# Monarch Butterfly Abundance Analysis: Wind and Environmental Factors

Analysis Report

2025-08-12

## Monarch Butterfly Abundance Analysis: LLM Implementation

### Part 1: Setup and Data Loading

```
library(tidyverse)
library(lubridate)
library(mgcv)  # for GAM models
library(corrplot)
library(knitr)
```

```
counts <- read_csv("../data/butterfly_abundance_index.csv")
```

```
Rows: 11885 Columns: 4
-- Column specification --------------------------------------------------------
Delimiter: ","
chr (2): deployment_id, image_filename
dbl (2): total_butterflies, butterflies_direct_sun

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.
```

```
deployments <- read_csv("../data/deployments.csv")
```

```
Rows: 19 Columns: 23
-- Column specification --------------------------------------------------------
```

```
Delimiter: ","
chr (14): camera_name, wind_meter_name, Deployed_time, Recovered_time, notes...
dbl  (8): height_m, horizontal_dist_to_cluster_m, view_direction, cluster_co...
lgl  (1): monarchs_present

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
temp <- read_csv("../data/temperature_data_2023.csv")
```

```
Rows: 56066 Columns: 6
-- Column specification --------------------------------------------------------
Delimiter: ","
chr (3): filename, deployment_id, extraction_status
dbl (3): timestamp, temperature, confidence

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
wind <- read_csv("../data/wind_all.csv")
```

```
Rows: 681013 Columns: 5
-- Column specification --------------------------------------------------------
Delimiter: ","
chr  (2): db_path, wind_meter_name
dbl  (2): speed, gust
dttm (1): time

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
cat("Data loaded successfully\n")
```

```
Data loaded successfully
```

```r
cat("Counts rows:", nrow(counts), "\n")
```

```
Counts rows: 11885
```

```r
cat("Deployments rows:", nrow(deployments), "\n")
```

```
Deployments rows: 19
```

```r
cat("Temperature rows:", nrow(temp), "\n")
```

```
Temperature rows: 56066
```

```r
cat("Wind rows:", nrow(wind), "\n")
```

```
Wind rows: 681013
```

---

### Part 2: Initial Data Cleaning and Merging

```r
# Filter and select deployments data
deployments_filtered <- deployments %>%
  filter(label_status == "Complete") %>%
  mutate(view_id = as.factor(view_id)) %>%
  select(deployment_id, view_id, wind_meter_name)

cat("Filtered deployments to", nrow(deployments_filtered), "complete records\n")
```

```
Filtered deployments to 12 complete records
```

```r
# Parse timestamps in counts data
counts_with_datetime <- counts %>%
  mutate(
    datetime_str = str_extract(image_filename, "\\d{14}"),
    datetime = ymd_hms(datetime_str)
  ) %>%
  select(-datetime_str)

cat("Parsed", sum(!is.na(counts_with_datetime$datetime)), "valid timestamps\n")
```

```
Parsed 11885 valid timestamps
```

```r
# Create master dataframe
master_df <- counts_with_datetime %>%
  left_join(deployments_filtered, by = "deployment_id") %>%
  left_join(temp, by = c("image_filename" = "filename"))

cat("Master dataframe created with", nrow(master_df), "rows\n")
```

```
Master dataframe created with 11885 rows
```

```r
cat("Columns:", paste(names(master_df), collapse = ", "), "\n")
```

```
Columns: deployment_id.x, image_filename, total_butterflies, butterflies_direct_sun, datetime
```

```r
# Check for completeness
complete_records <- master_df %>%
  filter(!is.na(view_id) & !is.na(temperature) & !is.na(datetime))
cat("Complete records with all key fields:", nrow(complete_records), "\n")
```

```
Complete records with all key fields: 11885
```

---

### Part 3: The Lagged Data Preparation Function

```r
prepare_lag_data <- function(master_df, wind_df, lag_minutes) {
  cat("Preparing data for", lag_minutes, "minute lag...\n")

  # Create time-shifted dataframes
  df_t <- master_df %>%
    filter(!is.na(datetime) & !is.na(view_id))

  df_t_lag <- df_t %>%
    rename(
      abundance_t_minus_1 = total_butterflies,
      datetime_t_minus_1 = datetime,
      butterflies_sun_t_minus_1 = butterflies_direct_sun,
      temperature_t_minus_1 = temperature
    ) %>%
```

```r
  select(view_id, datetime_t_minus_1, abundance_t_minus_1,
         butterflies_sun_t_minus_1, temperature_t_minus_1)

# Perform sliding window join
df_t_lag <- df_t_lag %>%
  mutate(datetime_expected = datetime_t_minus_1 + minutes(lag_minutes))

final_df <- df_t %>%
  left_join(
    df_t_lag,
    by = c("view_id" = "view_id", "datetime" = "datetime_expected")
  )

# Filter invalid pairs
final_df <- final_df %>%
  filter(!is.na(abundance_t_minus_1)) %>%
  filter(!(total_butterflies == 0 & abundance_t_minus_1 == 0))

cat("Valid pairs after filtering:", nrow(final_df), "\n")

if(nrow(final_df) == 0) {
  cat("No valid pairs found, returning empty dataframe\n")
  return(data.frame())
}

# More efficient calculation using data.table-style operations
cat("Calculating interval predictors efficiently...\n")

# Pre-sort and index the data for faster lookups
wind_sorted <- wind_df %>%
  arrange(wind_meter_name, time) %>%
  mutate(wind_id = row_number())

master_sorted <- master_df %>%
  arrange(view_id, datetime) %>%
  mutate(master_id = row_number())

# Calculate time delta first (fast operation)
final_df <- final_df %>%
  mutate(time_delta_mins = as.numeric(difftime(datetime, datetime_t_minus_1, units = "mins"

# Batch process by unique combinations to reduce redundant calculations
```

```r
unique_intervals <- final_df %>%
  select(view_id, wind_meter_name, datetime_t_minus_1, datetime) %>%
  distinct() %>%
  mutate(interval_id = row_number())

cat("Processing", nrow(unique_intervals), "unique intervals...\n")

# Initialize progress tracking
pb_intervals <- nrow(unique_intervals)

# Calculate metrics for each unique interval
interval_metrics <- list()

for(i in 1:nrow(unique_intervals)) {
  if(i %% max(1, floor(pb_intervals/10)) == 0) {
    cat("Progress:", round(100*i/pb_intervals, 1), "%\n")
  }

  interval <- unique_intervals[i, ]

  # Wind metrics for this interval
  wind_subset <- wind_sorted %>%
    filter(
      wind_meter_name == interval$wind_meter_name,
      time >= interval$datetime_t_minus_1,
      time <= interval$datetime
    )

  if (nrow(wind_subset) > 0) {
    wind_metrics <- list(
      mean_wind_speed = mean(wind_subset$speed, na.rm = TRUE),
      max_wind_speed = max(wind_subset$speed, na.rm = TRUE),
      sd_wind_speed = sd(wind_subset$speed, na.rm = TRUE),
      gust_factor = mean(wind_subset$gust, na.rm = TRUE) / mean(wind_subset$speed, na.rm =
      minutes_above_2mps = sum(wind_subset$speed > 2, na.rm = TRUE)
    )
  } else {
    wind_metrics <- list(
      mean_wind_speed = NA_real_,
      max_wind_speed = NA_real_,
      sd_wind_speed = NA_real_,
      gust_factor = NA_real_,
```

```
      minutes_above_2mps = NA_real_
    )
  }

  # Temperature and sunlight metrics for this interval
  temp_subset <- master_sorted %>%
    filter(
      view_id == interval$view_id,
      datetime >= interval$datetime_t_minus_1,
      datetime <= interval$datetime
    )

  if (nrow(temp_subset) > 0) {
    # Temperature metrics
    temp_metrics <- list(
      mean_temp = mean(temp_subset$temperature, na.rm = TRUE),
      max_temp = max(temp_subset$temperature, na.rm = TRUE),
      min_temp = min(temp_subset$temperature, na.rm = TRUE),
      sd_temp = sd(temp_subset$temperature, na.rm = TRUE)
    )

    # Sunlight metrics
    total_sun <- sum(temp_subset$butterflies_direct_sun, na.rm = TRUE)
    total_count <- sum(temp_subset$total_butterflies, na.rm = TRUE)
    if (total_count > 0) {
      overall_prop <- total_sun / total_count
      obs_props <- ifelse(temp_subset$total_butterflies > 0,
                          temp_subset$butterflies_direct_sun / temp_subset$total_butterflies
                          NA_real_)
      sun_metrics <- list(
        sunlight_proportion = overall_prop,
        max_prop_sunlight = max(obs_props, na.rm = TRUE),
        min_prop_sunlight = min(obs_props, na.rm = TRUE),
        sd_prop_sunlight = sd(obs_props, na.rm = TRUE)
      )
    } else {
      sun_metrics <- list(
        sunlight_proportion = NA_real_,
        max_prop_sunlight = NA_real_,
        min_prop_sunlight = NA_real_,
        sd_prop_sunlight = NA_real_
      )
```

```
    }
  } else {
    temp_metrics <- list(
      mean_temp = NA_real_,
      max_temp = NA_real_,
      min_temp = NA_real_,
      sd_temp = NA_real_
    )
    sun_metrics <- list(
      sunlight_proportion = NA_real_,
      max_prop_sunlight = NA_real_,
      min_prop_sunlight = NA_real_,
      sd_prop_sunlight = NA_real_
    )
  }

  # Store all metrics for this interval
  interval_metrics[[i]] <- c(
    interval_id = interval$interval_id,
    wind_metrics,
    temp_metrics,
    sun_metrics
  )
}

# Convert to dataframe
cat("Combining results...\n")
metrics_df <- do.call(rbind, lapply(interval_metrics, as.data.frame))

# Join back with original data
final_df_with_intervals <- final_df %>%
  left_join(unique_intervals, by = c("view_id", "wind_meter_name", "datetime_t_minus_1", "d
  left_join(metrics_df, by = "interval_id") %>%
  select(-interval_id)

# Remove rows with too many NAs in predictors
final_clean <- final_df_with_intervals %>%
  filter(!is.na(mean_wind_speed) & !is.na(mean_temp))

cat("Final dataset rows:", nrow(final_clean), "\n")
if(nrow(final_clean) > 0) {
  cat("Time delta range:", round(min(final_clean$time_delta_mins, na.rm = TRUE), 1),
```

```
        "to", round(max(final_clean$time_delta_mins, na.rm = TRUE), 1), "minutes\n")
  }

  return(final_clean)
}
```

---

## Part 4: Analysis and Modeling

### 30-Minute Lag Analysis

```
data_30m <- prepare_lag_data(master_df, wind, lag_minutes = 30)
```

```
Preparing data for 30 minute lag...
Valid pairs after filtering: 6888
Calculating interval predictors efficiently...
Processing 6888 unique intervals...
Progress: 10 %
Progress: 20 %
Progress: 30 %
Progress: 40 %
Progress: 49.9 %
Progress: 59.9 %
Progress: 69.9 %
Progress: 79.9 %
Progress: 89.9 %
Progress: 99.9 %
Combining results...
Final dataset rows: 5601
Time delta range: 30 to 30 minutes
```

### Exploratory Data Analysis (30 minutes)

```
cat("Data structure:\n")
```

```
Data structure:
```

```r
glimpse(data_30m)
```

```
Rows: 5,601
Columns: 30
$ deployment_id.x           <chr> "SC1", "SC1", "SC1", "SC1", "SC1", "SC1", "S~
$ image_filename            <chr> "SC1_20231117184501.JPG", "SC1_2023111718500~
$ total_butterflies         <dbl> 54, 54, 54, 54, 54, 54, 54, 54, 54, 54, 54, ~
$ butterflies_direct_sun    <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,~
$ datetime                  <dttm> 2023-11-17 18:45:01, 2023-11-17 18:50:01, 2~
$ view_id                   <fct> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1,~
$ wind_meter_name           <chr> "JazzPlay", "JazzPlay", "JazzPlay", "JazzPla~
$ deployment_id.y           <chr> "SC1", "SC1", "SC1", "SC1", "SC1", "SC1", "S~
$ timestamp                 <dbl> 2.023112e+13, 2.023112e+13, 2.023112e+13, 2.~
$ temperature               <dbl> 15, 16, 15, 16, 15, 16, 16, 15, 16, 15, 16, ~
$ confidence                <dbl> 0.4041873, 0.5208792, 0.6782328, 0.5311629, ~
$ extraction_status         <chr> "success", "success", "success", "success", ~
$ datetime_t_minus_1        <dttm> 2023-11-17 18:15:01, 2023-11-17 18:20:01, 2~
$ abundance_t_minus_1       <dbl> 54, 54, 54, 54, 54, 54, 54, 54, 54, 54, 54, ~
$ butterflies_sun_t_minus_1 <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,~
$ temperature_t_minus_1     <dbl> 15, 15, 15, 15, 15, 15, 15, 16, 15, 16, 15, ~
$ time_delta_mins           <dbl> 30, 30, 30, 30, 30, 30, 30, 30, 30, 30, 30, ~
$ mean_wind_speed           <dbl> 0.400000, 0.900000, 1.030769, 1.177778, 1.24~
$ max_wind_speed            <dbl> 0.4, 1.9, 1.9, 1.9, 1.9, 2.3, 2.7, 2.9, 2.9,~
$ sd_wind_speed             <dbl> 0.0000000, 0.6524678, 0.5297556, 0.5241863, ~
$ gust_factor               <dbl> 2.250000, 1.722222, 1.656716, 1.622642, 1.58~
$ minutes_above_2mps        <dbl> 0, 0, 0, 0, 0, 1, 3, 6, 7, 12, 16, 19, 22, 2~
$ mean_temp                 <dbl> 15.00000, 15.14286, 15.14286, 15.28571, 15.2~
$ max_temp                  <dbl> 15, 16, 16, 16, 16, 16, 16, 16, 16, 16, 16, ~
$ min_temp                  <dbl> 15, 15, 15, 15, 15, 15, 15, 15, 15, 15, 15, ~
$ sd_temp                   <dbl> 0.0000000, 0.3779645, 0.3779645, 0.4879500, ~
$ sunlight_proportion       <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,~
$ max_prop_sunlight         <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,~
$ min_prop_sunlight         <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,~
$ sd_prop_sunlight          <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,~
```

```r
cat("\n\nData summary:\n")
```

```
Data summary:
```

```r
summary(data_30m)
```

```
 deployment_id.x    image_filename     total_butterflies butterflies_direct_sun
 Length:5601        Length:5601        Min.   :  0.00    Min.   :  0.000
 Class :character   Class :character   1st Qu.: 14.00    1st Qu.:  0.000
 Mode  :character   Mode  :character   Median : 35.00    Median :  0.000
                                        Mean   : 66.52    Mean   :  1.726
                                        3rd Qu.: 73.00    3rd Qu.:  0.000
                                        Max.   :950.00    Max.   :150.000

    datetime                          view_id     wind_meter_name
 Min.   :2023-11-17 18:45:01.0   4      :2941   Length:5601
 1st Qu.:2023-12-11 22:00:01.0   2      :1039   Class :character
 Median :2023-12-30 03:50:01.0   1      : 640   Mode  :character
 Mean   :2023-12-29 12:47:03.9   5      : 444
 3rd Qu.:2024-01-23 16:50:01.0   11     : 288
 Max.   :2024-02-04 10:30:01.0   9      : 173
                                 (Other):  76
 deployment_id.y      timestamp          temperature     confidence
 Length:5601        Min.   :2.023e+13   Min.   : 4.0   Min.   :0.1222
 Class :character   1st Qu.:2.023e+13   1st Qu.:11.0   1st Qu.:0.4982
 Mode  :character   Median :2.023e+13   Median :13.0   Median :0.5951
                    Mean   :2.024e+13   Mean   :13.2   Mean   :0.5973
                    3rd Qu.:2.024e+13   3rd Qu.:15.0   3rd Qu.:0.6967
                    Max.   :2.024e+13   Max.   :33.0   Max.   :1.0000

 extraction_status  datetime_t_minus_1              abundance_t_minus_1
 Length:5601        Min.   :2023-11-17 18:15:01.0   Min.   :  0.00
 Class :character   1st Qu.:2023-12-11 21:30:01.0   1st Qu.: 14.00
 Mode  :character   Median :2023-12-30 03:20:01.0   Median : 35.00
                    Mean   :2023-12-29 12:17:03.9   Mean   : 66.69
                    3rd Qu.:2024-01-23 16:20:01.0   3rd Qu.: 73.00
                    Max.   :2024-02-04 10:00:01.0   Max.   :950.00

 butterflies_sun_t_minus_1 temperature_t_minus_1 time_delta_mins
 Min.   :  0.0             Min.   : 4.0          Min.   :30
 1st Qu.:  0.0             1st Qu.:11.0          1st Qu.:30
 Median :  0.0             Median :13.0          Median :30
 Mean   :  1.7             Mean   :13.2          Mean   :30
 3rd Qu.:  0.0             3rd Qu.:15.0          3rd Qu.:30
 Max.   :150.0             Max.   :33.0          Max.   :30
```

```
 mean_wind_speed  max_wind_speed  sd_wind_speed     gust_factor
 Min.   :0.0000   Min.   :0.000   Min.   :0.0000   Min.   :1.149
 1st Qu.:0.1567   1st Qu.:0.700   1st Qu.:0.1029   1st Qu.:1.346
 Median :0.7700   Median :1.200   Median :0.2312   Median :1.441
 Mean   :0.8902   Mean   :1.378   Mean   :0.2383   Mean   :  Inf
 3rd Qu.:1.3400   3rd Qu.:2.000   3rd Qu.:0.3354   3rd Qu.:1.590
 Max.   :5.3333   Max.   :6.700   Max.   :1.5024   Max.   :  Inf
                                  NA's   :1        NA's   :890
 minutes_above_2mps   mean_temp       max_temp       min_temp
 Min.   : 0.000     Min.   : 4.0   Min.   : 4.0   Min.   : 4.00
 1st Qu.: 0.000     1st Qu.:11.0   1st Qu.:11.0   1st Qu.:11.00
 Median : 0.000     Median :13.0   Median :13.0   Median :13.00
 Mean   : 2.949     Mean   :13.2   Mean   :13.5   Mean   :12.89
 3rd Qu.: 0.000     3rd Qu.:15.5   3rd Qu.:16.0   3rd Qu.:15.00
 Max.   :30.000     Max.   :33.0   Max.   :33.0   Max.   :33.00

    sd_temp       sunlight_proportion max_prop_sunlight min_prop_sunlight
 Min.   :0.0000   Min.   :0.00000     Min.   :0.000     Min.   :0.00000
 1st Qu.:0.0000   1st Qu.:0.00000     1st Qu.:0.000     1st Qu.:0.00000
 Median :0.0000   Median :0.00000     Median :0.000     Median :0.00000
 Mean   :0.3613   Mean   :0.07344     Mean   :0.104     Mean   :0.05038
 3rd Qu.:0.7071   3rd Qu.:0.00000     3rd Qu.:0.000     3rd Qu.:0.00000
 Max.   :4.9497   Max.   :1.00000     Max.   :1.000     Max.   :1.00000

 sd_prop_sunlight
 Min.   :0.00000
 1st Qu.:0.00000
 Median :0.00000
 Mean   :0.03026
 3rd Qu.:0.00000
 Max.   :0.70711
 NA's   :60
```

```r
# Check for data quality issues
cat("\n\nData quality checks:\n")
```

```
Data quality checks:
```

```r
cat("Total rows:", nrow(data_30m), "\n")
```

```
Total rows: 5601
```

```r
cat("Complete cases:", sum(complete.cases(data_30m)), "\n")
```

```
Complete cases: 4679
```

```r
cat("Rows with abundance_t_minus_1 > 0:", sum(data_30m$abundance_t_minus_1 > 0, na.rm = TRUE)
```

```
Rows with abundance_t_minus_1 > 0: 5566
```

```r
cat("Rows with total_butterflies > 0:", sum(data_30m$total_butterflies > 0, na.rm = TRUE), "
```

```
Rows with total_butterflies > 0: 5547
```

```r
cat("Unique view_ids:", length(unique(data_30m$view_id)), "\n")
```

```
Unique view_ids: 9
```

```r
# Check response variable distribution
cat("\nResponse variable (total_butterflies) distribution:\n")
```

```

Response variable (total_butterflies) distribution:
```

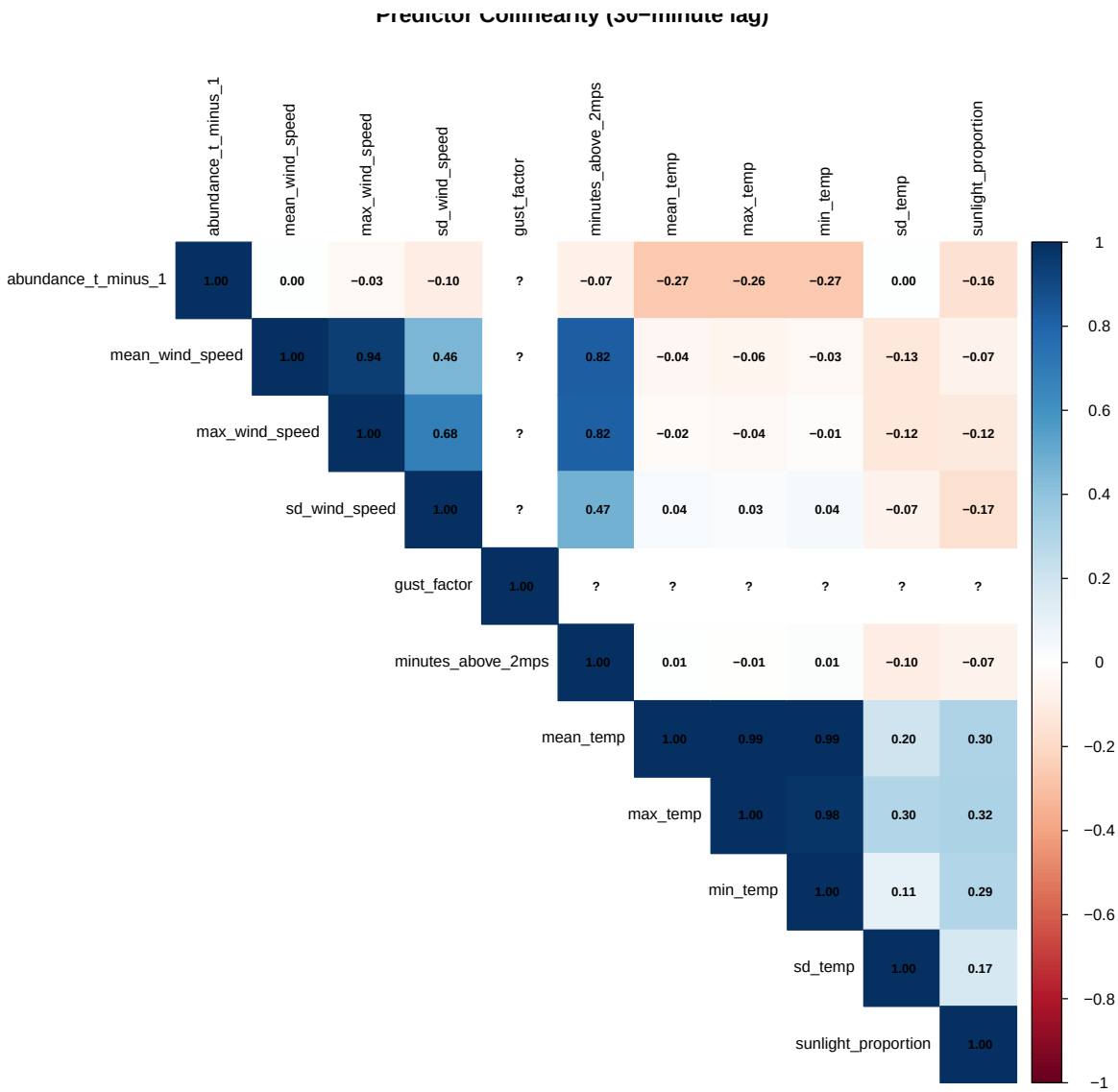
```r
table(data_30m$total_butterflies)
```

```
  0   1   2   3   4   5   6   7   8   9  10  11  12  13  14  15  16  17  18  19
 54 125 188 225 109 146 103  43  78  65  24  54  57  65 111 197 164  71  39 171
 20  21  22  23  24  25  26  27  28  29  30  31  32  33  34  35  36  37  38  39
 13  43  20  52 134  30  89  14  14   8  12  16  16  19 194 142  16  61   6  35
 40  41  42  43  44  45  46  47  48  49  50  51  52  53  54  55  56  57  58  59
 80  38  17  14  36  22  18  19  66  12  44  55 168  11 149  68  10  30   6   7
 60  61  62  63  64  65  66  67  68  69  70  71  72  73  74  75  76  77  78  79
  2  11   6  10   5   5   8 166   4   4   3   3  80   9   5  61   9   4  32   8
```

```
 80  81  82  83  84  85  86  87  88  89  90  91  92  93  94  95  96  97  98  99
  3   5   5   6   6  33   4   5   3   1 136  32   7   6  10   6   4   3  29   6
100 101 102 103 104 105 106 108 109 110 111 112 113 114 115 116 117 118 119 120
  3   5   4  10   5   8   7   7  28   1   5   1   3   1   3   1   2   3   4   4
121 123 124 125 126 127 128 129 130 131 132 133 134 135 137 139 140 141 143 144
  2   2   2   1   2   1   1  58   5   6   1   3   1   4   1   2   2   3   2   1
146 148 149 150 152 153 157 159 160 162 164 165 168 170 174 184 185 187 188 191
  2   1   2   1   4   4   2   1   1   1   3   1  26   1   1   1   1   1   1   1
192 193 196 197 198 199 200 201 202 203 204 205 206 207 208 209 210 211 212 213
  2   3   4   3   2   2   1   2  10   9  26   3  29   1   2   4  33  33  36  14
214 215 216 217 218 222 223 225 226 227 228 229 230 231 232 233 234 236 237 238
  1   3   3   2   1   2   5  30   3   1   2   4   1   3   4   9  12  57   4   2
242 243 244 245 247 248 249 250 251 252 253 254 255 256 257 262 263 264 266 267
 34  32   2   4   3   2   4   4   2  32  31   7   2   2   2   1  30   2   1   1
268 269 270 271 272 274 276 284 285 286 288 293 315 318 322 329 337 352 371 372
  1   1   2  28   1   1   3  30   1   2   3   1   1   1   1   1   1   1   1   1
374 497 499 528 551 569 619 740 950
  1   1  25   1   1   1   1   1  22
```

```r
# Select numeric predictors for correlation analysis
predictors_30m <- data_30m %>%
  select(abundance_t_minus_1, mean_wind_speed, max_wind_speed, sd_wind_speed,
         gust_factor, minutes_above_2mps, mean_temp, max_temp, min_temp, sd_temp,
         sunlight_proportion) %>%
  na.omit()

if (nrow(predictors_30m) > 1) {
  # Calculate correlation matrix
  cor_matrix_30m <- cor(predictors_30m, use = "complete.obs")

  # Check if correlation matrix has any NA/NaN/Inf values
  if (!any(is.na(cor_matrix_30m) | is.infinite(cor_matrix_30m))) {
    corrplot(cor_matrix_30m, method = "color", type = "upper",
             order = "hclust", tl.cex = 0.8, tl.col = "black",
             addCoef.col = "black", number.cex = 0.7,
             title = "Predictor Collinearity (30-minute lag)")
  } else {
    # If hclust fails, use original order
    corrplot(cor_matrix_30m, method = "color", type = "upper",
             order = "original", tl.cex = 0.8, tl.col = "black",
             addCoef.col = "black", number.cex = 0.7,
             title = "Predictor Collinearity (30-minute lag)")
```

```
  }
} else {
  cat("Not enough data for correlation analysis\n")
}
```



