To Appear: The International Journal Neural, Parallel & Scientific Computations,
Volume 16, 2008

A Temperature Forecasting Model for the Continental United States

Shou Hsing Shih and Chris P. Tsokos
Department of Mathematics and Statistics, University of South Florida
Tampa, Florida 33620

Abstract

Two major entities that play a major role in understanding Global Warming is temperature and Carbon Dioxide. The purpose of the present study is to utilize historical temperature in the Continental United States from 1895 to 2007 to develop a forecasting process to estimate future average monthly temperatures. In addition, we shall study through our modeling if there is a difference in the two methods that are being used to collect and massage the temperatures in the Continental United States.

Keywords – Time Series Forecasting, ARIMA, Multiplicative ARIMA, Global Warming, Temperature

1. INTRODUCTION

Temperature plays a very important role in Global Warming and its relation with Carbon Dioxide. The aim of the present study is to develop a statistical forecasting model for the temperature in the Continental United States. There are two methods being used in recording temperatures and we shall refer them as Version 1 and Version 2 data sets. Thus, an additional aim in the present study is to determine if the two methods of recording temperatures are indeed different. Version 1 data was collected by the United States Climate Division, USCD, and Version 2 data by the United States Historical Climatology Network, USHCN.

The Version 1 data set consists of monthly mean temperature and precipitation for all 344 climate divisions in the contiguous U. S. from January 1895 to June 2007. The data is adjusted for time of observation bias, however, no other adjustments are made for inhomogeneities. These inhomogeneities include changes in instrumentation, observer, and observation practices, station and instrumentation moves, and changes in station composition resulting from stations closing and opening over time within a division.

The Version 2 data set was first become available in July 2007, and it consists of data from a network of 1219 stations in the contiguous United States that were defined by scientists at the Global Change Research Program of the U. S. Department of Energy at National Climate Data Center. A methodology was developed and applied to test known station changes for their impact on the homogeneity, and necessary adjustments were made if the changes caused a statistically significant response in the time series. They claim that the data set is a consistent network through time, which minimizes any biasing due to network changes through time. For additional information concerning Version 1 of the data, see (Easterling & Peterson, 1995; Karl et al., 1986; Karl & Williams, 1987; Karl et al., 1988; Karl et al., 1990; Peterson & Easterling, 1994; Quayle et al., 1991). Information for Version 2 of the time series, see (Alexandersson & Moberg, 1997; Baker, 1975; Easterling et al., 1996; Easterling et al., 1999; Hughes et al., 1992; Karl et al., 1990; Karl et al., 1988; Karl et al., 1986; Karl & Williams, 1987; Lund & Reeves, 2002; Menne & Williams, 2005; Quinlan et al., 1987; Vose et al., 2003; Wang, 2003). Graphical presentations of both data sets are given by Figure 1.1 and 1.2.

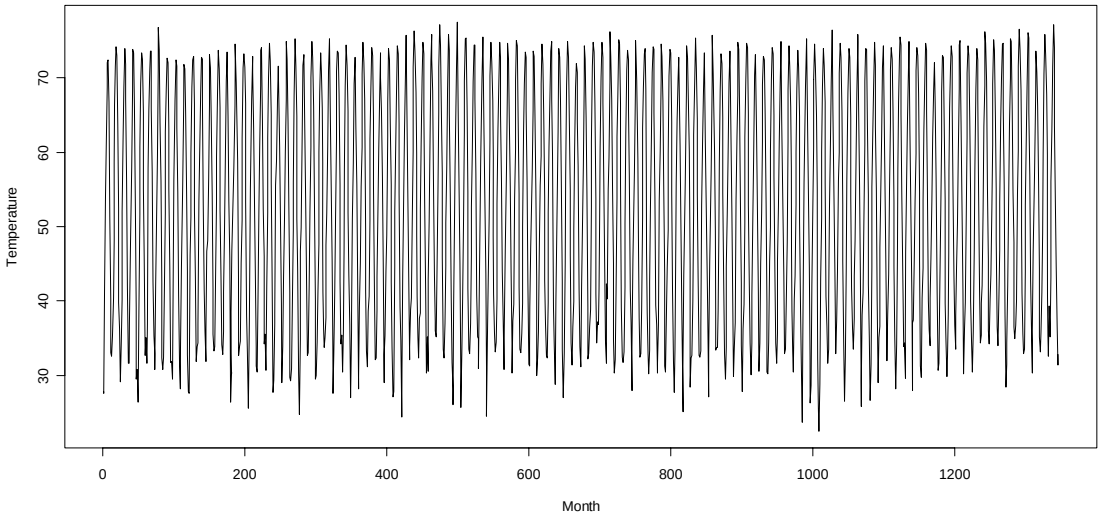

Figure 1.1 Time Series Plot for Monthly Temperature from the Continental United States 1895-2007 (Version 1 Dataset)

