BIRD ABUNDANCE AT BIRD FEEDERS IN RESPONSE TO TEMPERATURE, WIND SPEED AND PRECIPITATION DURING THE WINTER SEASON

A Thesis

presented to

the Faculty of California Polytechnic State University,

San Luis Obispo

In Partial Fulfillment

of the Requirements for the Degree

Master of Science in Computer Science

by

Siddhant Kahal

July 2018

COMMITTEE MEMBERSHIP

TITLE:	Bird Abundance at Bird Feeders in Response to Temperature, Wind Speed and Precipitation During the Winter Season
AUTHOR:	Siddhant Kahal
DATE SUBMITTED:	July 2018
COMMITTEE CHAIR:	Theresa Migler-VonDollen Ph.D. Professor of Computer Science
COMMITTEE MEMBER:	Alexander Dekhtyar, Ph.D. Professor of Computer Science
COMMITTEE MEMBER:	Clinton Francis, Ph.D. Assistant Professor of Biological Sciences

ABSTRACT

Bird Abundance at Bird Feeders in Response to Temperature, Wind Speed and Precipitation During the Winter Season

Siddhant Kahal

The goal of this project is to explore how 23 different bird species respond to 3 climatic attributes. These attributes are lower than average temperatures, wind speed and precipitation level. Information about the bird species and all of the data associated with them is provided by Project FeederWatch (PFW). This is a citizen based survey study that provides key information about bird species abundance through the use of backyard and community feeders. The study volunteers from across the United States and Canada monitor these bird feeders and note important information about the species such as the number of individuals seen. Other standard information is also included such as location data and date. An original data collection pipeline was developed for this study to append climate data from Weather Underground (WU) to the PFW bird feeder data. The final dataset helped to explore how exactly the birds are reacting to winter temperatures, wind speeds and rain levels. Our results indicate that birds species in general visit the bird feeders more often as temperatures dip below average. We found that the body mass of the bird plays no role in the number of visits. Birds don't seem to be significantly affected by precipitation or wind speed as our results indicate no relationship between these climatic factors and abundance at the feeders.

ACKNOWLEDGMENTS

Thanks to:

- Dr. Theresa Migler-VonDollen, for the excellent advice and encouragement throughout this project.
- Dr. Alexander Dekhtyar, for introducing me to this study and guidance in developing solutions for a variety of problems.
- Dr. Clinton Francis, for the expert advice in biology and data modeling.
- My family and friends, for all of the love and support over the years.
- Andrew Guenther, for uploading this template

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

Chapter 1

INTRODUCTION

Correlating climate data and biological data has resulted in many interesting findings, especially when concerning the reaction to climatic conditions. This project aims to explore how 23 different bird species respond to the climatic attributes of temperature variations, wind and precipitation variations. The bird species data was provided by Project FeederWatch (PFW), a Cornell Ornithology Lab study in which citizens monitor bird feeders from across the United States and Canada. The climate data that was appended to the PFW data set was provided by Weather Underground (WU). Our hypothesis is that the abundance at the feeders is largely due to the thermoregulation requirements. These requirements are further determined by the body mass of the birds, with smaller birds retaining less heat, thus requiring more food from the feeders to offset the caloric deficit. With respect to wind speed and precipitation variations, we expect to see an overall decline in the number of visits to the feeders as wind speed and precipitation levels increase. This is because flight conditions in these cases are not ideal.

The results indicate that our overall hypothesis in regards to temperature is correct. During colder than average periods of winter, more individuals of the species were spotted at the feeders. This pattern was present throughout most of the 23 species studied in this project. However our hypothesis regarding the body mass is incorrect as there is no evidence of a relationship between bird feeder visits and the thermoregulation needs of the birds in regards to their body mass. Additionally, our hypothesis regarding wind and precipitation is also wrong as the results showed no evidence of the bird species being significantly affected by wind and rain levels. Overall, this project proved the usefulness of correlating climate data with bird feeder data from PFW and thus more research is required to explore other effects. Currently, the largest hurdle for the continuation of this study is the Weather Underground limits on climate data. Overcoming this will allow for more bird species to be studied in regards to the responses the climate conditions.

Chapter 2

BACKGROUND

2.1 Climate and Biological Responses

Climate change is increasingly becoming an urgent issue as changing weather patterns have far reaching impacts on biological systems [28]. Before discussing any further, it is worthwhile to first define the key terms. Climate, in terms of this study, is defined as the prevailing weather conditions of the area. The measurements used to assess the climate for the project include temperature, precipitation and wind speed. There are other features, such as pressure, but that is ignored. Details about this are covered in the later sections.

Studies correlating climate and biological events provide key information on how animals and plants are reacting to current changes [33, 45]. Additionally, many of these studies offer a glimpse into biological impacts we can expect in the future. The large scale impacts of climate change on animal species can be seen through animal group movements. Changing weather patterns have been listed as one of the top factors in contributing to the decline of large scale animal migrations [60].

Animal sensitivity towards changing climate can also be observed on a smaller scale. For example, Edmun D. Brodie et al. found that garter snakes crawled more slowly, for shorter distances and performed fewer reversals of direction in cooler temperatures [36]. This indicates that garter snakes that are present in regions with cooler climates, thus lower average temperatures, are generally slower than those in warmer climates. Additionally, Canadian red squirrels are breeding earlier in the spring to take advantage of the earlier spruce cone production [33]. These studies indicate that sudden large-scale environmental changes in temperatures may impact every day life of the animals involved.

Animals' response to climate change is also largely determined by the availability of food resources. For birds, which are the focus of this study, these food resources are highly dependent on the surrounding vegetation. The bird species may rely on the plants themselves for food, feeding on fleshy fruits or seeds [29]. The birds may also prey upon the insects hiding among the foliage [42].

One study by Robinson et al. suggests that the vegetation affects the arthropod foraging behaviors of birds that can successfully exploit the surrounding habitat [56]. Changes in behavior may include perch selection and large-scale habitat selection. However, with climate change it is being observed that plants are in fact migrating at various rates around the globe [53]. For example, Cheatgrass invaded western North America for over the past century. Furthermore, this invasion has happened at various increased rates throughout the years rather than at a steady pace [53]. This changing vegetation may pose a serious challenge to bird species that are unable to adapt to the rapidly changing habitats.

The aim of this project is to specifically study bird behavior around bird feeders in relation to the following daily climatic factors: mean temperature, maximum temperature, minimum temperature, wind speed and precipitation levels. Feeding birds through bird feeders is a well known practice, but there is little known about consequences of such supplementary food sources [44]. Large scale surveys, such as Project Feederwatch, provide important data on different bird species' activities at the feeders. This data has the potential to be used in a variety of studies.

2.2 Project FeederWatch

Project FeederWatch (PFW) was developed through the Cornell Lab of Ornithology, and for 31 years enlisted sharp eyed amateur bird watchers to document bird species around various feeders [50]. For this study, only observations from the years 2007 and through 2012 were used. The citizens will note down important features such as the bird species and number of individuals seen. All observations are made from November through early April [25]. The citizen-wide approach allows for hundreds of thousands of observations from across the United States and Canada. However there are drawbacks, the citizens themselves are not formally trained in making and recording observations, thus not all of the submitted data is valid.

To filter out the incorrect data, Project FeederWatch has employed a rigorous validation process that ensures the end data is as accurate as possible. The review procedure involves a range of filtering methods. There are simple approaches, such as cross referencing the observations against a checklist of "allowed" species for each US and Canadian province. Then there are the other more involved approaches.

One such approach is aimed to solve the complex problem of filtering out incorrect plausible data, such as misidentifying an species as another species that also exists in the region. The proposed solution is to provide educational quizzes/games to the bird watchers so that their bird watching skills can be quantified. This information then can be used to filter out species observations that may be too challenging for the observer to identify [30]. Though this has not yet been fully implemented [12], it is evidence of the high level of effort being put in to validate the data.

For the reasons above, we are confident that the end bird feeder data provided by the citizens is accurate enough to be used in this project. Though there is data from Canada, for the scope of this project only observations from the continental United States were considered. This bird feeder data set is of such high quality that it has been used in a number of studies involving a variety of bird species [13].

2.3 Related Works

The Project FeederWatch data has the distinct advantage of being large in scale, due to the number of participants in the study, while containing important bird species details. This allows for both broad survey style studies that include multiple bird species and also more species specific projects.

One large scale study explored the continental dominance hierarchy of various bird species across North America. Elliot T. Miller et al. used the network of citizen scientists from Project FeederWatch to discover that hierarchal standing of a species was largely predicted by the body-mass [51]. Another study, more similar to the focus of this project, explored the reshuffling of North American winter bird communities. This project only focused on the eastern North American area, but the team found that shifting winter climate has proved to be advantageous for smaller bird species, giving them a chance to colonize new regions [54].

On a more species specific scale, Barry K. Hartup et al. studied the risk factors associated with mycoplasmal conjunctivitis in eastern House Finches and made the interesting observation that the type of bird feeder may affect the risk levels of contracting the infection [40]. And finally there is the urgent study involving the declining Evening Grosbeak populations. This project points out that number of individuals seen at the FeederWatch sites have declined by fifty percent [31]. The authors are urging for more studies to be conducted in order to determine the driving factors of this decline. The data set provided by this project may aid in determining if temperature, precipitation and wind are contributing factors.

2.4 Weather Underground

Locating the proper climate data was one of the greater challenges of this study, mainly due to limited location information for each PFW observation and the large number of total observations. The climate data source needed to be able to provide accurate weather and temperature related measurements for each bird feeder observation given only the latitude, longitude and U.S. state for the location. Additionally, there are thousands of observations for each of the bird species being studied, making the total number of observations in the hundreds of thousands.

The large amount of data processing made it a requirement for the climate data source to have a robust API through which an automated script can mine the data. This requirement made the majority of options available on the Internet unfeasible as many of the websites had specialized the human-focused web application tools for accessing their climate databases. For example, NOAA's site has the option ordering data sets by manually selecting the location and date through drop-down menus [7]. This process would be far too time intensive for this project. The use of the API is key to efficient access and Weather Underground is one of the few online resources that has the robust API tools.

Weather Underground (WU), like Project FeederWatch, is a citizen based data collection project. Weather Underground's network of 250,000+ personal weather stations provides accurate data for many of the major cities in the United States [19]. The best feature is that the historical and current measurements are widely accessible through their API. This makes incorporating the data mining process into the scripts very simple. Additionally, the tool is well documented which further eases the use of the API methods.

Just as with Project FeederWatch, when private citizens are providing the data great care must be taken in validating that data. Weather Underground performs a number of checks to ensure the climate data is of the highest quality possible [22]. An example of one such check is the temperature neighbor check. In this check, data is flagged if it differs too significantly from the neighboring temperature measurements.

WU also employs a team of meteorologists and climatologists for the development of their proprietary weather forecasting model [20]. This further supports the validity of the historical and current climate measurements as they qualify for use in the forecasting applications. For the reasons highlighted above, our assumption is that the climate data collected through Weather Underground is already validated and ready for use in the study. The details of how the WU's API was used in the scripts is covered in the Methodology chapter.

2.5 Focus of this Project

There are 2 large components to this project. The first portion is the climate data collection, in which climate data is appended for each PFW data point. This climate data includes the following daily measurements: maximum temperature, minimum temperature, mean temperature, precipitation level and wind speed. These parameters were deemed important by our domain expert, Professor Francis, in assessing the bird responses. Any data point for which climate data could not be found was removed from the final data set. These weather attributes are provided by Weather Underground and further details about the data collection are covered in the Methodology sections.

The second portion of the project seeks to explore how the number of individuals seen at the bird feeders change in regards to climatic factors such as temperature, wind speed and precipitation. This will be done through analysis of the constructed dataset, which includes the original PFW data and the appended climate data. We also hypothesize that the average body mass of the bird species plays a role in the relationship between abundance and the winter temperatures.

During the winter, food resources become important assets, as birds need more calories to maintain the appropriate body temperatures [46]. Additionally, smaller birds have higher caloric needs for temperature regulation as they do not have the advantage of a larger body mass [47]. The medium to large bird species will not need to feed as often, as they are not loosing as much heat to the surrounding environment [47]. Thus the smaller bird species are expected to visit the bird feeders more often to make-up for the calorie loss through body heating.

The bird species that are going to be focused will vary in size from small songbirds, such as the Dark-eyed Junco which has a body mass of 18 grams, to the larger Blue Jay which has a body mass of up to 100 grams [2]. A small bird species are defined to be, for this study, any species having a mean body mass of 18–30 grams. Medium bird species are defined to have a body mass range of 40–60 grams. Large bird species are defined to have a body mass range of 70–180 grams. Additionally, the bird species of this project are known to be frequent visitors, with exception of a few species, allowing for a larger data set. The bird species selection criteria is covered in further detail in the Bird Species Selection section.

We hypothesize that the number of the smaller bird species seen at the bird feeder will be greater when the temperature is lower than normal, as the birds are feeding more often. Additionally, we would expect to see the effect of the cold be less on the number of the larger to mid-sized bird species seen at the bird feeders.

Lastly, wind and rain levels have been shown to impact bird health and flight behavior [38]. It has even been observed that wet feathers can lead to torpor and eventual death [48]. There are other factors at play as well, such as danger detection. Small foraging birds rely on visual cues to alert them of danger, however in a windy environment there is too much visual stimuli and thus the responsiveness of the bird may be reduced [35]. These studies indicate that it is also worthwhile to explore how bird feeding changes around feeders in accordance to the wind and rain..

For the above climatic attributes, our hypothesis is that as flight conditions become unfavorable during days of higher than average precipitation or wind speed, fewer individuals are spotted at the feeders. We expect that the birds will wait to visit the bird feeders when conditions are more favorable, as there is evidence that birds take shelter in trees and vegetation during stormy conditions [48].

Chapter 3

METHODOLOGY

The most important components of this study are the data sets being constructed for each of the bird species. Before talking about the appended climate data and the bird species themselves, it is worthwhile to first go over the given attributes of the Project FeederWatch (PFW) data set. These initial attributes largely determined the format and characteristics of the climate data.

3.1 Given Attributes of the Project FeederWatch Data

There are a total of 18 different attributes for each tuple in the original FeederWatch data. The complete list of these features is below, and of course not all of them were applicable to this project. These descriptions were obtained through the guidelines provided by PFW. The actual guideline document can be found in Appendix B.

- Latitude: The latitude value of the observation location in decimals. No information was given about how this measurement was done or how accurate is it to the actual location. This is relevant to the project as this is used to establish a city location for a given observation. This city location is later used to fetch the climate data.
- Longitude: The longitude value of the observation location in decimals. No information was given about how this measurement was done or how accurate is it to the actual location. Similar to Latitude, this was relevant to the project for establishing a city location.
- ID: This is the identification number of the participant in Project FeederWatch.

This is a unique Cornell Lab of Ornithology identification number and it is not available to the public. This value served no purpose in this study.

- StatProv: This is the U.S. state or Canadian province of the observation location. We are only focused on the continental United States and this state information is vital. It is used to establish the city for a given observation.

- Entry Technique: These are the various methods of the identifying the latitude and longitude values for a given observation. This is not relevant to the focus of this study and thus not considered.

- FW Year: This is the FeederWatch season. The seasons run from November to April, thus only covering the winter season. Example, 'PFW–1992' indicates the season running from November 1991 to April 1992. These values were ignored for this project as the **Year** attribute below provided the year information for each PFW observation.

- Year: This is the year of the first day of the two-day count observation. This was relevant to this project as this was used to query the dated climate data.

- Month: The month of the first day of the two-day count observation, used to fetch the appropriate climate data. See Year.

- Day: The day of the observation. This is for the first day of the two-day count. This value was used to fetch the correct climate data. See Month and Year.

- NHalfDays: This is the number of the half days of the observation during the two-day count period and is used as a measure of observer effort. The half days range from 1 to 4. This quantification was taken into account in the data models used for analysis. See the Data Analysis section.

- Effort Hrs Atleast: Another measure of the observer's effort. This is a measure of how many hours the participant invested making the observations. This ranged from less than 1 hour to greater than 8 hours. This quantification was taken into account in the data models used for analysis. See the Data Analysis section.

