PCA with climate summary stats

Norah Saarman

2025-03-05

Load and Merge Data

Prepare for PCA, selecting only relevant columns

```r
# Load each data set
hybrid_data <- read_csv("../data/summary_stats_5y_hyb.csv")
```

```
## Rows: 43 Columns: 14
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## dbl (14): lat, lon, year, ID, mean_temp, mean_diurnal_range, max_temp_warmes...
##
## 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.
```

```r
pipiens_data <- read_csv("../data/summary_stats_5y_pip.csv")
```

```
## Rows: 413 Columns: 14
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## dbl (14): lat, lon, year, ID, mean_temp, mean_diurnal_range, max_temp_warmes...
##
## 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.
```

```r
quinquefasciatus_data <- read_csv("../data/summary_stats_5y_qui.csv")
```

```
## Rows: 340 Columns: 14
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## dbl (14): lat, lon, year, ID, mean_temp, mean_diurnal_range, max_temp_warmes...
##
## 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.
```

```r
# Add species labels to each dataset
hybrid_data <- hybrid_data %>% mutate(species = "Hybrid")
pipiens_data <- pipiens_data %>% mutate(species = "Culex pipiens")
quinquefasciatus_data <- quinquefasciatus_data %>% mutate(species = "Culex quinquefasciatus")

# Merge all data into one dataframe
climate_data <- bind_rows(hybrid_data, pipiens_data, quinquefasciatus_data)
```

```r
# Check structure
str(climate_data)
```

```
## tibble [796 x 15] (S3: tbl_df/tbl/data.frame)
##  $ lat                         : num [1:796] 37.2 35.4 36.3 36.8 37.3 ...
##  $ lon                         : num [1:796] -120 -119 -119 -120 -121 ...
##  $ year                        : num [1:796] 1991 1992 1992 1992 1992 ...
##  $ ID                          : num [1:796] 1 2 3 4 5 6 7 8 9 10 ...
##  $ mean_temp                   : num [1:796] 17.1 18.7 17.9 17.5 17.1 ...
##  $ mean_diurnal_range          : num [1:796] 15.8 15.2 15.3 15.4 14.6 ...
##  $ max_temp_warmest_month      : num [1:796] 42.9 43 43.4 42.5 41.5 ...
##  $ min_temp_coldest_month      : num [1:796] -8.26 -8.38 -7.2 -8.17 -8.21 ...
##  $ annual_range_temp           : num [1:796] 46.5 45.5 44.8 45.4 45.1 ...
##  $ annual_mean_precip          : num [1:796] 262 170 247 236 227 ...
##  $ highest_monthly_precip      : num [1:796] 89.5 69.4 98.9 88.8 80.5 ...
##  $ annual_avg_vapor_pressure   : num [1:796] 730 709 744 725 712 ...
##  $ highest_vapor_pressure_month: num [1:796] 1147 1084 1268 1202 1242 ...
##  $ lowest_vapor_pressure_month : num [1:796] 353 311 342 359 336 ...
##  $ species                     : chr [1:796] "Hybrid" "Hybrid" "Hybrid" "Hybrid" ...
```

```r
# Add a new column to classify data into three time periods
climate_data <- climate_data %>%
  mutate(time_period = case_when(
    year < 2000 ~ "Before 2000",
    year >= 2000 & year <= 2010 ~ "2000-2010",
    year > 2010 ~ "After 2010"
  ))

# Convert to factor to ensure correct ordering in plots
climate_data$time_period <- factor(climate_data$time_period, levels = c("Before 2000", "2000-2010", "Af

# Select only numeric climate variables (remove non-climate columns)
# Leaving latitude in as a proxy for autumn/spring photoperiod
climate_vars <- climate_data %>% select(-species, -lon, -year, -ID, -time_period)

# Standardize climate variables (PCA works best with scaled data)
climate_vars_scaled <- scale(climate_vars)
```

