ENVIRONMENTAL TENDENCIES OF SALT LAKE CITY

CORBIN APPLE, BRIDGET HYLAND, BEN STERLING

1. Introduction. It has been well established by The United States Environmental Protection Agency and other independent organizations that average temperature is increasing across the country. This study examines monthly temperature data from two stations: Rye Patch Dam, Nevada, and Salt Lake City International Airport, Utah. These stations were chosen because they are roughly at the same latitude (40.498°N and 40.790°N), longitude (118.316°W and 111.980°W), and elevation (1260.3m and 1287.8m, for Rye Patch Dam and Salt Lake City respectively). Our data is from the National Centers for Environmental Information [1], which contains monthly data about major meteorological parameters at many locations across the country. We chose Salt Lake City and Rye Patch Dam because of their similar geographical characteristics. Data for Rye Patch Dam begins in 1935, but data for Salt Lake City only reaches back to 1948, so we used only years from 1948 to 2020 in our analysis. The dataset includes many parameters, but of particular interest to us were average temperature, average precipitation, number of days with thunderstorms, and total minutes of sunshine.

The goal of our first hypothesis is to examine whether the effects of climate change statistically differ between these two locations; if they do, we may be able to infer that the climate change is predominantly man-made. We accomplished this by calculating the monthly temperature anomaly at each location and comparing the mean temperature anomalies of the two locations. The goal of our second hypothesis is to determine predictors of temperature. We use a multiple linear regression with temperature as the response variable and precipitation, days with thunderstorms, and minutes of sunshine as the dependent variables.

2. First Hypothesis. Climate change is a well-documented phenomenon: on average, the global temperature is increasing. In this section, we test the hypothesis that temperature increases in Salt Lake City and Rye Patch Dam are not equal.

Temperature anomaly is used to compare change in temperature over time. A mean temperature is calculated over a long period of time, and this long-term mean is subtracted from more recent observations. For example, it is known that the global average temperature was 13.9°C between 1901 and 2000. If the global average temperature in 2015 was 14.5°C, then the temperature anomaly for that year would be $14.5 - 13.9 = 0.6$°C. When analyzing temperature data over many years, it is advisable to use a temperature anomaly rather than absolute temperature because it eliminates seasonal variation within the year, allowing for a more significant result. This is an accepted and widely used method in climate analysis [DOES THIS CITATION WORK??] [2]

The GSOM records the monthly average temperature at each location, so we determined temperature anomaly monthly. We calculated a January anomaly by averaging the January temperatures of each year from 1948 to 1974. Then, we subtracted this long-term mean from each January temperature from 1975 to 2020. We did the same for the other eleven months and for both locations, yielding different anomalies for each. In our data files, this is column CG, labeled TAVG ADJ. We used this anomaly in our comparison in place of the absolute temperature. It served to reduce the variance of each sample to produce a meaningful result.

Because of the large number of observations in each sample, we invoked the central

limit theorem to approximate the distribution as normal. The Shapiro-Wilk test for normality gives a p-value of .01663 for the Salt Lake City sample and 0.0003392 for the Rye Patch Dam sample, which verifies that the normal approximation is appropriate ($P < 0.05$). The linear nature of the plots further supports our approximation.

Formally, our hypotheses are:

$$H_0 : \mu_{SLC} = \mu_{RPD} \ \ vs. \ \ H_a : \mu_{SLC} \neq \mu_{RPD}.$$

Because the two stations have similar geographical characteristics (i.e. latitude, longitude, and elevation), we determined that this is an observational matched pairs analysis and used the paired t-test at $\alpha = 0.05$ accordingly. As such, it was not necessary to determine equality of variance between the two samples. We determined the test statistic to be $t = -16.46$. This is less than the critical value $-t_{n-1,\alpha/2} = -1.648$ (where n = 487). Also, the p-value is 2.2×10^{-16}, which is less than the significance level $\alpha = 0.05$. Based on these results, we reject H_0 and conclude that the mean temperature anomaly at Salt Lake City is significantly greater than the mean temperature anomaly at Rye Patch Dam. In other words, since 1975, the temperature has increased more at Salt Lake City than at Rye Patch Dam. The reason for this is not known, but previous research suggests that Salt Lake City emits a relatively large amount of carbon pollution per capita, which hastens the effects of climate change there [3]. These results, and useful statistics, are summarized in Tables 1 and 2.