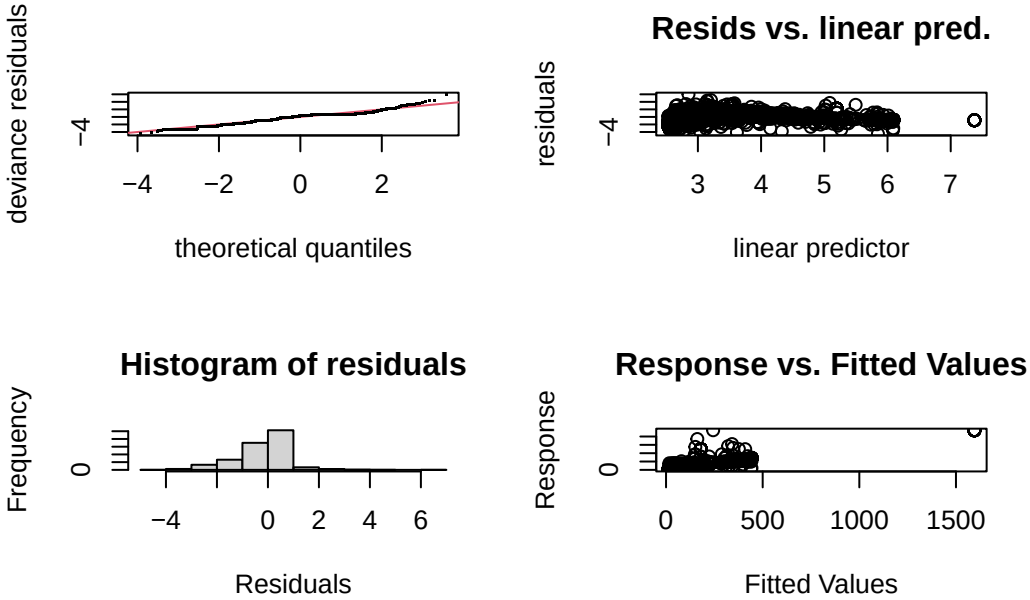
### Model Fitting (30 minutes)

```
# Define model formulas
model_formulas_30m <- list(
  "Baseline (abundance only)" = total_butterflies ~ s(abundance_t_minus_1, k = 4),
  "Wind speed" = total_butterflies ~ s(abundance_t_minus_1, k = 4) + s(mean_wind_speed, k = 4
  "Temperature" = total_butterflies ~ s(abundance_t_minus_1, k = 4) + s(mean_temp, k = 4),
  "Wind + Temperature" = total_butterflies ~ s(abundance_t_minus_1, k = 4) +
    s(mean_wind_speed, k = 4) + s(mean_temp, k = 4),
  "Wind + Temp + Sunlight" = total_butterflies ~ s(abundance_t_minus_1, k = 4) +
    s(mean_wind_speed, k = 4) + s(mean_temp, k = 4) + s(sunlight_proportion, k = 4),
  "Full model" = total_butterflies ~ s(abundance_t_minus_1, k = 4) +
    s(mean_wind_speed, k = 4) + s(mean_temp, k = 4) +
    s(sunlight_proportion, k = 4) + s(minutes_above_2mps, k = 4)
)

# Storage for results
models_30m <- list()
aic_values_30m <- numeric()

# Fit each model
for (i in seq_along(model_formulas_30m)) {
  model_name <- names(model_formulas_30m)[i]
  formula <- model_formulas_30m[[i]]

  cat("\n### Model", i, ":", model_name, "\n\n")

  tryCatch({
    # Check data before fitting
    model_vars <- all.vars(formula)
    model_data <- data_30m %>%
      select(all_of(model_vars), view_id) %>%
      na.omit()

    cat("Model data dimensions after removing NAs:", nrow(model_data), "x", ncol(model_data)
    cat("Variables in model:", paste(model_vars, collapse = ", "), "\n")

    # Check if we have enough data
    if(nrow(model_data) < 50) {
      cat("Warning: Very few observations (", nrow(model_data), ") for this model\n")
    }
```

```r
    # Check for sufficient variation in view_id
    view_counts <- table(model_data$view_id)
    cat("View IDs with observations:", length(view_counts), "\n")
    cat("Min/Max observations per view:", min(view_counts), "/", max(view_counts), "\n")

    # Fit GAM (simpler and faster than GAMM)
    gam_model <- gam(
      formula,
      family = nb(),
      data = model_data
    )

    cat("GAM fitted successfully\n")

    # Store model
    models_30m[[model_name]] <- gam_model
    aic_values_30m[model_name] <- AIC(gam_model)

    # Print summary
    cat("Model Summary:\n")
    print(summary(gam_model))

    # Diagnostic plots
    cat("\n\nDiagnostic plots:\n")
    gam.check(gam_model)

  }, error = function(e) {
    cat("Error fitting model:", e$message, "\n")
    cat("This might be due to insufficient data or convergence issues\n")
    aic_values_30m[model_name] <- NA
  })
}
```

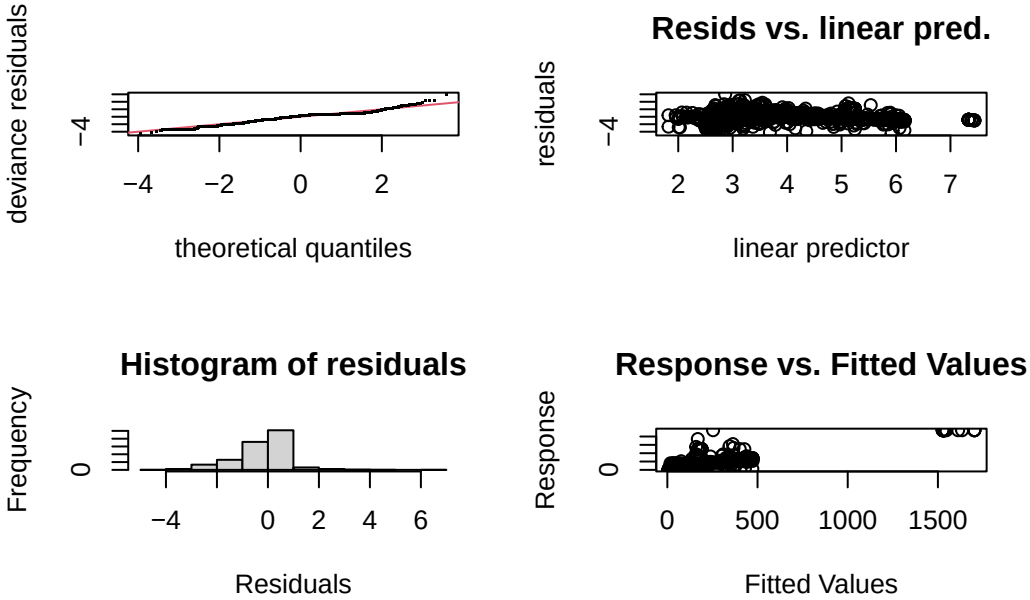
```
### Model 1 : Baseline (abundance only)

Model data dimensions after removing NAs: 5601 x 3
Variables in model: total_butterflies, abundance_t_minus_1
View IDs with observations: 9
Min/Max observations per view: 13 / 2941
GAM fitted successfully
Model Summary:
```

```
Family: Negative Binomial(3.394)
Link function: log

Formula:
total_butterflies ~ s(abundance_t_minus_1, k = 4)

Parametric coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) 3.601266   0.007723   466.3   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Approximate significance of smooth terms:
                         edf Ref.df Chi.sq p-value
s(abundance_t_minus_1) 2.998      3  20801  <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

R-sq.(adj) =  0.441   Deviance explained =   77%
-REML =  24558  Scale est. = 1         n = 5601


Diagnostic plots:
```



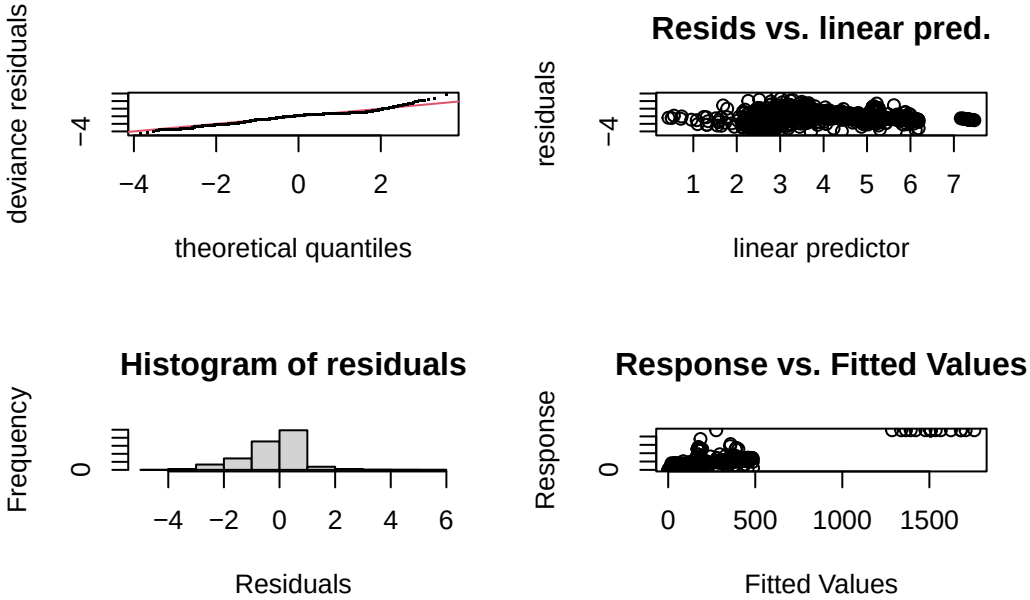

```
Method: REML   Optimizer: outer newton
full convergence after 7 iterations.
Gradient range [-0.0001256606,0.001236955]
(score 24558.26 & scale 1).
Hessian positive definite, eigenvalue range [0.9995636,1897.94].
Model rank =  4 / 4

Basis dimension (k) checking results. Low p-value (k-index<1) may
indicate that k is too low, especially if edf is close to k'.

                       k' edf k-index p-value
s(abundance_t_minus_1)  3   3    0.23  <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

### Model 2 : Wind speed

Model data dimensions after removing NAs: 5601 x 4
Variables in model: total_butterflies, abundance_t_minus_1, mean_wind_speed
View IDs with observations: 9
Min/Max observations per view: 13 / 2941
GAM fitted successfully
```

```
Model Summary:

Family: Negative Binomial(3.441)
Link function: log

Formula:
total_butterflies ~ s(abundance_t_minus_1, k = 4) + s(mean_wind_speed,
    k = 4)

Parametric coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) 3.599313   0.007679   468.7   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Approximate significance of smooth terms:
                         edf Ref.df   Chi.sq p-value
s(abundance_t_minus_1) 2.998  3.000 20617.92  <2e-16 ***
s(mean_wind_speed)     2.937  2.997    61.31  <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

R-sq.(adj) =  0.462   Deviance explained = 77.3%
-REML =  24536  Scale est. = 1         n = 5601


Diagnostic plots:
```



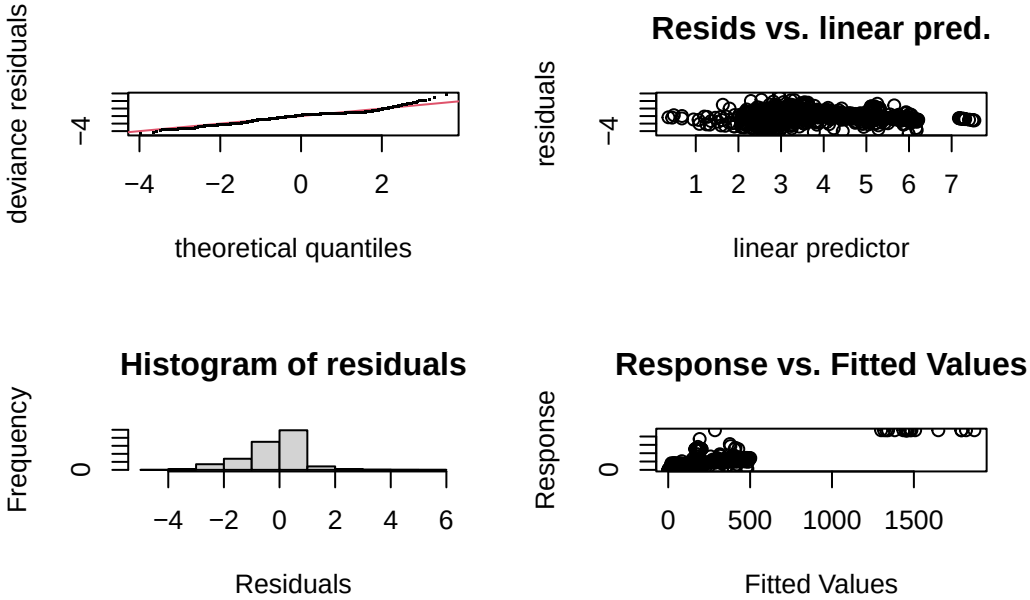

```
Method: REML   Optimizer: outer newton
full convergence after 8 iterations.
Gradient range [-5.585903e-05,0.0005515457]
(score 24536.02 & scale 1).
Hessian positive definite, eigenvalue range [0.8927119,1882.683].
Model rank =  7 / 7

Basis dimension (k) checking results. Low p-value (k-index<1) may
indicate that k is too low, especially if edf is close to k'.

                         k'  edf k-index p-value
s(abundance_t_minus_1) 3.00 3.00    0.25  <2e-16 ***
s(mean_wind_speed)     3.00 2.94    0.96    0.35
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

### Model 3 : Temperature

Model data dimensions after removing NAs: 5601 x 4
Variables in model: total_butterflies, abundance_t_minus_1, mean_temp
View IDs with observations: 9
Min/Max observations per view: 13 / 2941
```

```
GAM fitted successfully
Model Summary:

Family: Negative Binomial(3.548)
Link function: log

Formula:
total_butterflies ~ s(abundance_t_minus_1, k = 4) + s(mean_temp,
    k = 4)

Parametric coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) 3.594556   0.007588   473.7   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Approximate significance of smooth terms:
                         edf Ref.df  Chi.sq p-value
s(abundance_t_minus_1) 2.998  3.000 20367.5  <2e-16 ***
s(mean_temp)           2.963  2.999   210.9  <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

R-sq.(adj) =  0.509   Deviance explained = 77.9%
-REML =  24464  Scale est. = 1         n = 5601


Diagnostic plots:
```



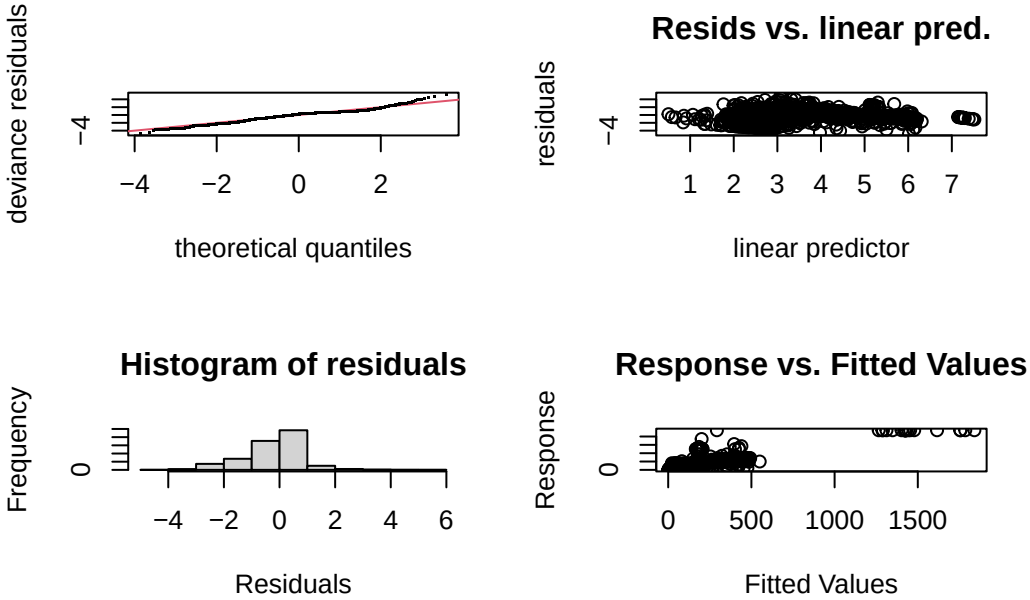

```
Method: REML   Optimizer: outer newton
full convergence after 7 iterations.
Gradient range [-0.0001647457,0.001617309]
(score 24463.81 & scale 1).
Hessian positive definite, eigenvalue range [0.9513535,1857.539].
Model rank =  7 / 7

Basis dimension (k) checking results. Low p-value (k-index<1) may
indicate that k is too low, especially if edf is close to k'.

                         k'  edf k-index p-value
s(abundance_t_minus_1) 3.00 3.00    0.27  <2e-16 ***
s(mean_temp)           3.00 2.96    0.90  <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

### Model 4 : Wind + Temperature

Model data dimensions after removing NAs: 5601 x 5
Variables in model: total_butterflies, abundance_t_minus_1, mean_wind_speed, mean_temp
View IDs with observations: 9
Min/Max observations per view: 13 / 2941
```

```
GAM fitted successfully
Model Summary:

Family: Negative Binomial(3.593)
Link function: log

Formula:
total_butterflies ~ s(abundance_t_minus_1, k = 4) + s(mean_wind_speed,
    k = 4) + s(mean_temp, k = 4)

Parametric coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  3.59283    0.00755   475.9   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Approximate significance of smooth terms:
                         edf Ref.df   Chi.sq p-value
s(abundance_t_minus_1) 2.998  3.000 20180.79  <2e-16 ***
s(mean_wind_speed)     2.899  2.992    54.53  <2e-16 ***
s(mean_temp)           2.963  2.999   207.17  <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

R-sq.(adj) =  0.516   Deviance explained = 78.1%
-REML =  24444  Scale est. = 1         n = 5601


Diagnostic plots:
```



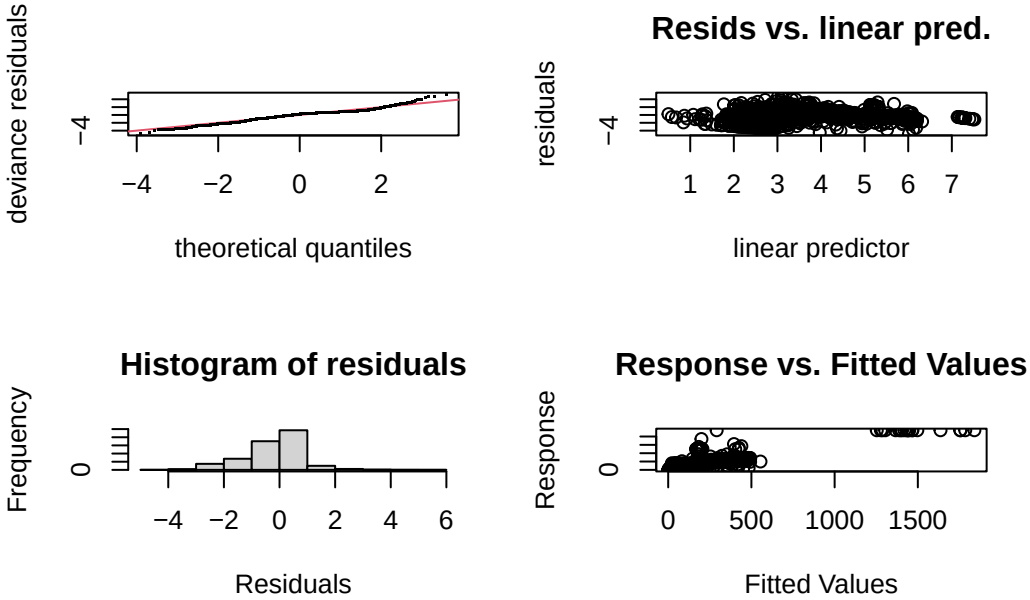

```
Method: REML   Optimizer: outer newton
full convergence after 8 iterations.
Gradient range [-0.0007605842,0.009544005]
(score 24443.8 & scale 1).
Hessian positive definite, eigenvalue range [0.8268327,1843.985].
Model rank =  10 / 10

Basis dimension (k) checking results. Low p-value (k-index<1) may
indicate that k is too low, especially if edf is close to k'.

                         k'  edf k-index p-value
s(abundance_t_minus_1) 3.00 3.00    0.27  <2e-16 ***
s(mean_wind_speed)     3.00 2.90    0.95    0.25
s(mean_temp)           3.00 2.96    0.90  <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

### Model 5 : Wind + Temp + Sunlight

Model data dimensions after removing NAs: 5601 x 6
Variables in model: total_butterflies, abundance_t_minus_1, mean_wind_speed, mean_temp, sunli
View IDs with observations: 9
```

```
Min/Max observations per view: 13 / 2941
GAM fitted successfully
Model Summary:

Family: Negative Binomial(3.723)
Link function: log

Formula:
total_butterflies ~ s(abundance_t_minus_1, k = 4) + s(mean_wind_speed,
    k = 4) + s(mean_temp, k = 4) + s(sunlight_proportion, k = 4)

Parametric coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) 3.588154   0.007442   482.2   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Approximate significance of smooth terms:
                         edf Ref.df   Chi.sq p-value
s(abundance_t_minus_1) 2.998  3.000 20026.80  <2e-16 ***
s(mean_wind_speed)     2.911  2.994    54.71  <2e-16 ***
s(mean_temp)           2.965  2.999   220.26  <2e-16 ***
s(sunlight_proportion) 2.918  2.994   167.94  <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

R-sq.(adj) =   0.55   Deviance explained = 78.7%
-REML =  24371  Scale est. = 1         n = 5601


Diagnostic plots:
```




```
Method: REML   Optimizer: outer newton
full convergence after 9 iterations.
Gradient range [-3.65158e-05,0.0003453744]
(score 24370.91 & scale 1).
Hessian positive definite, eigenvalue range [0.8462144,1812.133].
Model rank =  13 / 13

Basis dimension (k) checking results. Low p-value (k-index<1) may
indicate that k is too low, especially if edf is close to k'.

                         k'  edf k-index p-value
s(abundance_t_minus_1) 3.00 3.00    0.30  <2e-16 ***
s(mean_wind_speed)     3.00 2.91    0.96    0.44
s(mean_temp)           3.00 2.96    0.95    0.15
s(sunlight_proportion) 3.00 2.92    0.90  <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

### Model 6 : Full model

Model data dimensions after removing NAs: 5601 x 7
Variables in model: total_butterflies, abundance_t_minus_1, mean_wind_speed, mean_temp, sunli
```

```
View IDs with observations: 9
Min/Max observations per view: 13 / 2941
GAM fitted successfully
Model Summary:

Family: Negative Binomial(3.724)
Link function: log

Formula:
total_butterflies ~ s(abundance_t_minus_1, k = 4) + s(mean_wind_speed,
    k = 4) + s(mean_temp, k = 4) + s(sunlight_proportion, k = 4) +
    s(minutes_above_2mps, k = 4)

Parametric coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) 3.588079   0.007441   482.2   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Approximate significance of smooth terms:
                         edf Ref.df    Chi.sq  p-value
s(abundance_t_minus_1) 2.998  3.000 19966.525  < 2e-16 ***
s(mean_wind_speed)     2.343  2.659    36.285 5.02e-07 ***
s(mean_temp)           2.965  2.999   219.517  < 2e-16 ***
s(sunlight_proportion) 2.918  2.994   166.786  < 2e-16 ***
s(minutes_above_2mps)  1.002  1.005     8.689  0.00325 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

R-sq.(adj) =  0.546   Deviance explained = 78.7%
-REML =  24371  Scale est. = 1         n = 5601


Diagnostic plots:
```




```
Method: REML   Optimizer: outer newton
full convergence after 8 iterations.
Gradient range [-0.0006182826,0.0007920954]
(score 24370.82 & scale 1).
Hessian positive definite, eigenvalue range [0.0006164027,1812.22].
Model rank =  16 / 16

Basis dimension (k) checking results. Low p-value (k-index<1) may
indicate that k is too low, especially if edf is close to k'.

                         k'  edf k-index p-value
s(abundance_t_minus_1) 3.00 3.00    0.28  <2e-16 ***
s(mean_wind_speed)     3.00 2.34    0.95    0.09 .
s(mean_temp)           3.00 2.96    0.90  <2e-16 ***
s(sunlight_proportion) 3.00 2.92    0.89  <2e-16 ***
s(minutes_above_2mps)  3.00 1.00    0.97    0.69
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
# Model comparison table
comparison_30m <- data.frame(
```

```r
  Model = names(aic_values_30m),
  AIC = round(aic_values_30m, 2)
) %>%
  arrange(AIC)

cat("\n### Model Comparison (30-minute lag)\n")
```

```
### Model Comparison (30-minute lag)
```

```r
kable(comparison_30m)
```

| | Model | AIC |
|---|---|---|
| Full model | Full model | 48682.07 |
| Wind + Temp + Sunlight | Wind + Temp + Sunlight | 48683.39 |
| Wind + Temperature | Wind + Temperature | 48840.75 |
| Temperature | Temperature | 48891.39 |
| Wind speed | Wind speed | 49037.06 |
| Baseline (abundance only) | Baseline (abundance only) | 49093.12 |

