Time Series Analysis of New England Weather Data

Andrew Disher* Dr. Wanchunzi Yu †

Abstract

Climate change is an issue that has been at the forefront of public concern for the past few decades and understanding it as well as the direction it is heading is of the utmost importance. There has been extensive research in the past that has addressed this issue on a global scale, yet the resulting findings have been unable to resonate with and convince populations of the local implications of such research. The purpose of this study is to analyze weather data of the New England region in the United States of America to (1) explore trends in snowfall, snow depth, precipitation and average temperature, evaluate the significance of such trends and (2) create time series models that will accurately forecast future values that may be of some use in understanding the future of weather in the region. The structure of this study was organized according to each of the six states in New England, which include Rhode Island, Connecticut, Massachusetts, Maine, Vermont and New Hampshire. For the first of the two tasks, simple linear regression models were applied to each metric against time to better visualize the linear trends inherent in the data, which are valuable for explanatory purposes. As for the second task, various time series model types were tested, but ultimately ARIMA and seasonal ARIMA models were fit to the data in an attempt to predict future monthly values for each metric. The assumptions of these models were also checked via diagnostic measures such as the residual auto-correlation function (ACF), partial auto-correlation function (PACF), the Ljung-Box statistics for the residuals, as well as the usual diagnostics for normality and constant variance. Stationarity was also considered in the creation of these models, and as a result suitable models for prediction for certain weather metrics were unable to be acquired since they failed to meet this requirement. In total, 24 linear regression models and 24 time series models were created to better understand the nature and trajectory of climate change in New England.

Key Words: Time Series Analysis, Gloabal CLimate Change, Data Visualization

1. Introduction

Climate change presents many challenges to the world in the forms of glacier retreat, coral reef deterioration, rising overall temperatures, and sea levels rising causing a multitude of problems, one of which is the complete submergence of islands in the worlds oceans. However, as severe as the risk is for destruction caused by climate change throughout the world it is often difficult to see its immediate and thus potential effects in people's own local areas. In particular, the Northeastern region of the United States known as New England has seemingly sustained no massive wildfire breakouts like California and no extensive ecosystem breakdowns (yet) like in the case of Australia's coral reefs. The effects of climate change, however, don't have to be so dramatic in order for them to affect our livelihoods.

In the case of Massachusetts, the iconic cranberry growing industry, which has been the Bay state's primary agricultural product since the early 1800s, has suffered due to increased temperatures. Warmer Spring and Autumn seasons act as catalysts for the continued emergence of pests and fungi, both of which can substantially

*Undergraduate Student in the Dept. of Mathematics, Bridgewater State University, 24 Park Avenue, Bridgewater, MA 02325

†Dept. of Mathematics, Bridgewater State University, 24 Park Avenue, Bridgewater, MA 02325

reduce crop yields. Cranberries are thus more susceptible to rot, which is degrading the quality of cranberries being sold to distributors, costing farmers potential profit(Gardner, n.d.). Additionally, warmer Winters are resulting in a decrease in snowfall but more importantly in this case a lack of ice (Ellwood, Playfair, Polgar, & Primack, 2013). This denies the plants the usual cocoon-like protection provided by farmers when they flood the cranberry bogs during the winter for the purpose of creating an ice sheet above the crop (CCCGA, n.d.). Normally the ice sheet would isolate and protect the stems from damage during the winter months while they lay in a dormant state, but with the increasing temperatures the crops will be exposed. With the changes in the Massachusetts climate, cranberries' growing cycles have been interrupted as well. The plants begin to bloom with fruit when the temperature is warm enough, so the blooming period has been arriving sooner which causes added variability and complications to harvesting periods (Ellwood et al., 2013). With the added uncertainty and the real risk for substantially reduced profits, Massachusetts's farmers have begun to exit the industry altogether, which has begun to shift northwards to more suitable growing climates like the southern Canadian province Quebec(Ellwood et al., 2013).

Similar to Massachusetts, the state of Maine suffers from increased temperatures but also from a lack of snow cover that is the direct result of a warmer climate. This results in the decline of numerous tree species like the sugar maple, red maple, and birch trees since snowfall is important to their annual growth cycles . As with the Massachusetts cranberry crop, colder temperatures have historically deterred pests from damaging these deciduous tree species in Maine (Fernandez et al., 2020). These types of trees, particularly the sugar maple, are the main sources of sap refined into the maple syrup so coveted by the nation's consumers. Declines in these tree populations causes greater difficulty in meeting demand and higher prices for consumers. This not only has the potential to disrupt the natural ecosystem in Maine, but also directly impacts the American people's wallets.

Vermont is a state where more varied climatological impacts can be seen, including a similar impact on tree populations like members of the maple tree family as well as birch. Vermont's maple syrup industry is consistently the largest producer of maple syrup in the United States year after year, on average producing 3 to 4 times the amount of maple syrup Maine produces. In 2019, Maine produced approximately 580,000 gallons of the sweet syrup while Vermont achieved a staggering 2,070,000 gallons (of Agriculture, n.d.). However, climate change is expected to make hitting these target numbers more difficult, and it is often that experts claim the only reason production is maintained is that advancing technology makes up for fewer trees to tap for sap (McDonald & Schoen, n.d.). In addition to the negative impacts on maple syrup production, increased rainfall and variability with all weather patterns have made extreme precipitation weather events like storms more intense but also more frequent. In 2011, hurricane Irene passed over much of the northeast, but had its greatest impact when it landed in Vermont. Due to the existing vulnerabilities of Vermont's rivers and the added rerouting of rivers for irrigation purposes, the storm caused statewide flooding that resulted in $800 million of damages and lost income for farmers and residents at the expense of the state and federal taxpayer (State of Vermont, 2020).

In Connecticut, rising temperatures have been observed to be one of the greatest areas of concern regarding a changing climate. There are multiple reasons for this, including the effects that rising temperatures will have on the dairy industry in the state. Cows produce less milk when continuously exposed to a warmer climate, so

naturally increasing temperatures will contribute to decreases in profits in the dairy industry, which accounts for approximately 13 percent of the state's farm revenue and is valued at around $70 million. According to the Environmental Protection Agency (EPA), with increased temperatures there will be a subsequent increase in the formation of smog, a dangerous pollutant that contributes to respiratory problems, and the severity of the effects of ragweed, a plant that also damages respiratory health(Environmental Protection Agency, 2016a).

