Summer 2022

Analysis of Hawk Mountain Sanctuary Observation Data from 1976 through 2021

Dale E. Parson

Analysis of Hawk Mountain Sanctuary Observation Data from 1976 through 2021

Dale E. Parson, Ph.D., 2022
Professor of Computer Science and Information Technology
Kutztown University of Pennsylvania
parson@kutztown.edu
https://faculty.kutztown.edu/parson/

ACKNOWLEDGEMENTS

I would like to express gratitude to Dr. Laurie Goodrich, Sarkis Acopian Director of Conservation Science at Hawk Mountain Sanctuary [1], for working with me and my students from fall 2019 through the present in understanding the mission, activities, and raptor observation datasets of Hawk Mountain. I also thank Dr. Lorin Basden Arnold, provost and vice president for Academic Affairs, and the Kutztown University Research Committee for funding a summer 2022 work stipend and 2022-2023 student worker support for two graduate students to extend this work and two undergraduates to enhance graphical visualization software of raptor observations for use with the public at Hawk Mountain. Master thesis candidate Eric Burgos collected, cleaned, and formatted NOAA climate data from the Allentown Airport [2] for use in his thesis and supplied this data and support for the current study. Graduate student Faith Evans supplied the base map for Figure 10 by annotating a Google Map satellite view with Hawk Mountain landmarks as part of an independent study in animated data visualization in 2020. Dr. Shane Parson, a professional environmental engineer working largely with climate-related flood planning, pointed me at learning more about the *heat island effect* in relation to North Lookout versus Allentown Airport climate measure differences.

OBJECTIVES

The primary objective is to correlate climate change data to changes in raptor observations. Additional objectives include uncovering trends in climate observations at Hawk Mountain's North Lookout and the Allentown Airport throughout the observation period, and to examine trends in raptor observation properties independent of climate changes.

OUTLINE

This document is an edited set of notes for use by raptor specialists using Hawk Mountain Sanctuary observation data from 1934 through 2021. Analysis concentrates on 1976 through 2021 because of missing data in earlier years and accelerated climate trends in later years. It is best read in order because earlier sections explain data modeling techniques used throughout.

Section 1 is on The Data and its Preparation.

Section 2 is on Analysis of Weather Patterns.

Section 3 is on Analysis of Climate-to-Raptor Patterns.

Section 3 climate -> raptor analyses appear below with links here: Red-Tailed Hawk, Sharp-Shinned Hawk, American Kestrel, Bald Eagle, Golden Eagle, Broad-Winged Hawk, Total Raptors. Follow-up analyses by graduate students during the 2022-2023 academic year include Cooper's Hawk, Osprey, Northern Harrier, Northern Goshawk, and Rough-Legged Hawk.

The Reference Section is at the end.

1. THE DATA

1a. Data Sources

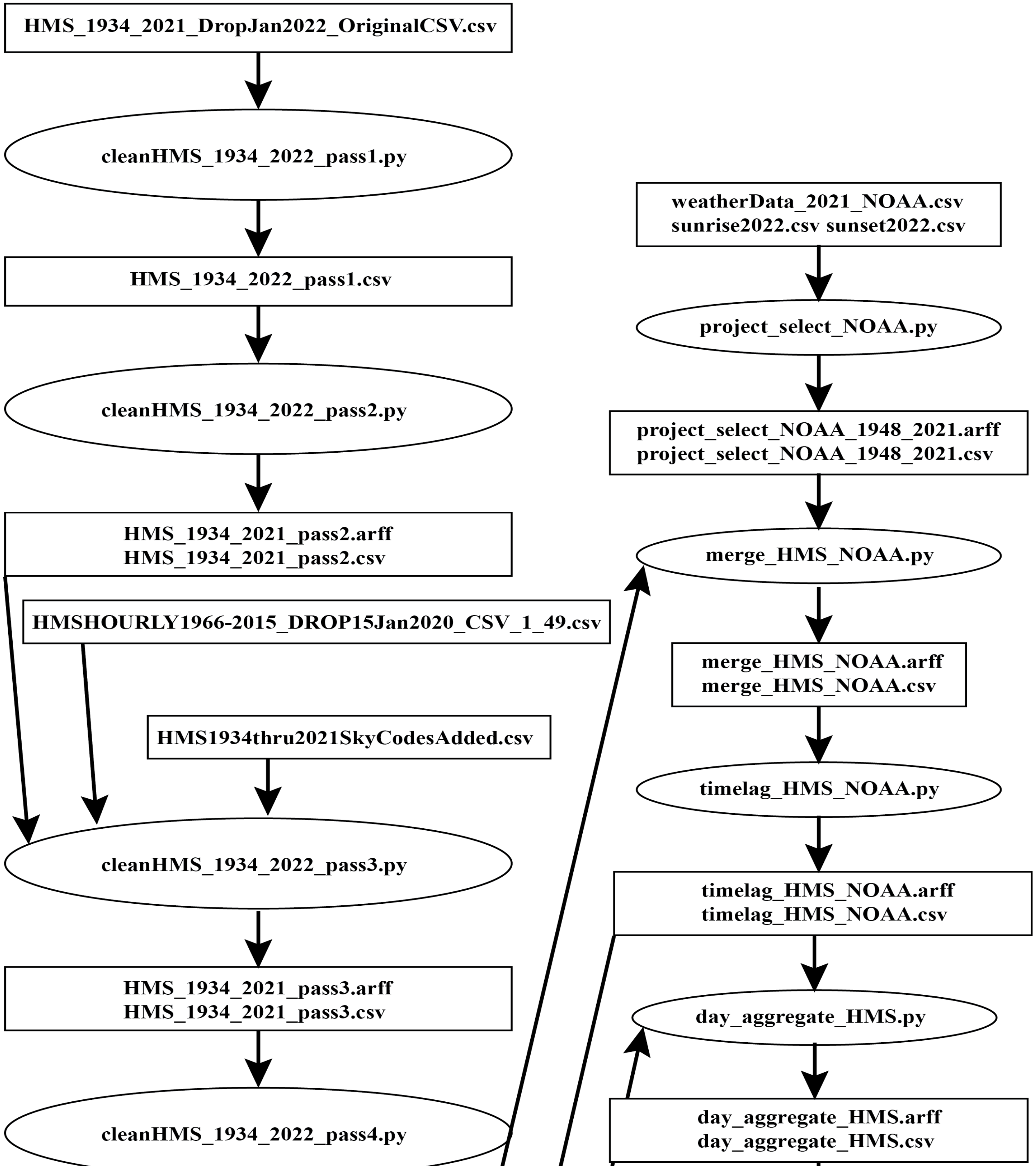
Dr. Goodrich sent climate and raptor observational data as an Excel spreadsheet for 2017-2018 in fall 2019 for use in CSC458 Data Mining and Predictive Analytics I at Kutztown, giving students and myself an in-person tutorial at the Hawk Mountain Visitor Center, followed by a walk to North Lookout, during the semester. In January 2020 Dr. Goodrich sent observation data for 1966 through 2015 for use by two graduate students doing an analysis project in CSC558 Data Mining and Predictive Analytics II in the spring. This dataset (hereafter *HMJan2020Data*) consists of 49 *attributes*, a.k.a. *properties* (Excel columns), including 31 raptor counts, 9 weather columns, and 9 columns of date / time and observer records. There are 48,649 *records*, a.k.a. *instances* (Excel rows) in this data. In January 2022 Dr. Goodrich sent observation data for 1934 through 2021 to extend the years of observation and extend and fill in missing values for the HMJan2020Data. This dataset (hereafter HMJan2022Data) consists of 94 attributes, a.k.a. properties (Excel columns), including 76 raptor counts (including sub-counts of immature, adult, unknown, and others), 8 weather columns, and 10 columns of date / time and observer records. There are 63,127 records, a.k.a. instances (Excel rows) in this data. Up through 1966 there is only one observation record per day. They become more numerous in 1967, settling on multiple 60-minute observation records. The standard observation period per Dr. Goodrich is August 15 through December 15 [3]. I have processed data from August 1 through December 31, discarding the rest for this study. On June 23, 2022 Dr. Goodrich sent an additional Excel file with SkyCode values starting in 2000 to augment missing data.

Master thesis candidate Eric Burgos collected, cleaned, and formatted NOAA climate data from the Allentown Airport [2] for use in his thesis and supplied this data for the current study on June 7, 2022.

I downloaded NOAA sunrise and sunset date-times on June 22, 2022 [4], using these to average NOAA Allentown Airport daylight records for the years of one-per-day Hawk Mountain observation records.

1b. Data Cleaning and Organization

This section gives a summary of data transformation passes using my custom Python scripts. These scripts and derived data are automated and can be deployed to any Linux machine with the necessary Python libraries installed [5, 6]. Much of this custom code development has taken place on a Windows 10 PC using the Linux-like Cygwin environment [7]. CPU-intensive analysis has also used a multiprocessor Linux server at Kutztown University. Detailed descriptions of nodes in Figure 1 appear below the figure. Understanding the details of data cleaning is not necessary for many readers. They can skip ahead to Section 2 Analysis of Weather Patterns.


HMS_1934_2021_DropJan2022_OriginalCSV.csv
cleanHMS_1934_2022_pass1.py
HMS_1934_2022_pass1.csv
cleanHMS_1934_2022_pass2.py
HMS_1934_2021_pass2.arff
HMS_1934_2021_pass2.csv
HMSHOURLY1966-2015_DROP15Jan2020_CSV_1_49.csv
HMS1934thru2021SkyCodesAdded.csv
cleanHMS_1934_2022_pass3.py
HMS_1934_2021_pass3.arff
HMS_1934_2021_pass3.csv
cleanHMS_1934_2022_pass4.py
weatherData_2021_NOAA.csv
sunrise2022.csv sunset2022.csv
project_select_NOAA.py
project_select_NOAA_1948_2021.arff
project_select_NOAA_1948_2021.csv
merge_HMS_NOAA.py
merge_HMS_NOAA.arff
merge_HMS_NOAA.csv
timelag_HMS_NOAA.py
timelag_HMS_NOAA.arff
timelag_HMS_NOAA.csv
day_aggregate_HMS.py
day_aggregate_HMS.arff
day_aggregate_HMS.csv

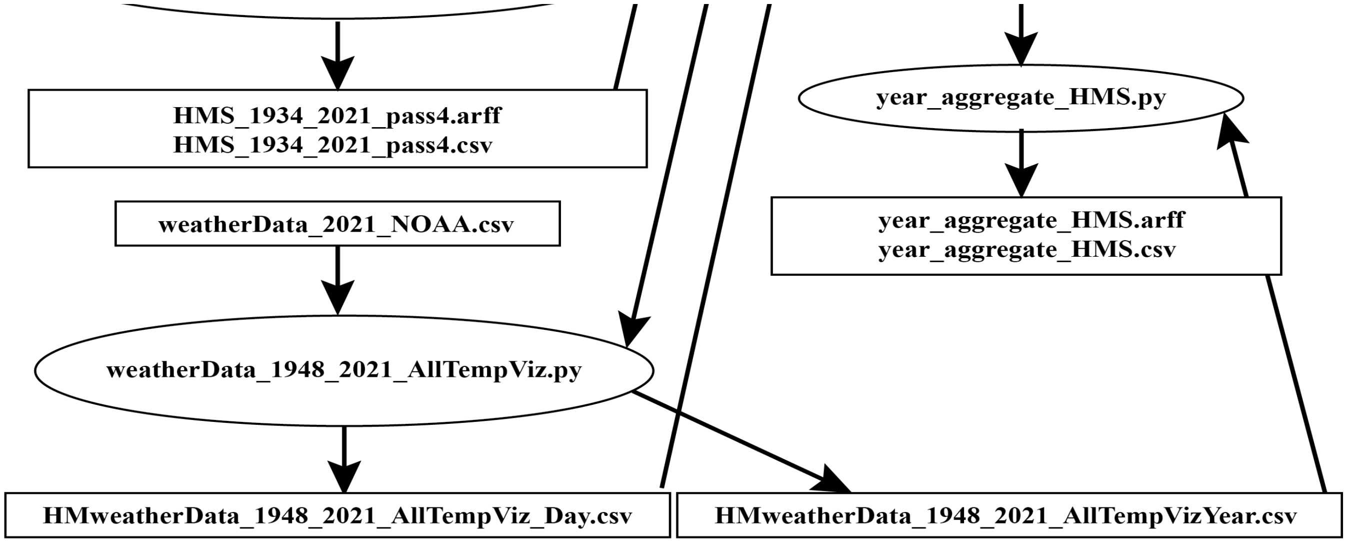

HMS_1934_2021_pass4.arff
HMS_1934_2021_pass4.csv
weatherData_2021_NOAA.csv
weatherData_1948_2021_AllTempViz.py
HMweatherData_1948_2021_AllTempViz_Day.csv
HMweatherData_1948_2021_AllTempVizYear.csv
year_aggregate_HMS.py
year_aggregate_HMS.arff
year_aggregate_HMS.csv

Figure 1: Nodes in the makefile-driven data cleaning flow. Rectangles are data stores and ellipses are custom scripts.

cleanHMS_1934_2022_pass1.py (115 non-blank lines consisting of 74 code and 41 comment lines)

```
# cleanHMS_1934_2022_pass1.py cleans data from Hawk Mtn January 2022,
# D. Parson, spring 2022. It throws out spare commentary rows at the
# top and bottom, combine the two-row headers into one, ignores blank
# rows, and writes pass1 cleaned data to a CSV file. Hawk Mtn's files
# included a few UnicodeDecodeError rows followed by a partial row,
# and this script just ignores those rows. No other rows are deleted.
# __main__ hard codes the input and output file names as
# ../Hawk Mountain Sanctuary 1934 2021 OriginalCSV.csv and
# HMS_1934_2021_pass1.csv respectively.

# Update 6/11/2022 we are skipping timezones in datetime objects because
# Weka doesn't understand them and chokes interpreting datetimes during
# invalid EDT transitions in March. Those transitions occur during
# non-raptor-observation hours but the first pass over NOAA weather data
# has those in. Also pass1 reading HMS CSV chokes on some non-UNICODE chars
# that don't affect acad or mcgonagall, so we have to patch around that.
```

cleanHMS_1934_2022_pass2.py (324 non-blank lines consisting of 193 code and 131 comment lines)

```
# cleanHMS_1934_2022_pass2.py cleans data from Hawk Mtn January 2022,
# D. Parson, spring 2022. pass2 uses the output of pass1 as input.

# Update 6/11/2022 we are skipping timezones in datetime objects because
# Weka doesn't understand them and chokes interpreting datetimes during
# invalid EDT transitions in March. Those transitions occur during
# non-raptor-observation hours but the first pass over NOAA weather data
# has those in. Also pass1 reading HMS CSV chokes on some non-UNICODE chars
# that don't affect acad or mcgonagall, so we have to patch around that.

# This pass2 does following:
# 1. When an _All column is 0 it sums the other columns for that species.
#    There is no GE_ALL in data so pass2 synthesizes one from other counts.
# 2. Makes a datetime column 0 and derives year, month, monthday, yearday,
#    hour (in day). Sort the output on datetime of measurement. It arrives
#    that way, but it is best to be sure.
#    HMS timestamps are always EST, never EDT, per Laurie Goodrich 6/9/2022.
#    NOAA weather data timestamps are UTC.
# 3. Removes Jan through July per Laurie Goodrich, keeping Autumn counts.
# 4. Tries to patch temperature C when it is > 2.5 pstdev from the mean
#    of all readings, less than -3.0 pstdev, or == 0. The 2.5 is to
#    limit below max August PA temp here (108F)
#    https://climate.met.psu.edu/data/state/staterecords.php
#    and min noted by Laurie Goodrich of -11C. These limits comes out to
#    -16C (3.2F) and 45.3C (113.54F) on this dataset. Zeroes are filtered
#    because there are a lot of consecutive zeroes that should say unrecorded
#    in the data. The patch in all three cases is to look at predecessor and
#    succssor temp on the same day. If both are non-0, use their mean, else
#    use the non-0 one. This approach patches 897/(897+11116) = 7.5% of all
#    FLAGGED temps, which is 897/56729 = 1.6% of all instances. If we start
#    computing at Aug 1976 we get 859/(859+4430) = 16.2% of all flagged
#    entries patched and 859/49427 = 1.7% of all instances. Others are flagged
#    as unknown.
# 5. If there are 0 total observers set that cell to unknown.
# 6. A duration of 0 gets set to unknown.
# 7. CloudCover up to this point is all 0.0:
#       datetime,CloudCover
#       ...
#       2006-08-18 11:00:00,0.0
#       2006-08-18 12:00:00,0.0
#       2006-08-18 13:00:00,0.0
#       2006-08-18 14:00:00,0.0
#       2006-08-18 15:00:00,0.0
#       2006-08-18 16:00:00,0.0
#       2006-08-19 06:00:00,100.0
#       2006-08-19 07:00:00,100.0
#       2006-08-19 08:00:00,100.0
#       2006-08-19 09:00:00,100.0
#       2006-08-19 10:00:00,100.0
#       2006-08-19 11:00:00,100.0
#       2006-08-19 12:00:00,100.0
#       2006-08-19 13:00:00,100.0
#       2006-08-20 06:00:00,50.0
#       2006-08-20 07:00:00,50.0
#       2006-08-20 08:00:00,100.0
#       2006-08-20 09:00:00,80.0
#       2006-08-20 10:00:00,80.0
#   Per Laurie Goodrich April 15, 2022, CloudCover should range [0,100]%,
#   so set leading zeroes and out-of-range values to unknown.
#   We wind up with 4 trailing, out-of-range audits:
#       < CLOUDCOVER AUDIT MARKING: 2006-08-18 16:00:00 value 0.0 as nan
#       47713d10348
#       < CLOUDCOVER AUDIT MARKING: 2007-10-10 14:00:00 value 127.0 as nan
#       49074d11708
#       < CLOUDCOVER AUDIT MARKING: 2017-09-26 14:00:00 value 127.0 as nan
#       49091d11724
#       < CLOUDCOVER AUDIT MARKING: 2017-11-11 16:00:00 value -3.0 as nan
#       49140d11772
#       < CLOUDCOVER AUDIT MARKING: 2018-09-23 10:00:00 value 127.0 as nan
#   Changed nan assignment to None on 7/25/2022, Weka thinks nan is valid value.

# 8. Writes both CSV and ARFF.

# Input date formats in Date and Start input columns:
# 7-Oct-1934 (year is always 4 digits; day is 1 or 2; month is 3 letters).
# 07:00 (hour and minute are two digits each).
# Output date formats previously used:
# @attribute datetime date "yyyy-MM-dd HH:mm"
# @attribute SunDate date yyyy-MM-dd
# @attribute Sunrise date HH:mm:ss
```

cleanHMS_1934_2022_pass3.py (623 non-blank lines consisting of 215 code and 408 comment lines)

```
# cleanHMS_1934_2022_pass3.py merges data from Hawk Mtn Jan2020 data drop
# into data from Jan2022 drop when it is missing from the latter.
# This pass was originally missing but investigation of missing data from
# the Jan2022 drop led to its creation, moving the original pass3 to pass4.

# Update 6/11/2022 we are skipping timezones in datetime objects because
# Weka doesn't understand them and chokes interpreting datetimes during
# invalid EDT transitions in March. Those transitions occur during
# non-raptor-observation hours but the first pass over NOAA weather data
# has those in. Also pass1 reading HMS CSV chokes on some non-UNICODE chars
# that don't affect acad or mcgonagall, so we have to patch around that.
# Finally the Python version currently running on the KU servers does not
# have the key= parameter for bisect_left so we have to kludge around that.
#
# NOTES from June 1 and 2, 2022 email exchange with Dr. Laurie Goodrich:
# I took another look at the autumn records from 1966 through 2015 in the
# Jan2020 data drop (those are the years in that data) vs autumn records
# from 1966 through 2015 in the Jan2022 data drop (1934-1965 and 2016-2021
# temporarily removed). In addition to FlightDIR and FlightHT discrepancies
# discussed below, the Jan2020 WindSpeed is missing 6% of the measures vs 49%
# missing in the Jan2022 data drop. WindDIR, in contrast, is missing 9%
# in Jan2020 and 3% in Jan2022.
#
# Should WindDIR be unknown when WindSpeed is recorded as 0?
# How about when WindSpeed is recorded as Unknown?
#
# I am going to patch Jan2020 FlightDIR and FlightHT from Jan2020 data into
# Jan2022 data when they are missing in Jan2022, noting that they will be
# Unknown when TOTAL raptors equal 0. (This latter 'noting' audit is in
# pass4 since FlightDIR and FlightHT are strings here.
# I will do the same for missing WindSpeed and WindDIR values. I can edit
# in an audit (in pass4) for the answer for the previous
# paragraph's question later.
#
# That leaves the missing SkyCode column attribute in the Jan2022 data drop.
# The Jan2020 data drop has SkyCode with only 8% missing values, but patching
# that into Jan2022 data leaves 1934-1965 and 2016-2021 with no SkyCode.
# I am just going to leave it out of the data merge for now.
#
# Update June 24, 2022 Dr. Laurie Goodrich just sent me
# HMS1934thru2021SkyCodesAdded.xlsx (HMS1934thru2021SkyCodesAdded.csv
# converted) with SkyCode in 2000-2021, and the NOAA data starts only at 1948,
# so patching in HMS1934thru2021SkyCodesAdded.csv in this script leaves us
# with missing SkyCode values for 1948-1965, so I am now including SkyCode.
# We get 56 years with SkyCode.
```

cleanHMS_1934_2022_pass4.py (490 non-blank lines consisting of 326 code and 164 comment lines)

```
# cleanHMS_1934_2022_pass4.py cleans data from Hawk Mtn January 2022,
# D. Parson, spring 2022. pass4 uses the output of pass3 as input.
# This pass4 does its steps in discrete functions in order to serve as
# the basis for CSC458 Assignment 1 in Fall 2022. D. Parson.
# It reads an ARFF file using arfflib_3_2.readARFF in order to preserve
# @attribute type information and writes both ARFF and CSV.
# It changes wind & flight directions to histograms for each of 16 directions,
# also to compass degrees, fixes letter transposes in direction data.
# It pulls leading numbers out of visibility & flight height fields that have
# trailing alphas. Data entry errors!

# Update 6/11/2022 we are skipping timezones in datetime objects because
# Weka doesn't understand them and chokes interpreting datetimes during
# invalid EDT transitions in March. Those transitions occur during
# non-raptor-observation hours but the first pass over NOAA weather data
# has those in. Also pass1 reading HMS CSV chokes on some non-UNICODE chars
# that don't affect acad or mcgonagall, so we have to patch around that.
# Update 6/17/2022 add single numeric attribute WindDegrees for converting
# Winddir to compass degrees.
```

Flight and wind direction processing store the cardinal compass points centered at the nearest 22.5 degree interval as a number, and also store directions as histograms of each of the 16 cardinal directions N through NNW. It was necessary to convert user-entered data such as the string "NE" into numeric data for analysis. Also, some records such as "NNW" were actually recorded by observers as "WNN", etc., so pass4 permutes all such text data and selects the closest canonical string for conversion to a centered compass degree and a tally in a histogram of directions. Aggregation of the 22.5 degree values into days and years uses the statistical mode to record the predominant value. Computing the mean would converge towards 180 degrees which would not be useful.

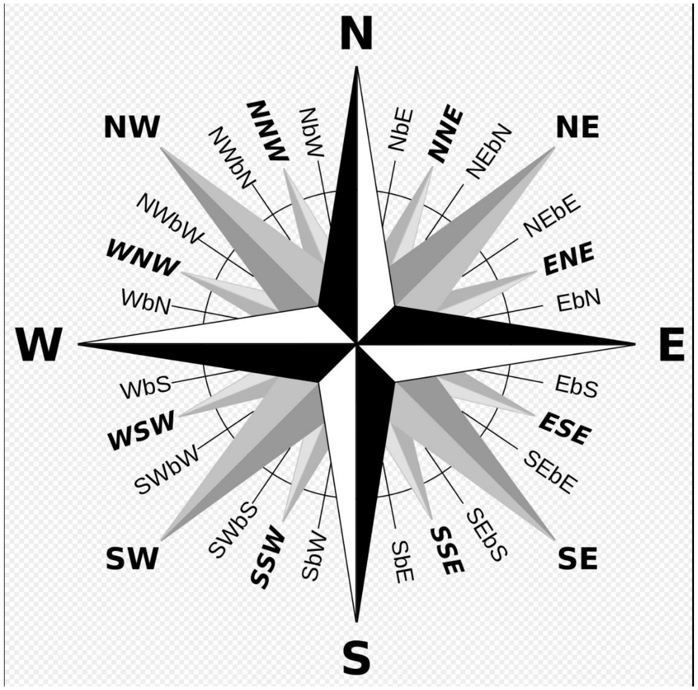


Figure 2: Illustration from https://en.wikipedia.org/wiki/Points_of_the_compass

merge_HMS_NOAA.py (666 non-blank lines consisting of 484 code and 180 comment lines)

```
__xlationTable__ = (
    # How to merge HMS and NOAA attributes into the output ARFF & CSV.
    # Each entry is [DB, attrName, howToAggregate, inputColumn, outputColumn]
    #   where DB is "HMS" or "NOAA", howToAggregate applies mainly to
    #   collapsing multiple NOAA rows into 1 HMS row when the latter is
    #   long duration, especially only 1 measure in the day in early years,
    #   with "copy" meaning copy the first one, "mean" meaning average
    #   them, "tally" for NOAA directions meaning add the 1's, "mode"
    #   for NOAA HourlyWindDirection to get the prevailing wind, or a
    #   function reference to a function that reduces a list of values
    #   from a column to a single, scalar value.
    ["HMS", "datetime", "copy", 0, 0, ('date',
            "'yyyy-MM-dd HH:mm:ss'", '%Y-%m-%d %H:%M:%S')],
    ...
```

timelag_HMS_NOAA.py (168 non-blank lines consisting of 115 code and 53 comment lines)

```
#* Author:            Dr. Parson
#* Creation Date:     June 26, 2022
#* Professor Name:    Dr. Parson
#* Filename:          timelag_HMS_NOAA.py
#* Purpose:           Create time lags of some attributes of
#*                    merge_HMS_NOAA.arff created by
#*                    merge_HMS_NOAA.py. The reason is that weather
#*                    *changes* can affect raptor behavior.
#*                    There are 1-, 2-, and 3-day lags of
#*                    HMtempC, Visibility and CloudCover from HM,
#*                    HourlyDryBulbTemperature,
#*                    HourlyWetBulbTemperature,
#*                    HourlyDewPointTemperature,
#*                    HourlyStationPressure,
#*                    HourlyRelativeHumidity, and
#*                    HourlyVisibility from Allentown Airport NOAA data.
#*                    The added attributes are deltas + or - from
#*                    previous readings, None when not both are available:
#*                    HMtempC_1, HMtempC_2, and HMtempC_3 for 1-,
#*                    2-, and 3-day lags, similar for other 2 HM and
#*                    6 NOAA attributes listed above, placed in front of
#*                    the raptor count attributes. Output is to
#*                    timelag_HMS_NOAA.arff and timelag_HMS_NOAA.csv.
```

weatherData_1948_2021_AllTempViz.py (156 non-blank lines consisting of 139 code and 17 comment lines)

```
# D. Parson 7/11/2022 weatherData_1948_2021_AllTempViz.py
# Compute stats on ALL weatherData_2021_NOAA.csv HourlyDryBulbTemperature
# Output as space-separated fields to stdout.
# UPDATE 7/13/2022:
#   1. Aggregate per-day for ALL data in weatherData_2021_NOAA.csv,
#      in addition to the Aug-Dec days of HM observations.
#   2. Add HourlyVisibility fields as
#   3. Write output as CSV files to weatherData_1948_2021_AllTempViz_Day.csv
#      and weatherData_1948_2021_AllTempViz_Year.csv for all hours in NOAA
#      data, and to HM_weatherData_1948_2021_AllTempViz_Day.csv
#      for HM observation days.
```

day_aggregate_HMS.py (579 non-blank lines consisting of 482 code and 97 comment lines)

```
#* Author:            Dr. Parson
#* Creation Date:     July 2, 2022
#* Professor Name:    Dr. Parson
#* Filename:          day_aggregate_HMS.py
#* Purpose:           Aggregate multiple observation rows into 1-per-day row.
#*                    Input HMS and NOAA already merged by merge_HMS_NOAA.py
#*                    into merge_HMS_NOAA.arff with select temperature,
#*              humidity, and barometric pressures time-lagged across
#*              24, 48, and 72 hours into deltas (differences)
#*              by timelag_HMS_NOAA.py into timelag_HMS_NOAA.arff,
#*              which is input to this script.
#*              Table DailyAggregateList at the bottom of this
#*              file gives the output attributes and their types of
#*              aggregation provided by these helper functions.
#*              Helpers do not include missing None values in
#*              their statistical analysis.
#*      Parameters for the following aggregation functions are as follows:
...
```

year_aggregate.py (1686 non-blank lines consisting of 1558 code and 128 comment lines)

```
#* Author:          Dr. Parson
#* Creation Date:   July 3, 2022
#* Professor Name:  Dr. Parson
#* Filename:        year_aggregate_HMS.py
#* Purpose:         Aggregate multiple day-rows into one-per-day year.
#*              Input taken from day_aggregate_HMS.py's
#*              day_aggregate_HMS.arff that aggregated one or more
#*              observations per day into per-day instances;
#*              current script aggregates those into per-year
#*              instances.
#*              Table YearlyAggregateList at the bottom of this
#*              file gives the output attributes and their types of
#*              aggregation provided by these helper functions.
#*              Helpers do not include missing None values in
#*              their statistical analysis.
#*      Parameters for the following aggregation functions are as follows:
...
# UPDATE 7/13/2022 we need to add entire-day (not just HM observation times)
# and entire-year (aggregation of entire-days for year) by integrating
# all hours for NOAA days from weatherData_2021_NOAA.csv into HM days in
# day_aggregate_HMS.py -> day_aggregate_HMS -> year_aggregate_HMS(.arff|.csv).
# UPDATE 7/14/2022 The above add just uses NOAA data for the HM observation
# days. We are keeping that but also adding the FULL annual NOAA records
# in weatherData_1948_2021_AllTempViz_Year.csv because mapping
# year -> dryTempC24mean in that file pulls in many more NOAA records
# than just the NOAA records for August thru December HM-observation days
# in day_aggregate_HMS.arff. Also putting out a second pair of output
# .csv|.arff files for years 1976 thru 2021 similar to day_aggregate_HMS.py.
# UPDATE 7/9/2022 find the mean, median, pstdv, min, and max for non-0
# raptor counts over the year in addition to tallies, in order to see
# changes in distributions.
# UPDATE 8/6/2022 add _1st thru _peak variations for HM wind measures to
# see if timing of wind directions (in addition to intensity) has changed.
```

Below is a summary of the 199 daily numeric attributes for August through December created by day_aggregate_HMS.py

Hawk Mountain North Lookout data were recorded only once (one row in the spreadsheet) per observation day from 1934 through early September 1966, settling into roughly hourly observations in 1967. Temperature Celsius (HMtempC below) was not recorded regularly until August 14, 1969, and not at all in earlier years. There is a gap in both HM and NOAAA Allentown Airport data from 1970 through 1972. Most of this analysis runs 1976 through 2021.

Date / time attributes:
datetime 'yyyy-MM-dd HH:mm:ss' (Date & time of start of Hawk Mtn. observation)
ESTdatetime 'yyyy-MM-dd HH:mm:ss' (Date & time of NOAA Allentown record)
year yearSince1976 month monthday yearday daySinceAug1
duration (sum of observation duration in minutes)

Climate attributes as measured at North Lookout August through December.
Each is the mean of that measure for the recorded day.
SkyCode (see HMANA_WindspdAltitudeSkycode.pdf)
HMtempC
WindSpd (converted to km/hour from HMANA_WindspdAltitudeSkycode.pdf))
Visibility
CloudCover
WindDegrees (mode for the day, rounded to the closest 22.5 degree point)

Histograms (daily tallies) of WindDegrees:
wndN wndNNE wndNE wndENE wndE wndESE wndSE wndSSE
wndS wndSSW wndSW wndWSW wndW wndWNW wndNW wndNNW wndUNK

NOAA data from Allentown Airport were recorded once every hour, on the hour, from 1948 through 1964. In 1965 recordings changed to once every three hours. There is no NOAA data for 1970 through 1972, and in 1973 they were recorded at least once every hour, sometimes more frequently. Unless otherwise documented, each NOAA reading pairs, 1-to-1, with the Hawk Mountain observation period (row in the HM data) that is closest in time on the same day. Each is the mean of that measure for the recorded day.

YEAR MONTH DAY
HourlyDryBulbTemperature
HourlyWetBulbTemperature
HourlyDewPointTemperature
HourlyWindSpeed
HourlyPrecipitation
HourlyStationPressure
HourlyRelativeHumidity
HourlyVisibility
HourlyWindDirection

Histograms (daily tallies) of HourlyWindDirection for matching HM observation times:
noaawdN noaawdNNE noaawdNE noaawdENE noaawdE noaawdESE noaawdSE noaawdSSE
noaawdS noaawdSSW noaawdSW noaawdWSW noaawdW noaawdWNW noaawdNW noaawdNNW noaawdUNK

Hour & minute of sunrise & sunset contain matching closest NOAA record to its HM counterpart.
sunriseH sunriseM sunsetH sunsetM

The following give the amount of change, positive or negative, from the same measurement, 24, 48, and 72 hours previously.

HMtempC_24 Visibility_24 CloudCover_24 HourlyDryBulbTemperature_24
HourlyWetBulbTemperature_24 HourlyDewPointTemperature_24
HourlyStationPressure_24 HourlyRelativeHumidity_24 HourlyVisibility_24
HMtempC_48 Visibility_48 CloudCover_48 HourlyDryBulbTemperature_48
HourlyWetBulbTemperature_48 HourlyDewPointTemperature_48
HourlyStationPressure_48 HourlyRelativeHumidity_48 HourlyVisibility_48
HMtempC_72 Visibility_72 CloudCover_72 HourlyDryBulbTemperature_72
HourlyWetBulbTemperature_72 HourlyDewPointTemperature_72
HourlyStationPressure_72 HourlyRelativeHumidity_72 HourlyVisibility_72

Number of HM records and HMtempC recordings for this day.
RecordCount
HMtempCCount

The "dryTemp24" recordings are all for all NOAA Allentown HourlyDryBulbTemperature recordings for the day, i.e., across all 24 hours. The timestamps are just for data verification. The mean through max are HourlyDryBulbTemperature measures. The dryViz24 is the mean of HourlyVisibility across all 24 hours, although not being used because of an apparent data error in HourlyVisibility.
dryTempC24year dryTempC24month dryTempC24day dryTempC24count
dryTempC24mean dryTempC24median dryTempC24pstdev dryTempC24min dryTempC24max
dryViz24

North Lookout Raptor Measurements
FlightDegrees (rounded to the closest 22.5 degree point)
Histograms (tallies) of WindDegrees:
fltN fltNNE fltNE fltENE fltE fltESE fltSE fltSSE
fltS fltSSW fltSW fltWSW fltW fltWNW fltNW fltNNW fltUNK
FlightHT (see HMANA_WindspdAltitudeSkycode.pdf)
Raptor Counts, where _All is the sum of the subcategories.
These measures are daily sums, not means.
AK_All AK_Female AK_Male AK_Unk
BE_Adult BE_All BE_Imm BE_Unk
BV_All
BW_Adult BW_All BW_Imm BW_Unk
CH_Adult CH_All CH_Imm CH_Unk
GE_All GE_Adult GE_Imm GE_Subadult GE_Unk
GY_All
ML_All ML_Fem_Imm ML_Male ML_Unk
MK_All
NG_Adult NG_All NG_Imm NG_Unk
NH_All NH_Brown NH_Female NH_Fem_Imm NH_Imm NH_Male NH_Unk
OS_Adult OS_All OS_fish OS_foraging OS_Imm OS_nofish OS_Unk
PG_Adult PG_All PG_Imm PG_Unk
RL_All RL_Dark RL_Light RL_Unk
RS_Adult RS_All RS_Imm RS_Unk
RT_Adult RT_All RT_Imm RT_Unk
SS_Adult SS_All SS_Imm SS_Unk

SW_All
TV_Adult TV_All TV_Imm TV_Unk
UA_All
UB_All
UE_All
UF_All
TOTAL (sum of raptors for each day)

Below is a summary of the 1191 annual numeric attributes, unless otherwise noted for August through December, created by year_aggregate_HMS.py

Take each of HMtempC, Visibility, CloudCover, SkyCode, FlightHT, WindSpd, HourlyDryBulbTemperature, HourlyWetBulbTemperature, HourlyDewPointTemperature, HourlyWindSpeed, HourlyPrecipitation, HourlyStationPressure, HourlyRelativeHumidity, HourlyVisibility from the above daily records and and compute the mean, median, population standard deviation, min, & max for the observation autumn.
Take dryTempC24mean, dryTempC24median, dryTempC24pstdv, dryTempC24min, dryTempC24max and compute the stats for the observation period. **Also compute for the entire year January - December, 24-hours**, dryTempC24meanFullNOAA, dryTempC24medianFullNOAA, dryTempC24pstdevFullNOAA, dryTempC24minFullNOAA, and dryTempC24maxFullNOAA, using all NOAA Allentown Airport measures for each year.

Wind direction, flight direction, and NOAA wind direction degree measures are the **statistical modes** -- most frequently occurring daily values -- for the year. No other average makes sense for cyclic degrees.

Wind direction, flight direction, and NOAA wind direction histograms -- daily tallies for N, NNE, ... NNW -- are **sums** (tallies) for each year.

For all daily raptor counts, compute the **sum** (tally), **mean**, **median**, **population standard deviation**, **min**, and **max** for the daily records.

For all non-zero daily raptor counts, compute the **day-of-year** of the **1st** sighting, the day of the **25th percentile** for the year arrived, the days of the **50th and 75th percentiles**, the **last** sighting, and the **day of the peak sighting** for that raptor.

These are a lot of attributes for the machine learning (ML) algorithms to sift through, but with only 46 years from 1976 through 2021, model derivation runs within a few hours.

