- fit(wf\_rf\_CO2, data = Spring\_Arrival)  
  
rf\_CO2\_fall\_fit <- fit(wf\_rf\_CO2, data = Fall\_Departure)

Source: [Article Notebook](https://andiehal.github.io/ESS-330-Final-Project-REDO/index.qmd.html)

### II. Predictions for Models

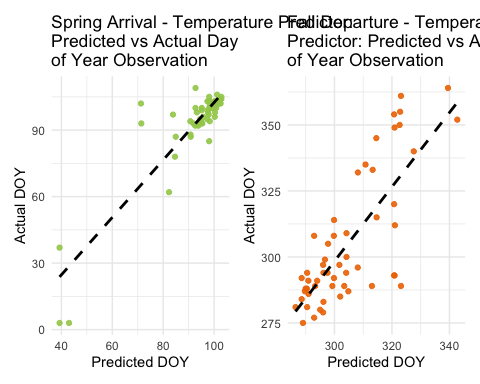
# Temperature Models  
temp\_spring\_preds <- predict(rf\_temp\_spring\_fit, Spring\_Arrival) |>  
 bind\_cols(Spring\_Arrival)  
  
temp\_fall\_preds <- predict(rf\_temp\_fall\_fit, Fall\_Departure) |>  
 bind\_cols(Fall\_Departure)  
  
# CO2 Models  
CO2\_spring\_preds <- predict(rf\_CO2\_spring\_fit, Spring\_Arrival) |>  
 bind\_cols(Spring\_Arrival)  
  
CO2\_fall\_preds <- predict(rf\_CO2\_fall\_fit, Fall\_Departure) |>  
 bind\_cols(Fall\_Departure)

Source: [Article Notebook](https://andiehal.github.io/ESS-330-Final-Project-REDO/index.qmd.html)

### III. Plotting Temperature Prediction Models

library(patchwork)   
  
# Spring Arrival  
p1 <- ggplot(temp\_spring\_preds, aes(x = .pred, y = DayOfYear)) +  
 geom\_point(alpha = 0.9, color = "#A2CD5A") +  
 geom\_smooth(method = "lm", se = FALSE, color = "black", linetype = "dashed") +  
 theme\_minimal() +  
 labs(title = "Spring Arrival - Temperature Predictor:\nPredicted vs Actual Day\nof Year Observation",  
 x = "Predicted DOY", y = "Actual DOY")  
  
# Fall Departure  
p2 <- ggplot(temp\_fall\_preds, aes(x = .pred, y = DayOfYear)) +  
 geom\_point(alpha = 0.9, color = "darkorange2") +  
 geom\_smooth(method = "lm", se = FALSE, color = "black", linetype = "dashed") +  
 theme\_minimal() +  
 labs(title = "Fall Departure - Temperature\nPredictor: Predicted vs Actual Day\nof Year Observation",  
 x = "Predicted DOY", y = "Actual DOY")  
  
  
p1 + p2

`geom\_smooth()` using formula = 'y ~ x'  
`geom\_smooth()` using formula = 'y ~ x'

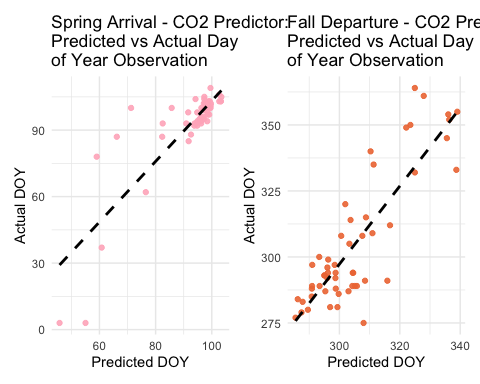


Source: [Article Notebook](https://andiehal.github.io/ESS-330-Final-Project-REDO/index.qmd.html)

### IV. Plotting CO2 Prediction Models

# Spring Arrival  
p3 <- ggplot(CO2\_spring\_preds, aes(x = .pred, y = DayOfYear)) +  
 geom\_point(alpha = 0.9, color = "pink1") +  
 geom\_smooth(method = "lm", se = FALSE, color = "black", linetype = "dashed") +  
 theme\_minimal() +  
 labs(title = "Spring Arrival - CO2 Predictor:\nPredicted vs Actual Day\nof Year Observation",  
 x = "Predicted DOY", y = "Actual DOY")  
  
# Fall Departure  
p4 <- ggplot(CO2\_fall\_preds, aes(x = .pred, y = DayOfYear)) +  
 geom\_point(alpha = 0.9, color = "sienna2") +  
 geom\_smooth(method = "lm", se = FALSE, color = "black", linetype = "dashed") +  
 theme\_minimal() +  
 labs(title = "Fall Departure - CO2 Predictor:\nPredicted vs Actual Day\nof Year Observation",  
 x = "Predicted DOY", y = "Actual DOY")  
  
  
p3 + p4

`geom\_smooth()` using formula = 'y ~ x'  
`geom\_smooth()` using formula = 'y ~ x'