	ΔT Salt Lake City	ΔT Rye Patch
μ	1.1229	0.1317
σ	1.9765	1.9226
n	487	487
$P_{Shapiro}$	0.0166	0.0003

TABLE 1
Temperature Difference Statistics

t	-16.46
df	486
P	2.2×10^{-16}
Conf. Interval	$(-1.1095, -0.8728)$

TABLE 2
Paired t Test Results for ΔT

We also present the Q-Q Plots in Figures 1 and 2 to visualize the data normality. The linear relationships suggest that both datasets are normal.

We then randomly selected 10% of the data from each station to remove from the dataset and treat as missing. Although our dataset contains ample missing values on its own, we removed more values in order to evalued the effects of missing data. Our analysis ignored months in which temperature data for Salt Lake City *or* Rye Patch Dam was missing. This resulted in a loss of total data points greater than 10%, leaving us with a new sample size (that is different for each run) for each station. Performing the Shapiro-Wilk test again, the distribution of the observations were still normal. The new test statistic for a sample run was -13.984, which is less than the critical value $-t_{n-1,\alpha/2} = -1.648$ (where n = 397). The p-value was 2.2×10^{-16},

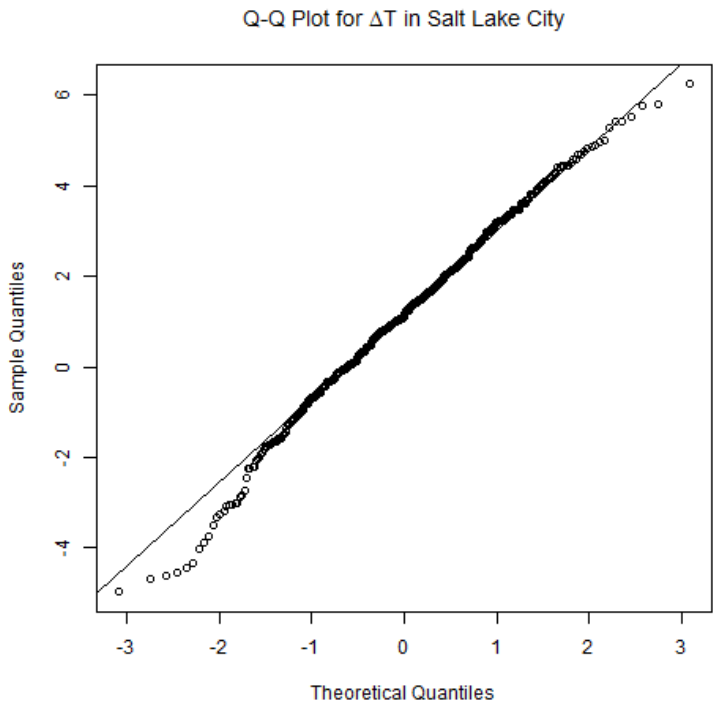

FIG. 1. *Q-Q Plot in Salt Lake City*

which is less than the significance level $\alpha = 0.05$ (note that the reported p-value is the same as before because the true value is below machine precision). Therefore, we reject H_0 and conclude that the mean temperature anomaly at Salt Lake City is greater than the mean temperature anomaly at Rye Patch Dam. The addition of more missing values reduced the sample size and (theoretically) resulted in a higher p-value, but did not change the outcome of the hypothesis test.