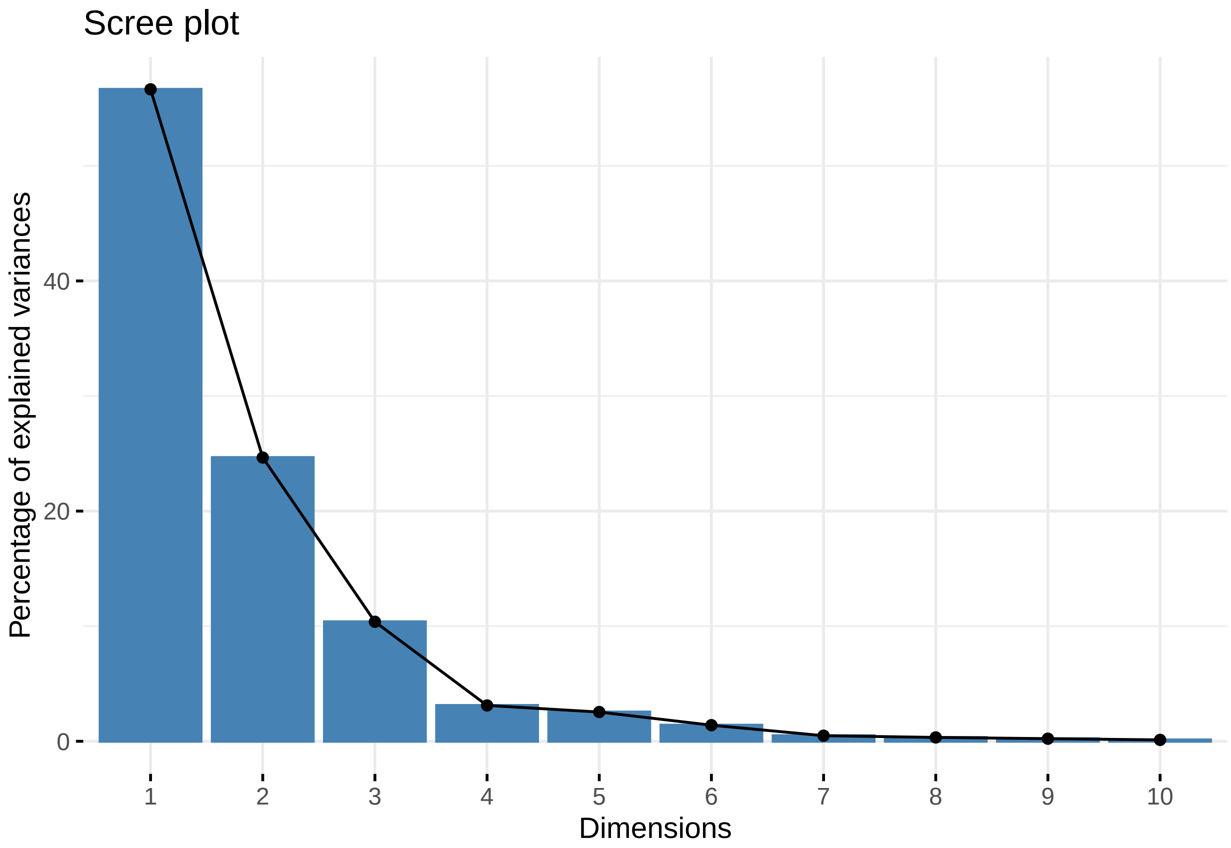
Perform PCA

```r
# Run PCA
pca_result <- prcomp(climate_vars_scaled, center = TRUE, scale. = TRUE)

# Check variance explained by PCs
summary(pca_result)
```

```
## Importance of components:
##                           PC1    PC2    PC3     PC4     PC5     PC6     PC7
## Standard deviation     2.4962 1.6468 1.0688 0.58538 0.52899 0.39262 0.23177
## Proportion of Variance 0.5665 0.2465 0.1038 0.03115 0.02544 0.01401 0.00488
## Cumulative Proportion  0.5665 0.8130 0.9169 0.94800 0.97344 0.98745 0.99234
##                            PC8     PC9    PC10    PC11
## Standard deviation     0.19245 0.15821 0.11665 0.09288
## Proportion of Variance 0.00337 0.00228 0.00124 0.00078
## Cumulative Proportion  0.99570 0.99798 0.99922 1.00000
```