Rhode Island is yet another state that will suffer particularly from increasing temperatures, on land and in the ocean. Fish species that are integral to the state's fishing industry, like cod and lobster, are expected to decline and migrate north to cooler waters. On land, the active season for mosquitoes is getting longer, making the possibility of contracting dangerous diseases like the West Nile virus, Eastern equine encephalitis (EEE), and Lyme disease ever more likely. In addition, the threats posed by invasive species that are damaging to the New England ecosystem are becoming more real by the day (Environmental Protection Agency, 2016b). One example is that of the Asian longhorn beetle, which can destroy entire forests and threatens the existence of millions of acres of America's hardwood trees. Forestry is an important industry in New England and sufficient tree populations are necessary for us to combat climate change as well (United States Department of Agriculture, n.d.). The Asian longhorn beetle poses a threat to both.

New Hampshire has been hit extremely hard by the decrease in snowfall over the past decade. The state heavily relies on the skiing industry's revenues, but with the decline in snowfall the region has lost an estimated 10 to 20 percent of its ski season days. This represents an estimated loss of $42 million to $84 million in direct and indirect spending in New Hampshire. New Hampshire is also famous for its beautiful sights and is an iconic destination for hikers and vacationing families looking to experience a taste of nature. However, according to a 2008 New Hampshire Department of Environmental Services report there has been substantial "dulling and browning of the foliage season due to tree die-offs, species substitution, and 'climate stressed' unhealthy trees." The report goes on to note that "New Hampshire foliage travelers on average spend a total of $292 million annually", suggesting that this is yet another way in which important revenue is likely to be lost due to a changing climate, hurting the state's economy and its residents (New Hampshire Department of Environmental Services, n.d.).

Keeping in mind the severity of the impacts of changing weather outlined above, the abilities to quantify trends in such weather, monitor them, and realistically predict how they will change in the future is imperative. This paper aims to use multiple statistical techniques, including linear regression and ARIMA time series analysis, to address these concerns. Six states are discussed, with four types of weather being examined for the years 1999-2018 per state: average temperature, snowfall, snow depth, and precipitation. Thus, 24 linear regression models and 24 ARIMA time series models were needed to adequately assess the weather of each state.

The structure of this paper is as follows. In section 2 the intuition behind the use of linear regression is explained and visuals for the data and each model are produced. In section 3 the ARIMA modeling is introduced and applied to the data, again for each weather type. The process for creating the ARIMA models is much more laborious, and is explained step by step. The resulting models will be compared with various criteria like the Akaike information criterion (AIC), root mean square error (RMSE), mean absolute error (MAE), and more. Lastly, the

forecasts of the chosen models will then be compared to the real data collected for the year of 2019. Without a loss of generality, we discuss in depth the model creation process for only the case of Massachusetts and discuss the results of the analyses of the other states later.

2. Trend Analysis

2.1 Notes on Data Preparation

It is important to note a few things before discussion of the data analyses begins. The data used were all taken from the Global Historical Climatology Network (GHCN) and United States Historical Climatology Network (USHCN) databases, provided by the National Oceanic and Atmospheric Administration (NOAA). All data was either obtained or coerced into a monthly format for this study. It is important to note that the data spanned the years between 1999 and 2018, inclusive. It made sense to include every month within this time frame for precipitation and average temperature since year round information about the metrics was desired for the study. However, it did not make good sense to include all of the months within the time frame for snow depth and snowfall for obvious reasons. The study only wished to analyze months that have historically yielded substantial amounts of snow, thus a new "period of interest" was used that included months between October and May, inclusive. This is an 8 month period, so the data sets for these two metrics consist of 160 observations, while the previous two metrics have 240. Lastly, it is important to explain the difference between the snowfall and snow depth metric, as the two terms are easily confused. As defined by the NOAA, snowfall refers to the amount of fresh snow that has fallen between time points, whereas snow depth is the the total amount of snow present at the time of measurement, including both old and new snowfall.

2.2 Linear Regression

Simple linear regression is a tool that has been used for many years and has been an invaluable way for scientists to explore relationships between two variables and to ascertain if they are correlated or not. The method simply regresses one variable against another on the two axes of the two-dimensional Cartesian plane and estimates an ideal equation that minimizes the mean square error between the data and the estimated equation line. As a result, an equation is obtained that most accurately represents the data. In this case, a regression of each weather metric was performed against time. The features of the regression equations that are most important to this study are the sign and magnitude of the slope coefficient found using the least-squares method. The regression equation used has the following form:

$$Y_i = \beta_0 + \beta_1 T_i + \varepsilon_i$$

where Y_i is the weather metric, T_i is time (in months), β_0 is the intercept coefficient, β_1 is the estimated slope coefficient, and ε_i is the error term. Linear regression is valuable since it only provides a representation of the linear association between two variables, and it is precisely a linear trend that will show if there is a general decrease/increase in the observed weather.

2.3 Trends in Massachusetts Weather

As mentioned above, the features of interest regarding the least squares regression line are the sign of the slope and its magnitude. These together represent the direction of the trend within the data and the severity of it. Below in Figure 1 are visualizations of the regressions of each weather metric against time.