3. Second Hypothesis. [TO DO: Add a section on missing values research]

The second hypothesis analyzes which environmental factors are the best predicters of temperature for Salt Lake City. According to the United States EPA ("EPA")[4], higher temperature results in either more or less percipitation and higher frequency of storms. This study performs a multiple linear regression of temperature against percipitation, number of thunderstorms per month, and minutes of sunlight per month. In conjuction with the EPA article noted above, we also considered two key characteristics when deciding which parameters we would inlcude in our study: (1) Linear relationship with temperature, and (2) Availability of data. For example, [TO CONFIRM WHY WE DID NOT INCLUDE, WHY DID WE NOT INCLUDE FOG? SEEMS PRETTY GOOD] we did not include total monthly snowfall or number of days with fog because they had a very weak linear realtionship (i.e. low R^2) with temperature, although we did have sufficient snowfall and number of days with fog data available. The parameter with the most significant linear relationship with temperature was, not suprisingly, minutes of sunlight. A major factor we had to consider here, however, was that sunlight data was only recorded from 1965 to 2004.

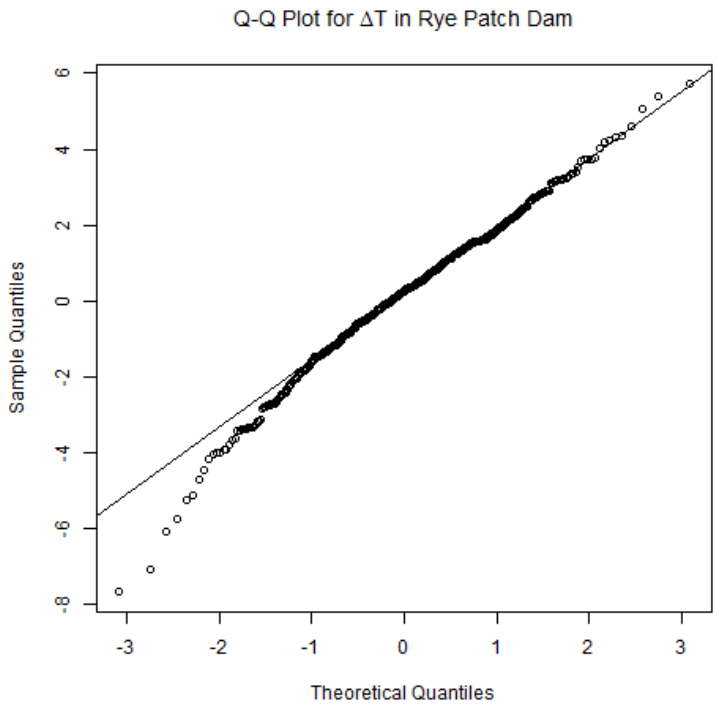

FIG. 2. *Q-Q Plot in Rye Patch*

Thus, in order to inlcude this parameter in our analysis, we needed to restrict our study down to this range (roughly 479 total months, or observations). Additionally, we limited our data used in our study to months in which we had available daya for all four parameters listed above. Therefore, when including temperature, percipation, and number of thunderstorms in our study, our sample size decreased from 479 months to 322 months because at least one of the these parameters was missing data for in the 479 months of data avialable between 1965 and 2004.

Turning specifically to our selected independent parameters, and as we conducted in our initial paramter selection process noted above, we compared temperature data by each independent variable, inlcuding a line of best fit. As expected, there is a strong positive correlation between minutes of sunlight and temperature, as seen in Figure 3. Similarly, albeit not as strong, there is also to be a positive correlation between number of thunderstorms and temperature, as seen in Figure 4. Lastly, the correlation between percipitation and temperature is negative, as seen in Figure 5, which is expected in some regions as stated by the US EPA[4].

