Air Temperature Trend Analysis on NY Weather Station Data

Kevin C. Zhou

Hunter College High School, 71 E 94th St., New York, New York, 10128, USA; kevinzhou@hunterschools.org

ABSTRACT: Climate change is a rapidly progressing problem. With temperatures increasing around the globe, big cities are often the places where one can see the most pronounced changes. Although populated locations, such as New York City, have been attempting to reduce pollution and building homes on higher elevation to account for future high sea level, the seasons are hotter than ever before. New York City's temperature changes were compared to smaller towns/cities within the state. The average temperature was calculated for each day using publicly available temperature data that stretched from the late 1800s to present day from the NOAA's National Centers for Environmental Information and then the temperature values were averaged. The average temperature was graphed for January, April, July, and October, encompassing the four seasons. The data was gathered from six weather stations, one of which was in Manhattan. Python and other applications, including Numpy, Bokeh Plotting, SciPy, were used to show both a linear and cubic polynomial relationship between the average temperature. The trends showed that overall, Manhattan had a more pronounced upward when compared to the data gathered from the stations in relatively less populated areas.

KEYWORDS: Mathematics; Other; Python; Modeling; Linear Regression; Climate change.

Introduction

Air temperature "describes the kinetic energy, or energy of motion, of the gases that make up air."[1] Since kinetic energy and temperature are directly proportional, an increase in the velocity of gas molecules means a higher the air temperature.[1] Air temperature is particularly significant because of its effect on factors that may affect plants' and animals' wellbeing and on other weather factors, including humidity, rate of evaporation, both the wind's speed and direction, and precipitation patterns.[1] While warmer temperatures are more ideal for reproduction, animals who thrive in colder temperatures could be negatively affected by significant temperature changes.[1] There is an interest in studying air temperature for several reasons, including analyzing how global warming and air pollution affects animals and plants through changes in temperature in different areas.[2] Complex animal behaviors can be affected by climate change.[3] Global warming has likely been facilitating temperature changes and this can have extremely negative effects in both the near and far future.[4] For example, severe weather can worsen and become commonplace.[4] The sea level can also rise, leading to negative effects.[4] Death rates will increase not only due to natural disasters, but also because of some people's inability to afford a heater or air conditioner on cold and warm days, respectively.[5] Often, the urban areas are much more humid than the rural areas and that has a large effect on all aspects of a person's life, including health.[6] Specifically, those most vulnerable include pregnant woman and children, who can be affected by OTC levels; "OTC refers to comfort level of a person with regards to an exposed outdoor environment."[7] Even in urban locations, the places that provided the most comfort are those most similar to natural landscapes, such as parks.[13]

In this paper, long-term data entailing air temperature is analyzed. In doing this research, more insight into how air temperature in areas in the northeastern part of the United States, specifically New York, has changed over long periods of time, is gained. Previous research from the United States Environmental Protection Agency (EPA) has shown that the average annual temperature within the US states (excluding Hawaii and Alaska) has fluctuated, but has a general upward trend over time, although some experience a lower increase in temperature than others. For example, the Northern and Western areas have experienced the most change, while the Southeast has had the least change. Past results will be reanalyzed, using publicly available data.[9]

There is a scientific phenomenon called the Urban Heat Island Effect where structures that are typical of urban areas, such as buildings and roads, take in and release more heat than natural places, such as forests.[10] It has a very close relationship with increasing urban heat, but also with other factors such as making "urban pollution more severe, affecting citizens' health."[11]

Methods

Study Area:

The state of New York was chosen as the study area, since there is a diverse demographic with areas that have large population and areas with a small population. Some important information relating to the data about the area is as follows: The average annual minimum temperature is 8.9 degrees Celsius and the average annual maximum temperature is 16.8 degrees Celsius.[12] The total area of the region is 54, 555 square miles (land area is 47, 126 m^2).[13]

Data:

Publicly available weather station data from the NOAA's National Centers for Environmental Information was used.[14] The data was recorded on a daily basis for long periods of time. It included maximum temperature, minimum tempratu-