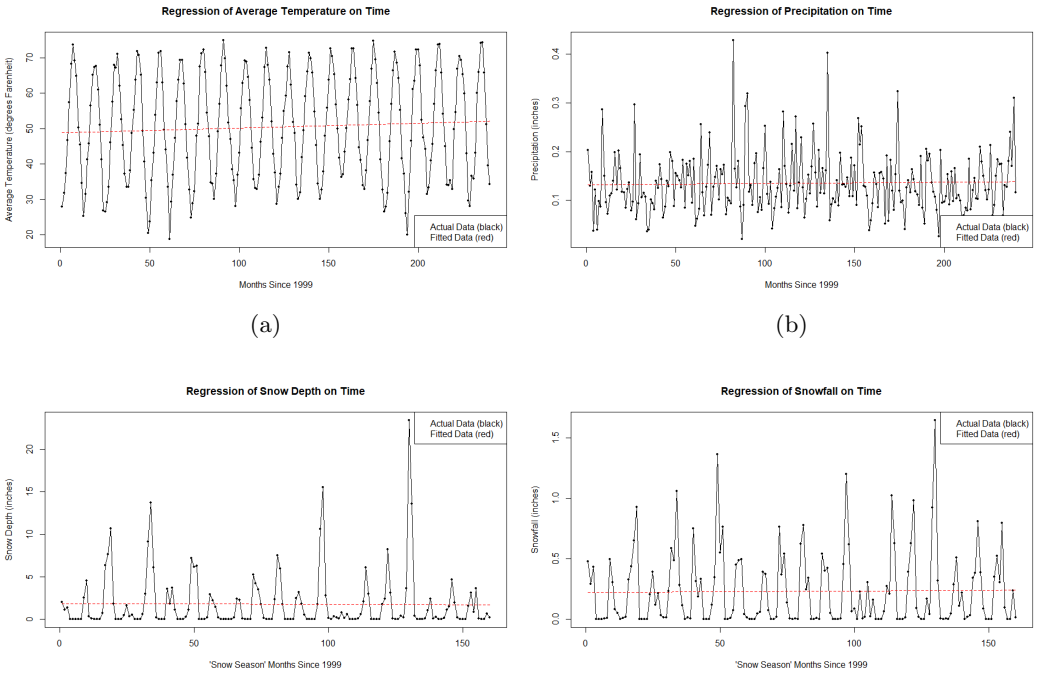


Figure 1: Regressions of Weathers Metrics against Time in Massachusetts

It can be seen that there is a general increasing trend in the amount of monthly average temperature in Massachusetts over the last 20 years, which corroborates the theory that the regional climate is warming. Monthly precipitation also exhibits a slight increasing trend, as does snowfall. The only metric that exhibits a decreasing trend is snow depth.

In Table 1, the equations for each linear regression relation can be seen. The slope coefficients for each type of weather tend to be small, however changes in climate tend to always be small and gradual over time. It is also important to keep in mind that small changes in weather tend to result in larger impacts on the stability of things like the health of wildlife in the corresponding area.

Weather Type	Regression Equation
Average Temperature	$\hat{Y} = 48.79 + 0.013T$
Precipitation	$\hat{Y} = .13 + 0.000026T$
Snowfall	$\hat{Y} = .22 + .00011T$
Snow Depth	$\hat{Y} = 1.85 - 0.00089T$

Table 1: Regression for Massachusetts Weather

Regarding the results of the linear regression lines for each of the other states, many of them corroborate the claims made about New England climate change and

reaffirm the expectations they had, whereas a few others seem to contradict. In Table 2, the outcomes of the regressions performed on each weather metric were colored either red or green to reflect an increasing trend (green) or a decreasing trend (red).

State	Precipitation	Snowfall	Snow Depth	Average Temperature
Connecticut	Decreasing	Increasing	Increasing	Increasing
Maine	Increasing	Increasing	Decreasing	Increasing
Massachusetts	Increasing	Increasing	Decreasing	Increasing
New Hampshire	Increasing	Increasing	Decreasing	Increasing
Rhode Island	Increasing	Increasing	Increasing	Increasing
Vermont	Increasing	Decreasing	Decreasing	Increasing

Table 2: Trends of New England Weather

Average temperature was the only metric for which all the outcomes of its regressions were consistent with their expectations. For the other three metrics, it appears that each have varying amounts of deference from their expectations. In general, there exists increasing trends in both average temperature and precipitation, with a minor degree of inconsistency arising from Connecticut's precipitation regression. However, the number of states in which snowfall is increasing is larger than what was expected. This could be due to other geographical factors that were not considered in this study, or it could be something more obvious. This study included data only tracing back to the year 1999, yet large scale human caused climate changing effects in the US, like the emission of greenhouse gases, date back to the the industrial revolution of the early 1800s. If a larger sample size was used in this study dating back farther in US history, the resulting trends of the linear regressions would likely be more pronounced.

3. Time Series Analysis

3.1 Motivation

Time series analysis is extremely valuable in many situations. People have a fascination about the future and have always looked for ways to predict what has yet to pass. Time series analysis is a tool that can produce such predictions under a given set of assumptions, making it extremely valuable to a great many fields, in this case climate change. Knowing what the temperature will be like in the future will serve to inform decision makers with quantitative information rather than a general qualitative assertion that temperature is on the rise.

The time series analysis was performed using the seasonal autoregressive integrated moving average (SARIMA) model, although other methods were attempted, like the Holt-Winters exponential smoothing model. In the end SARIMA was the easiest to implement and was needed to address the nonstationarity of the data.

3.2 Seasonal Autoregressive Integrated Moving Average Models

The seasonal autoregressive integrated moving average models are a subset of the many time series models and includes both autoregressive (AR) terms, moving average (MA) terms, and the use of differencing operators when the process is nonstationary (thus integrated). The model can be represented with the following notation:

$$y_t = \theta_0 + \sum_{i=1}^{p} \phi_i y_{t-i} + \sum_{i=1}^{P} \Phi_i y_{t-is} + \varepsilon_i + \sum_{i=1}^{q} \theta_i \varepsilon_{t-i} + \sum_{i=1}^{Q} \Theta_i \varepsilon_{t-is}$$

where y_t is the observation on the variable of interest at time point t, θ_0 is either the mean of the process or a deterministic trend constant, ϕ_i is the ith nonseasonal AR coefficient, Φ_i is the ith seasonal AR coefficient, ε_i is the ith error term, θ_i is the ith nonseasonal MA coefficient, Θ_i is the ith seasonal MA coefficient, and s is the periodicity/seasonality constant (Wei, 2006). For the subsequent models, the deterministic trend constant θ_0 was omitted for simplicity.

The intuition behind the ARIMA time series model is simple. It makes sense that, for example, information about whether it rained or not today and how much it rained will provide a decent idea about the rainfall the next day. This idea can be observed in the ARIMA model through the inclusion of the autoregressive terms, where y_{t-1}, y_{t-2}, ... , ε_{t-p} take on previous values of the series, which may be daily, weekly, monthly data and so on. Additionally, the moving average terms include the predictors ε_{t-1}, ε_{t-2}, ... , ε_{t-q}, which take values of the previous error terms of the series.