- BirdSpp: This is the name of the species seen at the feeder for a given observation. This is a crucial feature for this project. One of the main questions of this study deals with different sized bird species. The bird species attribute was used to filter out data only pertaining to the species of focus, as due to time constraints only a portion of the PFW data could be processed. These bird species were decided upon consultation with our field expert Professor Francis. More about this is covered in the Bird Species Selection section.

- NSeen: The number of individuals seen, this is the maximum number of the species in view at a single time over the two-day observation period. Another crucial attribute for the project. This data is relevant because it is a quantification of how many individuals of the bird species are using the bird feeders as a food source.

- Valid: This is a flag used to preserve data quality. In regards to this project, the data has already been filtered to only include valid and reviewed data. These data values are not relevant to this study.

- Reviewed: This is a flag used to preserve data quality. Again within the context of this project, the data is already reviewed. Thus, these values are not used. See Valid above.

- Loc ID: This is the unique identifier for the location of the observation. Note that participants, see ID above, can have multiple count locations or Loc ID's.

This data is not relevant.

- Sub ID: This is the submission identifier and it uniquely identifies the entire checklist submitted by a participant from the count period. Note, if the species observations are all on the same checklist, then they have the same Sub ID. This is not relevant for the project.

- Obs ID: This is the observation identifier and it uniquely identifies a single observation. All species reported on a single checklist receive different observation identifiers. This data is not relevant for this study.

To summarize, there are many attributes to the PFW data set, but only a handful are useful for this project. The latitude, longitude and states data are all key to this project as they are used to establish a U.S. city location for which climate data can be collected. Additionally, the year, month and day values are important as they ensure the climate data is correctly dated. As a reminder this a winter season study, and the season ranges from November to April. Nseen is the main quantification of the abundance of a bird species at a given feeder location and is used to quantify how many individuals are using the feeders. Note again, the count period for the number seen is 2 days. Lastly, there is bird species, and this is one of the most important attributes. It is covered in detail in the Bird Species Selection section.

The hours of effort, location id and number of half days are all factors taken into account for the data analysis and is covered in the later sections. These attributes are only relevant for results and conclusions portions of the project. The rest of the attributes included with the Project FeederWatch data are simply ignored as they are not relevant to the questions being tackled.

3.2 Bird Species Selection

When selecting bird species for this study, 3 main factors became the driving force. First, the bird species must be non-migratory during the winter season. This filtering criteria ensures that the bird species for this study remain in their respective regions and actually experience the colder climate of winter. The only exception to this requirement is the Evening Grosbeak. However, the species was chosen for this study because of the previous work done on it in regards to the population decline at the feeders.

Second, the bird species must be well represented in the PFW data set. Our criteria being that the bird species must at least have a few hundred observations. This will ensure that there is enough data to develop models for after the filtering and clean up steps. The exact details of these data processing steps are covered in the next few sections.

The third factor in the bird selection process was the body mass of the species. As a reminder the focus of the project revolves around how different massed birds react to cooler temperatures, with the main comparison being between smaller species and mid to large species seen at the feeders. Thus, the aim of the final selection was to have about the same number of smaller massed bird species as the mid to larger sized bird species. Our biology expert, Professor Francis, offered valuable consultation in selecting the species and also in ensuring the requirements above were met.

The final set of bird species that was selected is presented in Table 3.1. This table contains all of the bird species that are used for this study. The first criteria of the bird species being well-represented in the original PFW dataset is met as many of the species have tuples numbering in the thousands, with the only exception being Pine Grosbeak.

However, another issue arises with some species having too many tuples, as there are daily limitations on how much climate data can be collected through WU's API. For example, the number of observations for Mourning Dove is 153,417 and the time required for appending climate data to all of the observations would be in terms of weeks not hours given the limitations. This makes processing all of the observations for the 23 species impractical. For this reason only a maximum of 5,000 observations per each bird species were used for the climate data collection. These 5,000 observations were chosen at random. This allows us to obtain climate data for all of the 23 bird species, while still having enough observations for further data processing and modeling.

The second goal of including roughly equal numbers of small, mid and large sized bird species is also met as clearly shown by the average body mass measurements in the table. The mean body mass values, in grams, are presented in the column labeled Body Mass of Table 3.1 and these averages were calculated using values from the Cornell Ornithology Lab's website All About Birds and the bird mass handbook by John B. Dunning Jr. [2, 37]. The number of observations for each species in the original PFW dataset is presented in the column labeled PFW Observations.

The number of the observations per species that were used in the final data models and analysis are presented in the column Filtered Obs. The filtration and data processing methods used to obtain the final tuples in Filtered Obs. are covered in the later sections.

Table 3.1: Bird Species and the number of observations used in this study.

Bird Species	Body Mass (grams)	PFW Observations	Filtered Obs.
American Goldenfinch	15.5	133496	4054
Black-billed Magpie	177.5	3704	3287
Black-capped Chickadee	11.5	102562	2648

Table 3.1: Bird Species and the number of observations used in this study.

Bird Species	Body Mass (grams)	PFW Observations	Filtered Obs.
Blue Jay	85	106666	1264
Brown Creeper	7.5	7457	4770
Chestnut-backed Chickadee	9.5	11992	4643
Chipping Sparrow	13.5	13004	3158
Common Grackle	108	25920	2720
Common Redpoll	15.5	4545	3615
Dark-eyed Junco	24	159161	5000
Downy Woodpecker	24.5	137064	4699
European Starling	78	57676	4408
Evening Grosbeak	63.5	1595	1206
Hairy Woodpecker	67.5	50665	4691
Mountain Chickadee	11	5140	2995
Mourning Dove	121	153417	2876
Northern Mockingbird	51.5	26507	4792
Northern Cardinal	45	147259	2118
Pine Grosbeak	56.4	250	231
Pine Siskin	15	27250	2738
Red-bellied Woodpecker	73.5	91767	4685
Tufted Titmouse	22	113012	4811
White-throated Sparrow	27	61422	4517
All Species	NA	1441531	79926

3.3 Given Attributes of Weather Underground Data

There are over a hundred climate attributes provided by Weather Underground, ranging from the standard average temperature values to the very specific measurements of cooling days since the first day of the year [21]. There are far too many attributes to list and discuss individually, instead listed below are the climate measurements that were deemed important and relevant according to our domain expert Professor Francis. From this initial set of attributes the final 6 were eventually chosen for the actual study.

An important note, for the historical climate data, Weather Underground has 2 categories for each of the climate features, observations and daily summaries [21]. Observations are all of the direct climate measurements for the day. For example, the hourly temperature values, for which there will 24 measurements. While the daily summary is, as the name implies, a single summary value for the day. For example, the mean temperature calculation for the day or the average wind speed for the day, both are single values for the day. For the scope of this study, only daily summary values are useful as it is a single average measurement for the day, which can then be appended to the PFW observation with same day. The observations are ignored as there may be multiple observation values for one day. The listed climate features below are all from the daily summary category.

- date: This is the date of the observations. This is very relevant to the study as this ensures that the correctly dated climate values are collected for the PFW data.
- percipi: Precipitation in inches of rain. This feature is also relevant to this project. An important note, WU provides the precipitation measurements both in inches and millimeters, but only the inches value was collected. This is

because the data can be later converted if necessary.

- snowfalli: Snowfall in inches of snow. Though this is an important climatic factor, it is not going to be focused on in this project.

- humidity: Humidity represented as a percentage. Again, this may be important for other studies, but for this study these values are ignored.

- maxtempi: Maximum temperature value for the day in Fahrenheit. This is relevant to this project as one of the main questions is to analyze bird species abundance around bird feeders in regards to temperature. Note, Weather Underground does provide the temperature measurements in Celsius, but only the Fahrenheit values were collected. The idea being that the values can be easily converted in the future if required.

- mintempi: Minimum temperature value for the day in Fahrenheit. This is relevant to this study, see maxtempi.

- meantempi: Mean temperature value for the day in Fahrenheit. This is relevant to this project, see maxtempi.

- meanwspdi: Mean windspeed for the day in miles per hour. This is relevant to this study as one of the main questions is to determine whether wind speed has an affect on bird abundance at the bird feeders. Note, Weather Underground does provide the speed in kilometers per hour, but only the miles per hour values were collected. The values can be converted and appended later if required.

- minwspdi: Minimum wind speed for the day in miles per hour. This is not relevant to this study as the mean wind speed for the day is more useful for determining the wind conditions throughout the day, see meanwspdi.

- maxwspdi: Maximum wind speed for the day in miles per hour. This is not relevant to this study as the mean wind speed for the day is more useful for determining the wind conditions throughout the day, see meanspdi.

To summarize, of the subset of the climate features presented above the final group of relevant attributes are as follows: date, precipi, mintempi, maxtempi, meantempi, and meanwspdi. Eventually only the mean temperature was deemed relevant for the actual study in terms of the temperature measurements. The minimum and max temperature values were ignored for the analysis. The details of this will be covered in the Preprocessing for Data Models section. The date is vital to ensure the climate measurements are from the same date as the PFW observation. The precipi values were also deemed relevant for this project. Finally, only the mean wind speed values (meanwspdi) were used in the study, while the max and min wind speed values were ignored. Further information about the reasoning behind these selections is in the next section.

3.4 Script Implementation for City Attribute

Before the Weather Underground's API can be utilized, the PFW data needs additional location details besides just latitude, longitude and state values of the bird feeders. The majority of the population of United States lives within major city areas [58], as a result much of the data from PFW and WU are from these densely populated urban areas. The challenge now is to filter out all of the bird feeders that are located too far away from a major city.

The pseudo code for the following steps is presented in Algorithm 1. The first step is to load the original PFW data into a data frame which can then be logically manipulated. Next, there is the loop that iterates through each row of the data frame. For each row, or observation, the latitude, longitude and state values become the parameters for the method getNearestCity(). This method is then called and the return value is stored in the variable city.

The getNearestCity() method is key to this procedure as this determines whether the bird feeder location is close enough to a city. This function contains the top 1,000 U.S. cities and their latitude, longitude and state values in a key value pair data structure. With the given bird feeder location parameters, the distances between the feeder and the nearest major cities of the state are calculated.

The city with the minimum distance is then returned, as this will provide the nearest location for a Weather Underground weather station. If the bird feeder is more than 40 kilometers away from the city, then that city is not considered and ignored. If no suitable major city is found, then getNearestCity() will return NULL and that row is deleted from the data frame. Figure 3.1 is a generalized visual representation of the city selection process.

The last step simply involves converting the data frame to a CSV file and exporting it with the correct file name. At this stage, all of the PFW observations now have a city attribute, for which climate data can acquired through Weather Underground. The actual implementation, in the Python programming language, of Algorithm 1 is presented in Appendix C.

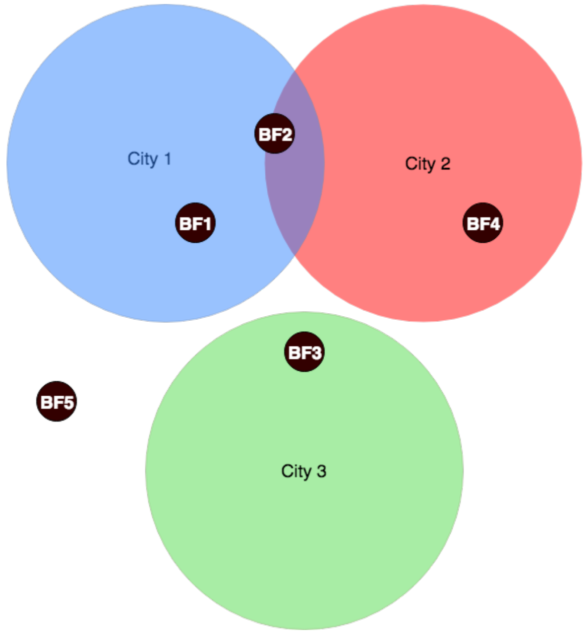

Figure 3.1: This figure illustrates how Algorithm 1 selects a city for a given bird feeder of PFW. The the black circles labeled with BF represents the location of a bird feeder (BF). The colored circles represent a distance radius of 40 kilometers from the city center. For example, the green circle represents distances that are within 40 kilometers from the center of City 3. Given the situation above, BF1 and BF2 would be assigned City 1. Note, even though BF2 is in the range for City 2, it is closer to City 1 and it is picked by the algorithm. BF3 would be assigned City 3 and BF4 would be assigned City 2. Lastly, BF5 is not within 40 kilometers from any major city so all observations with BF5 will be deleted from the final data set as the city is needed for acquiring the climate data. This will be covered in detail in the next section.

Algorithm 1: Steps for appending city data to PFW where possible.

```
Input  : Original Project FeederWatch Data
Output: Project FeederWatch Data with cities appended.
dataFrame ← "PFWData.csv"
dataFrame.addColumns("city")
foreach Row r in dataFrame do
    /* getNearestCity method returns the city which is at most
       40km away from the lat/long coordinates.  Returns NULL if
       no city in the state is found.                           */
    city ← getNearestCity(r.latitude, r.longitude, r.state)
    if city not NULL then
        r.append("city", city)
    else
        dataFrame.deleteRow(r)
return dataFrame.toCSV("PWFDataCities.csv")
```

3.5 Script Implementation for Climate Data Collection

Before discussing the climate data collection script, it is worthwhile to point out that the bird count of PFW is performed over the course of 2 days. However, the original PFW data provides only the date for the first count day. As such, climate data for that day is collected and later used for analysis. The assumption being that the climate data applies to both day 1 and day 2 of the count days.

The climate data collection portion is relatively straightforward now that there is the city feature along with year, month and day. Algorithm 2 illustrates the steps in a more formal pseudo code fashion. The first step, as in Algorithm 1, is to load the

CSV file into a data frame. Afterwards, empty columns are added for the climate attributes and they are named appropriately. Once again, there is a loop that iterates through each tuple, or row, of the data frame.

For each row, measurements of mean temperature, minimum temperature, maximum temperature, mean wind speed and precipitation level are collected through WU's API link. The API link, once loaded, returns a JSON object that contains the climate attributes described above. In order to construct this link, the current row's city and date is required, as shown with the constructApiLink() function in Algorithm 2.

The constructApiLink() method is quite simple as it only takes the values provided in the parameters and constructs a correctly formatted string, which is the API link. This API link, specific to the provided city and date, is then returned. The link is then loaded through the method loadURL() and the returned JSON object is stored in the variable climateJSON.

The final steps of the loop are to check whether the desired climate attribute exists in climateJSON. If this is true then that value is appended to the appropriate column of the current row. Lines 8 and 9 of Algorithm 2 illustrate this by first checking whether maxTemp exists in the JSON, and if the condition is true, the maxTemp value is appended.

The last step of the algorithm is to convert the data frame to the CSV format and output the file. As mentioned in the Bird Species Selection section there are limitations to how many WU API calls the script can make in one day. For this reason the script processed data in subsets of 500 observations. The actual implementation for processing a subset of 500, using the Python programming language, is presented in Appendix D.

Algorithm 2: Steps for appending Weather Underground climate data.

Input : Original Project FeederWatch Data with cities.

Output: Project FeederWatch Data with climate data appended.

```
dataFrame ← "PFWDataCities.csv"
dataFrame.addColumns("meanTemp", "maxTemp", "minTemp",
  "windSpeed", "precip")
foreach Row r in dataFrame do
    apiLink ← constructApiLink(r.date, r.city)
    climateJSON ← loadURL(apiLink)
    if meanTemp in climateJSON then
        r.append("meanTemp", meanTemp)
    if maxTemp in climateJSON then
        r.append("maxTemp", maxTemp)
    if minTemp in climateJSON then
        r.append("minTemp", minTemp)
    if windSpeed in climateJSON then
        r.append("windSpeed", windSpeed)
    if precip in climateJSON then
        r.append("precip", precip)
return dataFrame.toCSV("PWFDataClimate.csv")
```

3.6 Preprocessing for Data Models

To allow for the best data models and insights as possible, additional processing and filtering steps were performed on the dataset. The first step was to filter out bird feeder locations that may not experience low enough winter temperatures to warrant additional feeder visits. To achieve this all bird feeder locations with latitude values below thirty-eight degrees were removed, as these locations may be too mild in terms of winter climate. Figure 3.2 provides a visual representation of the locations of the bird feeders after the latitude filtration step.