1c. Data Storage Formats

The file name extensions in Figure 1 above are ".arff" and ".csv". The former is the Attribute-Relation File Format of the Weka toolkit [8] used in the summer 2022 analyses of these data. ARFF files [9] include numeric and non-numeric type declarations such as datetime for the attributes listed in the previous section. Otherwise they are identical to comma-separated-value (CSV) files in using a row of comma-separated-values for each record of attributes such as a daily or yearly aggregate. My Python libraries include a custom library for manipulating ARFF data structures and translating them from and to CSV formats. Providing CSV files accommodates using other tools such as Microsoft Excel and the widely used Python CSV library.

2. ANALYSIS OF WEATHER PATTERNS

2A. Building and Running the Model Learners

Readers new to data science need to be aware that there is typically a trade-off between accuracy and intelligibility in machine learning (ML) models. On June 27, 2016, Dr. Oskars Rieksts of Kutztown University (now retired) and I met with Dr. Roland Dunbrack of the Dunbrack Lab of the Fox Chase Cancer Center in Philadelphia to discuss biomedical applications of data science [10, 11]. While most of the discussion of mapping genetic structures to probabilities of types of cancers was over my head, when the discussion turned to ML models, I was at home. One of the modeling techniques used in their lab was **RandomForest**, which can produce extremely accurate but effectively undecipherable models. Clearly, in biomedical applications, accuracy can be life-saving. However, for the present study, intelligibility with acceptable accuracy is the approach used in ML model selection discussed next. I did experiment with using **RandomTree**, which often gave accurate results with this dataset, but RandomTree constructed huge decision trees that were essentially point-by-point memorization of the current data. In data science jargon they were **over-fit** to the dataset used as model training data. The modeling techniques discussed in this section show trends over time in an intelligible way.

This first part of analysis has been straightforward. Python script **year_date2weather.py** reads file **year_aggregate_HMS_1976_2021_kupapcsit01.arff**, which is **year_aggregate_HMS.arff** of Figure 1 filtered to 1976 through 2021 as extracted on Kutztown Linux server kupapcsit01. The year filtering is to eliminate early years with missing Hawk Mountain temperature and other missing data. This script then deletes attributes except the year and one of the climate attributes listed in Section 1b above, and invokes Weka [8] in command line mode to extract a machine learning model and test it. Year serves as the control variable and the single climate attribute as the experimental variable. The ML Regressors that build variants of linear regression formulas are the following.

LinearRegression maps year to the climate attribute. When there are multiple control variables, LinearRegression selects the most predictive ones to use in model construction. Since **year_date2weather.py** uses only year as the singleton control variable, LinearRegression is similar to SimpleLinearRegression. Discussion below examines some linear regression formulas derived from this dataset.

SimpleLinearRegression uses only one control variable at a time, in this case year.

M5P builds a model tree, which is a decision tree that makes decision about what linear regression formulas to use based on dominant attribute values in the decision tree. It is useful when correlations are collections of linear relationships rather than a single linear relationship of attributes. This script uses two variants of M5P, one of which constrains the size of the decision tree in the interest of intelligibility.

REPTree is a simple decision tree builder, and **DecisionTable** extracts a tables that maps control variables to experimental variable values. **SMOreg** derives non-linear data partitions called *vectors* to separate data subsets with linear relationships. **RandomTree** builds a complex and over-fit decision tree. This script employs them because they run quickly and may uncover correlations not found by variants of linear regression. There are 46 records in the yearly aggregations, one for each year 1976 through 2021, and 1191 attributes as previously discussed.

Script **year_date2weather.py** iterates over mapping yearSince1976 to one ClimateAttribute at a time, and applies 8 ML model learners (2 variants of M5P), invoking multiple Weka processes in parallel on a multiprocessor server.

Python script **day_date2weather.py** reads file **day_aggregate_HMS_1976_2021_kupapcsit01.arff**, which is **day_aggregate_HMS.arff** of Figure 1 filtered to 1976 through 2021 as extracted on Kutztown Linux server kupapcsit01. This script maps the ordered pair (yearSince1976, daySinceAug1) -> ClimateAttribute in a manner identical to yearly analysis. August 1 through December 31 would yield (153 days X 46 years) = 7,038 records; there are in fact 5,642 records of the 199 attributes previously discussed for daily aggregation because not every day in August through December had observations. Daily climate analysis uses the same 8 ML modelers used for yearly analysis discussed above.

2B. Analysis of Yearly and Daily Models for Observation Days

Correlation Coefficient (CC) is a measure of how well a ML model's prediction match actual values. "A correlation coefficient of zero indicates that no linear relationship exists between two continuous variables, and a correlation coefficient of −1 or +1 indicates a perfect linear relationship." [12] The present discussion focuses on dissecting models with high absolute CCs relative to other models for these data, skipping detailed discussion of CC mathematics. The scripts of the previous section extract statistical accuracy measures for their model executions and store them in CSV files for visual examination. There is a very large amount of model data extracted when running these models. In each analysis case it has been necessary to write additional Python code to summarize model structures and accuracy.

Models relating yearSince1976 and (yearSince1976, daySinceAug1) to climate aspects find that temperature correlations top out the CC list.

	Climate Attribute	Model	CC
1	dryTempC24mean	REPTree	0.8924

2	dryTempC24mean	RandomTree	0.889
3	HourlyDryBulbTemperature	REPTree	0.8844
4	dryTempC24mean	M5P	0.8818
5	dryTempC24median	REPTree	0.8811
6	dryTempC24mean	m5pNode500	0.88
7	HourlyDryBulbTemperature	RandomTree	0.879
8	dryTempC24mean	LinearRegression	0.8777
9	dryTempC24mean	SMOreg	0.8777
10	dryTempC24mean	DecisionTable	0.8766
11	HMtempC	REPTree	0.8765
34	HMtempC	LinearRegression	0.8639

Table 1: Leading Contenders for (yearSince1976, daySinceAug1) to Climate Modeling

CCs in the range 0.86 to 0.89 are very good. Here is a leading subset of the 979 tree nodes in that top row's model:

```
REPTree
============

daySinceAug1 < 71.5
|   daySinceAug1 < 45.5
|   |   daySinceAug1 < 37.5
|   |   |   yearSince1976 < 37.5
|   |   |   |   daySinceAug1 < 17.5
|   |   |   |   |   yearSince1976 < 36.5
|   |   |   |   |   |   yearSince1976 < 15.5
|   |   |   |   |   |   |   daySinceAug1 < 9.5
|   |   |   |   |   |   |   |   yearSince1976 < 8.5
|   |   |   |   |   |   |   |   |   daySinceAug1 < 8.5 : 25.39 (12/2.78) [4/0.85]
|   |   |   |   |   |   |   |   |   daySinceAug1 >= 8.5 : 23.25 (2/0.04) [1/0.33]
...
```

Such model trees are too detailed for uncovering general trends and they are over-fit to the specific data details. We want to start by looking at HM North Lookout data with intelligible models and high CCs. Model 34 fits the bill.

MODEL 1

```
Linear Regression Model
HMtempC =
      0.0198 * yearSince1976 +
     -0.2033 * daySinceAug1 +
     27.5322
Correlation coefficient                  0.8639
Mean absolute error                      3.3556
Root mean squared error                    4.2343
```

Those error measures are in domain units (degrees C in this model), with "Root mean squared error" emphasizing outliers. We will concentrate on CC for now.

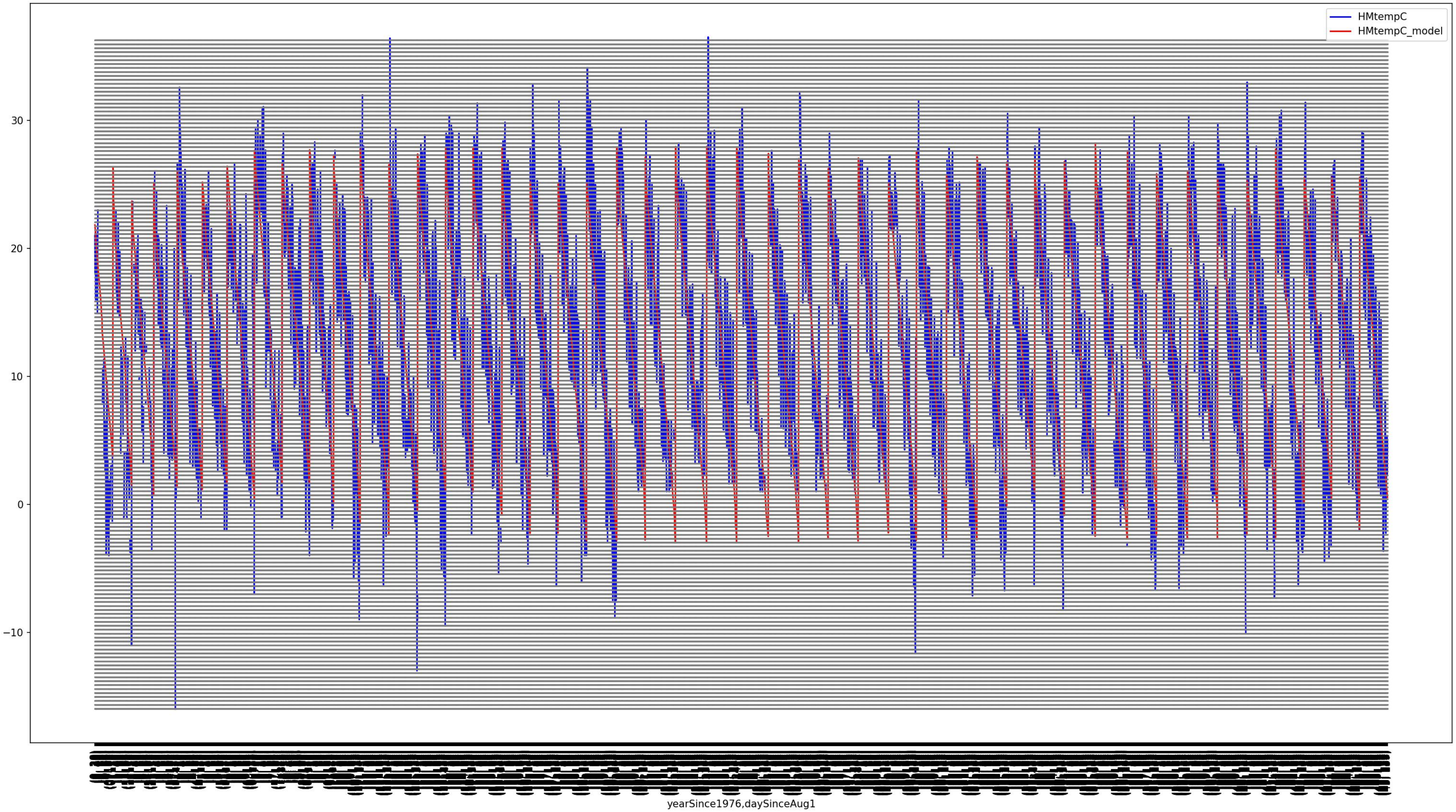


Figure 3: and for HMtempC HMtempC_model = 0.0198 * yearSince1976 + -0.2034 * daySinceAug1 + 27.5346 1976 thru 2021

Zooming in on the first year of 1976 and the second last year of 2020 makes the above graph intelligible.

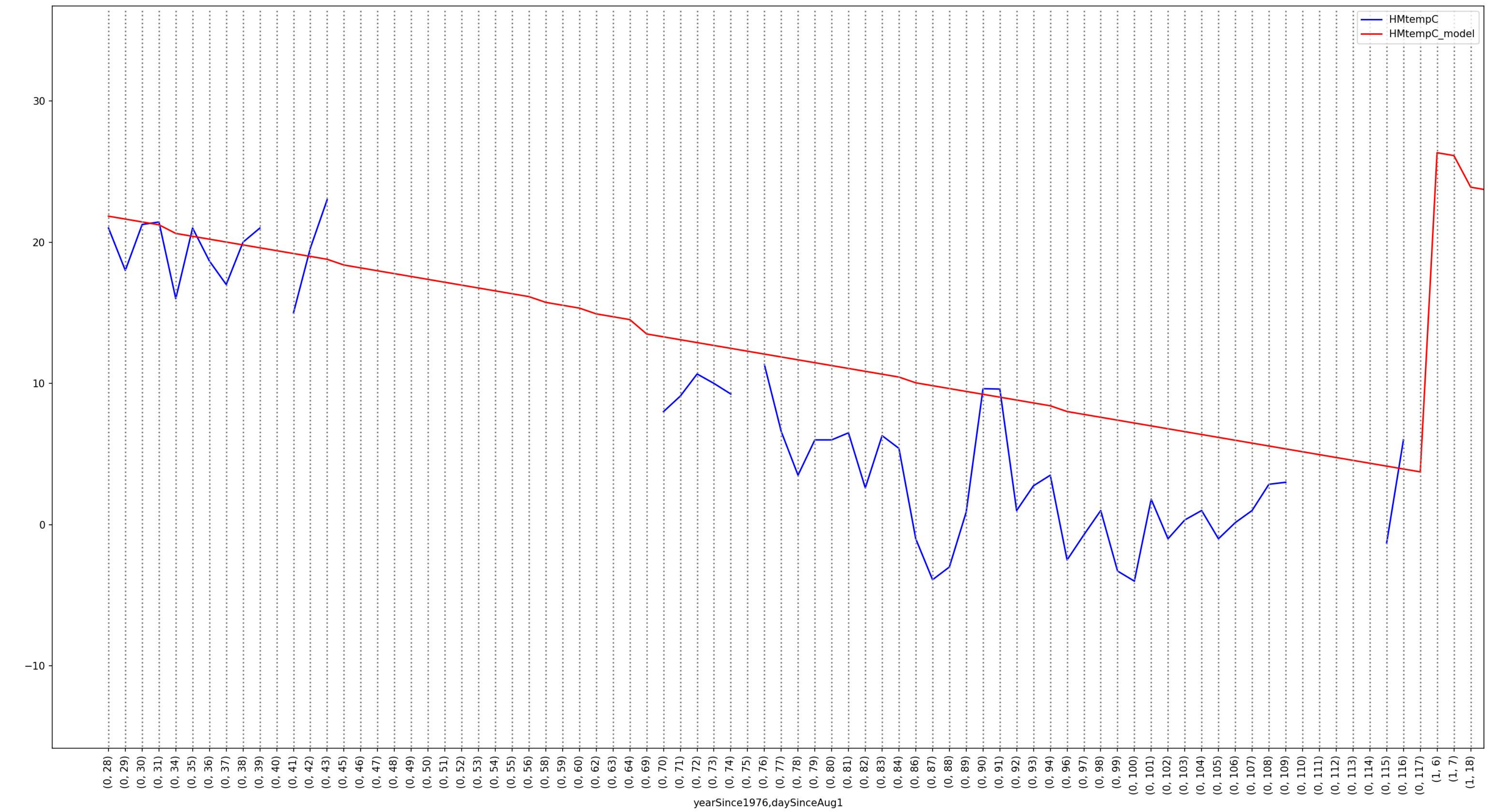


Figure 4: HMtempC and HMtempC_model = 0.0198 * yearSince1976 + -0.2034 * daySinceAug1 + 27.5346 for 1976

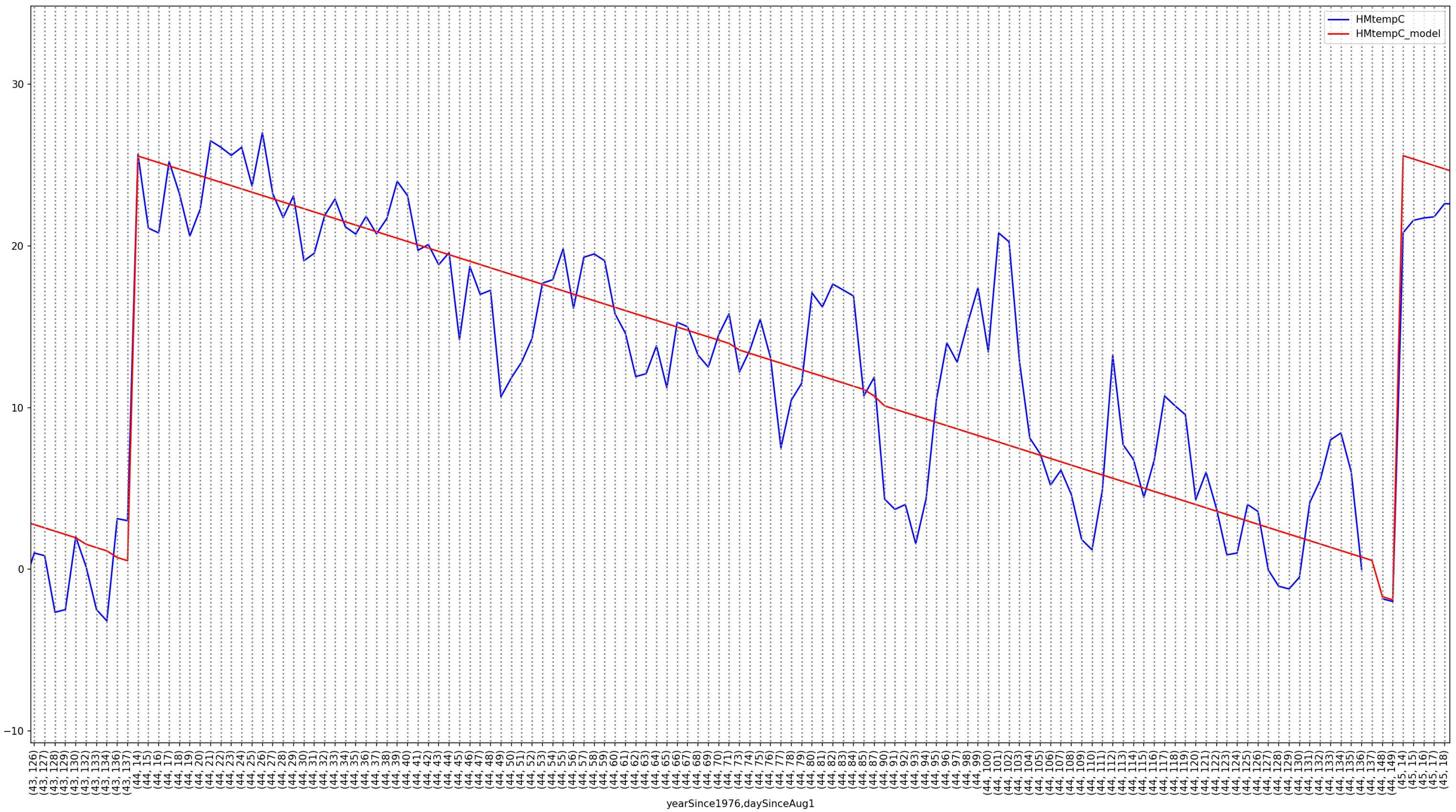


Figure 5: HMtempC and HMtempC_model = 0.0198 * yearSince1976 + -0.2034 * daySinceAug1 + 27.5346 for 2020

The sawtooth waveforms of Figures 4 and 5 illustrate the annual fall in average daily recorded temperature from the first HM observation in August through the last one in December. The blue line shows the actual HMtempC and the red waveform shows the HMtempC_model approximation that gives the CC of 0.8639 for the 34th model of Table 1. The number that establishes consistent year-to-year warming for North Lookout is the 0.0198 * yearSince1976 term. Based on daytime HM observation data, the temperature has been rising almost 0.02 C each year, 0.92 C across 46 years.

Attribute **dryTempC24mean** is the NOAA dry bulb temperature recorded for Allentown Airport for each full day including nighttime during days of HM observations August through December. Here is the LinearRegression model for dryTemp24Cmean as a function of (yearSince1976, daySinceAug1), with a CC of 0.8777 that is slightly better than the CC of 0.8639 of HMtempC's model.

MODEL 2

```
Linear Regression Model
dryTempC24mean =
     0.0317 * yearSince1976 +
    -0.1886 * daySinceAug1 +
    25.6809
Correlation coefficient                 0.8777
Mean absolute error                     3.067
Root mean squared error                   3.8482
```

The annual ramp of 0.0317 C per year X 46 years yields a 1.4582 increase over 46 years at the Allentown Airport. Inclusion of nighttime recordings and increases in temperature due to pollution are factors. The Lehigh Valley suffers from a *Heat Island Effect*.

"Heat islands are urbanized areas that experience higher temperatures than outlying areas. Structures such as buildings, roads, and other infrastructure absorb and re-emit the sun's heat more than natural landscapes such as forests and water bodies. Urban areas, where these structures are highly concentrated and greenery is limited, become "islands" of higher temperatures relative to outlying areas. Daytime temperatures in urban areas are about 1–7°F higher than temperatures in outlying areas and nighttime temperatures are about 2-5°F higher." [13] The Lehigh Valley experiences substantial pollution, possibly due to the large amount of diesel traffic. "Lehigh University professor Ben Felzer, a climate and biogeochemical modeler, believes public focus should be as much on health as on warming. ... The trapping of pollutants is enhanced in the Lehigh Valley due to inversions caused by the topography that prevent air from adequately mixing. So, understanding the local meteorology is crucial," Felzer said. "The Lehigh Valley sits between Blue Mountain to the north and South Mountain to the south, creating a dip in elevation that allows air to linger." [14]

Another model worth examination in this daily HMtempC discussion is the M5P model tree with configure parameters constrained so that a minimum of 500 records must be applied to each linear regression formula in the decision tree. Allowing fewer that 500 increases accuracy at the cost of intelligibility and likely over-fitting to the data.

MODEL 3

```
M5 pruned model tree:
daySinceAug1 <= 69.5 :
|   daySinceAug1 <= 44.5 : LM1 (1336/43.923%)
|   daySinceAug1 >  44.5 : LM2 (1091/46.866%)
daySinceAug1 >  69.5 :
|   daySinceAug1 <= 101.5 : LM3 (1372/51.929%)
|   daySinceAug1 >  101.5 : LM4 (1419/55.291%)
LM num: 1
HMtempC =
    0.0003 * yearSince1976
    - 0.1335 * daySinceAug1
    + 26.2001
LM num: 2
HMtempC =
    0.0003 * yearSince1976
    - 0.212 * daySinceAug1
    + 28.6385
LM num: 3
HMtempC =
    0.0405 * yearSince1976
    - 0.191 * daySinceAug1
    + 25.6877
LM num: 4
HMtempC =
    0.0003 * yearSince1976
    - 0.1511 * daySinceAug1
    + 21.7777
Number of Rules : 4
Correlation coefficient                  0.8647
Mean absolute error                      3.3489
Root mean squared error                    4.2236
```

This model shows a increase of 0.0405 C per year from daySinceAug1 70 through 101, i.e., October 10 through November 10, which is the main HM observation period, with increases of 0.0003 C per year during the remaining days and higher constant coefficients during the earlier days of the observation periods. Note the 4 regions for distinct linear expressions fitted to the HMtempC values for 2020 in Figure 6.

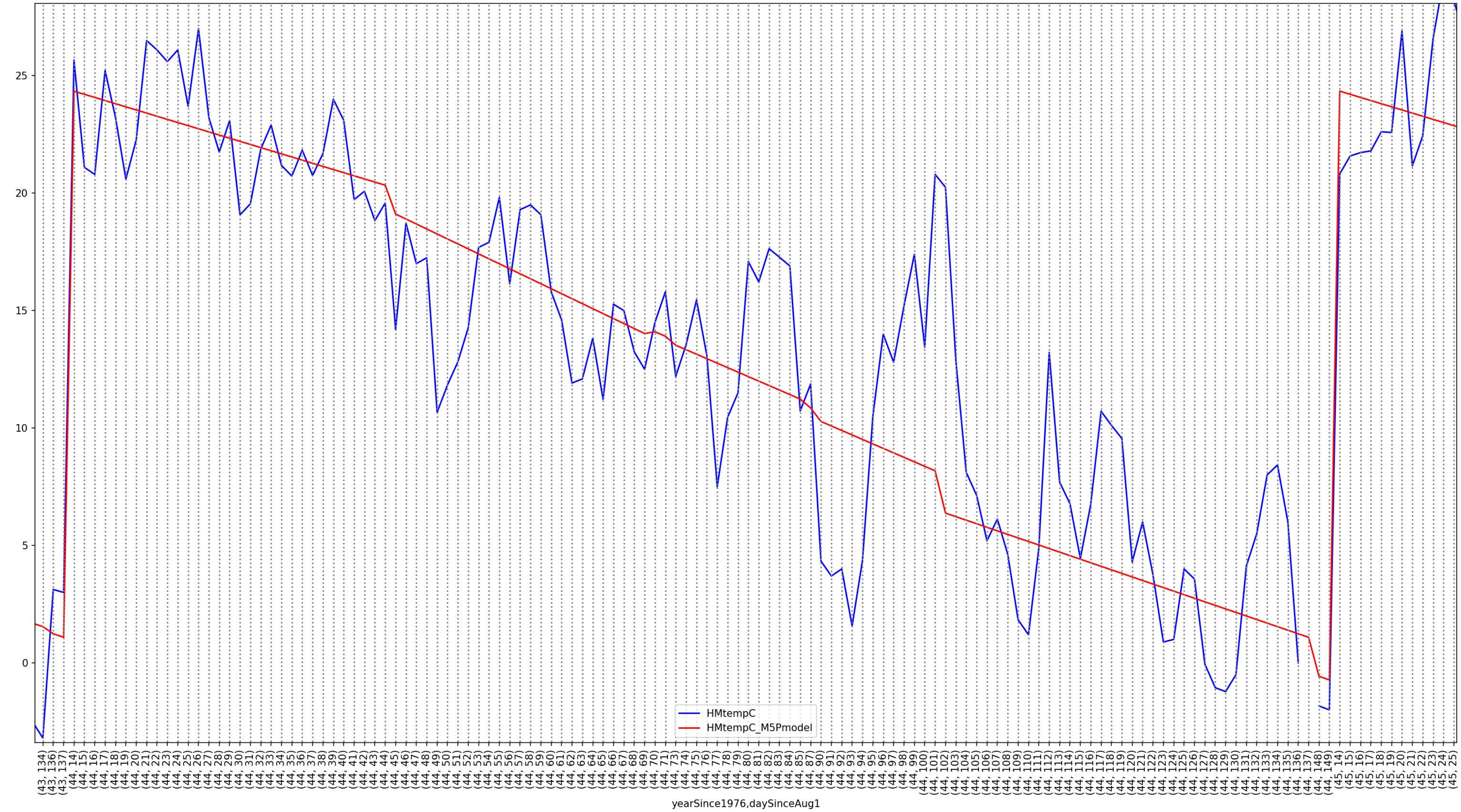


Figure 6: HMtempC and M5P model tree given above for 2020 HMtempC_M5Pmodel

These analyses of day-by-day temperature changes for the August - December observation period beg the question of what happens when we use the 46-record yearly aggregated data in year_aggregate_HMS.arff for 1976 through 2021. Here are the CCs for mapping yearSince1976 to 4 temperature attributes taken one at a time.

Temp Attribute	CC
HMtempC_mean of daily records Aug-Dec	-0.3251
HourlyDryBulbTemperature_mean (NOAA) of daily records Aug-Dec	-0.275
dryTempC24mean of all NOAA records (24 hours) for observation days	-0.3408
dryTempC24meanFullNOAA all NOAA dry bulb records for entire year	0.5216

Table 2: yearSince1976 to temperature correlation

To review, HMtempC_mean is the mean of all HM recorded HMtempC readings for the Aug-Dec observation period, HourlyDryBulbTemperature_mean is the mean of all NOAA temperatures recorded in a 1-to-1 correspondence with HM observation times, dryTempC24mean is the mean of all NOAA temperatures (24 hours) recorded only on HM Aug-Dec observation days, and dryTempC24meanFullNOAA is the mean of all NOAA dry bulb temperatures for every day and every hourly recording throughout the year.

The full year of dryTempC24meanFullNOAA measurements contributes to its CC, which at 0.5216 is respectable but much lower than the per-day CCs.

MODEL 4

```
Linear Regression Model
dryTempC24meanFullNOAA =
     0.0319 * yearSince1976 +
    10.6484
Correlation coefficient                  0.5216
Mean absolute error                      0.5193
Root mean squared error                     0.6379
```

I experimented with exponential formulas using yearSince1976 raised to some constant power, and polynomial formulas using some constant base raised to yearSince1976 power. I was able to get a marginal gain in CC and marginal losses with other error measures by raising 1.009 to the yearSince1976 power, but the changes were too small and the curve headed into an extreme upper area to justify its use. Figure 7 shows the above year-by-year Linear Regression Model in red plotted against the NOAA all-year temperature means in blue.

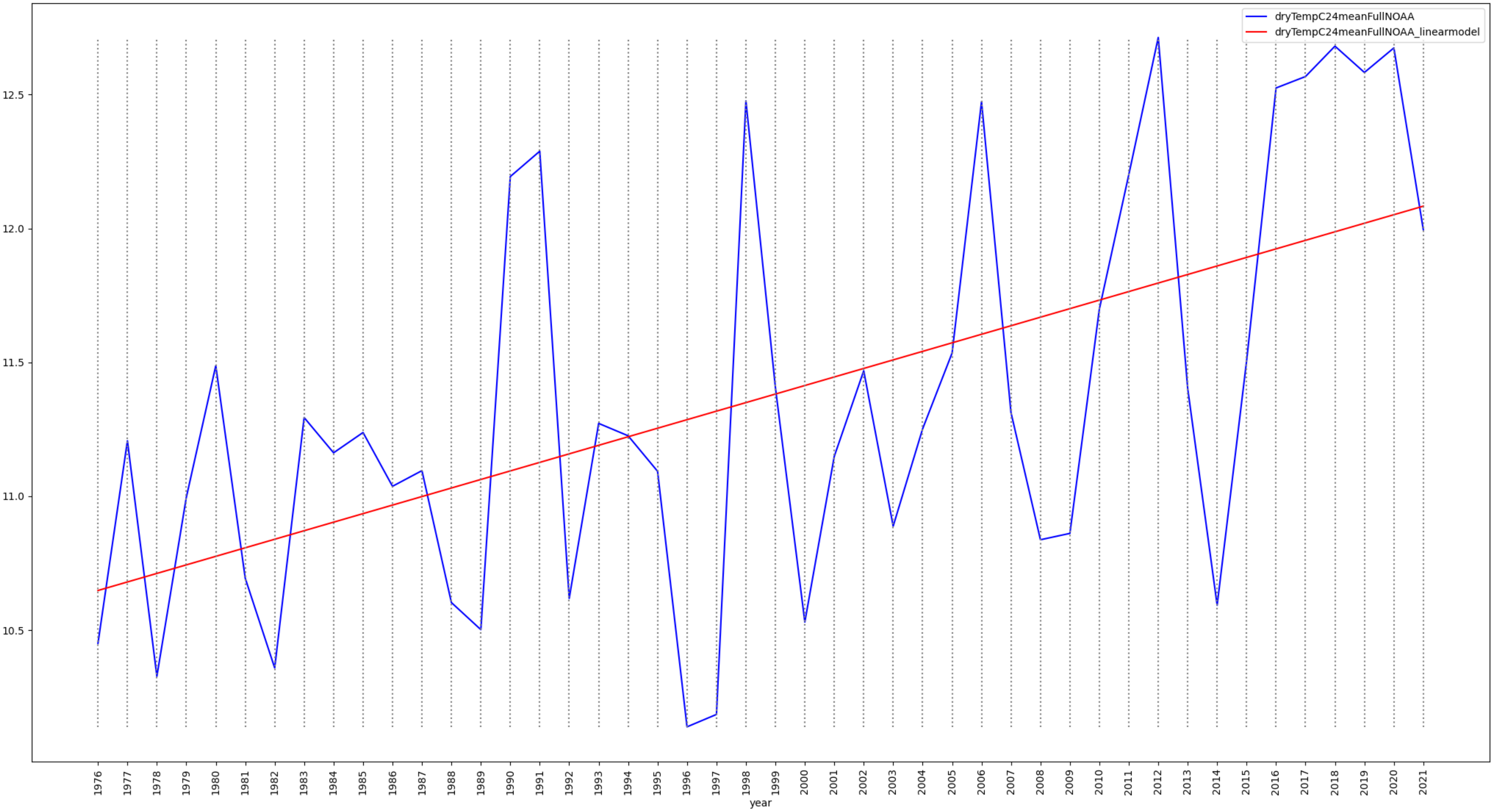


Figure 7: 1976 - 2021 Allentown Airport tempC mean and above Linear Regression MODEL 4

Again, with these year-round, round-the-clock temperature data measured at Allentown Airport but not at Hawk Mountain, they suffer from the Heat Island Effect compared to North Lookout. Given a CC of 0.8639, we can assume the HM model

HMtempC = 0.0198 * yearSince1976 + -0.2033 * daySinceAug1 + 27.5322, Correlation coefficient 0.8639

is accurate for our purposes of this analysis. There has been an average annual increase of about 0.02 degrees C at North Lookout over the last 46 years.

In a July 20, 2022 email Dr. Goodrich suggested examining temperature trends in earlier years. Because we are missing temperature data for 1970 through 1972 for both HM and Allentown, I derived a dataset for Allentown spanning 1948 through 1969, Trial and error led to identical models for Linear Regression, Simple Linear Regression, and M5P, after removing 1948 through 1951, leaving the 18 years from 1952 through 1969 in the data. This is that model for full years of temperature data from Allentown Airport.

MODEL 5

```
Linear Regression Model
dryTempC24meanFullNOAA =
    -0.0674 * yearSince1948 +
    11.2943
Correlation coefficient            0.5224
Mean absolute error                0.3848
Root mean squared error              0.4802
```

Temperatures trended down at the rate of -0.0674 degrees C per year from 1952 through 1969 compared to trending up by 0.0319 degrees C per year from 1976 through 2021. The error measures in these two formulas are very close, with the 1952-1969 measures slightly more accurate. Unfortunately, HM temperatures began to be recorded only in 1967, so correlating early raptor counts with HM temperatures is not possible.