3.3 Building an ARIMA model for Average Temperature

A total of twenty-eight ARIMA models were created to forecast weather for this study, but only that for average temperature in Massachusetts will be outlined in this paper. The original time series can be seen in Figure 2.

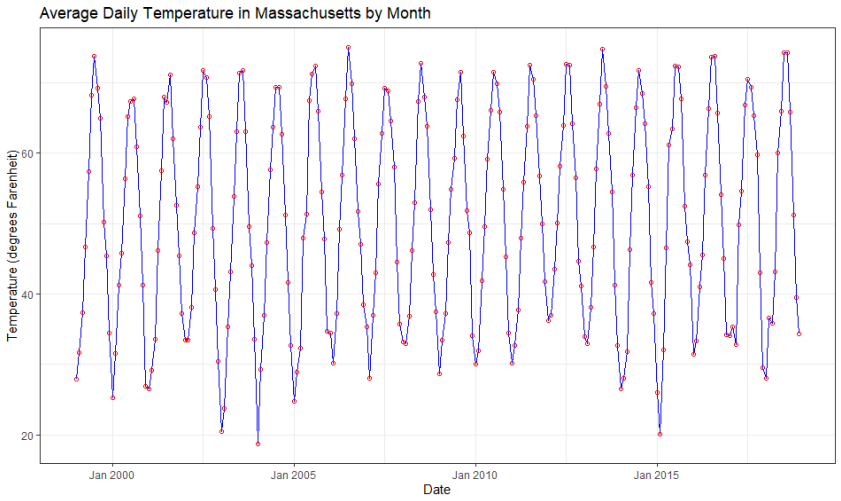
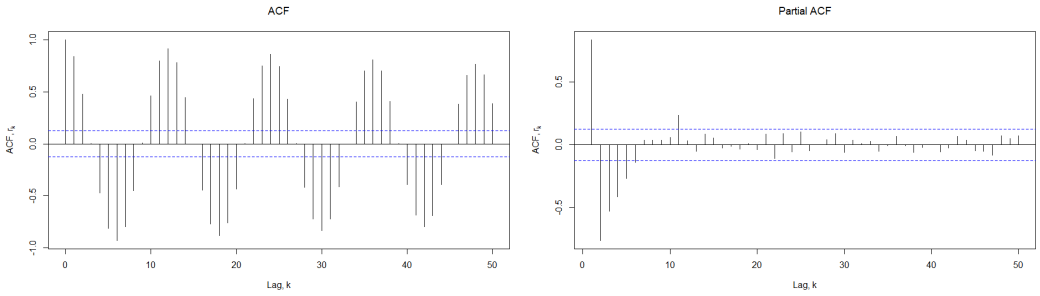


Figure 2: Time Series for Average Temperature in Massachusetts

Plotting the original time series is important because it can yield revealing information. It can be seen that the time series exhibits a constant mean with a small amount of variation from year to year. Another important observation is that the series tends to follow a similar pattern from year to year, with temperatures falling in the winter months and rising again in the summer months. This is to be expected, and this seasonality is characteristic of many weather time series.

Another important tool that is used in time series analysis is the sample autocorrelation function (ACF). This function takes all of the values of the time series and a lag value k as inputs and outputs the correlation between values of the series

k lags apart. These values can represented by y_t and y_{t+k}, and for every observation at time point t of the series the observation pairs k lags apart are regressed and their correlation computed. This is where the term autocorrelation comes from, since it finds the correlation between a time series and a lagged version of itself in time (Montgomery, Jennings, & Kulahci, 2016). The function can be seen in figure 3, along with another tool that is known as the sample partial autocorrelation function (PACF).



(a) (b)

Figure 3: Sample Autocorrelation and Partial Autocorrelation Functions for Average Temperature in Massachusetts

The goal of plotting the ACF is to examine the autocorrelation that the time series process exhibits at each of its time lags. Before proceeding, it is necessary to define what stationarity of a time series is, as it is an important attribute that is required for producing accurate forecasting models. Regarding this definition, Montgomery et al. (2016) state that "If a time series is stationary this means that the joint probability distribution of any two observations, say y_t and y_{t+k}, is the same for any two time periods t and $t+k$ that are separated by the same interval k" (p. 36). This is referred to as strict stationarity and does not commonly occur in practice with real world data. Instead, it is sufficient to show that the time series has a finite mean and variance, which is referred to as weak stationarity.

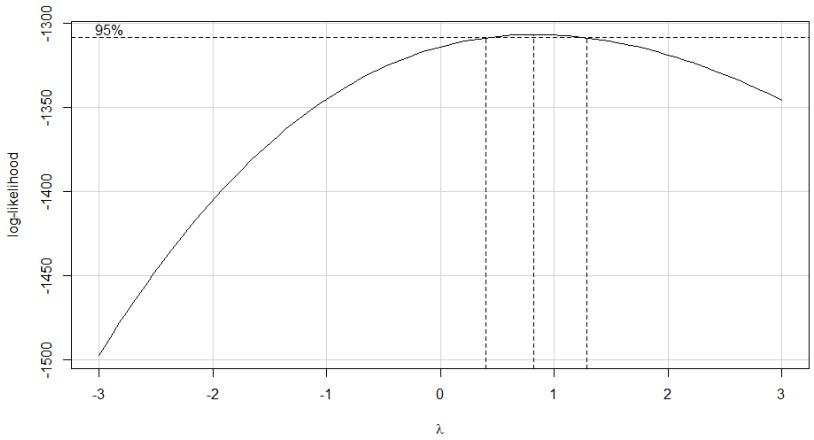


Figure 4: Box-Cox Transformation Lambdas plotted against their Log-Liklihood values

Examination of Figure 2 has already shown that the average temperature time series exhibits a near constant mean and variance. Time series commonly exhibit non-constant variance issues, which can be corrected most of the time by applying a Box-Cox transformation. Checking this assumption is worthwhile since non-constant variance has severe consequences regarding the accuracy of forecast intervals. By viewing the plot in Figure 4, it clear that the optimal λ value for a Box-Cox transformation is close to 1, so no variance-corrective transformation should be performed. However, there also appears to be a clear seasonality, which can be confirmed by the periodic spikes every twelve lags in the ACF plot in Figure 3. Note that the periodic spikes all seem to protrude from the bounds of the significance limits in the plot as well, signifiying that there exists a detectable amount of serial correlation at those lags. There are multiple ways of addressing/removing seasonality, but the one employed here is known as seasonal differencing.