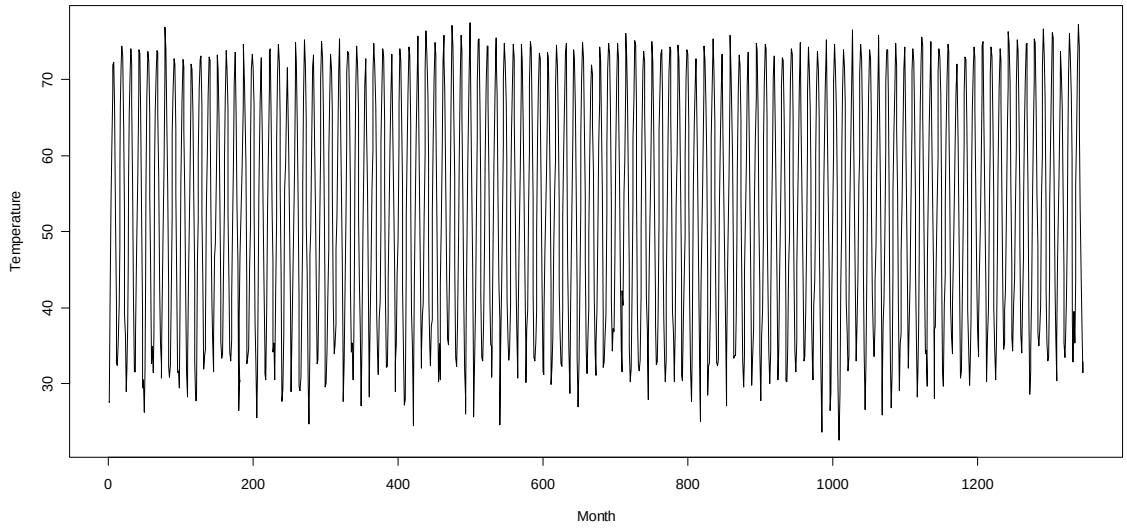

Figure 2.2 Time Series Plot for Monthly Temperature from the Continental United States 1895-2007 (Version 2 Dataset)

2. ANALYTICAL PROCEDURE

The multiplicative seasonal autoregressive integrated moving average, ARIMA model is defined by

$$\Phi_P(B^s)\phi_p(B)(1-B)^d(1-B^s)^D x_t = \theta_q(B)\Gamma_Q(B^s)\varepsilon_t \qquad (2.1)$$

where p is the order of the autoregressive process, d is the order of regular differencing, q is the order of the moving average process, P is the order of the seasonal autoregressive process, D is the order of the seasonal differencing, Q is the order of the seasonal moving average process, and the subindex s refers to the seasonal period. We shall denote the subject model by $\text{ARIMA}(p,d,q)\times(P,D,Q)_s$, and $\phi_p(B),\theta_q(B),\Phi_P(B^s),\Gamma_q(B^s)$ defined as follows:

$$\phi_p(B) = (1-\phi_1 B-\phi_2 B^2-\ldots-\phi_p B^p)$$

$$\theta_q(B) = (1-\theta_1 B-\theta_2 B^2-\ldots-\theta_q B^q)$$

$$\Phi_P(B^s) = 1-\Phi_1 B^s-\Phi_2 B^{2s}-\ldots-\Phi_P B^{Ps}$$

and

$$\Gamma_Q(B^s) = 1-\Gamma_1 B^s-\Gamma_2 B^{2s}-\ldots-\Gamma_Q B^{Qs}.$$

The order of the multiplicative ARIMA model determines the structure of the model, and it is essential to have a good methodology in terms of developing the

forecasting model. In the present study, we start with addressing the issue of the seasonal subindex *s*. After we examine the original data, shown by Figure 1.1 and 1.2, we have reason to believe the monthly temperature of the Continental United States behaves as a periodic function with a cycle of 12 months. Hence, we let the seasonal subindex $s = 12$. In time series analysis, one cannot proceed with a model building procedure without confirming the stationarity of a given stochastic realization, thus, we test the overall stationarity of the series by using the method introduced by Kwiatkowski, D., Phillips, P. C. B., Schmidt, P., and Shin, Y in 1992, (Kwiatkowski et al., 1992).