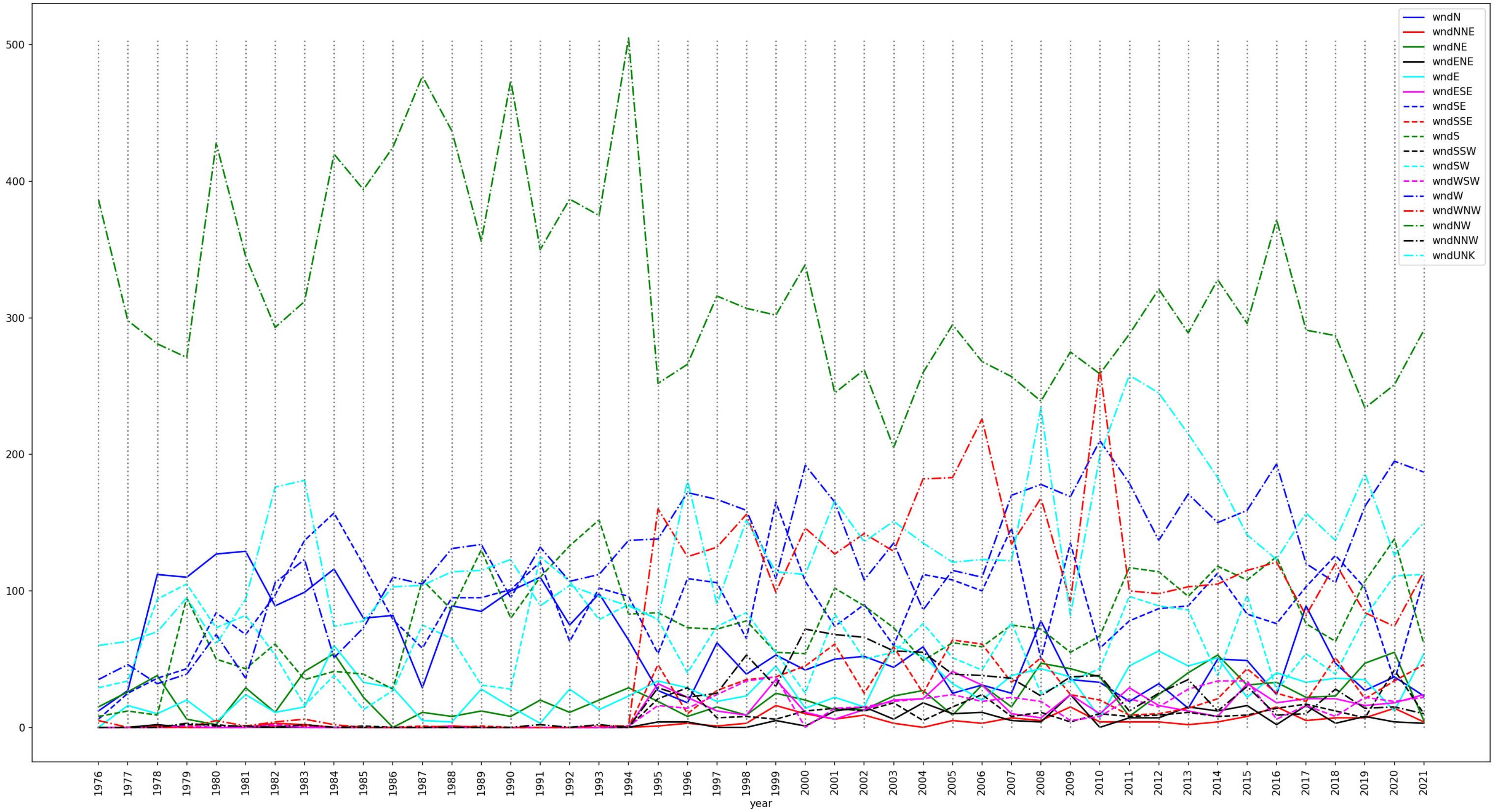
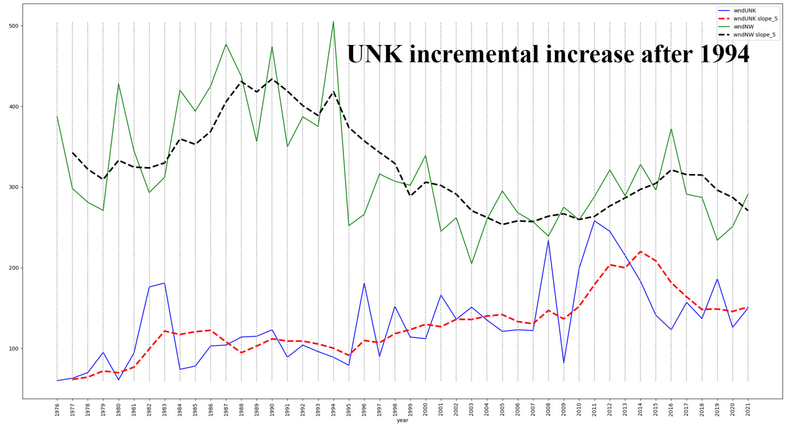
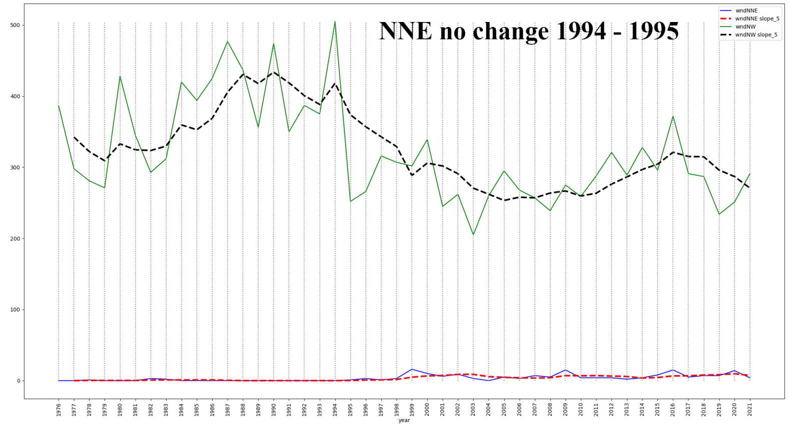


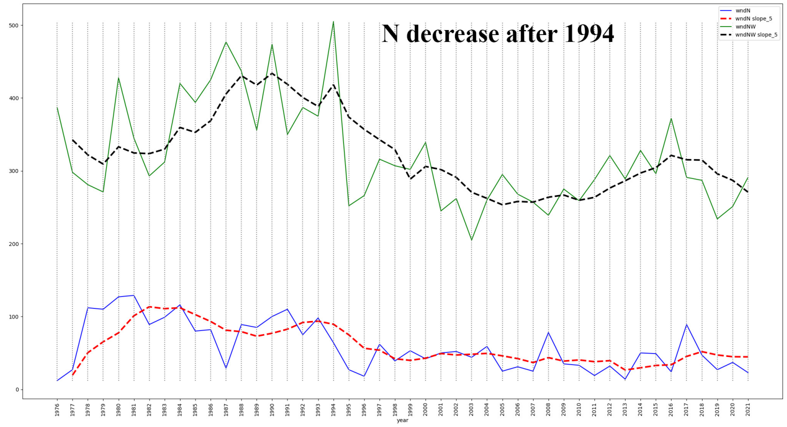
Figure 8: 1976 - 2021 Annual Wind Direction Tallies Aug-Dec at North Lookout

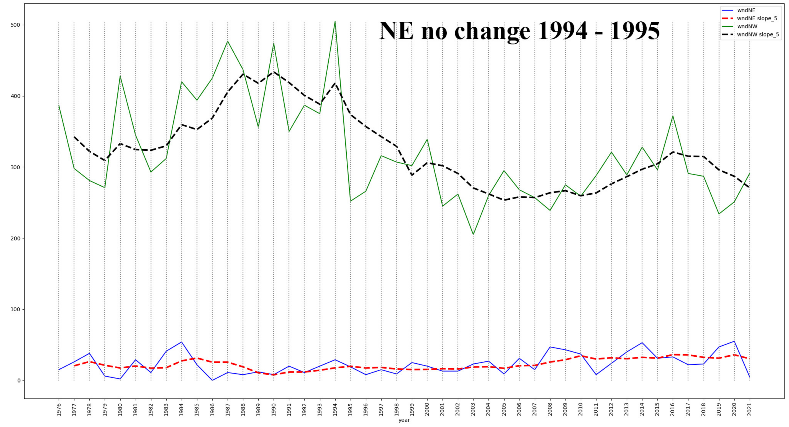
Winds are clearly important to raptor flight. Figure 8 shows annual tallies (sums) of wind direction counts at Hawk Mountain, including UNKnown for low and mixed wind conditions. Northwest wndNW tallies dominate all years except 2008 (236 for NW versus 234 for UNKnown) and 2010 (259 for NW versus 263 for WNW). The change that really stands out is the apparently permanent decline in NW wind count from 1994 (NW count = 505 for 124 observations Aug-Dec) to 1995 (NW count = 252 for 132 observations Aug-Dec). Increases from years before 1994 include UNK, W, and WNW.

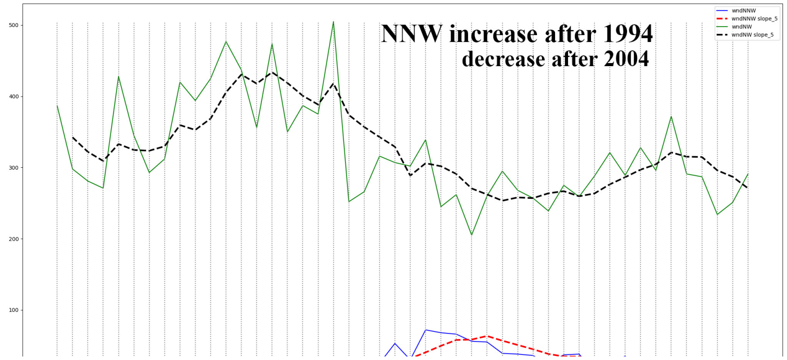
Figure 9 breaks the wind directions of Figure 8 into pairs consisting of NNE, NE, ENE, E, ESE, SE, SSE, and S (going down the right side), and SSW, SW, WSW, W, WNW, NNW, N, and UNK (going up the left side), with NW appearing in each graph. The slope_5 lines show five-year running averages of their counts. Going from 1994 to 1995 when NW shows its substantial decrease in count, WNW shows the most pronounced increase and N shows decrease. ESE, SSE, SSW, WSW, and NNW show small but steady increases. W shows a steady incremental increase since 1981

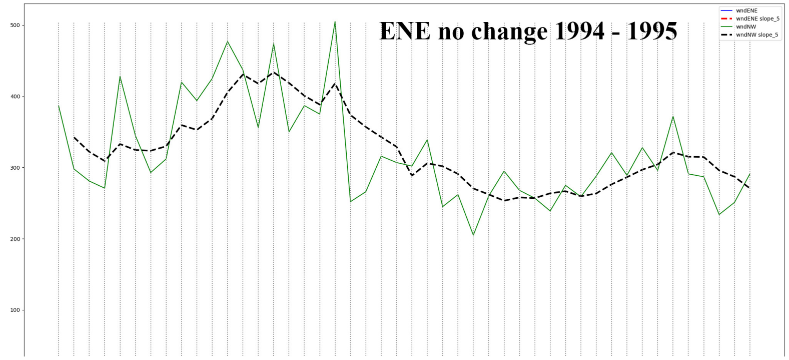

UNK incremental increase after 1994

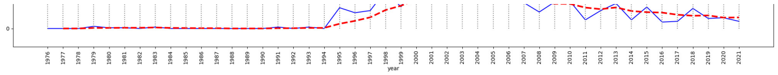

NNE no change 1994 - 1995

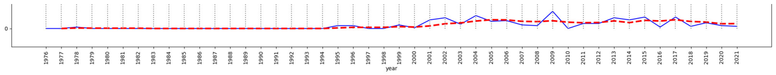

N decrease after 1994

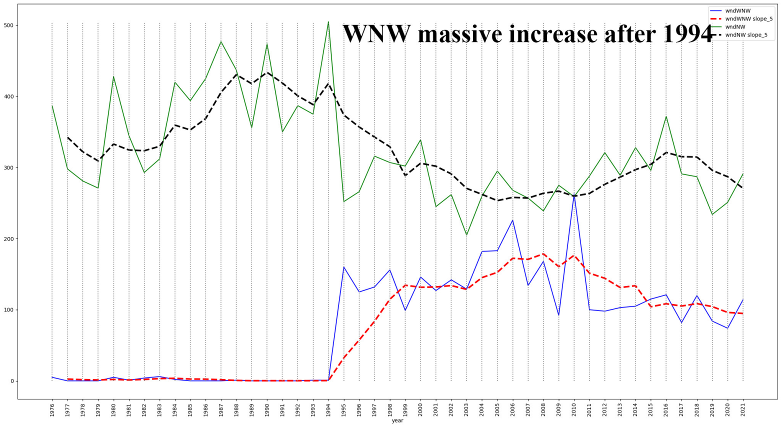

NE no change 1994 - 1995

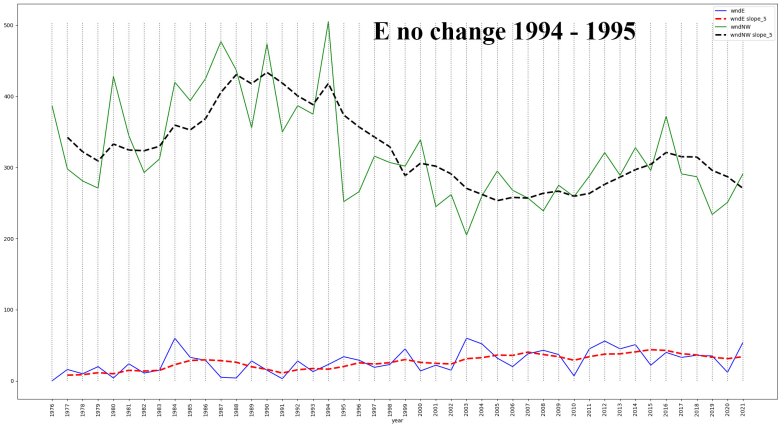

NNW increase after 1994
decrease after 2004

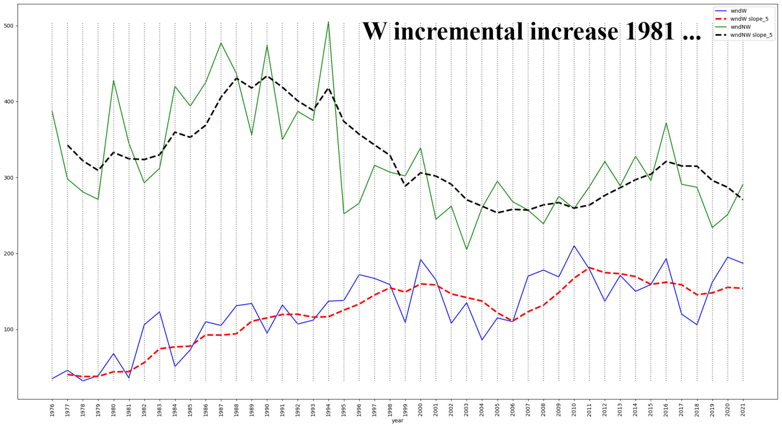

ENE no change 1994 - 1995

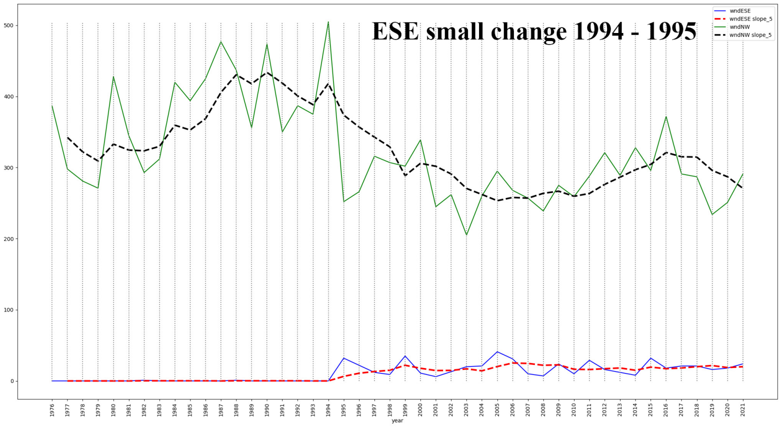


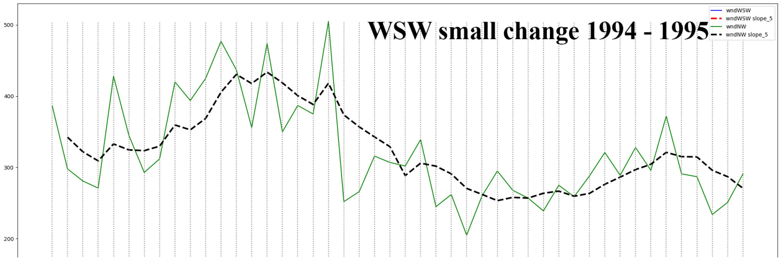


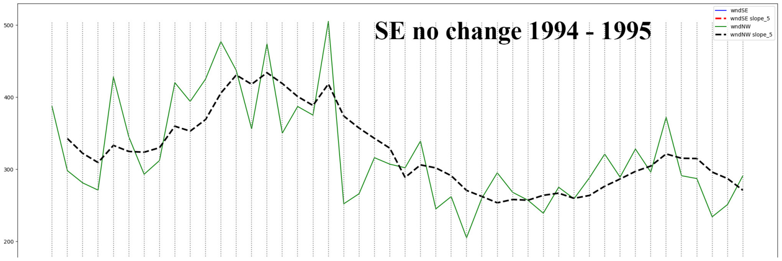

WNW massive increase after 1994


E no change 1994 - 1995


W incremental increase 1981 ...


ESE small change 1994 - 1995


WSW small change 1994 - 1995


SE no change 1994 - 1995

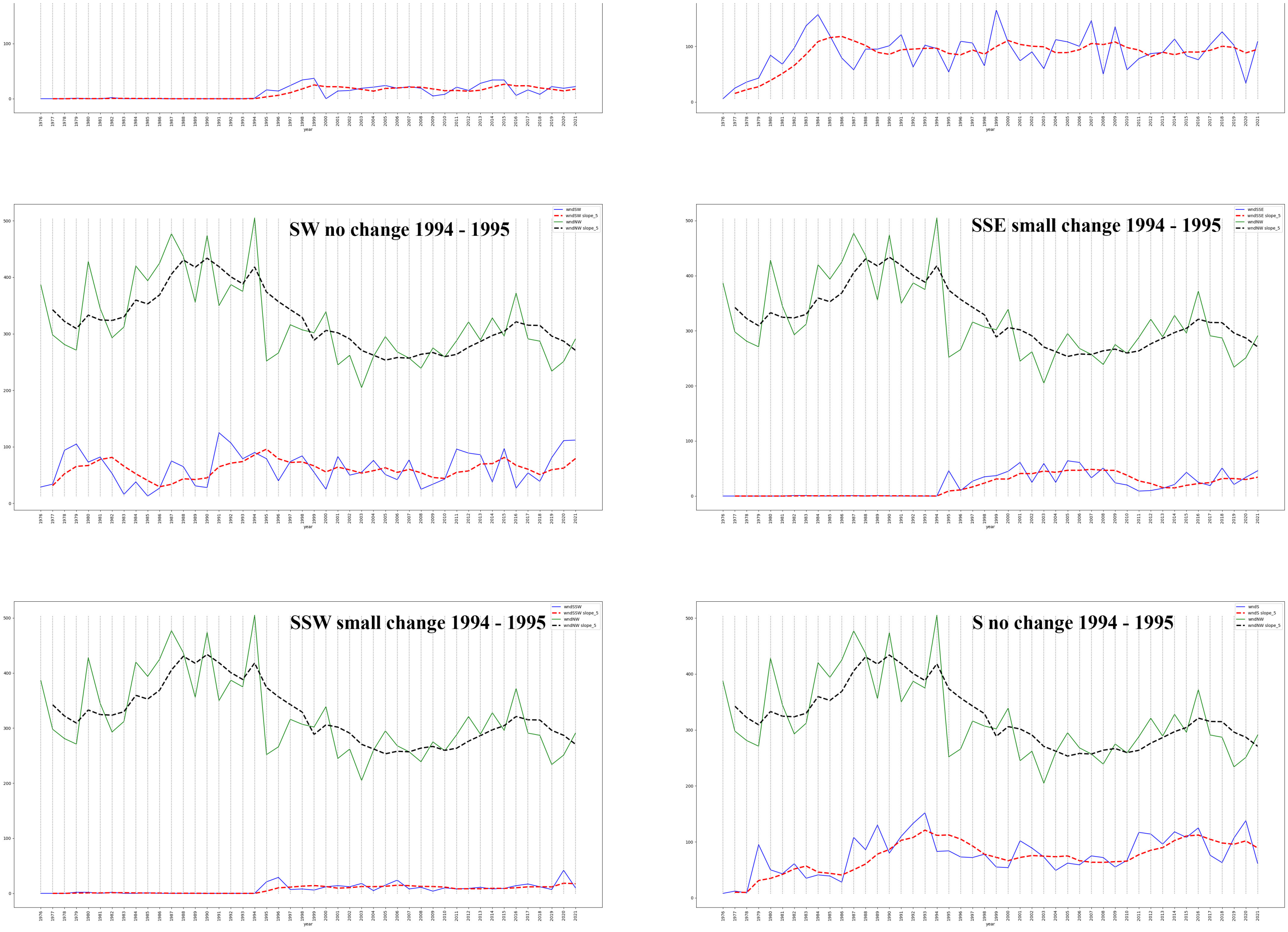


Figure 9: Pairwise 1976 - 2021 Annual Wind Direction Tallies Aug-Dec at North Lookout

Overall, wind directions are distributing away from the predominant NW during the fall observations. Figure 10 illustrates the post-1994 shifts using red arrows for increases and blue for decreases. NNW wind plays a prominent role in Section 3 analyses. It increased from 1994 through 2004 and has decreased consistently starting in 2005.

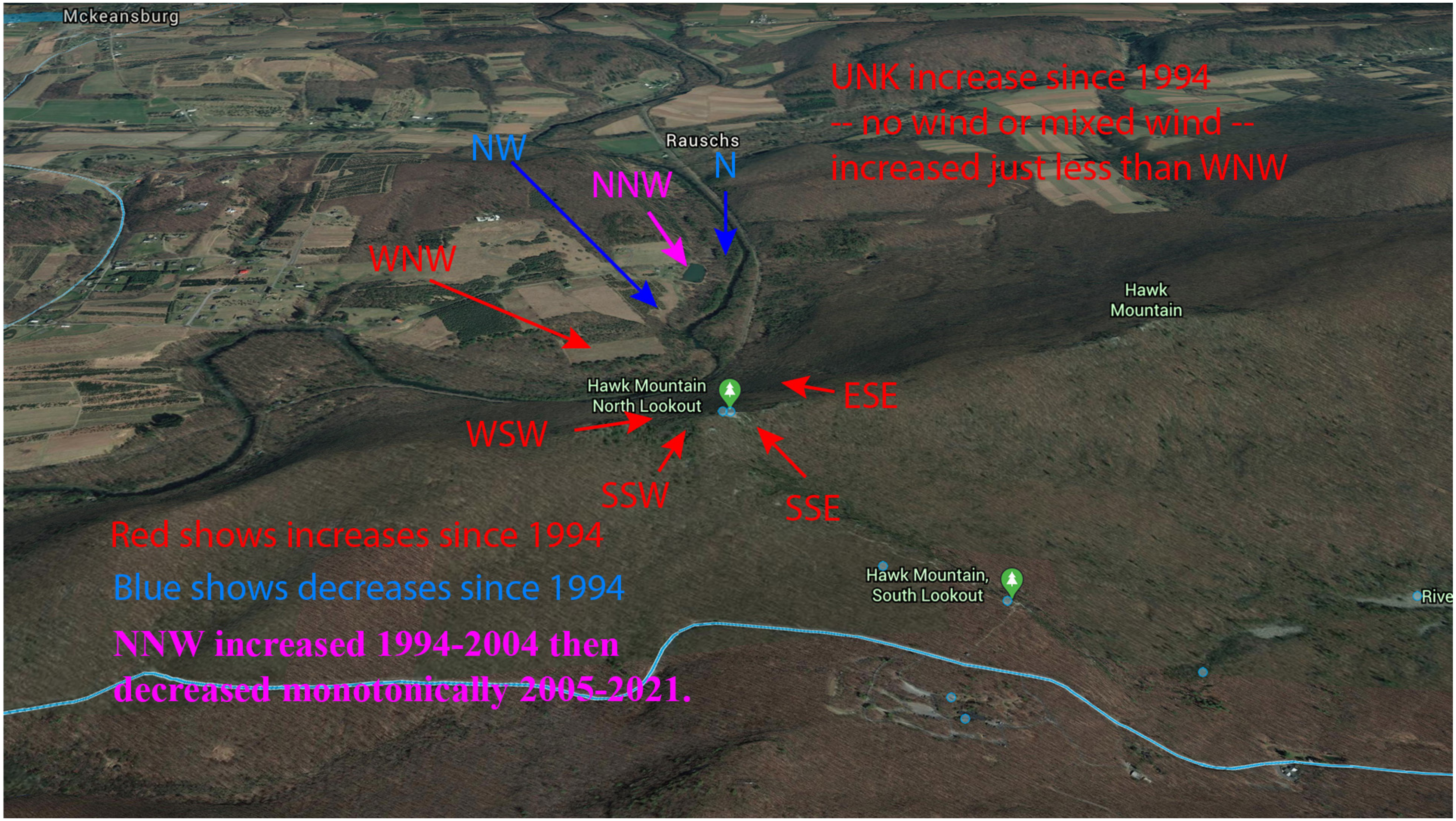

Mckeansburg
Rauschs
UNK increase since 1994
-- no wind or mixed wind --
increased just less than WNW
NW
NNW
N
WNW
Hawk Mountain
Hawk Mountain North Lookout
ESE
WSW
SSW
SSE
Red shows increases since 1994
Blue shows decreases since 1994
NNW increased 1994-2004 then decreased monotonically 2005-2021.
Hawk Mountain, South Lookout
Rive

Figure 10: Post-1994 Shifts in Wind Direction Tallies.

Figure 11 shows the most pronounced changes in raptor flight direction for Aug-Dec 1979 through 2021. Other measures are down in the noise. Note that a wind direction measure is in the direction from which the wind is arriving while a flight direction measure is in the direction to which raptors are flying. There is no pronounced permanent shift from 1994 to 1995. SW flight is trending downward from earlier years with an upward trend in UNK (mixed per observation) and WSW flights.

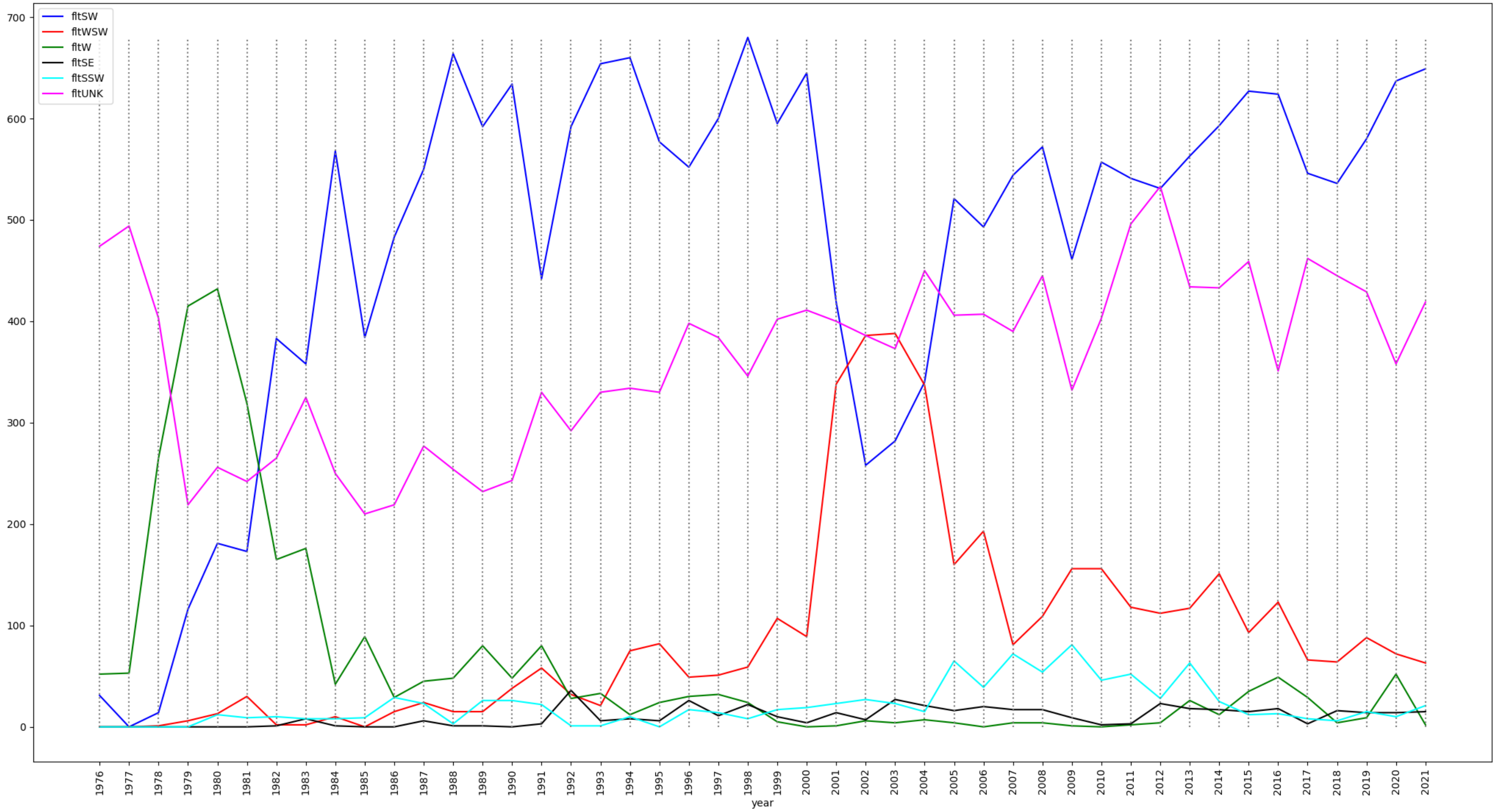


Figure 11: HM Flight Direction Tallies for the Six Most Pronounced Changes in 1976 - 2021 Observations

Figure 12 shows North Lookout wind speed mean, it population standard deviation, and its maximum divided by 5 for scaling to fit the display trending down since the 1990s. The dashed lines are --year running averages. Observers record wind speed in discrete groups as documented here, with measures in Figure 12 converted to the mean of each measurement group in km/h. On July 18 Dr. Goodrich wrote [15]: "I also know that in recent years, 2015 or after, some counters may be underestimating wind speed ... they started using a wind gauge ... maybe have people record wind in two ways so we can assess how much off, it would be only off by maybe 5-10 mph, Before 2010 or so we always used Beaufort wind scale, which is observer estimated based on tree movements. As rough as this sounds , I feel it does a better job of estimating winds than a hand held gauge that is subject to observer error and friction at ground level, tree blockage etc." Figure 12 shows a descending slope after 2010 that was already in effect after 2003.

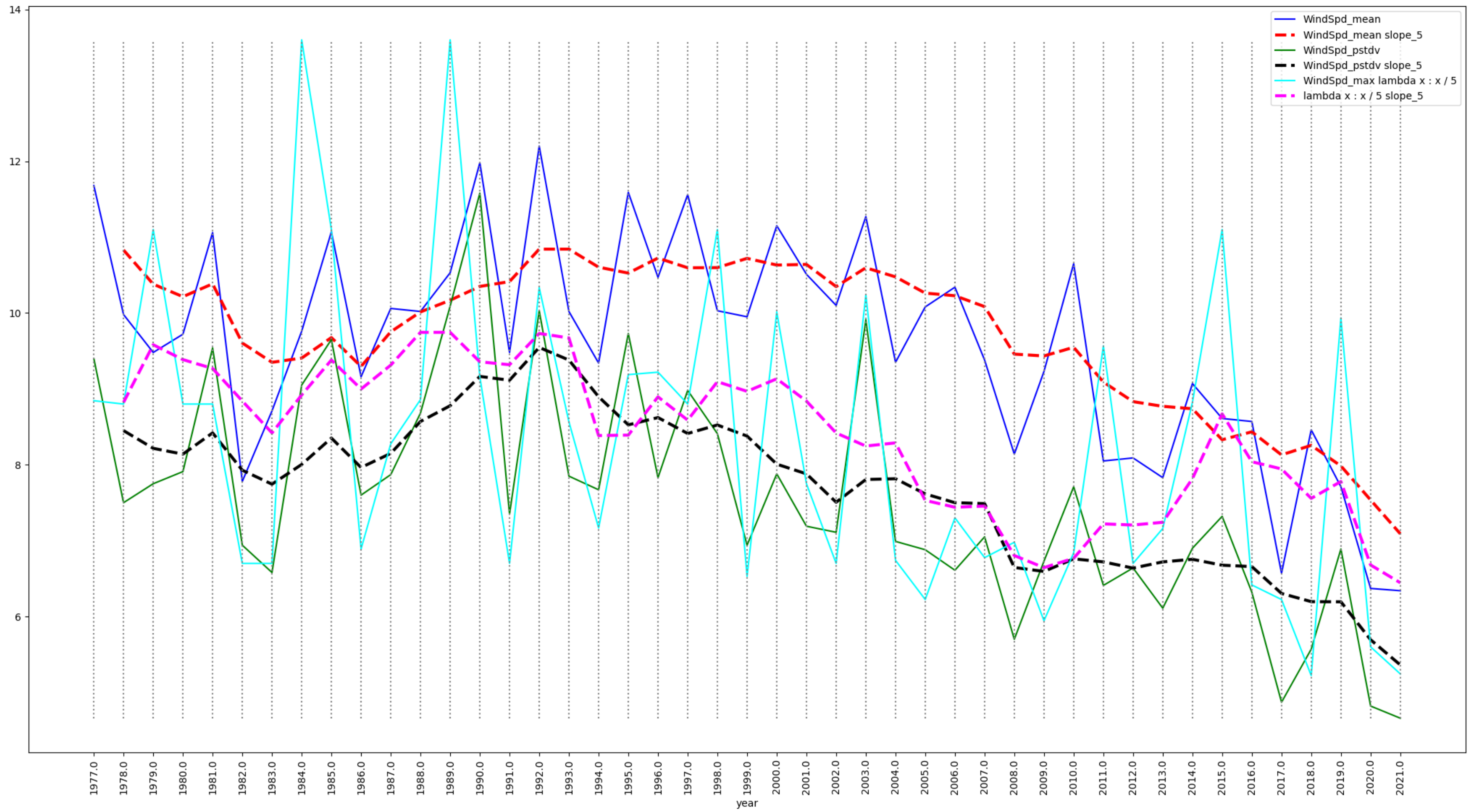


Figure 12: North Lookout Wind Speed Mean, Population Standard Deviation, and Max / 5 and Their 5-year Average Slopes

Figure 13 shows the Allentown Airport wind speed parameters paired one-to-one with the HM observations. Mean wind speed shows similar values and a decline starting earlier. Standard deviation is lower and declining and maximum speed is scaled down only by 2.5, both unsurprisingly in relation to North Lookout deviations and maximums at the top of a mountain. Figure 13 correlates to the decline in Figure 12. Other NOAA measures including percent humidity and barometric pressure (not graphed) do not show consistent changes from 1976 through 2021. Precipitation shows a minor uptick in recent years.

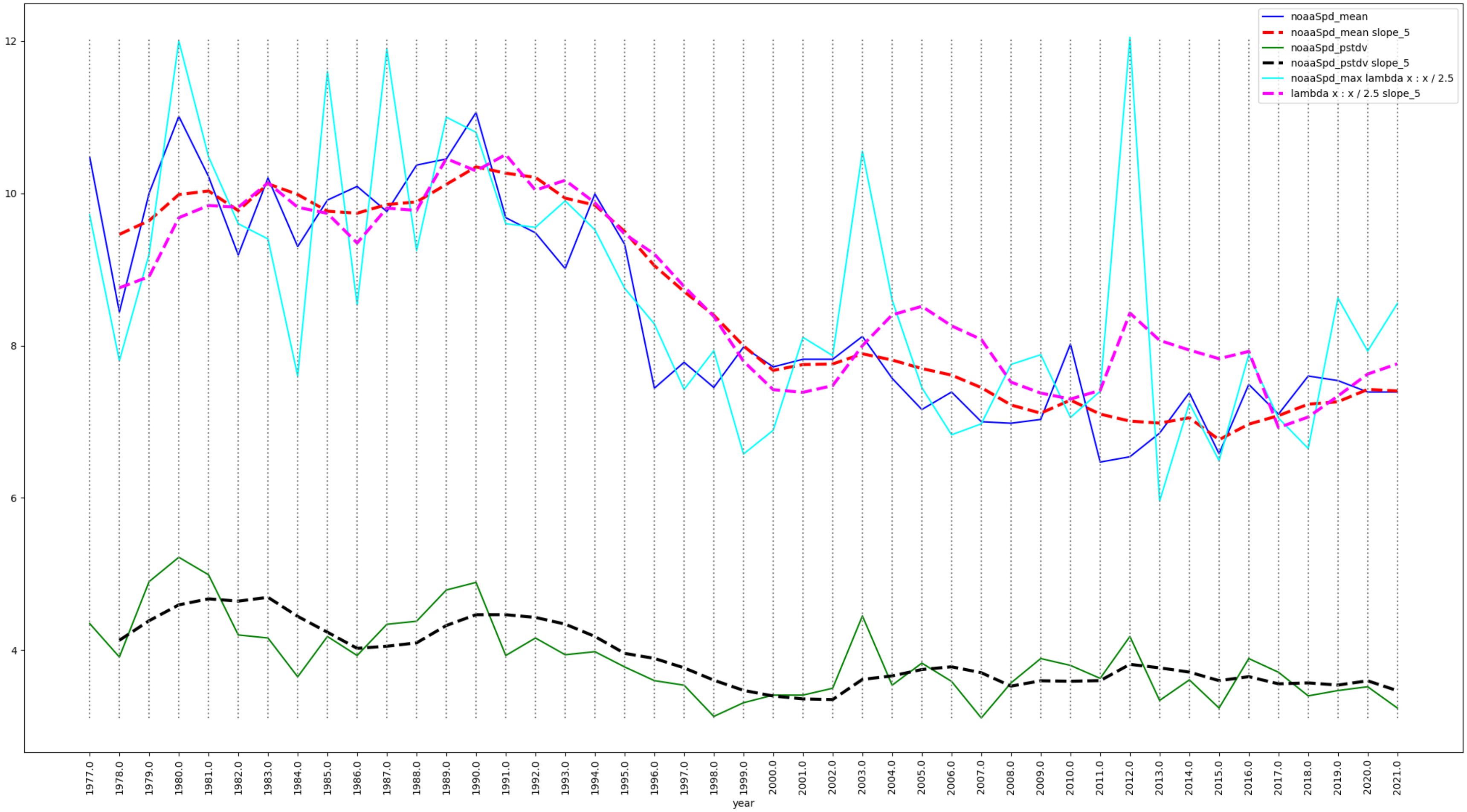


Figure 13: Allentown Airport Wind Speed Mean, Population Standard Deviation, and Max / 2.5 and Their 5-year Average Slopes

Figure 14 is the final graph devoted to weather measurements during the fall observation period. All values are normalized to the range [0.0, 1.0] so they fit in the same Y scale, with their legends giving their actual lower and upper values. All except HourlyVisibility are North Lookout measures which also show their 5-year running averages. HM Visibility trends downward in recent years and Flight Height oscillates (the key for its values is here). Mean CloudCover measures in percents start only in 2007 and trend sharply upward. These time-limited CloudCover data correlate strongly with some raptor counts examined in Section 3. Finally, NOAA mean HourlyVisibility shows an unaccounted and suspicious decline from 1995 to 1996 that does not correlate with HM Visibility. It seems very likely to be a data error.