3.4 Seasonal and Non-Seasonal Differencing

Non-seasonal differencing is a process that is used for removing a positive or negative linear trend in time series data. It subtracts the previous value of the time series from the current value, and it can be expressed by

$$w_t = y_t - y_{t-1},$$

where w_t is the newly differenced value of the series. This operation is done to every value of the series. Naturally, the first value of the series will have no previous value and must be dropped from the newly differenced series such that the new series has one less observation than the orginal. However, standard differencing is not helpful here since there is no linear trend to be removed, but a seasonal trend. Instead, a differencing operation will be performed that can be expressed by

$$z_t = y_t - y_{t-s},$$

where z_t is the seasonally differenced value of the series, and s is the lag interval with which the seasonality repeats. Recall that the ACF showed that there are significant spikes that repeat every twelve lags, indicating a monthly seasonality. This makes sense since the data are monthly observations and weather occurs in cycles. Thus, the value of s here would be 12 and the seasonal differencing operation will subtract from each value of the series the value that occurred 12 time steps (months in this case) previously. As before, the resulting time series will have fewer data points, here a difference of twelve.

After performing this operation, it is necessary to plot the differenced time series, as well as its accompanying ACF and PACF plots. The differenced time series no longer exhibits a clear, predictable seasonality. Instead, the tendency for the series to commit to bouts of increasing and decreasing behavior is much more sporadic. Note also the ACF and PACF of this series in Figure 6. The persistent oscillating pattern in the ACF has mostly been removed. There remains autocorellation in the time series, although to a lesser extent than before. yet it cannot be ignored, and at this stage a SARIMA model must be fit.

3.5 Model Fitting

Now that the average temperature time series has been rendered a minimum of weakly stationary, fitting a SARIMA model is approporiate. In addition to aiding

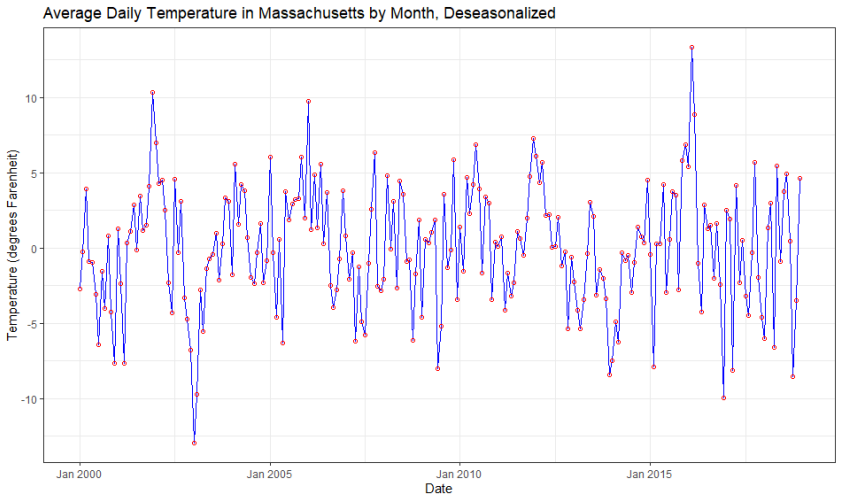
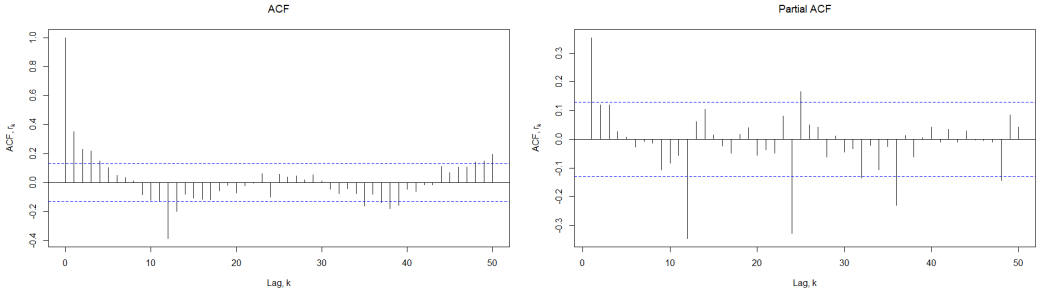


Figure 5: Seasonally Differenced Time Series for Average Temperature



(a) (b)

Figure 6: Sample Autocorrelation and Partial Autocorrelation Functions for Seasonally Differenced Time Series

the modeler in making a nonstaionary series stationary, the ACF and the PACF are also valuable tools in determining the orders of an ARIMA model. There is a tendency for certain patterns that arise in the ACF and PACF that correspond with certain ARIMA models. Therefore, by carefully observing the two correlation based functions it is possible to determine the correct type and number of terms(AR, MA, etc.) to include in the model.

In general, it should be noted that the ACF is used for determining the order of the moving average and seasonal moving average terms and the PACF is used for determining the order of the autoregrressive and seasonal autoregressive terms. There are a few other rules that govern reading the plots, and they can be seen in Table 3. There are also similar rules that govern the interpretation of seasonal patterns which will be discussed after determining the nonseasonal term orders.

Process	ACF	PACF
AR(p)	Tails off after exponential decay or damped sine wave	Cuts off after lag p
MA(q)	Cuts off after lag q	Tails off after exponential decay or damped sine wave
ARMA(p,q)	Tails off after lag $(q - p)$	Tails off after lag $(p - q)$

Table 3: ARIMA Model Identification Rules

Beginning with the ACF of the seasonally differenced time series found in Figure 6, it can be observed that there is an exponential decay pattern that results in the ACF dropping off quite quickly. This and the fact that the PACF cuts off after the very first lag imply that the nonseasonal portion of the model is an AR(1) process. A nonseasonal moving average term is not expected to be included at this point.