Next, two additional attributes were added to aid in the next data processing steps. The first added attribute was the average temperature for a given winter season. For example, the winter of 2012 may have had an average season temperature of 35 degrees Fahrenheit. However, during the actual season, daily temperatures may have varied from 30 degrees to 45 degrees.

The second attribute appended was the average precipitation levels for the winter season of the year. This is very similar to the average season temperature, but this value represents the mean precipitation levels for the PFW season. For example, the winter of 2007 may have had an average of 4 inches of rain, while actual daily measurements may have varied greatly. There were no seasonal averages for wind speed at the time of the study, so the actual daily average wind speed values were used for the data models. The wind speed values were scaled from 0 to 1 to allow for proper comparisons in the models. All of the seasonal averages were obtained through the "WorldClim" database [24]. WorldClim provides climate data for minimum, mean, and maximum temperature and precipitation for the years 1970–2000 [23]. The spatial resolution of the climate data ranges from 1 square kilometer to 340 square kilometers [23]. As with the Weather Underground data, this climate data is already validated and ready for use in this project.

Finally, with the added two attributes it was possible to calculate and append the anomalies in regards to the mean daily temperatures and precipitation levels. It is worth pointing out again that only the mean temperature values are used, while the minimum and max temperature values for the observation are ignored. This is because, we are interested in reactions from species when temperatures were especially low, thus requiring more feeder visits. Deviations of this mean temperature value from the winter season's average temperature gives a better indication of how cold it was on the day of the observation.

The anomalies for mean temperature were calculated through the following steps. The average season temperature was subtracted from the collected mean temperature for the observation. The resulting value was either negative, positive or zero. A negative anomaly value represents a lower than average winter temperature while a positive value represents a higher than average temperature. And a value of 0 indicates an average winter temperature. These steps were performed for every row and the resulting value was appended as the temperature anomaly for that row.

The anomalies for precipitation value were calculated in the same manor as the temperature. The anomalies allows for more accurate comparisons between the number of individuals observed at the feeders and the deviations from the average winter temperature and precipitation levels. This is because the members of a bird species that live in colder or rainier regions are more accustomed to those conditions than individuals of the same species in warmer and drier regions. So by comparing the abundance to the deviations from the seasonal averages, we get a better sense of how the bird species are reacting to the changes in the regional climates.

3.7 Linear-mixed Effect Models

Once the data sets were constructed, the next challenge was to pick a valid and useful model. Here Professor Francis was of great help as he had the most expertise in which data model was applicable. The linear-mixed effect model was chosen as it allows us to observe relationships among multiple variables and grouping factors in the data

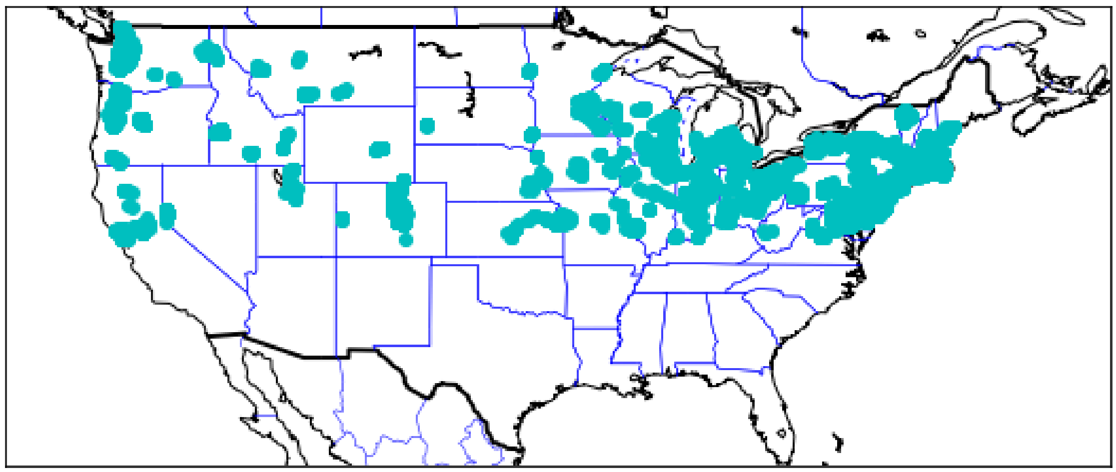

Figure 3.2: The cyan colored points on the United States map represents all of the actual bird feeders used in this study. As discussed earlier, all the feeder locations are above the 38 degree latitude boundary. This avoids all observations from locations that may have mild winters. For example, birds in southern California and Texas are going to experience warmer winters than the birds in New York or Michigan simply due to geographic location. Since one of the goals of this project is to explore the reactions of bird species to cold temperatures, it is advantageous to only include geographic areas which experience relatively cold winters.

set, more on this later. The linear-mixed effect model is a few degrees higher in complexity than the simple linear model, so it is worthwhile to cover the basics of these models before continuing any further.

Lets start at the beginning with the basic linear model. A hypothetical relationship of interest can be represented as this formula:

$$y \sim x$$

This formula reads as “y as a function of x” or it can also be referred to as “y predicted by x”. For this paper, the two terms are considered equivalent and is used interchangeably. The y in this case is considered a dependent variable. The variables to the right of the $\sim$ symbol are referred to as the predictor or independent variables.

In the formula above the only predictor variable is the x, since it is a simple linear relationship. The predictor variables can also be referred to as fixed effects, but this is not relevant yet.

In real world studies, no relationship is just defined by 1 effect as with y being completely determined by x. This is far too deterministic. Other effects at play must be accounted for in the model. There may be many "random factors" at play. To account for this lets add another term to the formula:

$$y \sim x + \epsilon$$

This new additive factor, ϵ, it accounts for all of the random error. This term stands for everything that effects y that is not x. In other words, from our perspective the ϵ accounts for all the effects that are uncontrollable by us. This formula is the schematic depiction of the model that would be built using a statistical tool. Now that the simple formula is defined lets take steps towards building the linear-mixed model formula.

First lets add another fixed effect, say "a", resulting in this formula:

$$y \sim x + a + \epsilon$$

It is important to note that "a" is added with plus sign, indicating an additive relationship. An additive effect is such that the affect of a on y is not dependent on the affect of x on y. The variables are independent from each other in this sense. This brings up one of the important assumptions of the linear model. The linear model assumes that all of the predictor variables are independent from each other. Logically, this makes sense as if there are 2 or more similarly effecting predictor variables, then it become difficult to identify which predictor is playing the larger role [8].

Now it is time to introduce the random effects, and in order to illustrate this concept lets make the example more concrete and relevant. Though the formula below appears similar to the actual study, it is completely hypothetical and only for explanatory purposes. Lets say that we wanted to explore the relationship between the number of birds seen at the feeders, mean temperature and precipitation. Also assume that there are different bird species in the data set, this is important for the random effects. The formula looks like this now:

$$Nseen \sim temp + precip + \epsilon$$

Recall that one of the major assumptions for using the linear model is the independence of the predictors, but multiple responses from the same bird species violates this assumption. This is because individuals from the same species are going to behave similarly. The similarity within species is going to be an idiosyncratic factor that affects all responses from the same species. If this is unaccounted for then the responses will be rendered inter-dependent and not independent.

These species specific differences are accounted for in the linear model with the assumption of random intercepts for each species. Essentially this means that each bird species is assigned a different intercept value when the groupings are accounted for, but the line estimates for the groups will have the same slope value as the overall line model for the fixed effects [14]. These intercept estimations are calculated for each species groupings. These intercepts are the points where the line estimates for each of the species crosses the y-axis. In other words, when temp and precip have the values of 0 for their linear models.

To account for the grouping of species, a random effect is added to the formula:

$$Nseen \sim temp + precip + (1|species) + \epsilon$$

The bird species term appears to be complicated, but that is only due to the syntax. This syntax mirrors the R programming language, which was used to construct the actual models for this study. This is covered in great detail in the later sections. The "species" term translates to "assume a different intercept for each bird species". The "1" stands for the intercept [8]. The formula above represents a linear-mixed effect model

Now perhaps it becomes more clear why this model is named "mixed". In the earlier models only the fixed effects, or predictor variables, were taken into account. And then there was a generic error term that added. Now essentially there is more structure to that epsilon error term through the addition of a random effect. However, the ϵ is still present in the formula above because there are differences present within the species groups. The epsilon errors were calculated for all of our actual models, but the values were not used in the analysis. Instead only the defined random effects were used in the final analysis.

3.8 R Programming Language and Linear Modeling

For this project all of the data modeling was done through the use of the R programming language [55]. This is because R has a robust set of packages that provides powerful statistical analysis. Additionally, our biology expert, Professor Francis, is well versed in the use of R for biological studies.

The formulas presented in the above sections are just schematic descriptions. The lme4 package of R allows us to actually build the linear models and analyze the relationships [27]. The relationships are analyzed through the model summary, which is presented through a table. It is best to illustrate this with the continued example of the hypothetical bird species study formula:

$$Nseen \sim temp + precip + (1|species) + \epsilon$$

The R code for constructing and summarizing the model is presented below. Notice the lack of the ϵ, or the random error variable. The lmer method automatically accounts for the random error, without the explicit declaration in the parameters. However, the random error values are not of particular importance in this study and are ignored for the remainder of the analysis.

```
model <- lmer(NSeen ~ temp + precip + (1|species))
summary(model)
```

The output of the "summary" method is presented in Tables 3.2 and 3.3. As a reminder no actual model was built as there is no data set, the tables are only for explanatory purposes. Lets first focus on Table 3.2. This table shows all of the fixed effects and their relationships to the abundance of the bird species. The "Intercept" row covers the y intercept of the estimated line, the standard error and the t value. The intercept is estimated through the line models when all of the predictor variables are 0. Recall, the predictor variables are "temp" and "precip". The intercept estimate is not applicable to this study. The "Estimate" column presents the slope estimates for each of the fixed effects. For example, there is a positive slope estimate of 2.02 associated with temp, indicating that as the temp value increased so did the Nseen value. "SE" is the column representing the standard error and finally there is the "t value" column. The standard error is for the slope estimate value and is used to calculate the t value. The t value is very crucial for this study and is covered in great detail in the Significance and t value section.

The other table that is included with model summary is the Random Effects, represented by Table 3.3. It is important to note the variance strengths, shown in the "Variance" column, among the different random effects. In this case we only have one random effect, but nonetheless it has caused some variance among the data. This indicates that it was useful to account for this in the data model. The other column is the "Std. Dev.". This is measure of how much variability in the dependent measure there is due to bird species [8]. For this study, a simple check was done with every constructed model to make sure the Random Effects table showed variances for the variables chosen. This is covered in more detail in the Data Models for this Study section.

Table 3.2: Model Summary for the hypothetical situation of modeling bird abundance, temperature and precipitation.

Fixed Effects	Estimate	SE	t value
Intercept	7.09	4.8	1.477
temp	2.02	1.6	1.263
precip	-4.32	2.47	-1.749

Table 3.3: Random effects for the hypothetical situation of modeling bird abundance, temperature and precipitation.

Groups	Name	Variance	Std.Dev.
species	Intercept	2.24019	1.49009

3.9 Significance and t value

The t value is very important for the analysis portion of this study. This value is calculated by dividing the "Estimate" by the "SE". The resulting value is a measure of the fixed effect strength as well as the significance of the relationship presented by the model. The t value is a more reliable measure of the effect strength than just the slope magnitude because it takes into account the standard error. The t value is a major component in comparing how different climatic factors affect the bird abundance at the feeders and at what strengths.

The second importance of the t value comes in the form of significance. We used t values as a proxy for assessing the strength of the effect of predictor variables because the lme4 package does not calculate p-values. In essence, when the magnitude of t is greater than or equal to 2, the standard error of the effect size will be less than 1/2 the absolute value of the effect. In other words, this means the 95 percent confidence interval would never overlap zero with t greater than 2.

The confidence intervals were calculated for each of the models constructed using the R method confint. This method takes the constructed model as a parameter and returns the lower and upper bounds of the confidence interval. The effects were considered significant, if this range did not overlap zero. As explained in the previous paragraph, this non-overlap of zero is obtained if the absolute t value is greater than or equal to 2. As such, from this point on the significance of the effect is only going to be quantified by the t value.

3.10 Assumptions

As with any other model, the linear-mixed model comes with assumptions that have to be satisfied in order for the linear model to be meaningful [9]. To check that the assumptions are satisfied, a simple test can be performed with the residual values. A residual value is the difference between the observed dependent variable and the predicted value from the line estimate [15]. If the histogram of the residuals follows an approximately normal curve then one can assume the assumptions are met as the distance of the data points from the line estimate follow a roughly normal curve when plotted [3, 10].

The residual histograms were inspected individually by eye, as quantitative methods in R were not advanced enough to detect approximate normal distributions. Too many of the rough normal distributions were rejected by these methods, thus it was decided to check every histogram by eye. In addition to the histograms, Q-Q plots for the residual values were also created and inspected. A Q-Q plot is a scatter-plot created by plotting 2 sets of quantiles, or percentiles, against one another [18]. If both sets of quantities came from the same distribution, then we should see the points forming a line [18]. For this study the residual values need to follow an approximately normal distribution, as such the residual quantiles are plotted against the normal distribution quantiles. If the residuals follow an approximately normal distribution then the points should roughly follow a line. These specific Q-Q plots are referred to as Q-Q normal plots. One histogram and one Q-Q normal plot were constructed and inspected for each of the data models built. This process is covered in detail in the next section.

3.11 Data Models for this Study

The R code for modeling the actual constructed data set is below. Notice that the syntax is very similar to the formulas discussed before. Besides the 3 flags at the end of the parameters, the formula schema is almost identical in format to the previous schematics.

```r
model <- lmer(log(NSeen) ~ temp_anomalies + scale(wspeed_kmph)
                 + precip_anomalies + (1|EFFORT_HRS_ATLEAST) + (1|LOC_ID)
                 + (1|NHalfDays) + (1|city), REML=F, data=dat1,
                 na.action="na.fail")
```

The linear mixed-effect model is constructed using the lmer function, with the returned data model stored in the variable model. One model was built for each of the 23 bird species, using the same R code and methods. The results of these species specific linear models are discussed in the Results section. The key components of the lmer function are the parameters, as they determine which variables to build the linear models with.

In addition to the 23 models for each of the species, one other general model was built to encompass all of the bird species. The only difference in construction of this model is the inclusion an additional random effect (1|BirdSpp), or bird species. This allows the data points to be grouped by bird species. The code for constructing the generalized linear-mixed effect model is below.

```
model <- lmer(log(NSeen) ~ temp_anomalies + scale(wspeed_kmph)
                  + precip_anomalies + (1|EFFORT_HRS_ATLEAST) + (1|LOC_ID)
    + (1|NHalfDays) + (1|city) + (1|BirdSpp), REML=F, data=dat2, na.
    action="na.fail")
```

The first variable used in the method is the log(NSeen) and this is simply the the natural log of number of individuals seen of that bird species, as discussed in the Given Attributes of the Project FeederWatch Data section. This is the dependent variable in the linear models and this followed by the $\sim$ symbol [6]. That symbol indicates the term “function of” and is followed by the independent variables to the right. These independent variables are also known as predictor variables.

One may wonder why the natural log of the number of individuals at the feeders is used in the models. In order to illustrate why the natural log of the Nseen value was used instead of just the Nseen value, it is best to walk through a concrete example. Let’s focus on the data set for the species White-throated Sparrow. The histogram of the values of Nseen for this species is presented in Figure 3.3. It is very clear that there is a left skew to the distribution. This results in a non-normal distribution of the residuals, as presented in Figure 3.4. This histogram does not exhibit an approximate normal distribution, in regards to this study. This violates the major assumption of using the linear mixed-effect model, as this model assumes an approximately normal distribution of the residuals.