DOI: 10.36838/v3i4.6

re, and temperature observed at the time (in degrees Celsius). The six data sets were chosen because of how large the range of data was and also how complete the data was. Although some data points are missing, they include as much of the long-term data as could be found in the study area. The weather stations are in the state of New York, with one in Manhattan and others in Upstate New York: Angelica, Delhi, Alfred, Alcove Dam, and Addison (Figure 1). The data was cleaned in order to do analysis by making all dates the same format (Year-Day-Month). The Angelica station dataset had 97% coverage, the Delhi station dataset had 84% coverage, the Alfred station dataset had 88% coverage, the Alcove Dam station had 95% coverage, the Addison station dataset had 83% coverage, and the Manhattan station data had 100% coverage (coverage refers to the percentage of the data that is intact from the original).[14] These six data sets were chosen for several reasons. First, they all have relatively high coverage percentages, which means that they are mostly intact in terms of the data. Second, they cover a long stretch of data (since the late 1800s and early 1900s), which was needed so the analysis could be done effectively. Third, Manhattan is the only data set in which there is a population significantly higher than the other five locations in order to test if population would affect long-term temperature trends. The number of data stations picked was also limited so that the data would be manageable.

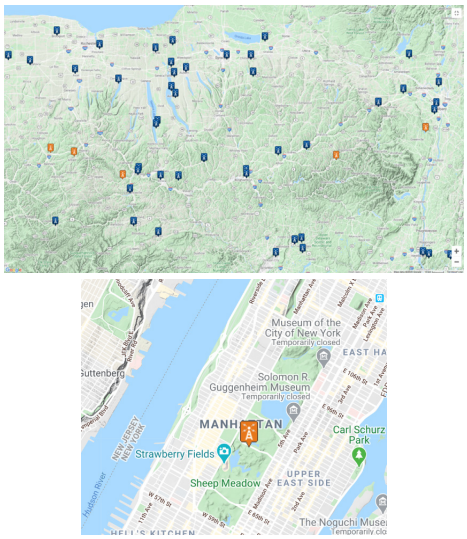

Figure 1: Map of stations in which the selected ones are highlighted in orange.

The population statistics of the areas around each weather station are as follows:[15]

Angelica (2019) – 818
Addison (2019) – 1652
Delhi (2019) – 2949
Alfred (2019) – 3991
Manhattan, NYC (2020) – 1,631,990[16]
Alcove, zip code 12007 – 60[17]

Hypothesis:

It was hypothesized that the air temperature of the data would increase over time in general and the data that came from the farthest north part of New York would show the most significant change.

Methods

It was decided to choose one month in each season for further analysis. Therefore, January, April, July, and October were chosen representing winter, spring, summer, and fall, respectively. Each graph was the average of the TMAX and TMIN for each month. Each graph had trendlines of form

$y=ax^3+bx^2+cx+d$ and $y=ax+b$.

Pandas was used for importing, preparing, and analyzing the data. The data was in a Comma Separated Version (CSV) file format. Out of all the data columns that existed in the original data set only, TMAX, TMIN, and DATE were imported.[18]

Variables were set for TMAX and TMIN to create an average for each day in each data set. Again, Pandas was used to group the averages by month. Finally, it was coded so a particular month could be selected to graph.

Furthermore, Numpy was used to create the trendlines ($y = ax+b$ and $y = ax^3+bx^2+cx+d$), with numpy.polyfit for the graphs. The variable x represents the date exactly in the middle of each month in which the average daily temperature on that day was recorded. For example, it could be January 15, 1899. Polyfit is a function that is built into numpy with several parameters. If it is a line, only 2 parameters are used, but if it is a cubic, there needs to be 4 parameters.[19] Bokeh is used to show the graph, along with the trendline. The numpy function, is finite, and was used to compensate for missing data so that trendlines (of which one has the equation of a polynomial and the other the form of a linear equation) could be properly produced and graphed.[20]

Simple linear regression is an assumption that data can be modeled through a linear relationship ($y= ax+b$). Polynomial regression is similar, except the relationship can be assumed to be of a trendline of the form of a standard polynomial like a cubic. It uses the method of least squares. The matrix operations used in this study were meant to organize and standardize the data so it could be used more easily. A matrix is an array of rows and columns that contain variables (in this case numbers; years and temperature) which I treated as a single identity. Numpy.array was used to make the data into matrices. SciPy was a tool also utilized to perform the T-test for the linear regression and retrieve the p-values to validate the equations.[21]

The T-test is a way to indicate the statistical significance of the results by looking at the coefficients gotten through simple linear regression. The overall method is to sum up the deviation for each value of the data set, $e_i = y_i - a^*x_i - b$, and determines if the standard deviation is small enough and determine the accuracy of the linear equation.[22]

Besides the T-test algorithm, the validity of the results was tested in Excel by importing the data and using the data analysis regression function.