As for the seasonal portion of the model, the way in which the ACF and PACF are used is relatively the same in that we apply the same rules mentioned in the chart above, except that we will restrict our scope to only the seasonal lags (the ones that occur every twelve lags). The ACF seems to cut off after one seasonal lag, which can be observed as the large spike at lag twelve, whereas the PACF seasonal lags seem to follow a pattern of exponential decay. This implies that a seasonal MA(1) term must be included.

Now that the terms have been identified, we can state our tentative model as an $\text{ARIMA}(1,0,0)(0,1,1)_{12}$ model. However, it is important to recognize that this model identification process tends to be mildly subjective and should be used as a way to acquire a rough idea of what the model should be, which is why this model is tentative. A more accurate way to find the most appropriate model is to compare the tentative model's selection criteria values with those of other models that either increase or decrease the orders of the terms by one. Criteria used for model selection include Mean Error (ME), Root Mean Square Error (RMSE), Mean Absolute Error (MAE), Mean Percentage Error (MPE), Mean Absolute Percentage Error (MAPE), Mean Absolute Scaled Error (MASE), Akaike Information Criterion (AIC), and the Bayesian Information Criterion (BIC). A useful to make these comparisons is to create a chart that lists each model that was fit to the data and their corresponding values for each of the model identification criteria.

The model selection criteria all follow the same general rule that the lowest value indicates the best model, except for Mean Error and Mean Percentage Error, for which a value closer to 0 is best since the the error in either the negative or positive direction are both possible and undesirable.

According to Table 4, the model with the most supporting model selection criteria (seen as the numbers highlighted in green) appears to be the

Model	ME	RMSE	MAE	MPE	MAPE	MASE	AIC	BIC
$ARIMA(0,0,0)(0,1,0)_{12}$	0.001	3.928	3.024	-0.562	7.212	0.413	1284.549	1288.030
$ARIMA(0,0,0)(0,1,1)_{12}$	0.341	3.093	2.323	0.128	5.631	0.318	1192.530	1199.491
$ARIMA(1,0,0)(0,1,1)_{12}$	0.244	2.824	2.189	-0.018	5.216	0.299	1156.589	1167.031
$ARIMA(1,0,1)(0,1,1)_{12}$	0.188	2.788	2.150	-0.151	5.101	0.294	1154.069	1167.991

Table 4: Model Identification Criteria Comparison

$ARIMA(1,0,1)(0,1,1)_{12}$, which differs from our tentative model by including a nonseasonal moving average component. Therefore, we will choose this as our final model and examine the diagnostic plots to ensure that the assumptions for the ARIMA model have been met.

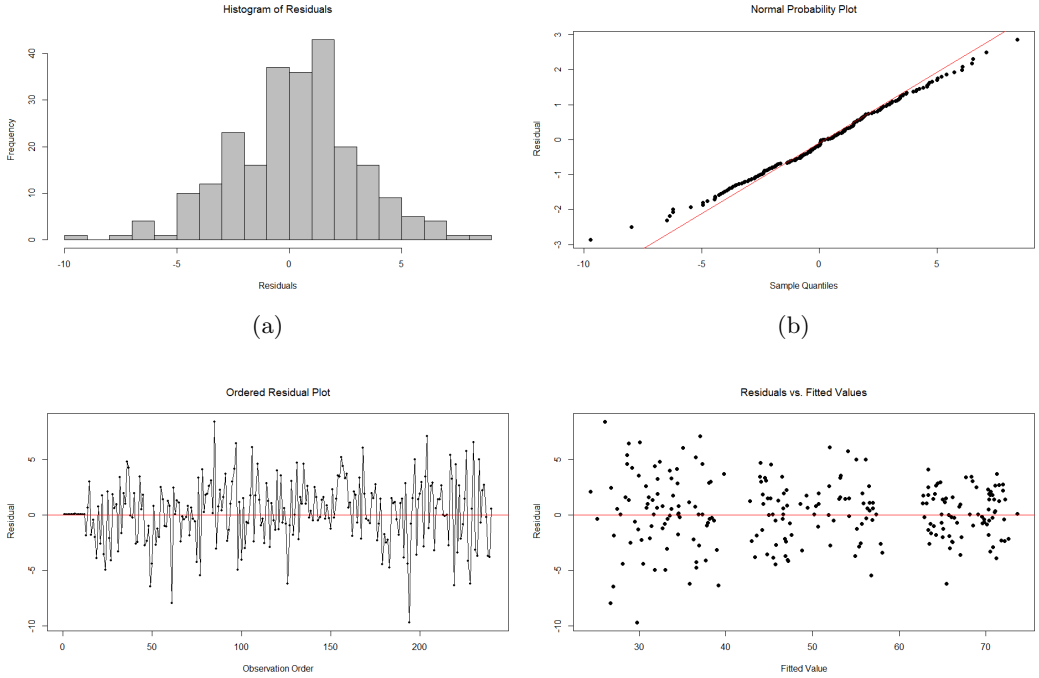


Figure 7: Model Diagnostics 1

The diagnostics plots in Figure 7, specifically the histogram of model residuals and the normal probability plot, suggest that the normality assumption is satisfied. There do seem to be some heavy tails in the normal probability plot, but the normality assumption is robust and can be roughly assumed. According to the ordered residual plot and the plot of the residuals against fitted values, the second assumption of constant variance is upheld as well.