Once the order of the differencing is identified, it is common for one ARIMA $(p,d,q)\times(P,D,Q)_s$ model that we have several sets of (p,q,P,Q) that are all adequately representing a given set of time series. Akaile's information criterion, AIC, (Akaile, 1974), was first introduced by Akaile in 1974 plays a major role in our model selecting process. We shall choose the set of (p,q,P,Q) that produces the smallest AIC. Another important aspect in our model selection process is to determine the seasonal differencing, *D*, the goal is to select a smaller AIC without complicating the selected model. Hence, we only compute the AIC for both $D = 0$ and $D = 1$ based on our previous selection of the orders (p,d,q,P,Q), and choose the model with smaller AIC to be our final model.

Below we summarize the model identifying procedure:

- Determine the seasonal period *s*.
- Check for stationarity of the given time series $\{x_t\}$ by determining the order of differencing *d*, where $d = 0,1,2,...$ according to KPSS test, until we achieve stationarity.
- Deciding the order *m* of the process, for our case, we let $m = 5$ where $p + q + P + Q = m$.
- After (d,m) being selected, listing all possible configurations of (p,q,P,Q) for $p + q + P + Q \le m$.
- For each set of (p,q,P,Q), estimates the parameters for each model, that is, $\phi_1, \phi_2, ..., \phi_p, \theta_1, \theta_2, ..., \theta_q, \Phi_1, \Phi_2, ..., \Phi_P, \Gamma_1, \Gamma_2, ..., \Gamma_Q$.
- Compute the AIC for each model, and choose the one with smallest AIC.

- After (p,d,q,P,Q) is selected, we determine the seasonal differencing filter by selecting the smaller AIC between the model with $D=0$ and $D=1$.
- Our final model will have identified the order of (p,d,q,P,D,Q).

In order to determine how good our proposed model is, we shall define several statistical criteria that we shall use to evaluate the subject forecasting model. The residuals of the model, $r_t = x_t - \hat{x}_t$, where x_t and $\hat{x}_t$ are the actual value and predicted value, respectively. Mean of the residuals, $\bar{r} = \frac{\sum_{t=1}^{n} r_t}{n}$. Variance of the residuals, $S_r^{\,2} = \frac{\sum_{t=1}^{n} (r_t - \bar{r})}{n-1}$. Standard deviation of the residuals, $S_r = \sqrt{S_r^{\,2}}$. Standard error of the residuals, $SE = S_r^{\,2} \Big/ \sqrt{n}$. Mean square error, $MSE = \frac{\sum_{t=1}^{n} r_t^{\,2}}{n}$.

3. DEVELOPMENT OF FORECASTING MODELS

The historical temperature data for the continental United States that we shall use are shown by Figure 1 and 2. A visual inspection does not show any obvious trends being present. Thus, we let the seasonal period $s=12$. Following the step-by-step procedure we described above, we found that the model best characterizes the average monthly temperature of the Continental United States for both Version 1 and 2 is a ARIMA$(2,1,1)\times(1,1,1)_{12}$ process, analytical given by

$$(1-\Phi_1 B^{12})(1-\phi_1 B-\phi_2 B^2)(1-B)(1-B^{12})x_t = (1-\theta_1 B)(1-\Gamma_1 B^{12})\varepsilon_t \qquad (3.1)$$

Expanding both sides of the above ARIMA, we have

$$\begin{aligned}&[1-(1+\phi_1)B+(\phi_1-\phi_2)B^2+\phi_2 B^3-(1+\Phi_1)B^{12}+(1+\phi_1+\Phi_1+\phi_1\Phi_1)B^{13}\\&+(\phi_2+\phi_2\Phi_1-\phi_1-\phi_1\Phi_1)B^{14}-(\phi_2+\phi_2\Phi_1)B^{15}+\Phi_1 B^{24}-(\phi_1+\Phi_1)B^{25}\\&+(\phi_1\Phi_1-\phi_2\Phi_1)B^{26}+\phi_2\Phi_1 B^{27}]x_t = (1-\theta_1 B-\Gamma_1 B^{12}+\theta_1\Gamma_1 B^{13})\varepsilon_t\end{aligned}$$