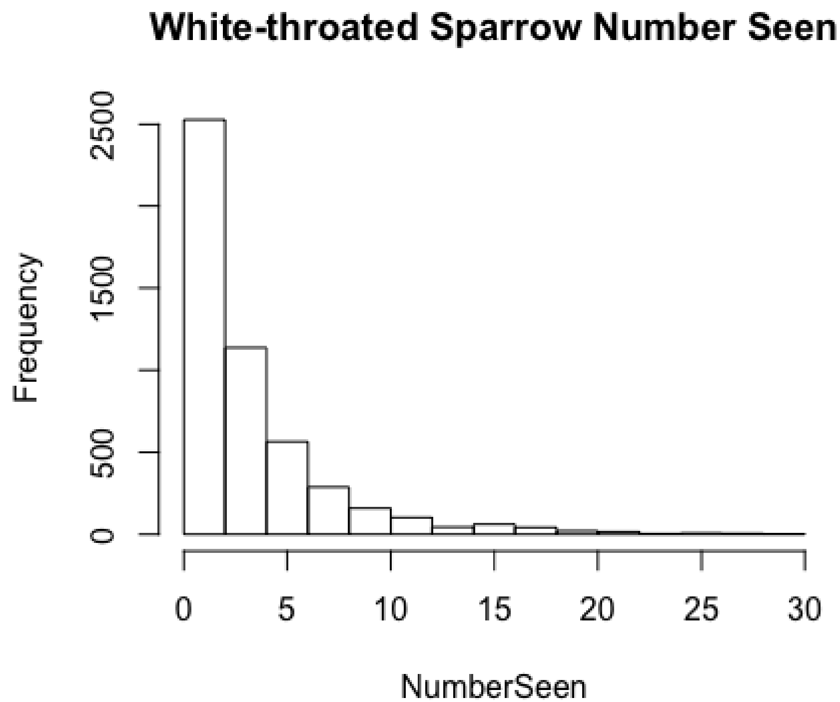

Figure 3.3: The histogram of the Nseen values of the White-throated Sparrow.

The histogram of the natural log of the Nseen is presented in Figure 3.5. There is still a left skew present, however there is a better distribution of values. The histogram of the residuals from the linear mixed-effect model with the natural log of Nseen is presented in Figure 3.6. This time the residual values follow an approximately normal distribution, thus allowing us to draw insights from the linear mixed-effect model.

The predictor variables for the models of this study are temp_anomalies, precepi_anomalies and scale(wspeed_kmph). These variables can also be referred to as fixed effects as they are the main focus of the analysis. Note the scale of the wind speed value is used for the models. This is because the standardized values of the wind speed allow for better comparisons against the other independent variables. Recall, the units of the wind speed values are miles per hour and the other variables,

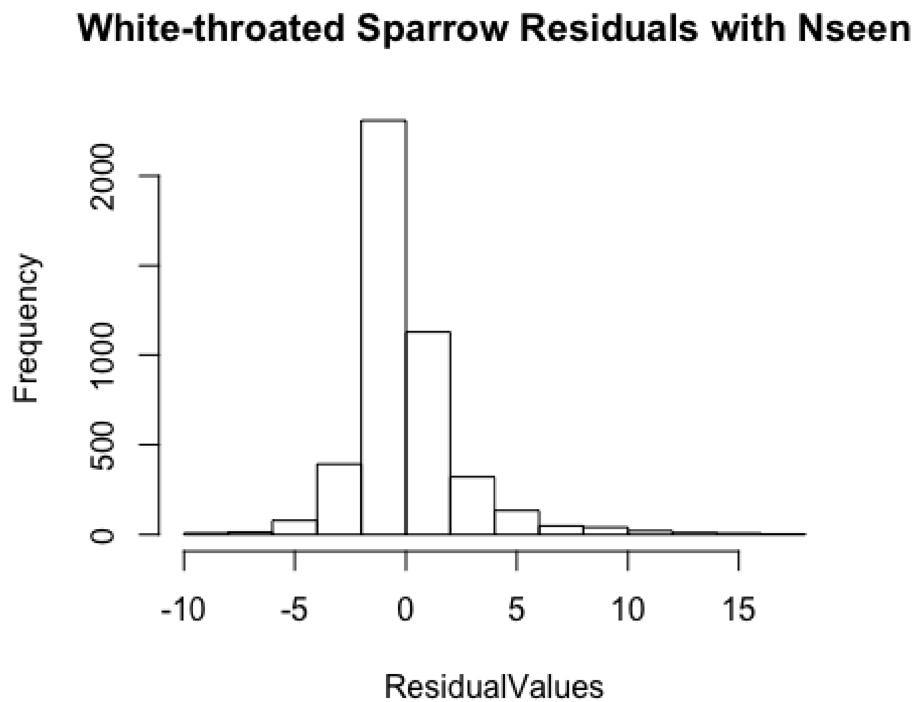

Figure 3.4: The histogram of the residual values of the linear model constructed with the Nseen values. This is for the White-throated Sparrow.

such as temperature anomalies, are just the scaled values.

These fixed effects are added together by the addition sign to indicate they are additive properties. Recall, the addition sign indicates that the variables are to be modeled as additive effects. These effects are such that there is no dependability among the additive properties [1]. In other words the effect of temp_anomalies on log(NSeen) does not depend on the value of scale(wspeed_kmph) or any of the other additive effects.

The last additive effects to cover are the random effects, otherwise known as grouping variables. The random effects allow the data models to account for the grouping of certain observations. For example, the syntax (1|city) tells the method lmer to fit a linear model with a varying-intercept group effect using the variable

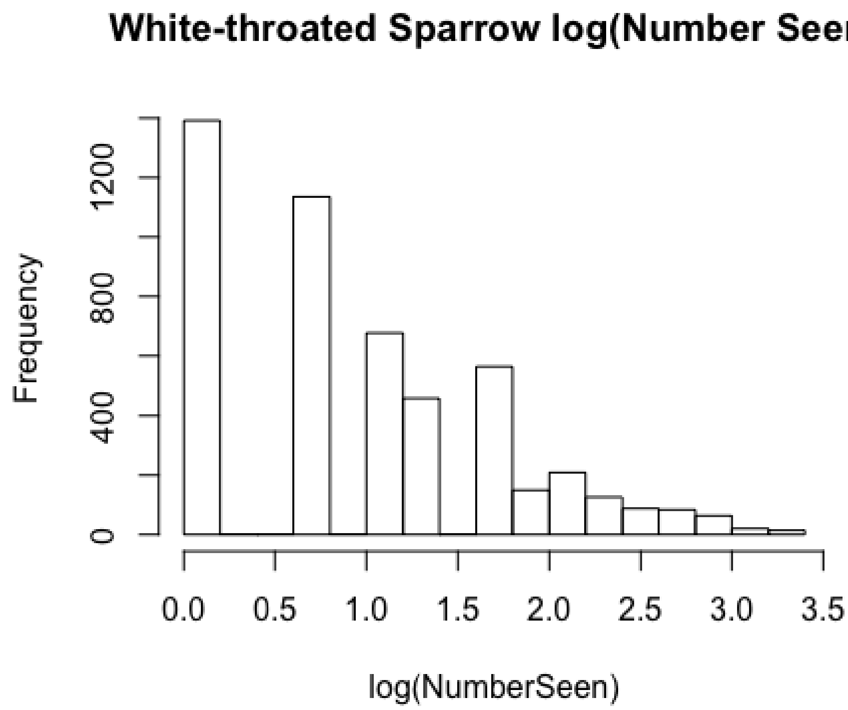

Figure 3.5: The histogram of the natural log of Nseen values of the White-throated Sparrow.

city. In other words it specifies that all observations that have the same city attribute belong to the same group, as the weather data may have been collected from the same weather stations. As such, those tuples should not be considered independent from each other and the linear model is adjusted accordingly. Since the focus of this project is body mass and the influences of climate conditions specifically, the exact influence of city and other random effects on the log(Nseen) is ignored.

To summarize, the grouping variables are (1|city), (1|NHalfDays), (1|LOC_ID) and (1|EFFORT_HRS_ATLEAST). The effort of hours (NHalfDays) and number of half days (EFFORT_HRS_ATLEAST) are groupings based on the effort of the observer making that particular observation. This is important to take into account as different observers of PFW have varying degrees of precision in making a bird feeder observations. Additionally, one observer may submit multiple observations. Those

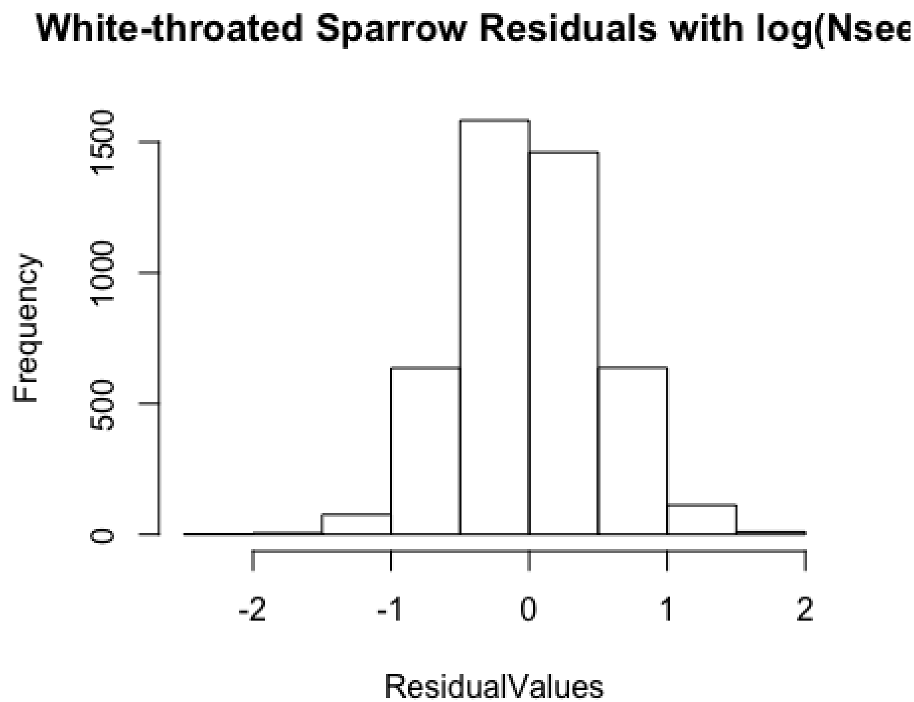

Figure 3.6: The histogram of the residual values of the linear model constructed with the log(Nseen)values. This is for the White-throated Sparrow.

observations need to be grouped.

Location ID (LOC_ID) is the location grouping. Previously the city grouping was discussed, this grouping is based on the location of the actual bird feeder. This is also important as there may be multiple observations from a single bird feeder and thus those observations should be viewed as a group. As a reminder, this location ID was already provided in the original PFW data set.

Table 3.4: Random effects for the linear-mixed model of the White-throated Sparrow.

Groups	Name	Variance	Std.Dev.
LOC_ID	Intercept	0.24015	0.4901
city	Intercept	0.0834	0.2889

Table 3.4: Random effects for the linear-mixed model of the White-throated Sparrow.

Groups	Name	Variance	Std.Dev.
NHalfDays	Intercept	0.002411	0.0491
EFFORT_HRS_ATLEAST	Intercept	0.002052	0.0453

To clearly illustrate the effects of the grouping variables, the random effects table for the previous example of White-throated Sparrow is presented in Table 3.4. The most important attribute is the Variance, as this determines how much influence the random effect variable has on the model. The higher the variance, the stronger the effect of that variable. It is clear from the table that the LOC_ID and the city has the most influence, while (1|EFFORT_HRS_ATLEAST) and (1|NHalfDays) has less influence, but there is influence nonetheless. For this project all of the Random Effects tables were inspected to ensure there was at least some amount of variation among the groups. A full list of the Random Effects table for each of the 23 species is presented in Appendix A.

The (1|BirdSpp) term is for the generalized model as it groups observations by species name. Meaning, data points that have the same bird species name are considered non-independent from each other and thus are treated accordingly for the model. The idea here being to observe any large-scale trends that occur when considering all the species' datasets together as one large dataset.

Finally, the last few parameters are flags and the source of the data. The REML=F flag forces the model to use the maximum likelihood for parameter estimates rather than restricted maximum likelihood (REML). The relevance of this is that if flag was the default value of REML then it may produce a non-reliable model. This is especially true when you compare models with different effects, which is true for this analysis.

The data= term sets the data source. The data source for the 23 models were the 23 data frames, one for each bird species. Data frames are a form of data representation that allows the code methods to process it. In other words, raw data of many formats are converted to data frames to be used in code functions. For the generalized model, the data source was the combined data frame which contained all of tuples from each of the 23 species. The data frames contain all of the attributes required for constructing the models. In the end, a total of 23 linear-mixed models were constructed and analyzed, one for each bird species. In addition, a general model was constructed with all of the species combined.

The na.action= “na.fail” flag sets the action for the method when a null value in the parameters of lmer is encountered. The “na.fail” ensures that the linear-mixed model object is only returned by lmer if there are no null values in the method arguments. All of the null values were removed from the constructed datasets during the preprocessing steps before the lmer method was actually called, but the flag ensures that no null values were overlooked.

3.12 Model Outputs

The returned object from the method lmer is the constructed linear-mixed model. The R summary method is then called on this returned model. This summary method returns the slope estimates, standard errors for the slope estimates and the t value for each of the independent variables. As reminder, the t value is the slope estimate divided by the standard error value. Also recall that the independent variables for this study are the temperature anomalies, precipitation anomalies and scaled wind speeds. The full summary table for the White-throated Sparrow linear-mixed model, which includes all of the attributes for each fixed effect, is presented in Table 3.5.

Table 3.5: Model Summary for White-throated Sparrow. Recall that fixed effects is another name for independent variables.

Fixed Effects	Estimate	SE	t value
Intercept	7.09E-01	4.88E-02	14.522
Temp. Anom.	-2.02E-02	1.66E-03	-12.183
scale(wind speed)	5.18E-03	1.03E-02	0.503
Precip. Anom.	4.32E-03	6.47E-03	0.667

The slope estimate, labeled "Estimate", of the linear model is the effect of the climate attributes on the natural log of the number seen of the species. For example, lets just focus on the temperature anomalies. If the the slope is a negative value then more individuals of the bird species were seen at the feeders when the mean daily temperatures were lower than the average winter temperature for that year. If the slope estimate is a positive value, then as the temperature anomalies got more negative the less the individuals visit the bird feeders. In other words, as the temperature increased more individuals of the species were observed. Recall that slope magnitude is not an accurate measure of effect strength. The effect strength is measured by the t value.

The t value is also analogous to the statistical p-value in terms of model significance, except for the fact that models must have a value of above 2 or less than -2 in order to be considered significant. With p-value, results are considered significant if the p-value is below 0.05 [11]. For this project, the t value is also a measure of how strong the effect of that particular climate attribute is on the number of individuals seen at the feeders. For example, lets take the anomalies from precipitation levels. If the absolute value of the t value for this fixed effect is lower than 2, then the relationship is not significant enough to warrant insights. If the t value is above 2 then we can safely draw insights from the relationship as the effect of precipitation is strong enough [16].

The magnitude of the absolute value of the t value is also significant, as the greater the value the stronger the effect [17]. For example, lets say one species has a t value of 2.5 while another bird species has a t value of 4.5. The second species is more effected by the anomalies in precipitation than the first species.

3.13 Residual Values

In order to use the linear-models for analysis, the main assumption is that the residual values must have an approximately normal distribution. A histogram for each of the 23 bird species' data models and the generalized data model was created and inspected to ensure normal distribution. This process was done through visual inspection. This conclusion was reached after talks with Professor Francis, who has years of experience modeling biological data.

Two examples of what we determined to be approximate normal distributions are presented in Figures 3.7 and 3.8. These are the residual values for the bird species White-throated Sparrow and Northern Cardinal, respectively. Finally, Figures 3.9 and 3.10 present the histograms of the residuals that we deemed not approximately normal. They are from the species Hairy Woodpecker and Red-bellied Woodpecker, respectively.