Source: [Article Notebook](https://andiehal.github.io/ESS-330-Final-Project-REDO/index.qmd.html)

## **Part 6. Results Discussion**

To assess the relationship between climate variables and Wood Thrush migration phenology, we implemented and evaluated six machine learning models: linear regression, random forest, and XGBoost—each applied separately to CO₂ and temperature predictor datasets. Performance was measured using 5-fold cross-validation due to the size of the datasets and runtime constraints. With this we had 4 total final datasets, separated by Spring/Fall and Temperature/CO₂.

The **random forest model with temperature** as the predictor emerged as the strongest model across both spring arrival and fall departure dates, showing the **lowest RMSE** and **highest R² values**, indicating superior predictive performance and robustness.

### **I. Spring Arrival Predictions**

#### Temperature Models -

The Temperature Spring model’s predictions showed a **tight clustering along the 1:1 line**, especially near the top-right of the plot. This indicates that as temperatures have risen, **Wood Thrushes are arriving later than expected**. This mismatch in timing may suggest a lag in response to changing environmental cues or a disruption in optimal conditions for early arrival.

The higher accuracy in spring predictions may be because spring arrival is tightly regulated by temperature-dependent cues (such as snow melt, insect emergence, or vegetation growth), making it more sensitive to annual temperature changes.

#### CO₂ Models -

CO₂’s Spring model had similar results, with a clustering of points in the top right hand corner and fairly close to the actual dates. This is sensible with how CO₂ comes as a direct impact onto temperature and can provide data that states that as they move in similar directions, they both have impacts on migration periods and dates.

The accuracy is reliable, however, when compared to the temperature model, they have differences in how many points are along this line or close to. In conclusion for spring arrival, CO₂ and temperature have a higher reliability to help predict dates with temperature being more accurate.

### **II. Fall Departure Predictions**

#### Temperature Models -

Fall departure predictions demonstrated more **scatter and variability** compared to spring arrival. While the random forest model still performed best, the weaker R² values suggest that **fall migration timing is less directly tied to temperature or CO₂ alone**. The wider spread of points around the 1:1 line, particularly in the lower-left region, implies that **birds are departing earlier than expected** in many cases.

This might indicate that other variables—such as food availability, photoperiod (day length), or precipitation patterns—play a larger role in influencing fall departure behavior.

#### CO₂ Models -

While temperature-based models consistently outperformed CO₂-based models, the **patterns in both sets of predictions were similar**, suggesting that CO₂ indirectly influences migration phenology through its effect on global temperatures.

Visualizations comparing CO₂ and temperature models showed nearly **parallel trends**, with later spring arrivals and earlier fall departures becoming more common over time—consistent with broader ecological impacts of climate change.

### III. Overall Analysis Conclusion

From our multiple model simulations, we concluded that temperature has the bigger impact directly on migration dates. While CO₂ did have similar results to temperature, it had much more variability and less accuracy. This may be because of CO₂ potentially being an indirect impact on migration patterns, but can help to predict these dates when added to temperature models.

**Future studies** could include more sophisticated models on computers that can run more in-depth and larger datasets. Along with more skill knowledge of RStudio and how to better create models that both predict and demonstrate correlation between predictor and response variables.

**Some limitations** that we encountered throughout this project was mostly through data analysis and cleaning. Our climate data was particular to how it was supposed to be cleaned and took many hours of excel refining to be usable in RStudio. Another major roadblock that we encountered and was a large portion of our challenges was the rendering of our files to be able to display them as a website through Github pages. With multiple meetings dedicated to resolving this issue, we still faced many errors with long rendering periods, models taking hours to process, and seemingly endless errors. However, we preserved through it all to create a simple yet replicable method to help understand the correlation between temperature, CO₂, and migration dates.

**The significance of this study** relates back to our original problem that inspired our target. Avian species are one of the longest standing species groups, dating back to before records existed. With their adaptable and resilient nature, they are ideal for studies regarding changes to key factors that can drastically harm their populations. Our goal with this study was to begin a coding process that can be replicated and used across multiple species of different conservation statuses. While this report may need more refining and different methods to achieve a successful and accurate depiction of the impact on these variables, it begins the discussion of how humans are not the only species to be impacted by CO₂ emissions and temperature variance.

**Our take away** from this is that even as junior data analysts, we can create studies that can be built upon and developed to create solutions or incentives for policy management of wildlife and emissions. We hoped to contribute to the wider ecological research surrounding conservation of native and important species.