As an additional preliminarly analysis, in order to obtain a strong multiple linear regression model to estimate temperature, we sought to use parameters that are not highly correlated with each other, as including additional variables that are highly correlated is more likely to simply increase model complexity as opposed to imporve the model fit. Figure 6, which is a correlation matrix of the three independent parameters, suggests that the independent variables of sunlight, days of thunderstorm, and percipitation are not highly correlated, further supporting our decision to include

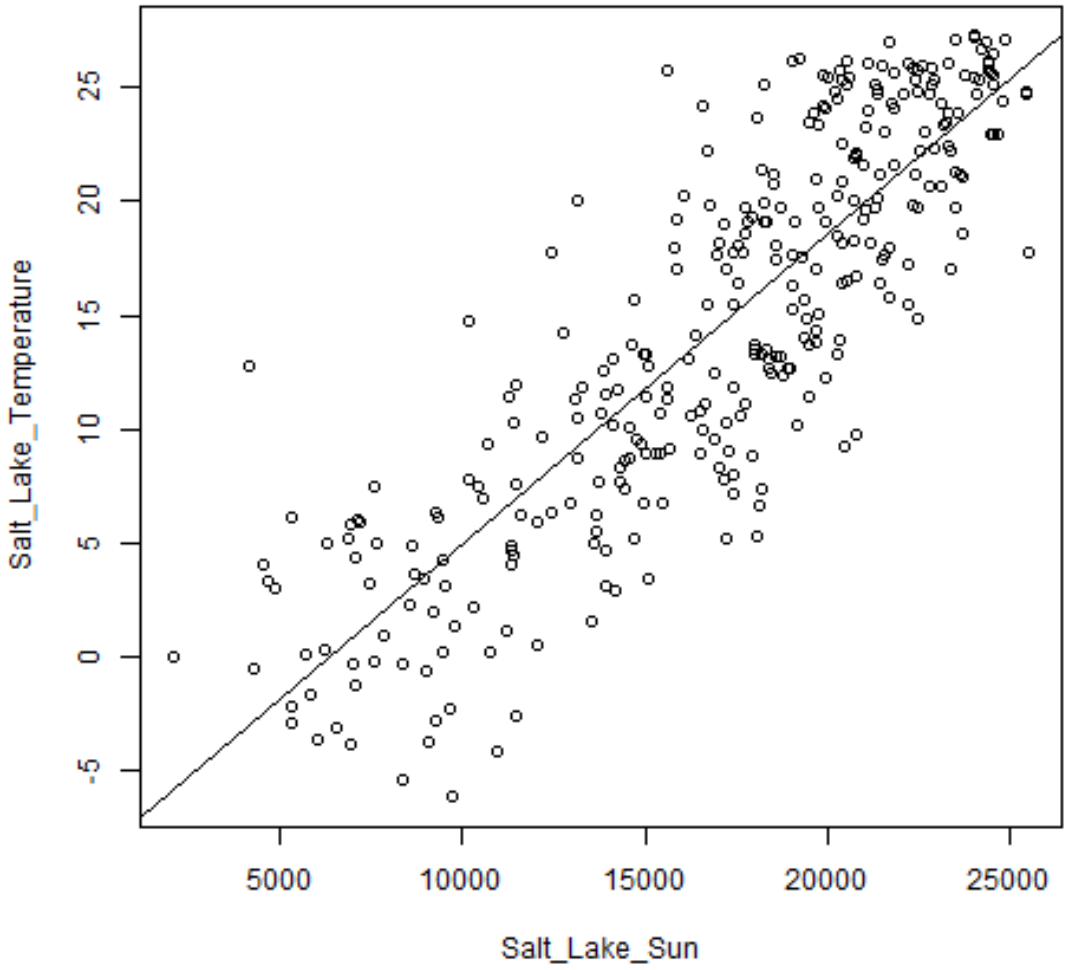

FIG. 3. *Salt Lake City Temperature versus Sun*

them as parameters in our model.

[TO DO: NEED THE R OUTPUT FROM THE MULT LIN REG TO TALK ABOUT IT IN A PARAGRAPH] At this point, we conducted the analysis on the selected paramters. Per Figure [ADD TABLE OF SUMMARY FOR FIT OUTPUT], we can see the the model has an adjusted R^2 of $\approx 80.0\%$, meaning that roughly 80.0% of the variance in temperature can be explained by the three predictors. Using the coefficients from the table, we may produce a formula for the fitted values, as seen in equation [ADD THE Y HAT EQUATION, Y_HAT = B_0 + X_1*B_1 + X_2*B_2 + X_3*B_3].