Results and Discussion

In each of the January graphs, there was variation in the trend in Upstate NY locations. For the station in Angelica, NY (Figure 2-13), the curve started in 1893 and went up, then went down, starting around 1930, and then up again in around 2000. There was an overall downward trend, from

approximately -4.8 °C to -5.2 °C. For the station in Delhi, NY (Figure 11), the graph goes in an upward direction, from -7.2 °C to -2.8 °C, going gradually down from just after 1920 to around the mid-1990s where the curve started to go up. For the station in Alfred, NY (Figure 12), the curve started in 1893 and went down, then gradually up, starting in around 1940 and went back down around 2000. There is a slight downward trend of an estimated 0.1 °C. For the station in Alcove Dam, NY (Figure 13), there seems to be several ups and downs, but the graph's trendline goes upwards generally from -6.8 °C to -5.4 °C. The curve goes down gradually at first but then starts to go back up around the mid-1970s. For the station in Addison, NY (Figure 9), it goes down from 1893 to around the 1990s and goes up, with a slight downward trend from -3.6 °C to -4.8 °C.

Table 1: Summary of trendline equations and p-values. Manhattan has larger positive slopes and smaller p-values..

	Angelica	Delhi	Alfred	Alcove Dam	Addison	Manhattan
January	y= -0.00771 * x + -4.52 pvalue=0.258	y= -0.0157 * x + -4.71 pvalue=0.170	y= -0.000809 * x + -5.33 pvalue=0.914	y= 0.0177 * x + -6.62 pvalue=0.204	y= -0.0108 * x + -3.59 pvalue=0.127	y= 0.0115 * x + -0.875 pvalue=0.0140
April	y= 0.00215 * x + 6.4 pvalue=0.595	y= 0.00275 * x + 6.42 pvalue=0.686	y= 0.00438 * x + 6.08 pvalue=0.948	y= 0.0109 * x + 6.59 pvalue=0.208	y= -0.00561 * x + 7.82 pvalue=0.178	y= 0.0218 * x + 9.05 pvalue=3.00e-15
July	y= -0.00482 * x + 20.0 pvalue=0.143	y= -0.00813 * x + 20.1 pvalue=0.0835	y= -0.00192 * x + 19.6 pvalue=0.533	y= -0.00226 * x + 20.9 pvalue=0.700	y= -0.0115 * x + 21.5 pvalue=0.000247	y= 0.00967 * x + 23.9 pvalue=6.23e-06
October	y= -0.00274 * x + 9.33 pvalue=0.515	y= -0.00773 * x + 9.5 pvalue=0.245	y= -0.00153 * x + 9.2 pvalue=0.741	y= -0.00467 * x + 9.57 pvalue=0.575	y= -0.0068 * x + 10.5 pvalue=0.129	y= 0.0115 * x + 13.1 pvalue=9.96e-05

First location of focus: Angelica, NY.

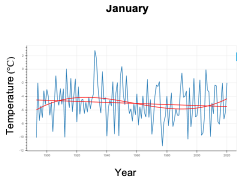

Figure 2: January temperature data in Angelica, NY.

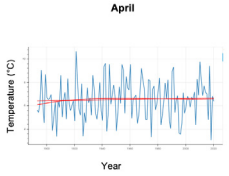

Figure 3: April temperature data in Angelica, NY.

y= -0.00771 * x + -4.52
y= 1.32e-05 * x^3 + -0.00257 * x^2 + 0.126 * x + -5.99

y= 0.00215 * x + 6.4
y= 1.17e-06 * x^3 + -0.000299 * x^2 + 0.0232 * x + 6.08

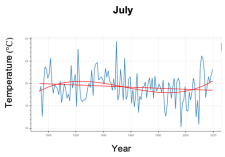

Figure 4: July temperature data in Angelica, NY.

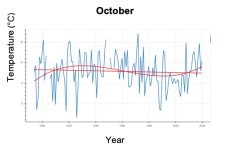

Figure 5: October temperature data in Angelica, NY.

y= -0.00482 * x + 20.0
y= 7.79e-06 * x^3 + -0.00144 * x^2 + 0.0654 * x + 19.3

y= -0.00274 * x + 9.33
y= 9.08e-06 * x^3 + -0.00181 * x^2 + 0.0956 * x + 8.18

Second location of focus: Manhattan

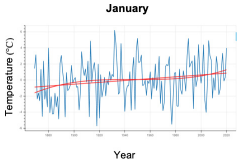

Figure 6: January temperature data in Manhattan, NY.