---

### 1-Hour Lag Analysis

```r
data_60m <- prepare_lag_data(master_df, wind, lag_minutes = 60)
```

```
Preparing data for 60 minute lag...
Valid pairs after filtering: 6895
Calculating interval predictors efficiently...
Processing 6895 unique intervals...
Progress: 10 %
Progress: 20 %
Progress: 30 %
Progress: 40 %
Progress: 50 %
Progress: 60 %
Progress: 69.9 %
```

```
Progress: 79.9 %
Progress: 89.9 %
Progress: 99.9 %
Combining results...
Final dataset rows: 5618
Time delta range: 60 to 60 minutes
```

### Exploratory Data Analysis (1 hour)

```r
cat("Data structure:\n")
```

```
Data structure:
```

```r
glimpse(data_60m)
```

```
Rows: 5,618
Columns: 30
$ deployment_id.x           <chr> "SC1", "SC1", "SC1", "SC1", "SC1", "SC1", "S~
$ image_filename            <chr> "SC1_20231117184501.JPG", "SC1_2023111718500~
$ total_butterflies         <dbl> 54, 54, 54, 54, 54, 54, 54, 54, 54, 54, 54, ~
$ butterflies_direct_sun    <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,~
$ datetime                  <dttm> 2023-11-17 18:45:01, 2023-11-17 18:50:01, 2~
$ view_id                   <fct> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1,~
$ wind_meter_name           <chr> "JazzPlay", "JazzPlay", "JazzPlay", "JazzPla~
$ deployment_id.y           <chr> "SC1", "SC1", "SC1", "SC1", "SC1", "SC1", "S~
$ timestamp                 <dbl> 2.023112e+13, 2.023112e+13, 2.023112e+13, 2.~
$ temperature               <dbl> 15, 16, 15, 16, 15, 16, 16, 15, 16, 15, 16, ~
$ confidence                <dbl> 0.4041873, 0.5208792, 0.6782328, 0.5311629, ~
$ extraction_status         <chr> "success", "success", "success", "success", ~
$ datetime_t_minus_1        <dttm> 2023-11-17 17:45:01, 2023-11-17 17:50:01, 2~
$ abundance_t_minus_1       <dbl> 54, 54, 54, 54, 54, 54, 54, 54, 54, 54, 54, ~
$ butterflies_sun_t_minus_1 <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,~
$ temperature_t_minus_1     <dbl> 15, 15, 15, 15, 15, 15, 15, 15, 15, 15, 15, ~
$ time_delta_mins           <dbl> 60, 60, 60, 60, 60, 60, 60, 60, 60, 60, 60, ~
$ mean_wind_speed           <dbl> 0.400000, 0.900000, 1.030769, 1.177778, 1.24~
$ max_wind_speed            <dbl> 0.4, 1.9, 1.9, 1.9, 1.9, 2.3, 2.7, 2.9, 2.9,~
$ sd_wind_speed             <dbl> 0.0000000, 0.6524678, 0.5297556, 0.5241863, ~
$ gust_factor               <dbl> 2.250000, 1.722222, 1.656716, 1.622642, 1.58~
$ minutes_above_2mps        <dbl> 0, 0, 0, 0, 0, 1, 3, 6, 7, 12, 16, 20, 25, 3~
$ mean_temp                 <dbl> 15.00000, 15.07692, 15.07692, 15.15385, 15.1~
```

```
$ max_temp                  <dbl> 15, 16, 16, 16, 16, 16, 16, 16, 16, 16, 16, ~
$ min_temp                  <dbl> 15, 15, 15, 15, 15, 15, 15, 15, 15, 15, 15, ~
$ sd_temp                   <dbl> 0.0000000, 0.2773501, 0.2773501, 0.3755338, ~
$ sunlight_proportion       <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,~
$ max_prop_sunlight         <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,~
$ min_prop_sunlight         <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,~
$ sd_prop_sunlight          <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,~
```

```r
cat("\n\nData summary:\n")
```

```
Data summary:
```

```r
summary(data_60m)
```

```
 deployment_id.x    image_filename     total_butterflies butterflies_direct_sun
 Length:5618        Length:5618        Min.   :  0.00    Min.   :  0.000
 Class :character   Class :character   1st Qu.: 14.00    1st Qu.:  0.000
 Mode  :character   Mode  :character   Median : 35.00    Median :  0.000
                                       Mean   : 66.13    Mean   :  1.688
                                       3rd Qu.: 72.00    3rd Qu.:  0.000
                                       Max.   :950.00    Max.   :150.000

    datetime                          view_id     wind_meter_name
 Min.   :2023-11-17 18:45:01.00   4      :2931   Length:5618
 1st Qu.:2023-12-11 14:07:31.00   2      :1055   Class :character
 Median :2023-12-29 19:05:01.00   1      : 648   Mode  :character
 Mean   :2023-12-29 09:10:30.23   5      : 449
 3rd Qu.:2024-01-23 13:12:31.00   11     : 287
 Max.   :2024-02-04 10:50:01.00   9      : 175
                                  (Other):  73
 deployment_id.y      timestamp          temperature      confidence
 Length:5618        Min.   :2.023e+13   Min.   : 4.00   Min.   :0.1222
 Class :character   1st Qu.:2.023e+13   1st Qu.:11.00   1st Qu.:0.4982
 Mode  :character   Median :2.023e+13   Median :13.00   Median :0.5949
                    Mean   :2.024e+13   Mean   :13.22   Mean   :0.5971
                    3rd Qu.:2.024e+13   3rd Qu.:15.00   3rd Qu.:0.6962
                    Max.   :2.024e+13   Max.   :33.00   Max.   :1.0000

 extraction_status  datetime_t_minus_1               abundance_t_minus_1
```

```
Length:5618        Min.   :2023-11-17 17:45:01.00   Min.   :  0.00
Class :character   1st Qu.:2023-12-11 13:07:31.00   1st Qu.: 14.00
Mode  :character   Median :2023-12-29 18:05:01.00   Median : 35.00
                   Mean   :2023-12-29 08:10:30.23   Mean   : 66.44
                   3rd Qu.:2024-01-23 12:12:31.00   3rd Qu.: 73.00
                   Max.   :2024-02-04 09:50:01.00   Max.   :950.00

butterflies_sun_t_minus_1 temperature_t_minus_1 time_delta_mins
Min.   :  0.000           Min.   : 4.00         Min.   :60
1st Qu.:  0.000           1st Qu.:11.00         1st Qu.:60
Median :  0.000           Median :13.00         Median :60
Mean   :  1.712           Mean   :13.23         Mean   :60
3rd Qu.:  0.000           3rd Qu.:15.00         3rd Qu.:60
Max.   :150.000           Max.   :33.00         Max.   :60

mean_wind_speed  max_wind_speed  sd_wind_speed     gust_factor
Min.   :0.0000   Min.   :0.000   Min.   :0.0000   Min.   :1.166
1st Qu.:0.1871   1st Qu.:0.800   1st Qu.:0.1436   1st Qu.:1.352
Median :0.7608   Median :1.300   Median :0.2701   Median :1.444
Mean   :0.8873   Mean   :1.526   Mean   :0.2783   Mean   :  Inf
3rd Qu.:1.3113   3rd Qu.:2.100   3rd Qu.:0.3822   3rd Qu.:1.586
Max.   :5.2900   Max.   :6.700   Max.   :1.3876   Max.   :  Inf
                                 NA's   :1        NA's   :687
minutes_above_2mps   mean_temp        max_temp        min_temp
Min.   : 0.00      Min.   : 4.00   Min.   : 4.00   Min.   : 4.00
1st Qu.: 0.00      1st Qu.:11.00   1st Qu.:11.00   1st Qu.:11.00
Median : 0.00      Median :13.00   Median :13.00   Median :12.00
Mean   : 5.86      Mean   :13.22   Mean   :13.77   Mean   :12.69
3rd Qu.: 1.00      3rd Qu.:15.38   3rd Qu.:16.00   3rd Qu.:15.00
Max.   :60.00      Max.   :32.14   Max.   :33.00   Max.   :30.00

   sd_temp       sunlight_proportion max_prop_sunlight min_prop_sunlight
Min.   :0.0000   Min.   :0.000000    Min.   :0.00000   Min.   :0.0000
1st Qu.:0.0000   1st Qu.:0.000000    1st Qu.:0.00000   1st Qu.:0.0000
Median :0.5345   Median :0.000000    Median :0.00000   Median :0.0000
Mean   :0.5121   Mean   :0.073082    Mean   :0.12364   Mean   :0.0415
3rd Qu.:0.5774   3rd Qu.:0.004453    3rd Qu.:0.01923   3rd Qu.:0.0000
Max.   :5.5076   Max.   :1.000000    Max.   :1.00000   Max.   :1.0000

sd_prop_sunlight
Min.   :0.00000
1st Qu.:0.00000
Median :0.00000
```

```
 Mean   :0.03808
 3rd Qu.:0.00000
 Max.   :0.70711
 NA's   :49
```

```r
# Check for data quality issues
cat("\n\nData quality checks:\n")
```

```
Data quality checks:
```

```r
cat("Total rows:", nrow(data_60m), "\n")
```

```
Total rows: 5618
```

```r
cat("Complete cases:", sum(complete.cases(data_60m)), "\n")
```

```
Complete cases: 4898
```

```r
cat("Rows with abundance_t_minus_1 > 0:", sum(data_60m$abundance_t_minus_1 > 0, na.rm = TRUE
```

```
Rows with abundance_t_minus_1 > 0: 5560
```

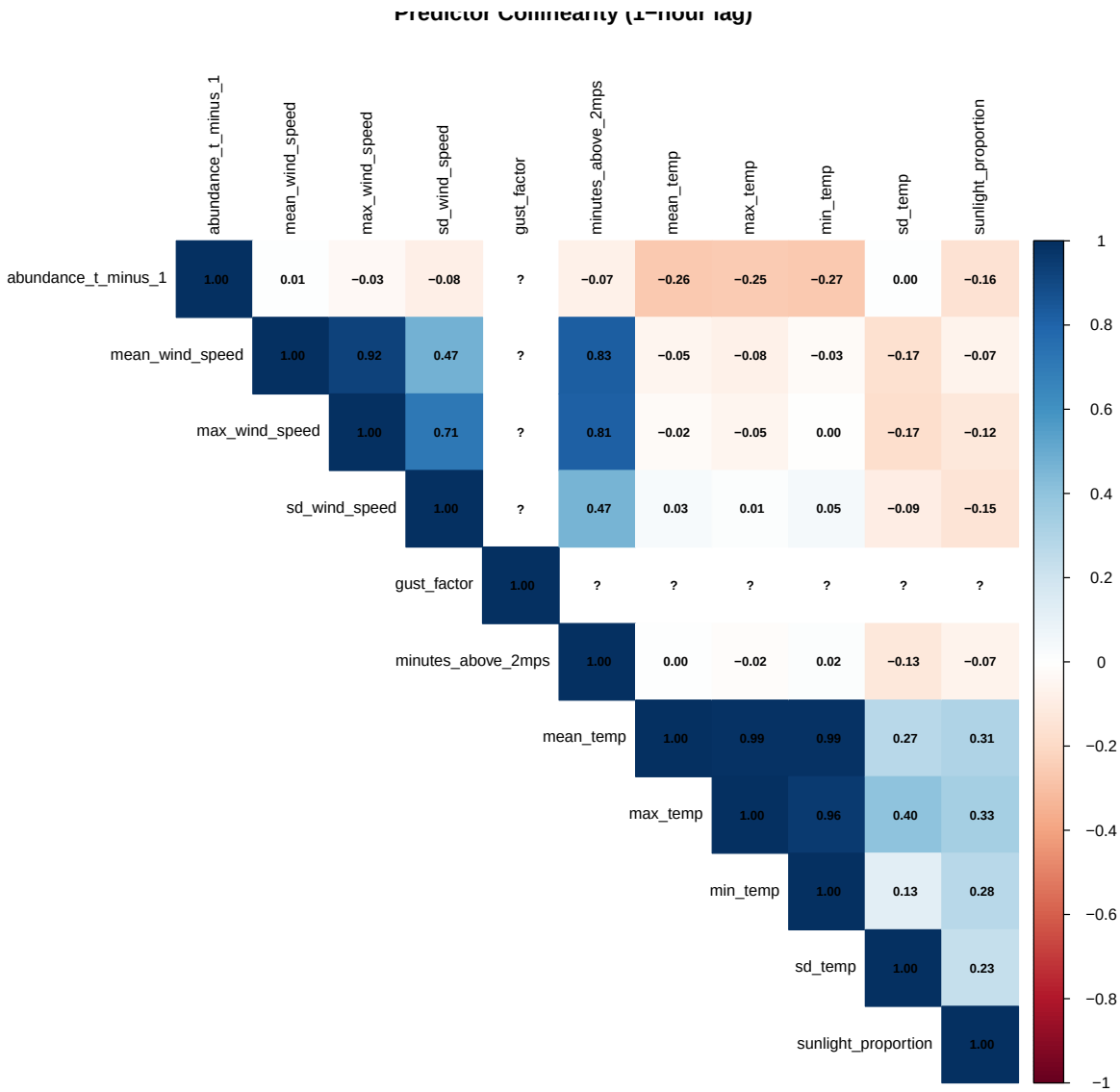
```r
cat("Rows with total_butterflies > 0:", sum(data_60m$total_butterflies > 0, na.rm = TRUE), "
```

```
Rows with total_butterflies > 0: 5529
```

```r
cat("Unique view_ids:", length(unique(data_60m$view_id)), "\n")
```

```
Unique view_ids: 9
```

```r
# Select numeric predictors for correlation analysis
predictors_60m <- data_60m %>%
  select(abundance_t_minus_1, mean_wind_speed, max_wind_speed, sd_wind_speed,
         gust_factor, minutes_above_2mps, mean_temp, max_temp, min_temp, sd_temp,
         sunlight_proportion) %>%
  na.omit()

if (nrow(predictors_60m) > 1) {
  cor_matrix_60m <- cor(predictors_60m, use = "complete.obs")
  if (!any(is.na(cor_matrix_60m) | is.infinite(cor_matrix_60m))) {
    corrplot(cor_matrix_60m, method = "color", type = "upper",
             order = "hclust", tl.cex = 0.8, tl.col = "black",
             addCoef.col = "black", number.cex = 0.7,
             title = "Predictor Collinearity (1-hour lag)")
  } else {
    corrplot(cor_matrix_60m, method = "color", type = "upper",
             order = "original", tl.cex = 0.8, tl.col = "black",
             addCoef.col = "black", number.cex = 0.7,
             title = "Predictor Collinearity (1-hour lag)")
  }
} else {
  cat("Not enough data for correlation analysis\n")
}
```



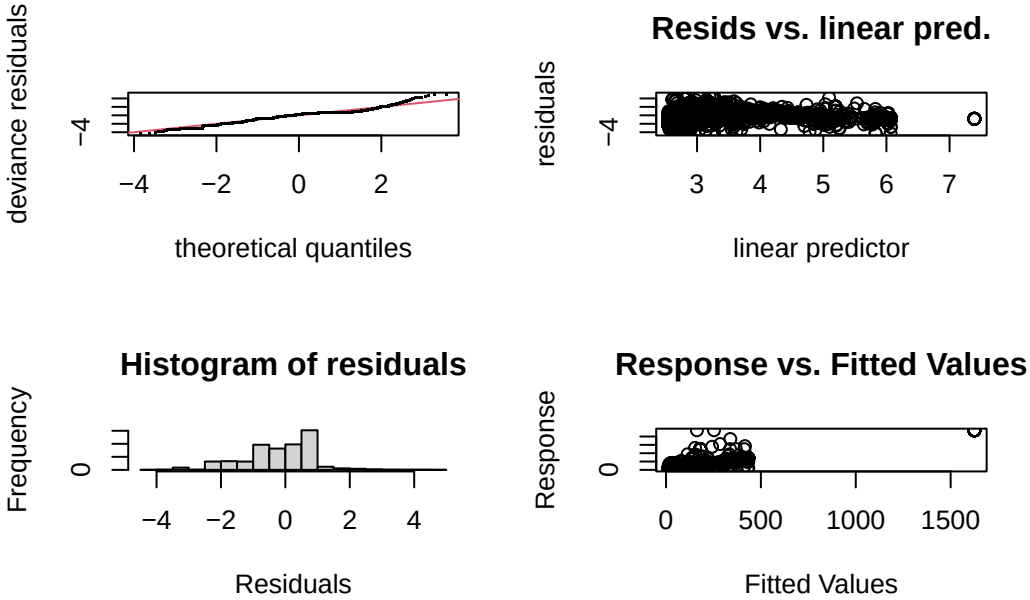

## Model Fitting (1 hour)

```
# Use same model formulas
model_formulas_60m <- model_formulas_30m

# Storage for results
models_60m <- list()
```

```r
aic_values_60m <- numeric()

# Fit each model
for (i in seq_along(model_formulas_60m)) {
  model_name <- names(model_formulas_60m)[i]
  formula <- model_formulas_60m[[i]]

  cat("\n### Model", i, ":", model_name, "\n\n")

  tryCatch({
    # Check data before fitting
    model_vars <- all.vars(formula)
    model_data <- data_60m %>%
      select(all_of(model_vars), view_id) %>%
      na.omit()

    cat("Model data dimensions after removing NAs:", nrow(model_data), "x", ncol(model_data)

    # Fit GAM
    gam_model <- gam(
      formula,
      family = nb(),
      data = model_data
    )

    cat("GAM fitted successfully\n")

    # Store model
    models_60m[[model_name]] <- gam_model
    aic_values_60m[model_name] <- AIC(gam_model)

    # Print summary
    cat("Model Summary:\n")
    print(summary(gam_model))

    # Diagnostic plots
    cat("\n\nDiagnostic plots:\n")
    gam.check(gam_model)

  }, error = function(e) {
    cat("Error fitting model:", e$message, "\n")
    aic_values_60m[model_name] <- NA
```

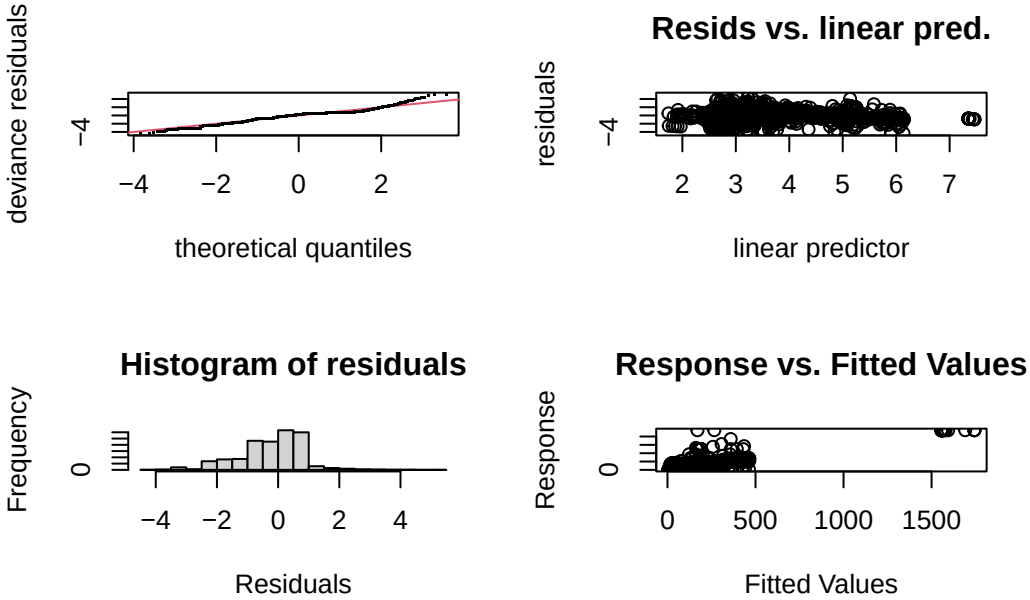
```
  })
}
```

```
### Model 1 : Baseline (abundance only)

Model data dimensions after removing NAs: 5618 x 3
GAM fitted successfully
Model Summary:

Family: Negative Binomial(2.848)
Link function: log

Formula:
total_butterflies ~ s(abundance_t_minus_1, k = 4)

Parametric coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) 3.603399   0.008335   432.3   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Approximate significance of smooth terms:
                          edf Ref.df Chi.sq p-value
s(abundance_t_minus_1) 2.998      3  17359  <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

R-sq.(adj) =  0.436   Deviance explained = 73.6%
-REML =  24938  Scale est. = 1         n = 5618


Diagnostic plots:
```



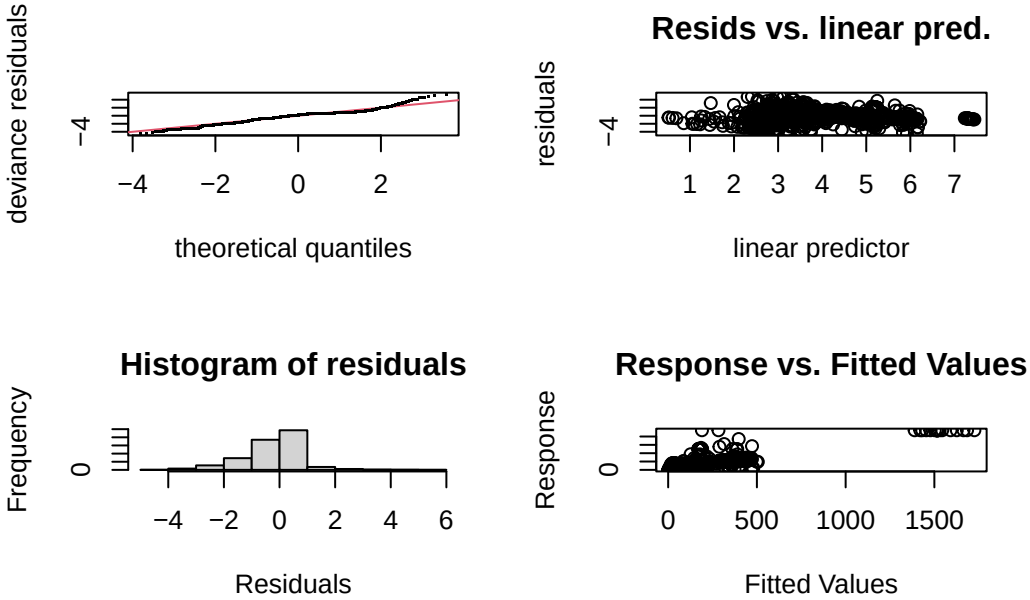

```
Method: REML   Optimizer: outer newton
full convergence after 7 iterations.
Gradient range [-0.0001393679,0.00164137]
(score 24938.2 & scale 1).
Hessian positive definite, eigenvalue range [0.9996031,2023.384].
Model rank =  4 / 4

Basis dimension (k) checking results. Low p-value (k-index<1) may
indicate that k is too low, especially if edf is close to k'.

                       k' edf k-index p-value
s(abundance_t_minus_1)  3   3    0.38  <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

### Model 2 : Wind speed

Model data dimensions after removing NAs: 5618 x 4
GAM fitted successfully
Model Summary:

Family: Negative Binomial(2.887)
```

```
Link function: log

Formula:
total_butterflies ~ s(abundance_t_minus_1, k = 4) + s(mean_wind_speed,
    k = 4)

Parametric coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) 3.600972   0.008288   434.5   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Approximate significance of smooth terms:
                         edf Ref.df   Chi.sq p-value
s(abundance_t_minus_1) 2.998  3.000 17175.47  <2e-16 ***
s(mean_wind_speed)     2.937  2.997    68.65  <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

R-sq.(adj) =  0.461   Deviance explained = 73.9%
-REML =  24914  Scale est. = 1         n = 5618


Diagnostic plots:
```