Lastly, Q-Q normal plots for the residual values were also created and inspected to ensure the approximate normal distributions. For these scatter plots, if the points approximately form a line, then one can assume the residual values follow a rough normal distribution. An example of a Q-Q normal plot representing an approximately normal distribution is presented in Figure 3.11. This plot is for the White-throated Sparrow. And finally, a Q-Q normal plot representing a non-normal distribution is

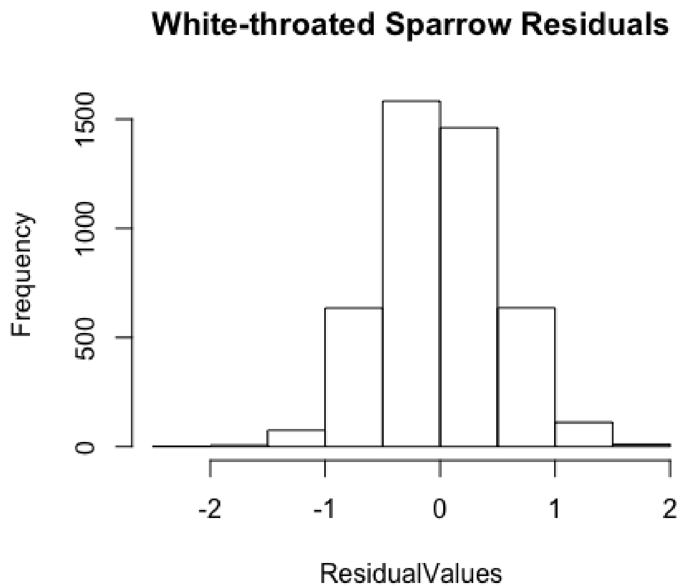

Figure 3.7: Residual values from the White-throated Sparrow data.

presented in Figure 3.12. This plot is for the Hairy Woodpecker.

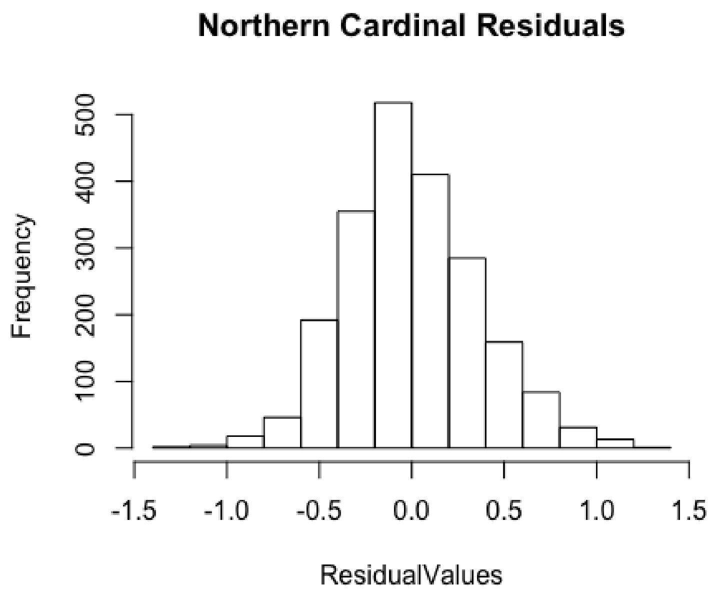

Figure 3.8: Residual values from the Northern Cardinal data.

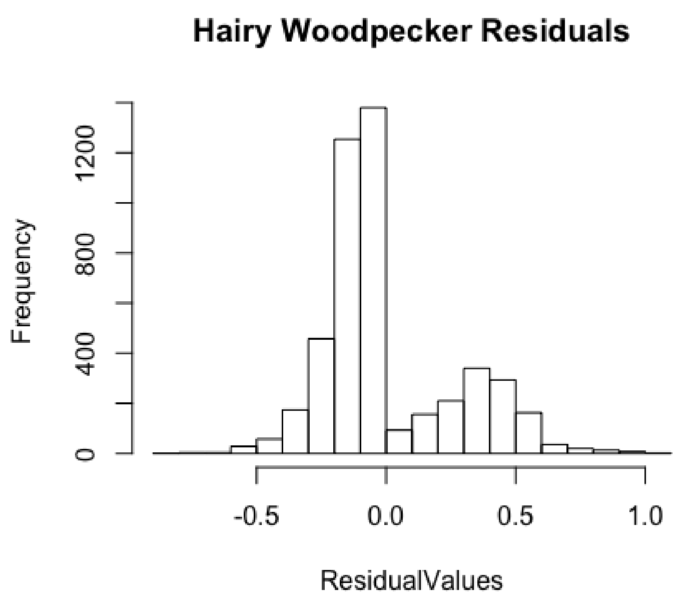

Figure 3.9: Residual values from the Hairy Woodpecker data.

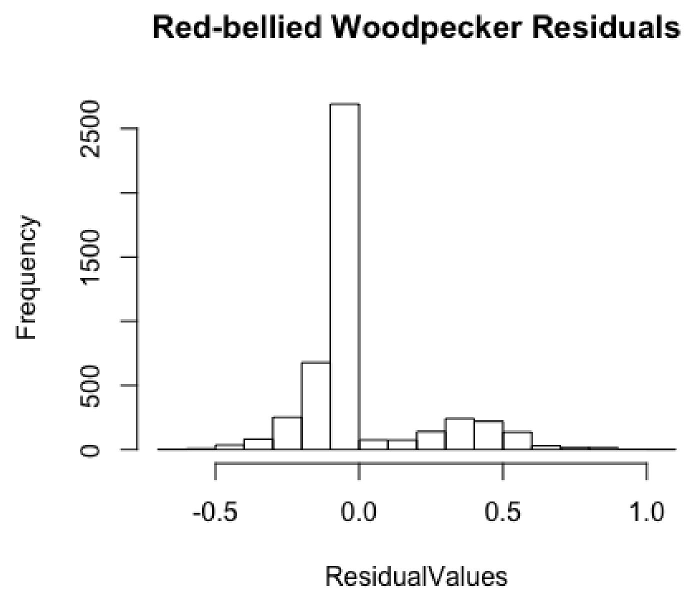

Figure 3.10: Residual values from the Red-bellied Woodpecker data.

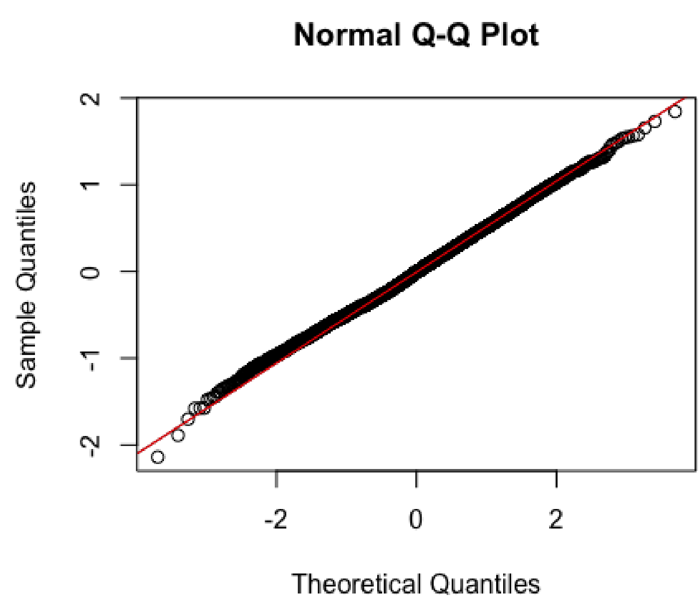

Figure 3.11: Residual values from the White-throated Sparrow data. Observe how the majority of the points follow the red line.

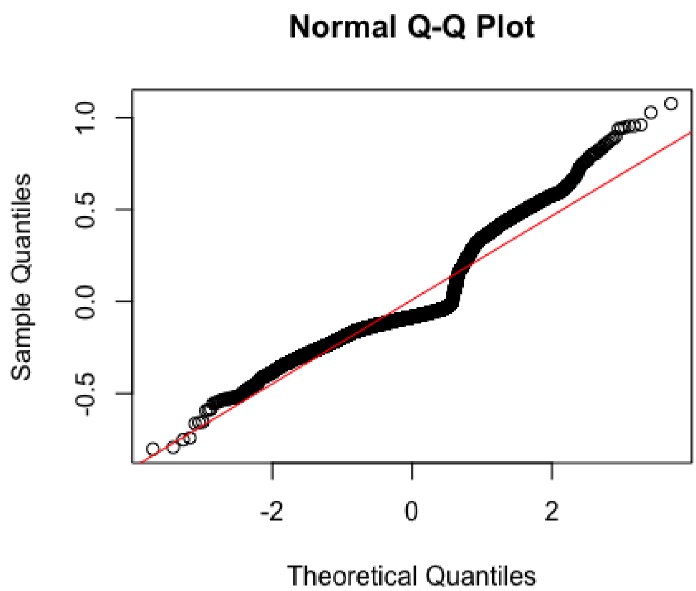

Figure 3.12: Residual values from the Hairy Woodpecker data. Observe the major deviations from the red line.

Chapter 4

RESULTS

4.1 Responses to Temperature Anomalies

As a reminder, our hypothesis for the temperature anomalies was that the body mass of the species largely determined the number of individuals seen at the feeders during periods of cold temperatures. The larger massed birds have an easier time retaining heat and thus would not be required to visit the bird feeders as often. The smaller birds have to eat more food in order to make up for the calories that are lost trying to stay warm during the colder days. We ignored the evolutionary and behavior differences among the species and instead focused on the thermoregulation requirements of the birds in regards to the body mass. Thus the body mass values of the species play a key role in the results and further analysis.

We expected to see that as the body mass of the bird species increases so does the slope estimate of the model, as the slope becomes less negative. The sign of the slope estimate is key. With the t value, we expected the opposite relationship, as the body mass increases the t value decreases in magnitude, as this indicates there is less of an effect of temperature variation.

Table 4.1: Linear-mixed Models: log(Nseen) as a function of temperature anomalies. Slope Estimates and t-values for all of the bird species for the temperature anomalies analysis. The rows are in increasing order by body mass of the species.

Bird Species	Body Mass (grams)	Slope Estimate	t-value	Sig.
Brown Creeper	7.5	-0.0006486	-1.215	not sig.
Chestnut-backed Chickadee	9.5	-0.00862	-4.276	sig.

Table 4.1: Linear-mixed Models: log(Nseen) as a function of temperature anomalies. Slope Estimates and t-values for all of the bird species for the temperature anomalies analysis. The rows are in increasing order by body mass of the species.

Bird Species	Body Mass (grams)	Slope Estimate	t-value	Sig.
Mountain Chickadee	11	-0.002494	-2.003	sig.
Black-capped Chickadee	11.5	-0.001069	-0.624	not sig.
Chipping Sparrow	13.5	0.001731	1.274	not sig.
Pine Siskin	15	-0.008119	-2.653	sig.
American Goldenfinch	15.5	-0.006899	-3.334	sig.
Common Redpoll	15.5	-0.003651	-1.336	not sig.
Tufted Titmouse	22	-0.004384	-4.018	sig.
Dark-eyed Junco	24	-0.028332	-15.38	sig.
Downy Woodpecker	24.5	-0.0032084	-3.35	sig.
White-throated Sparrow	27	-0.021324	-13.44	sig.
Northern Cardinal	45	-0.017232	-8.11	sig.
Northern Mockingbird	51.5	0.0033151	6.43	sig.
Pine Grosbeak	56.4	-0.02334	-2.929	sig.
Evening Grosbeak	63.5	0.01041	2.001	sig.
Hairy Woodpecker	67.5	-0.0010992	-1.475	not sig.
Red-bellied Woodpecker	73.5	-0.0008638	-1.402	not sig.
European Starling	78	-0.015726	-7.305	sig.
Blue Jay	85	-0.015082	-5.624	sig.
Common Grackle	108	0.015218	5.003	sig.
Mourning Dove	121	-0.016746	-6.754	sig.
Black-billed Magpie	177.5	-0.003586	-2.368	sig.
All species	NA	-7.38E-03	-15.31	sig.

Table 4.1 presents the slope estimates for the linear-models constructed with temperature deviations as the independent variable and the natural log of number seen as the dependent variable. The t values are also presented in the column labeled t-value. The majority of the observations are considered significant enough to draw insights from, with only 6 of the 23 observations having absolute t values lower than 2.

In order to ensure the validity of our analysis, all non significant rows were ignored in the constructed data set. These ignored tuples all had an absolute t value of less than 2, indicating that the relationships were not strong enough to support insights. All of the remaining tuples contain statistically valid data, which can be confidently analyzed. The scatter plot of the slope estimates and body mass values from Table 4.1 is presented in Figure 4.1.

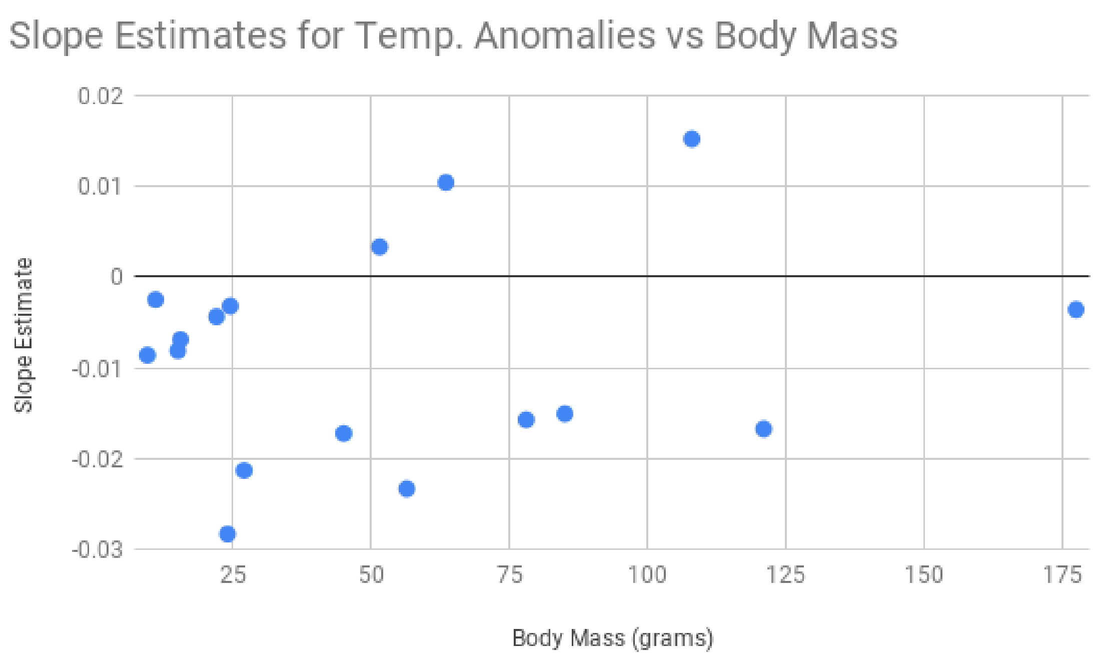

Figure 4.1: The plot shows the relationship between the slope estimates for the linear-mixed models with temperature anomalies being the independent variable.

From the scatter plot in Figure 4.1 it becomes clear that many of the bird species tended to visit the bird feeders more often as the temperature fell below average. Only 3 bird species, all of them with mid range body mass, have positive slopes. Interestingly, the Evening Grosbeak is one of those species. This stands out because this species is also suffering range-wide decline in population at the feeders [31]. The largest of the bird species analyzed, with a body mass of 177.5 grams, has a negative slope. Thus in a larger sense our hypothesis was correct, bird species in general are more abundant at the feeders during especially at colder temperatures.

However, there is yet no support from this study that body mass in fact determines the number of individuals seen. This is made clear by the scatter plot. The smallest massed species, those with body masses from 9.5 grams to 24.5 grams, have negative slope estimate values that are almost equivalent to the negative slope estimates of the largest species, with a body mass of 177.5 grams.

For a visual representation of the data points being analyzed in these models refer to Figures 4.2 and 4.3. These are for the species American Goldfinch and Evening Grosbeak, respectively. The figures plot the relationship between the natural log of Nseen and temperature anomalies. There are far too many data points in the plots to observe any trends by eye, thus the use of the models throughout this analysis. However, it is still worthwhile to scan the data point distributions that are present.