[WOULD WE LIKE TO INLCUDE A PARAGRAPH ON CONFIDENCE OR

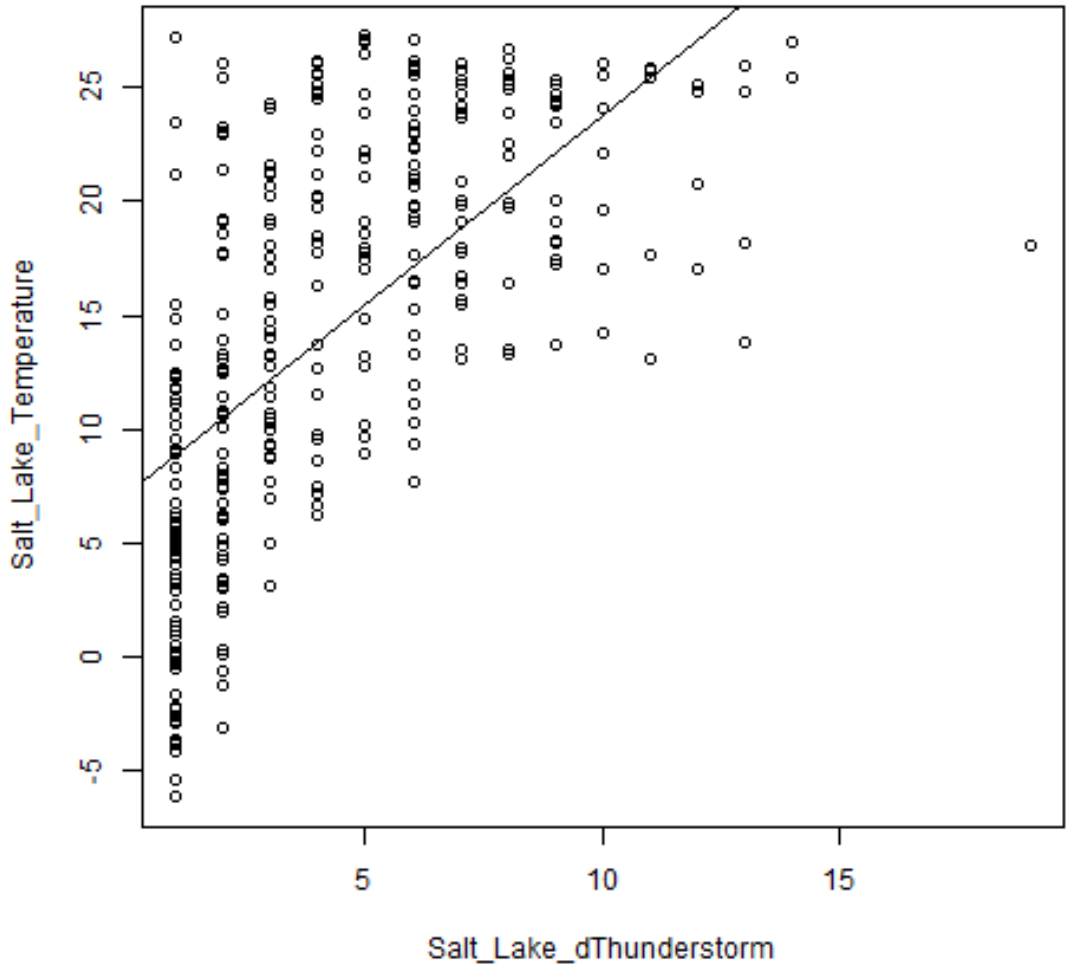

FIG. 4. *Salt Lake City Temperature versus Days of Thunderstorms*

PREDICTION INTERVALS?]