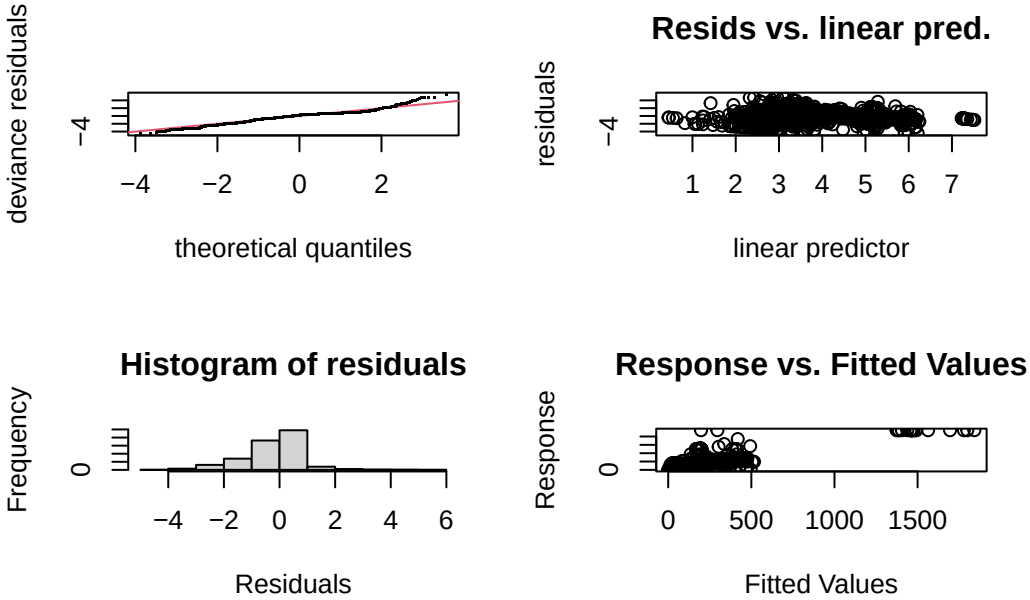

```
Method: REML   Optimizer: outer newton
full convergence after 8 iterations.
Gradient range [-2.943872e-05,0.0003456825]
(score 24913.81 & scale 1).
Hessian positive definite, eigenvalue range [0.8921423,2008.74].
Model rank =  7 / 7

Basis dimension (k) checking results. Low p-value (k-index<1) may
indicate that k is too low, especially if edf is close to k'.

                         k'  edf k-index p-value
s(abundance_t_minus_1) 3.00 3.00    0.38  <2e-16 ***
s(mean_wind_speed)     3.00 2.94    0.93    0.01 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

### Model 3 : Temperature

Model data dimensions after removing NAs: 5618 x 4
GAM fitted successfully
Model Summary:
```

```
Family: Negative Binomial(2.956)
Link function: log

Formula:
total_butterflies ~ s(abundance_t_minus_1, k = 4) + s(mean_temp,
    k = 4)

Parametric coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  3.59660    0.00821   438.1   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Approximate significance of smooth terms:
                         edf Ref.df  Chi.sq p-value
s(abundance_t_minus_1) 2.998  3.000 16937.5  <2e-16 ***
s(mean_temp)           2.972  2.999   195.3  <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

R-sq.(adj) =  0.522   Deviance explained = 74.4%
-REML =  24856  Scale est. = 1         n = 5618


Diagnostic plots:
```



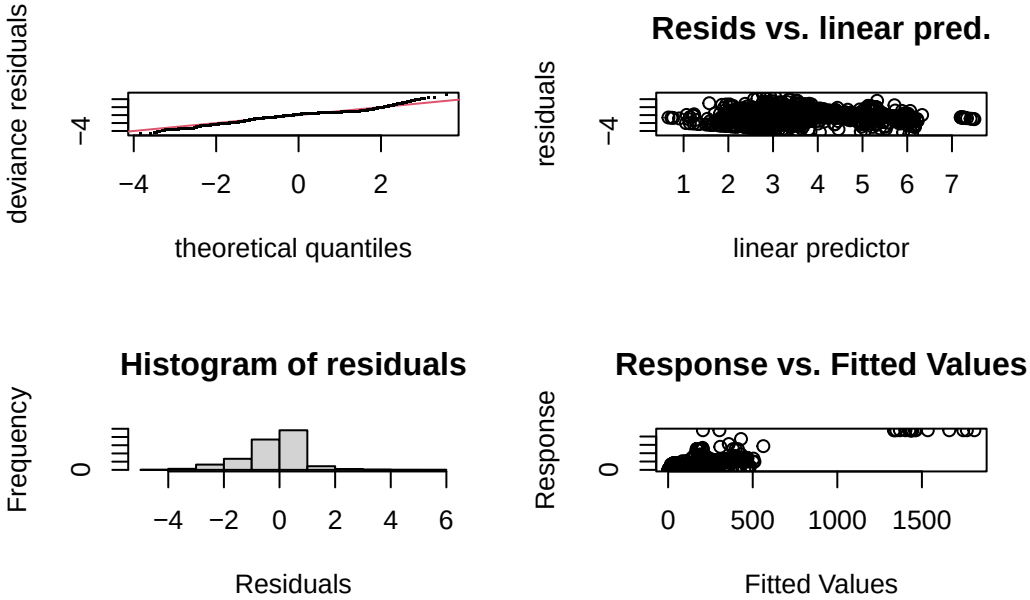

```
Method: REML   Optimizer: outer newton
full convergence after 7 iterations.
Gradient range [-0.0001803716,0.002133872]
(score 24856.13 & scale 1).
Hessian positive definite, eigenvalue range [0.9541734,1988.328].
Model rank =  7 / 7

Basis dimension (k) checking results. Low p-value (k-index<1) may
indicate that k is too low, especially if edf is close to k'.

                        k'  edf k-index p-value
s(abundance_t_minus_1) 3.00 3.00    0.39  <2e-16 ***
s(mean_temp)           3.00 2.97    0.92  <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

### Model 4 : Wind + Temperature

Model data dimensions after removing NAs: 5618 x 5
GAM fitted successfully
Model Summary:
```

```
Family: Negative Binomial(2.993)
Link function: log

Formula:
total_butterflies ~ s(abundance_t_minus_1, k = 4) + s(mean_wind_speed,
    k = 4) + s(mean_temp, k = 4)

Parametric coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) 3.594481   0.008169     440   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Approximate significance of smooth terms:
                          edf Ref.df   Chi.sq p-value
s(abundance_t_minus_1) 2.998  3.000 16752.00  <2e-16 ***
s(mean_wind_speed)     2.879  2.989    58.41  <2e-16 ***
s(mean_temp)           2.973  2.999   190.73  <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

R-sq.(adj) =   0.53   Deviance explained = 74.7%
-REML =  24835  Scale est. = 1         n = 5618


Diagnostic plots:
```



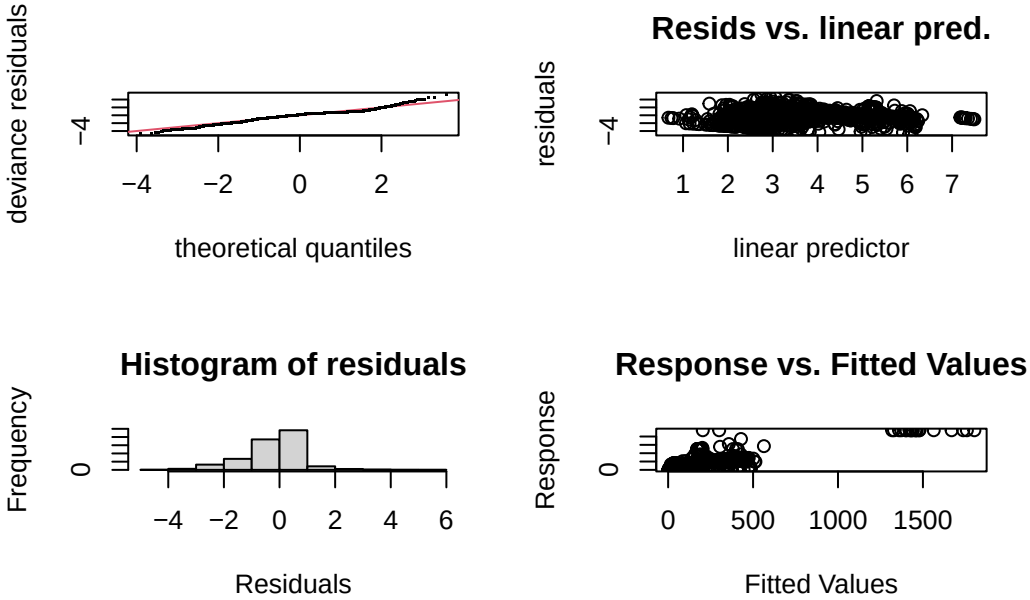

```
Method: REML   Optimizer: outer newton
full convergence after 7 iterations.
Gradient range [-0.00162929,0.01929681]
(score 24834.71 & scale 1).
Hessian positive definite, eigenvalue range [0.7919262,1975.37].
Model rank =  10 / 10

Basis dimension (k) checking results. Low p-value (k-index<1) may
indicate that k is too low, especially if edf is close to k'.

                         k'  edf k-index p-value
s(abundance_t_minus_1) 3.00 3.00    0.39  <2e-16 ***
s(mean_wind_speed)     3.00 2.88    0.93   0.015 *
s(mean_temp)           3.00 2.97    0.92  <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

### Model 5 : Wind + Temp + Sunlight

Model data dimensions after removing NAs: 5618 x 6
GAM fitted successfully
Model Summary:
```

```
Family: Negative Binomial(3.125)
Link function: log

Formula:
total_butterflies ~ s(abundance_t_minus_1, k = 4) + s(mean_wind_speed,
    k = 4) + s(mean_temp, k = 4) + s(sunlight_proportion, k = 4)

Parametric coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) 3.587528   0.008023   447.2   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Approximate significance of smooth terms:
                         edf Ref.df   Chi.sq p-value
s(abundance_t_minus_1) 2.998  3.000 16782.84  <2e-16 ***
s(mean_wind_speed)     2.897  2.992    57.91  <2e-16 ***
s(mean_temp)           2.973  2.999   208.09  <2e-16 ***
s(sunlight_proportion) 2.933  2.996   218.30  <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

R-sq.(adj) =  0.559   Deviance explained = 75.6%
-REML =  24739  Scale est. = 1         n = 5618


Diagnostic plots:
```




```
Method: REML   Optimizer: outer newton
full convergence after 8 iterations.
Gradient range [-0.0009857475,0.01095602]
(score 24738.78 & scale 1).
Hessian positive definite, eigenvalue range [0.8232855,1937.787].
Model rank =  13 / 13

Basis dimension (k) checking results. Low p-value (k-index<1) may
indicate that k is too low, especially if edf is close to k'.

                         k'  edf k-index p-value
s(abundance_t_minus_1) 3.00 3.00    0.41  <2e-16 ***
s(mean_wind_speed)     3.00 2.90    0.93    0.02 *
s(mean_temp)           3.00 2.97    0.89  <2e-16 ***
s(sunlight_proportion) 3.00 2.93    0.90  <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

### Model 6 : Full model

Model data dimensions after removing NAs: 5618 x 7
GAM fitted successfully
```

```
Model Summary:

Family: Negative Binomial(3.129)
Link function: log

Formula:
total_butterflies ~ s(abundance_t_minus_1, k = 4) + s(mean_wind_speed,
    k = 4) + s(mean_temp, k = 4) + s(sunlight_proportion, k = 4) +
    s(minutes_above_2mps, k = 4)

Parametric coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) 3.587296   0.008019   447.3   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Approximate significance of smooth terms:
                         edf Ref.df    Chi.sq p-value
s(abundance_t_minus_1) 2.998  3.000 16548.582  <2e-16 ***
s(mean_wind_speed)     2.479  2.784    38.291  <2e-16 ***
s(mean_temp)           2.973  2.999   208.019  <2e-16 ***
s(sunlight_proportion) 2.933  2.996   217.921  <2e-16 ***
s(minutes_above_2mps)  1.954  2.317     8.271  0.0394 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

R-sq.(adj) =  0.555   Deviance explained = 75.7%
-REML =  24738  Scale est. = 1         n = 5618


Diagnostic plots:
```




```
Method: REML   Optimizer: outer newton
full convergence after 9 iterations.
Gradient range [-1.252413e-05,0.0001421771]
(score 24738.3 & scale 1).
Hessian positive definite, eigenvalue range [0.1007015,1935.989].
Model rank =  16 / 16

Basis dimension (k) checking results. Low p-value (k-index<1) may
indicate that k is too low, especially if edf is close to k'.

                         k'  edf k-index p-value
s(abundance_t_minus_1) 3.00 3.00    0.40  <2e-16 ***
s(mean_wind_speed)     3.00 2.48    0.92  <2e-16 ***
s(mean_temp)           3.00 2.97    0.89  <2e-16 ***
s(sunlight_proportion) 3.00 2.93    0.91  <2e-16 ***
s(minutes_above_2mps)  3.00 1.95    0.95    0.31
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
# Model comparison table
comparison_60m <- data.frame(
```

```r
  Model = names(aic_values_60m),
  AIC = round(aic_values_60m, 2)
) %>%
  arrange(AIC)

cat("\n### Model Comparison (1-hour lag)\n")
```

```
### Model Comparison (1-hour lag)
```

```r
kable(comparison_60m)
```

| | Model | AIC |
|---|---|---|
| Full model | Full model | 49417.42 |
| Wind + Temp + Sunlight | Wind + Temp + Sunlight | 49420.09 |
| Wind + Temperature | Wind + Temperature | 49623.71 |
| Temperature | Temperature | 49676.53 |
| Wind speed | Wind speed | 49793.65 |
| Baseline (abundance only) | Baseline (abundance only) | 49853.76 |

---

### 2-Hour Lag Analysis

```r
data_120m <- prepare_lag_data(master_df, wind, lag_minutes = 120)
```

```
Preparing data for 120 minute lag...
Valid pairs after filtering: 6912
Calculating interval predictors efficiently...
Processing 6912 unique intervals...
Progress: 10 %
Progress: 20 %
Progress: 30 %
Progress: 40 %
Progress: 50 %
Progress: 60 %
Progress: 70 %
```

```
Progress: 80 %
Progress: 90 %
Progress: 100 %
Combining results...
Final dataset rows: 5650
Time delta range: 120 to 120 minutes
```

### Exploratory Data Analysis (2 hours)

```r
cat("Data structure:\n")
```

```
Data structure:
```

```r
glimpse(data_120m)
```

```
Rows: 5,650
Columns: 30
$ deployment_id.x           <chr> "SC1", "SC1", "SC1", "SC1", "SC1", "SC1", "S~
$ image_filename            <chr> "SC1_20231117184501.JPG", "SC1_2023111718500~
$ total_butterflies         <dbl> 54, 54, 54, 54, 54, 54, 54, 54, 54, 54, 54, ~
$ butterflies_direct_sun    <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,~
$ datetime                  <dttm> 2023-11-17 18:45:01, 2023-11-17 18:50:01, 2~
$ view_id                   <fct> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1,~
$ wind_meter_name           <chr> "JazzPlay", "JazzPlay", "JazzPlay", "JazzPla~
$ deployment_id.y           <chr> "SC1", "SC1", "SC1", "SC1", "SC1", "SC1", "S~
$ timestamp                 <dbl> 2.023112e+13, 2.023112e+13, 2.023112e+13, 2.~
$ temperature               <dbl> 15, 16, 15, 16, 15, 16, 16, 15, 16, 15, 16, ~
$ confidence                <dbl> 0.4041873, 0.5208792, 0.6782328, 0.5311629, ~
$ extraction_status         <chr> "success", "success", "success", "success", ~
$ datetime_t_minus_1        <dttm> 2023-11-17 16:45:01, 2023-11-17 16:50:01, 2~
$ abundance_t_minus_1       <dbl> 39, 41, 39, 68, 73, 72, 61, 49, 57, 51, 54, ~
$ butterflies_sun_t_minus_1 <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,~
$ temperature_t_minus_1     <dbl> 15, 15, 15, 15, 15, 15, 16, 15, 15, 15, 15, ~
$ time_delta_mins           <dbl> 120, 120, 120, 120, 120, 120, 120, 120, 120,~
$ mean_wind_speed           <dbl> 0.400000, 0.900000, 1.030769, 1.177778, 1.24~
$ max_wind_speed            <dbl> 0.4, 1.9, 1.9, 1.9, 1.9, 2.3, 2.7, 2.9, 2.9,~
$ sd_wind_speed             <dbl> 0.0000000, 0.6524678, 0.5297556, 0.5241863, ~
$ gust_factor               <dbl> 2.250000, 1.722222, 1.656716, 1.622642, 1.58~
$ minutes_above_2mps        <dbl> 0, 0, 0, 0, 0, 1, 3, 6, 7, 12, 16, 20, 25, 3~
$ mean_temp                 <dbl> 15.04, 15.08, 15.08, 15.12, 15.12, 15.16, 15~
```

```
$ max_temp                  <dbl> 16, 16, 16, 16, 16, 16, 16, 16, 16, 16, 16, ~
$ min_temp                  <dbl> 15, 15, 15, 15, 15, 15, 15, 15, 15, 15, 15, ~
$ sd_temp                   <dbl> 0.2000000, 0.2768875, 0.2768875, 0.3316625, ~
$ sunlight_proportion       <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,~
$ max_prop_sunlight         <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,~
$ min_prop_sunlight         <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,~
$ sd_prop_sunlight          <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,~
```

```r
cat("\n\nData summary:\n")
```

```
Data summary:
```

```r
summary(data_120m)
```

```
 deployment_id.x    image_filename     total_butterflies butterflies_direct_sun
 Length:5650        Length:5650        Min.   :  0.00    Min.   :  0.000
 Class :character   Class :character   1st Qu.: 13.00    1st Qu.:  0.000
 Mode  :character   Mode  :character   Median : 34.00    Median :  0.000
                                       Mean   : 65.33    Mean   :  1.647
                                       3rd Qu.: 72.00    3rd Qu.:  0.000
                                       Max.   :950.00    Max.   :150.000

    datetime                          view_id     wind_meter_name
 Min.   :2023-11-17 18:45:01.00   4      :2921   Length:5650
 1st Qu.:2023-12-11 03:07:31.00   2      :1080   Class :character
 Median :2023-12-29 07:50:01.00   1      : 662   Mode  :character
 Mean   :2023-12-29 03:20:27.64   5      : 460
 3rd Qu.:2024-01-23 08:12:31.00   11     : 287
 Max.   :2024-02-04 10:50:01.00   9      : 171
                                  (Other):  69
 deployment_id.y      timestamp          temperature      confidence
 Length:5650        Min.   :2.023e+13   Min.   : 4.00   Min.   :0.1222
 Class :character   1st Qu.:2.023e+13   1st Qu.:11.00   1st Qu.:0.4967
 Mode  :character   Median :2.023e+13   Median :13.00   Median :0.5934
                    Mean   :2.024e+13   Mean   :13.26   Mean   :0.5960
                    3rd Qu.:2.024e+13   3rd Qu.:15.00   3rd Qu.:0.6951
                    Max.   :2.024e+13   Max.   :33.00   Max.   :1.0000

 extraction_status  datetime_t_minus_1               abundance_t_minus_1
```

```
Length:5650        Min.   :2023-11-17 16:45:01.00   Min.   :  0.00
Class :character   1st Qu.:2023-12-11 01:07:31.00   1st Qu.: 14.00
Mode  :character   Median :2023-12-29 05:50:01.00   Median : 35.00
                   Mean   :2023-12-29 01:20:27.64   Mean   : 65.93
                   3rd Qu.:2024-01-23 06:12:31.00   3rd Qu.: 73.75
                   Max.   :2024-02-04 08:50:01.00   Max.   :950.00

butterflies_sun_t_minus_1 temperature_t_minus_1 time_delta_mins
Min.   :  0.000           Min.   : 4.00         Min.   :120
1st Qu.:  0.000           1st Qu.:11.00         1st Qu.:120
Median :  0.000           Median :13.00         Median :120
Mean   :  1.707           Mean   :13.25         Mean   :120
3rd Qu.:  0.000           3rd Qu.:15.00         3rd Qu.:120
Max.   :150.000           Max.   :30.00         Max.   :120

mean_wind_speed  max_wind_speed sd_wind_speed     gust_factor
Min.   :0.0000   Min.   :0.00   Min.   :0.0000   Min.   :1.181
1st Qu.:0.2429   1st Qu.:1.00   1st Qu.:0.2049   1st Qu.:1.357
Median :0.7521   Median :1.50   Median :0.3170   Median :1.447
Mean   :0.8837   Mean   :1.71   Mean   :0.3288   Mean   :  Inf
3rd Qu.:1.3033   3rd Qu.:2.30   3rd Qu.:0.4395   3rd Qu.:1.571
Max.   :5.0092   Max.   :6.70   Max.   :1.2932   Max.   :  Inf
                                NA's   :1        NA's   :498
minutes_above_2mps   mean_temp        max_temp        min_temp
Min.   :  0.00     Min.   : 4.20   Min.   : 5.00   Min.   : 4.00
1st Qu.:  0.00     1st Qu.:11.00   1st Qu.:12.00   1st Qu.:10.00
Median :  0.00     Median :13.00   Median :14.00   Median :12.00
Mean   : 11.58     Mean   :13.26   Mean   :14.22   Mean   :12.35
3rd Qu.:  4.00     3rd Qu.:15.48   3rd Qu.:16.00   3rd Qu.:15.00
Max.   :120.00     Max.   :29.85   Max.   :33.00   Max.   :28.00

   sd_temp       sunlight_proportion max_prop_sunlight min_prop_sunlight
Min.   :0.0000   Min.   :0.00000     Min.   :0.00000   Min.   :0.00000
1st Qu.:0.4472   1st Qu.:0.00000     1st Qu.:0.00000   1st Qu.:0.00000
Median :0.5477   Median :0.00000     Median :0.00000   Median :0.00000
Mean   :0.7375   Mean   :0.07285     Mean   :0.15567   Mean   :0.03173
3rd Qu.:0.8944   3rd Qu.:0.01923     3rd Qu.:0.07868   3rd Qu.:0.00000
Max.   :6.8775   Max.   :1.00000     Max.   :1.00000   Max.   :1.00000

sd_prop_sunlight
Min.   :0.00000
1st Qu.:0.00000
Median :0.00000
```

```
 Mean   :0.04904
 3rd Qu.:0.00221
 Max.   :0.70711
 NA's   :37
```

```r
# Check for data quality issues
cat("\n\nData quality checks:\n")
```

```
Data quality checks:
```

```r
cat("Total rows:", nrow(data_120m), "\n")
```

```
Total rows: 5650
```

```r
cat("Complete cases:", sum(complete.cases(data_120m)), "\n")
```

```
Complete cases: 5122
```

```r
cat("Rows with abundance_t_minus_1 > 0:", sum(data_120m$abundance_t_minus_1 > 0, na.rm = TRU
```

```
Rows with abundance_t_minus_1 > 0: 5552
```

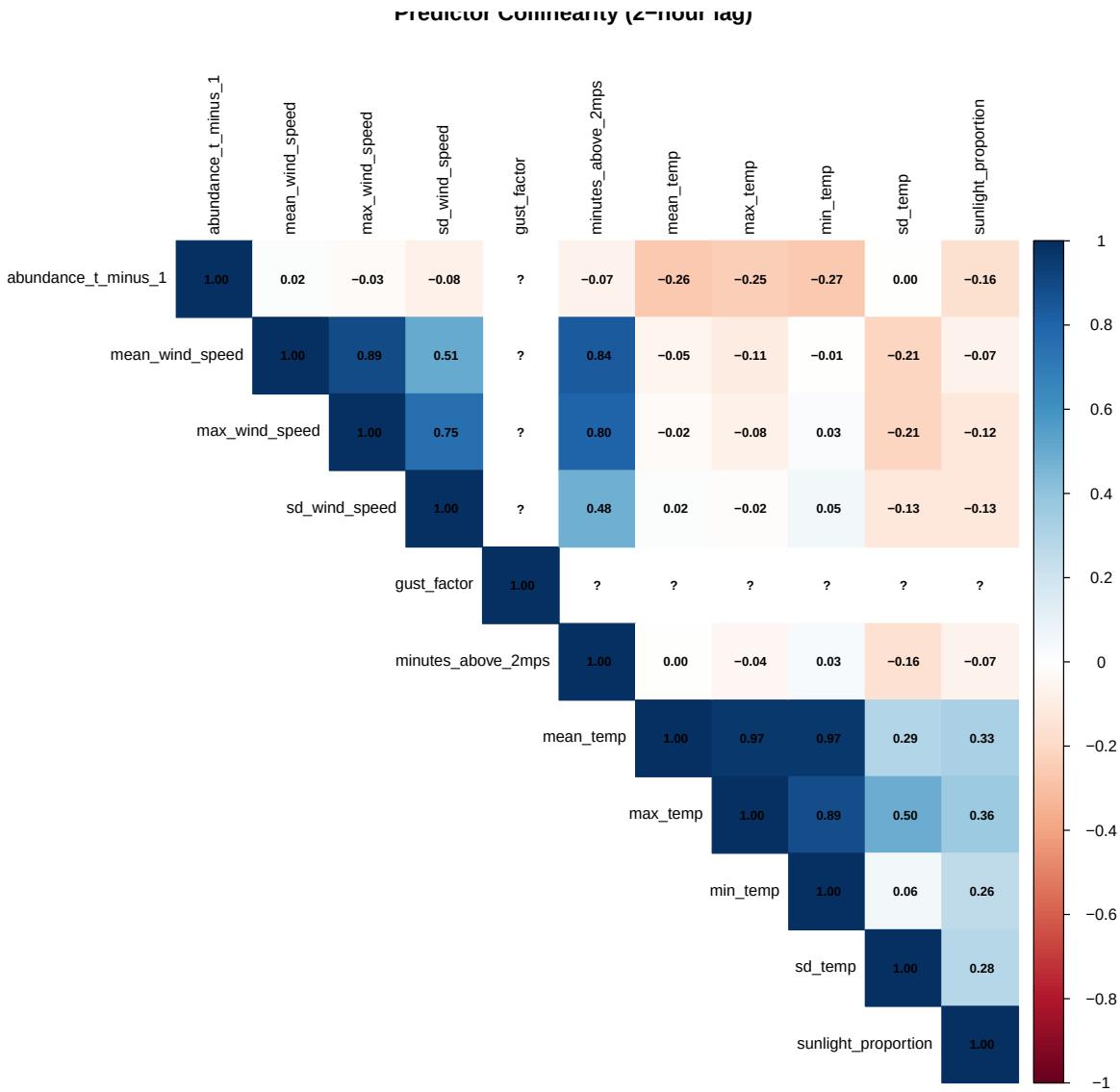
```r
cat("Rows with total_butterflies > 0:", sum(data_120m$total_butterflies > 0, na.rm = TRUE), 
```

```
Rows with total_butterflies > 0: 5499
```

```r
cat("Unique view_ids:", length(unique(data_120m$view_id)), "\n")
```

```
Unique view_ids: 9
```

```r
# Select numeric predictors for correlation analysis
predictors_120m <- data_120m %>%
  select(abundance_t_minus_1, mean_wind_speed, max_wind_speed, sd_wind_speed,
         gust_factor, minutes_above_2mps, mean_temp, max_temp, min_temp, sd_temp,
         sunlight_proportion) %>%
  na.omit()

if (nrow(predictors_120m) > 1) {
  cor_matrix_120m <- cor(predictors_120m, use = "complete.obs")
  if (!any(is.na(cor_matrix_120m) | is.infinite(cor_matrix_120m))) {
    corrplot(cor_matrix_120m, method = "color", type = "upper",
             order = "hclust", tl.cex = 0.8, tl.col = "black",
             addCoef.col = "black", number.cex = 0.7,
             title = "Predictor Collinearity (2-hour lag)")
  } else {
    corrplot(cor_matrix_120m, method = "color", type = "upper",
             order = "original", tl.cex = 0.8, tl.col = "black",
             addCoef.col = "black", number.cex = 0.7,
             title = "Predictor Collinearity (2-hour lag)")
  }
} else {
  cat("Not enough data for correlation analysis\n")
}
```