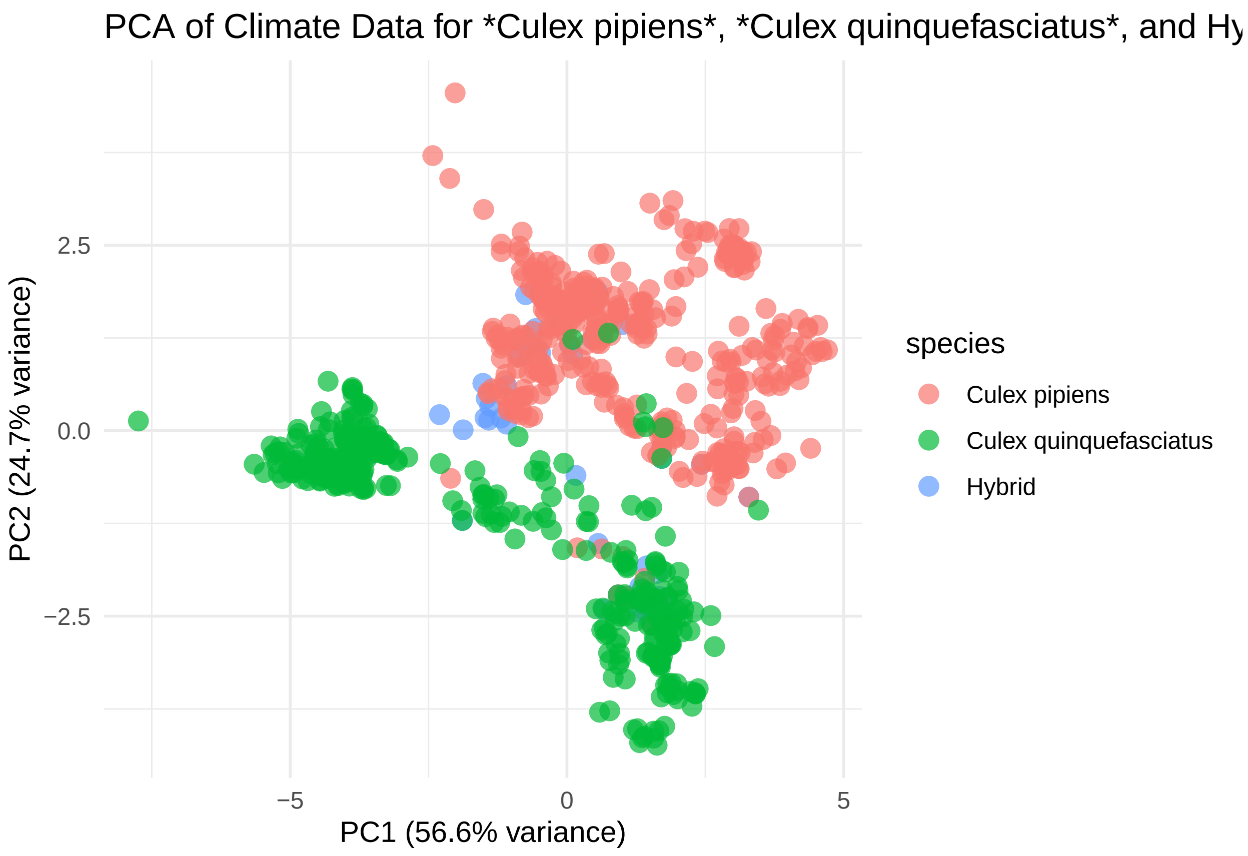
```
# Scree plot to visualize explained variance
fviz_eig(pca_result)
```