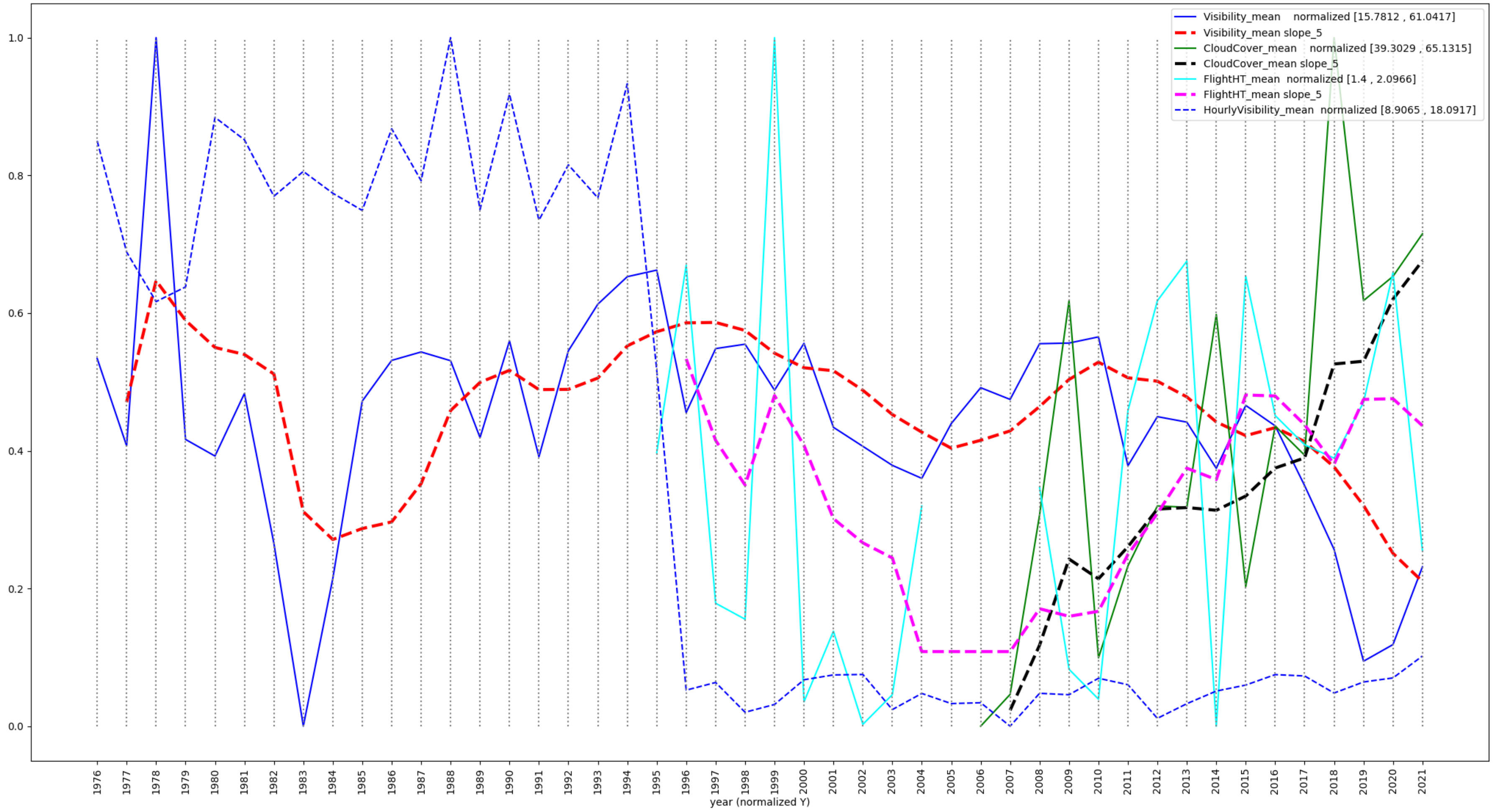


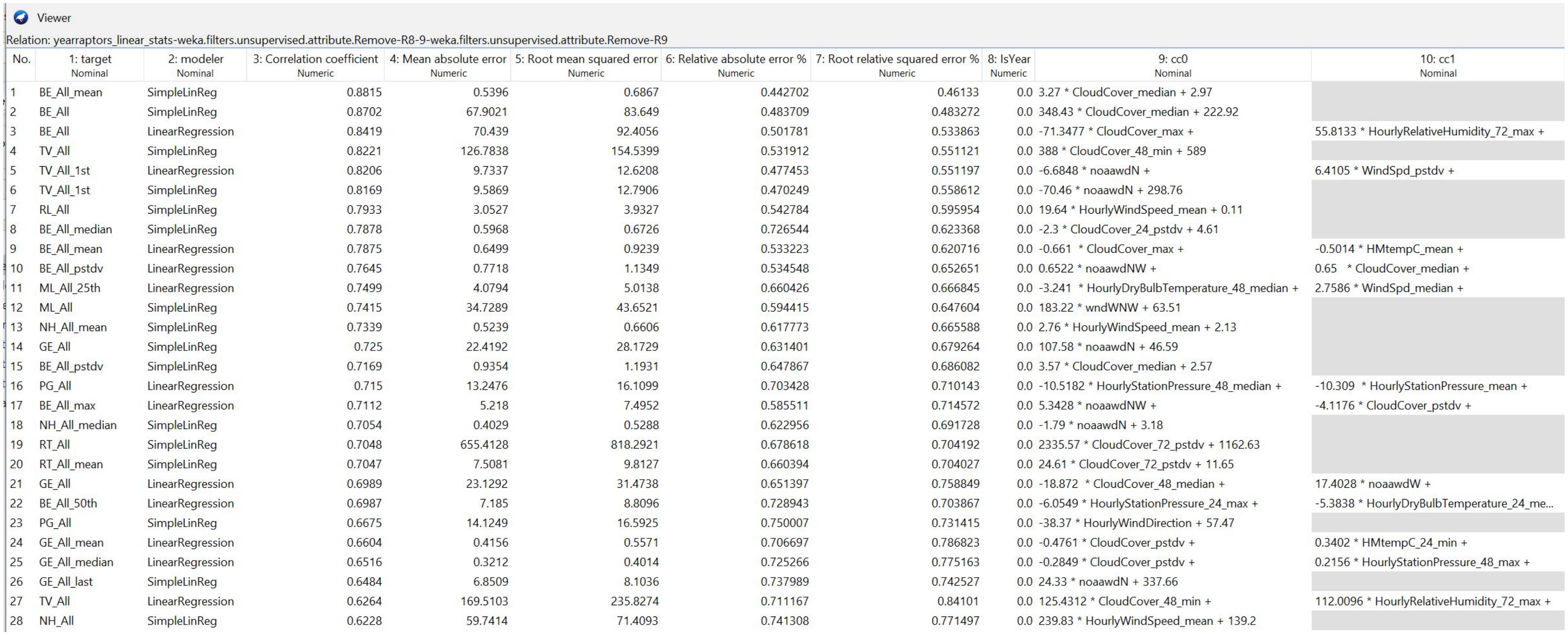
Figure 14: Hawk Mountain Mean Visibility, Mean Cloud Cover, Mean Flight Height, Their 5-year Average Slopes, and NOAA HourlyVisibility

3. ANALYSIS OF CLIMATE-TO-RAPTOR PATTERNS

3A. Building, Running, and Understanding the Models

The combinatorics of analyzing year-to-climate properties and (year, day)-to-climate properties of the previous section are simpler than those of climate-to-raptor properties. The former associate only one year or one (year, day) pair (control variable) with one climate attribute (experimental variable) per linear or other model, while climate-to-raptor analysis associates 227 climate and related attributes (227 control variables) with 935 raptor attributes, one at a time, in model construction and testing. Simplifying the raptor properties to only the "_All" subset (the sum of all subcategories such as Adult, Immature, etc.) reduces the raptor attribute attribute count from 935 to 253. It was necessary to write and extend an additional Python script to analyze all AttributeSelection, LinearRegression, and SimpleLinearRegression model runs to extract features including climate attributes frequently used to associate with raptors, climate attributes with large linear formula coefficients, the coefficients themselves, and large correlation coefficients (CCs). This Python code appears in Appendix A.

Figures 15 and 16 shows the Weka [8] output tables of this code for analyzing yearly aggregated data 1976 through 2021, which also rewrites the order of LinearRegression formula coefficients in descending order by magnitude. All attributes except the target raptor attribute are normalized into the range [0.0, 1.0] by taking each value's (value-minvalue)/(maxvalue-minvalue). Normalization puts all LinearRegression formula coefficients (multipliers) on the same scale so an analyst can compare them for importance. Normalization does not change LinearRegression accuracy or error measures. There are also dual runs for each model, one that includes the year attribute and one that does not, to avoid simple memorization of data by the models. The analyses of this section use no-year models unless otherwise noted.

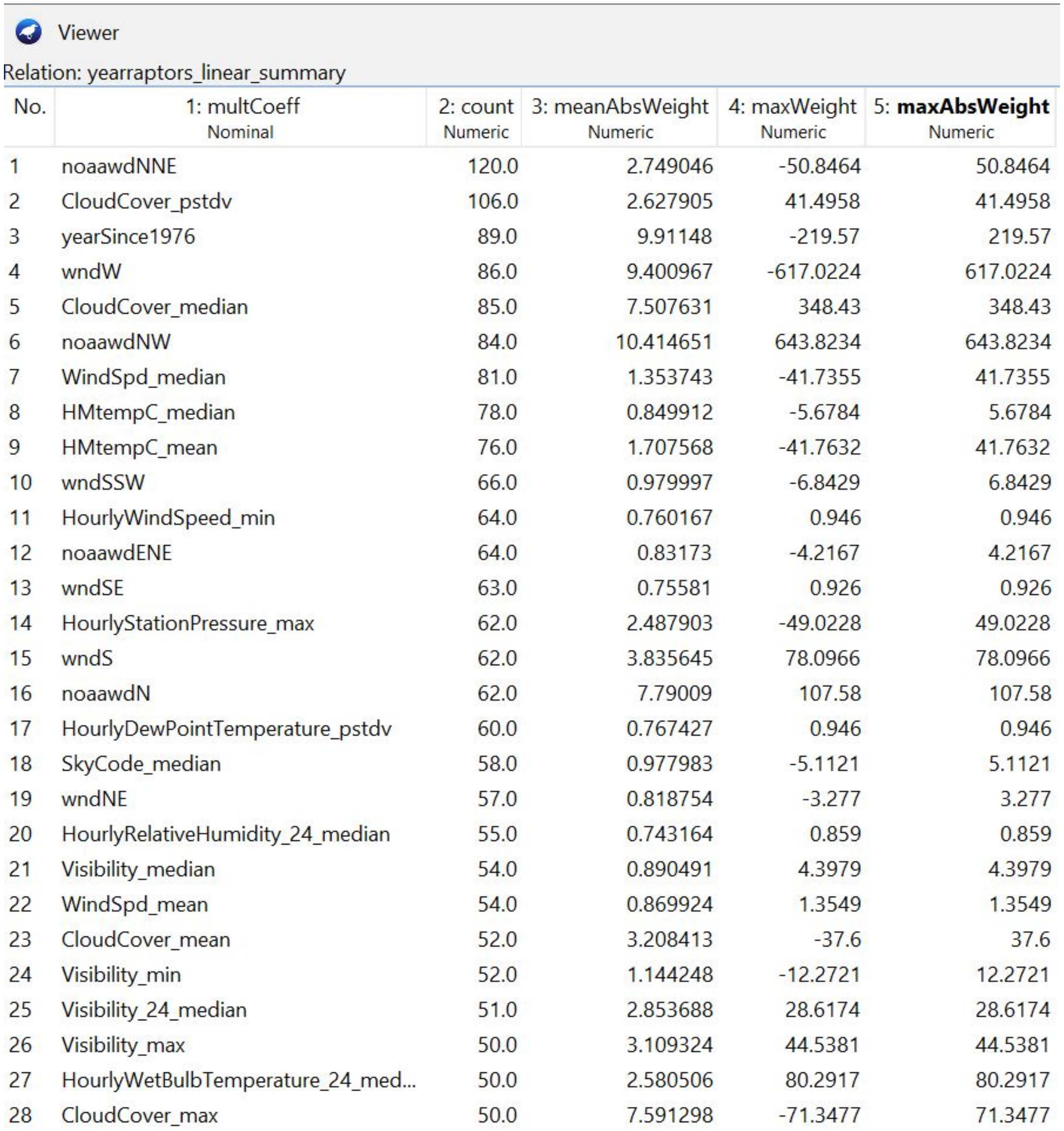

Viewer

Relation: yearraptors_linear_stats-weka.filters.unsupervised.attribute.Remove-R8-9-weka.filters.unsupervised.attribute.Remove-R9

No.	1: target	2: modeler	3: Correlation coefficient	4: Mean absolute error	5: Root mean squared error	6: Relative absolute error %	7: Root relative squared error %	8: IsYear	9: cc0	10: cc1
	Nominal	Nominal	Numeric	Numeric	Numeric	Numeric	Numeric	Numeric	Nominal	Nominal
1	BE_All_mean	SimpleLinReg	0.8815	0.5396	0.6867	0.442702	0.46133	0.0	3.27 * CloudCover_median + 2.97	
2	BE_All	SimpleLinReg	0.8702	67.9021	83.649	0.483709	0.483272	0.0	348.43 * CloudCover_median + 222.92	
3	BE_All	LinearRegression	0.8419	70.439	92.4056	0.501781	0.533863	0.0	-71.3477 * CloudCover_max +	55.8133 * HourlyRelativeHumidity_72_max +
4	TV_All	SimpleLinReg	0.8221	126.7838	154.5399	0.531912	0.551121	0.0	388 * CloudCover_48_min + 589	
5	TV_All_1st	LinearRegression	0.8206	9.7337	12.6208	0.477453	0.551197	0.0	-6.6848 * noaawdN +	6.4105 * WindSpd_pstdv +
6	TV_All_1st	SimpleLinReg	0.8169	9.5869	12.7906	0.470249	0.558612	0.0	-70.46 * noaawdN + 298.76	
7	RL_All	SimpleLinReg	0.7933	3.0527	3.9327	0.542784	0.595954	0.0	19.64 * HourlyWindSpeed_mean + 0.11	
8	BE_All_median	SimpleLinReg	0.7878	0.5968	0.6726	0.726544	0.623368	0.0	-2.3 * CloudCover_24_pstdv + 4.61	
9	BE_All_mean	LinearRegression	0.7875	0.6499	0.9239	0.533223	0.620716	0.0	-0.661 * CloudCover_max +	-0.5014 * HMtempC_mean +
10	BE_All_pstdv	LinearRegression	0.7645	0.7718	1.1349	0.534548	0.652651	0.0	0.6522 * noaawdNW +	0.65 * CloudCover_median +
11	ML_All_25th	LinearRegression	0.7499	4.0794	5.0138	0.660426	0.666845	0.0	-3.241 * HourlyDryBulbTemperature_48_median +	2.7586 * WindSpd_median +
12	ML_All	SimpleLinReg	0.7415	34.7289	43.6521	0.594415	0.647604	0.0	183.22 * wndWNW + 63.51	
13	NH_All_mean	SimpleLinReg	0.7339	0.5239	0.6606	0.617773	0.665588	0.0	2.76 * HourlyWindSpeed_mean + 2.13	
14	GE_All	SimpleLinReg	0.725	22.4192	28.1729	0.631401	0.679264	0.0	107.58 * noaawdN + 46.59	
15	BE_All_pstdv	SimpleLinReg	0.7169	0.9354	1.1931	0.647867	0.686082	0.0	3.57 * CloudCover_median + 2.57	
16	PG_All	LinearRegression	0.715	13.2476	16.1099	0.703428	0.710143	0.0	-10.5182 * HourlyStationPressure_48_median +	-10.309 * HourlyStationPressure_mean +
17	BE_All_max	LinearRegression	0.7112	5.218	7.4952	0.585511	0.714572	0.0	5.3428 * noaawdNW +	-4.1176 * CloudCover_pstdv +
18	NH_All_median	SimpleLinReg	0.7054	0.4029	0.5288	0.622956	0.691728	0.0	-1.79 * noaawdN + 3.18	
19	RT_All	SimpleLinReg	0.7048	655.4128	818.2921	0.678618	0.704192	0.0	2335.57 * CloudCover_72_pstdv + 1162.63	
20	RT_All_mean	SimpleLinReg	0.7047	7.5081	9.8127	0.660394	0.704027	0.0	24.61 * CloudCover_72_pstdv + 11.65	
21	GE_All	LinearRegression	0.6989	23.1292	31.4738	0.651397	0.758849	0.0	-18.872 * CloudCover_48_median +	17.4028 * noaawdW +
22	BE_All_50th	LinearRegression	0.6987	7.185	8.8096	0.728943	0.703867	0.0	-6.0549 * HourlyStationPressure_24_max +	-5.3838 * HourlyDryBulbTemperature_24_me...
23	PG_All	SimpleLinReg	0.6675	14.1249	16.5925	0.750007	0.731415	0.0	-38.37 * HourlyWindDirection + 57.47	
24	GE_All_mean	LinearRegression	0.6604	0.4156	0.5571	0.706697	0.786823	0.0	-0.4761 * CloudCover_pstdv +	0.3402 * HMtempC_24_min +
25	GE_All_median	LinearRegression	0.6516	0.3212	0.4014	0.725266	0.775163	0.0	-0.2849 * CloudCover_pstdv +	0.2156 * HourlyStationPressure_48_max +
26	GE_All_last	SimpleLinReg	0.6484	6.8509	8.1036	0.737989	0.742527	0.0	24.33 * noaawdN + 337.66	
27	TV_All	LinearRegression	0.6264	169.5103	235.8274	0.711167	0.84101	0.0	125.4312 * CloudCover_48_min +	112.0096 * HourlyRelativeHumidity_72_max +
28	NH_All	SimpleLinReg	0.6228	59.7414	71.4093	0.741308	0.771497	0.0	239.83 * HourlyWindSpeed_mean + 139.2	

Figure 15: Linear Model Properties for the Top Contenders by CC for No-Year Models

Figure 15 shows the first 28 rows of model properties, out of 612 total rows with a CC of at least 0.5, sorted primarily with no-year at the top and then by CC in descending order. There are 10 linear formula multipliers cc0 through cc9 tagged to each full LinearRegression model, with ccN multipliers descending in weight from left to right in this table. Figure 15 shows only cc0 and cc1. SimpleLinearRegression only ever uses one cc0 multiplier.



Relation: yearraptors_linear_summary

No.	1: multCoeff Nominal	2: count Numeric	3: meanAbsWeight Numeric	4: maxWeight Numeric	5: **maxAbsWeight** Numeric
1	noaawdNNE	120.0	2.749046	-50.8464	50.8464
2	CloudCover_pstdv	106.0	2.627905	41.4958	41.4958
3	yearSince1976	89.0	9.91148	-219.57	219.57
4	wndW	86.0	9.400967	-617.0224	617.0224
5	CloudCover_median	85.0	7.507631	348.43	348.43
6	noaawdNW	84.0	10.414651	643.8234	643.8234
7	WindSpd_median	81.0	1.353743	-41.7355	41.7355
8	HMtempC_median	78.0	0.849912	-5.6784	5.6784
9	HMtempC_mean	76.0	1.707568	-41.7632	41.7632
10	wndSSW	66.0	0.979997	-6.8429	6.8429
11	HourlyWindSpeed_min	64.0	0.760167	0.946	0.946
12	noaawdENE	64.0	0.83173	-4.2167	4.2167
13	wndSE	63.0	0.75581	0.926	0.926
14	HourlyStationPressure_max	62.0	2.487903	-49.0228	49.0228
15	wndS	62.0	3.835645	78.0966	78.0966
16	noaawdN	62.0	7.79009	107.58	107.58
17	HourlyDewPointTemperature_pstdv	60.0	0.767427	0.946	0.946
18	SkyCode_median	58.0	0.977983	-5.1121	5.1121
19	wndNE	57.0	0.818754	-3.277	3.277
20	HourlyRelativeHumidity_24_median	55.0	0.743164	0.859	0.859
21	Visibility_median	54.0	0.890491	4.3979	4.3979
22	WindSpd_mean	54.0	0.869924	1.3549	1.3549
23	CloudCover_mean	52.0	3.208413	-37.6	37.6
24	Visibility_min	52.0	1.144248	-12.2721	12.2721
25	Visibility_24_median	51.0	2.853688	28.6174	28.6174
26	Visibility_max	50.0	3.109324	44.5381	44.5381
27	HourlyWetBulbTemperature_24_med...	50.0	2.580506	80.2917	80.2917
28	CloudCover_max	50.0	7.591298	-71.3477	71.3477

Figure 16: Summary Statistics for the Linear Model Climate Attributes of Figure 15

This Python script pulls out a **count** of how many times each climate attribute appears in linear regression formulas, i.e., how many formulas contain it. It also pulls out the **mean of the absolute value of that attribute's multipliers** across all formulas, the **maximum-magnitude signed value of that multiplier**, and the **absolute value of the latter** for sorting. Figure 16 shows the first 28 rows, sorted by count, of 212 such summary analysis rows.

3B. Analyzing Yearly Climate to Raptor Associations

NOTE TO THE READER: The analysis up to the SS (sharp-shinned) raptors was exploratory in the sense that I was exploring modeling techniques and correlations by a trial-and-error approach. Sharp-shinned analysis is more concise because I used techniques that gave clear results. Also, I summarize some modeling algorithms in this section leading up to SS.

The steps in this section are guided somewhat by the tables of Figures 15, 16, and their underlying models, and partly by suggestions from Dr. Goodrich. It is essential to note that this analysis can establish correlation but typically does not establish causation. Where one climate attribute correlates with a raptor attribute I strive to uncover additional climate attributes correlated with that initial climate attribute. The fact that we are looking at data captured at a single physical location that does not include winter, spring, or summer raptor locations and behaviors means that causation may be difficult to ferret out.

The first two rows of Figure 15 lead to these two Simple Linear Regression models.

MODEL 6

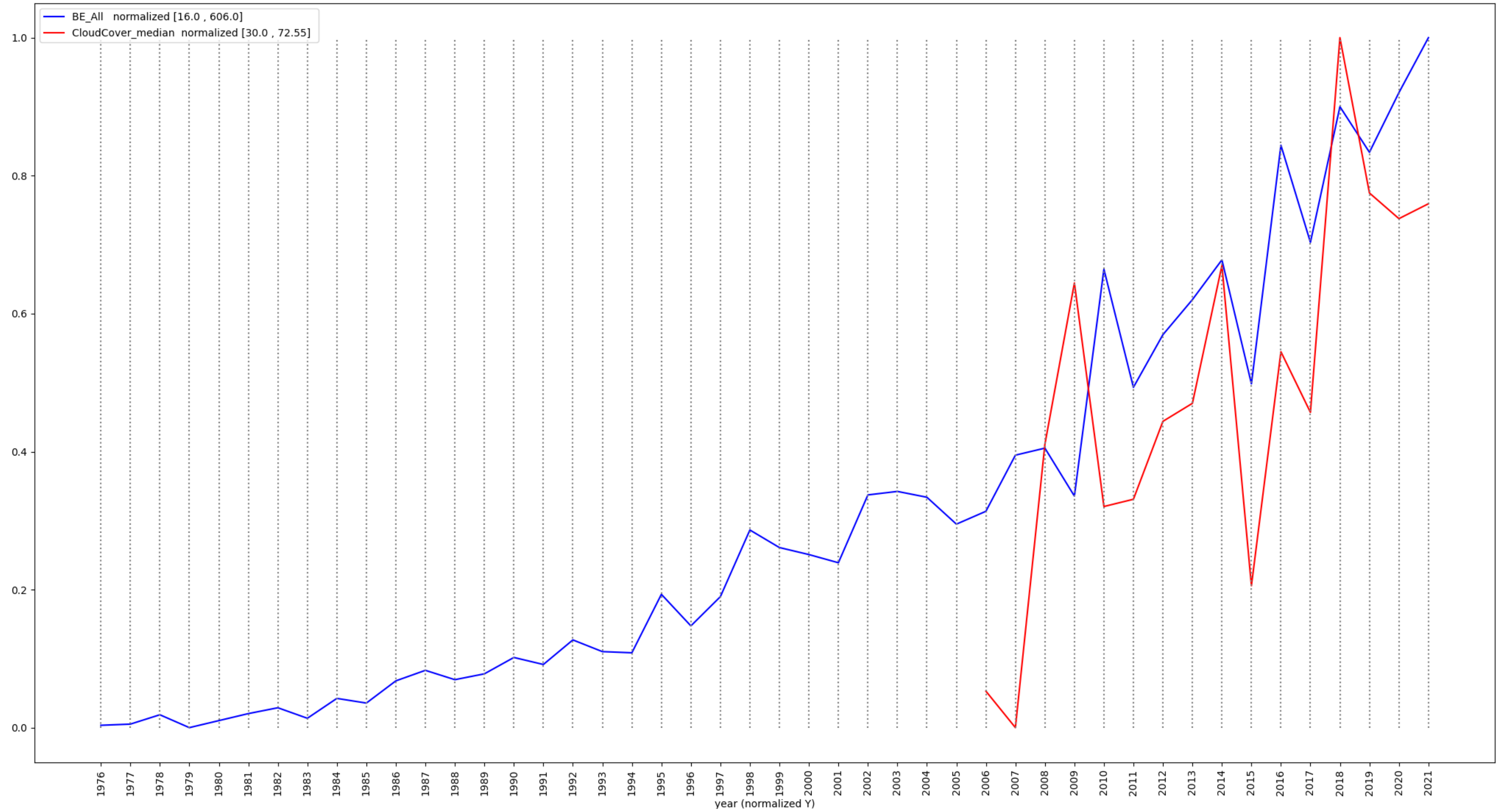
```
BE_All_mean = 3.27 * CloudCover_median + 2.97, Predicting 2.02 if attribute value is missing.
Correlation coefficient                  0.8815
Mean absolute error                      0.5396
Root mean squared error                    0.6867
```

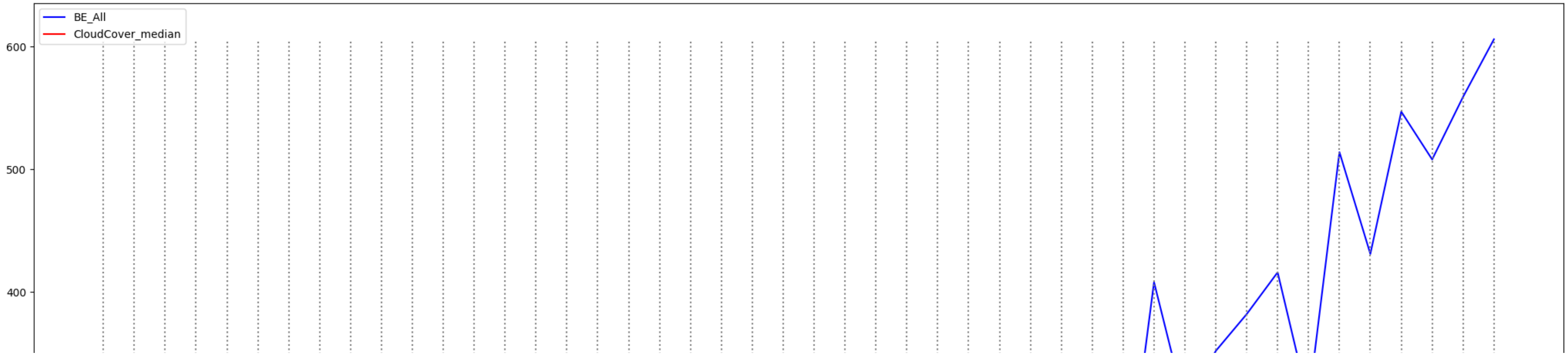
MODEL 7

```
BE_All = 348.43 * CloudCover_median + 222.92, Predicting 92.53 if attribute value is missing.
Correlation coefficient                  0.8702
Mean absolute error                     67.9021
Root mean squared error                   83.649
```

BE_All_mean is simply BE_All (the sum of all bald eagle observations) divided by the number of observations in the year. Hence, the CCs are close, but the CloudCover multiplier and the error measures scale according to the attribute being predicted, BE_All or BE_All_mean. BE_All is easier to interpret from a graph since it is simply the count of BEs for each year. Figure 17 graphs BE_All and CloudCover_median in two ways. The top graph normalizes their values into the range [0.0, 1.0] to ease inspection of peak and valley alignment in time. The bottom graphs shows their actual values for each year -- full count for BE and median value for CloudCover -- to show the massive growth in BE counts since 1976. An important thing to notice is the extremely high correlation of peaks and valleys between these two attributes starting in 2011. It appears that the highly correlated upward slopes of CloudCover_median and BE_All serve as a time keeper for tracking the years (Model 16 discussion expands on the climate-attribute **time keeper** concept), while the positively correlated peaks and valleys of Figure 17 indicate true climate effects on BE_All. These could be direct contributions of CloudCover or contributions of related climate attributes.

The recovery of the bald eagle in PA is of course an exciting story, "So, much eagle habitat was compromised early as this state was colonized and its population grew. Timbering also eliminated many good nesting sites because eagles generally nest in large trees. But the primary reason for the eagle's decline over the last century was the effect of the pesticide DDT and its derivatives on eagle reproduction. It accumulated in eagles and caused their eggs to be too thin to withstand the eagle's weight during incubation. As a result, the bald eagle population plummeted. In 1972, the use of this DDT and other harmful pesticides that bio-accumulated in birds was banned in the United States. The drastic decline of bald eagles and other birds eventually bottomed out. Clean water regulations and heightened environmental awareness during this period also contributed to better fish populations and spurred the recovery on." [16]


BE_All normalized [16.0 , 606.0]
CloudCover_median normalized [30.0 , 72.55]
1.0
0.8
0.6
0.4
0.2
0.0
1976
1977
1978
1979
1980
1981
1982
1983
1984
1985
1986
1987
1988
1989
1990
1991
1992
1993
1994
1995
1996
1997
1998
1999
2000
2001
2002
2003
2004
2005
2006
2007
2008
2009
2010
2011
2012
2013
2014
2015
2016
2017
2018
2019
2020
2021
year (normalized Y)


BE_All
CloudCover_median
600
500
400

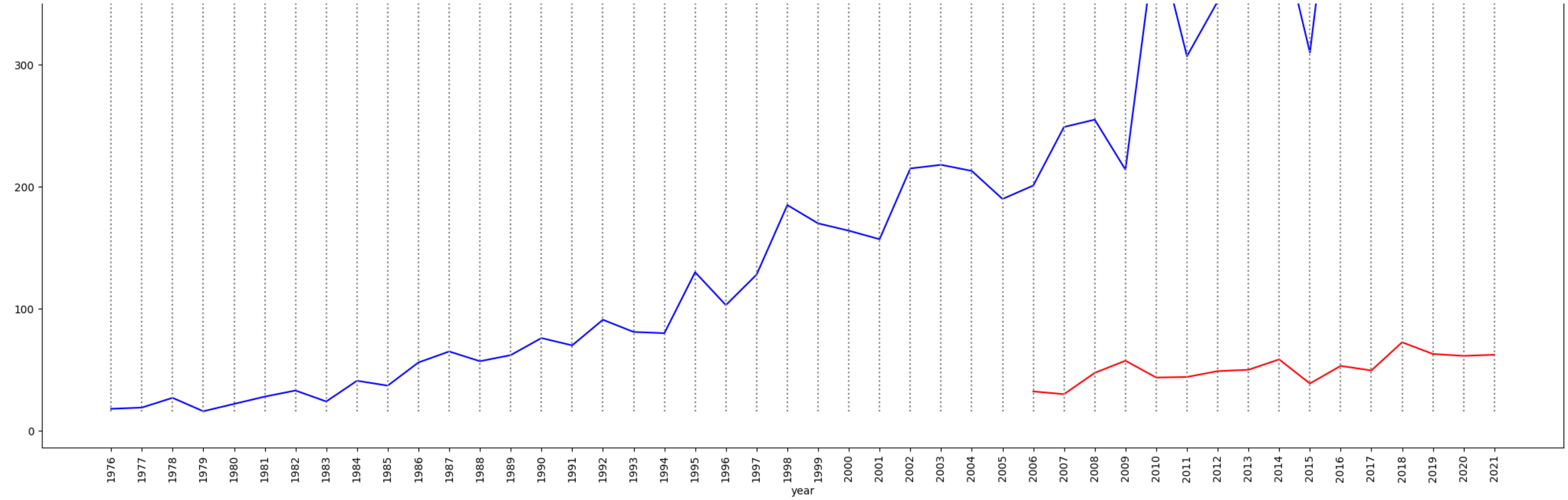


Figure 17: Normalized BE_All tallies and CloudCover Percentages (top) and Raw Numbers (bottom) for HM Observation Days

CloudCover does not appear in the HM dataset until 2006, and it would be really nice to both get more data and also to correlate it to other climate attributes. The reason for "Predicting 2.02 if attribute value is missing" in the BE_All_mean SimpleLinearRegression model and "Predicting 92.53 if attribute value is missing" model is that CloudCover is missing before 2005 when the BE values were consistently low. 92.53 is just the mean BE_all value between 1976 and 2005. On reason for these models' accuracy is the consistently low BE values in early years.

In an attempt to find a replacement for CloudCover I analyzed mapping 38 normalized North Lookout climate measures to CloudCover_median. Here is the LinearRegression model for the 16 years 2006 through 2021 when CloudCover was recorded.

MODEL 8

```
Linear Regression Model
CloudCover_median =
    -5.8135 * HMtempC_mean +
    -4.7909 * Visibility_mean +
     9.7847 * SkyCode_mean +
    -5.7107 * WindSpd_mean +
   -10.6196 * HMtempC_median +
    -3.7118 * Visibility_median +
     4.3707 * SkyCode_median +
    -6.2348 * WindSpd_median +
     7.2354 * HMtempC_pstdv +
    -8.4994 * Visibility_pstdv +
    -9.3012 * WindSpd_pstdv +
   -11.6912 * WindSpd_min +
     3.1    * HMtempC_max +
     4.7114 * SkyCode_max +
    -1.6948 * WindSpd_max +
     8.436  * wndNE +
     3.6759 * wndENE +
    -4.0914 * wndSE +
     7.1938 * wndSSE +
    -4.6624 * wndS +
    -7.6931 * wndSSW +
    -1.161  * wndSW +
     4.2785 * wndW +
     9.688  * wndNW +
    63.3077
Correlation coefficient                  0.4519
Mean absolute error                      9.5347
Root mean squared error                   12.0404
```

Since all non-CloudCover_median attributes are normalized into the range [0.0, 1.0] the weights operate on the same scale. Higher magnitude weights mean more importance. Here is the SimpleLinearRegression model for the same data.

MODEL 9

```
CloudCover_median = -25.2 * Visibility_pstdv + 68.3
Predicting 0 if attribute value is missing.
Correlation coefficient                  0.5763
Mean absolute error                      7.5679
Root mean squared error                   9.3368
```

M5P model tree yields this linear regression model with no decision tree.

MODEL 10

```
M5 pruned model tree:
LM num: 1
CloudCover_median =
   17.4692 * SkyCode_mean
   - 30.0622 * Visibility_pstdv
   + 64.4891
Number of Rules : 1
Correlation coefficient                  0.5303
Mean absolute error                      8.1035
Root mean squared error                   9.7222
```

SkyCode mean and median appear in both the LinearRegression and M5P models, which makes sense, since increasing SkyCode numbers 0 through 9 correspond to increasing cloud cover (reference here). This stage of the analysis is a search for missing CloudCover years. It does not explain why cloudier conditions would correspond with an increase in BE counts. The multipliers for the Visibility measures in the above models are all negative, meaning BE_All goes up as Visibility goes down. A strong LinearRegression multiplier for wndNW and a strong negative multiplier for wndSSW are intuitive, but the strong positive multiplier for wndSSE seems counterintuitive for the winds at North Lookout.

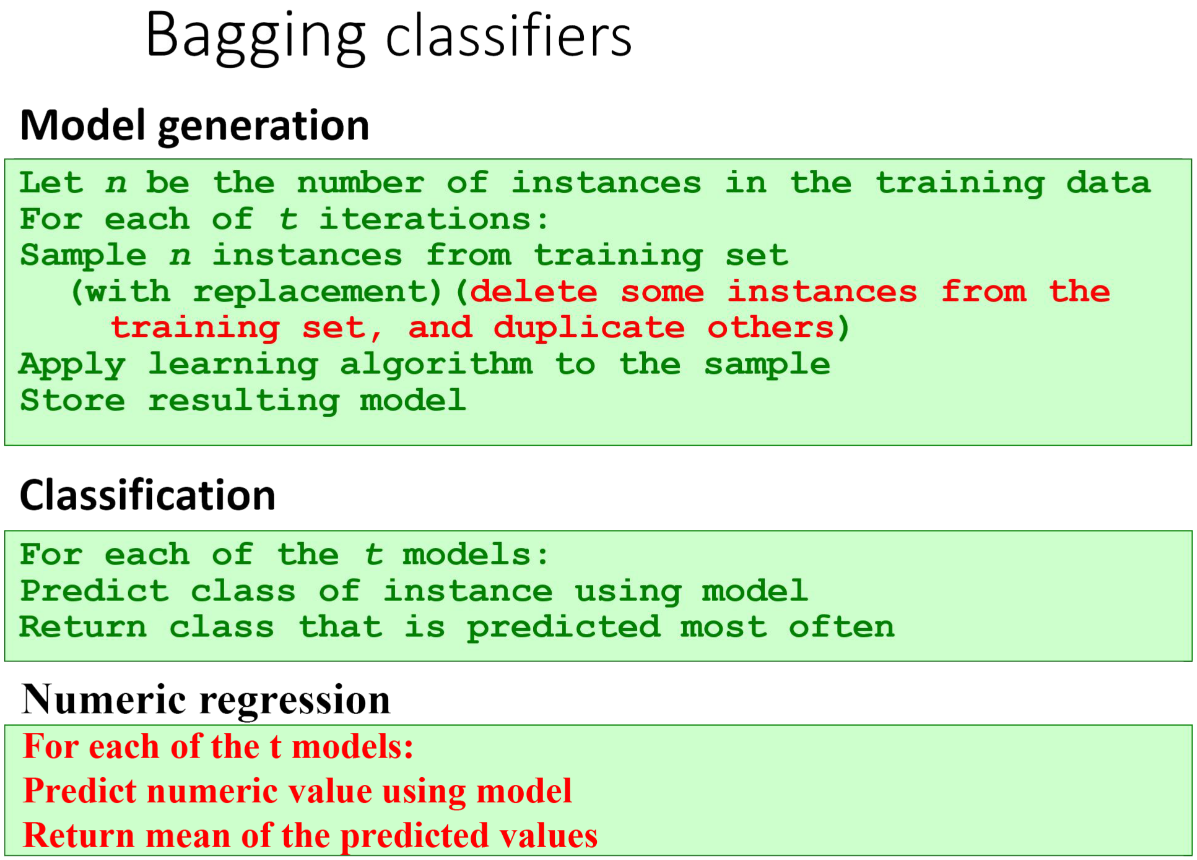
I ran an ensemble Bagging regressor using several underlying base regressors and got the following error measures when using 10 LinearRegression models.

MODEL 11

```
Bagging with 10 iterations and base learner Linear Regression Model
Correlation coefficient                  0.7144
Mean absolute error                      6.6577
```

Root mean squared error 8.5975

The CC of 0.7144 is a 24% improvement over the CC of 0.5763 of SimpleLinearRegression. The problem for intelligibility is that this Bagging run built 10 LinearRegression models. For each it took out some random number of years, duplicated others to get back to 16, and had LinearRegression build another model. It then averaged the resulting predictions. Bagging can help smooth over outliers in the data by increasing the probability of using more typical data, but manually interpreting 10 linear models built using 12 attributes mapped to CloudCover_median passes a point of diminishing returns. Figure 18 [17] summarizes Bagging for numeric regression with my additions in red.