In addition to checking normality and constant variance, it is important to ensure that the model has effectively eliminated all or most of the serial correlation in the residuals and is stationary. The residual ACF and residual PACF in Figure 8 show the first 25 lags in the model residuals, only one of which is significant. Despite barely extending past the significance limits, this one significant lag is most likely a random shock in the data and it is unlikely to indicate remaining serial correlation. Ljung-Box statistics were computed for each residual ACF lag to test for signifi-

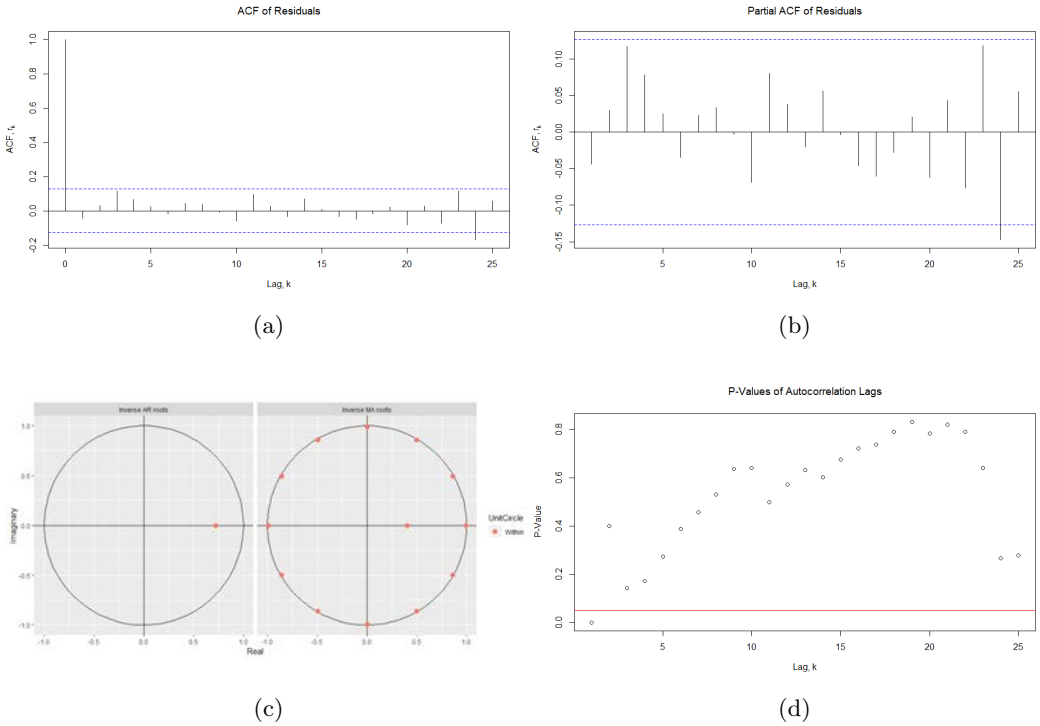


Figure 8: Model Diagnostics 2

cance, and the graph in Figure 8(d) plots the p-values of these statistics against the lag number.These p-values were tested against a $\alpha = .05$ threshold, which is represented by a red line. This indicates further that no significant correlation is present. The unit root graph shows the inverse roots of the moving average and autoregressive terms in the model. The inverse AR root is clearly inside the unit circle, indicating that the model's AR term is stationary. The inverse MA roots are less clearly inside the unit circle, but they do in fact lie within the circle, indicating that the model's MA terms are invertible.

3.6 Model Forecasting

Now that model assumptions have been upheld through diagnostics, we can now use our model to forecast future observations. Using the ARIMA$(1,0,1)(0,1,1)_{12}$ model we decided on, we plot the model's fitted values overlaid with the original time series

Judging from the the time series in Figure 9, the model captures the trend of the temperature data rather well. In fact, most of the SARIMA models that were compared earlier through the various criteria were able to capture the trend of the data, and the choice of our model was simply one that minimized in sample error. In practice, most of these models would have sufficed. The forecasts are in blue, and can be seen in closer detail in Figure 10. Since the data used to train the model was data from 1999 to 2018, temperature data from 2019 was used to evaluate the accuracy of the model. All of the 2019 data falls comfortably within the forecast intervals, indicating that the model is a functional tool for forecasting the average daily temperature by month for the state of Massachusetts.

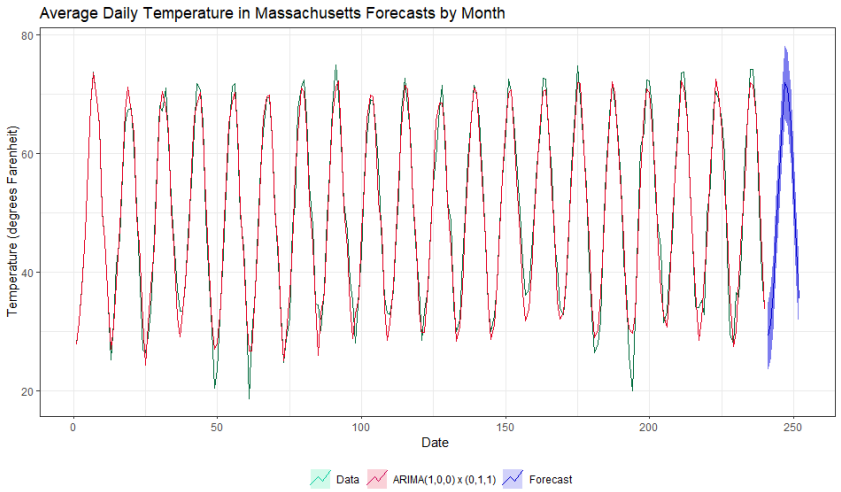


Figure 9: Average Temperature Data and Model Overlay with Forecasts

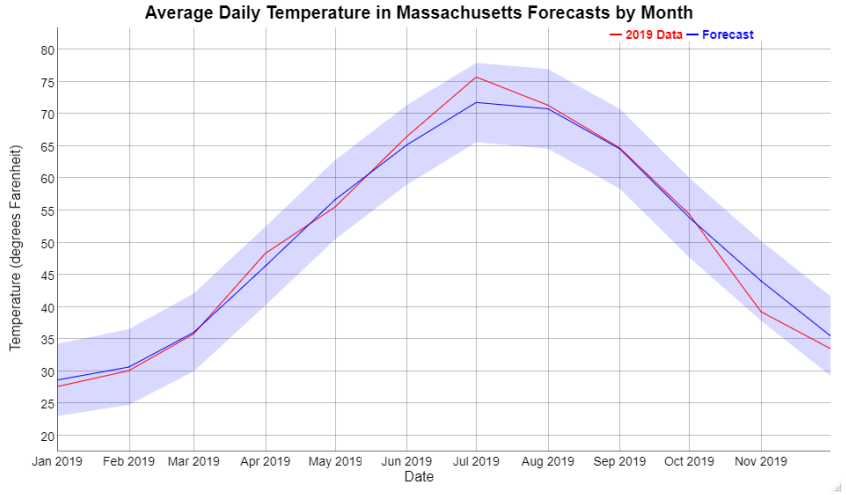


Figure 10: Forecasts, Forecast intervals and 2019 Data Comparison

4. Concluding Comments

Using linear regression, general linear trends were computed and observed for average temperature, snowfall, snow depth, and precipitation. The different increasing/decreasing trends that were observed were purely for exploratory data analysis and for informing the average person. Unfortunately, the stochastic nature of the weather data results in the linear regression models being unacceptable for significance testing since the models diagnostics indicated substantial non-normality.