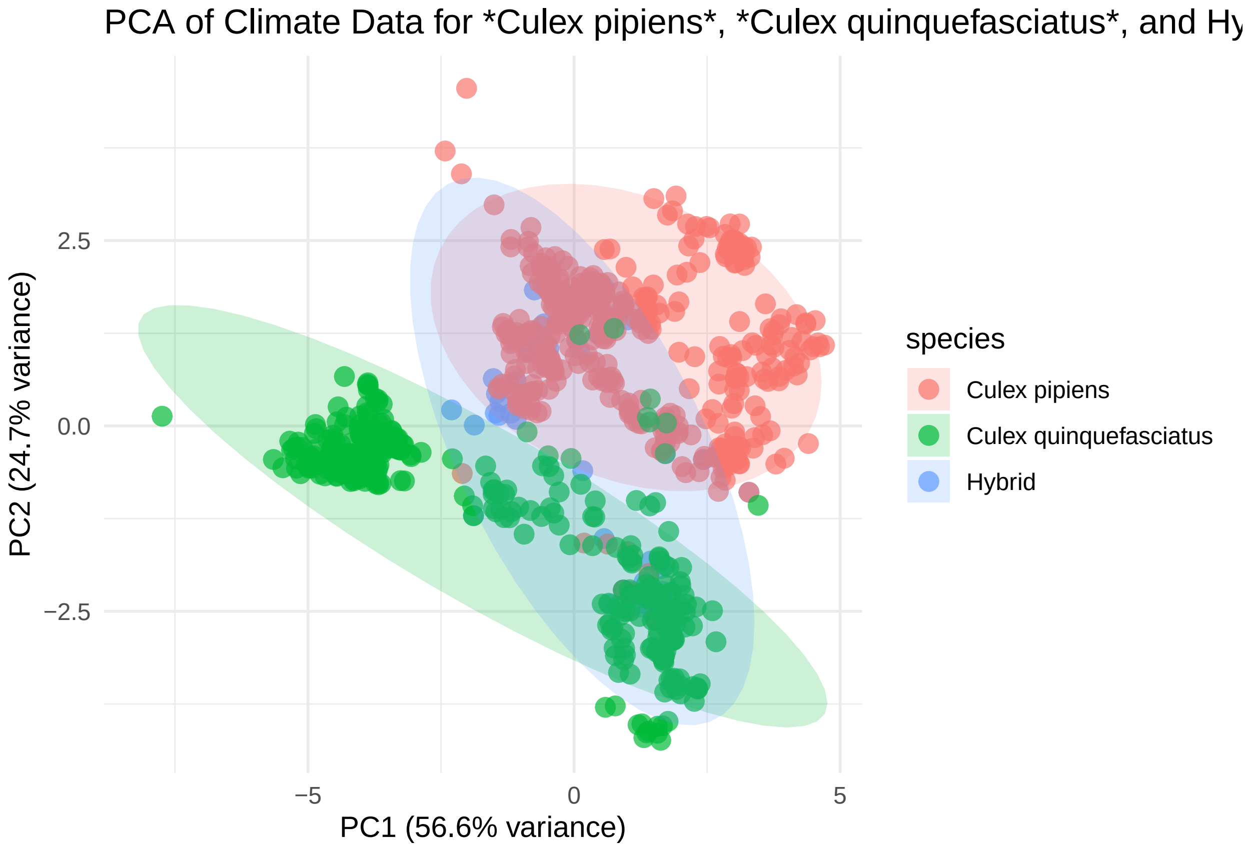
```
# Add PCA results to a new dataframe
pca_data <- as.data.frame(pca_result$x)
pca_data$species <- climate_data$species  # Retain species labels
pca_data$time_period <- climate_data$time_period  # Add time period labels
```

Visualize PCA results by species

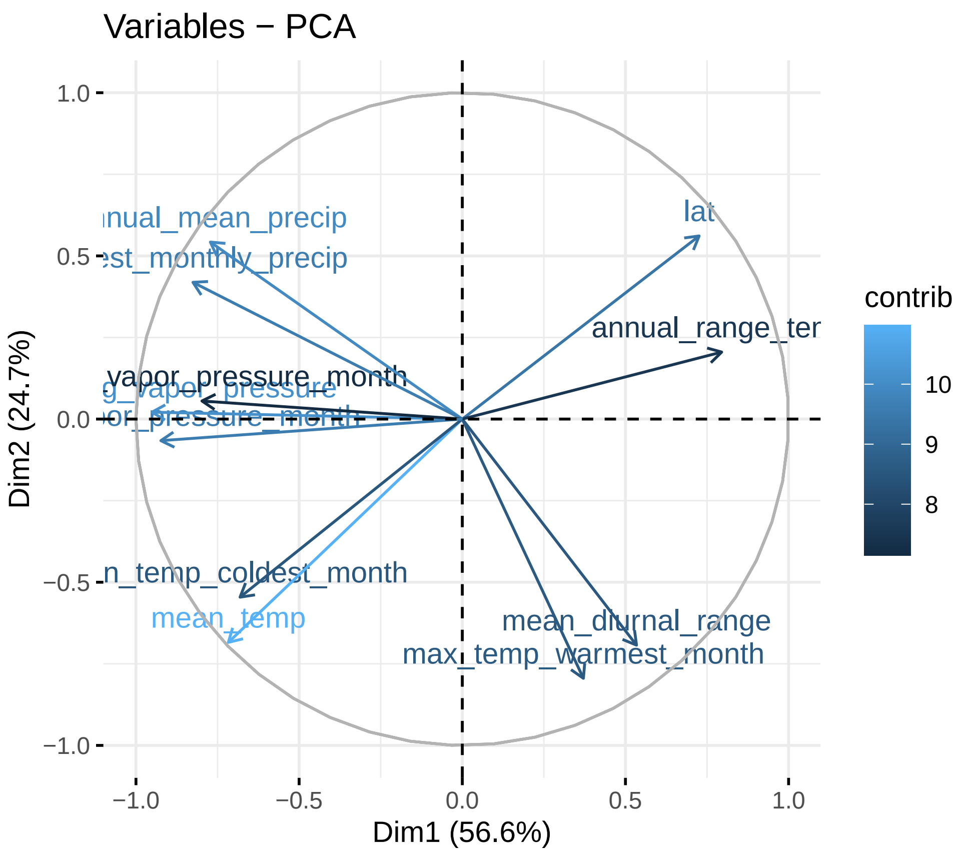
```
# PCA Scatter Plot
ggplot(pca_data, aes(x = PC1, y = PC2, color = species)) +
  geom_point(size = 3, alpha = 0.7) +
  labs(title = "PCA of Climate Data for *Culex pipiens*, *Culex quinquefasciatus*, and Hybrids",
       x = paste0("PC1 (", round(summary(pca_result)$importance[2,1] * 100, 1), "% variance)"),
       y = paste0("PC2 (", round(summary(pca_result)$importance[2,2] * 100, 1), "% variance)")) +
  theme_minimal() +
  scale_color_manual(values = c("#F8766D", "#00BA38", "#619CFF"))  # Adjust colors
```