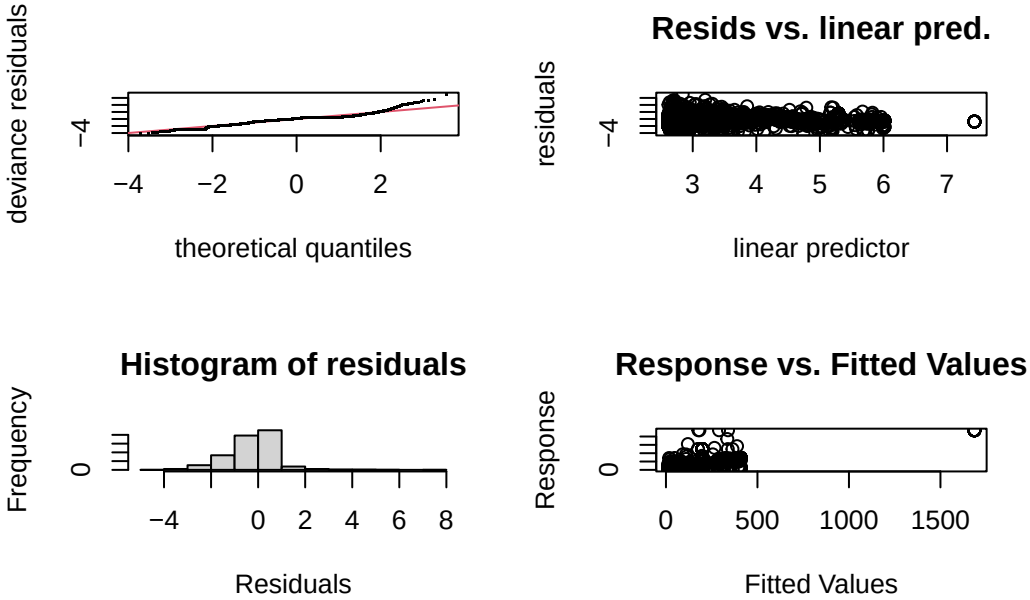

## Model Fitting (2 hours)

```
# Use same model formulas
model_formulas_120m <- model_formulas_30m

# Storage for results
models_120m <- list()
```

```r
aic_values_120m <- numeric()

# Fit each model
for (i in seq_along(model_formulas_120m)) {
  model_name <- names(model_formulas_120m)[i]
  formula <- model_formulas_120m[[i]]

  cat("\n### Model", i, ":", model_name, "\n\n")

  tryCatch({
    # Check data before fitting
    model_vars <- all.vars(formula)
    model_data <- data_120m %>%
      select(all_of(model_vars), view_id) %>%
      na.omit()

    cat("Model data dimensions after removing NAs:", nrow(model_data), "x", ncol(model_data)

    # Fit GAM
    gam_model <- gam(
      formula,
      family = nb(),
      data = model_data
    )

    cat("GAM fitted successfully\n")

    # Store model
    models_120m[[model_name]] <- gam_model
    aic_values_120m[model_name] <- AIC(gam_model)

    # Print summary
    cat("Model Summary:\n")
    print(summary(gam_model))

    # Diagnostic plots
    cat("\n\nDiagnostic plots:\n")
    gam.check(gam_model)

  }, error = function(e) {
    cat("Error fitting model:", e$message, "\n")
    aic_values_120m[model_name] <- NA
```

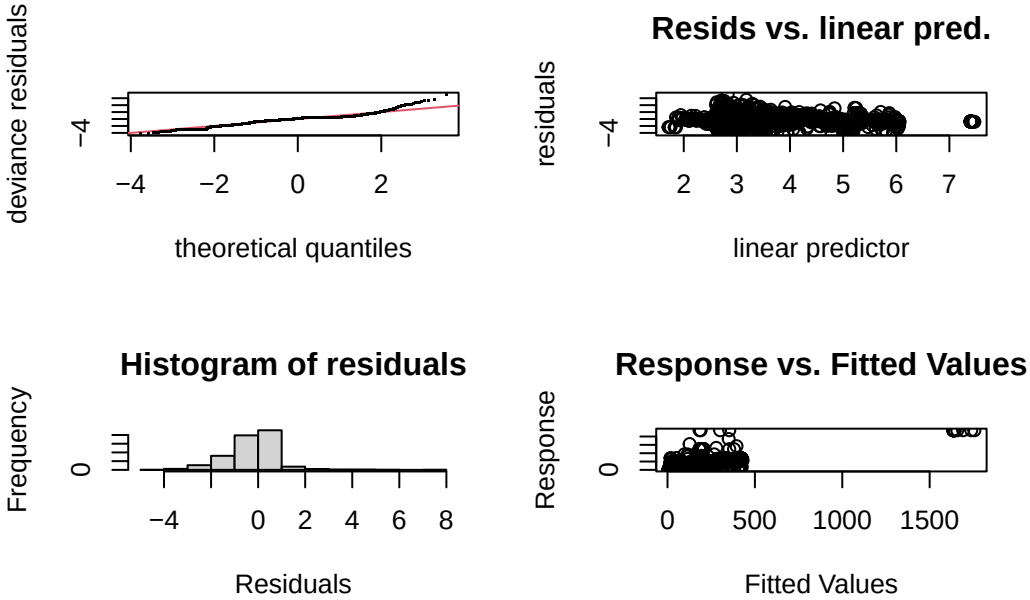
```
  })
}
```

```
### Model 1 : Baseline (abundance only)

Model data dimensions after removing NAs: 5650 x 3
GAM fitted successfully
Model Summary:

Family: Negative Binomial(2.078)
Link function: log

Formula:
total_butterflies ~ s(abundance_t_minus_1, k = 4)

Parametric coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) 3.621855   0.009583     378   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Approximate significance of smooth terms:
                         edf Ref.df Chi.sq p-value
s(abundance_t_minus_1) 2.997      3  11906  <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

R-sq.(adj) =  0.456   Deviance explained =   66%
-REML =  25652  Scale est. = 1         n = 5650


Diagnostic plots:
```



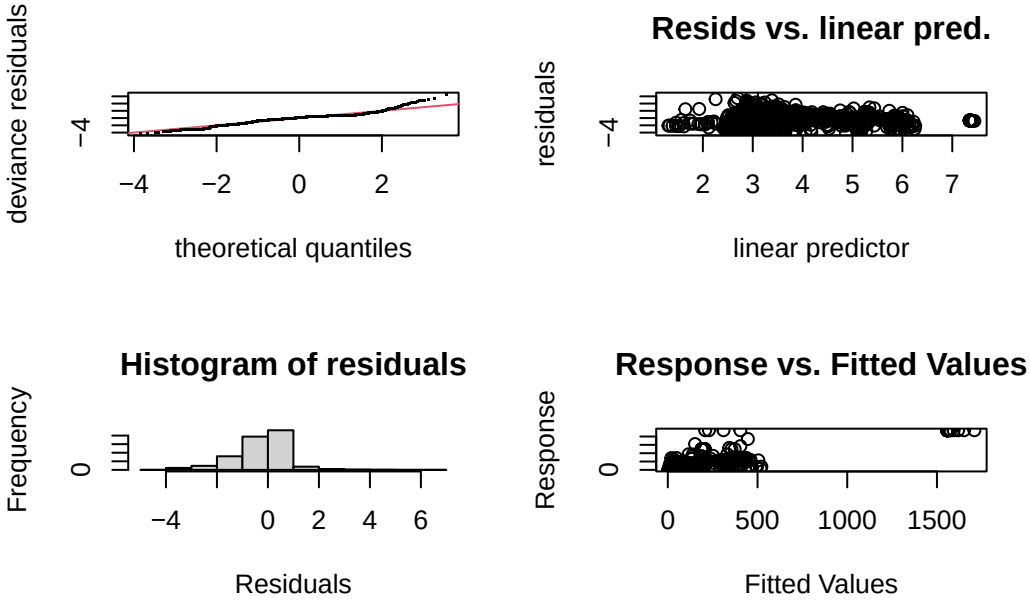

```
Method: REML   Optimizer: outer newton
full convergence after 7 iterations.
Gradient range [-6.682563e-05,0.001205509]
(score 25652.2 & scale 1).
Hessian positive definite, eigenvalue range [0.9978899,2267.259].
Model rank =  4 / 4

Basis dimension (k) checking results. Low p-value (k-index<1) may
indicate that k is too low, especially if edf is close to k'.

                       k' edf k-index p-value
s(abundance_t_minus_1)  3   3    0.53  <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

### Model 2 : Wind speed

Model data dimensions after removing NAs: 5650 x 4
GAM fitted successfully
Model Summary:

Family: Negative Binomial(2.097)
```

```
Link function: log

Formula:
total_butterflies ~ s(abundance_t_minus_1, k = 4) + s(mean_wind_speed,
    k = 4)

Parametric coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) 3.619456   0.009545   379.2   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Approximate significance of smooth terms:
                         edf Ref.df   Chi.sq p-value
s(abundance_t_minus_1) 2.997  3.000 11682.45  <2e-16 ***
s(mean_wind_speed)     2.890  2.991    54.17  <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

R-sq.(adj) =   0.48   Deviance explained = 66.3%
-REML =  25635  Scale est. = 1         n = 5650


Diagnostic plots:
```



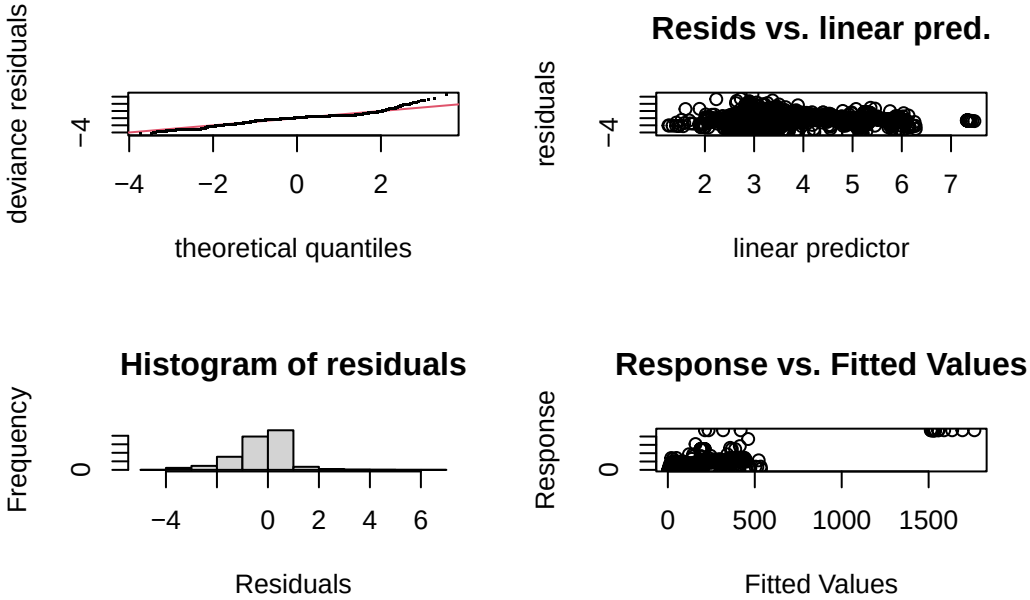

```
Method: REML   Optimizer: outer newton
full convergence after 7 iterations.
Gradient range [-0.0002371538,0.004192237]
(score 25635.12 & scale 1).
Hessian positive definite, eigenvalue range [0.8267913,2256.827].
Model rank =  7 / 7

Basis dimension (k) checking results. Low p-value (k-index<1) may
indicate that k is too low, especially if edf is close to k'.

                         k'  edf k-index p-value
s(abundance_t_minus_1) 3.00 3.00    0.54  <2e-16 ***
s(mean_wind_speed)     3.00 2.89    0.90  <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

### Model 3 : Temperature

Model data dimensions after removing NAs: 5650 x 4
GAM fitted successfully
Model Summary:
```

```
Family: Negative Binomial(2.149)
Link function: log

Formula:
total_butterflies ~ s(abundance_t_minus_1, k = 4) + s(mean_temp,
    k = 4)

Parametric coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) 3.613795   0.009443   382.7   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Approximate significance of smooth terms:
                         edf Ref.df  Chi.sq p-value
s(abundance_t_minus_1) 2.997      3 11864.4  <2e-16 ***
s(mean_temp)           2.976      3   189.2  <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

R-sq.(adj) =   0.53   Deviance explained =   67%
-REML =  25573  Scale est. = 1         n = 5650


Diagnostic plots:
```



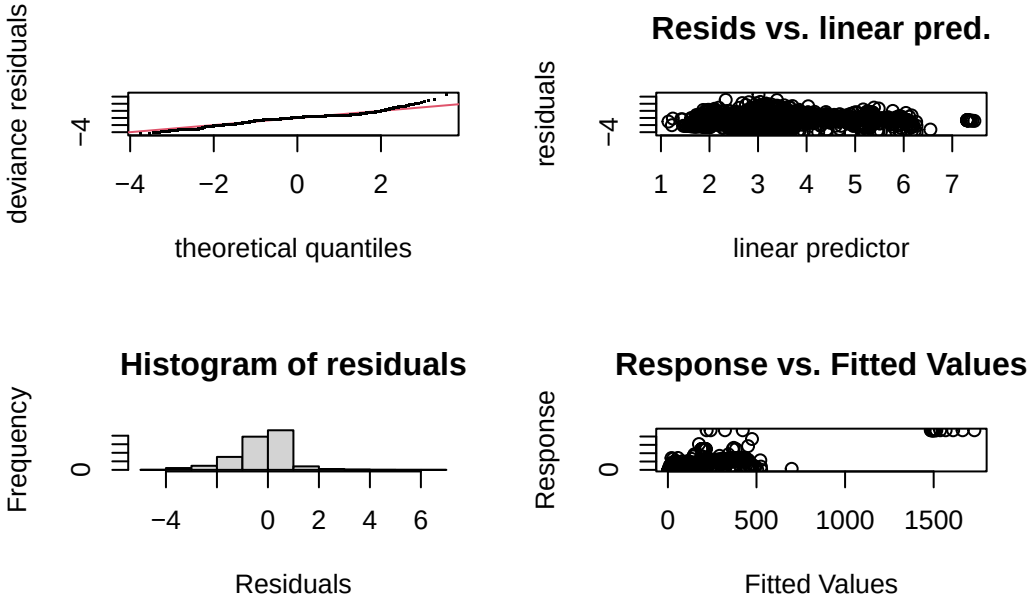

```
Method: REML   Optimizer: outer newton
full convergence after 7 iterations.
Gradient range [-0.0001337358,0.002243531]
(score 25572.56 & scale 1).
Hessian positive definite, eigenvalue range [0.95088,2235.148].
Model rank =  7 / 7

Basis dimension (k) checking results. Low p-value (k-index<1) may
indicate that k is too low, especially if edf is close to k'.

                         k'  edf k-index p-value
s(abundance_t_minus_1) 3.00 3.00    0.52  <2e-16 ***
s(mean_temp)           3.00 2.98    0.86  <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

### Model 4 : Wind + Temperature

Model data dimensions after removing NAs: 5650 x 5
GAM fitted successfully
Model Summary:
```

```
Family: Negative Binomial(2.166)
Link function: log

Formula:
total_butterflies ~ s(abundance_t_minus_1, k = 4) + s(mean_wind_speed,
    k = 4) + s(mean_temp, k = 4)

Parametric coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) 3.611827   0.009412   383.7   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Approximate significance of smooth terms:
                          edf Ref.df   Chi.sq p-value
s(abundance_t_minus_1) 2.997  3.000 11667.95  <2e-16 ***
s(mean_wind_speed)     2.596  2.887    41.72  <2e-16 ***
s(mean_temp)           2.977  3.000   188.61  <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

R-sq.(adj) =  0.537   Deviance explained = 67.3%
-REML =  25558  Scale est. = 1         n = 5650


Diagnostic plots:
```



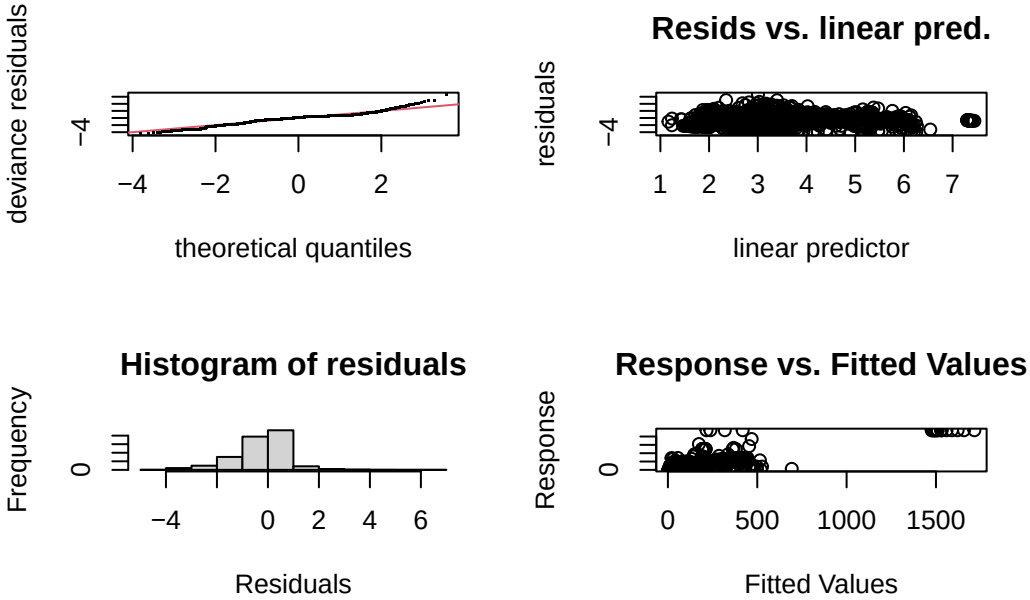

```
Method: REML   Optimizer: outer newton
full convergence after 7 iterations.
Gradient range [1.147519e-06,0.0001002677]
(score 25558.09 & scale 1).
Hessian positive definite, eigenvalue range [0.5236651,2226.502].
Model rank =  10 / 10

Basis dimension (k) checking results. Low p-value (k-index<1) may
indicate that k is too low, especially if edf is close to k'.

                         k'  edf k-index p-value
s(abundance_t_minus_1) 3.00 3.00    0.51  <2e-16 ***
s(mean_wind_speed)     3.00 2.60    0.92  <2e-16 ***
s(mean_temp)           3.00 2.98    0.90  <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

### Model 5 : Wind + Temp + Sunlight

Model data dimensions after removing NAs: 5650 x 6
GAM fitted successfully
Model Summary:
```

```
Family: Negative Binomial(2.287)
Link function: log

Formula:
total_butterflies ~ s(abundance_t_minus_1, k = 4) + s(mean_wind_speed,
    k = 4) + s(mean_temp, k = 4) + s(sunlight_proportion, k = 4)

Parametric coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) 3.599236   0.009196   391.4   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Approximate significance of smooth terms:
                         edf Ref.df   Chi.sq p-value
s(abundance_t_minus_1) 2.997  3.000 11876.80  <2e-16 ***
s(mean_wind_speed)     2.695  2.934    39.76  <2e-16 ***
s(mean_temp)           2.972  2.999   188.50  <2e-16 ***
s(sunlight_proportion) 2.933  2.996   314.52  <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

R-sq.(adj) =  0.545   Deviance explained = 68.8%
-REML =  25425  Scale est. = 1         n = 5650


Diagnostic plots:
```




```
Method: REML   Optimizer: outer newton
full convergence after 15 iterations.
Gradient range [-8.872229e-06,0.001428982]
(score 25424.55 & scale 1).
Hessian positive definite, eigenvalue range [0.6383153,2179.041].
Model rank =  13 / 13

Basis dimension (k) checking results. Low p-value (k-index<1) may
indicate that k is too low, especially if edf is close to k'.

                         k'  edf k-index p-value
s(abundance_t_minus_1) 3.00 3.00    0.51  <2e-16 ***
s(mean_wind_speed)     3.00 2.69    0.93    0.03 *
s(mean_temp)           3.00 2.97    0.88  <2e-16 ***
s(sunlight_proportion) 3.00 2.93    0.88  <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

### Model 6 : Full model

Model data dimensions after removing NAs: 5650 x 7
GAM fitted successfully
```

```
Model Summary:

Family: Negative Binomial(2.294)
Link function: log

Formula:
total_butterflies ~ s(abundance_t_minus_1, k = 4) + s(mean_wind_speed,
    k = 4) + s(mean_temp, k = 4) + s(sunlight_proportion, k = 4) +
    s(minutes_above_2mps, k = 4)

Parametric coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) 3.598547   0.009186   391.7   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Approximate significance of smooth terms:
                         edf Ref.df   Chi.sq  p-value
s(abundance_t_minus_1) 2.997  3.000 11589.09  < 2e-16 ***
s(mean_wind_speed)     2.632  2.894    28.54 2.91e-06 ***
s(mean_temp)           2.971  2.999   185.98  < 2e-16 ***
s(sunlight_proportion) 2.935  2.996   316.19  < 2e-16 ***
s(minutes_above_2mps)  2.268  2.631    15.04  0.00909 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

R-sq.(adj) =  0.541   Deviance explained = 68.9%
-REML =  25422  Scale est. = 1         n = 5650


Diagnostic plots:
```




```
Method: REML   Optimizer: outer newton
full convergence after 9 iterations.
Gradient range [-0.0001926551,0.00133617]
(score 25421.97 & scale 1).
Hessian positive definite, eigenvalue range [0.577564,2175.759].
Model rank =  16 / 16

Basis dimension (k) checking results. Low p-value (k-index<1) may
indicate that k is too low, especially if edf is close to k'.

                        k'  edf k-index p-value
s(abundance_t_minus_1) 3.00 3.00    0.50  <2e-16 ***
s(mean_wind_speed)     3.00 2.63    0.93   0.015 *
s(mean_temp)           3.00 2.97    0.85  <2e-16 ***
s(sunlight_proportion) 3.00 2.93    0.89  <2e-16 ***
s(minutes_above_2mps)  3.00 2.27    0.93   0.030 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
# Model comparison table
comparison_120m <- data.frame(
```

```r
  Model = names(aic_values_120m),
  AIC = round(aic_values_120m, 2)
) %>%
  arrange(AIC)

cat("\n### Model Comparison (2-hour lag)\n")
```

```
### Model Comparison (2-hour lag)
```

```r
kable(comparison_120m)
```

| | Model | AIC |
|---|---|---|
| Full model | Full model | 50786.61 |
| Wind + Temp + Sunlight | Wind + Temp + Sunlight | 50796.94 |
| Wind + Temperature | Wind + Temperature | 51075.30 |
| Temperature | Temperature | 51111.31 |
| Wind speed | Wind speed | 51239.45 |
| Baseline (abundance only) | Baseline (abundance only) | 51283.38 |

---

### 4-Hour Lag Analysis

```r
data_240m <- prepare_lag_data(master_df, wind, lag_minutes = 240)
```

```
Preparing data for 240 minute lag...
Valid pairs after filtering: 6894
Calculating interval predictors efficiently...
Processing 6894 unique intervals...
Progress: 10 %
Progress: 20 %
Progress: 30 %
Progress: 40 %
Progress: 50 %
Progress: 60 %
Progress: 70 %
```

```
Progress: 80 %
Progress: 89.9 %
Progress: 99.9 %
Combining results...
Final dataset rows: 5663
Time delta range: 240 to 240 minutes
```