Bagging classifiers

Model generation

```
Let n be the number of instances in the training data
For each of t iterations:
Sample n instances from training set
   (with replacement)(delete some instances from the
     training set, and duplicate others)
Apply learning algorithm to the sample
Store resulting model
```

Classification

```
For each of the t models:
Predict class of instance using model
Return class that is predicted most often
```

Numeric regression

```
For each of the t models:
Predict numeric value using model
Return mean of the predicted values
```

Figure 18: Bagging Ensemble Models for Discrete Classification and Numeric Regression

After inspecting the strongest climate attribute-to-CloudCover_median associations, I selected 2, SkyCode_median and WindSpd_median for all 46 years, and ran their Normalized [0.0, 1.0] values through three ML algorithms to predict BE_All.

MODEL 12
SimpleLinearRegression on WindSpd_median
BE_All = -424.17 * WindSpd_median + 478.87, Predicting 0 if attribute value is missing.
Correlation coefficient 0.6157
Mean absolute error 101.5793
Root mean squared error 133.2505

MODEL 13
LinearRegression Model
BE_All =
   228.9845 * SkyCode_median +
  -422.6738 * WindSpd_median +
   379.3568
Correlation coefficient   0.6654
Mean absolute error   99.7265
Root mean squared error   126.3827

MODEL 14
M5 pruned model tree:
LM num: 1
BE_All =
   228.9845 * SkyCode_median
   - 422.6738 * WindSpd_median
   + 379.3568
Number of Rules : 1
Correlation coefficient   0.6654
Mean absolute error   99.7265
Root mean squared error   126.3827

These CCs are less than the CC of 0.8702 at the top of Figure 15 mostly because CloudCover data start only in 2006 and so SimpleLinearRegression uses the consistently low mean value of BE_All across the early years before CloudCover. SkyCode and WindSpeed span all 46 years and so SimpleLinearRegression attempts to use those early values when BE numbers had not begun their dramatic recovery. Limiting the data to the 16 years 2006 through 2021 reduces the CC because the peaks and valleys come to dominate the error differences. LinearRegression and M5P both come in with CC=0.2534. SimpleLinearRegression does a little better with 16 years of normalized data.

MODEL 15
SimpleLinearRegression on WindSpd_median
BE_All = -222.75 * WindSpd_median + 499.18
Predicting 0 if attribute value is missing.
Correlation coefficient   0.4255
Mean absolute error   104.9936
Root mean squared error   115.1737

Much of this analysis has to do with finding reasonable substitutes for CloudCover, since its data begins only at 2006. Here is an imperfect but useful model for estimating CloudCover during the years prior to 2006, using normalized values for SkyCode and WindSpd. SkyCode is similar to CloudCover and so correlates positively. WindSpd consistently correlates negatively, indicating fewer clouds and fewer bald eagles on very windy days.

MODEL 16
LinearRegression Model
CloudCover_median =
     0.5721 * SkyCode_median +
    -0.1475 * WindSpd_median +
     1.2329
Correlation coefficient   0.5964
Mean absolute error   0.1603

Root mean squared error	0.2156

It appears these models are using CloudCover and SkyCode primarily as **time keepers**. BE_All goes up as they go up because BE_All has increased, mostly due to DDT elimination, in the same years that they have increased. The consistent yearly decrease in WindSpd illustrated in Figure 12, having negative multipliers in Models 12 through 16, may also serve as a time keeper. These attributes correlate temporally with BE_All but they do not cause its consistent increase. However, Figure 17 below Model 7 shows BE_All peaks and valleys aligning with those of CloudCover_median, indicating a second-order, positive correlation. The red-tailed hawk analysis of Figure 20 below indicates that CloudCover and WindSpd peak values correlate positively with RT_All on a per-year basis, not merely consistently across years, thereby differing from the time keeper interpretation. It is important to note that WndSpd_median in Figure 20 correlates positively with RT_All, while Models 12 through 16 show negative correlation with BE_All, so for BE_All it is primarily a time keeper. In summary, negative correlation of WindSpd with BE_All and other increasing raptor counts serves primarily as a time keeper, while positive correlation of WindSpd with raptor counts that includes aligned per-year peaks in both serves as genuine climate correlation.

The climate-raptor analyses to this point concentrating on BE_All and CloudCover attributes constitute **bottom-up analyses**, a.k.a. **data-driven analyses**, because they use the data to drive detection of correlations. A complementary approach uses **top-down analyses**, a.k.a., **goal-driven analyses**, to search for anticipated correlations. The latter approach requires problem domain expertise, in this case the expertise of Dr. Goodrich. The next round of analyses are guided by Dr. Goodrich's suggestions and requests.

On July 20, 2022 Dr. Goodrich emailed me: "The RTHA and SSHA might be good to start with on raptor species. BW is such an anomaly as they push through more with date in mind than weather but still worth a look. Also I think our biggest flights are on easterly winds of some kind as it pushes the kettles off the piedmont plains to mountains. I might do **total raptors, rtha, ssha, and kestrel** to start. And maybe **golden eagles (GE_All) for a late season** bird and **bald eagle (BE_All). BW** as well but will show different pattern I suspect." RTHA are red-tail hawks (**RT_All** in the data); SSHA are sharp-shinned hawks (**SS_All**); American kestrel (**AK_All**); broad-wing hawks which normally have very high peak numbers on the peak day (**BW_All**).

Dr. Goodrich had previously emailed on March 23, 2022: "To-date, studies on raptor migration and climate change have mostly focused on documenting changes in timing of migration (e.g.,Therrien et al 2018) or changes in wintering latitude (e.g., Paprocki et al. 2017). However, trends in raptor counts at watch sites show unexplained declines in numbers of some species, most notably at Hawk Mountain, Pennsylvania, the longest-running migration count in the Americas. The Raptor Population Index analyses (https://www.rpi-project.org) show significant declines in Sharp-shinned and **Cooper's hawks**, red-tailed hawks, **osprey**, and a variety of other species. All of these species are known to concentrate along Appalachian ridges under certain wind conditions and have peak flights associated with cold front passage (Maransky et al 1997, Allen et al 1996,,and others). The strength of winds has also been shown to be correlated with numbers of eagles sighted at Hawk Mountain (Laurie Goodrich, Hawk Mountain News, pers. comm).

There is widespread concern about declines in Broad-winged Hawk, Sharp-shinned Hawk and other migrants (rpi-project.org; species assessments). If climate change is changing local weather patterns as some authors suggest (e.g. frequency of fronts or strong winds) (REF?) , raptor migrants may be shifting their migration strategy and showing declining counts at watch sites. Because most raptors are not monitored well in breeding season, understanding the trends observed in migration counts is a high priority. Because Hawk Mountain has the best long running dataset, it may be the best source of understanding how local weather may have changed and how that change may effect raptor numbers sighted. A longitudinal study of weather at the site and raptor sightings can confirm weather factors associated with large counts of raptors, by species and if the lower counts are related to changes in weather factors.

Please add that the trends in raptor numbers are well detailed at www.rpi-project.org and significant ten year (2990-2019) and long term declines are being observed in osprey, **northern harrier**, sharp-shinned hawk, Coopers hawk, **northern goshawk**, red-tailed hawk, **red-shouldered hawk**, broad-winged hawk, **rough-legged hawk**, American kestrel. The fact that most species of raptors are showing declines suggests either that most are short-stopping north of Hawk Mountain or weather changes are occurring to influence concentrations, or that all are declining. Christmas Bird Counts show some states and provinces may have declining populations in some species but many are stable. Therefore, we are eager to examine if migration behavior could be changing due to shifts in winds or other weather variables."

The following climate -> raptor analyses appear below with links here: Red-Tailed Hawk, Sharp-Shinned Hawk, American Kestrel, Bald Eagle, Golden Eagle, Broad-Winged Hawk, Total Raptors. Follow-up analyses by graduate students during the 2022-2023 academic year include Cooper's Hawk, Osprey, Northern Harrier, Northern Goshawk, and Rough-Legged Hawk.

Red-Tailed Hawks (RT)

The Linear Regression model considered first correlates all North Lookout and Allentown Airport climate measures to RT_All, the sum of RT counts for each year, using Linear Regression from 1976 through 2021 on all attributes normalized to the range [0.0, 1.0] to non-normalized RT_All counts including yearSince1976. As explained previously, normalization puts all expression multipliers on a single scale for comparison of magnitude. Here are the accuracy measures.

MODEL 17

Correlation coefficient	0.511
Mean absolute error	835.4733
Root mean squared error	1125.2418

LinearRegression.txt including yearSince1976 and all attributes except RT_All normalized into the range [0.0, 1.0]
RT_All = 1017.3812 * HourlyDryBulbTemperature_24_min + 735.8144 * CloudCover_72_pstdv + -671.6571 * HMtempC_24_min + 648.763 * HourlyRelativeHumidity_pstdv + 643.8234 * noaawdNW + -617.0224 * wndW + 475.3436 * HourlyWindSpeed_pstdv + -474.6915 * HourlyRelativeHumidity_72_pstdv + 470.9802 * HourlyWetBulbTemperature_72_median + 460.529 * HourlyRelativeHumidity_48_mean + -450.9171 * dryTempC24pstdev + -447.6376 * HourlyRelativeHumidity_24_min + -444.9649 * HourlyRelativeHumidity_24_median + 443.9384 * HourlyStationPressure_24_median + -422.2674 * **yearSince1976** + 405.2646 * HourlyRelativeHumidity_min + -403.6422 * HourlyPrecipitation_mean + -403.5317 * HourlyDewPointTemperature_48_min + 402.4825 * CloudCover_max + 400.3726 * Visibility_max + 397.9899 * HourlyStationPressure_24_mean + 391.5557 * CloudCover_48_pstdv + 378.2416 * HMtempC_max + -377.5195 * HourlyStationPressure_max + -373.0364 * HourlyStationPressure_min + -369.9472 * HourlyDewPointTemperature_min + 369.756 * dryTempC24maxFullNOAA + 366.4538 * HourlyStationPressure_72_median + -361.6279 * HMtempC_pstdv + 356.8084 * wndSSW + -348.6117 * noaawdENE + -347.3991 * Visibility_24_max + 345.1377 * CloudCover_24_median + -343.8892 * HourlyStationPressure_72_max + 336.1095 * HMtempC_48_max + 323.4141 * WindSpd_median + 323.3924 * HourlyStationPressure_72_mean + -319.056 * Visibility_24_pstdv + 316.1437 * HourlyRelativeHumidity_48_median + 308.9925 * HourlyDryBulbTemperature_48_pstdv + -307.1109 * HourlyWetBulbTemperature_max + 305.5581 * HourlyDewPointTemperature_48_pstdv + -304.0721 * HourlyDryBulbTemperature_72_median + 298.523 * HourlyRelativeHumidity_48_pstdv + -292.3686 * SkyCode_median + -287.964 * CloudCover_24_min + 286.0883 * HMtempC_median + -285.9316 * Visibility_72_min + -279.8569 * HourlyPrecipitation_pstdv + 276.188 * HourlyWindSpeed_max + 272.6111 * HMtempC_24_pstdv + 271.1781 * HourlyStationPressure_24_max + -269.0669 * CloudCover_mean + -268.2325 * HourlyDryBulbTemperature_72_min + 263.4729 * HourlyStationPressure_median + -255.7089 * Visibility_24_mean + -251.2272 * HourlyDewPointTemperature_72_min + -250.9826 * HourlyDewPointTemperature_48_median + -250.3956 * noaawdESE + 249.0219 * HourlyDewPointTemperature_48_max + -248.9752 * HourlyWetBulbTemperature_min + 243.8099 * HourlyDryBulbTemperature_72_pstdv + -243.0199 * HourlyDewPointTemperature_mean + -242.427 * wndS + -241.2425 * HMtempC_min + -236.1699 * HourlyDewPointTemperature_24_mean + -235.2734 * HMtempC_24_mean + -234.436 * HourlyRelativeHumidity_median + 233.574 * HourlyStationPressure_pstdv + -232.0542 * wndNE + 230.3717 * wndNNW + -226.5066 * Visibility_48_pstdv + 220.0377 * dryTempC24medianFullNOAA + -219.8205 * HourlyRelativeHumidity_72_max + -218.5341 * Visibility_48_min + 217.1015 * HourlyWindSpeed_mean + 214.3785 * HourlyDryBulbTemperature_48_max + -212.975 * HourlyDryBulbTemperature_24_pstdv + -212.8371 * HourlyRelativeHumidity_72_min + 212.2967 * Visibility_48_median + -210.3944 * HourlyWetBulbTemperature_pstdv + -207.0666 * HourlyDewPointTemperature_24_min + 193.474 * noaawdWNW + -181.5483 * noaawdN + -181.3676 * HourlyDryBulbTemperature_pstdv + 180.5762 * HourlyWindDirection + -171.1405 * HourlyRelativeHumidity_72_mean + 168.381 * HourlyDryBulbTemperature_48_min + 167.8806 * wndESE + 165.8407 * SkyCode_mean + 165.7518 * HMtempC_72_median + 164.6565 * noaawdNNW + 157.6631 * dryTempC24pstdevFullNOAA + -156.2897 * Visibility_72_median + 155.5467 * HourlyWindSpeed_median + -154.8183 * HourlyDewPointTemperature_24_pstdv + 136.4734 * noaawdSSE + -123.26 * Visibility_min +
3475.4356

I have given this full model just to point out its complexity. I sometimes tell my students that, if you are interested in understanding a model and not strictly in using it, data mining the model can be as complicated as data mining the problem domain data. The above Linear Regression is an example.

MODEL 18

Next comes **Simple Linear Regression** (one attribute mapped to RT_All) with these accuracy measures.

Correlation coefficient	0.7048
Mean absolute error	655.4128
Root mean squared error	818.2921

SimpleLinearRegression including yearSince1976 and all attributes except RT_All normalized into the range [0.0, 1.0].
RT_All = 2335.57 * CloudCover_72_pstdv + 1162.63, Predicting 3778.17 if attribute value is missing.

RT_All = 3776.17 is simply the mean of annual counts for 1976 through 2005, the 30 missing years for HM CloudCover data. **CloudCover_72_pstdv** is the population standard deviation of the unsigned magnitude of CloudCover per-day changes from the measurement 72 hours earlier. All such statistical measures _24, _48, and _72 including _mean, _median, and _pstdv aggregate the unsigned magnitude of day-to-day changes because _mean and _median converge to 0 when retaining the sign. The _min and _max statistics retain the up-or-down numeric sign.

MODEL 19

The **M5P** model tree excluding yearSince1976 with its size constrained by configuration parameters gives a readable model using normalized climate attributes. These are its accuracy measures.

```
Correlation coefficient                0.5204
Mean absolute error                  813.005
Root mean squared error                978.7382
```

M5P pruned model tree not including yearSince1976 and all attributes except RT_All normalized into the range [0.0, 1.0]

```
CloudCover_max <= 0.969 :
|   noaawdSSE <= 0.427 : LM1 (14/20.367%)
|   noaawdSSE >  0.427 :
|   |   HourlyStationPressure_max <= 0.757 : LM2 (13/35.092%)
|   |   HourlyStationPressure_max >  0.757 : LM3 (4/33.336%)
CloudCover_max >  0.969 : LM4 (15/47.871%)
LM num: 1
RT_All =
    968.2647 * HourlyStationPressure_pstdv
    + 553.5098 * HourlyDryBulbTemperature_min
    - 912.0292 * HourlyStationPressure_max
    + 807.4307 * noaawdSSE
    + 583.8608 * HourlyRelativeHumidity_48_mean
    + 2637.6208
LM num: 2
RT_All =
    342.9928 * HourlyStationPressure_pstdv
    + 293.2709 * HourlyDryBulbTemperature_min
    - 1303.9584 * HourlyStationPressure_max
    + 769.3624 * noaawdSSE
    + 1189.6687 * HourlyRelativeHumidity_48_mean
    + 3204.6518
LM num: 3
RT_All =
    342.9928 * HourlyStationPressure_pstdv
    + 293.2709 * HourlyDryBulbTemperature_min
    - 1502.825 * HourlyStationPressure_max
    + 769.3624 * noaawdSSE
    + 996.1399 * HourlyRelativeHumidity_48_mean
    + 3297.4794
LM num: 4
RT_All =
    -3416.1578 * HourlyWetBulbTemperature_mean
    - 942.1215 * HourlyStationPressure_max
    + 615.4329 * noaawdSSE
    + 895.2533 * HourlyRelativeHumidity_48_mean
    + 5300.7813
```

Number of Rules : 4

There are a few important things to note about this readable M5P model of climate -> RT_All. First, it is using a derived measure of CloudCover again. The closest HM attribute to CloudCover is SkyCode as documented here.

SKY CODES
0 Clear; 0-15% cloud cover
1 Partly cloudy; 16-50%cover
2 Mostly cloudy; 51-75% cover
3 Overcast; 76-100% cover
4 Wind-driven sand, dust,snow
5 Fog or haze
6 Drizzle
7 Rain
8 Snow
9 Thunderstorm, with or without precipitation

SkyCode does not have the resolution of CloudCover which is in the range [0.0, 100.0] percent. It is easier to estimate a cloud cover percentage than to pigeon hole the cloud cover percentage into the ranges listed for SkyCode. It is easier to make mistakes in recording SkyCode. **My suggestion is to ensures cloud cover measures in all future observations, given its importance to analysis despite 30 years of missing data out of 46 years**.

MODEL 20

Second, M5P's decision tree makes its primary decision based on Cloud_Cover_max for each year. For the 16 years recorded out of 46, 15 had Cloud_Cover_max of 100% and 1 had a value of 98%, giving a mean of 99.875 and a median and mode of 100% for years with Cloud_Cover data. The M5P test "CloudCover_max <= 0.969" will never pass, given a minimum Cloud_Cover_max of 0.98 for one year. This appears to be a dead-wood artifact of removing year and yearSince1976 as attributes from model construction in the interest of using climate-only data. Leaving yearSince1976 in the climate data results in this M5P model with these accuracy measures.

```
Correlation coefficient              0.6899
Mean absolute error                668.8454
Root mean squared error              849.0912
M5 pruned model tree:
(using smoothed linear models)
```

M5P pruned model tree including yearSince1976 and all attributes except RT_All normalized into the range [0.0, 1.0]

```
yearSince1976 <= 0.678 :
|   HourlyRelativeHumidity_24_max <= 0.495 :
|   |   HourlyRelativeHumidity_24_mean <= 0.406 : LM1 (13/28.313%)
|   |   HourlyRelativeHumidity_24_mean >  0.406 : LM2 (7/27.365%)
|   HourlyRelativeHumidity_24_max >  0.495 : LM3 (11/46.805%)
yearSince1976 >  0.678 : LM4 (15/33.019%)
LM num: 1
RT_All =
    -915.5041 * yearSince1976
    + 124.8347 * WindSpd_median
    - 323.6489 * HourlyWetBulbTemperature_mean
    - 631.293 * HourlyRelativeHumidity_24_mean
    + 1063.9538 * HourlyRelativeHumidity_24_max
    + 3847.3844
LM num: 2
RT_All =
    -915.5041 * yearSince1976
    - 353.1478 * WindSpd_median
    - 323.6489 * HourlyWetBulbTemperature_mean
    - 718.3598 * HourlyRelativeHumidity_24_mean
    + 1063.9538 * HourlyRelativeHumidity_24_max
    + 4140.3149
LM num: 3
RT_All =
    -915.5041 * yearSince1976
    - 475.3912 * WindSpd_median
    - 323.6489 * HourlyWetBulbTemperature_mean
    - 420.0649 * HourlyRelativeHumidity_24_mean
    - 596.0296 * HourlyDryBulbTemperature_72_median
    + 1280.1403 * HourlyRelativeHumidity_24_max
    + 4733.1695
LM num: 4
RT_All =
    -1403.7729 * yearSince1976
    - 3677.8552 * HourlyWetBulbTemperature_mean
    + 673.7698 * HourlyRelativeHumidity_24_max
    + 6059.9419
Number of Rules : 4
```

The "yearSince1976 <= 0.678" test picks up all years 1976 through 2005 without CloudCover data plus 2006. The highest-magnitude multiplier "-1403.779" for yearSince1976 in expression 4 for 2007-2021 reflects the more rapid decline in RT_All in the final 15 years of the data. Expression LM4 for both models give negative weight to HourlyWetBulbTemperature_mean, reflecting the correlation between rising temperatures and falling RT_All counts.

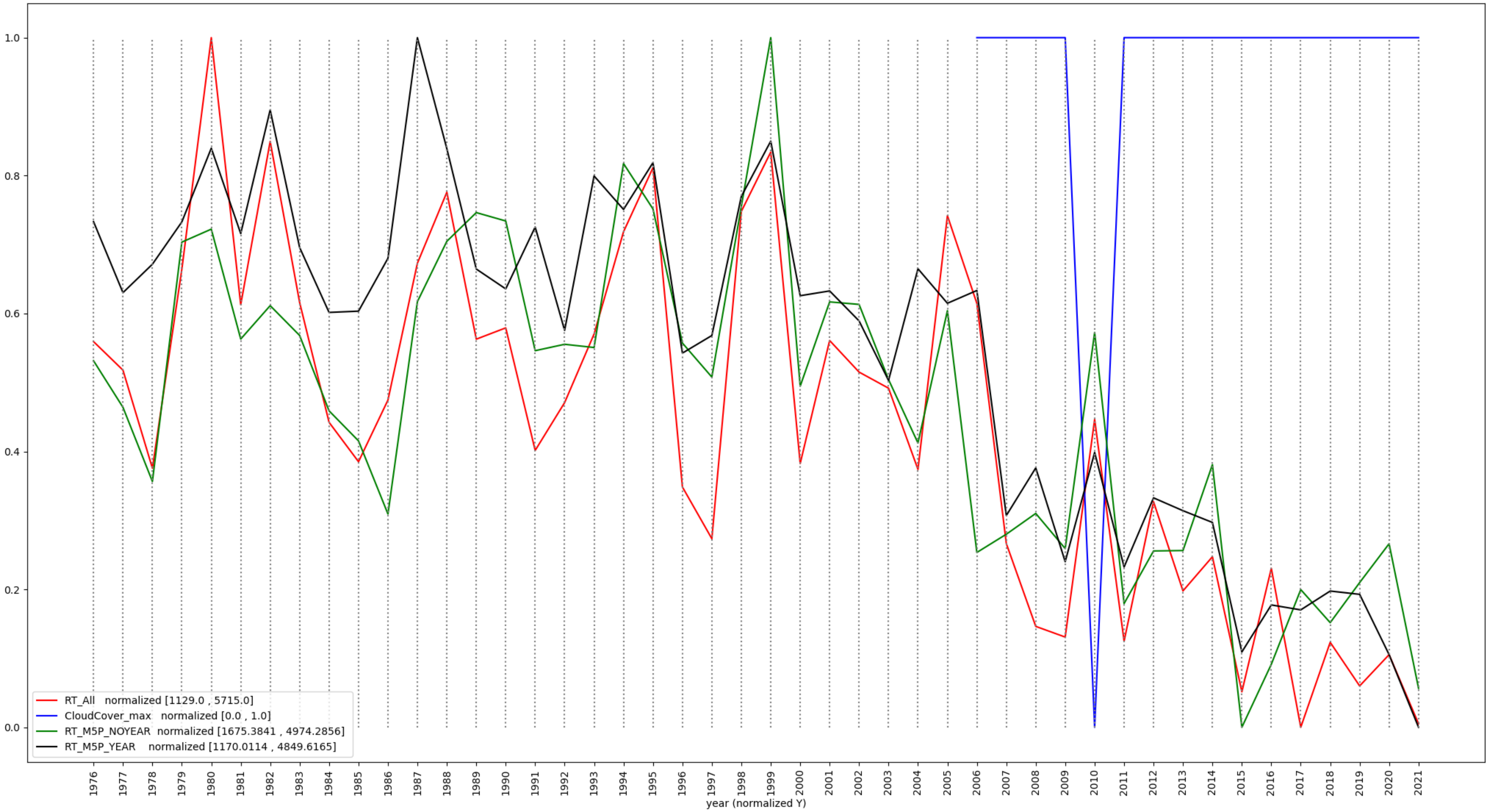


Figure 19: The Two Most Recent M5P models (Models 19 and 20) for RT_All above Including NOAA data With and Without yearSince1976

Figure 19 shows normalized (with [lower, upper] actual data range in the legend) RT_All counts in red, CloudCover_max at [98%, 100%] before normalization, the above M5P model excluding year timestamps with a CC of 0.5204 in green, and the most recent M5P model including year timestamps with a CC of 0.6899 in black. The way to interpret these graphs is to look for the peaks and the valleys (a.k.a. troughs) in the red RT_All time series and align them visually with peaks and valleys in the M5P models. RT_M5P_NOYEAR aligns well with RT_All in some years, RT_M5P_YEAR aligns well in some, and both align well in some years. A perfect model would superimpose on the RT_All red line. Neither does.

A CC of 0.6869 for the black-line model is respectable but hard to interpret. There are a lot of terms in that decision tree and those expressions, including a lot of Allentown Airport data. With these thoughts in mind, I spent some time looking at RT-correlation graphs using only the HM wind and cloud cover observation data graphed in Section 2. Figure 20 illustrates the highlights. From 2013 through 2021 the peaks and valleys for RT_All, CloudCover_median, WindSpd_median, and wndNW (wind from the NW tally) line up almost perfectly. The blue, green, and black lines in Figure 20 are normalized observation counts, not models.

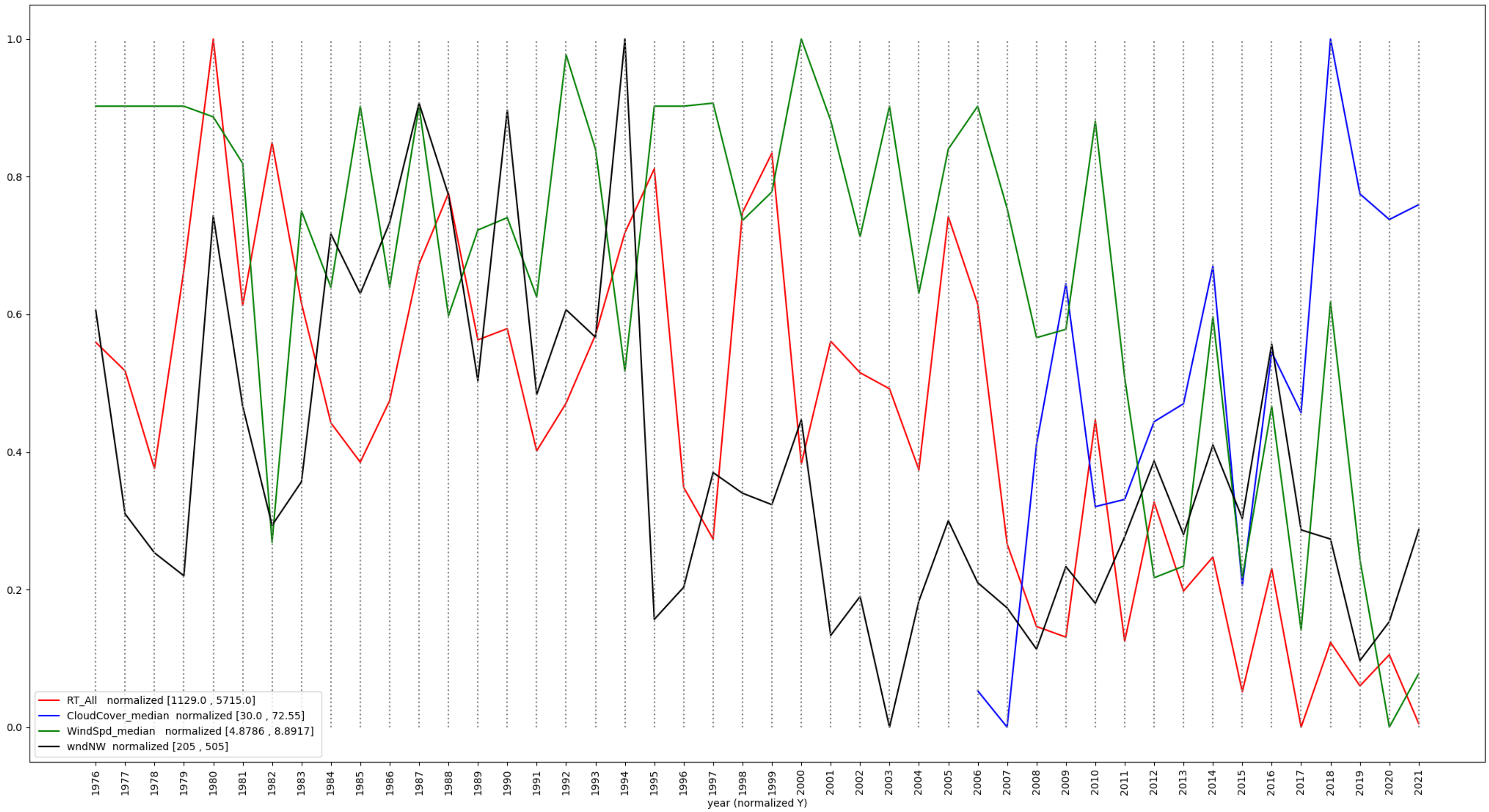


Figure 20: Annual RT_All Tallies in red and associated normalized CloudCover_median, WindSpd_median, and wndNW North Lookout Measures

In addition to peak and valley alignments by year, Figure 21 shows declining parallel slope alignment of RT_All and WindSpd_median for 5-year running averages.

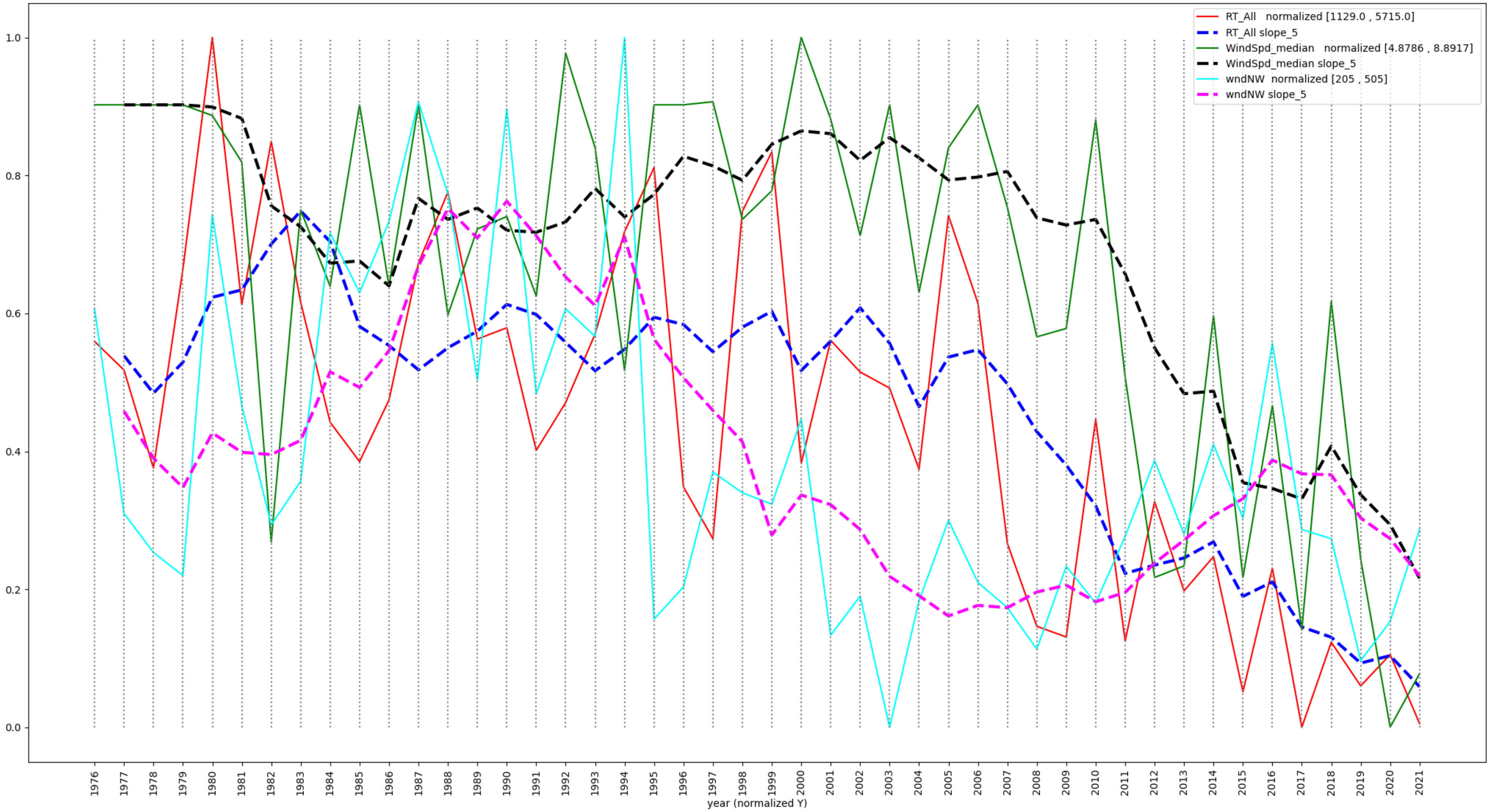


Figure 21: Parallel Declines in RT_All and WindSpd_median since 2000 and Near-Parallel Declines in wndNW tallies using rolling 5-year averages

The annual increase in HM temperature in Figure 22 has a much less pronounced 5-year slope that those of Figure 21 and RT_All of Figure 22. Figures 3 through 5 of Section 2 illustrate an average annual HM increase of 0.02 degrees C. Allentown Airport shows a sharper annual increase of 0.03 degrees C for the Allentown Heat Island, which is still a gradual increase over its range.

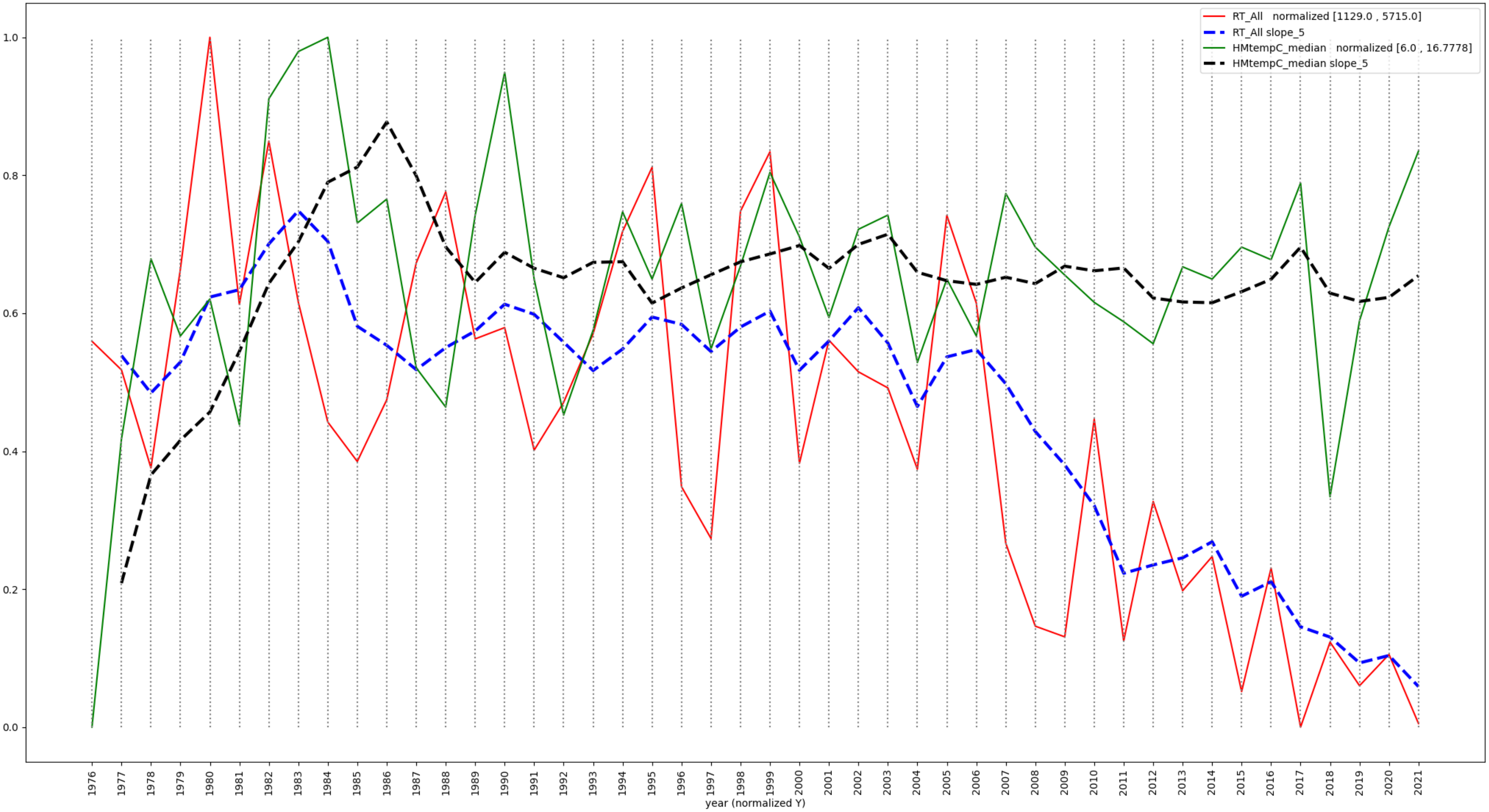


Figure 22: 5-Year Rolling Averages for RT_All tallies and Temperature Celsius measurements at North Lookout

This is about as close as this study is likely to come to uncovering causation in addition to correlation, at least for RT_All tallies. Mean temperature is not increasing at the rate of RT_All decline at North lookout, while WindSpd_median and wndNW are. Furthermore, the peaks and valleys of these wind attributes align well with the annual peaks and valleys of RT_All. I defer to the experts, but clearly we do not have the stronger NW and NNW wind tunnel to North Lookout that existed in the past, indicating a stronger effect for diverging RT migration paths than for wintering further north.