Now that we have completed out multiple linear regression analysis on the three selected parameters, we turn to a final model comparison analysis. Here, we will conduct a best subset analysis to locate the best model at each model size, including 1, 2 and 3 parameters. After locating the best model at each parameter size, we will conduct a model comparison analysis, analyzing the adjusted R^2, AIC, and BIC values for each model.

The best subset selection process considers all 3 choose k [PUT IN MATH FORM] at each k number of parameters, k = 1, 2, 3. For each k, the best subset method will select the model with the lowest SSE. As can been seen in Figure [ADD FIGURE

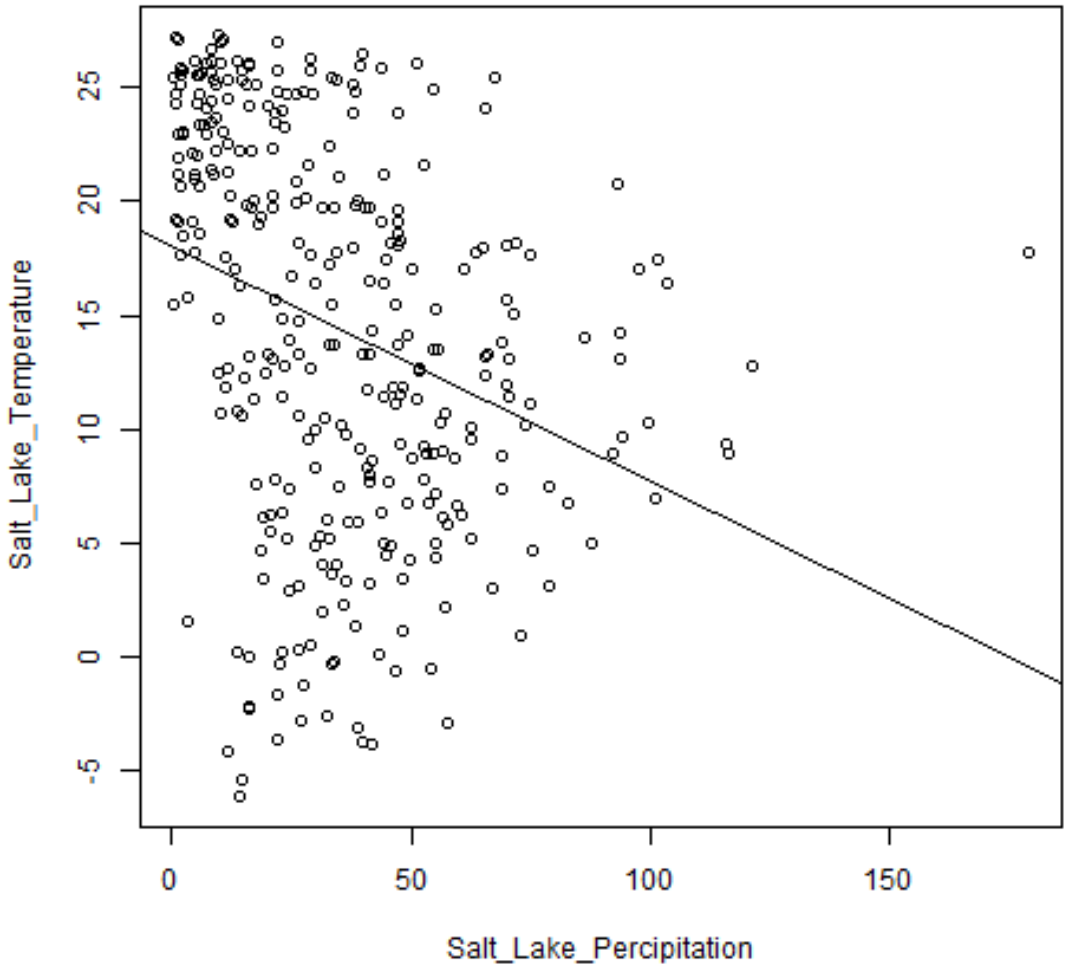

FIG. 5. *Salt Lake City Temperature versus Percipitation*

WITH *'S HERE], the best models for k = 1, 2, and 3 are sunlight, sunlihgt + days of thunderstorms, and sunlight + days of thunderstorms + percipation, respectively.