### Exploratory Data Analysis (4 hours)

```r
cat("Data structure:\n")
```

```
Data structure:
```

```r
glimpse(data_240m)
```

```
Rows: 5,663
Columns: 30
$ deployment_id.x           <chr> "SC1", "SC1", "SC1", "SC1", "SC1", "SC1", "S~
$ image_filename            <chr> "SC1_20231117184501.JPG", "SC1_2023111718500~
$ total_butterflies         <dbl> 54, 54, 54, 54, 54, 54, 54, 54, 54, 54, 54, ~
$ butterflies_direct_sun    <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,~
$ datetime                  <dttm> 2023-11-17 18:45:01, 2023-11-17 18:50:01, 2~
$ view_id                   <fct> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1,~
$ wind_meter_name           <chr> "JazzPlay", "JazzPlay", "JazzPlay", "JazzPla~
$ deployment_id.y           <chr> "SC1", "SC1", "SC1", "SC1", "SC1", "SC1", "S~
$ timestamp                 <dbl> 2.023112e+13, 2.023112e+13, 2.023112e+13, 2.~
$ temperature               <dbl> 15, 16, 15, 16, 15, 16, 16, 15, 16, 15, 16, ~
$ confidence                <dbl> 0.4041873, 0.5208792, 0.6782328, 0.5311629, ~
$ extraction_status         <chr> "success", "success", "success", "success", ~
$ datetime_t_minus_1        <dttm> 2023-11-17 14:45:01, 2023-11-17 14:50:01, 2~
$ abundance_t_minus_1       <dbl> 90, 80, 95, 58, 56, 56, 64, 40, 48, 59, 79, ~
$ butterflies_sun_t_minus_1 <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,~
$ temperature_t_minus_1     <dbl> 16, 16, 16, 16, 15, 16, 15, 15, 15, 15, 15, ~
$ time_delta_mins           <dbl> 240, 240, 240, 240, 240, 240, 240, 240, 240,~
$ mean_wind_speed           <dbl> 0.400000, 0.900000, 1.030769, 1.177778, 1.24~
$ max_wind_speed            <dbl> 0.4, 1.9, 1.9, 1.9, 1.9, 2.3, 2.7, 2.9, 2.9,~
$ sd_wind_speed             <dbl> 0.0000000, 0.6524678, 0.5297556, 0.5241863, ~
$ gust_factor               <dbl> 2.250000, 1.722222, 1.656716, 1.622642, 1.58~
$ minutes_above_2mps        <dbl> 0, 0, 0, 0, 0, 1, 3, 6, 7, 12, 16, 20, 25, 3~
$ mean_temp                 <dbl> 15.12245, 15.12245, 15.10204, 15.10204, 15.0~
```

```
$ max_temp                  <dbl> 16, 16, 16, 16, 16, 16, 16, 16, 16, 16, 16, ~
$ min_temp                  <dbl> 15, 15, 15, 15, 15, 15, 15, 15, 15, 15, 15, ~
$ sd_temp                   <dbl> 0.3312007, 0.3312007, 0.3058389, 0.3058389, ~
$ sunlight_proportion       <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,~
$ max_prop_sunlight         <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,~
$ min_prop_sunlight         <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,~
$ sd_prop_sunlight          <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,~
```

```r
cat("\n\nData summary:\n")
```

```
Data summary:
```

```r
summary(data_240m)
```

```
 deployment_id.x    image_filename     total_butterflies butterflies_direct_sun
 Length:5663        Length:5663        Min.   :  0.00    Min.   :  0.000
 Class :character   Class :character   1st Qu.: 12.00    1st Qu.:  0.000
 Mode  :character   Mode  :character   Median : 34.00    Median :  0.000
                                       Mean   : 64.02    Mean   :  1.585
                                       3rd Qu.: 72.00    3rd Qu.:  0.000
                                       Max.   :950.00    Max.   :150.000

    datetime                          view_id     wind_meter_name
 Min.   :2023-11-17 18:45:01.00   4      :2895   Length:5663
 1st Qu.:2023-12-10 17:30:01.00   2      :1115   Class :character
 Median :2023-12-28 22:50:01.00   1      : 674   Mode  :character
 Mean   :2023-12-28 21:36:27.57   5      : 464
 3rd Qu.:2024-01-23 05:05:01.00   11     : 287
 Max.   :2024-02-04 10:50:01.00   9      : 165
                                  (Other):  63
 deployment_id.y      timestamp          temperature     confidence
 Length:5663        Min.   :2.023e+13   Min.   : 4.0   Min.   :0.1222
 Class :character   1st Qu.:2.023e+13   1st Qu.:11.0   1st Qu.:0.4971
 Mode  :character   Median :2.023e+13   Median :13.0   Median :0.5940
                    Mean   :2.024e+13   Mean   :13.3   Mean   :0.5959
                    3rd Qu.:2.024e+13   3rd Qu.:15.0   3rd Qu.:0.6945
                    Max.   :2.024e+13   Max.   :33.0   Max.   :1.0000

 extraction_status  datetime_t_minus_1               abundance_t_minus_1
```

```
Length:5663        Min.   :2023-11-17 14:45:01.00   Min.   :  0.0
Class :character   1st Qu.:2023-12-10 13:30:01.00   1st Qu.: 14.0
Mode  :character   Median :2023-12-28 18:50:01.00   Median : 35.0
                   Mean   :2023-12-28 17:36:27.57   Mean   : 65.2
                   3rd Qu.:2024-01-23 01:05:01.00   3rd Qu.: 75.0
                   Max.   :2024-02-04 06:50:01.00   Max.   :950.0

butterflies_sun_t_minus_1 temperature_t_minus_1 time_delta_mins
Min.   :  0.000           Min.   : 4.00         Min.   :240
1st Qu.:  0.000           1st Qu.:11.00         1st Qu.:240
Median :  0.000           Median :13.00         Median :240
Mean   :  1.668           Mean   :13.32         Mean   :240
3rd Qu.:  0.000           3rd Qu.:15.00         3rd Qu.:240
Max.   :150.000           Max.   :31.00         Max.   :240

mean_wind_speed  max_wind_speed  sd_wind_speed     gust_factor
Min.   :0.0000   Min.   :0.000   Min.   :0.0000   Min.   :  1.205
1st Qu.:0.3140   1st Qu.:1.200   1st Qu.:0.2626   1st Qu.:  1.363
Median :0.7368   Median :1.800   Median :0.3848   Median :  1.442
Mean   :0.8833   Mean   :1.952   Mean   :0.3932   Mean   :  1.633
3rd Qu.:1.2775   3rd Qu.:2.500   3rd Qu.:0.5087   3rd Qu.:  1.559
Max.   :4.4979   Max.   :6.700   Max.   :1.5738   Max.   :102.000
                                 NA's   :1        NA's   :318
minutes_above_2mps   mean_temp         max_temp     min_temp
Min.   :  0.00     Min.   : 4.556   Min.   : 5   Min.   : 4.00
1st Qu.:  0.00     1st Qu.:11.160   1st Qu.:12   1st Qu.:10.00
Median :  0.00     Median :13.000   Median :15   Median :12.00
Mean   : 22.85     Mean   :13.300   Mean   :15   Mean   :11.76
3rd Qu.: 12.00     3rd Qu.:15.444   3rd Qu.:17   3rd Qu.:14.00
Max.   :240.00     Max.   :28.333   Max.   :33   Max.   :26.00

   sd_temp       sunlight_proportion max_prop_sunlight min_prop_sunlight
Min.   :0.0000   Min.   :0.00000     Min.   :0.0000    Min.   :0.00000
1st Qu.:0.5000   1st Qu.:0.00000     1st Qu.:0.0000    1st Qu.:0.00000
Median :0.8327   Median :0.00000     Median :0.0000    Median :0.00000
Mean   :1.1158   Mean   :0.06910     Mean   :0.2060    Mean   :0.01889
3rd Qu.:1.3944   3rd Qu.:0.03934     3rd Qu.:0.2308    3rd Qu.:0.00000
Max.   :7.3314   Max.   :1.00000     Max.   :1.0000    Max.   :1.00000

sd_prop_sunlight
Min.   :0.00000
1st Qu.:0.00000
Median :0.00000
```

```
 Mean   :0.06631
 3rd Qu.:0.05261
 Max.   :0.70711
 NA's   :21
```

```r
# Check for data quality issues
cat("\n\nData quality checks:\n")
```

```
Data quality checks:
```

```r
cat("Total rows:", nrow(data_240m), "\n")
```

```
Total rows: 5663
```

```r
cat("Complete cases:", sum(complete.cases(data_240m)), "\n")
```

```
Complete cases: 5324
```

```r
cat("Rows with abundance_t_minus_1 > 0:", sum(data_240m$abundance_t_minus_1 > 0, na.rm = TRU
```

```
Rows with abundance_t_minus_1 > 0: 5508
```

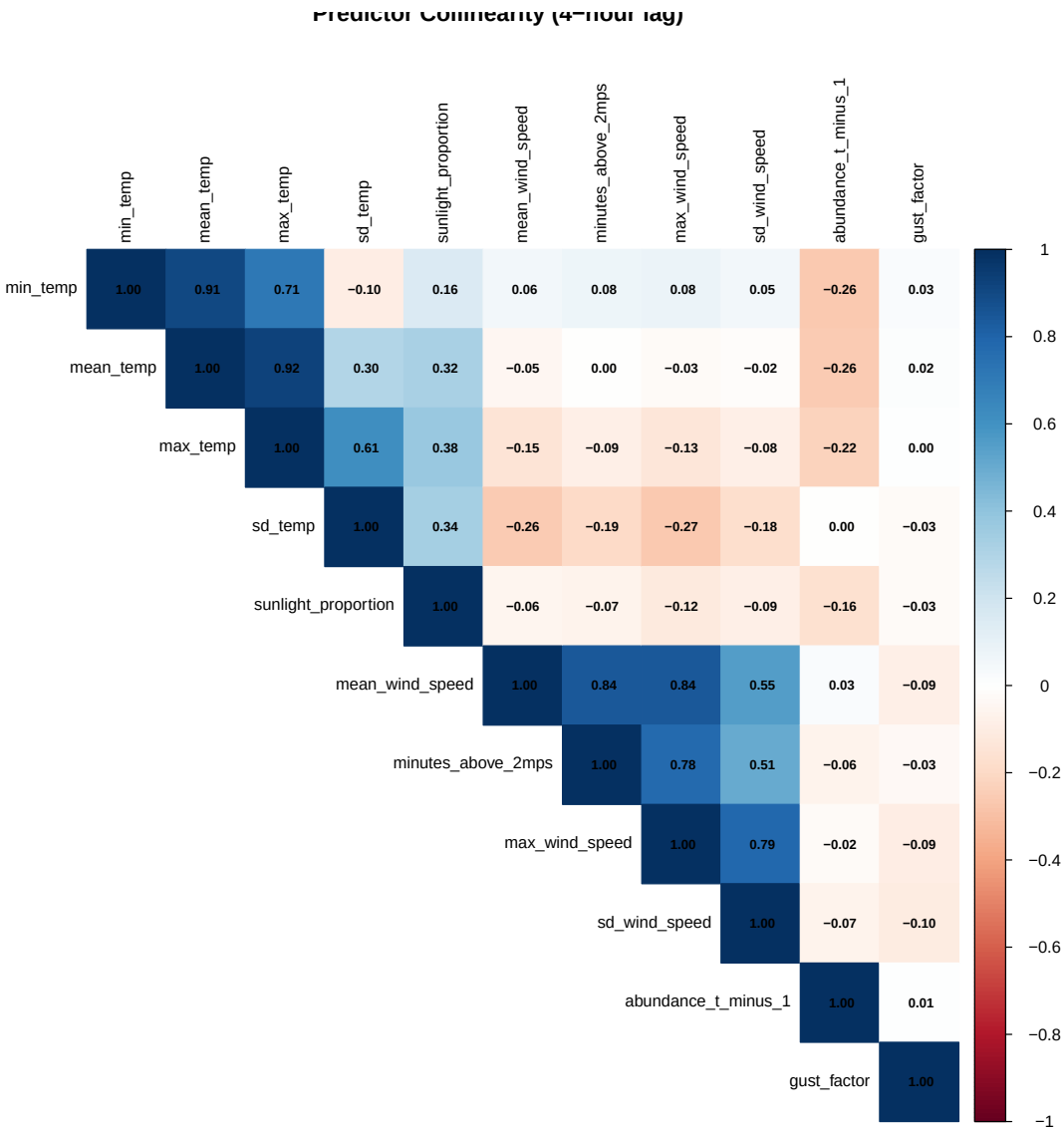
```r
cat("Rows with total_butterflies > 0:", sum(data_240m$total_butterflies > 0, na.rm = TRUE), 
```

```
Rows with total_butterflies > 0: 5437
```

```r
cat("Unique view_ids:", length(unique(data_240m$view_id)), "\n")
```

```
Unique view_ids: 9
```

```r
# Select numeric predictors for correlation analysis
predictors_240m <- data_240m %>%
  select(abundance_t_minus_1, mean_wind_speed, max_wind_speed, sd_wind_speed,
         gust_factor, minutes_above_2mps, mean_temp, max_temp, min_temp, sd_temp,
         sunlight_proportion) %>%
  na.omit()

if (nrow(predictors_240m) > 1) {
  cor_matrix_240m <- cor(predictors_240m, use = "complete.obs")
  if (!any(is.na(cor_matrix_240m) | is.infinite(cor_matrix_240m))) {
    corrplot(cor_matrix_240m, method = "color", type = "upper",
             order = "hclust", tl.cex = 0.8, tl.col = "black",
             addCoef.col = "black", number.cex = 0.7,
             title = "Predictor Collinearity (4-hour lag)")
  } else {
    corrplot(cor_matrix_240m, method = "color", type = "upper",
             order = "original", tl.cex = 0.8, tl.col = "black",
             addCoef.col = "black", number.cex = 0.7,
             title = "Predictor Collinearity (4-hour lag)")
  }
} else {
  cat("Not enough data for correlation analysis\n")
}
```



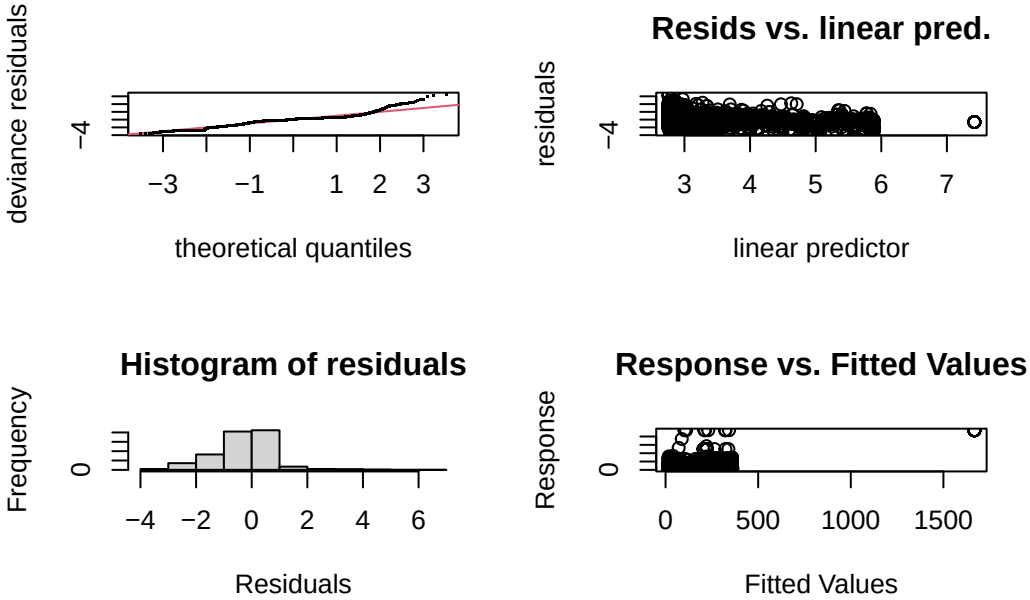

## Model Fitting (4 hours)

```
# Use same model formulas
model_formulas_240m <- model_formulas_30m

# Storage for results
models_240m <- list()
```

```r
aic_values_240m <- numeric()

# Fit each model
for (i in seq_along(model_formulas_240m)) {
  model_name <- names(model_formulas_240m)[i]
  formula <- model_formulas_240m[[i]]

  cat("\n### Model", i, ":", model_name, "\n\n")

  tryCatch({
    # Check data before fitting
    model_vars <- all.vars(formula)
    model_data <- data_240m %>%
      select(all_of(model_vars), view_id) %>%
      na.omit()

    cat("Model data dimensions after removing NAs:", nrow(model_data), "x", ncol(model_data)

    # Fit GAM
    gam_model <- gam(
      formula,
      family = nb(),
      data = model_data
    )

    cat("GAM fitted successfully\n")

    # Store model
    models_240m[[model_name]] <- gam_model
    aic_values_240m[model_name] <- AIC(gam_model)

    # Print summary
    cat("Model Summary:\n")
    print(summary(gam_model))

    # Diagnostic plots
    cat("\n\nDiagnostic plots:\n")
    gam.check(gam_model)

  }, error = function(e) {
    cat("Error fitting model:", e$message, "\n")
    aic_values_240m[model_name] <- NA
```

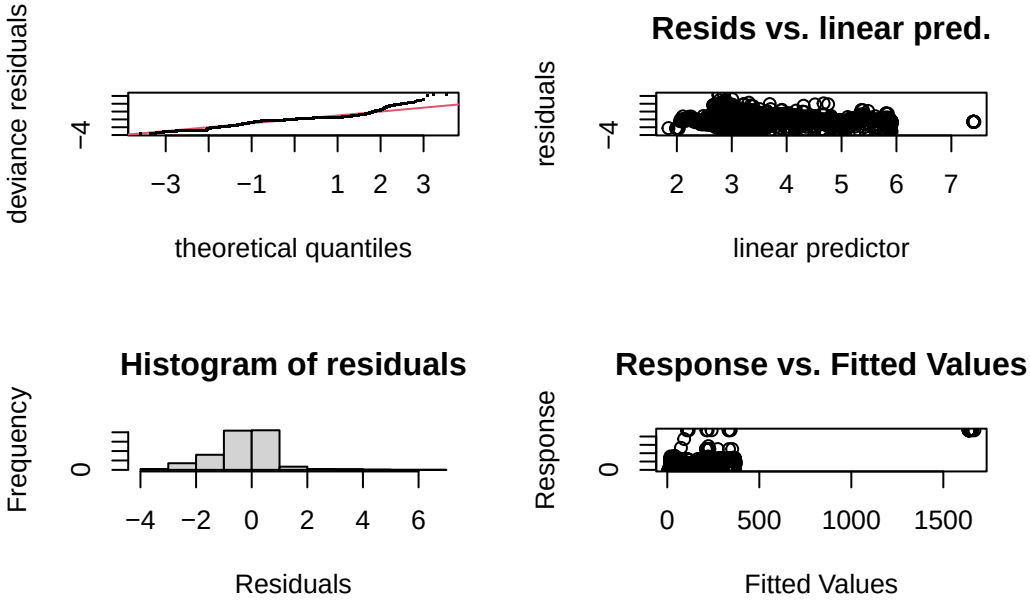
```
  })
}
```

```
### Model 1 : Baseline (abundance only)

Model data dimensions after removing NAs: 5663 x 3
GAM fitted successfully
Model Summary:

Family: Negative Binomial(1.476)
Link function: log

Formula:
total_butterflies ~ s(abundance_t_minus_1, k = 4)

Parametric coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  3.66365    0.01121   326.7   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Approximate significance of smooth terms:
                         edf Ref.df Chi.sq p-value
s(abundance_t_minus_1) 2.993      3   7128  <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

R-sq.(adj) =  0.531   Deviance explained = 55.1%
-REML =  26330  Scale est. = 1         n = 5663


Diagnostic plots:
```



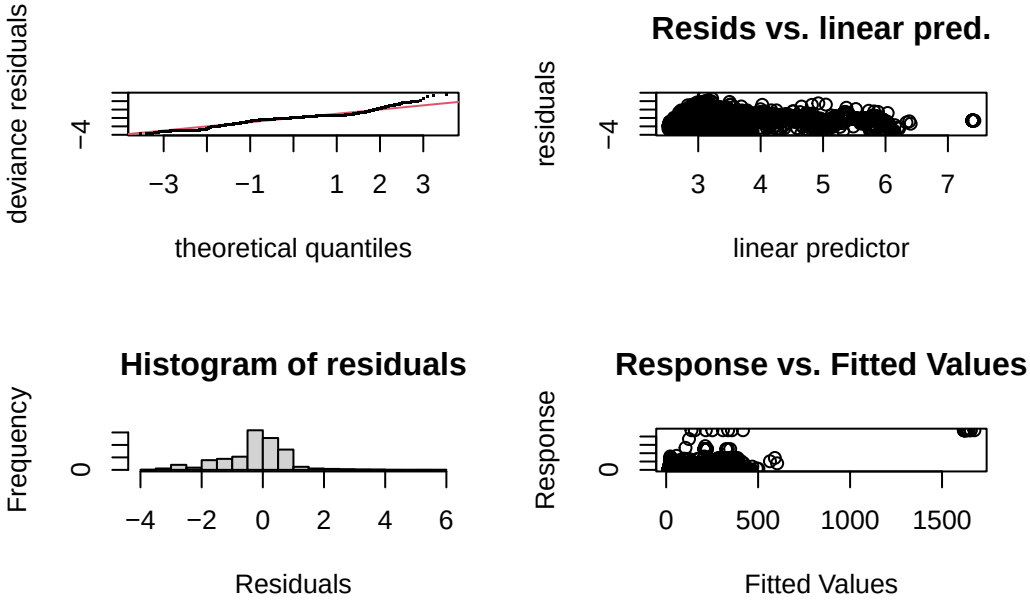

```
Method: REML   Optimizer: outer newton
full convergence after 6 iterations.
Gradient range [-0.000312979,0.01757181]
(score 26330.22 & scale 1).
Hessian positive definite, eigenvalue range [1.01065,2537.847].
Model rank =  4 / 4

Basis dimension (k) checking results. Low p-value (k-index<1) may
indicate that k is too low, especially if edf is close to k'.

                         k'  edf k-index p-value
s(abundance_t_minus_1) 3.00 2.99    0.67  <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

### Model 2 : Wind speed

Model data dimensions after removing NAs: 5663 x 4
GAM fitted successfully
Model Summary:

Family: Negative Binomial(1.489)
```

```
Link function: log

Formula:
total_butterflies ~ s(abundance_t_minus_1, k = 4) + s(mean_wind_speed,
    k = 4)

Parametric coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  3.66032    0.01117   327.7   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Approximate significance of smooth terms:
                         edf Ref.df  Chi.sq p-value
s(abundance_t_minus_1) 2.993  3.000 6996.63  <2e-16 ***
s(mean_wind_speed)     2.831  2.979   59.54  <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

R-sq.(adj) =   0.54   Deviance explained = 55.5%
-REML =  26311  Scale est. = 1         n = 5663


Diagnostic plots:
```



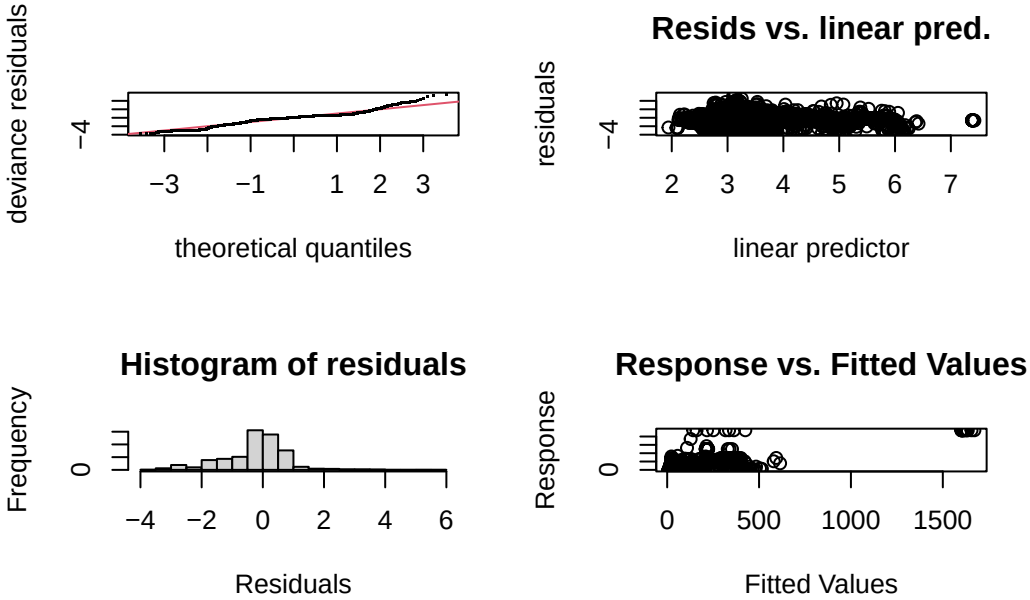

```
Method: REML   Optimizer: outer newton
full convergence after 7 iterations.
Gradient range [-1.937476e-05,0.0008045332]
(score 26311.3 & scale 1).
Hessian positive definite, eigenvalue range [0.7897879,2527.983].
Model rank =  7 / 7

Basis dimension (k) checking results. Low p-value (k-index<1) may
indicate that k is too low, especially if edf is close to k'.

                        k'  edf k-index p-value
s(abundance_t_minus_1) 3.00 2.99    0.64  <2e-16 ***
s(mean_wind_speed)     3.00 2.83    0.87  <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

### Model 3 : Temperature

Model data dimensions after removing NAs: 5663 x 4
GAM fitted successfully
Model Summary:
```

```
Family: Negative Binomial(1.543)
Link function: log

Formula:
total_butterflies ~ s(abundance_t_minus_1, k = 4) + s(mean_temp,
    k = 4)

Parametric coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  3.64829    0.01099     332   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Approximate significance of smooth terms:
                         edf Ref.df Chi.sq p-value
s(abundance_t_minus_1) 2.994  3.000 7736.8  <2e-16 ***
s(mean_temp)           2.952  2.998  242.9  <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

R-sq.(adj) =  0.511   Deviance explained =   57%
-REML =  26211  Scale est. = 1         n = 5663


Diagnostic plots:
```