Another simple linear model was constructed with just the body mass and slope estimate data. The code for constructing this model using R programming language is presented below. The independent variable for this model is the natural log of the body mass and the dependent variable is the slope estimate for the temperature anomalies. The natural log is used again in order to make the left skew less prominent. This allows for a more reliable linear model.

```r
model <- lm(tempAnomaliesSlopeEstimates ~ log(bodyMass))
```

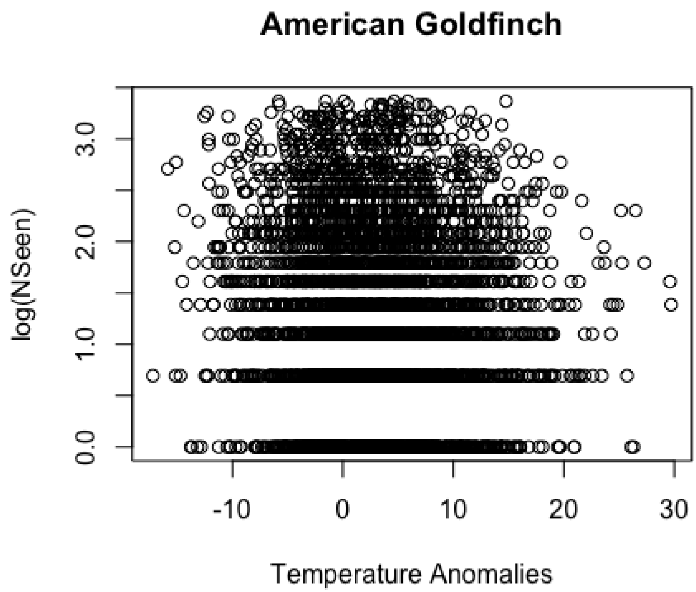

Figure 4.2: Data points for the American Goldfinch.

According to our hypothesis there should be a positive relationship or slope for this simple model. However, the relationship presented in the linear model above is not significant enough to warrant insight. It has a t value of 0.326, which is far below 2, indicating there is a very weak effect of body mass on the slope estimates. This is further evidence that body mass does not determine the frequency of feeder visits due to cold temperatures.

The second value of importance is the t value for the species linear-mixed models. As a reminder the magnitude of the t-value determined the strength of the effect of the temperature anomalies on the abundance at the feeders. Table 4.1 also presents the body mass of the species and the t-values. The negative and positive signs on the values are only related to the slope intercept sign, and serve no purpose in assessing the strength of the effect, thus only the absolute values are used for the analysis.

We expected to observe a negative relationship between the body mass and the strength of the effect, otherwise known as the magnitude of the t value. In other

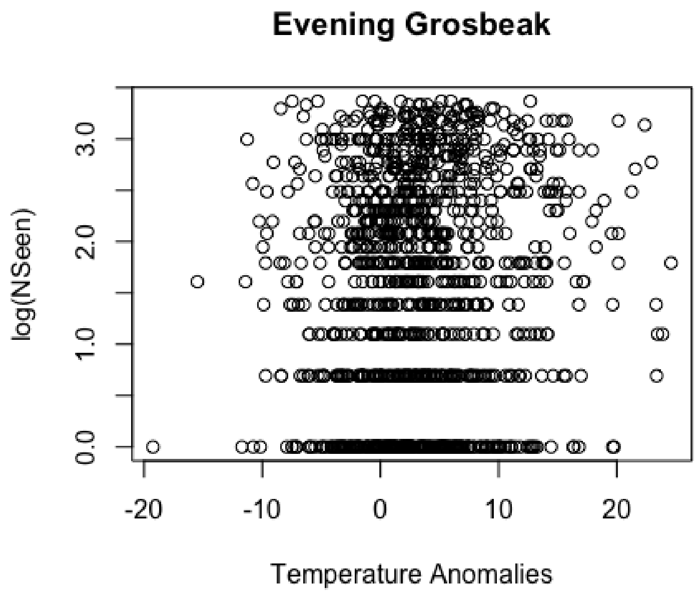

Figure 4.3: Data points for the Evening Grosbeak.

words, as the body mass increases the t values, or the effect of temperature anomalies, decreases, but not dipping below 2. The scatter plot for the absolute t values values is presented in Figure 4.2.

The first interesting aspect of the scatter plot are the 3 outliers. The smallest, in terms of body mass, of the 3 is the Dark-eyed Junco with a t value magnitude of 15.38 and a body mass of 24 grams. Next is the White-throated sparrow with a t value of 13.44 and mass of 27 grams. Lastly, there is the Northern Cardinal with t value of 8.11 and mass of 45 grams. These three species are especially effected by fluctuations in temperature when it comes to visiting feeders for food.

The second interesting characteristic of the scatter plot, besides the outliers, is that all of the species for this study, regardless of the body mass, were about equally effected by fluctuations in temperatures. This is supported by the fact that the majority of the t values fall within 3 units of 5. A clear example is again the Black-billed Magpie with the body mass of 177.5 grams. It has the t value magnitude of

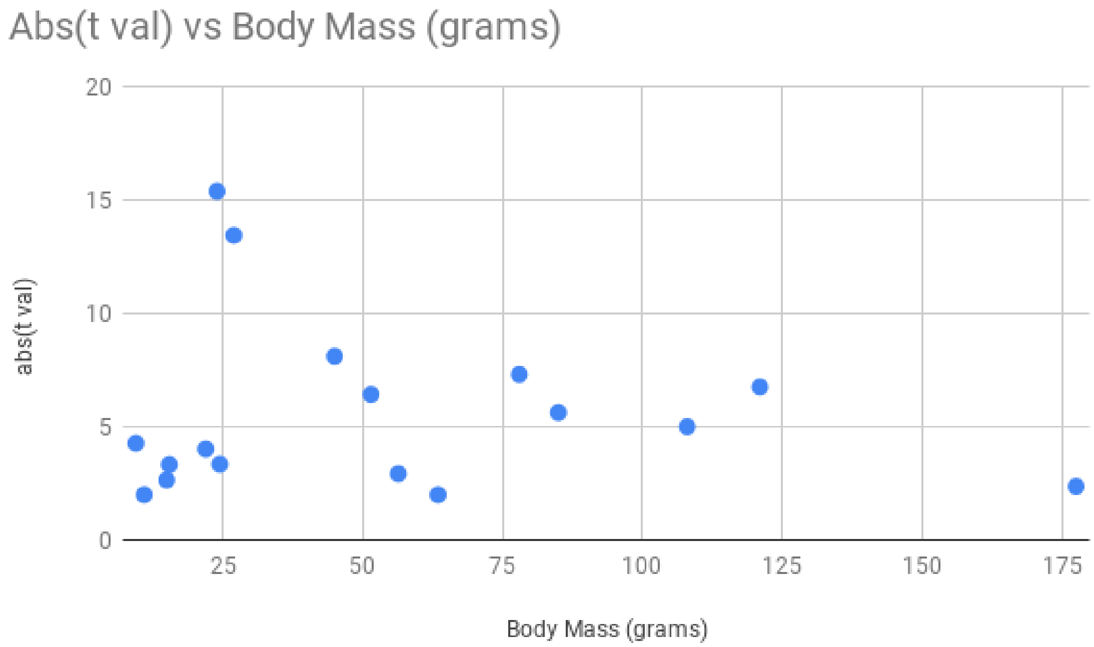

Figure 4.4: The plot shows the relationship between the t value magnitudes for the linear-mixed models with temperature anomalies being the independent variable.

2.368, which is in the same range as some of the smaller bird species. This is contrary to what we expected for this group of species, as there is no negative relationship.

The last row of Table 4.1 with the Bird Species attribute of All species, presents the summary of the linear-mixed model built with all of the same variables as the previous 23 species. Except there was one key difference, the bird species was also considered a random effect. Recall, that the data source for this was the data frame of all of the individuals bird species' data frames combined. This model allows us to observe the general responses to temperature anomalies across all of the species for the study.

With this generalized linear model, the slope estimate for all bird species is negative. This is further support for the idea that many of the species visit the bird feeders more often during colder than average days. Additionally, the magnitude of the t value is well above 2, indicating that the effect of temperature anomalies is strong.

4.2 Responses to Wind Speed

For the analysis with mean wind speed for the day, we expected a species wide decline in abundance at the feeders as the wind speed increased in magnitude. The reason for this is that the conditions for flight are not ideal on windy days. The body mass values of the species are not necessary for this analysis. From the results, there is no evidence of this effect, in fact there is evidence of the opposite effect. When analyzing the bird species individually there appears to be little effect of wind, however with the generalized linear-mixed model for all species there appears to be a positive relationship.

Table 4.2: Linear-mixed Models: log(Nseen) as a function of the scaled wind speed values.

Bird Species	Body Mass (grams)	Slope Estimate	t-value	Sig.
Brown Creeper	7.5	-0.0042868	-1.33	not sig.
Chestnut-backed Chickadee	9.5	0.004216	0.548	not sig.
Mountain Chickadee	11	1.08E-02	1.173	not sig.
Black-capped Chickadee	11.5	-0.0151193	-1.393	not sig.
Chipping Sparrow	13.5	0.014614	1.447	not sig.
Pine Siskin	15	2.88E-04	0.015	not sig.
American Goldenfinch	15.5	0.053091	3.959	sig.
Common Redpoll	15.5	-0.01463	-0.872	not sig.
Tufted Titmouse	22	-5.78E-03	-0.826	not sig.
Dark-eyed Junco	24	0.045943	5.581	sig.

Table 4.2: Linear-mixed Models: log(Nseen) as a function of the scaled wind speed values.

Bird Species	Body Mass (grams)	Slope Estimate	t-value	Sig.
Downy Woodpecker	24.5	0.0158335	2.527	sig.
White-throated Sparrow	27	5.18E-03	0.503	not sig.
Northern Cardinal	45	3.33E-02	2.29	sig.
Northern Mockingbird	51.5	2.03E-03	0.651	not sig.
Pine Grosbeak	56.4	-1.80E-02	-0.319	not sig.
Evening Grosbeak	63.5	3.69E-02	1.296	not sig.
Hairy Woodpecker	67.5	1.08E-03	0.221	not sig.
Red-bellied Woodpecker	73.5	8.49E-03	2.028	sig.
European Starling	78	2.68E-02	1.856	not sig.
Blue Jay	85	-0.014708	-0.819	not sig.
Common Grackle	108	0.03435	1.738	not sig.
Mourning Dove	121	-1.62E-02	-1.008	not sig.
Black-billed Magpie	177.5	0.021912	2.007	sig.
All species	NA	1.31E-02	5.054	sig.

Table 4.2 presents the slope estimates and t values for the 23 species and the All species linear-mixed model. Many of the models are not significant, with t value magnitudes below 2. However, there is one species with an unusually high absolute t value of 5.581. This species is the Dark-eyed Junco.

It is interesting that the Dark-eyed Junco so far is the only species particularly affected by wind and temperature anomalies. The effect of wind is also interesting as the slope estimate is a positive value. This indicates that as wind speed increased, more individuals were seen at the feeders. This is the opposite of what we were expecting. Perhaps it becomes more difficult to forage for food elsewhere on windy days, making the feeders an easier option. More research is required for this inquiry.

This pattern continues with the All species model. The slope estimate for all of the species is a positive value, moreover the magnitude of the t value is 5.05. This is far above the requirement for significance, indicating the wind speed generally has a strong effect on the species for this study.

4.3 Responses to Precipitation Anomalies

Our hypothesis for precipitation anomalies was very similar to the wind speed analysis. We expected that as the precipitation levels became higher than average the flight conditions became less ideal, even life threatening. It is possible for birds to suffer torpor and even death when the feathers become too wet [48]. This would result in fewer individuals visiting the feeders. The body mass values of the species is not necessary for this analysis. The results, however, did not support our hypothesis.

Table 4.3: Linear-mixed Models: log(Nseen) as a function of precipitation anomalies.

Bird Species	Body Mass (grams)	Slope Estimate	t-value	Sig.
Brown Creeper	7.5	0.0010779	0.24	not sig.
Chestnut-backed Chickadee	9.5	0.007065	0.975	not sig.
Mountain Chickadee	11	3.59E-02	1.548	not sig.
Black-capped Chickadee	11.5	-0.0001098	-0.297	not sig.
Chipping Sparrow	13.5	-0.016613	-1.394	not sig.
Pine Siskin	15	4.89E-02	1.253	not sig.
American Goldenfinch	15.5	0.034164	1.481	not sig.
Common Redpoll	15.5	0.035895	1.135	not sig.
Tufted Titmouse	22	-5.01E-03	-0.576	not sig.

Table 4.3: Linear-mixed Models: log(Nseen) as a function of precipitation anomalies.

Bird Species	Body Mass (grams)	Slope Estimate	t-value	Sig.
Dark-eyed Junco	24	0.050267	4.129	sig.
Downy Woodpecker	24.5	0.0108151	1.575	not sig.
White-throated Sparrow	27	4.32E-03	0.667	not sig.
Northern Cardinal	45	2.75E-02	1.087	not sig.
Northern Mockingbird	51.5	-2.92E-03	-1.733	not sig.
Pine Grosbeak	56.4	-3.84E-02	-0.256	not sig.
Evening Grosbeak	63.5	1.95E-02	0.668	not sig.
Hairy Woodpecker	67.5	1.91E-02	2.032	sig.
Red-bellied Woodpecker	73.5	3.37E-03	0.505	not sig.
European Starling	78	-1.42E-05	-0.001	not sig.
Blue Jay	85	-0.025149	-0.779	not sig.
Common Grackle	108	0.005112	0.198	not sig.
Mourning Dove	121	3.40E-03	0.119	not sig.
Black-billed Magpie	177.5	0.03558	1.535	not sig.
All species	NA	1.11E-05	2.60E-02	not sig.

Table 4.3 presents the slope estimates and t values for precipitation anomalies as the independent variable. The values are for the 23 species linear-mixed models and the all species general linear-mixed model. One can clearly observe that many of the bird species are not affected by precipitation levels when it came to visiting feeders during the winter. The generalized linear-mixed model is also insignificant. There are even more bird species not significantly effected than with the wind speed analysis.

The most interesting aspect of these results is once again the Dark-eyed Junco. This species exhibits a positive slope estimate, meaning that as precipitation levels increased, more individuals were spotted at the bird feeders. The t value of 4.129 is also well above 2, indicating the effect of precipitation anomalies is strong. It appears that the Dark-eyed Junco is more sensitive to the climate attributes of temperature variations, wind speed and precipitation variations than the other 22 birds analyzed. Furthermore it is the only species to have significant linear-mixed models with all of 3 independent variables.

4.4 Validation

There are 2 threats to validity given the analysis of the data presented above. The first threat is regarding the generalization of the results to other bird species. In other words, can the insights for the results be applied to other bird species? As mentioned in the earlier section of Bird Species Selection, great care was taken in choosing the appropriate bird species for this project. We worked closely with Professor Francis to select species that are not migratory during winter, with exception of the Evening Grosbeak, allowing us to only study birds that experience winter temperatures. Additionally, a variety of body mass values are represented through the selected species allowing for a wider analysis in regards to the affect of body mass on feeder visits. Lastly, there are total of 23 selected species, which after consultations with Professor Francis was deemed a large enough sample size for this study. This sample size provides an adequate estimate of the non-migratory U.S. winter bird communities in the regions of focus, given the limitations of the Weather Underground API.

The second threat to validity targets the internal analysis of the responses of the bird species. In other words, is the analysis from the data models of temperature anomalies, wind speed and precipitation anomalies accurate? This threat is disarmed with use of the t values. Only models with t value magnitudes greater than or equal to 2 were considered for analysis, as these models exhibit relationships significant enough to warrant insight. Models not meeting this criteria were ignored, as the insights drawn from them may not be valid. Additionally, in order to use the linear-mixed models certain assumptions must be satisfied. Great care was taken to make sure the assumptions were not violated. These checks were performed through the use of histograms and Q-Q normal plots of the residual values, as previously described in the Assumptions and Residual sections.