Simplify it, we get

$$\begin{aligned}&x_t-(1+\phi_1)x_{t-1}+(\phi_1-\phi_2)x_{t-2}+\phi_2 x_{t-3}-(1+\Phi_1)x_{t-12}+(1+\phi_1+\Phi_1+\phi_1\Phi_1)x_{t-13}\\&+(\phi_2+\phi_2\Phi_1-\phi_1-\phi_1\Phi_1)x_{t-14}-(\phi_2+\phi_2\Phi_1)x_{t-15}+\Phi_1 x_{t-24}-(\phi_1+\Phi_1)x_{t-25}\\&+(\phi_1\Phi_1-\phi_2\Phi_1)x_{t-26}+\phi_2\Phi_1 x_{t-27} = \varepsilon_t-\theta_1\varepsilon_{t-1}-\Gamma_1\varepsilon_{t-12}+\theta_1\Gamma_1\varepsilon_{t-13}\end{aligned}$$

Thus, the one-step ahead forecasting model for Version 1 data is given by

$$\begin{aligned}\hat{x}_t &= 1.0941x_{t-1} - 0.057x_{t-2} - 0.0371x_{t-3} + 0.9954x_{t-12} - 1.0891x_{t-13} + \\ &\quad 0.0567x_{t-14} + 0.0369x_{t-15} + 0.0046x_{t-24} + 0.0895x_{t-25} - 0.0004x_{t-26} + \\ &\quad 0.00017x_{t-27} - 0.9861\varepsilon_{t-1} - 0.9742\Gamma_1\varepsilon_{t-12} + 0.9607\varepsilon_{t-13}\end{aligned} \tag{3.2}$$

and the one-step ahead forecasting model for Version 2 data is given by

$$\begin{aligned}\hat{x}_t &= 1.0952x_{t-1} - 0.0556x_{t-2} - 0.0396x_{t-3} + 0.9964x_{t-12} - 0.9009x_{t-13} + \\ &\quad 0.0554x_{t-14} + 0.0395x_{t-15} + 0.0036\Phi_1x_{t-24} + 0.0916x_{t-25} + 0.0002x_{t-26} + \\ &\quad 0.00014x_{t-27} - 0.9855\varepsilon_{t-1} - 0.9741\varepsilon_{t-12} + 0.9599\varepsilon_{t-13}\end{aligned} \tag{3.3}$$

Note the closeness of the two forecasting models.

4. EVALUATION OF THE PROPOSED MODELS

We begin by forecasting for the last one hundred observations the monthly average temperature in the Continental United States for both Version 1 and 2, using the models given by expression 3.2 and 3.3. A graphical presentation of the results are presented below by Figure 4.1 and 4.2.