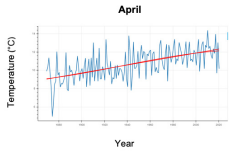

Figure 7: April temperature data in Manhattan, NY.

y= 0.0115 * x + -0.875
y= 3.35e-06 * x^3 + -0.000799 * x^2 + 0.0633 * x + -1.59

y= 0.0218 * x + 9.05
y= -7.28e-07 * x^3 + 0.000144 * x^2 + 0.0151 * x + 9.09

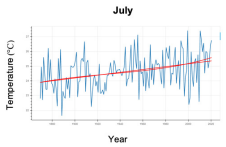

Figure 8: July temperature data in Manhattan, NY.

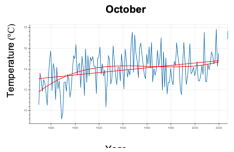

Figure 9: October temperature data in Manhattan, NY.

y= 0.00967 * x + 23.9
y= 6.61e-07 * x^3 + -0.000121 * x^2 + 0.0143 * x + 23.9

y= 0.0115 * x + 13.1
y= 3.52e-06 * x^3 + -0.000981 * x^2 + 0.0871 * x + 11.8

Other locations in January: Addison, Delhi, Alfred, Alcove Dam

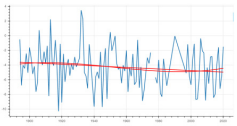

Figure 10: January temperature in Addison, NY.

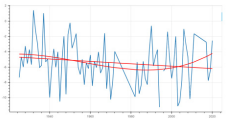

Figure 11: January temperature in Delhi, NY.

y= -0.0157 * x + -4.71
y= 1.6e-05 * x^3 + -0.00147 * x^2 + -0.00382 * x + -4.31

y= -0.000809 * x + -5.33
y= 8e-06 * x^3 + -0.00146 * x^2 + 0.0684 * x + -5.96

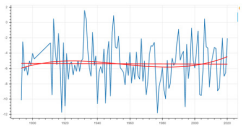

Figure 12: January temperature in Alfred, NY.

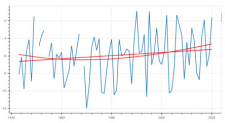

Figure 13: January temperature in Alcove Dam, NY.

y= 0.0177 * x + -6.62
y= -3.74e-06 * x^3 + 0.00116 * x^2 + -0.0517 * x + -5.82

y= -0.0108 * x + -3.59
y= 4.41e-06 * x^3 + -0.00077 * x^2 + 0.0224 * x + -3.84

But in Manhattan, the overall trend is a more obvious upward change from -0.8 °C to 0.8 °C. It goes up from around 1869 to 1940 and then goes down slightly until 1980, where it begins to go up again (Figure 6).

There is similar behavior in all the other months for each location as demonstrated in Table 1 and Figures 2-13.

Conclusion

Overall, the hypothesis was partially supported. There was no overall general upwards trend over time in all the stations, except Manhattan. There was much variation in the Upstate New York locations, according to the graphs with trendlines of form y= ax+b. Each station behaved relatively similarly to each other in that there was an approximately equal number of trendlines going in a downward direction as the number of trendlines going in an upward direction. There was also a mix of how extreme the upward trend or downward trend was. However, stations that had more fully complete data sets (not missing any data) were more likely to have less of a large absolute value in beginning point temperature and end point temperatures. One interesting observation noticed was that the April graphs tended to have fewer extreme trends. Additionally, most of the graphs started to trend upward very recently (as indicated by the polynomial trend lines). It is possible that the exponential population growth within the last century, and

DOI: 10.36838/v3i4.6

also the rise of innovational technology, lead to increases in carbon emissions and therefore, higher temperatures for any areas, urban or not.

On the other hand, Manhattan graphs all had an upward trend that was very noticeable, which can be attributed to climate change due to the large population compared to the other areas. The large population likely causes higher emission of greenhouses gases and other negative effects that cause the temperature to increase noticeably over long periods of time. Carbon emissions can often build up over time in local environments and then, the effects spread to larger areas.

Furthermore, based on p-value of T-test analysis, Manhattan p-values were all below 0.05, indicating statistical significance while all the other locations p-values were generally above that threshold, which shows insignificance. One interesting thing to note is that the July values were the lowest, which could be due to the fact that more people are outside, furthering the congestion of greenhouse gases in the atmosphere.

To build off of this, there are several ways to continue this research. One is to expand the number of data sets and/or the area covered by the stations. In this paper, the data was only limited to the New York region and five data sets. Another is to have more specific graphs, on trends for one specific year or decade. A third, which is slightly controversial, is to "fill the gap", as in make the graph complete. It can make for more complete graphs, but not necessarily more accurate or more precise.