Chapter 5

CLOSING DISCUSSION

5.1 Project Summary

To summarize, the general idea of our hypothesis is correct in regards to the responses to temperature anomalies. Many of the bird species for this study tended to visit the bird feeders more during colder than average temperatures. There is no evidence yet of body mass determining the response and the strength of the effect of temperature anomalies.

For the wind speed analysis, many of the species display no significant relationship. However, the general model for all species indicates that overall as the wind speed increased, more individuals visited the feeders. This general model has a significant t value of 5.054. Precipitation anomalies has an even less significant effect on the abundance of birds at the feeders when compared to wind speed. Our hypothesis on both of these climatic attributes were wrong, as we expected less birds at the feeders during periods of heavy rain or high wind.

The Dark-eyed Junco stands out as a species because it has a strong relationship with all 3 of the climatic attributes focused on. Another stand out species in this study is the Evening Grosbeak. As mentioned in the Background sections, the population of this species is on the decline in certain areas. Our study indicates that the Evening Grosbeak is significantly impacted by temperature. However, it, along with only 2 other species, displayed a positive slope estimate value. Meaning that as the temperature gets warmer, more individuals are seen at the feeders. It is possible that the areas in which the Evening Grosbeak are declining are too cold. More definitive research needs to be done, especially to account for the migratory behavior of this species, before a conclusion can be reached.

5.2 Future Work

This project further demonstrates the value of correlating climate data with biological data, such as the Project FeederWatch dataset. As discussed in the Methodology sections there was a limitation on how much of the PFW data set could processed. This limitation was mainly due to the Weather Underground's API, as it only allowed a set amount of calls to collect the climate data. The goal for the future is to collect climate data for all of the original PFW data set. For this study only a sample of the original data set was used. In order to append all the climate data, additional licenses or a different type of WU license must purchased in order to allow for more API calls per day.

This project has also revealed interesting species specific patterns to climate, such as with the Dark-eyed Junco and Evening Grosbeak. Though this study did not take into account any phylogeny or evolutionary history of the species, it is worthwhile to explore these ideas in the context of climate. Adding climate data to all of the PFW tuples will definitely aid in answering the next questions, and for that reason it should be among the first objectives to tackle. Once that is achieved, there is the potential to conduct many more studies. These projects can focus both on the species specific scale and the larger scale, involving many different bird species.

BIBLIOGRAPHY

[1] Additive models vs. interaction models. https://chemicalstatistician.wordpress.com/2014/03/07/applied-statistics-lesson-of-the-day-additive-models-vs-interaction-models-in-2-factor-experimental-designs/ (Accessed on 04/15/2018).

[2] All about birds. https://www.allaboutbirds.org/ (Accessed on 04/15/2018).

[3] Assumptions of linear regression. http://www.statisticssolutions.com/assumptions-of-linear-regression/ (Accessed on 06/06/2018).

[4] Cal Poly Github. http://www.github.com/CalPoly.

[5] Clinton francis. https://bio.calpoly.edu/content/clinton-francis (Accessed on 04/15/2018).

[6] Getting started with mixed effect models in r. https://www.jaredknowles.com/journal/2013/11/25/getting-started-with-mixed-effect-models-in-r (Accessed on 04/15/2018).

[7] Noaa data tools: Local climatological data (lcd). https://www.ncdc.noaa.gov/cdo-web/datatools/lcd (Accessed on 03/10/2018).

[8] P-value. http://www.bodowinter.com/tutorial/bw_LME_tutorial2.pdf (Accessed on 06/03/2018).

[9] P-value. http://www.bodowinter.com/tutorial/bw_LME_tutorial1.pdf (Accessed on 06/03/2018).

[10] P-value. https://www.st-andrews.ac.uk/media/capod/students/mathssupport/mixedeffectsknir.pdf (Accessed on 06/03/2018).

[11] P-value. http://blog.minitab.com/blog/understanding-statistics/what-can-you-say-when-your-p-value-is-greater-than-005 (Accessed on 06/03/2018).

[12] Photo quiz 5: On your field marks, get set, go! https://www.allaboutbirds.org/photo-quiz-5-on-your-field-marks-get-set-go/ (Accessed on 04/15/2018).

[13] *Project FeederWatch - Scientific Publications.* https://feederwatch.org/explore/scientific-publications/ (Accessed on 04/05/2018).

[14] Random intercept models. http://www.bristol.ac.uk/cmm/learning/videos/random-intercepts.html#exvar (Accessed on 06/06/2018).

[15] Residual value definition. http://stattrek.com/statistics/dictionary.aspx?definition=residual (Accessed on 06/06/2018).

[16] T-value. https://www.allbusiness.com/barrons_dictionary/dictionary-t-value-4942040-1.html (Accessed on 05/15/2018).

[17] T-value interpretation. http://blog.minitab.com/blog/statistics-and-quality-data-analysis/what-is-a-t-test-and-why-is-it-like-telling-a-kid-to-clean-up-that-mess-in-the-kitchen (Accessed on 05/15/2018).

[18] Understanding q-q plots.
https://data.library.virginia.edu/understanding-q-q-plots/
(Accessed on 06/16/2018).

[19] Weather underground: About our data.
https://www.wunderground.com/about/data (Accessed on 04/05/2018).

[20] Weather underground: About us.
https://www.wunderground.com/about/our-company (Accessed on 05/03/2018).

[21] Weather underground: Api documentation.
https://www.wunderground.com/weather/api/d/docs?d=data/history
(Accessed on 04/05/2018).

[22] Weather underground: Quality control improvements.
https://www.wunderground.com/blog/PWSmet/pws-updates-quality-control-improvements.html (Accessed on 05/03/2018).

[23] Worldclim. http://worldclim.org/version2 (Accessed on 05/25/2018).

[24] Worldclim homepage. http://worldclim.org (Accessed on 05/25/2018).

[25] *Project Overview - FeederWatch*, 2018.
https://feederwatch.org/about/project-overview/ (Accessed on 04/24/2018).

[26] T. Alerstam. Wind as selective agent in bird migration. *Ornis Scandinavica*, pages 76–93, 1979.

[27] D. Bates, M. Mächler, B. Bolker, and S. Walker. Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, 67(1):1–48, 2015.

[28] J. K. Baum and B. Worm. Cascading top-down effects of changing oceanic predator abundances. *Journal of Animal Ecology*, 78(4):699–714, 2009.

[29] C. E. Bock and L. W. Lepthien. Synchronous eruptions of boreal seed-eating birds. *The American Naturalist*, 110(974):559–571, 1976.

[30] D. N. Bonter and C. B. Cooper. Data validation in citizen science: a case study from project feederwatch. *Frontiers in Ecology and the Environment*, 10(6):305–307, 2012.

[31] D. N. Bonter and M. G. Harvey. Winter survey data reveal rangewide decline in evening grosbeak populations. *The Condor*, 110(2):376–381, 2008.

[32] J. M. Bower and H. Bolouri. *Computational modeling of genetic and biochemical networks*. MIT press, 2001.

[33] W. E. Bradshaw and C. M. Holzapfel. Evolutionary response to rapid climate change. *Science(Washington)*, 312(5779):1477–1478, 2006.

[34] M. C. Brittingham and S. A. Temple. Does winter bird feeding promote dependency?(¿ promueve dependencia la alimentación de aves durante el invierno?). *Journal of Field Ornithology*, pages 190–194, 1992.

[35] J. M. Carr and S. L. Lima. High wind speeds decrease the responsiveness of birds to potentially threatening moving stimuli. *Animal Behaviour*, 80(2):215–220, 2010.

[36] E. D BRODIE III and N. H. RUSSELL. The consistency of individual differences in behaviour: temperature effects on antipredator behaviour in garter snakes. *Animal behaviour*, 57(2):445–451, 1999.

[37] J. B. Dunning Jr. *CRC handbook of avian body masses*. CRC press, 2007.

[38] B. Erni, F. Liechti, L. G. Underhill, and B. Bruderer. Wind and rain govern the intensity of nocturnal bird migration in central europe-a log-linear regression analysis. *Ardea*, 90(1):155–166, 2002.

[39] T. C. Grubb. Weather-dependent foraging behavior of some birds wintering in a deciduous woodland. *The Condor*, 77(2):175–182, 1975.

[40] B. K. Hartup, H. O. Mohammed, G. V. Kollias, and A. A. Dhondt. Risk factors associated with mycoplasmal conjunctivitis in house finches. *Journal of wildlife diseases*, 34(2):281–288, 1998.

[41] W. M. Hochachka, A. A. Dhondt, et al. Spatial variation in an avian host community: implications for disease dynamics. *EcoHealth*, 6(4):540–545, 2009.

[42] R. T. Holmes, J. C. Schultz, and P. Nothnagle. Bird predation on forest insects: an exclosure experiment. *Science*, 206(4417):462–463, 1979.

[43] IPCC. *Climate Change 2014–Impacts, Adaptation and Vulnerability: Regional Aspects*. Cambridge University Press, 2014.

[44] D. N. Jones and S. James Reynolds. Feeding birds in our towns and cities: a global research opportunity. *Journal of avian biology*, 39(3):265–271, 2008.

[45] A. S. Jump and J. Penuelas. Running to stand still: adaptation and the response of plants to rapid climate change. *Ecology Letters*, 8(9):1010–1020, 2005.

[46] S. C. Kendeigh. Effect of temperature and season on energy resources of the english sparrow. *The Auk*, 66(2):113–127, 1949.

[47] S. C. Kendeigh. Energy requirements for existence in relation to size of bird. *The Condor*, 72(1):60–65, 1970.

[48] R. Kennedy. Direct effects of rain on birds: a review. *British Birds*, 63:401–414, 1970.

[49] G. Körtner, R. M. Brigham, and F. Geiser. Metabolism: winter torpor in a large bird. *Nature*, 407(6802):318, 2000.

[50] R. Louv and J. W. Fitzpatrick. *Citizen science: Public participation in environmental research*. Cornell University Press, 2012.

[51] E. T. Miller, D. N. Bonter, C. Eldermire, B. G. Freeman, E. I. Greig, L. J. Harmon, C. Lisle, W. M. Hochachka, and D. Stephens. Fighting over food unites the birds of north america in a continental dominance hierarchy. *Behavioral Ecology*, 28(6):1454–1463, 2017.

[52] J. Pinheiro, D. Bates, S. DebRoy, D. Sarkar, and R Core Team. *nlme: Linear and Nonlinear Mixed Effects Models*, 2018. R package version 3.1-137.

[53] L. F. Pitelka and P. M. W. Group. Plant migration and climate change: a more realistic portrait of plant migration is essential to predicting biological responses to global warming in a world drastically altered by human activity. *American Scientist*, pages 464–473, 1997.

[54] K. Princé and B. Zuckerberg. Climate change in our backyards: the reshuffling of north america's winter bird communities. *Global change biology*, 21(2):572–585, 2015.

[55] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2017.

[56] S. K. Robinson and R. T. Holmes. Foraging behavior of forest birds: the relationships among search tactics, diet, and habitat structure. *Ecology*, 63(6):1918–1931, 1982.

[57] C. Strong, B. Zuckerberg, J. L. Betancourt, and W. D. Koenig. Climatic dipoles drive two principal modes of north american boreal bird irruption. *Proceedings of the National Academy of Sciences*, 112(21):E2795–E2802, 2015.

[58] P. Sutton, D. Roberts, C. Elvidge, and H. Meij. A comparison of nighttime satellite imagery and population density for the continental united states. *Photogrammetric engineering and remote sensing*, 63(11):1303–1313, 1997.

[59] M. A. Waller and S. E. Fawcett. Data science, predictive analytics, and big data: a revolution that will transform supply chain design and management. *Journal of Business Logistics*, 34(2):77–84, 2013.

[60] D. S. Wilcove and M. Wikelski. Going, going, gone: is animal migration disappearing. *PLoS biology*, 6(7):e188, 2008.

[61] B. Zuckerberg, D. N. Bonter, W. M. Hochachka, W. D. Koenig, A. T. DeGaetano, and J. L. Dickinson. Climatic constraints on wintering bird distributions are modified by urbanization and weather. *Journal of Animal Ecology*, 80(2):403–413, 2011.

APPENDICES

Appendix A

RANDOM EFFECTS TABLES FOR THE BIRD SPECIES

Note the "Residual" row presents the random error, or ϵ. These values were ignored for analysis.

Table A.1: American Goldfinch

Groups	Name	Variance	Std. Dev.
LOC_ID	Intercept	0.2769126	0.52622
city	Intercept	0.0170802	0.13069
NHalfDays	Intercept	0.0009424	0.0307
Effort_Hrs_Atleast	Intercept	0.0091336	0.09557
Residual		0.5117337	0.71536

Table A.2: Black-billed Magpie

Groups	Name	Variance	Std. Dev.
LOC_ID	Intercept	0.18718	0.4326
city	Intercept	0.02261	0.1504
NHalfDays	Intercept	0	0
Effort_Hrs_Atleast	Intercept	0.01063	0.1031
Residual		0.25656	0.5065

Table A.3: Black-capped Chickadee

Groups	Name	Variance	Std. Dev.
LOC_ID	Intercept	0.17376	0.41684
city	Intercept	0.034414	0.18551
NHalfDays	Intercept	0.001475	0.03841
Effort_Hrs_Atleast	Intercept	0.00497	0.07049
Residual		0.173698	0.41677

Table A.4: Blue Jay

Groups	Name	Variance	Std. Dev.
LOC_ID	Intercept	0.1797844	0.42401
city	Intercept	0.0164977	0.12844
NHalfDays	Intercept	0.0002742	0.01656
Effort_Hrs_Atleast	Intercept	0.0076209	0.0873
Residual		0.2386128	0.48848

Table A.5: Brown Creeper

Groups	Name	Variance	Std. Dev.
LOC_ID	Intercept	0.0055615	0.07458
city	Intercept	0.0006456	0.02541
NHalfDays	Intercept	0	0
Effort_Hrs_Atleast	Intercept	0	0
Residual		0.0408037	0.202

Table A.6: Chestnut-backed Chickadee

Groups	Name	Variance	Std. Dev.
LOC_ID	Intercept	0.16533	0.4066
city	Intercept	0.01781	0.13346
NHalfDays	Intercept	0	0
Effort_Hrs_Atleast	Intercept	0.00689	0.08301
Residual		0.20546	0.45327

Table A.7: Chipping Sparrow

Groups	Name	Variance	Std. Dev.
LOC_ID	Intercept	0.089829	0.29972
city	Intercept	0.007325	0.08559
NHalfDays	Intercept	0.001693	0.04115
Effort_Hrs_Atleast	Intercept	0	0
Residual		0.244769	0.49474

Table A.8: Common Grackle

Groups	Name	Variance	Std. Dev.
LOC_ID	Intercept	0.171613	0.41426
city	Intercept	0.003542	0.05951
NHalfDays	Intercept	0	0
Effort_Hrs_Atleast	Intercept	0.006067	0.07789
Residual		0.829976	0.91103

Table A.9: Common Redpoll

Groups	Name	Variance	Std. Dev.
LOC_ID	Intercept	0.175402	0.41881
city	Intercept	0.129613	0.36002
NHalfDays	Intercept	0.002685	0.05182
Effort_Hrs_Atleast	Intercept	0	0
Residual		0.808522	0.89918

Table A.10: Dark-eyed Junco

Groups	Name	Variance	Std. Dev.
LOC_ID	Intercept	0.2346261	0.48438
city	Intercept	0.0455615	0.21345
NHalfDays	Intercept	0.0008802	0.02967
Effort_Hrs_Atleast	Intercept	0.0136565	0.11686
Residual		0.4199296	0.64802

Table A.11: Downy Woodpecker

Groups	Name	Variance	Std. Dev.
LOC_ID	Intercept	0.081578	0.28562
city	Intercept	0.010689	0.10339
NHalfDays	Intercept	0.001016	0.03187
Effort_Hrs_Atleast	Intercept	0.00653	0.08081
Residual		0.120721	0.34745

Table A.12: European Starlin

Groups	Name	Variance	Std. Dev.
LOC_ID	Intercept	0.195598	0.44226
city	Intercept	0.005131	0.07163
NHalfDays	Intercept	0	0
Effort_Hrs_Atleast	Intercept	0.012898	0.11357
Residual		0.694701	0.83349