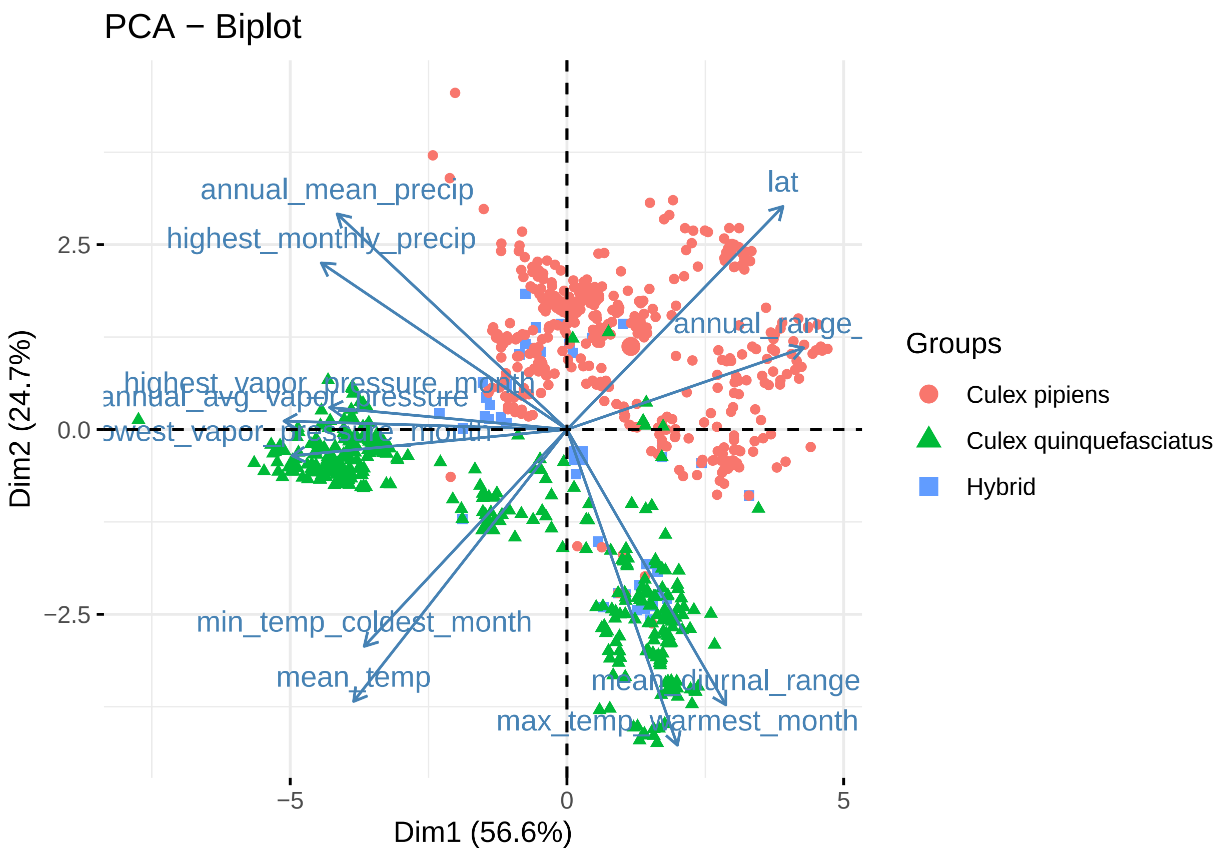
```
# PCA Scatter Plot with 95% confidence ellipses
ggplot(pca_data, aes(x = PC1, y = PC2, color = species, fill = species)) +
  geom_point(size = 3, alpha = 0.7) +  # Scatter plot points
  stat_ellipse(type = "t", level = 0.95, alpha = 0.2, geom = "polygon", color = NA) +  # 95% Confidence
  labs(title = "PCA of Climate Data for *Culex pipiens*, *Culex quinquefasciatus*, and Hybrids",
       x = paste0("PC1 (", round(summary(pca_result)$importance[2,1] * 100, 1), "% variance)"),
       y = paste0("PC2 (", round(summary(pca_result)$importance[2,2] * 100, 1), "% variance)")) +
  theme_minimal() +
  scale_color_manual(values = c("#F8766D", "#00BA38", "#619CFF")) +  # Adjust point colors
  scale_fill_manual(values = c("#F8766D", "#00BA38", "#619CFF")) +  # Adjust ellipse colors
  theme(legend.position = "right")
```



```
# Variable contributions to PCs
fviz_pca_var(pca_result, col.var = "contrib")
```



```r
# Biplot with species grouping
fviz_pca_biplot(pca_result, label="var", habillage = climate_data$species)
```