Below are three linear models for the North Lookout attributes **yearSince1976, HMtempC_median, CloudCover_median, WindSpd_median, wndENE, wndWNW, wndNW, wndNNW**, and **RT_All**, with all except RT_All normalized to [0.0, 1.0] for their respective [min, max] values from 1976 through 2021. These models do no use HMtempC_median, wnENE, or wndWNW as factors, despite the substantial growth in WNW since 1994 seen in Figure 9. Basically, wind directions other than NW and NNW, including wndUNK, seem to act as random, uncorrelated directions for RT flight. Attribute **wndUNK** indicates "no direction", indicative of no wind or mixed wind directions (likely mild breezes) during the observation window.

MODEL 21

Simple linear regression on CloudCover_median

```
RT_All = -1453.92 * CloudCover_median + 2721.9, Predicting 3778.17 if attribute value is missing.
Correlation coefficient               0.73
Mean absolute error                 632.78
Root mean squared error               782.7286
```

MODEL 22

Linear Regression Model

```
RT_All =
  -3634.4268 * yearSince1976 +
   2270.8468 * wndWNW +
   1348.4813 * wndNW +
   3778.1611
Correlation coefficient               0.6668
Mean absolute error                 710.4056
Root mean squared error               869.1728
```

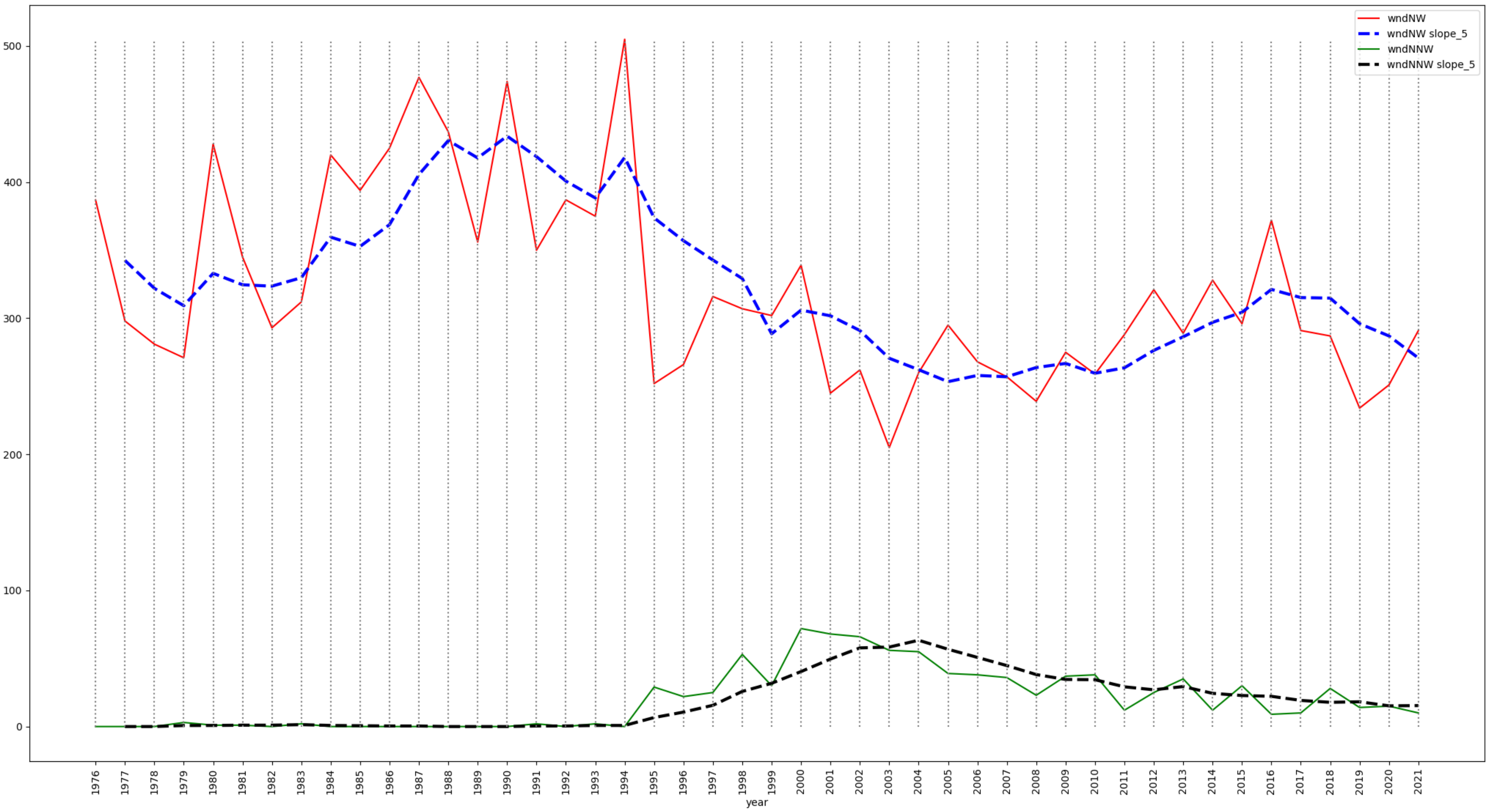
MODEL 23

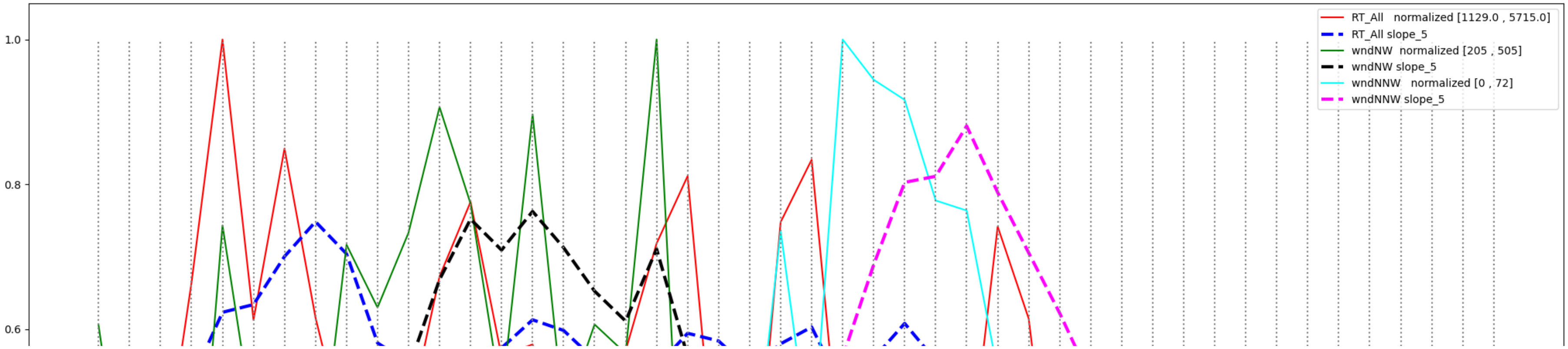
M5 pruned model tree:

```
yearSince1976 <= 0.678 : LM1 (31/68.389%)
yearSince1976 >  0.678 : LM2 (15/36.876%)
LM num: 1
RT_All =
    -1211.0951 * yearSince1976
    + 518.7268 * wndWNW
    + 4032.3638
LM num: 2
RT_All =
    -1857.0125 * yearSince1976
    + 686.2995 * WindSpd_median
    + 795.381 * wndWNW
    + 2935.1756
Number of Rules : 2
Correlation coefficient               0.7106
Mean absolute error                 657.7596
Root mean squared error               805.8803
```

The normalized yearSince1976 value of 0.678 corresponds to the boundary between 2006 and 2007. M5P expression LM1 applies to 1976-2006, and LM2 applies to 2007-2021. Declining wind speed contributes to the more recent LM2, the yearSince1976 slope falls off faster, and wndWNW carries more weight than wndNW in these formula because of its

larger multiplier. The top graph in Figure 23 shows that wndNNW is falling more consistently than wndNW since its peak in 2004. The bottom graph shows the 5-year running average slope in normalized wndNNW more closely matches the slope of decline of RT_All than does wndNW, hence the tighter correlation in these models.


wndNW
wndNW slope_5
wndNNW
wndNNW slope_5
500
400
300
200
100
0
1976
1977
1978
1979
1980
1981
1982
1983
1984
1985
1986
1987
1988
1989
1990
1991
1992
1993
1994
1995
1996
1997
1998
1999
2000
2001
2002
2003
2004
2005
2006
2007
2008
2009
2010
2011
2012
2013
2014
2015
2016
2017
2018
2019
2020
2021
year


RT_All normalized [1129.0 , 5715.0]
RT_All slope_5
wndNW normalized [205 , 505]
wndNW slope_5
wndNNW normalized [0 , 72]
wndNNW slope_5
1.0
0.8
0.6

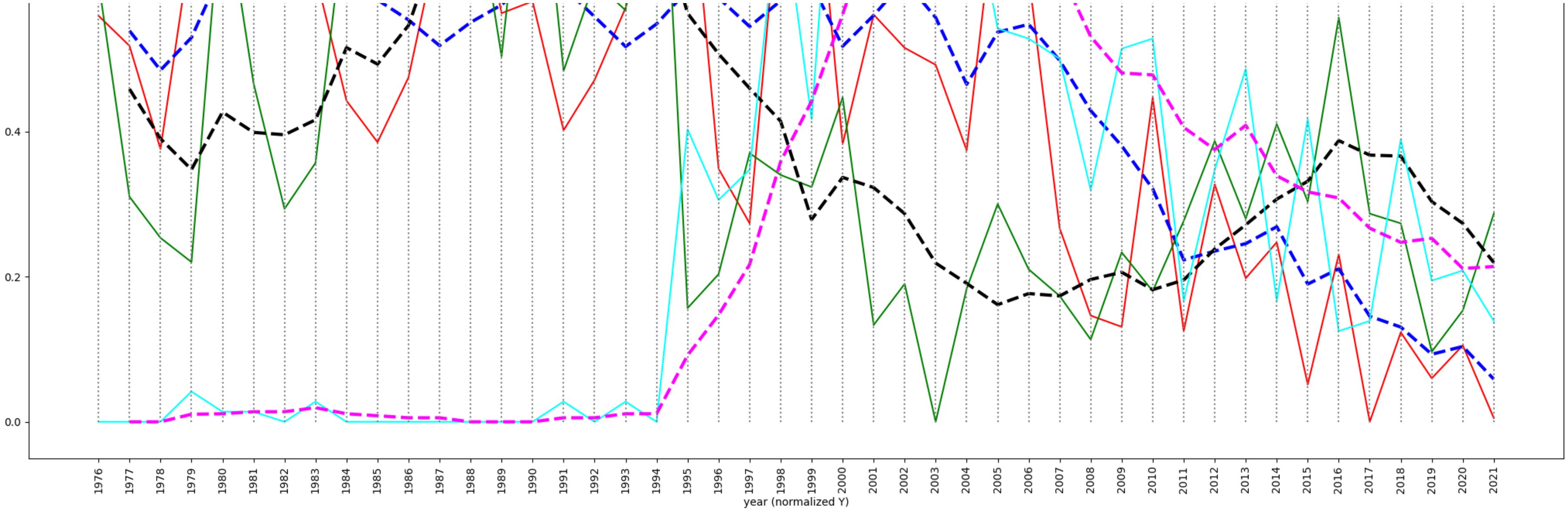


Figure 23: Trends in wndNW and wndNNW and the stronger correlation of the latter's 5-year running average decline to RT_All.

Finally for RT, Figure 24 shows that the day-of-year of the first sighting (_1st in the legend), days of the 25%, 50%, and 75% of the red-tails (_25th, _50th, and _75th), and the day of the peak count (_peak) have not changed incrementally since 1976. RT_last, the day of the last sighting for the year, and its 5-year running average show an increase that has tailed off (pardon the pun).

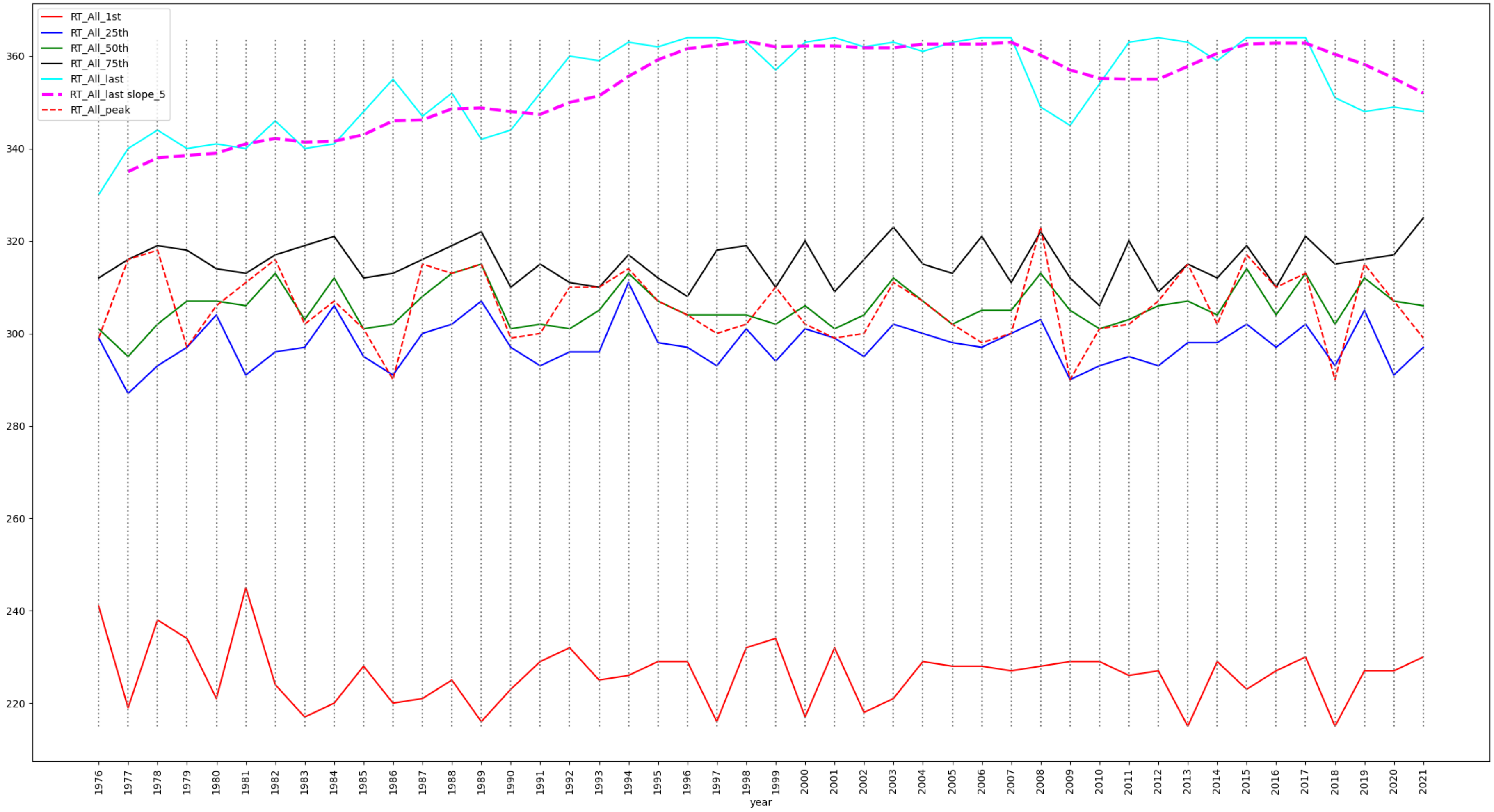


Figure 24: Day-of-Year for First, 25%, 50%, and 75% of RT Count Sightings, Day of Final Sighting, and Day of Peak Sighting for the Year.

Sharp-Shinned Hawks (SS)

Before considering models let us take a look at graphs of primary climate measures from North Lookout examined for red-tail hawks. Figure 25 shows that, unlike RT_All's parallel slope with wndNNW at the bottom of Figure 23, SS_All's decline closely parallels that of wndNW. HMtempC_median (not shown) does not show as steep of an uphill slope as any downhill in Figure 25. Most (but not all) peaks and valleys of SS_All and wndNW align as well

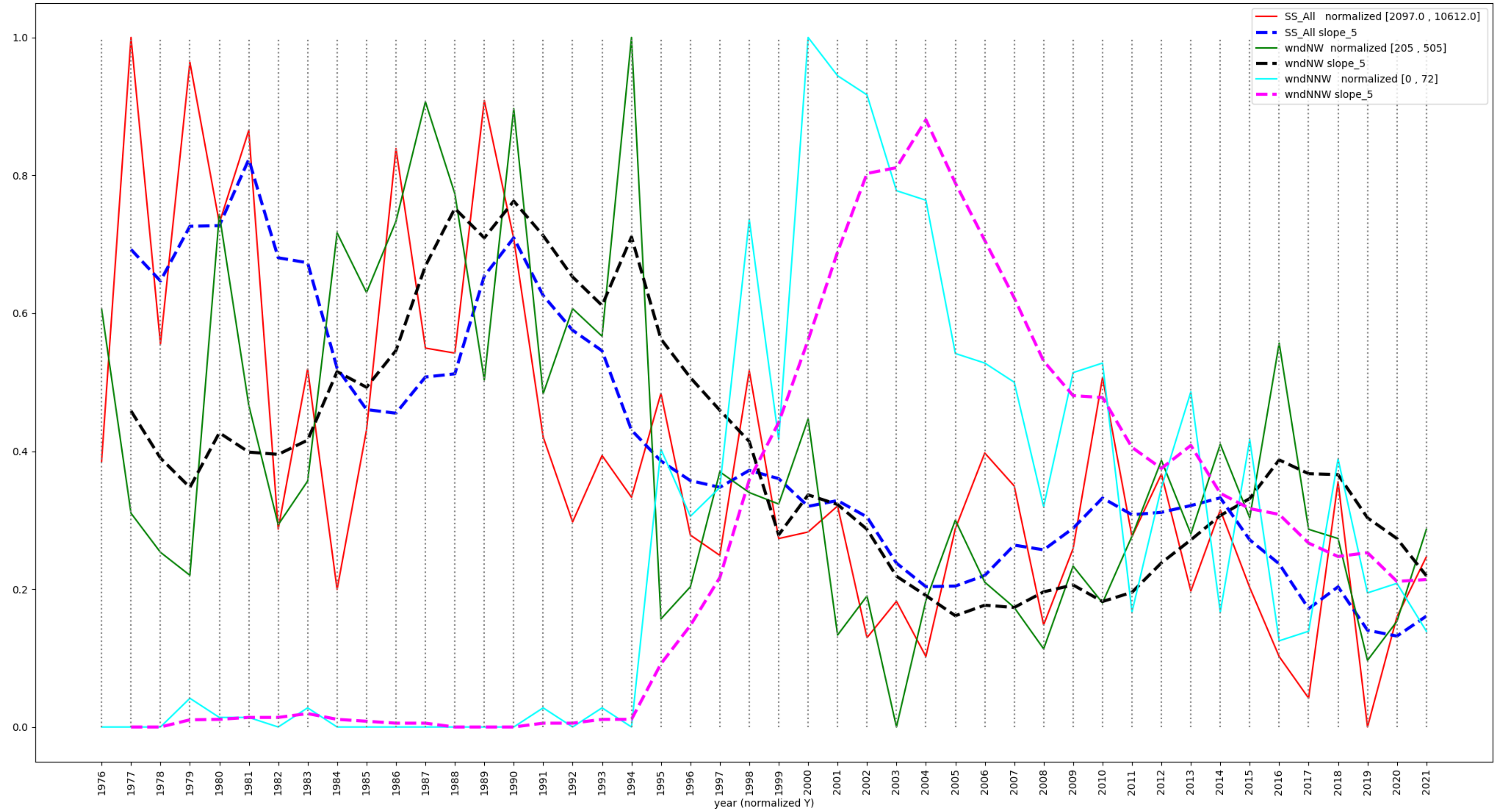
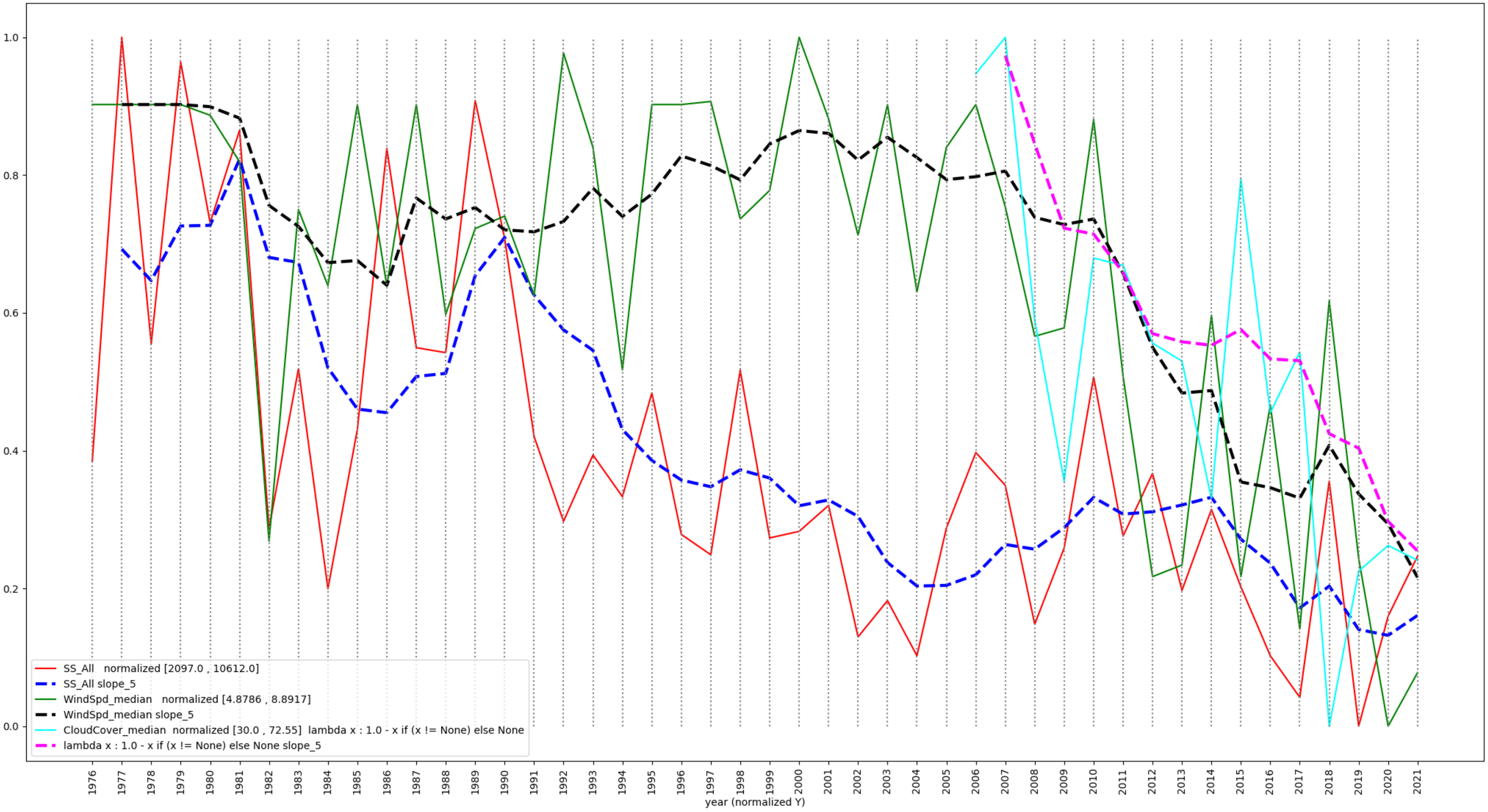
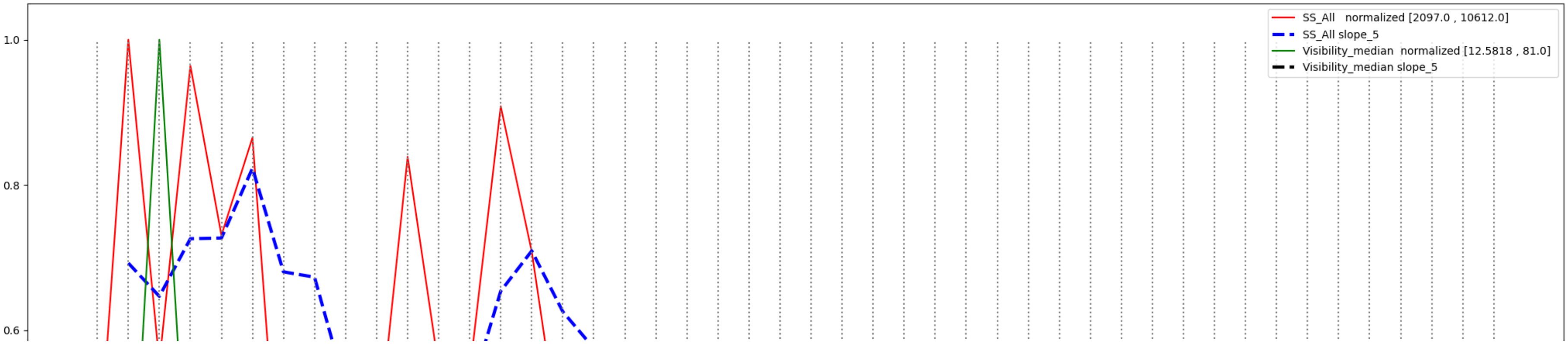


Figure 25: SS_All's 5-year Running Average Decline Tracks wndNW Decline - they converge with wndNNW around 2014

The top graph in Figure 26 shows the 5-year average slope decline of SS_All tracking the inverse-decline of CloudCover_median, converging with the declining slope of WindSpd_median at 2014. The inverse-decline of normalized CloudCover_median is (1.0 - its normalized value). This re-normalization inverts the direction of the CloudCover_median slope so the min value of 30.0% is at the top of the slope and the max value of 72.55% is at the bottom. Substituting SkyCode_median for CloudCover_median does not align as well because of inconsistencies in that low-resolution measure.


1.0
0.8
0.6
0.4
0.2
0.0
SS_All normalized [2097.0 , 10612.0]
SS_All slope_5
WindSpd_median normalized [4.8786 , 8.8917]
WindSpd_median slope_5
CloudCover_median normalized [30.0 , 72.55] lambda x : 1.0 - x if (x != None) else None
lambda x : 1.0 - x if (x != None) else None slope_5
1976
1977
1978
1979
1980
1981
1982
1983
1984
1985
1986
1987
1988
1989
1990
1991
1992
1993
1994
1995
1996
1997
1998
1999
2000
2001
2002
2003
2004
2005
2006
2007
2008
2009
2010
2011
2012
2013
2014
2015
2016
2017
2018
2019
2020
2021
year (normalized Y)


1.0
0.8
0.6
SS_All normalized [2097.0 , 10612.0]
SS_All slope_5
Visibility_median normalized [12.5818 , 81.0]
Visibility_median slope_5

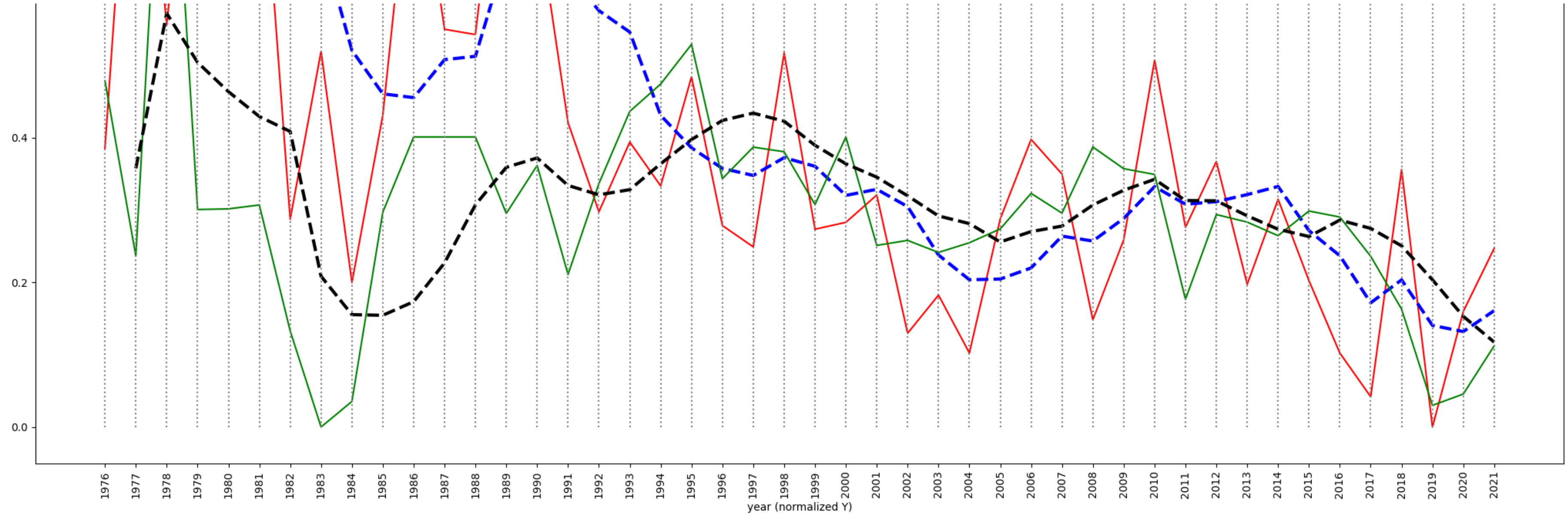


Figure 26: SS_ALL as a Function of WindSpd_median and Inverted CloudCover_median (top) and Visibility_median (bottom)

The bottom graph in Figure 26 shows the declining slope of SS_All matching that of Visibility_median. That decline hints at hard-to-see sharp-shinned hawks, but it is equally likely a function of increasing CloudCover. Peaks align only sometimes in the bottom of Figure 26, so CloudCover and Visibility may be acting as time keepers.

I examined three linear models for the North Lookout attributes yearSince1976, HMtempC_median, CloudCover_median, WindSpd_median, wndENE, wndWNW, wndNW, wndNNW, and SS_All, the same as used for RT_All analysis, with all but SS_All normalized to [0.0, 1.0]. Simple Linear Regression, Linear Regression, and M5P relied exclusively on correlating yearSince1976 to SS_All, so I removed yearSince1976 and tried again. Mapping only the year to the SS count amounts to extraction of the SS_All slope in Figure 26. The resulting models give almost the same CC.

MODEL 24

```
Instances:    46
Attributes:   8
              HMtempC_median
              CloudCover_median
              WindSpd_median
              wndENE
              wndWNW
              wndNW
              wndNNW
              SS_All
```

Simple Linear Regression on wndENE

```
SS_All = -4223.09 * wndENE + 6239.8, Predicting 0 if attribute value is missing.
Correlation coefficient                  0.358
Mean absolute error                   1576.7458
Root mean squared error                  1989.6326
```

MODEL 25

Linear Regression Model

```
SS_All =
   3924.1695 * WindSpd_median +
  -3321.1986 * wndNNW +
   3708.5498
Correlation coefficient                0.5636
Mean absolute error                 1410.895
Root mean squared error              1720.5199
```

MODEL 26

M5 pruned model tree:

```
(using smoothed linear models)
LM1 (46/73.697%)
LM num: 1
SS_All =
    2819.8671 * WindSpd_median
    - 1769.7587 * wndENE
    - 2736.3734 * wndWNW
    + 4702.091
Number of Rules : 1
Correlation coefficient                0.5656
Mean absolute error                 1408.2835
Root mean squared error              1717.2739
```

While these models do not explicitly use the wndNW -> SS_All correlation diagrammed in Figure 25, Linear Regression and M5P use WindSpd_median of the top of Figure 26 as the primary positive correlating attribute, with negative multipliers for wndNNW (Linear Regression), wndENE and wndWNW (M5P) to compensate when stronger winds are in non-wndNW directions. After getting these results I tried adding all 16 wind directions plus wndUNK back into the data and extracting these models again. They came out with lower accuracy and do not appear here. Non-wndNW wind directions act as noise in the models, presumably diverging SS flight away from North Lookout.

Figure 27 shows that the day-of-year of the first sighting (_1st in the legend), days of the 25%, 50%, and 75% of the sharp-shins (_25th, _50th, and _75th), and the days of the peak count (_peak) and final sighting (_last) have not changed incrementally since 1976.

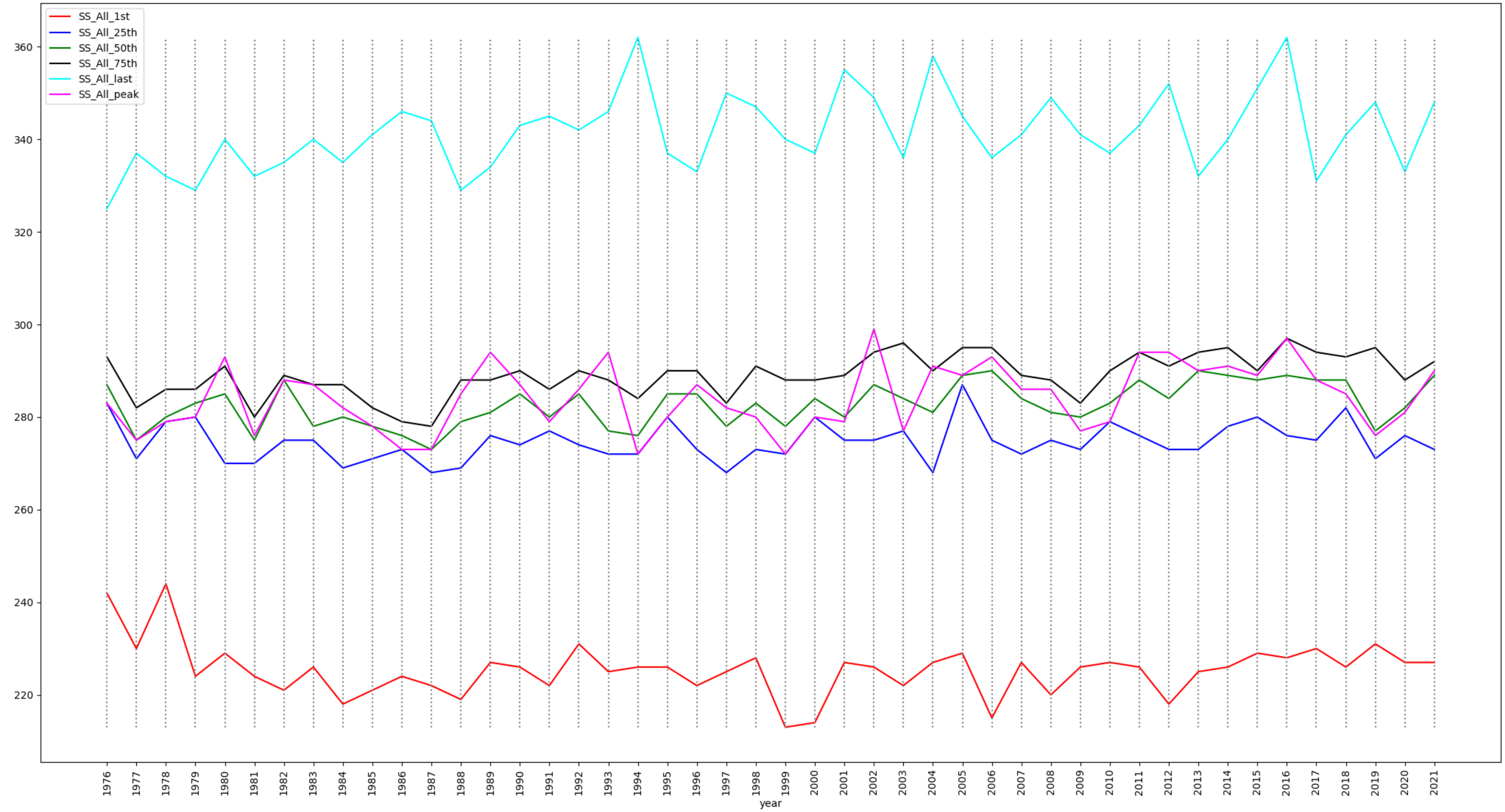


Figure 27: Day-of-Year for First, 25%, 50%, and 75% of SS Count Sightings, Day of Final Sighting, and Day of Peak Sighting for the Year.

Given the correlation of the declining wind speed at North Lookout to declining RT_All and SS_All tallies of Figures 21 and 26, it is necessary to do a literature search of recent research on declining wind speeds, a.k.a. Terrestrial Stilling (TS) in eastern North America.

"Plain Language Summary
The 10-m wind speed (NWS) is a key parameter in meteorology and climate science, whose changes could affect the natural environment and human society. Previous studies reported that the 10 m wind speed over land had weakened over the past several decades, a phenomenon termed terrestrial stilling (TS). However, recent studies based on observation and reanalyzes showed that the TS started to recover during the last decade, which had caused renewed interest in wind speed changes in the future. This study indicates that the NWS in the Northern Hemisphere (NH) mid-latitudes will likely continue to weaken during the 21st century forced by increased greenhouse gases. However, if the world cuts carbon emissions substantially, the decline of NWS will be interrupted and reversed after the mid-21st century. Future changes in NWS show seasonal differences, with the largest (smallest) decreasing trends of NWS in summer (winter). In this study, we propose that elevated upper-air warmings over the mid-latitudes could also play a key role in reducing the NWS. This study is useful for understanding the changes in hydrological cycles, air pollution, and wind energy development, particularly for countries in the NH mid-latitudes.

...
Previous studies have mainly attributed the surface TS to increased surface roughness/friction (e.g., Earth's greening and urbanization) and the amplified warming in high latitudes. Although these two drivers can to some extent explain the TS, they seem unsuccessful to explain the recent reversal of TS, and are difficult to account for the changes in the upper tropospheric winds." [18]

After the above analyses I added the **day** of the 1st, 25%, 50%, 75%, last, and peak count for all wind directions wndN through wndUNK, examined wind times, and looked for correlation to RT and SS counts and times. Nothing clear emerged.