Now that we have our three best models at each k, we will calculate the following metrics for each model: (1) Adjusted R^2; (2) AIC; and, (3) BIC. The adjusted R^2 will show what percentage of the variance in the model is explained by the parameters, adjusting downward as model complexity increases. AIC and BIC are both penalized-likelihood criteria, where the BIC penalizes model compexity a bit more than AIC. Overall, the best model of the three will have the great adjusted R^2 and the lowest AIC and BIC values. As seen in Table [ADD TABLE OF THE VALUES HERE], the full model, or the one including all three independent paramters, is the best overall

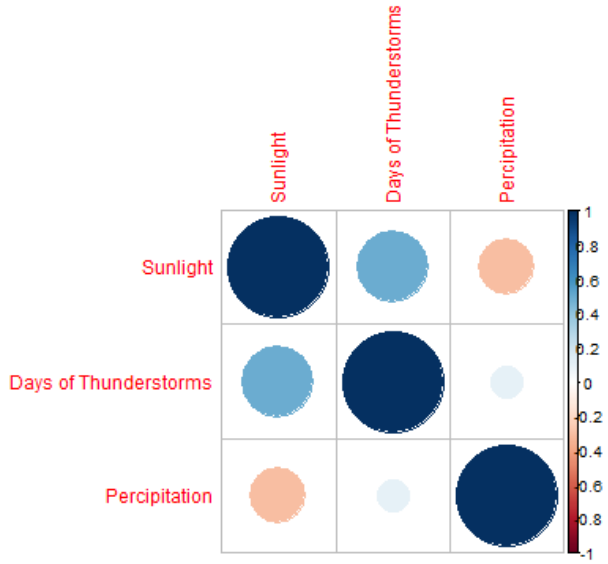

FIG. 6. *Correlation plot of chosen dependent variables*

model to choose to estimate temperature.

	ΔT Salt Lake City	ΔT Rye Patch
μ	1.1229	0.1317
σ	1.9765	1.9226
n	487	487
$P_{Shapiro}$	0.0166	0.0003

TABLE 3
Temperature Difference Statistics

t	-16.46
df	486
P	2.2×10^{-16}
Conf. Interval	$(-1.1095, -0.8728)$

TABLE 4
Paired t Test Results for ΔT

4. R Code Compilation Instructions. To run the R scripts, it is imperative to do so from the /data subdirectory as both reference data from a relative path. The team ran the scripts from Terminal with the "Rscript" command instead of RStudio;

they could possibly be run from RStudio, but we did not test it with this.

If not installed already, one can install the libraries used by the scripts by entering the following commands in Terminal:

- R
- >install.packages("ggpubr")
- >install.packages("ggplot2")
- >install.packages("latex2exp")
- >install.packages("car")
- >install.packages("corrplot")
- >install.packages("leaps")

One can run the scripts by executing the following commands in Terminal:

- cd <Base Directory>/AMS_572_Project/data
- Rscript Hypothesis_1.R
- Rscript Hypothesis_2.R

REFERENCES

[1] J. H. Lawrimore, R. Ray, S. Applequist, B. Korzeniewski, and M. J. Menne, *Global summary of the month (gsom) dataset, version 1*, 2016, https://doi.org/10.7289/V5QV3JJ5 (accessed 2020-11-14).

[2] NASA, *The exploring the environment project*, http://ete.cet.edu/gcc/?/globaltemp_anomalies/.

[3] The White House, *The threat of carbon pollution: Utah*, https://obamawhitehouse.archives.gov/sites/default/files/docs/state-reports/climate/Utah%20Fact%20Sheet.pdf.

[4] United States Environmental Protection Agency, *What climate change means for utah*, https://19january2017snapshot.epa.gov/sites/production/files/2016-09/documents/climate-change-ut.pdf.