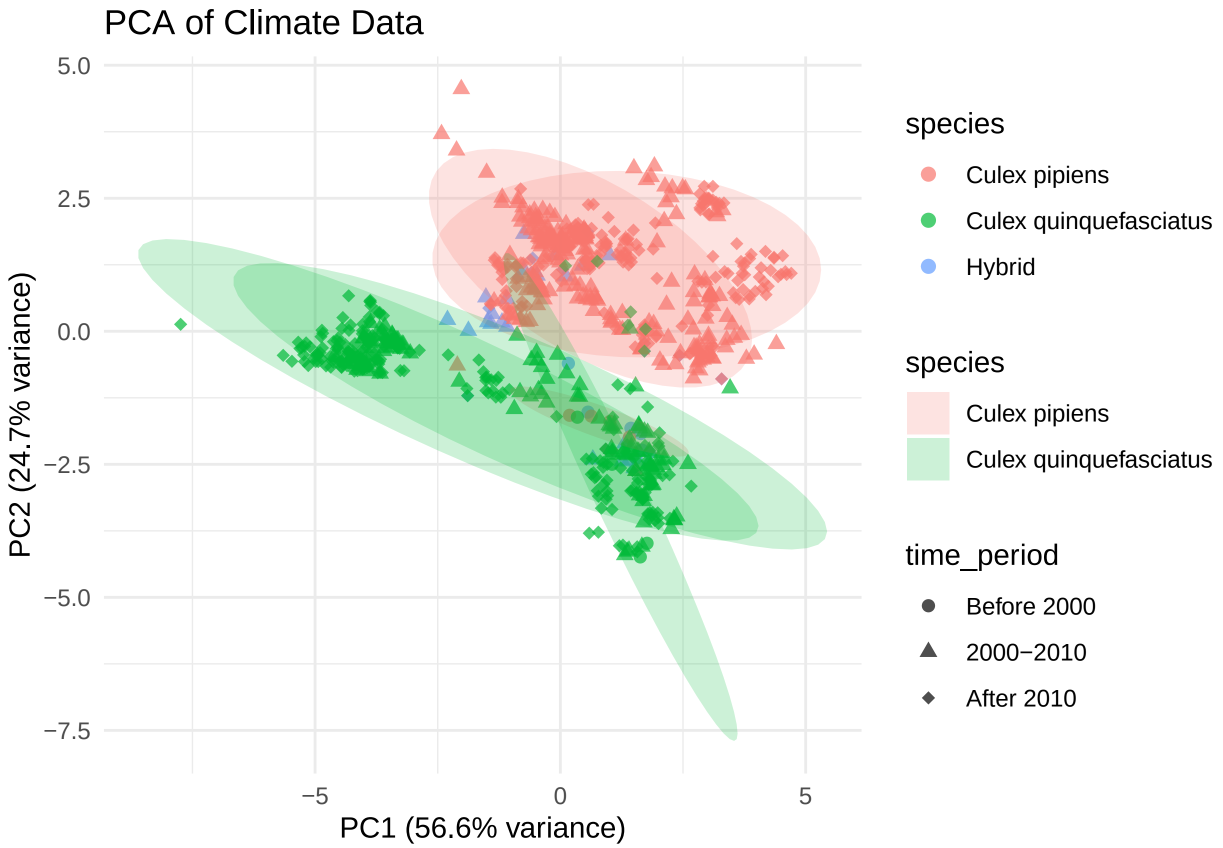
Vizualize PCA results by species and Time Period