Aknowledgements

I would like to thank Mohammad Abouali, Adjunct Faculty at San Diego State University, for teaching me the necessary knowledge to analyze the data sets and produce the necessary graphs and trendlines.

References

1. Fondriest Staff, "What is Air Temperature?," Fondriest, last modified August 12, 2010, accessed May 19, 2020, https://www.fondriest.com/news/airtemperature.htm.
2. Denchak, Melissa. "Are the Effects of Global Warming Really that Bad?" NRDC. Last modified March 15, 2016. https://www.nrdc.org/stories/are-effects-global-warming-really-bad.
3. Gordo, Oscar, and Hideyuki Doi. "Drivers of population variability in phenological responses to climate change in Japanese birds." Climate Research 54, no. 2 (September 2012): 95-112.
4. Denchak, Melissa. "Are the Effects of Global Warming Really that Bad?" NRDC. Last modified March 15, 2016. https://www.nrdc.org/stories/are-effects-global-warming-really-bad.
5. Denchak, Melissa. "Are the Effects of Global Warming Really that Bad?" NRDC. Last modified March 15, 2016. https://www.nrdc.org/stories/are-effects-global-warming-really-bad.
6. Lukić, Milica, and Jelena Milovanović. "UTCI BASED ASSESSMENT OF URBAN OUTDOOR THERMAL COMFORT IN BELGRADE, SERBIA." International Scientific Conference on Information Technology and Data Related Research, 2020, 70-77.
7. Saun, Fong Chng, and Logaraj Ramakreshnan. "Evaluation of secondary school student's outdoor thermal comfort during peak urban heating hours in Greater Kuala Lumpur." Journal of Health and Translational Medicine, 2020.
8. Aram, Farshid, Ester Higueras Garcia, Sephideh Baghaee, Ebrahim Solgi, Amir Mosavi, and Shahab S. Band. "How Parks Provide Thermal Comfort Perception in the Metropolitan Cores; a Case Study in Madrid Mediterranean Climatic Zone." Climate Risk Management, 2020.
9. NOAA, "Climate Change Indicators: U.S. and Global Temperature," United States Environmental Protection Agency, last modified August 2016, accessed October 29, 2020, https://www.epa.gov/climate-indicators/climate-change-indicators-us-and-global-temperature.
10. "Learn About Heat Islands," United States Environmental Protection Agency, Accessed November 8, 2020. https://www.epa.gov/heatislands/learn-about-heat-islands.
11. Yang, Li & Qian, Feng & Song, De-Xuan & Zheng, Ke-Jia. (2016). Research on Urban Heat-Island Effect. Procedia Engineering. 169. 11-18. 10.1016/j.proeng.2016.10.002.
12. "Average Day And Night Temperature In New York (New York State) In Celsius." Weather & Climate. https://weather-and-climate.com/average-monthly-min-max-Temperature,New-York,United-States-of-America.
13. Infoplease Staff, "Land and Water Area of States," Infoplease, last modified February 11, 2017, accessed May 19, 2020, https://www.infoplease.com/us/states/land-and-water-area-of-states.
14. "Learn About Heat Islands," United States Environmental Protection Agency, Accessed November 8, 2020. https://www.epa.gov/heatislands/learn-about-heat-islands.
15. United States Census Bureau. Accessed October 29, 2020. https://www.census.gov/.
16. "Manhattan Population 2020." World Population Review. Last modified 2020. https://worldpopulationreview.com/boroughs/manhattan-population.
17. "Stats and Demographics for the 12007 ZIP Code." United States Zipcodes. https://www.unitedstateszipcodes.org/12007/.
18. "User Guide." pandas. https://pandas.pydata.org/docs/user_guide/index.html.
19. "numpy.polyfit," NumPy, accessed October 29, 2020. https://numpy.org/doc/stable/reference/generated/numpy.polyfit.html.
20. "bokeh.plotting," bokeh, Accessed October 29, 2020. https://docs.bokeh.org/en/latest/docs/reference/plotting.html.
21. "NumPy Reference,", Scipy.org, Accessed November 8, 2020. https://docs.scipy.org/doc/numpy-1.17.0/reference/.
22. "Simple Linear Regression Analysis, "Accessed November 8, 2020. http://reliawiki.org/index.php/Simple_Linear_Regression_Analysis.

Author

Kevin Zhou is a senior from Hunter College High School in New York City, NY. He is passionate about climate change and is a STEM student. He also loves volunteering and making a positive change in communities. He is a senior volunteer for Queens Youth Justice Center and a podcast producer for City Atlas.