Table A.13: Evening Grosbeak

Groups	Name	Variance	Std. Dev.
LOC_ID	Intercept	0.42765	0.65395
city	Intercept	0.05215	0.22836
NHalfDays	Intercept	0	0
Effort_Hrs_Atleast	Intercept	0.00264	0.05138
Residual		0.71113	0.84329

Table A.14: Hairy Woodpecker

Groups	Name	Variance	Std. Dev.
LOC_ID	Intercept	0.0356293	0.18876
city	Intercept	0.0038428	0.06199
NHalfDays	Intercept	0.0003211	0.01792
Effort_Hrs_Atleast	Intercept	0.0023614	0.04859
Residual		0.0813611	0.28524

Table A.15: Mountain Chickadee

Groups	Name	Variance	Std. Dev.
LOC_ID	Intercept	0.0356293	0.18876
city	Intercept	0.0038428	0.06199
NHalfDays	Intercept	0.0003211	0.01792
Effort_Hrs_Atleast	Intercept	0.0023614	0.04859
Residual		0.0813611	0.28524

Table A.16: Mourning Dove

Groups	Name	Variance	Std. Dev.
LOC_ID	Intercept	0.265376	0.51515
city	Intercept	0.006261	0.07912
NHalfDays	Intercept	0.00337	0.05806
Effort_Hrs_Atleast	Intercept	0.017	0.13038
Residual		0.4813	0.69376

Table A.17: Northern Mockingbird

Groups	Name	Variance	Std. Dev.
LOC_ID	Intercept	8.12E-03	0.0901
city	Intercept	1.02E-03	0.03192
NHalfDays	Intercept	5.82E-05	0.00763
Effort_Hrs_Atleast	Intercept	0.00E+00	0
Residual		3.79E-02	0.19456

Table A.18: Northern Cardinal

Groups	Name	Variance	Std. Dev.
LOC_ID	Intercept	0.272015	0.52155
city	Intercept	0.059878	0.2447
NHalfDays	Intercept	0.009218	0.09601
Effort_Hrs_Atleast	Intercept	0.008608	0.09278
Residual		0.225365	0.47473

Table A.19: Pine Grosbeak

Groups	Name	Variance	Std. Dev.
LOC_ID	Intercept	0.1908	0.4368
city	Intercept	0.03666	0.1915
NHalfDays	Intercept	0	0
Effort_Hrs_Atleast	Intercept	0	0
Residual		0.54211	0.7363

Table A.20: Pine Siskin

Groups	Name	Variance	Std. Dev.
LOC_ID	Intercept	0.2292297	0.47878
city	Intercept	0.0280484	0.16748
NHalfDays	Intercept	0	0
Effort_Hrs_Atleast	Intercept	0.0006648	0.02578
Residual		0.6927975	0.83234

Table A.21: Red-bellied Woodpecker

Groups	Name	Variance	Std. Dev.
LOC_ID	Intercept	0.026744	0.16354
city	Intercept	0.002125	0.0461
NHalfDays	Intercept	0	0
Effort_Hrs_Atleast	Intercept	0.002155	0.04642
Residual		0.057358	0.23949

Table A.22: Tufted Titmouse

Groups	Name	Variance	Std. Dev.
LOC_ID	Intercept	0.109101	0.33031
city	Intercept	0.015438	0.12425
NHalfDays	Intercept	0.001416	0.03763
Effort_Hrs_Atleast	Intercept	0.00652	0.08075
Residual		0.160669	0.40084

Table A.23: White-throated Sparrow

Groups	Name	Variance	Std. Dev.
LOC_ID	Intercept	0.240152	0.4901
city	Intercept	0.083445	0.2889
NHalfDays	Intercept	0.002411	0.0491
Effort_Hrs_Atleast	Intercept	0.002052	0.0453
Residual		0.320554	0.5662

Appendix B

GUIDELINES FOR PFW DATA

Quick guide to FeederWatch data

Project FeederWatch is a citizen science program operated by the Cornell Lab of Ornithology in cooperation with Bird Studies Canada. More information: www.feederwatch.org

For more details about the data and how they can be accurately used in a biologically relevant manner, contact David Bonter (dnb23@cornell.edu) or Wes Hochachka (wmh6@cornell.edu).

Note that the FeederWatch protocol is a repeated measures design. Participants (ID) are reporting from the same location (LOC_ID) as often as 21 times each winter with many people reporting for many years. As such, the data are amenable to occupancy modeling.

Data quality: All FeederWatch reports (starting in 2006) are passed through geographically and temporally explicit filters to flag records that are unexpected for the location/month. Records that do not trigger a flag are given a '1' in the Valid field and enter the database without review. Records that trigger the flag ('0' in the Valid field) and are reviewed by a biologist who may clear the flag, ask the participant for supporting information, or reject the report. Reports marked Valid = 1 and Reviewed = 1 have been reviewed by a biologist and approved. Reports marked as Invalid in the database are not included in this file. [This is a similar review system as the one used by eBird.] Undoubtedly, some of the reports in the database involve incorrect identifications or reports by participants who do not completely follow the FeederWatch protocol. As with all datasets, use with caution. See Bonter & Cooper 2012. Data validation in citizen science: A case study from Project FeederWatch. *Frontiers in Ecology and the Environment* 10:305-307.

FeederWatch Protocol: In brief, participants (N ~ 20,000 in 2014) count birds at their own bird feeders (usually) as often as once/week from early November to early April annually. The count period is 2 days long. Participants report the MAXIMUM number of each species in view at any one time during the 2-day count period. This methodology avoids double-counting of the same individual birds. The complete protocol is available here: http://feederwatch.org/about/detailed-instructions/

Data fields included in the .csv file:

ID = The participants' unique Cornell Lab of Ornithology identification number. **Note that these ID numbers should not be shared or made publically available as they could, in theory, be used to access personal information of participants.** They should simply be used as a unique identifier of the individuals making observations.

Latitude/Longitude: In decimal degrees. Note that these locations are identified by the participants with varying degrees of accuracy. Prior to 2000, all sites were given the latitude and longitude of the centroid of the ZIP code (identified as "POSTCODE LAT/LONG LOOKUP" in the ENTRY_TECHNIQUE field). Since our online data entry system was developed (late 1999), we've maintained a series of different systems for identifying the count locations that are far more accurate. Contact David Bonter for more details.

StatProv = US State or Canadian Province of count location.

ENTRY_TECHNIQUE: See Latitude/Longitude above.

FW_YEAR = The FeederWatch season. Seasons run from November to April. For instance, the FW_YEAR "PFW_1992" indicates the season running from November 1991 to April 1992.

Month/Day/Year = First date of the 2-day observation period.

NHalfDays = Number of Half Days of observation during the 2-day count period. This is a measure of observer effort. Categorical (1, 2, 3, or 4 half days).

EFFORT_HRS_ATLEAST = Another measure of observer effort. Categorical (< 1, 1-4 hours, 4-8 hours, > 8 hours). Researchers at the Cornell Lab have found both measures of effort are informative in models.

BirdSpp = 6-letter species code. Contact David Bonter if the key is needed.

NSeen = Number of individuals seen (remember that this is the maximum number of the species in view at any one time over the 2-day observation period).

Valid/Reviewed. See above.

LOC_ID = Unique identifier of the location of observation. Note that participants (ID) can have multiple count locations (LOC_ID).

SUB_ID = Submission Identifier. This is a unique identifier of the entire checklist submitted by a participant from a count period. All species observations submitted on the same checklist have the same SUB_ID.

OBS_ID = Observation identifier. This is a unique identifier for the single observation. All species reported on a single checklist receive different OBS_IDs.

Appendix C

SCRIPT IMPLEMENTATION USING PYTHON FOR APPENDING CITIES

```
from multiprocessing import Process
import netCDF4 as nc
import numpy as np
import pandas as pd
import sys
import json
import time
import matplotlib.pyplot as plt
from urllib.request import urlopen
from math import sin, cos, sqrt, atan2, radians

# CLASS FOR ORGANIZING THE TOP CITIES
class city:
    def __init__(self, lat, lon, name):
        self.lat = lat
        self.lon = lon
        self.name = name

# CALCULATING THE DISTANCE GIVEN 2 LAT LONG POINTS
def calcDistance(lat1, lon1, lat2, lon2):
    R = 6373.0

    latitude1 = radians(lat1)
    longitude1 = radians(lon1)
    latitude2 = radians(lat2)
    longitude2 = radians(lon2)

    dlon = longitude2 - longitude1
```

```
    dlat = latitude2 - latitude1

    a = sin(dlat / 2)**2 + cos(latitude1) * cos(latitude2) * sin(dlon / 2)**2
    c = 2 * atan2(sqrt(a), sqrt(1 - a))

    return (R * c)

# ATTACH A CITY TO THE ROW IF POSSIBLE
def processRow(lat, lon, cities):
    distance = sys.maxsize
    city = ''
    for c in cities:
        temp = calcDistance(lat, lon, c.lat, c.lon)

        # find the closes city
        if (temp < distance):
            city = c.name
            distance = temp

    # IF THE POINT IS WITHIN 40KM OF CITY
    if (distance <= 40):
        return city
    else:
        return ''


# FIND ALL ROWS CLOSE TO A MAJOR CITY IN THE STATE
def processState(s, states_dict):
  curr_state = ''
  s['city'] = ''
  # FOR EACH ROW FIND A CITY IF POSSIBLE
```

```
  for index,row in s.iterrows():
    curr_state = row['StatProv']
    lat = row['LATITUDE']
    lon = row['LONGITUDE']
    cities = states_dict[row['StatProv']]
    city = processRow(lat, lon, cities)
    s.set_value(index, 'city', city)

  # DELETING ROWS WITH NO CITY
  for i,r in s.iterrows():
    if(r['city'] == ''):
      s.drop(i, inplace=True)

  if (curr_state != ''):
    file_name = curr_state + '.csv'
    # create the csv for state with the city added
    s.to_csv(file_name, sep=',')


# MAIN METHOD
if __name__ == '__main__':
  states_dict = {}

  statesConv = {
      'AL':'"AL"',
      'AK':'"AK"',
      'AZ':'"AZ"',
      'AR':'"AR"',
      'CA':'"CA"',
      'CO':'"CO"',
      'CT':'"CT"',
      'DE':'"DE"',
```

```
        'FL': '"FL"',
        'GA': '"GA"',
        'HI': '"HI"',
        'ID': '"ID"',
        'IL': '"IL"',
        'IN': '"IN"',
        'IA': '"IA"',
        'KS': '"KS"',
        'KY': '"KY"',
        'LA': '"LA"',
        'ME': '"ME"',
        'MD': '"MD"',
        'MA': '"MA"',
        'MI': '"MI"',
        'MN': '"MN"',
        'MS': '"MS"',
        'MO': '"MO"',
        'MT': '"MT"',
        'NE': '"NE"',
        'NV': '"NV"',
        'NH': '"NH"',
        'NJ': '"NJ"',
        'NM': '"NM"',
        'NY': '"NY"',
        'NC': '"NC"',
        'ND': '"ND"',
        'OH': '"OH"',
        'OK': '"OK"',
        'OR': '"OR"',
        'PA': '"PA"',
        'RI': '"RI"',
        'SC': '"SC"',
```

```
        'SD':'"SD"',
        'TN':'"TN"',
        'TX':'"TX"',
        'UT':'"UT"',
        'VT':'"VT"',
        'VA':'"VA"',
        'WA':'"WA"',
        'WV':'"WV"',
        'WI':'"WI"',
        'WY':'"WY"',
        'DC':'"DC"'
}

major_cities = pd.read_csv('topCities.csv')

# Change name of file
# birdfeeder_df = pd.read_csv('PFW2011-12_Subset.csv')

done = []
state_dfs = []

# ORGANIZING BIRDFEEDER DATA BY STATE

for index,row in major_cities.iterrows():
    curr_state = row['state_abbrv']
    if curr_state not in done:
        done.append(curr_state)
        state = statesConv[curr_state]
        string = 'StatProv==' + state
        state_subset = birdfeeder_df.query(string).copy()
        state_dfs.append(state_subset)

```

```
        # UNCOMMENT THE CODE BELOW IF YOU WANT STATE SPECIFIC CSV's of
  the PFW data

    # if not state_subset.empty:
    #   csv_title = 'PFW2011-12_' + curr_state + '.csv'
    #   state_subset.to_csv(csv_title, sep=',')


# CREATING A HASHMAP (STATE : CITES LIST)

for index,row in major_cities.iterrows():
  x = city(row['lat'], row['long'], row['City'])
  if (row['state_abbrv'] in states_dict):
    states_dict[row['state_abbrv']].append(x)
  else :
    states_dict[row['state_abbrv']] = []
    states_dict[row['state_abbrv']].append(x)

# MAIN LOOP FOR GOING THROUGH ALL OF THE STATES
# Parallel processes
i = 0
while i < len(state_dfs):
  s1 = state_dfs[i]
  i += 1

  p1 = Process(target=processState, args=(s1,states_dict))
  p1.start()
```

Appendix D

SCRIPT IMPLEMENTATION USING PYTHON FOR COLLECTING WEATHER UNDERGROUND CLIMATE DATA

```
from multiprocessing import Process
import netCDF4 as nc
import numpy as np
import pandas as pd
import sys
import json
import time
import math
import os.path
import matplotlib.pyplot as plt
from urllib.request import urlopen
from math import sin, cos, sqrt, atan2, radians

# MAIN METHOD
if __name__ == '__main__':

  # Change to correct filename
  bf_df = pd.read_csv('houfin_data/houfin_cities_0.csv')

  counter = 0

  # Make sure the license number is correct
  str1 = 'http://api.wunderground.com/api/753a4a7523f842e3/history_'

  bf_df['temp_mean_F'] = np.nan
  bf_df['temp_max_F'] = np.nan
  bf_df['temp_min_F'] = np.nan
```

```
bf_df['wspeed_mph'] = np.nan
bf_df['percip_inches'] = np.nan


for index,row in bf_df.iterrows():
  # Making only 10 api calls per minute

  # 10 sec over 1 minute (just for buffer, really only requires 1
 minute)
  if (counter >= 10):
    time.sleep(70)
    counter = 0


  date = row['Year']
  date *= 100
  date += row['Month']
  date *= 100
  date += row['Day']
  date_str = str(date)

  str2 = '/q/' + row['StatProv'] + '/' + row['city'] + '.json'

  # get the json file that conatins the weather data for that day
  api_link = str1 + date_str + str2
  f = urlopen(api_link)
  json_string = f.read()
  parsed_json = json.loads(json_string)

  # getting temp_f (temperature in F)
  if ('history' in parsed_json):
    if ('dailysummary' in parsed_json['history']):
```

```
      if (len(parsed_json['history']['dailysummary']) == 1):
        if ('meantempi' in parsed_json['history']['dailysummary'][0]):
          temp = parsed_json['history']['dailysummary'][0]['meantempi']
          bf_df.loc[index,'temp_mean_F'] = temp

        if ('maxtempi' in parsed_json['history']['dailysummary'][0]):
          temp_max = parsed_json['history']['dailysummary'][0]['maxtempi']
          bf_df.loc[index,'temp_max_F'] = temp_max

        if ('mintempi' in parsed_json['history']['dailysummary'][0]):
          temp_min = parsed_json['history']['dailysummary'][0]['mintempi']
          bf_df.loc[index,'temp_min_F'] = temp_min

        if ('meanwindspdi' in parsed_json['history']['dailysummary'][0]):
          wspeed = parsed_json['history']['dailysummary'][0]['meanwindspdi']
          bf_df.loc[index,'wspeed_mph'] = wspeed

        if ('precipi' in parsed_json['history']['dailysummary'][0]):
          percip = parsed_json['history']['dailysummary'][0]['precipi']
          bf_df.loc[index,'percip_inches'] = percip

    counter += 1


# Outputing the final csv with temperatures
# Make sure the title is correct (change #)
```

```
bf_df.to_csv('houfin_data/houfin_climate_0.csv', sep=',')
```