Before 2000 2000-2010 After 2010

```
# Define colors by species
species_colors <- c("Culex pipiens" = "#F8766D",
           "Culex quinquefasciatus" = "#00BA38",
           "Hybrid" = "#619CFF")

# Define different shapes for time periods
time_shapes <- c("Before 2000" = 16,
                 "2000-2010" = 17,
                 "After 2010" = 18)  # Different point symbols

ggplot(pca_data, aes(x = PC1, y = PC2,
      color = species, shape = time_period)) +
      geom_point(size = 2, alpha = 0.7) + # Points by species,
      stat_ellipse(data = pca_data %>% filter(species != "Hybrid"),
               aes(fill = species), type = "t", level = 0.95,
               geom = "polygon", alpha = 0.2,
               color = NA) +  # ellipses by species
      labs(title = "PCA of Climate Data",
       x = paste0("PC1 (",round(summary(pca_result)$importance[2,1] * 100, 1), "% variance)"),
       y = paste0("PC2 (", round(summary(pca_result)$importance[2,2] * 100, 1), "% variance)")) +
      theme_minimal() +
      scale_color_manual(values = species_colors) +  # by species
      scale_fill_manual(values = species_colors) +  # by species
      scale_shape_manual(values = time_shapes) +  # by time period
      theme(legend.position = "right")
```