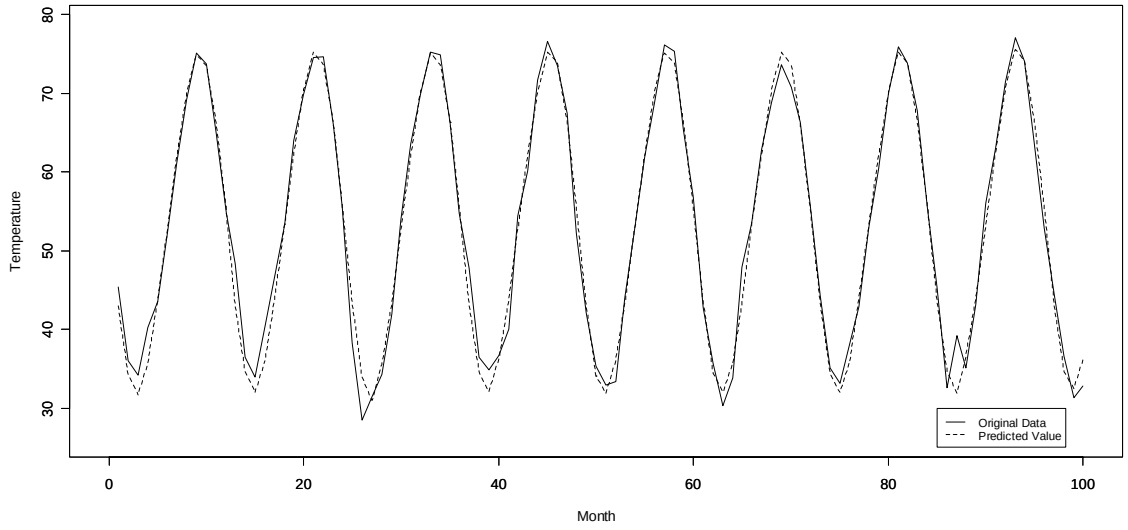

Figure 4.1 Monthly Temperature VS. Our Predicted Values for the Last 100 Observations (Version 1 Dataset)

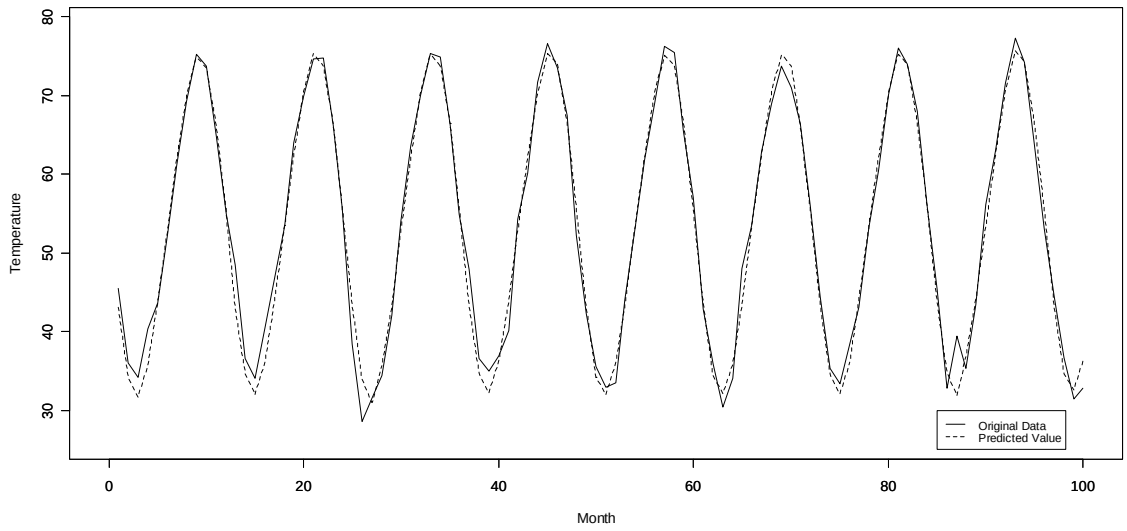

Figure 4.2 Monthly Temperature VS. Our Predicted Values for the Last 100 Observations (Version 2 Dataset)

As can be observed that both models are similar and the one-step ahead forecasting is quite good, except the temperature of January 2006 took an unexpected turn. We identify this inconsistency a possible outlier.

We proceed to calculate the residuals estimates, $r_t = x_t - \hat{x}_t$, for both forecasting process given by (3.2) and (3.3). The results are graphically presented below by Figure 4.3 and 4.4.