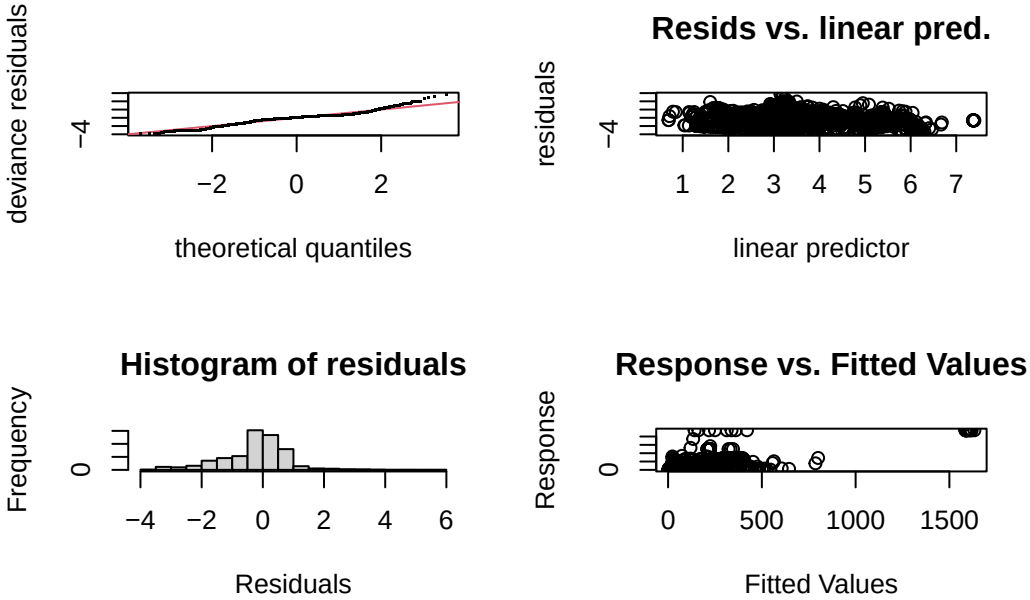

```
Method: REML   Optimizer: outer newton
full convergence after 8 iterations.
Gradient range [-0.0007511866,0.0200039]
(score 26211.32 & scale 1).
Hessian positive definite, eigenvalue range [0.9081812,2496.277].
Model rank =  7 / 7

Basis dimension (k) checking results. Low p-value (k-index<1) may
indicate that k is too low, especially if edf is close to k'.

                        k'  edf k-index p-value
s(abundance_t_minus_1) 3.00 2.99    0.64  <2e-16 ***
s(mean_temp)           3.00 2.95    0.85  <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

### Model 4 : Wind + Temperature

Model data dimensions after removing NAs: 5663 x 5
GAM fitted successfully
Model Summary:
```

```
Family: Negative Binomial(1.556)
Link function: log

Formula:
total_butterflies ~ s(abundance_t_minus_1, k = 4) + s(mean_wind_speed,
    k = 4) + s(mean_temp, k = 4)

Parametric coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  3.64540    0.01095   332.9   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Approximate significance of smooth terms:
                         edf Ref.df Chi.sq p-value
s(abundance_t_minus_1) 2.994  3.000 7619.5  <2e-16 ***
s(mean_wind_speed)     2.562  2.869   51.0  <2e-16 ***
s(mean_temp)           2.958  2.999  241.2  <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

R-sq.(adj) =  0.508   Deviance explained = 57.3%
-REML =  26194  Scale est. = 1         n = 5663


Diagnostic plots:
```



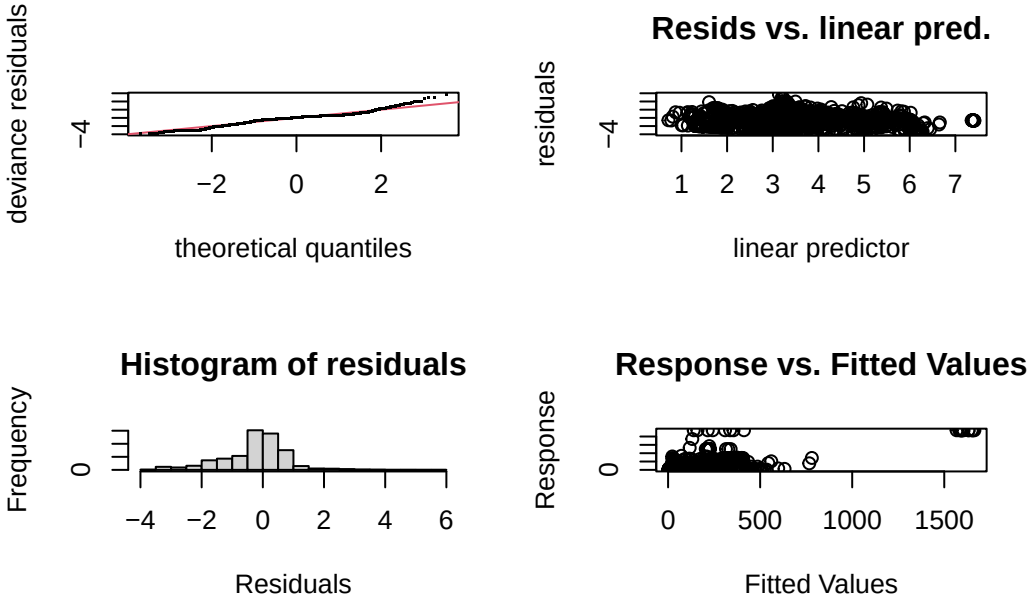

```
Method: REML   Optimizer: outer newton
full convergence after 8 iterations.
Gradient range [-0.0001329546,0.003192056]
(score 26194.01 & scale 1).
Hessian positive definite, eigenvalue range [0.7880008,2487.571].
Model rank =  10 / 10

Basis dimension (k) checking results. Low p-value (k-index<1) may
indicate that k is too low, especially if edf is close to k'.

                         k'  edf k-index p-value
s(abundance_t_minus_1) 3.00 2.99    0.66  <2e-16 ***
s(mean_wind_speed)     3.00 2.56    0.88  <2e-16 ***
s(mean_temp)           3.00 2.96    0.84  <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

### Model 5 : Wind + Temp + Sunlight

Model data dimensions after removing NAs: 5663 x 6
GAM fitted successfully
Model Summary:
```

```
Family: Negative Binomial(1.666)
Link function: log

Formula:
total_butterflies ~ s(abundance_t_minus_1, k = 4) + s(mean_wind_speed,
    k = 4) + s(mean_temp, k = 4) + s(sunlight_proportion, k = 4)

Parametric coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  3.62249    0.01063   340.7   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Approximate significance of smooth terms:
                         edf Ref.df  Chi.sq p-value
s(abundance_t_minus_1) 2.994  3.000 7851.19  <2e-16 ***
s(mean_wind_speed)     2.678  2.927   52.15  <2e-16 ***
s(mean_temp)           2.934  2.997  252.64  <2e-16 ***
s(sunlight_proportion) 2.934  2.996  474.05  <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

R-sq.(adj) =  0.484   Deviance explained = 59.9%
-REML =  26014  Scale est. = 1         n = 5663


Diagnostic plots:
```




```
Method: REML   Optimizer: outer newton
full convergence after 7 iterations.
Gradient range [-1.83353e-05,0.0004124176]
(score 26013.73 & scale 1).
Hessian positive definite, eigenvalue range [0.85644,2424.168].
Model rank =  13 / 13

Basis dimension (k) checking results. Low p-value (k-index<1) may
indicate that k is too low, especially if edf is close to k'.

                         k'  edf k-index p-value
s(abundance_t_minus_1) 3.00 2.99    0.65  <2e-16 ***
s(mean_wind_speed)     3.00 2.68    0.90  <2e-16 ***
s(mean_temp)           3.00 2.93    0.88  <2e-16 ***
s(sunlight_proportion) 3.00 2.93    0.86  <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

### Model 6 : Full model

Model data dimensions after removing NAs: 5663 x 7
GAM fitted successfully
```

```
Model Summary:

Family: Negative Binomial(1.671)
Link function: log

Formula:
total_butterflies ~ s(abundance_t_minus_1, k = 4) + s(mean_wind_speed,
    k = 4) + s(mean_temp, k = 4) + s(sunlight_proportion, k = 4) +
    s(minutes_above_2mps, k = 4)

Parametric coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  3.62134    0.01062   341.1   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Approximate significance of smooth terms:
                         edf Ref.df  Chi.sq p-value
s(abundance_t_minus_1) 2.995  3.000 7639.58 < 2e-16 ***
s(mean_wind_speed)     2.794  2.962   39.97 < 2e-16 ***
s(mean_temp)           2.930  2.996  251.95 < 2e-16 ***
s(sunlight_proportion) 2.932  2.996  478.66 < 2e-16 ***
s(minutes_above_2mps)  2.279  2.652   18.24 0.00214 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

R-sq.(adj) =  0.474   Deviance explained =   60%
-REML =  26009  Scale est. = 1         n = 5663


Diagnostic plots:
```




```
Method: REML   Optimizer: outer newton
full convergence after 7 iterations.
Gradient range [-0.000543766,0.01970619]
(score 26009.4 & scale 1).
Hessian positive definite, eigenvalue range [0.6096667,2420.569].
Model rank =  16 / 16

Basis dimension (k) checking results. Low p-value (k-index<1) may
indicate that k is too low, especially if edf is close to k'.

                         k'  edf k-index p-value
s(abundance_t_minus_1) 3.00 2.99    0.62  <2e-16 ***
s(mean_wind_speed)     3.00 2.79    0.88  <2e-16 ***
s(mean_temp)           3.00 2.93    0.87  <2e-16 ***
s(sunlight_proportion) 3.00 2.93    0.88  <2e-16 ***
s(minutes_above_2mps)  3.00 2.28    0.93    0.38
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
# Model comparison table
comparison_240m <- data.frame(
```

```
  Model = names(aic_values_240m),
  AIC = round(aic_values_240m, 2)
) %>%
  arrange(AIC)

cat("\n### Model Comparison (4-hour lag)\n")
```

```
### Model Comparison (4-hour lag)
```

```
kable(comparison_240m)
```

| | Model | AIC |
|---|---|---|
| Full model | Full model | 51965.27 |
| Wind + Temp + Sunlight | Wind + Temp + Sunlight | 51979.85 |
| Wind + Temperature | Wind + Temperature | 52351.12 |
| Temperature | Temperature | 52392.53 |
| Wind speed | Wind speed | 52595.21 |
| Baseline (abundance only) | Baseline (abundance only) | 52641.53 |

---

### Part 5: Full Day Lag Analysis

```
cat("Preparing full day lag data...\n")
```

```
Preparing full day lag data...
```

```
# Group by view_id and date
full_day_data <- master_df %>%
  filter(!is.na(datetime) & !is.na(view_id)) %>%
  mutate(date = as.Date(datetime)) %>%
  group_by(view_id, date) %>%
  arrange(datetime) %>%
  mutate(
    # Find first observation with butterflies > 0
    has_butterflies = total_butterflies > 0,
```

```r
    is_first_with_butterflies = has_butterflies & !lag(has_butterflies, default = FALSE)
  ) %>%
  filter(any(is_first_with_butterflies)) %>%  # Keep only days with at least one butterfly ob
  summarise(
    # t-1 observation (first with butterflies)
    datetime_t_minus_1 = datetime[which(is_first_with_butterflies)[1]],
    abundance_t_minus_1 = total_butterflies[which(is_first_with_butterflies)[1]],

    # t observation (last of day)
    datetime = last(datetime),
    total_butterflies = last(total_butterflies),

    # Keep wind meter name for joining
    wind_meter_name = first(wind_meter_name),

    .groups = "drop"
  ) %>%
  # Calculate time delta and filter
  mutate(
    time_delta_hours = as.numeric(difftime(datetime, datetime_t_minus_1, units = "hours"))
  ) %>%
  filter(time_delta_hours > 4)

cat("Full day pairs found:", nrow(full_day_data), "\n")
```

```
Full day pairs found: 125
```

```r
cat("Time delta range:", round(min(full_day_data$time_delta_hours), 1),
    "to", round(max(full_day_data$time_delta_hours), 1), "hours\n")
```

```
Time delta range: 7 to 23.9 hours
```

```r
# Calculate environmental predictors for full day intervals
full_day_with_predictors <- full_day_data %>%
  rowwise() %>%
  mutate(
    # Wind metrics
    wind_data = list({
      wind_subset <- wind %>%
        filter(
```

```
        wind_meter_name == .data$wind_meter_name,
        time >= datetime_t_minus_1,
        time <= datetime
      )
    if (nrow(wind_subset) > 0) {
      list(
        mean_wind_speed = mean(wind_subset$speed, na.rm = TRUE),
        max_wind_speed = max(wind_subset$speed, na.rm = TRUE),
        sd_wind_speed = sd(wind_subset$speed, na.rm = TRUE),
        gust_factor = mean(wind_subset$gust, na.rm = TRUE) / mean(wind_subset$speed, na.rm
        minutes_above_2mps = sum(wind_subset$speed > 2, na.rm = TRUE)
      )
    } else {
      list(
        mean_wind_speed = NA_real_,
        max_wind_speed = NA_real_,
        sd_wind_speed = NA_real_,
        gust_factor = NA_real_,
        minutes_above_2mps = NA_real_
      )
    }
  }),
  mean_wind_speed = wind_data$mean_wind_speed,
  max_wind_speed = wind_data$max_wind_speed,
  sd_wind_speed = wind_data$sd_wind_speed,
  gust_factor = wind_data$gust_factor,
  minutes_above_2mps = wind_data$minutes_above_2mps,

  # Temperature metrics
  temp_data = list({
    temp_subset <- master_df %>%
      filter(
        view_id == .data$view_id,
        datetime >= datetime_t_minus_1,
        datetime <= .data$datetime
      )
    if (nrow(temp_subset) > 0) {
      list(
        mean_temp = mean(temp_subset$temperature, na.rm = TRUE),
        max_temp = max(temp_subset$temperature, na.rm = TRUE),
        min_temp = min(temp_subset$temperature, na.rm = TRUE),
        sd_temp = sd(temp_subset$temperature, na.rm = TRUE)
```

```
      )
    } else {
      list(
        mean_temp = NA_real_,
        max_temp = NA_real_,
        min_temp = NA_real_,
        sd_temp = NA_real_
      )
    }
  }),
  mean_temp = temp_data$mean_temp,
  max_temp = temp_data$max_temp,
  min_temp = temp_data$min_temp,
  sd_temp = temp_data$sd_temp,

  # Sunlight metrics
  sun_data = list({
    sun_subset <- master_df %>%
      filter(
        view_id == .data$view_id,
        datetime >= datetime_t_minus_1,
        datetime <= .data$datetime
      )
    if (nrow(sun_subset) > 0) {
      total_sun <- sum(sun_subset$butterflies_direct_sun, na.rm = TRUE)
      total_count <- sum(sun_subset$total_butterflies, na.rm = TRUE)
      if (total_count > 0) {
        overall_prop <- total_sun / total_count
        # Calculate per-observation proportions
        obs_props <- ifelse(sun_subset$total_butterflies > 0,
                            sun_subset$butterflies_direct_sun / sun_subset$total_butterflies
                            NA_real_)
        list(
          sunlight_proportion = overall_prop,
          max_prop = max(obs_props, na.rm = TRUE),
          min_prop = min(obs_props, na.rm = TRUE),
          sd_prop = sd(obs_props, na.rm = TRUE)
        )
      } else {
        list(
          sunlight_proportion = NA_real_,
          max_prop = NA_real_,
```

```
            min_prop = NA_real_,
            sd_prop = NA_real_
          )
        }
      } else {
        list(
          sunlight_proportion = NA_real_,
          max_prop = NA_real_,
          min_prop = NA_real_,
          sd_prop = NA_real_
        )
      }
    }),
    sunlight_proportion = sun_data$sunlight_proportion,
    max_prop_sunlight = sun_data$max_prop,
    min_prop_sunlight = sun_data$min_prop,
    sd_prop_sunlight = sun_data$sd_prop
  ) %>%
  ungroup() %>%
  select(-wind_data, -temp_data, -sun_data)

# Filter for complete cases
data_full_day <- full_day_with_predictors %>%
  filter(!is.na(mean_wind_speed) & !is.na(mean_temp))

cat("Final full day dataset rows:", nrow(data_full_day), "\n")
```

```
Final full day dataset rows: 125
```

### Exploratory Data Analysis (Full Day)

```
cat("Data structure:\n")
```

```
Data structure:
```

```
glimpse(data_full_day)
```

```
Rows: 125
```

```
Columns: 21
$ view_id             <fct> 1, 1, 1, 1, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2,~
$ date                <date> 2023-11-17, 2023-11-18, 2023-11-19, 2023-11-20, 2~
$ datetime_t_minus_1  <dttm> 2023-11-17 11:40:01, 2023-11-18 00:00:01, 2023-11~
$ abundance_t_minus_1 <dbl> 121, 54, 19, 2, 436, 499, 206, 1, 2, 168, 236, 26,~
$ datetime            <dttm> 2023-11-17 23:55:01, 2023-11-18 23:55:01, 2023-11~
$ total_butterflies   <dbl> 54, 19, 2, 0, 499, 206, 0, 2, 168, 236, 26, 15, 37~
$ wind_meter_name     <chr> "JazzPlay", "JazzPlay", "JazzPlay", "JazzPlay", "S~
$ time_delta_hours    <dbl> 12.25000, 23.91667, 23.83333, 23.83333, 9.00000, 2~
$ mean_wind_speed     <dbl> 2.2716738, 1.4147838, 0.1985714, 1.1108467, 0.1808~
$ max_wind_speed      <dbl> 4.2, 5.0, 2.0, 4.3, 1.8, 2.3, 2.8, 3.4, 3.3, 3.0, ~
$ sd_wind_speed       <dbl> 0.7711548, 0.8540580, 0.3589021, 0.7926403, 0.3206~
$ gust_factor         <dbl> 1.401285, 1.440162, 2.217672, 1.541010, 2.034462, ~
$ minutes_above_2mps  <int> 280, 683, 0, 137, 0, 5, 25, 85, 156, 173, 195, 138~
$ mean_temp           <dbl> 13.25099, 13.18943, 13.02735, 12.97943, 13.02589, ~
$ max_temp            <dbl> 37, 37, 37, 37, 37, 37, 37, 37, 37, 37, 37, 37, 37~
$ min_temp            <dbl> 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3,~
$ sd_temp             <dbl> 3.951748, 3.971907, 4.000563, 4.047503, 4.219018, ~
$ sunlight_proportion <dbl> 0.02921473, 0.03045164, 0.03255667, 0.03318173, 0.~
$ max_prop_sunlight   <dbl> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1,~
$ min_prop_sunlight   <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,~
$ sd_prop_sunlight    <dbl> 0.2328945, 0.2368612, 0.2461221, 0.2520011, 0.2523~
```

```r
cat("\n\nData summary:\n")
```

```
Data summary:
```

```r
summary(data_full_day)
```

```
    view_id        date             datetime_t_minus_1
 2      :55   Min.   :2023-11-17   Min.   :2023-11-17 11:40:01.00
 4      :52   1st Qu.:2023-12-22   1st Qu.:2023-12-22 00:20:01.00
 1      : 4   Median :2024-01-05   Median :2024-01-05 00:00:01.00
 9      : 4   Mean   :2024-01-03   Mean   :2024-01-03 06:08:56.68
 5      : 3   3rd Qu.:2024-01-19   3rd Qu.:2024-01-19 00:20:01.00
 11     : 3   Max.   :2024-02-04   Max.   :2024-02-04 00:00:01.00
 (Other): 4
 abundance_t_minus_1    datetime                      total_butterflies
 Min.   :  1.00      Min.   :2023-11-17 23:55:01.0   Min.   :  0.0
```

```
 1st Qu.:  6.00      1st Qu.:2023-12-22 23:50:01.0   1st Qu.:  3.0
 Median : 37.00      Median :2024-01-05 23:36:04.0   Median : 26.0
 Mean   : 78.16      Mean   :2024-01-04 03:15:34.0   Mean   : 77.6
 3rd Qu.:109.00      3rd Qu.:2024-01-19 23:50:01.0   3rd Qu.: 91.0
 Max.   :499.00      Max.   :2024-02-04 10:50:01.0   Max.   :950.0

 wind_meter_name    time_delta_hours mean_wind_speed  max_wind_speed
 Length:125         Min.   : 7.00    Min.   :0.1012   Min.   :1.600
 Class :character   1st Qu.:23.50    1st Qu.:0.3617   1st Qu.:2.900
 Mode  :character   Median :23.50    Median :0.6096   Median :3.400
                    Mean   :21.11    Mean   :0.7048   Mean   :3.836
                    3rd Qu.:23.50    3rd Qu.:0.8764   3rd Qu.:4.300
                    Max.   :23.92    Max.   :2.3941   Max.   :9.700

 sd_wind_speed     gust_factor    minutes_above_2mps   mean_temp
 Min.   :0.2398   Min.   :1.346   Min.   :   0.0     Min.   : 9.686
 1st Qu.:0.5133   1st Qu.:1.476   1st Qu.:  35.0     1st Qu.:12.820
 Median :0.6241   Median :1.518   Median : 177.0     Median :13.019
 Mean   :0.6488   Mean   :1.542   Mean   : 406.6     Mean   :13.124
 3rd Qu.:0.7849   3rd Qu.:1.594   3rd Qu.: 558.0     3rd Qu.:13.658
 Max.   :1.4544   Max.   :2.218   Max.   :3876.0     Max.   :14.397

    max_temp        min_temp        sd_temp       sunlight_proportion
 Min.   :11.00   Min.   :3.000   Min.   :0.5333   Min.   :0.002424
 1st Qu.:37.00   1st Qu.:3.000   1st Qu.:4.3022   1st Qu.:0.033951
 Median :37.00   Median :3.000   Median :4.5068   Median :0.034954
 Mean   :35.66   Mean   :3.768   Mean   :4.4176   Mean   :0.055235
 3rd Qu.:37.00   3rd Qu.:5.000   3rd Qu.:4.6698   3rd Qu.:0.074669
 Max.   :37.00   Max.   :9.000   Max.   :5.4154   Max.   :0.154394

 max_prop_sunlight min_prop_sunlight sd_prop_sunlight
 Min.   :0.2000    Min.   :0         Min.   :0.03306
 1st Qu.:1.0000    1st Qu.:0         1st Qu.:0.24936
 Median :1.0000    Median :0         Median :0.25270
 Mean   :0.9931    Mean   :0         Mean   :0.25218
 3rd Qu.:1.0000    3rd Qu.:0         3rd Qu.:0.25648
 Max.   :1.0000    Max.   :0         Max.   :0.29015
```

```r
# Check for data quality issues
cat("\n\nData quality checks:\n")
```

```
Data quality checks:
```

```r
cat("Total rows:", nrow(data_full_day), "\n")
```

```
Total rows: 125
```

```r
cat("Complete cases:", sum(complete.cases(data_full_day)), "\n")
```

```
Complete cases: 125
```

```r
cat("Rows with abundance_t_minus_1 > 0:", sum(data_full_day$abundance_t_minus_1 > 0, na.rm =
```

```
Rows with abundance_t_minus_1 > 0: 125
```

```r
cat("Rows with total_butterflies > 0:", sum(data_full_day$total_butterflies > 0, na.rm = TRU
```

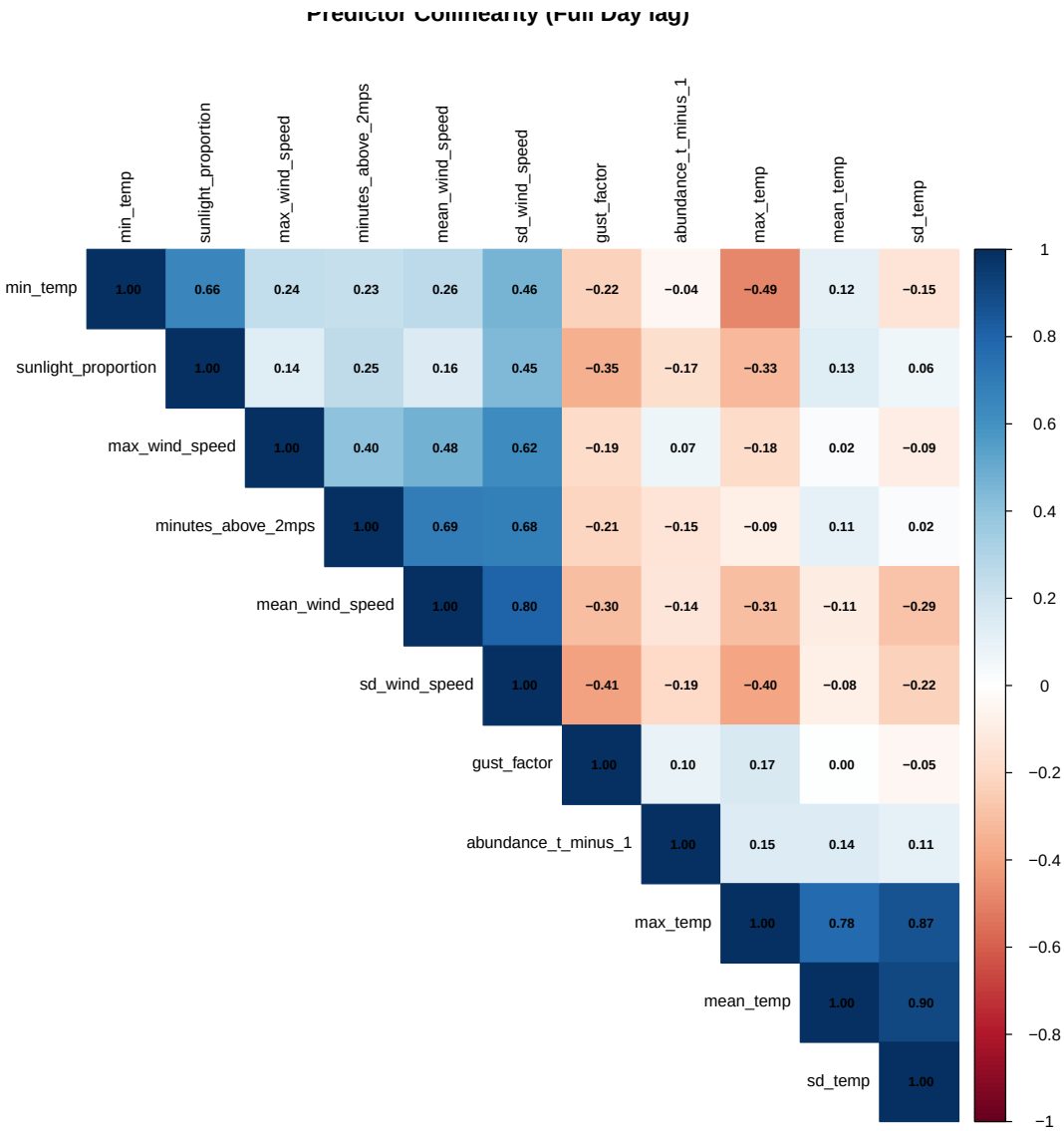
```
Rows with total_butterflies > 0: 108
```

```r
cat("Unique view_ids:", length(unique(data_full_day$view_id)), "\n")
```

```
Unique view_ids: 9
```

```r
# Select numeric predictors for correlation analysis
predictors_full_day <- data_full_day %>%
  select(abundance_t_minus_1, mean_wind_speed, max_wind_speed, sd_wind_speed,
         gust_factor, minutes_above_2mps, mean_temp, max_temp, min_temp, sd_temp,
         sunlight_proportion) %>%
  na.omit()

if (nrow(predictors_full_day) > 1) {
  cor_matrix_full_day <- cor(predictors_full_day, use = "complete.obs")
  if (!any(is.na(cor_matrix_full_day) | is.infinite(cor_matrix_full_day))) {
    corrplot(cor_matrix_full_day, method = "color", type = "upper",
             order = "hclust", tl.cex = 0.8, tl.col = "black",
             addCoef.col = "black", number.cex = 0.7,
             title = "Predictor Collinearity (Full Day lag)")
```

```r
  } else {
    corrplot(cor_matrix_full_day, method = "color", type = "upper",
             order = "original", tl.cex = 0.8, tl.col = "black",
             addCoef.col = "black", number.cex = 0.7,
             title = "Predictor Collinearity (Full Day lag)")
  }
} else {
  cat("Not enough data for correlation analysis\n")
}
```