Using fviz_pca_biplot and splitting by species and time period

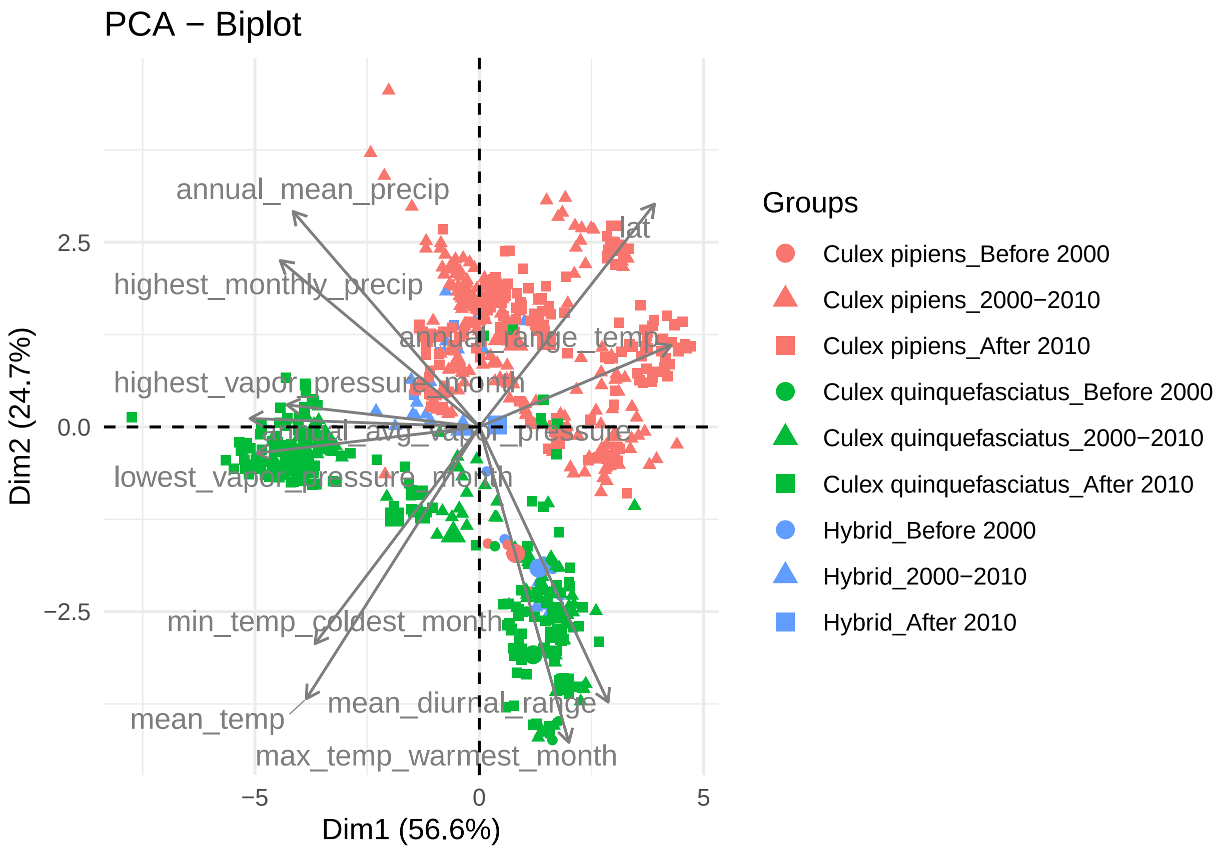
```
# Create a new categorical variable for species + time period
pca_data <- pca_data %>%
  mutate(species_time = paste(species, time_period, sep = "_"))

# Ensure the new variable is a factor
pca_data$species_time <- factor(pca_data$species_time,
  levels = c("Culex pipiens_Before 2000", "Culex pipiens_2000-2010",
  "Culex pipiens_After 2010", "Culex quinquefasciatus_Before 2000",
  "Culex quinquefasciatus_2000-2010", "Culex quinquefasciatus_After 2010",
  "Hybrid_Before 2000", "Hybrid_2000-2010", "Hybrid_After 2010"))

# Define colors (repeating species colors for each time period)
species_colors <- c("Culex pipiens_Before 2000" = "#F8766D",
  "Culex pipiens_2000-2010" = "#F8766D",
  "Culex pipiens_After 2010" = "#F8766D",
  "Culex quinquefasciatus_Before 2000" = "#00BA38",
  "Culex quinquefasciatus_2000-2010" = "#00BA38",
  "Culex quinquefasciatus_After 2010" = "#00BA38",
  "Hybrid_Before 2000" = "#619CFF",
  "Hybrid_2000-2010" = "#619CFF",
  "Hybrid_After 2010" = "#619CFF"
)

fviz_pca_biplot(pca_result,
                label = "var",  # Show variable labels
                habillage = pca_data$species_time,  # Trick habillage
                addEllipses = FALSE,  # No automatic ellipses
                repel = TRUE,  # Avoid overlapping labels
                col.var = "grey50"  # Change variable arrows to grey
) +
  scale_shape_manual(values = c(16, 17, 15, 16, 17, 15, 16, 17, 15)) +  # Circle, Triangle, Square
  scale_color_manual(values = species_colors) +  # Ensure consistent species colors
  theme_minimal() +
  theme(legend.position = "right")
```

```
## Scale for shape is already present.
## Adding another scale for shape, which will replace the existing scale.
```


PCA − Biplot
Dim2 (24.7%)
Dim1 (56.6%)
annual_mean_precip
highest_monthly_precip
lat
annual_range_temp
highest_vapor_pressure_month
annual_avg_vapor_pressure
lowest_vapor_pressure_month
min_temp_coldest_month
mean_diurnal_range
mean_temp
max_temp_warmest_month
Groups
● Culex pipiens_Before 2000
▲ Culex pipiens_2000−2010
■ Culex pipiens_After 2010
● Culex quinquefasciatus_Before 2000
▲ Culex quinquefasciatus_2000−2010
■ Culex quinquefasciatus_After 2010
● Hybrid_Before 2000
▲ Hybrid_2000−2010
■ Hybrid_After 2010