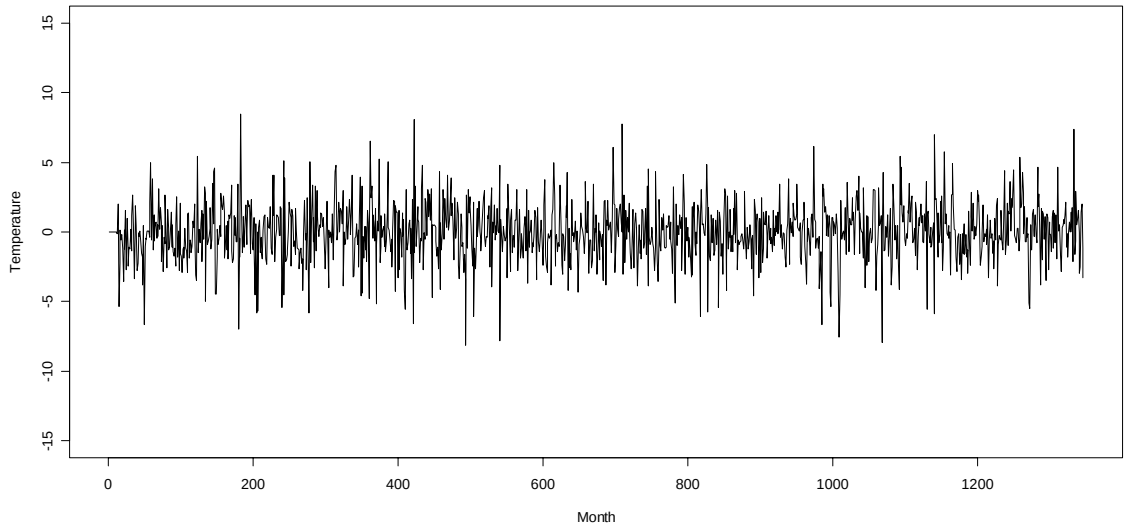

Figure 4.3 Residual Plot for Monthly Temperature on Continental United States 1895-2007 (Version 1 Dataset)

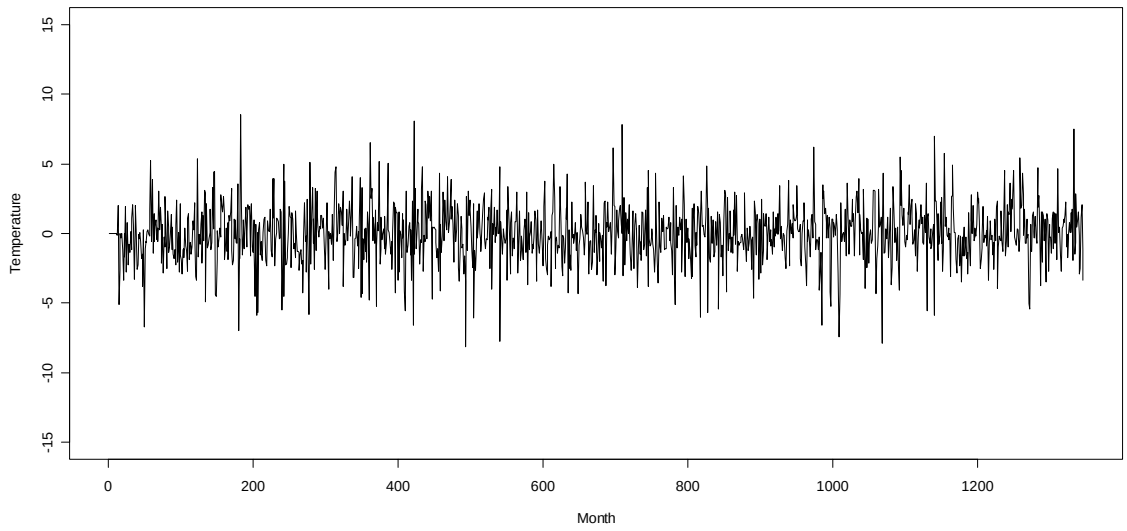

Figure 4.4 Residual Plot for Monthly Temperature on Continental United States 1895-2006 (Version 2 Dataset)

We observe that the residuals are quite small and isolating around the zero axis as expected. It indicates that both models are good models in predicting the Version 1 and Version 2 of the time series.

Next, we evaluate the mean of the residuals, $\bar{r}$, the variance, S_r^2, the standard deviation, S_r, standard error, *SE*, and the mean square error, *MSE*. The results are presented below by Table 4.1 and 4.2, for Version 1 and Version 2 data, respectively.

Table 4.1 Basic Evaluation Statistics (Version 1 Dataset)

$\bar{r}$	S_r^2	S_r	*SE*	*MSE*
-0.008512476	4.331902	2.081322	0.05673052	4.328756

Table 4.2 Basic Evaluation Statistics (Version 2 Dataset)

$\bar{r}$	S_r^2	S_r	*SE*	*MSE*
-0.01310953	4.323726	2.079357	0.05667696	4.320685

We observe that all evaluation criteria support the quality of the proposed forecasting model. We can also conclude the similarity of the two models. Thus, it raises the question

is the effort to collect two data sets implement two different procedures by two agencies necessary?