The SARIMA model created for average temperature was able to successfully capture the trends within the data and forecast future observations. Despite the success of such a model being fit to the average temperature data, SARIMA modeling techniques were unsuccessful in capturing the trends of the snow depth, snowfall, and precipitation data. Future considerations for this study would be to apply other time series modeling techniques, such as Holt-Winters exponential smoothing for the snowfall and snow depth data since they follow a clear seasonal trend.

Precipitation, due to its nonseasonal nature (the New England precipitation data exhibit far more consistency throughout the year than do the other weather

State	Precipitation	Average Temperature	Snowfall	Snow Depth
MA	$(0,0,0)(0,0,0)$	$(1,0,1)(0,1,1)_{12}$	$(0,0,0)(0,1,1)_8$	$(1,0,0)(0,1,1)_8$
CT	$(0,0,0)(0,0,0)$	$(1,0,1)(0,1,1)_{12}$	$(0,0,0)(1,1,1)_8$	$(0,0,0)(1,1,1)_8$
ME	$(0,0,0)(0,0,0)$	$(1,0,1)(0,1,1)_{12}$	$(0,0,0)(1,1,1)_8$	$(1,0,0)(0,1,1)_8$
VT	$(1,0,0)(1,0,0)$	$(1,0,0)(0,1,1)_{12}$	$(0,0,0)(2,1,0)_8$	$(1,0,0)(3,1,0)_8$
RI	$(0,0,0)(0,0,0)$	$(1,0,0)(0,1,1)_{12}$	$(0,0,0)(0,1,1)_8$	$(1,0,0)(3,1,0)_8$
NH	$(0,0,0)(0,0,0)$	$(1,0,1)(0,1,1)_{12}$	$(0,0,0)(2,1,1)_8$	$(1,0,0)(3,1,0)_8$

Table 5: All (S)ARIMA Models for Massachusetts and other States

data mentioned), would benefit more from standard exponential smoothing techniques. Additionally, there is another technique known as Improved Complete Ensemble Empirical Mode Decomposition with Adaptive Noise (ICEEMDAN). It is an improved variation of Empirical Mode Decomposition, which was initially an important aspect of the Hilbert-Huang Transform, and it may be able to more accurately capture the trends within the highly volatile precipitation data.

Although this paper discusses only the modeling of the average temperature data for Massachusetts in depth, the models for the other time series for Massachusetts and the other states can be found in Table 5. Additionally, the graphs and code for all time series and linear regression models created can be found on GitHub through the link https://github.com/AndrewDisher/New-England-Weather.

References

CCCGA. (n.d.). *How cranberries grow: Winter.* Retrieved 10/03/2020, from https://www.cranberries.org/how-cranberries-grow/winter

Ellwood, E., Playfair, S., Polgar, C., & Primack, R. (2013, 09). Cranberry flowering times and climate change in southern massachusetts. *International journal of biometeorology*, *58*. doi: 10.1007/s00484-013-0719-y

Environmental Protection Agency. (2016a). *What climate change means for connecticut.* Retrieved 10/03/2020, from https://19january2017snapshot.epa.gov/sites/production/files/2016-09/documents/climate-change-ct.pdf

Environmental Protection Agency. (2016b). *What climate change means for rhode island.* Retrieved 10/03/2020, from https://19january2017snapshot.epa.gov/sites/production/files/2016-09/documents/climate-change-ri.pdf

Fernandez, I., Birkel, S., Schmitt, C., Simonson, J., Lyon, B., Pershing, A., ... Mayewski, P. (2020). *Maine's climate future—2020 update.* Retrieved from https://climatechange.umaine.edu/wp-content/uploads/sites/439/2020/02/2020_Maines-Climate-Future-508-ADA-compliant.pdf

Gardner, S. (n.d.). *Climate change comes to the cranberry bog.* Retrieved 10/03/2020, from https://www.marketplace.org/2012/11/19/climate-change-comes-cranberry-bog/

McDonald, J., & Schoen, J. W. (n.d.). *Vermont's maple syrup industry is recovering from decades of decline. climate change could put that at risk.* Retrieved 10/03/2020, from https://www.cnbc.com/2019/09/28/climate-change-could-hurt-vermonts-maple-syrup-industry-recovery.html

Montgomery, D. C., Jennings, C. L., & Kulahci, M. (2016). *Introduction to time series analysis and forecasting, 2nd edition.* 111 River St. Hoboken,NJ 07030 USA: John Wiley and Sons, Inc.

New Hampshire Department of Environmental Services. (n.d.). *Global climate change and its impact on new hampshire.* Retrieved 10/03/2020, from https://www.des.nh.gov/organization/commissioner/pip/factsheets/ard/documents/ard-23.pdf

of Agriculture, U. S. D. (n.d.). *United states maple syrup production.* Retrieved 10/03/2020, from

State of Vermont. (2020). *Climate change in vermont.* Retrieved 10/03/2020, from https://climatechange.vermont.gov/our-changing-climate/what-it-means/flooding

United States Department of Agriculture. (n.d.). *Asian longhorned beetle.* Retrieved 10/03/2020, from https://www.aphis.usda.gov/aphis/resources/pests-diseases/hungry-pests/the-threat/asian-longhorned-beetle/asian-longhorned-beetle

Wei, W. S. (2006). *Time series analysis; univariate and multivariate methods, 2nd edition.* 1 Lake Street Upper Saddle River, NJ 07458 United States: Pearson Education, Inc.