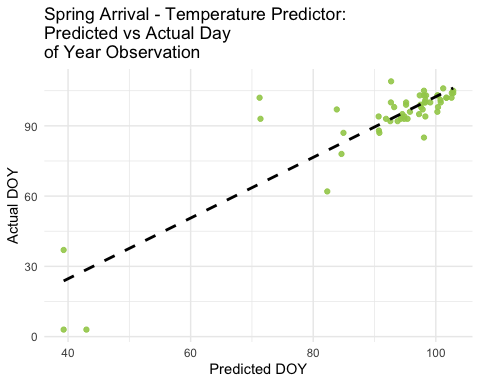
### **IV. Visual Summary**

We generated the following visualizations to illustrate key trends:

* **Figure 1**: Spring Arrival vs. Annual Temperature

print(p1)

`geom\_smooth()` using formula = 'y ~ x'

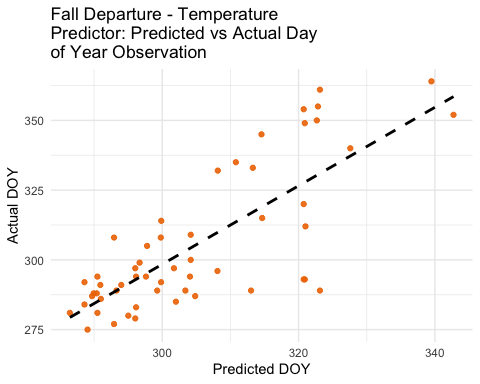


Source: [Article Notebook](https://andiehal.github.io/ESS-330-Final-Project-REDO/index.qmd.html)

* **Figure 2**: Fall Departure vs. Annual Temperature

print(p2)

`geom\_smooth()` using formula = 'y ~ x'

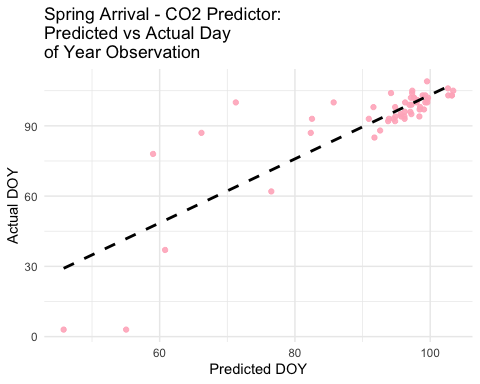


Source: [Article Notebook](https://andiehal.github.io/ESS-330-Final-Project-REDO/index.qmd.html)

* **Figure 3**: Spring Arrival vs. Annual CO₂

print(p3)

`geom\_smooth()` using formula = 'y ~ x'

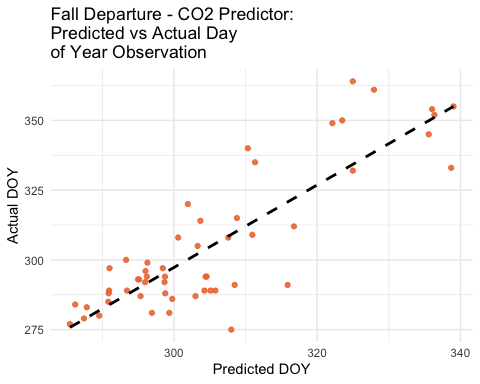


Source: [Article Notebook](https://andiehal.github.io/ESS-330-Final-Project-REDO/index.qmd.html)

* **Figure 4**: Fall Departure vs. Annual CO₂

print(p4)

`geom\_smooth()` using formula = 'y ~ x'



Source: [Article Notebook](https://andiehal.github.io/ESS-330-Final-Project-REDO/index.qmd.html)

Each figure includes:

* A 1:1 reference line for perfect prediction
* Color scaling by year or observation density (if used)
* Interpretation notes on clustering trends

These visualizations confirm that the **relationship between climate variables and migration phenology is strong**, particularly for temperature and spring migration events.

Friedlingstein, P. et al. (2023). Global carbon budget 2023. *Earth System Science Data*, *15*, 123–145. <https://doi.org/10.5194/essd-15-123-2023>

Lang, W. et al. (2024). Phenological divergence between plants and animals under climate change. *Nature Ecology and Evolution*, *9*, 261–272. https://doi.org/<https://doi.org/10.1038/s41559-024-02597-0>

NOAA National Centers for Environmental Information. (2024). *Monthly global climate report for annual 2023*. <https://www.ncei.noaa.gov/access/monitoring/monthly-report/global/202313>

Ornithology, C. L. of. (2024). *Wood thrush life history*. <https://www.allaboutbirds.org/guide/Wood_Thrush/lifehistory>