American Kestrel (AK)

AK_All and related AK attributes do not correlate closely to climate change factors that correlate reasonably well with RT_All and SS_All counts discussed above. Uncovering the causes for declines in AK numbers in North America is an active area of research for raptor biologists. The following quotation is from a November 2020 summary. [19]

"Thankfully, some researchers are starting to break through the fog after analyzing prior data. The American Kestrel Partnership [20], a project of The Peregrine Fund [21], is one such group of scientists. Now, it believes the key to understanding the kestrel’s decline lies in their wintering grounds or during migration."

After sifting through numerous models relating climate attributes to AK_All, this is the best I could find. All attributes except AK_All are normalized to [0.0, 1.0] for multiplier comparison.

MODEL 27

```
Attributes:   44
              year
              yearSince1976
              HMtempC_median
              Visibility_median
              CloudCover_median
              FlightHT_median
              WindSpd_median
              WindDegrees
              FlightDegrees
              wndN through wndUNK
              fltN through fltUNK
              AK_All
```

M5P:

```
year <= 0.567 : LM1 (26/85.754%) (Read this as 1976 through 2001.)
year >  0.567 : LM2 (20/43.939%)  (Read this as 2002 through 2021.)
LM num: 1
AK_All =
    -174.155 * year  (year 0.0 is 1976 and year 1.0 is 2021)
    - 76.173 * HMtempC_median
    + 54.8792 * wndNNW
    + 216.2394 * fltSW
    + 517.3267
LM num: 2
AK_All =
    -204.0102 * year  (year 0.0 is 1976 and year 1.0 is 2021)
    - 89.2313 * HMtempC_median
    + 193.2234 * wndNNW
```

```
   + 268.0409 * fltS
   + 146.0159 * fltSW
   + 347.0721
Number of Rules : 2
Correlation coefficient                  0.3798
Mean absolute error                    125.3798
Root mean squared error                  154.9207
```

Unfortunately, including year in the data simply restates the fact that AK_All is declining over the years. Removing year from the data results in a very poor M5P model with a CC of 0.1629.

MODEL 28

M5P

```
LM1 (46/78.751%)
LM num: 1
AK_All =
   210.5895 * WindSpd_median
   - 118.6502 * wndNE
   - 173.4154 * wndENE
   + 134.9931 * wndS
   + 369.5796
Number of Rules : 1
Correlation coefficient                  0.1629
Mean absolute error                    144.3497
Root mean squared error                  187.102
```

Of note is the fact that attributes WindSpd_median and CloudCover_median, which correlate with RT_All and SS_All as discussed previously, and which are available to AK_All Models above, do not appear as coefficients in these models. They are not sufficiently predictive of AK_All.

Matching West Nile Virus counts in PA humans [22] (2001 : 3, 2002 : 62, 2003 : 237, 2004 : 15, 2005 : 25, 2006 : 9, 2007 : 10, 2008 : 14, 2009 : 0, 2010 : 28, 2011 : 6, 2012 : 60, 2013 : 11, 2014 : 13, 2015 : 30, 2016 : 16, 2017 : 20, 2018 : 130, 2019 : 7, 2020 : 11) to AK counts shows no correlation. A study conducted at Hawk Mountain in 2009 concludes, "Our data indicated that WNV prevalence has declined in this population of kestrels over time." [23] While "the key to understanding the kestrel's decline lies in their wintering grounds or during migration" [19], the current climate -> raptor data are insufficient for finding significant factors in AK_All decline. Graphing Model 17's attributes makes nothing clear. Day-of-year time of the final AK sighting has increased slightly in recent years but no other time measures show consistent change.

A Hawk Mountain web page on American Kestrels states, "Recent increases in the numbers of **Sharp-shinned Hawks** and, in particular, **Cooper's Hawks**, two species that prey on kestrels, appear to be linked to at least some of the declines in the Northeast." [24] Mapping other raptor counts in this dataset to AK_All does not show a negative correlation for growth in SS_All and CH_All.

MODEL 29

Instances: 46
Attributes: 16
BE_All
BV_All
BW_All
CH_All
GE_All
ML_All

```
              NG_All
              NH_All
              OS_All
              PG_All
              RL_All
              RS_All
              RT_All
              SS_All
              TV_All
              AK_All
Linear Regression Model
AK_All =
    296.8874 * CH_All +
   -184.6674 * GE_All +
    176.209  * PG_All +
    228.0296 * RT_All +
    257.4603 * SS_All +
    169.2327
Correlation coefficient                  0.7008
Mean absolute error                     88.3201
Root mean squared error                113.5863
```

M5P gives a similar single-rule model with these attributes. Increases in SS_All and CH_All do not correlate with decreases in AK_All that would indicate predation. The negative correlation of GE_All to AK_All is likely due to the fact that GE_All, like BE_All, is on the rise. There is no evidence in this data that a peak year for GE aligns with a valley year for AK. Figure 28 shows that AK_All's presumed predators SS_All and CH_All have similar downward 5-year running average counts, and that AK peak and valley years align well with the other two species.

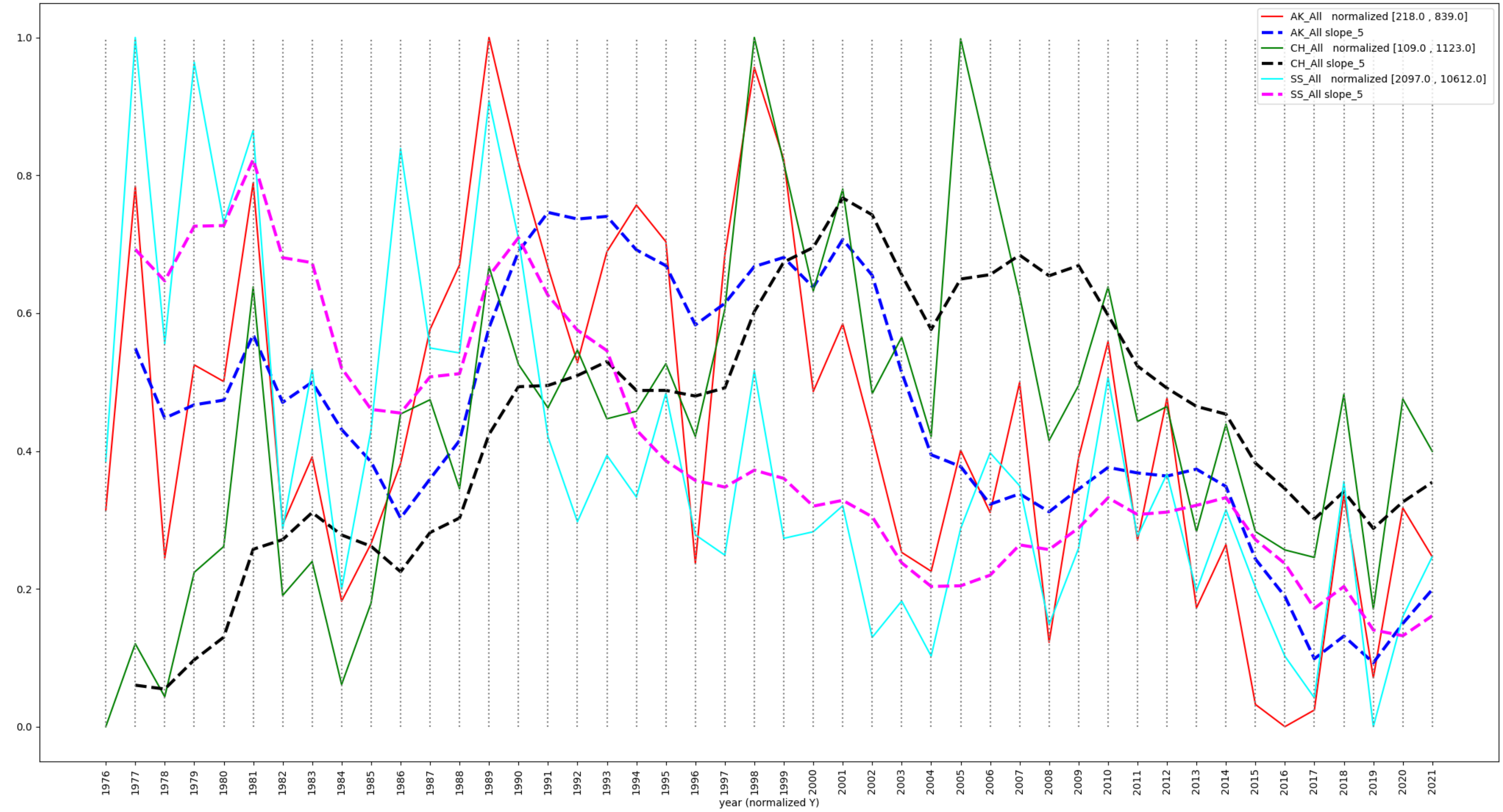


Figure 28: AK_All, CH_All, and SS_All Show Aligned Peaks and Valleys and Similar Downward Slopes in 5-year Average Counts.

Bald Eagles (BE)

A bald eagle flew over our back yard today while I was watering the tomato plants, so this is a good day to start this section. As cited at the start of Section 3, "But the primary reason for the eagle's decline over the last century was the effect of the pesticide DDT and its derivatives on eagle reproduction. It accumulated in eagles and caused their eggs to be too thin to withstand the eagle's weight during incubation. As a result, the bald eagle population plummeted. In 1972, the use of this DDT and other harmful pesticides that bio-accumulated in birds was banned in the United States. The drastic decline of bald eagles and other birds eventually bottomed out. Clean water regulations and heightened environmental awareness during this period also contributed to better fish populations and spurred the recovery on." [16]

The problem with trying to come up with any climate -> BE_All correlation is that the models are dominated by long-term trends in BE_All increases matched with possibly unrelated long-term trends in climate change. BE_All's negative correlation with WindSpd_median in Models 12 through 16 contrasts with WindSpd_median's positive correlations with RT_All, SS_All, and AK_All in Models 23, 25, 26, and 28 above. This negative correlation is likely related more to BE_All's growth in numbers due to elimination of DDT that stands in contrast to incremental reduction in WindSpd_median during the same time frame that it does to a WindSpd_median

-> BE_All negative causal relationship.

Modeling with yearSince1976 in the attributes gives most of the weight to that attribute because of the increased growth in BE_All from 1976 through 2021. Linear Regression with all attributes normalized to [0.0, 1.0] except BE_All gives the most intelligible models.

<u>MODEL 30</u>

Instances: 46
Attributes: 23
yearSince1976
HMtempC_median
Visibility_median
CloudCover_median
WindSpd_median
wndN
wndNNE
wndNE
wndENE
wndE
wndESE
wndSE
wndSSE
wndS
wndSSW
wndSW
wndWSW
wndW
wndWNW
wndNW
wndNNW
wndUNK
BE_All

Linear Regression Model

BE_All =
666.2806 * yearSince1976 +
121.9853 * CloudCover_median +
-59.2479 * WindSpd_median +
-38.2811 * wndN +
56.2614 * wndNNE +
-137.9384 * wndENE +
-84.5218 * wndESE +
-56.377 * wndSE +
-45.9711 * wndSSE +
-54.4672 * wndS +
-137.9858 * wndW +
87.1175 * wndWNW +
-51.5316 * wndNW +
-52.3749 * wndNNW +
47.1238

Correlation coefficient 0.9469

Mean absolute error 44.3011
Root mean squared error 56.4245

Taking yearSince1976 out of the attributes in order to use climate attributes gives the following model. Also removed were CloudCover_median and WindSpd_median because as established for Models 12 through 16 they correlate too closely to yearSince1976 as time keepers.

MODEL 31
Linear Regression Model
BE_All =
-216.5707 * Visibility_median +
-133.2518 * wndN +
280.2081 * wndNNE +
137.3128 * wndENE +
121.9965 * wndE +
-268.8294 * wndESE +
122.578 * wndSSE +
182.0644 * wndSW +
473.3243 * wndWNW +
-417.7585 * wndNNW +
95.4506
Correlation coefficient 0.5409
Mean absolute error 136.669
Root mean squared error 175.9519

In contrast to models for RT_All and SS_All, wndNNW gets a negative multiplier in the above models, several easterly wind directions get positive multipliers, and these models do not use wndNW. Visibility_median is inversely related to CloudCover_median of Model 30 and thus has a negative multiplier. Plotting normalized BE_All against the positively correlated wnd attributes in Model 31, highlighted in bold (wndNNE. wndENE, wndE, wndSSE, wndSW, wndWNW), shows no consistent alignment of peaks and valleys except for wndWNW. Two of the positively correlated directions wndSSE and wndWNW, show time-based increases in Figures 9 and 10, with WNW showing massive growth after 1994. Given the seemingly random mix of positive wnd correlations in Model 31, it appears again that the model selects these attributes as timestamps to correlate with the no-DDT-related growth of BE_All. That time-related growth dominates all models. Figure 29, however, shows close alignment of BE_All and wndWNW peaks and valleys. That alignment is the only reliable climate -> BE_All correlation coming out of these BE_All models. Not only is the growth of wndWNW seen in Figures 9 and 10 a time keeper, but peak and valley alignment shows additional correlation of wndWNW and BE_all; wndNW appears to have mixed correlation with BE_All, countered somewhat by its function as a time keeper.

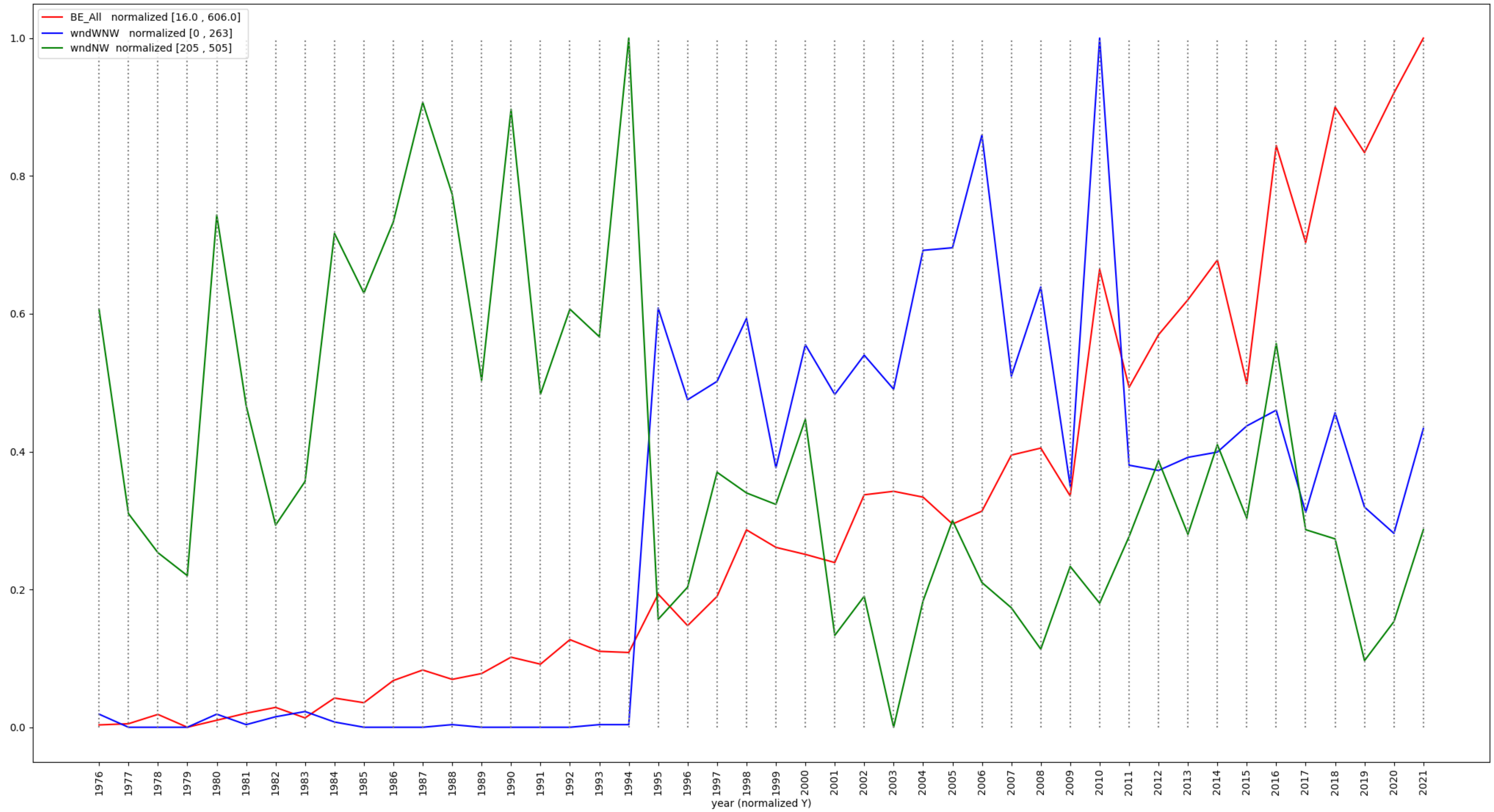


Figure 29: Close Correlation of Peaks and Valleys Between BE_All and wndWNW in Many Recent Years, Some correlation with wndNW

Figure 30 shows the day-of-year milestones for BE have mostly stayed consistent, with the exception of day of the peak count, which varies anywhere from below the arrival of 25% of the eagles, even touching the first sighting in 2001, to greater than the 75% count in 1997 and 2016. The day of the 50% count has increased slightly in recent years.

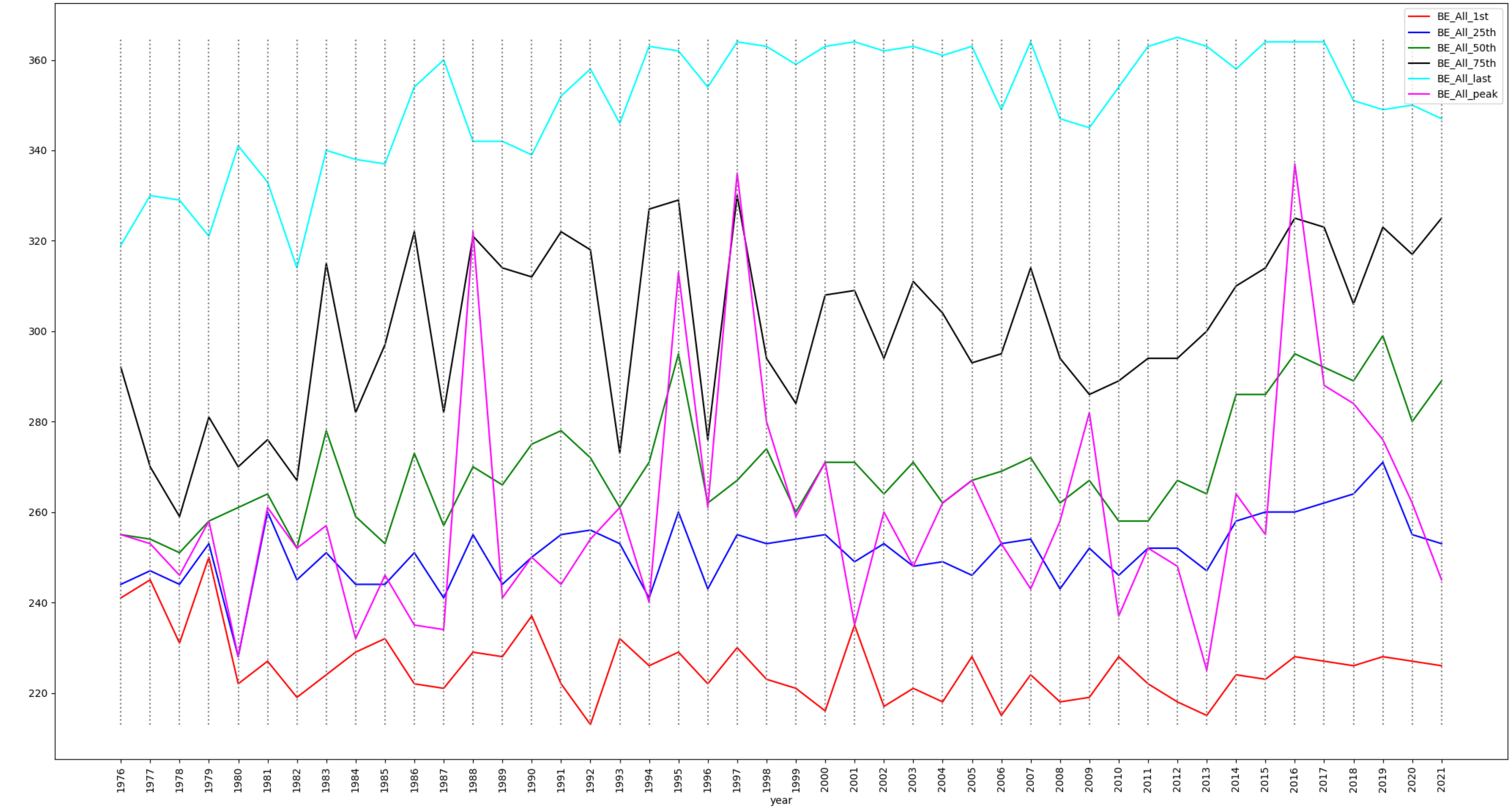


Figure 30: Day-of-Year for First, 25%, 50%, and 75% of BE Count Sightings, Day of Final Sighting, and Day of Peak Sighting for the Year.

Golden Eagles (GE)

Golden Eagle counts show some similarities to Bald Eagle counts and some important differences.

MODEL 32

Instances: 46
Attributes: 23
yearSince1976
HMtempC_median
Visibility_median
CloudCover_median
WindSpd_median

wndN
wndNNE
wndNE
wndENE
wndE
wndESE
wndSE
wndSSE
wndS
wndSSW
wndSW
wndWSW
wndW
wndWNW
wndNW
wndNNW
wndUNK
GE_All

Linear Regression Model

GE_All =

```
    38.9584 * yearSince1976 +
    28.3072 * wndSSE +
    33.99   * wndWSW +
    46.3362 * wndWNW +
    56.3899 * wndNW +
    42.1242 * wndUNK +
     5.1951
Correlation coefficient                  0.7752
Mean absolute error                     20.4135
Root mean squared error                 26.7513
```

Taking yearSince1976 out for similar reasons gives this model.

MODEL 33

Linear Regression Model

GE_All =

```
    31.7557 * wndSSE +
    33.2664 * wndSSW +
    43.1414 * wndWSW +
    49.2779 * wndWNW +
    68.7674 * wndNW +
    57.3074 * wndUNK +
     3.5213
Correlation coefficient                  0.7757
Mean absolute error                     21.9186
Root mean squared error                 27.2456
```

All wnd multipliers are positive with wndNW and wndNNW dominant; wndUNK's positive correlation is likely another time keeper correlated to year. Taking that out gives a slightly less accurate model that now uses negatively-correlated WindSpd_median as a time keeper.

MODEL 34
Linear Regression Model
GE_All =

-25.5706 * WindSpd_median +
20.8163 * wndSSE +
29.6806 * wndS +
39.4178 * wndWSW +
87.0055 * wndWNW +
46.4188 * wndNW +
34.9316

Correlation coefficient	0.7577
Mean absolute error	21.1305
Root mean squared error	27.6934

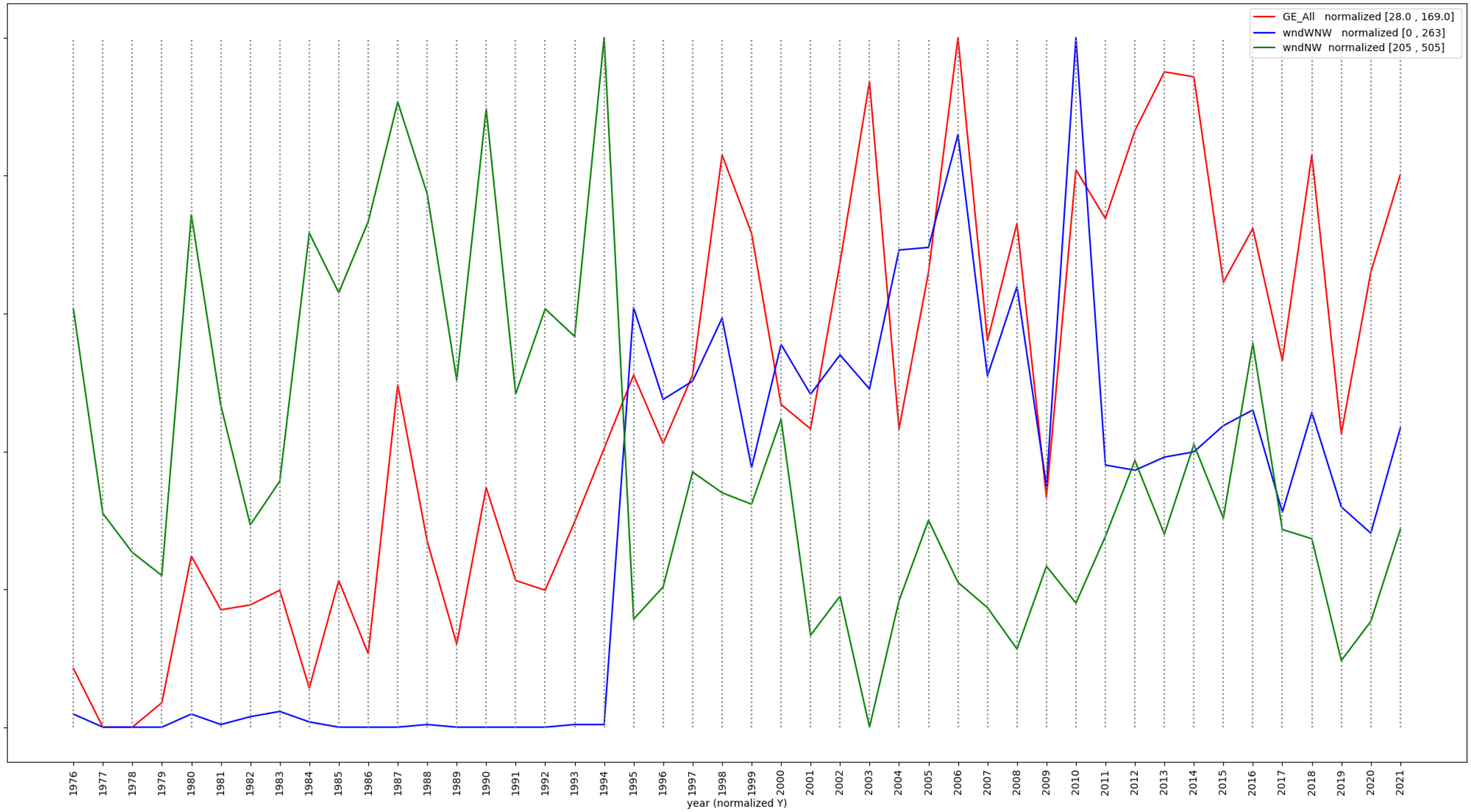


Figure 31: Alignment and Misalignment of GE_All Peaks and Valleys to wndWNW and wndNW.

Figure 31 shows that starting in 2006, GE_All's peaks align with whichever has the higher peak, wndWNW or wndNW, for each year. Before 2006 alignment is mixed. Note the growth in GE_All from 1976 through 2021, although not quite as pronounced as that of BE_All in Figure 29, where BE_All grew from 16 sightings in 1976 to 606 in 2021, a 3687.5% increase. GE_All increased from a minimum of 28 sightings to 169 sightings, a 503.6% increase.

Figure 32 shows that the day-of-year of the first GE sighting has trended somewhat earlier and that of the final sighting slightly later in recent years. Other day-of-arrival have stayed consistent.

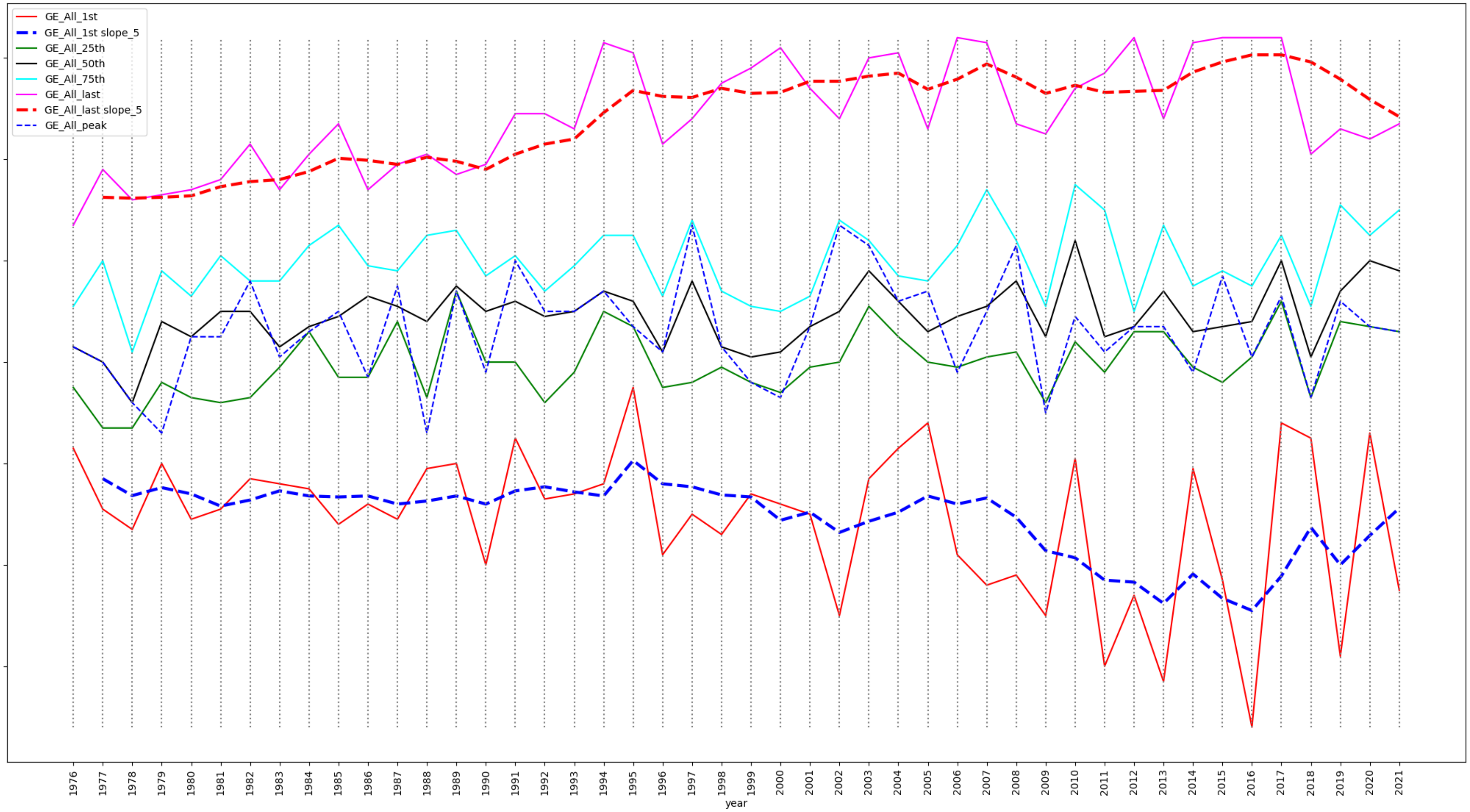


Figure 32: Day-of-Year for First, 25%, 50%, and 75% of GE Count Sightings, Day of Final Sighting, and Day of Peak Sighting for the Year.

Broad-Winged Hawk (BW)

The BW_All analysis is starting at 1979 because the spike in 1978 in Figure 33, verified by Dr. Goodrich, is an accurate outlier that does not correlate with more recent trends of interest.

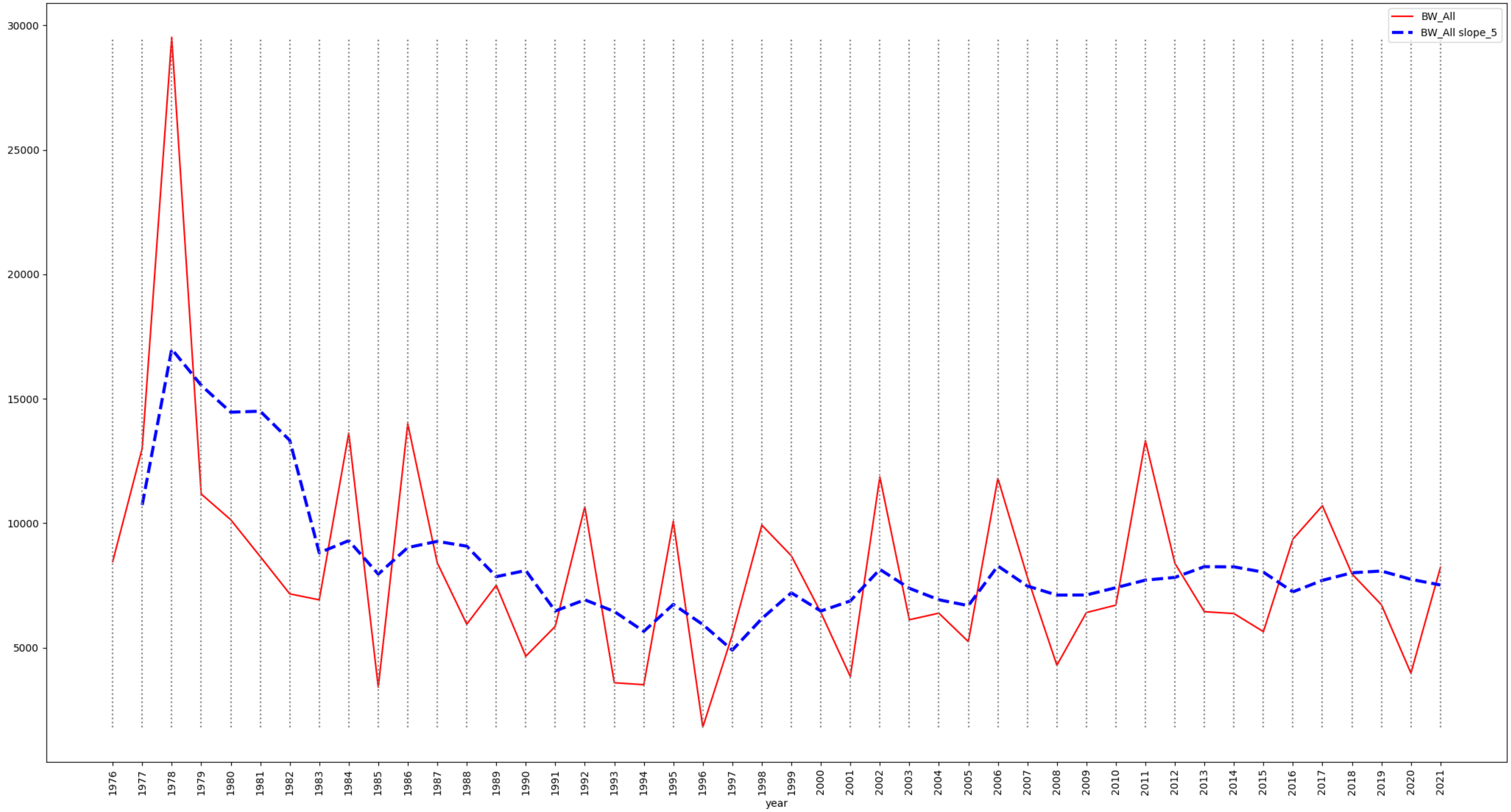


Figure 33: BW_All Counts with a Spike in 1978 to be Excluded from Analysis

For the first time in this set of analyses, no good climate -> BW_All correlations appear. These are the best models that appear in the semi-automated correlation of 379 weather attributes to BW_All without year; adding year adds no change.

MODEL 35
Simple Linear Regression
BW_All =
13913.3 * HMtempC_48_mean + 4254.25, Predicting 0 if attribute value is missing.

```
Correlation coefficient          0.1322
Mean absolute error           3171.3505
Root mean squared error          4538.6597
```

HMtempC_48_mean is the average of the unsigned change in HM temperature C from the recording 48 hours earlier. It is a measure of volatility of temperature change over 48 hours in day-to-day observation.

Linear Regression uses 163 of the 379 weather attributes with no better result.

MODEL 36 (partial)
Linear Regression
BW_All =

```
    -1810.0952 * Visibility_24_median +
    1671.3806 * Visibility_72_pstdv +
    1654.356  * HMtempC_48_mean +
    1568.7942 * wndW_1st +
    1562.0333 * wndNE_75th +
... (most terms elided)
    269.9119 * HourlyDewPointTemperature_24_min +
    -251.1375 * HourlyWindSpeed_mean +
    9490.1511
Correlation coefficient                  0.1364
Mean absolute error                   3719.1128
Root mean squared error                 4782.4621
```

MODEL 37

A hand-written model that simply predicts the mean of BW_All (7565.093) from 1979 through 2021 gives a useless Linear Regression result.

```
BW_All = 7565.093
Correlation coefficient                 -0.3506
Mean absolute error                   2404.4464
Root mean squared error                 2956.9442
```

Using the following 21 weather attributes ...

```
            HMtempC_median
            Visibility_median
            CloudCover_median
            WindSpd_median
            wndN
            wndNNE
            wndNE
            wndENE
            wndE
            wndESE
            wndSE
            wndSSE
            wndS
            wndSSW
            wndSW
            wndWSW
            wndW
            wndWNW
            wndNW
            wndNNW
            wndUNK
```

There is not one BW_All (count) attribute including BW_All_mean (day-by-day mean of BW count for August through December), BW_All_median, BW_All_pstdv, BW_All_min, or BW_All_max (again, daily stats for each year 1979-2021) that gives any correlation with the above 21 weather attributes or any additional weather attributes.

However, there are two day-of-year arrival attributes that are worth attention.