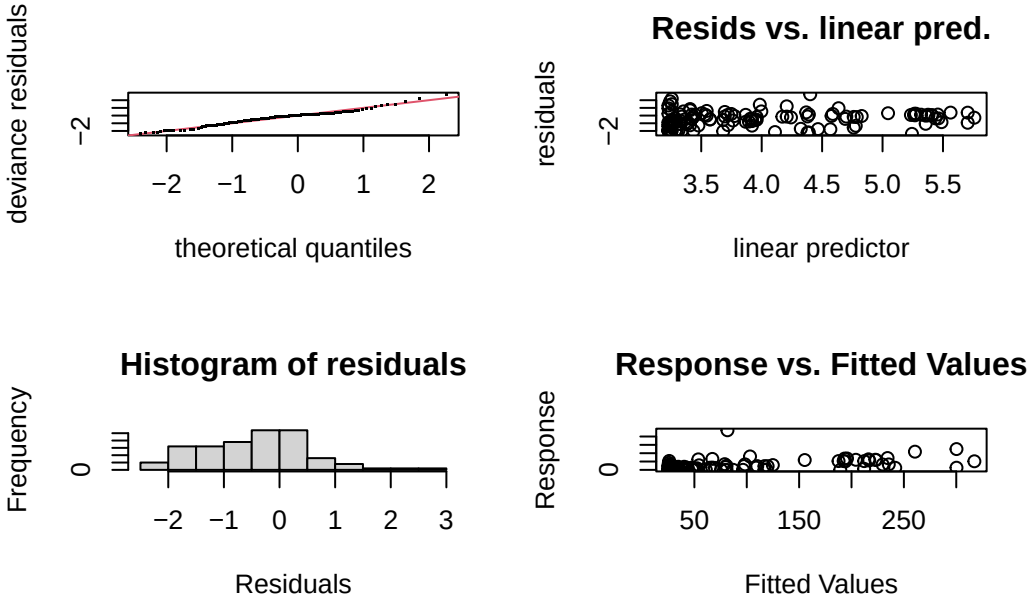

## Model Fitting (Full Day)

```
# Use same model formulas
model_formulas_full_day <- model_formulas_30m

# Storage for results
models_full_day <- list()
```

```r
aic_values_full_day <- numeric()

# Fit each model
for (i in seq_along(model_formulas_full_day)) {
  model_name <- names(model_formulas_full_day)[i]
  formula <- model_formulas_full_day[[i]]

  cat("\n### Model", i, ":", model_name, "\n\n")

  tryCatch({
    # Check data before fitting
    model_vars <- all.vars(formula)
    model_data <- data_full_day %>%
      select(all_of(model_vars), view_id) %>%
      na.omit()

    cat("Model data dimensions after removing NAs:", nrow(model_data), "x", ncol(model_data)

    # Fit GAM
    gam_model <- gam(
      formula,
      family = nb(),
      data = model_data
    )

    cat("GAM fitted successfully\n")

    # Store model
    models_full_day[[model_name]] <- gam_model
    aic_values_full_day[model_name] <- AIC(gam_model)

    # Print summary
    cat("Model Summary:\n")
    print(summary(gam_model))

    # Diagnostic plots
    cat("\n\nDiagnostic plots:\n")
    gam.check(gam_model)

  }, error = function(e) {
    cat("Error fitting model:", e$message, "\n")
    aic_values_full_day[model_name] <- NA
```

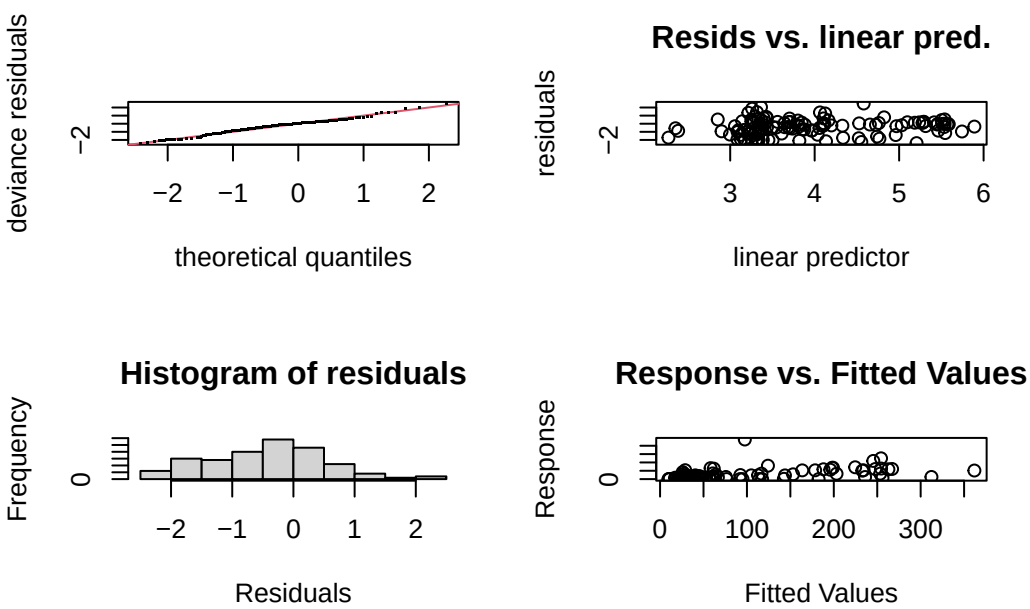
```
  })
}
```

```
### Model 1 : Baseline (abundance only)

Model data dimensions after removing NAs: 125 x 3
GAM fitted successfully
Model Summary:

Family: Negative Binomial(0.471)
Link function: log

Formula:
total_butterflies ~ s(abundance_t_minus_1, k = 4)

Parametric coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)    4.001      0.131   30.53   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Approximate significance of smooth terms:
                         edf Ref.df Chi.sq p-value
s(abundance_t_minus_1) 2.147  2.545  38.62  <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

R-sq.(adj) =  0.278   Deviance explained = 21.8%
-REML =    609  Scale est. = 1         n = 125


Diagnostic plots:
```



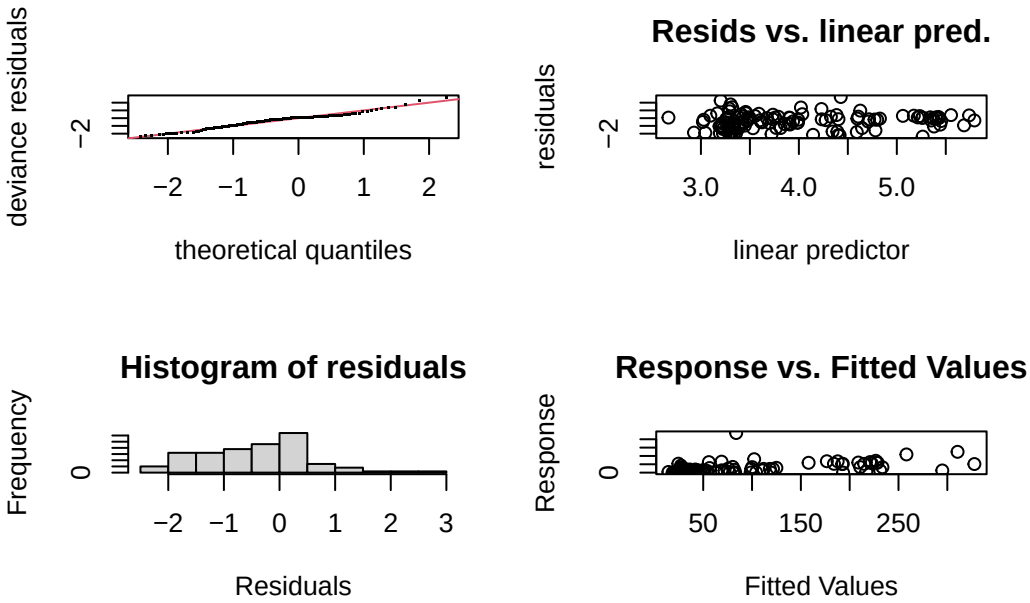

```
Method: REML   Optimizer: outer newton
full convergence after 4 iterations.
Gradient range [1.191524e-05,0.0003513138]
(score 609 & scale 1).
Hessian positive definite, eigenvalue range [0.4385614,67.78837].
Model rank =  4 / 4

Basis dimension (k) checking results. Low p-value (k-index<1) may
indicate that k is too low, especially if edf is close to k'.

                         k'  edf k-index p-value
s(abundance_t_minus_1) 3.00 2.15    0.91    0.87

### Model 2 : Wind speed

Model data dimensions after removing NAs: 125 x 4
GAM fitted successfully
Model Summary:

Family: Negative Binomial(0.482)
Link function: log
```

```
Formula:
total_butterflies ~ s(abundance_t_minus_1, k = 4) + s(mean_wind_speed,
    k = 4)

Parametric coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)   3.9555     0.1296   30.52   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Approximate significance of smooth terms:
                         edf Ref.df Chi.sq p-value
s(abundance_t_minus_1) 2.229  2.617 42.202  <2e-16 ***
s(mean_wind_speed)     2.109  2.488  5.316   0.196
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

R-sq.(adj) =   0.25   Deviance explained = 24.6%
-REML = 608.98  Scale est. = 1         n = 125


Diagnostic plots:
```



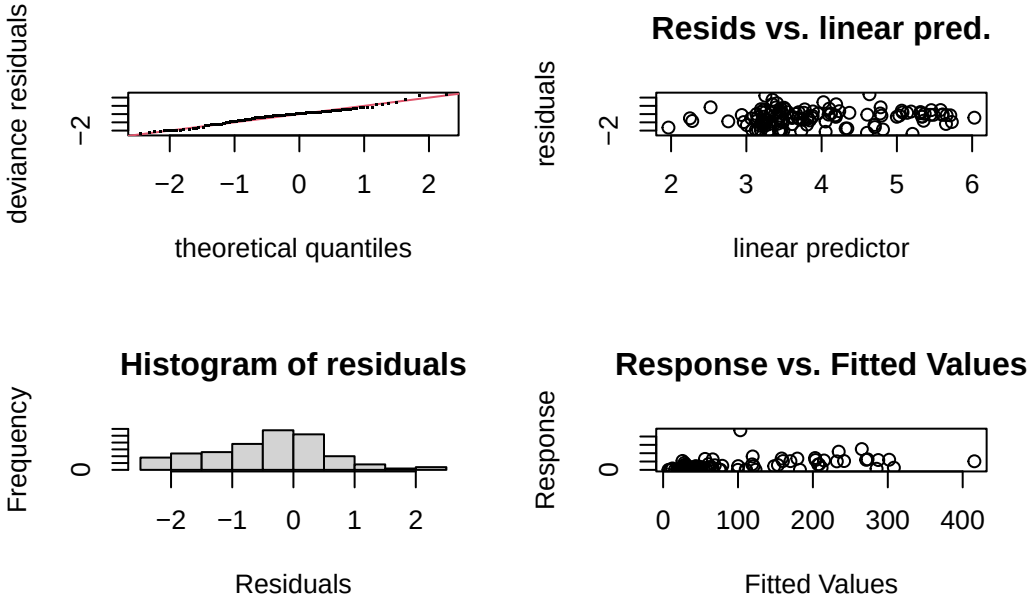
```
Method: REML   Optimizer: outer newton
full convergence after 4 iterations.
Gradient range [1.303456e-06,0.0001086969]
(score 608.9845 & scale 1).
Hessian positive definite, eigenvalue range [0.2204153,65.85178].
Model rank =  7 / 7

Basis dimension (k) checking results. Low p-value (k-index<1) may
indicate that k is too low, especially if edf is close to k'.

                         k'  edf k-index p-value
s(abundance_t_minus_1) 3.00 2.23    0.92    0.91
s(mean_wind_speed)     3.00 2.11    0.76    0.19

### Model 3 : Temperature

Model data dimensions after removing NAs: 125 x 4
GAM fitted successfully
Model Summary:

Family: Negative Binomial(0.472)
Link function: log

Formula:
total_butterflies ~ s(abundance_t_minus_1, k = 4) + s(mean_temp,
    k = 4)

Parametric coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)   3.9843     0.1309   30.43   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Approximate significance of smooth terms:
                         edf Ref.df Chi.sq p-value
s(abundance_t_minus_1) 2.110  2.509 36.333 5.4e-07 ***
s(mean_temp)           1.574  1.883  1.711   0.461
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

R-sq.(adj) =  0.276   Deviance explained = 22.8%
-REML = 609.65  Scale est. = 1         n = 125
```

```
Diagnostic plots:
```



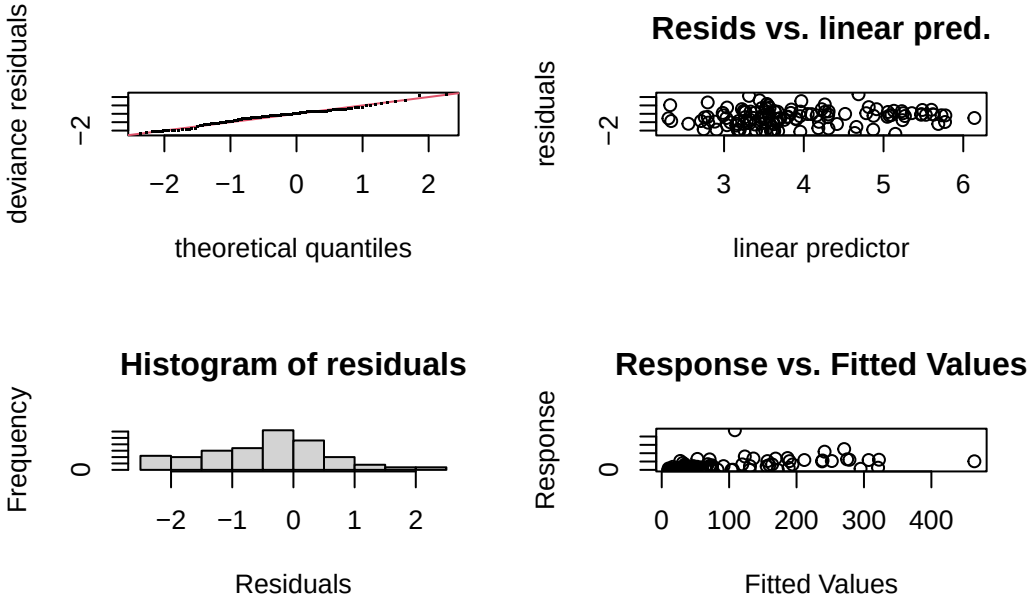

```
Method: REML   Optimizer: outer newton
full convergence after 6 iterations.
Gradient range [-3.845342e-10,4.72743e-09]
(score 609.6528 & scale 1).
Hessian positive definite, eigenvalue range [0.1457567,66.75074].
Model rank =  7 / 7

Basis dimension (k) checking results. Low p-value (k-index<1) may
indicate that k is too low, especially if edf is close to k'.

                         k'  edf k-index p-value
s(abundance_t_minus_1) 3.00 2.11    0.91    0.91
s(mean_temp)           3.00 1.57    0.80    0.36

### Model 4 : Wind + Temperature

Model data dimensions after removing NAs: 125 x 5
```

```
GAM fitted successfully
Model Summary:

Family: Negative Binomial(0.485)
Link function: log

Formula:
total_butterflies ~ s(abundance_t_minus_1, k = 4) + s(mean_wind_speed,
    k = 4) + s(mean_temp, k = 4)

Parametric coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)   3.9350     0.1293   30.43   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Approximate significance of smooth terms:
                         edf Ref.df Chi.sq p-value
s(abundance_t_minus_1) 2.148  2.543 40.596  <2e-16 ***
s(mean_wind_speed)     2.234  2.600  5.889   0.197
s(mean_temp)           1.663  1.985  1.723   0.425
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

R-sq.(adj) =  0.233   Deviance explained = 25.9%
-REML = 609.61  Scale est. = 1         n = 125


Diagnostic plots:
```



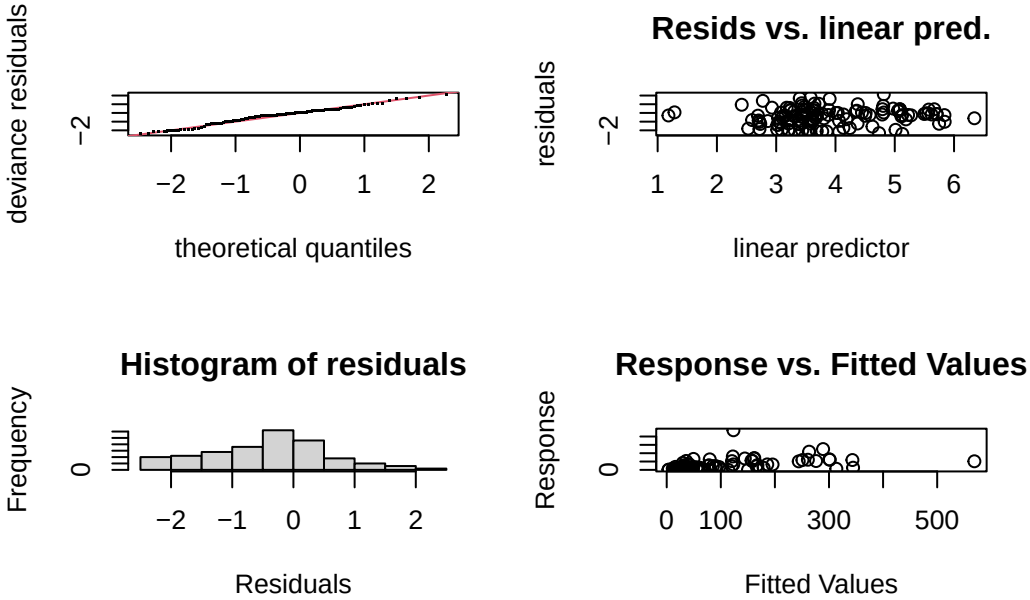

```
Method: REML   Optimizer: outer newton
full convergence after 6 iterations.
Gradient range [-7.722735e-07,1.471887e-06]
(score 609.6081 & scale 1).
Hessian positive definite, eigenvalue range [0.1717193,64.71343].
Model rank =  10 / 10

Basis dimension (k) checking results. Low p-value (k-index<1) may
indicate that k is too low, especially if edf is close to k'.

                         k'  edf k-index p-value
s(abundance_t_minus_1) 3.00 2.15    0.92    0.92
s(mean_wind_speed)     3.00 2.23    0.76    0.14
s(mean_temp)           3.00 1.66    0.82    0.40

### Model 5 : Wind + Temp + Sunlight

Model data dimensions after removing NAs: 125 x 6
GAM fitted successfully
Model Summary:

Family: Negative Binomial(0.487)
```

```
Link function: log

Formula:
total_butterflies ~ s(abundance_t_minus_1, k = 4) + s(mean_wind_speed,
    k = 4) + s(mean_temp, k = 4) + s(sunlight_proportion, k = 4)

Parametric coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)    3.928      0.129   30.45   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Approximate significance of smooth terms:
                         edf Ref.df Chi.sq  p-value
s(abundance_t_minus_1) 2.051  2.452 35.742 4.65e-07 ***
s(mean_wind_speed)     2.258  2.632  6.525   0.1519
s(mean_temp)           1.001  1.002  0.006   0.9439
s(sunlight_proportion) 1.000  1.001  3.941   0.0472 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

R-sq.(adj) =  0.225   Deviance explained = 26.4%
-REML = 609.25  Scale est. = 1         n = 125


Diagnostic plots:
```




```
Method: REML   Optimizer: outer newton
full convergence after 8 iterations.
Gradient range [-0.0002461131,0.0001473324]
(score 609.2549 & scale 1).
Hessian positive definite, eigenvalue range [9.885752e-05,64.70001].
Model rank =  13 / 13

Basis dimension (k) checking results. Low p-value (k-index<1) may
indicate that k is too low, especially if edf is close to k'.

                         k'  edf k-index p-value
s(abundance_t_minus_1) 3.00 2.05    0.91    0.78
s(mean_wind_speed)     3.00 2.26    0.74    0.13
s(mean_temp)           3.00 1.00    0.83    0.46
s(sunlight_proportion) 3.00 1.00    0.85    0.64

### Model 6 : Full model

Model data dimensions after removing NAs: 125 x 7
GAM fitted successfully
Model Summary:
```

```
Family: Negative Binomial(0.496)
Link function: log

Formula:
total_butterflies ~ s(abundance_t_minus_1, k = 4) + s(mean_wind_speed,
    k = 4) + s(mean_temp, k = 4) + s(sunlight_proportion, k = 4) +
    s(minutes_above_2mps, k = 4)

Parametric coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)    3.898      0.128   30.45   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Approximate significance of smooth terms:
                         edf Ref.df Chi.sq p-value
s(abundance_t_minus_1) 1.958  2.359 35.974  <2e-16 ***
s(mean_wind_speed)     2.088  2.436  7.097  0.0672 .
s(mean_temp)           1.000  1.001  0.014  0.9083
s(sunlight_proportion) 1.000  1.000  2.808  0.0938 .
s(minutes_above_2mps)  1.556  1.880  3.958  0.1258
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

R-sq.(adj) =  0.176   Deviance explained = 28.2%
-REML = 608.24  Scale est. = 1         n = 125


Diagnostic plots:
```




```
Method: REML   Optimizer: outer newton
full convergence after 10 iterations.
Gradient range [-0.0001569656,2.032617e-06]
(score 608.2351 & scale 1).
Hessian positive definite, eigenvalue range [2.438795e-05,63.45634].
Model rank =  16 / 16

Basis dimension (k) checking results. Low p-value (k-index<1) may
indicate that k is too low, especially if edf is close to k'.

                         k'  edf k-index p-value
s(abundance_t_minus_1) 3.00 1.96    0.91   0.830
s(mean_wind_speed)     3.00 2.09    0.74   0.065 .
s(mean_temp)           3.00 1.00    0.86   0.670
s(sunlight_proportion) 3.00 1.00    0.87   0.685
s(minutes_above_2mps)  3.00 1.56    0.95   0.935
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
# Model comparison table
comparison_full_day <- data.frame(
```

```
  Model = names(aic_values_full_day),
  AIC = round(aic_values_full_day, 2)
) %>%
  arrange(AIC)

cat("\n### Model Comparison (Full Day lag)\n")
```

```
### Model Comparison (Full Day lag)
```

```
kable(comparison_full_day)
```

| | Model | AIC |
|---|---|---|
| Wind speed | Wind speed | 1219.14 |
| Full model | Full model | 1219.16 |
| Baseline (abundance only) | Baseline (abundance only) | 1219.51 |
| Wind + Temp + Sunlight | Wind + Temp + Sunlight | 1219.72 |
| Wind + Temperature | Wind + Temperature | 1220.70 |
| Temperature | Temperature | 1221.24 |

---

## Final Summary

```
cat("Analysis Complete\n")
```

```
Analysis Complete
```

```
cat("=================\n\n")
```

```
=================
```

```r
# Combine all AIC comparisons
all_comparisons <- bind_rows(
  comparison_30m %>% mutate(Lag = "30 minutes"),
  comparison_60m %>% mutate(Lag = "1 hour"),
  comparison_120m %>% mutate(Lag = "2 hours"),
  comparison_240m %>% mutate(Lag = "4 hours"),
  comparison_full_day %>% mutate(Lag = "Full day")
) %>%
  select(Lag, Model, AIC) %>%
  pivot_wider(names_from = Lag, values_from = AIC)

cat("AIC Values Across All Time Lags:\n")
```

```
AIC Values Across All Time Lags:
```

```r
kable(all_comparisons, digits = 2)
```

| Model | 30 minutes | 1 hour | 2 hours | 4 hours | Full day |
|---|---|---|---|---|---|
| Full model | 48682.07 | 49417.42 | 50786.61 | 51965.27 | 1219.16 |
| Wind + Temp + Sunlight | 48683.39 | 49420.09 | 50796.94 | 51979.85 | 1219.72 |
| Wind + Temperature | 48840.75 | 49623.71 | 51075.30 | 52351.12 | 1220.70 |
| Temperature | 48891.39 | 49676.53 | 51111.31 | 52392.53 | 1221.24 |
| Wind speed | 49037.06 | 49793.65 | 51239.45 | 52595.21 | 1219.14 |
| Baseline (abundance only) | 49093.12 | 49853.76 | 51283.38 | 52641.53 | 1219.51 |

```r
# Find best model for each lag
best_models <- data.frame(
  Lag = c("30 minutes", "1 hour", "2 hours", "4 hours", "Full day"),
  Best_Model = c(
    comparison_30m$Model[1],
    comparison_60m$Model[1],
    comparison_120m$Model[1],
    comparison_240m$Model[1],
    comparison_full_day$Model[1]
  ),
  Best_AIC = c(
    comparison_30m$AIC[1],
    comparison_60m$AIC[1],
    comparison_120m$AIC[1],
```

```r
    comparison_240m$AIC[1],
    comparison_full_day$AIC[1]
  )
)

cat("\n\nBest Model for Each Time Lag:\n")
```

```
Best Model for Each Time Lag:
```

```r
kable(best_models, digits = 2)
```

| Lag | Best_Model | Best_AIC |
|---|---|---|
| 30 minutes | Full model | 48682.07 |
| 1 hour | Full model | 49417.42 |
| 2 hours | Full model | 50786.61 |
| 4 hours | Full model | 51965.27 |
| Full day | Wind speed | 1219.14 |