MODEL 38

M5 pruned model tree using normalized non-target attributes:

```
M5 pruned model tree:

wndNNE <= 0.406 :
|   wndW <= 0.405 : LM1 (9/90.35%)
|   wndW >  0.405 : LM2 (24/56.314%)
wndNNE >  0.406 :
|   wndN <= 0.296 : LM3 (7/56.6%)
|   wndN >  0.296 : LM4 (3/22.99%)

LM num: 1
BW_All_25th =
    0.8167 * WindSpd_median
    - 0.5892 * wndN
    + 0.844 * wndNNE
    + 0.9406 * wndE
    + 0.5744 * wndESE
    + 1.7983 * wndW
    + 1.4907 * wndNW
    + 252.5733
LM num: 2
BW_All_25th =
    0.5026 * WindSpd_median
    - 0.5892 * wndN
    + 0.844 * wndNNE
    + 0.5788 * wndE
    + 0.5744 * wndESE
    + 1.1066 * wndW
    + 1.173 * wndNW
    + 254.0794
LM num: 3
BW_All_25th =
    1.6842 * wndN
    + 1.6205 * wndNNE
    + 1.1028 * wndESE
    + 1.2763 * wndNW
    + 255.4164
LM num: 4
BW_All_25th =
    2.3098 * wndN
    + 1.6205 * wndNNE
```

```
   + 1.1028 * wndESE
   + 1.2763 * wndNW
   + 255.5385
Correlation coefficient          0.3519
Mean absolute error              1.6042
Root mean squared error            2.0982
```

BW_All_25th is the day-of-year when 25% of BW for a year have been observed. Model 38 uses normalized values for all attributes except BW_All_25th to support visual comparison of multiplier importance. Note that for low wndNNE and wndW (bottom 40% of their respective range, linear formulas 1 and 2), WindSpd_median has a high correlation with BW_All_25th, while for formulas 3 and 4, wndN and wndNNE have the strongest correlation.

Model 38's mean absolute error (MAE) of 1.6042 means that the model's prediction of the day-of-arrival of 25% of the BW is off by 1.6 days on average. Root mean squared error (RMSE) emphasizes outliers. It is a modest 2 days.

Model 39 predicts the day-of-year of the peak count of BW_All using the same normalized weather attributes as Models 37 and 38.

MODEL 39

Linear Regression Model

BW_All_peak =

```
     -6.7911 * Visibility_median +
     -4.9887 * wndN +
     -3.391  * wndNE +
      2.8295 * wndE +
     -3.8133 * wndSSE +
     -3.5749 * wndSW +
      4.6463 * wndWSW +
      4.722  * wndWNW +
     10.082  * wndNW +
      3.8047 * wndNNW +
     -5.6636 * wndUNK +
    261.4441
Correlation coefficient          0.3064
Mean absolute error              3.1971
Root mean squared error            4.1096
```

Finally, estimating BW_All daily counts using 5385 daily-aggregate (not yearly) normalized records from 1979 through 2021 yields a more accurate model than yearly-aggregate Models 35 and 36.

MODEL 40

```
Attributes:   24 (includes BW_All)
              HMtempC
              WindSpd
              WindDegrees
              wndN
              wndNNE
              wndNE
              wndENE
              wndE
              wndESE
              wndSE
              wndSSE
              wndS
              wndSSW
              wndSW
              wndWSW
              wndW
              wndWNW
              wndNW
              wndNNW
              wndUNK
              Visibility
              CloudCover
              FlightDegrees
              BW_All
Linear Regression Model
BW_All =
    221.7585 * HMtempC +
    208.0692 * wndN +
    228.0709 * wndNNE +
    208.8093 * wndNE +
    213.1463 * wndE +
    350.8032 * wndESE +
    243.8315 * wndSE +
    660.3895 * wndSSE +
    248.8221 * wndS +
    153.1684 * wndSW +
    135.74   * wndW +
    104.9282 * wndWNW +
    169.8283 * wndNW +
    169.0287 * wndNNW +
    123.7978 * wndUNK +
     51.9091 * Visibility +
  -207.8387
Correlation coefficient                  0.2326
Mean absolute error                    103.393
Root mean squared error                  273.6361
```

Daily-aggregate data give a more accurate prediction for BW_All than yearly-aggregate data (CC of 0.2326 versus Model 35's 0.1322 and Model 36's 0.1364), but it is still far from perfect. All wnd measures have positive multipliers, with none for wndENE, wndSSW, and wndWSW. Adding (yearSince1976, daySinceAug1) gives a Linear Regression model with a comparable CC of 0.2353. A daily-aggregate M5P model with 8 linear expressions at the leaves of its decision tree has a CC of 0.5176, but most of the added complexity uses daySinceAug1 as a time keeper. Removing yearSince1976 and daySinceAug1 drops the CC back down to 0.2398.

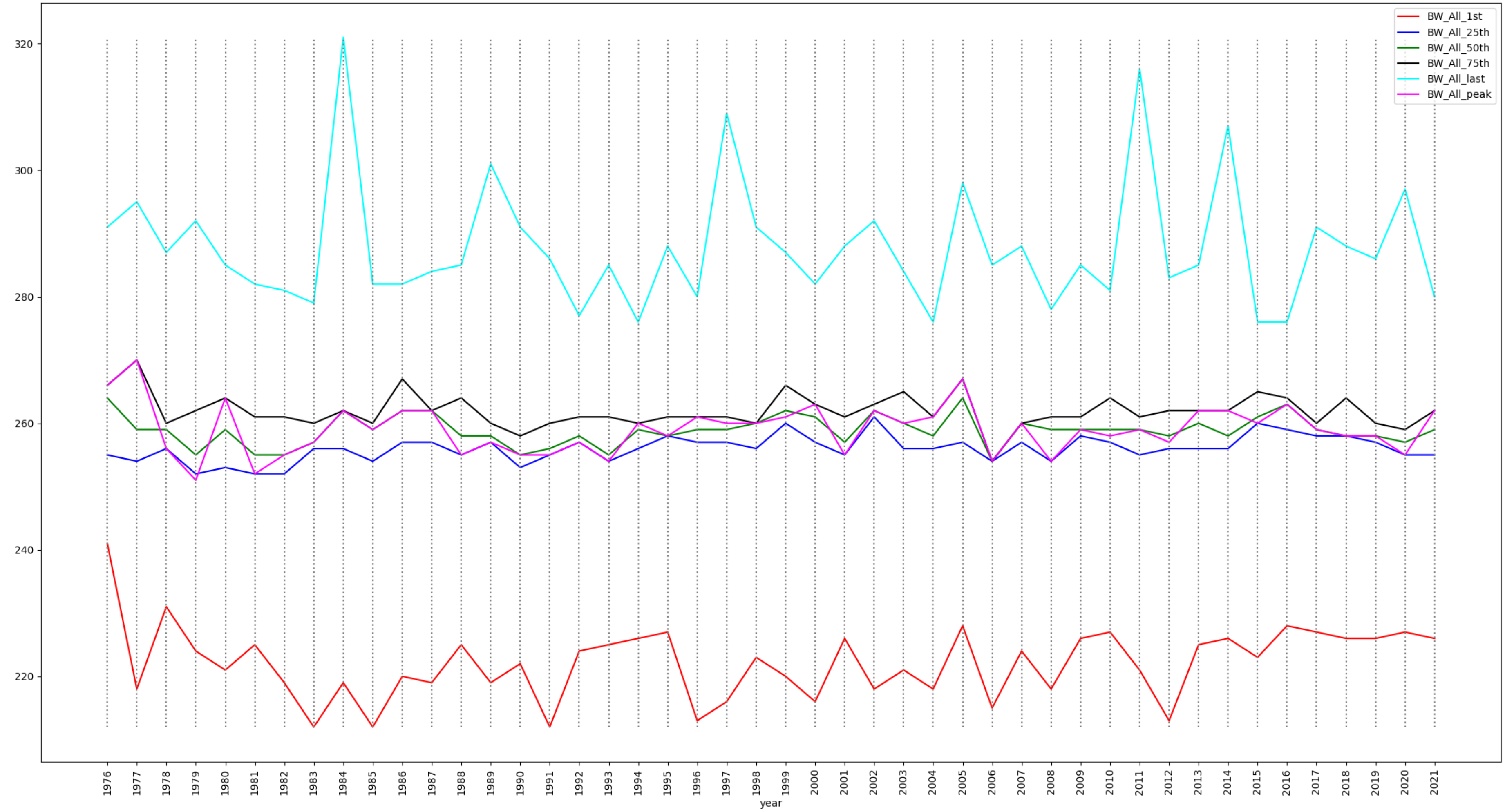


Figure 34: Day-of-Year for First, 25%, 50%, and 75% of BW Count Sightings, Day of Final Sighting, and Day of Peak Sighting for the Year.

Figure 34 shows fluctuations of BW 1st and last sightings but no long-term trend slopes. Figure 35 shows the day-of-year of the modeled BW peak observation with wndNW and wndNNW, two of two strongest multipliers in Model 39. The other two strongly correlated wind directions, wndWNW and wndWSW, do not align well at peaks and valleys. Figure 35 shows close alignment of BW_All_peak and wndNW from 1979 through 1994. After 1994 BW_All_peak aligns some years with wndNW, some years with wndNNW, and some years with their combination.

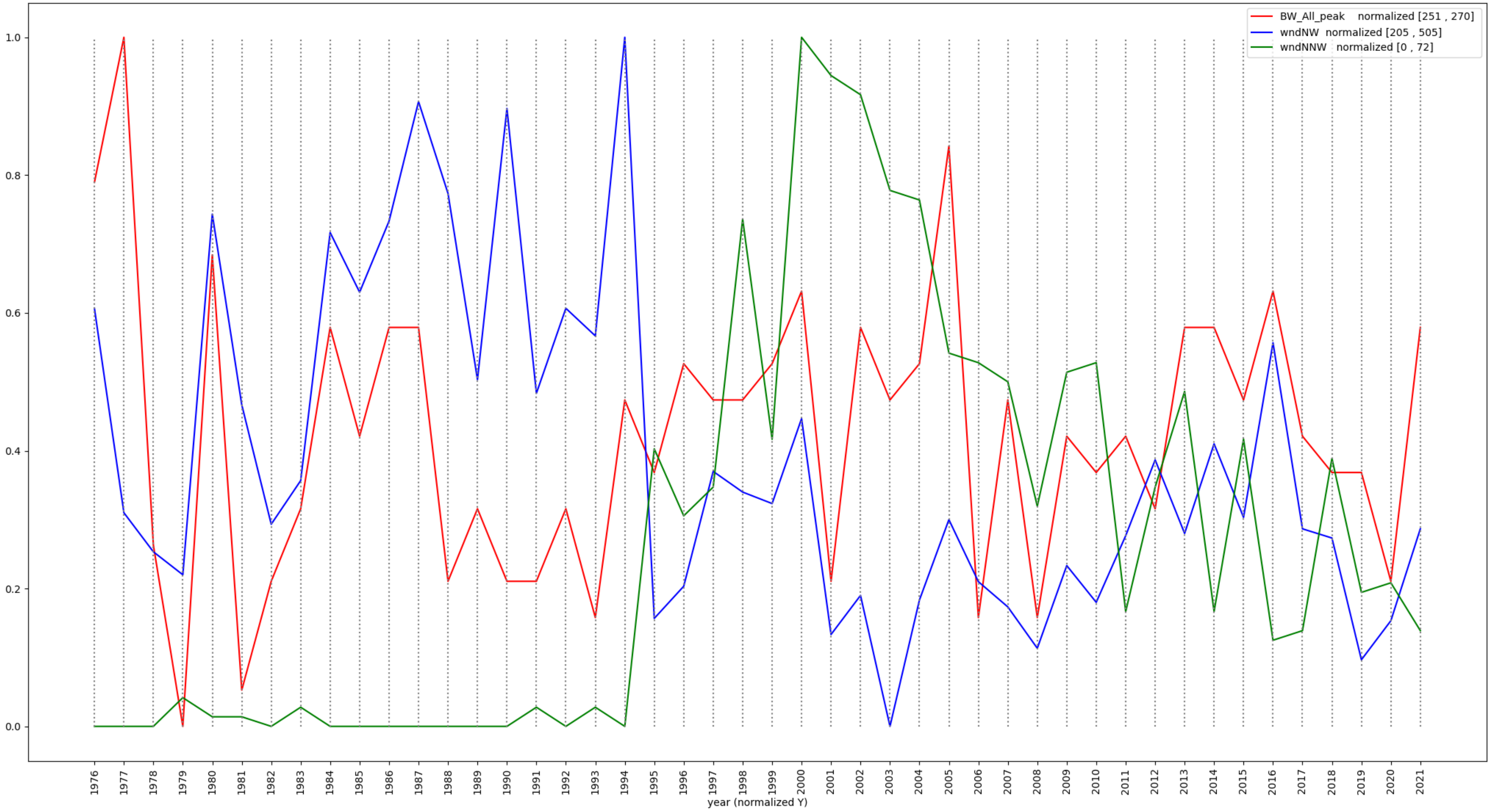


Figure 35: Alignment of Normalized BW_All_peak Day of Observation with wndNW and wndNNW.

Total Raptors (TOTAL)

Because TOTAL raptor attributes aggregate all individual raptor counts, times, and other properties, we can expect climate -> TOTAL correlations to be averaged on the low side.

Here are the models of interest that emerge from the semi-automated analysis of all climate attributes.

MODEL 41 (year not included)
Simple Linear Regression
TOTAL = -11768.69 * wndW + 26043.38, Predicting 0 if attribute value is missing. (Note: windW is not missing from any year.)
Correlation coefficient 0.4193
Mean absolute error 4043.726

```
Root mean squared error             4986.5752
```

MODEL 42 (year included)
M5P
TOTAL =

```
    -7500.1043 * yearSince1976
    + 6241.1935 * wndS_1st
    + 8253.2054 * HMtempC_48_median
    + 6839.7607 * dryTempC24maxFullNOAA
    + 15793.7803
Number of Rules : 1
Correlation coefficient               0.3847
Mean absolute error                4183.4244
Root mean squared error             5250.0484
```

Excluding year from Model 42's data yields a model with CC = 0.2832 that is very complicated to interpret. All attributes except TOTAL are normalized in the range [0.0, 1.0]. In Model 42 yearSince1976 is explicitly a time keeper showing decline of TOTAL raptors since 1976 but not relating to climate. Model 42's positive correlation to HMtempC_48_median relates volatility of temperature at HM over 48-hour day-to-day differentials to TOTAL count. Maximum temperature for each year (hot summers) correlates positively with TOTAL for unknown reasons. The first day of a wndS recording correlates positively and wndW's annual tally correlates negatively as seen in these models.

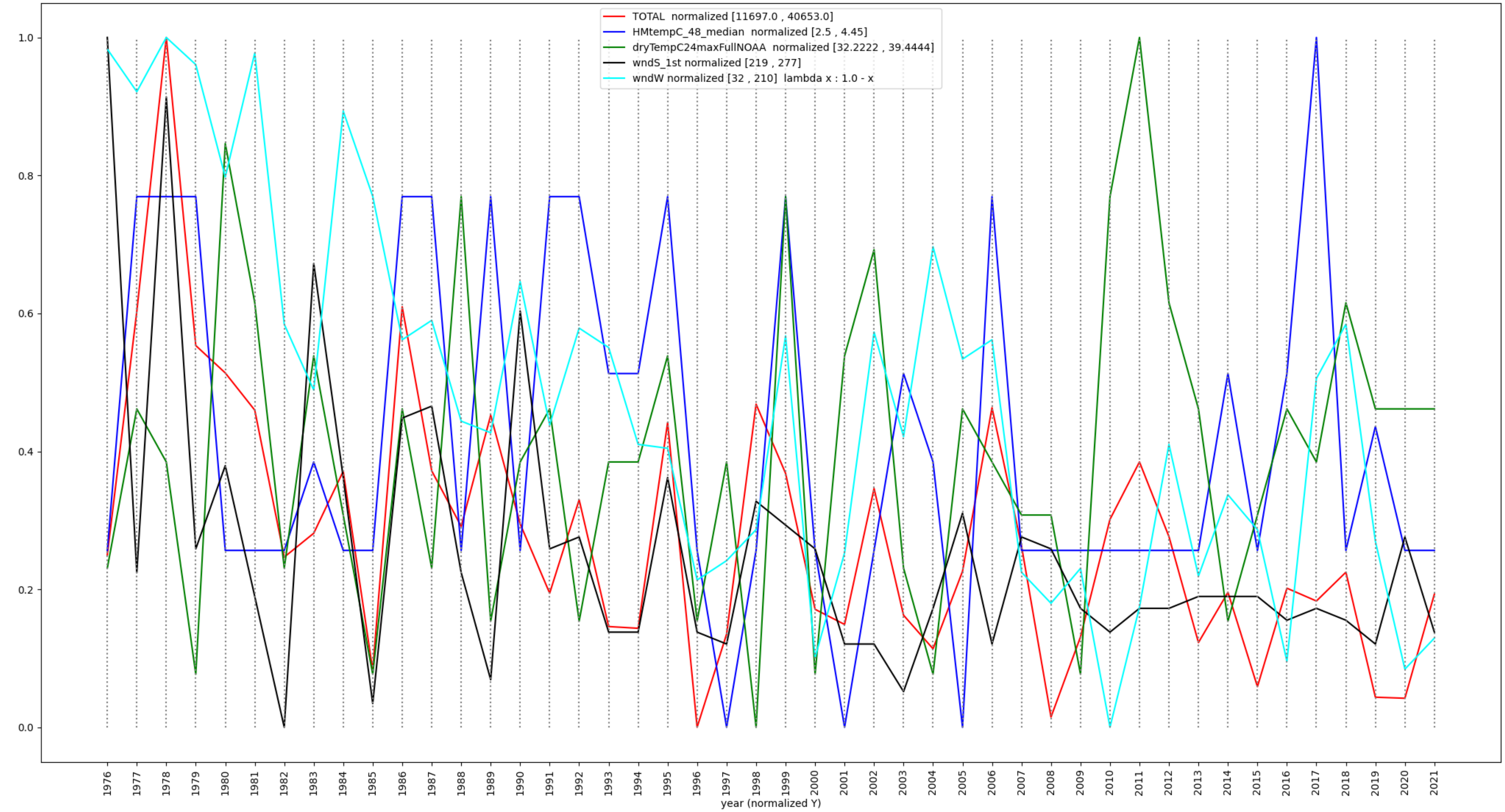


Figure 36: Normalized Climate Attributes Aligned with TOTAL Annual Counts

Several of these are measures of weather **volatility**. HMtempC_48_median is the amount of 48-hour temperature fluctuation plus or minus on observation days August through December -- it is unsigned so as not to converge to 0. Attribute dryTempC24maxFullNOAA is the **maximum** of dry bulb temperature readings at Allentown Airport across 24 hours and 365 days for each year. Attribute wndS_1st with a positive correlation is the day-of-year that the first wndS observation is recorded at HM, and wndW with a negative correlation is the sum of wndW observations. Figure 36 compensates for the negative correlation with the expression "1.0 - normalized wndW", which puts the max count of 210 at the bottom (0.0) and the min count of 32 at the top (1.0). TOTAL's peaks in red align with one or more of these peaks surprisingly well, as do its valleys. My amateur's interpretation is that TOTAL correlates well with volatile weather (HMTempC_48_mean and dryTempC24maxFullNOAA). It also correlates well with late-arriving wndS_1st but I do not conjecture why.

Model 43 on normalized non-Allentown Airport data gives is similar to Model 41 with a negative correlation on wndW.

MODEL 43 (year included)
Instances: 46

```
Attributes:   23
              yearSince1976
              HMtempC_median
              Visibility_median
              CloudCover_median
              WindSpd_median
              wndN
              wndNNE
              wndNE
              wndENE
              wndE
              wndESE
              wndSE
              wndSSE
              wndS
              wndSSW
              wndSW
              wndWSW
              wndW
              wndWNW
              wndNW
              wndNNW
              wndUNK
              TOTAL
```

Simple Linear Regression on wndW

```
TOTAL = -11768.69 * wndW + 26043.38, Predicting 0 if attribute value is missing.
Correlation coefficient                  0.4686
Mean absolute error                   3757.5171
Root mean squared error                4843.8401
```

Model 44 is somewhat more complicated and informative but less accurate. It shows a positive correlation of annual TOTAL with WindSpd_median.

MODEL 44 (year included)

```
TOTAL =
  12962.4411 * HMtempC_median +
   6918.5304 * WindSpd_median +
   5011.2858 * wndNNE +
 -11135.599  * wndSE +
  -8773.289  * wndSSW +
   4987.8884 * wndSW +
 -11976.5453 * wndW +
   8718.5663 * wndWNW +
  -5700.1333 * wndNNW +
  16139.2758
Correlation coefficient                  0.3244
Mean absolute error                   4625.2664
Root mean squared error                5726.0578
```

Single expression M5P is lies between Models 43 and 44 in detail and accuracy.

MODEL 45 (year included)

```
M5 pruned model tree:
LM num: 1
TOTAL =
   8231.4284 * HMtempC_median
   + 11425.7208 * Visibility_median
   - 5979.0017 * wndSE
   - 10834.1285 * wndW
   + 19862.3684
Correlation coefficient                0.458
Mean absolute error                 3941.4218
Root mean squared error               4919.3504
```

Attributes wndW and wndSE have the strongest and most consistent negative correlation for wind direction to TOTAL in Models 44 and 45. The others in Model 44 are probably attempts at fine tuning the model.

Analyzing daily aggregates in an attempt to correlate _24, _48, and _72 hour changes in climate variables to increases or decreases in TOTAL does not improve this analysis. Including year and day achieves slightly better than Model 45's CC with models that are too complex to interpret, with too many linear regression terms and M5P expressions. Removing year and day degrades those daily aggregate models well below Models 44 and 45 in terms of CC.

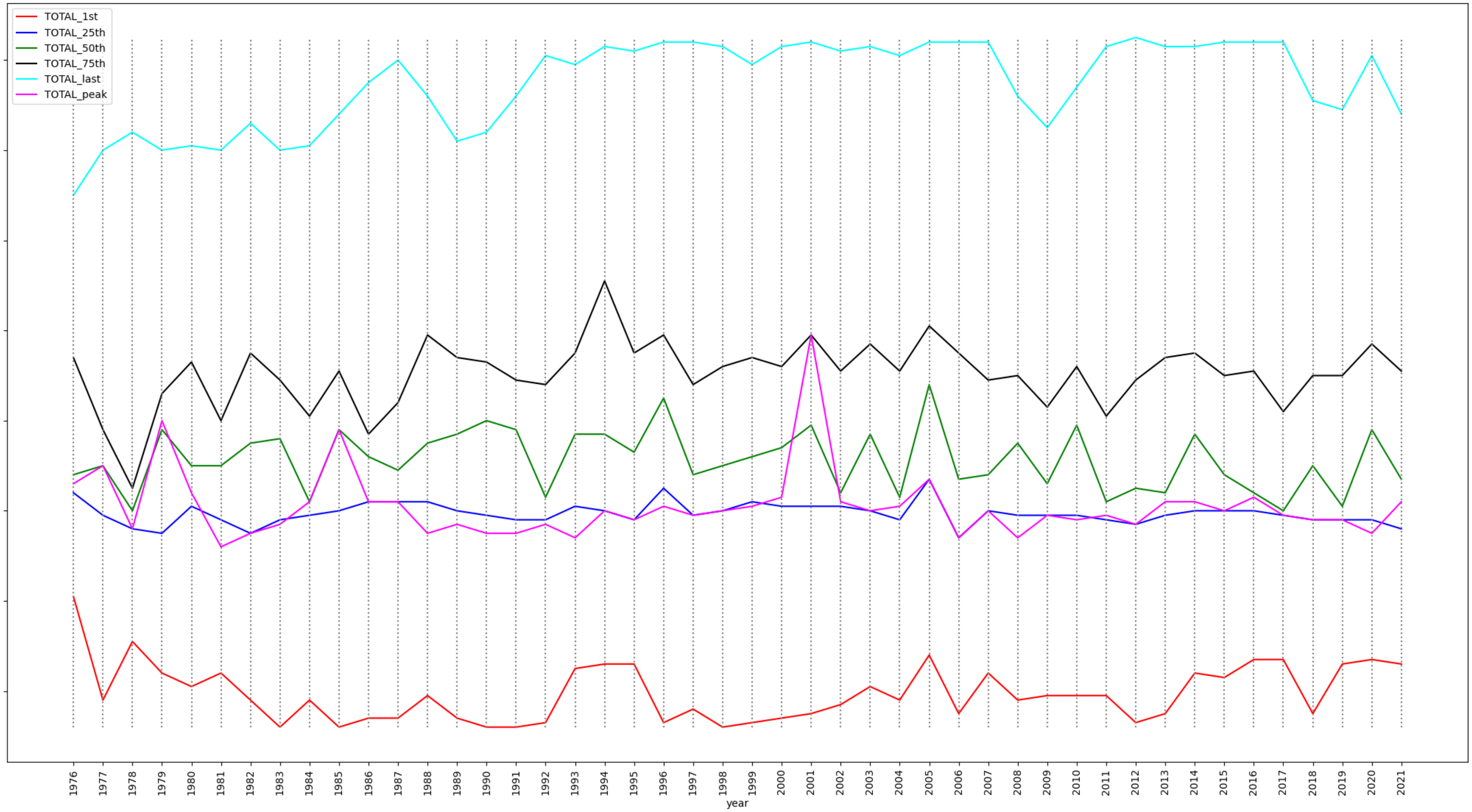


Figure 37: Day-of-Year for First, 25%, 50%, and 75% of TOTAL Count Sightings, Day of Final Sighting, and Day of Peak Sighting for the Year.

Other than a few transients such as 2001's alignment of TOTAL_peak with TOTAL_75th, there are no real trends in TOTAL day-of-year timing in Figure 37 after the early 1990s.

Cooper's Hawk (CH)

Osprey (OS)

Northern Harrier (NH)

Northern Goshawk (NG)

Rough-Legged Hawk (RL)

REFERENCES

1. Dr. Laurie Goodrich, Ph.D., Sarkis Acopian Director of Conservation Science at Hawk Mountain Sanctuary, https://www.hawkmountain.org/about/community/staff/laurie-goodrich.

2. National Oceanic and Atmospheric Administration (NOAA), National Centers for Environmental Information, Allentown Lehigh Valley International Airport data, 1948 through 2021.
https://www.ncdc.noaa.gov/cdo-web/datasets/GHCND/stations/GHCND:USW00014737/detail

3. Dr. Laurie Goodrich, email message regarding observation periods, July 12, 2022.

4. NOAA Solar Calculator, https://gml.noaa.gov/grad/solcalc/.

5. IPython: Interactive Computing, https://ipython.org/.

6. NumPy: Python numeric libraries, https://numpy.org/.

7. Cygwin: Get that Linux feeling on Windows, https://www.cygwin.com/.

8. Weka 3: Machine Learning Software in Java, https://www.cs.waikato.ac.nz/ml/weka/.

9. Attribute-Relation File Format (ARFF), https://www.cs.waikato.ac.nz/~ml/weka/arff.html.

10. Dunbrack Lab, Computational Structural Biology, http://dunbrack.fccc.edu/lab/.

11. Roland L. Dunbrack, Jr, Ph.D., Director - Organic Synthesis Facility, Director - Molecular Modeling Facility,
Adjunct Professor - University of Pennsylvania School of Medicine, Adjunct Associate Professor -
Drexel University College of Medicine, https://www.foxchase.org/roland-dunbrack-jr.

12. M. Mukaka, "A guide to appropriate use of Correlation coefficient in medical research", Malawi Medical Journal, 2012 Sep; 24(3): 69–71.
https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3576830/

13. U.S. Environmental Protection Agency, "Heat Island Effect". https://www.epa.gov/heatislands.

14. N. Malhotra, "The tipping point: The Lehigh Valley faces environmental and public health crises," *The Brown and White*, Lehigh University, January 20, 2021.
https://thebrownandwhite.com/2021/01/20/the-tipping-point-the-lehigh-valley-faces-environmental-and-public-health-crises/

15. Dr. Laurie Goodrich, email message regarding wind speed measurements, July 18, 2022.

16. Pennsylvania Game Commission, "Bald Eagle Species Profile", https://www.pgc.pa.gov/Wildlife/EndangeredandThreatened/Pages/BaldEagle.aspx.

17. Witten, Frank, and Hall, *Data Mining: Practical Machine Learning Tools and Techniques*, 3rd Edition, Morgan Kaufmann, 2011.

18. Deng, K., Liu, W., Azorin-Molina, C., Yang, S., Li, H., Zhang, G., et al. "2022 Terrestrial stilling projected to continue in the Northern Hemisphere mid-latitudes." *Earth's Future*, 10, e2021EF002448, July 7, 2022.
https://doi.org/10.1029/2021EF002448

19. C. Witko, Audubon, "As American Kestrels Mysteriously Decline, Researchers Look to Their Migration for Clues", November 20, 2020.
https://www.audubon.org/news/as-american-kestrels-mysteriously-decline-researchers-look-their-migration-clues

20. American Kestrel Partnership, https://kestrel.peregrinefund.org/home

21. The Peregrine Fund, https://peregrinefund.org/

22. CDC, West Nile Virus Disease Cases by State, Final Annual Maps & Data for 1999-2020,
https://www.cdc.gov/westnile/statsmaps/finalmapsdata/index.html

23. D. Medica and K. Bildstein, "Annual Variation in West Nile Virus Antibodies in American Kestrels in Eastern Pennsylvania", *The Journal of Raptor Research* 43(4):301-207, 2009.
https://www.hawkmountain.org/download/?id=4979&dl=1

24. Hawk Mountain Sanctuary, American Kestrels are the smallest and most colorful falcons in North America.
https://www.hawkmountain.org/raptors/american-kestrel

APPENDIX A: extractLinearCoefficients.py

extractLinearCoefficients.py resides in directory analysis/scripts. There is also some manual, ad hoc analysis code and data in dorecotory analysis_scripts/adhoc.

```
# extractLinearCoefficients.py D. Parson 7/21/2022
# Pull & order the LinearRegression and SimpleLinearRegression
# coefficients from Weka output, ordered by magnitude, not signed.
# Extended 7/25/2022 to extract attributeSelection attributes
import os
import sys
import csv
import numpy as np

def __extract_CCetc__(statslist, targetattr, algorithm, IsYear):
    # OFFSETS INTO OUTCSV ROW:
    # [0] targetattr from input parameter
    # [1] algorithm from input parameter
    # [10] 'Time taken to perform cross-validation: 4 seconds\n',
    # [2] 'Correlation coefficient                  -0.0195\n',
    # [3] 'Mean absolute error                      33.8613\n',
    # [4] 'Root mean squared error                    47.1405\n',
    # [5] 'Relative absolute error                  99.4741 %\n',
    # [6] 'Root relative squared error              100.0018 %\n',
    # [7] 'Total Number of Instances                2052     \n',
    # [8] 'Ignored Class Unknown Instances                3590     \n',
    # [9] added for raptor correlating with year (1) or without (0)
    # [11]..[20]    are top 10 linear coefficients in order of magnitude
    buildrow = {
        'Correlation coefficient'                :   2,
        'Mean absolute error'                    :   3,
        'Root mean squared error'                :   4,
        'Relative absolute error'                :   5,
        'Root relative squared error'            :   6,
        'Total Number of Instances'              :   7,
        'Ignored Class Unknown Instances'        :   8,
        'IsYear'                                 :   9,
        'Time taken to perform cross-validation' :   10
    }
    csvoutrow = [targetattr, algorithm] + [None for i in range(2, 21)]
    csvoutrow[9] = IsYear           # runWeka parameter
    keys = buildrow.keys()
    def filtfunc(ele): return ele.strip() != ''
    for scanix in range(-1, -len(statslist), -1):
        instat = statslist[scanix].strip()
        donescan = False
        for k in keys:
            if k in instat:
                paramix = buildrow[k]
                fieldlist = list(filter(filtfunc, instat.split(' ')))
                if paramix == 10:
                    csvoutrow[paramix] = float(fieldlist[-2].strip()) # seconds
                    donescan = True     # quit looking backward
                elif paramix == 5 or paramix == 6:
                    csvoutrow[paramix] = round(
                        float(fieldlist[-2].strip())/100.0,6)         # %
                else:
                    csvoutrow[paramix] = float(fieldlist[-1].strip())
                if (isinstance(csvoutrow[paramix], float)
                        and np.isnan(csvoutrow[paramix])):
                    csvoutrow[paramix] = None
                break           # quit iterating over keys after key found
        if donescan:
            break           # don't scan any more lines
    return csvoutrow

if __name__ == '__main__':
    mincc = 0.5     # minimum CC of interest
    basename = os.path.basename(os.getcwd())
    csvname = basename + '_linear_stats.csv'
    isnew = not os.path.exists(csvname)
    csvf = open(csvname, 'a')
    statscsv = csv.writer(csvf, delimiter=',', quotechar='"')
    if isnew:
        statscsv.writerow(['target', 'modeler', 'Correlation coefficient',
            'Mean absolute error', 'Root mean squared error',
            'Relative absolute error %', 'Root relative squared error %',
            'Total Number of Instances', 'Ignored Class Unknown Instances',
            'IsYear', 'Seconds taken to perform cross-validation',
            'cc0', 'cc1', 'cc2', 'cc3', 'cc4', 'cc5',
            'cc6', 'cc7', 'cc8', 'cc9'])
    forcefile = False
    filelist = os.listdir('.')
    def sortkey(weightrow): return abs(weightrow[0])  # (coefficient, string)
    if len(sys.argv) > 1:
        try:
            mincc = float(sys.argv[1])
            if len(sys.argv) > 2:
                filelist = sys.argv[2:]
                forcefile = True    # report explicit ones no matter what CC
        except ValueError:
            filelist = sys.argv[1:]
            forcefile = True
    for readname in filelist:
        if (readname.endswith('LinearRegression.txt')
                or readname.endswith('SimpleLinReg.txt')
                or readname.endswith('attributeSelection.txt')):
            print("PROCESSING", readname)
            IsYear = 0 if ('NOyear' in readname) else 1
            algorithm = ''
            if 'LinearRegression' in readname:
                algorithm = 'LinearRegression'
            elif 'SimpleLinReg' in readname:
                algorithm = 'SimpleLinReg'
            else:
                algorithm = 'attributeSelection'
            first_ = readname.find('_')
            last_ = readname.find('_NORM')
            targetattr = readname[first_ + 1:last_]
            if not ('_All' in targetattr):
                continue    # Skip subtypes for now.
            f = open(readname, 'r')
            lines = f.readlines()
            f.close()
            coeffs = []
            if (readname.endswith('LinearRegression.txt')
                    or  readname.endswith('SimpleLinReg.txt')):
                target = ''
                tail = ''
                datarow = __extract_CCetc__(lines, targetattr, algorithm,
                    IsYear)
                if (datarow[2] == None or
                        ((not forcefile) and datarow[2] < mincc)):
                    continue    # Skip this sub-par correlation coefficient
                interesting = False
                for line in lines:
                    l = line.strip()
                    if l.endswith(' ='):
                        target = l
                        interesting = True
                    elif '*' in l:
                        postnum = l.index(' ')
                        num = float(l[0:postnum])
                        coeffs.append((num, l))
                        interesting = True
                    elif interesting:
                        tail += l.strip() + '\n'
                writename = readname[0:-4] + '_COEFFICIENTS.txt'
                w = open(writename, 'w')
                w.write('PROCESSING FILE ' + readname + '\n')
                w.write(target + '\n')
                coeffs.sort(key=sortkey, reverse=True)
                # sort by magnitude, not -
                outcolumn = 11
                for co, s in coeffs:
                    w.write('    ' + s.strip() + '\n')
                    if outcolumn < len(datarow):
                        datarow[outcolumn] = s
                        outcolumn += 1
                w.write('\n' + tail)
                w.close()
            else:       # attributeSelection.txt
                datarow = [targetattr, 'attributeSelection']        \
                    + [None for i in range(2, 21)]
                # 2 thru 10 not used for attribute selection
                outcolumn = 11
                merit = 0.0
                foundAttrs = False
                for line in lines:
                    l = line.strip()
                    if 'Merit of best subset found' in l:
                        lastspace = l.rfind(' ')
                        merit = float(l[lastspace+1:])
                        if (not forcefile) and merit < mincc:
                            break       # Ignore this input file
                    elif 'Selected attributes' in l:
                        foundAttrs = True
                        continue
                    elif foundAttrs and l != '' and outcolumn < len(datarow):
                        datarow[outcolumn] = str(merit) + ' * ' + l
                        outcolumn += 1
                if (not forcefile) and merit < mincc:
                    continue       # Ignore this input file
            statscsv.writerow(datarow)
```

```
        csvf.flush()
        # csvf.close()  # This blows up strangely on some input files.
# It is run-time inefficient to do following each run but requires less
# coding than writing a new script. Only needed for linear models,
csvf = open(csvname, 'r')
rdr = csv.reader(csvf)
hdg = rdr.__next__()
cc0ix = hdg.index('cc0')
multipliers = {}        # map coefficient name to count
for row in rdr:
    for col in range(cc0ix, len(row)):
        l = row[col]
        if l == None:
            break       # Exit col loop.
        try:
            startix = l.index('*')
            plusix = l.index('+') if ('+' in l) else len(l)
            cname = l[startix+1:plusix].strip()
            weight = float(l[0:startix].strip())
        except Exception:
            # flaky format, skip it
            continue
        if cname in multipliers.keys():
            absweight = abs(weight)
            multipliers[cname][0] = multipliers[cname][0] + 1   # count
            multipliers[cname][1] = multipliers[cname][1] + absweight
            absold = abs(multipliers[cname][2])
            if absweight > absold: # Compare by magnitude, save the sign.
                multipliers[cname][2] = weight
        else:
            multipliers[cname] = [1, abs(weight), weight]
csvf.close()
lmult = []
for m in multipliers.keys():
    # [weight, cname, sumOfWeights (will be averaged later), maxWeight]
    row = [multipliers[m][0], m, multipliers[m][1], multipliers[m][2]]
    lmult.append(row)
lmult.sort(key=lambda row : row[0],reverse=True)
rf = open(basename + '_linear_summary.csv', 'w')
rf.write('multCoeff,count,meanAbsWeight,maxWeight,maxAbsWeight\n')
for weightrow in lmult:
    rf.write(weightrow[1] + ',' + str(weightrow[0]) + ','
        + str(round(weightrow[2]/weightrow[0],6)) + ','
        + str(round(weightrow[3],6))
        + ',' + str(round(abs(weightrow[3]),6)) + '\n')
rf.close()
```