We have demonstrated that our proposed models are capable of representing the past monthly average temperature of the Continental United States, it is also essential to show that these models are also capable of forecasting the future values of the temperature. Therefore, we hide the last 12 months of the temperature, restructure the models (3.2) and (3.3) and try to predict the following months only using the previous information. For example, we used the first 1334 observations $\{x_1, x_2, ..., x_{1334}\}$ to forecast $\hat{x}_{1335}$. Then we use the observations $\{x_1, x_2, ..., x_{1335}\}$ to forecast $\hat{x}_{1336}$, and continue this process until we obtain the forecasting values of the last 12 observations, that is, $\{\hat{x}_{1335}, \hat{x}_{1336}, ..., \hat{x}_{1346}\}$. Table 4.3, gives the actual, forecasting and residual data for the subject 12 months.

Table 4.3 (Version 1 Dataset)

	Original Values	Forecast Values	Residuals
March 2006	43.31	44.0291	-0.7191
April 2006	56.03	53.1361	2.89395
May 2006	63.06	62.5318	0.52821
June 2006	71.44	70.6153	0.82467
July 2006	77.1	75.5855	1.51453
August 2006	74.1	74.2054	-0.1054
September 2006	63.69	66.6904	-3.0004
October 2006	52.97	55.4991	-2.5291
November 2006	44.68	43.2673	1.41275
December 2006	36.64	34.6357	2.00433
January 2007	31.39	32.58	-1.19
February 2007	32.86	36.2024	-3.3424

Figure 4.5 below gives a graphical presentation of the information presented in Table 4.3 for Version 1 observed time series.

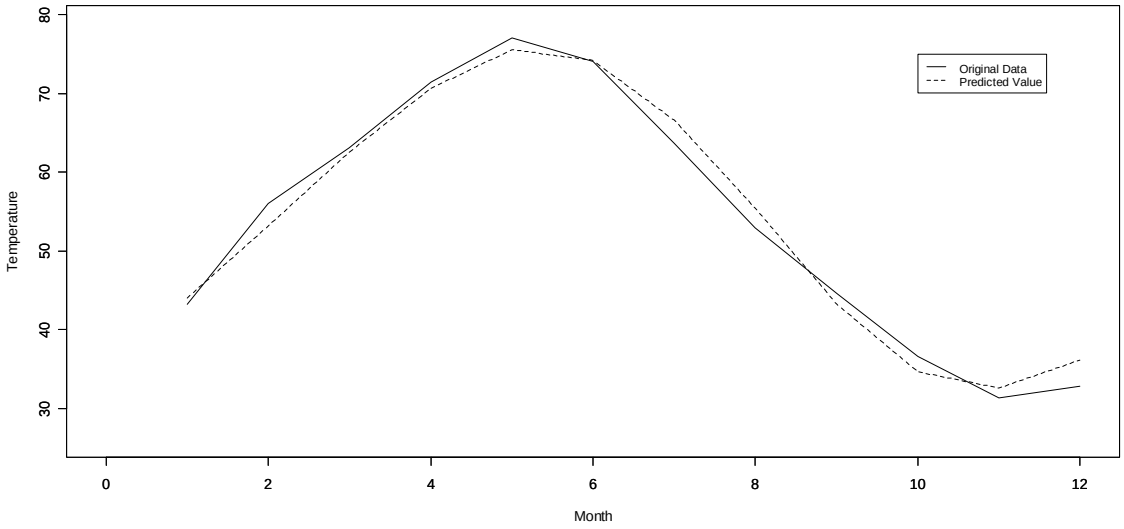

Figure 4.5. Monthly Temperature VS. Our Predicted Values for the Last 12 Observations (Version 1 Dataset)

Similarly, for Version 2 of the data set, we have calculated the estimates presented by Table 4.4.

Table 4.4

	Original Values	Forecast Values	Residuals
March 2006	43.45	44.1812	-0.7312
April 2006	56.12	53.2506	2.86942
May 2006	63.12	62.6351	0.48486
June 2006	71.55	70.7152	0.83478
July 2006	77.22	75.6947	1.52532
August 2006	74.19	74.3167	-0.1267
September 2006	63.86	66.8069	-2.9469
October 2006	53.13	55.6137	-2.4837
November 2006	44.58	43.3947	1.18529
December 2006	36.79	34.7224	2.06761
January 2007	31.46	32.6854	-1.2254
February 2007	32.86	36.3025	-3.4425

A graphical presentation of the results given in Table 4.4 are given below by Figure 4.6.

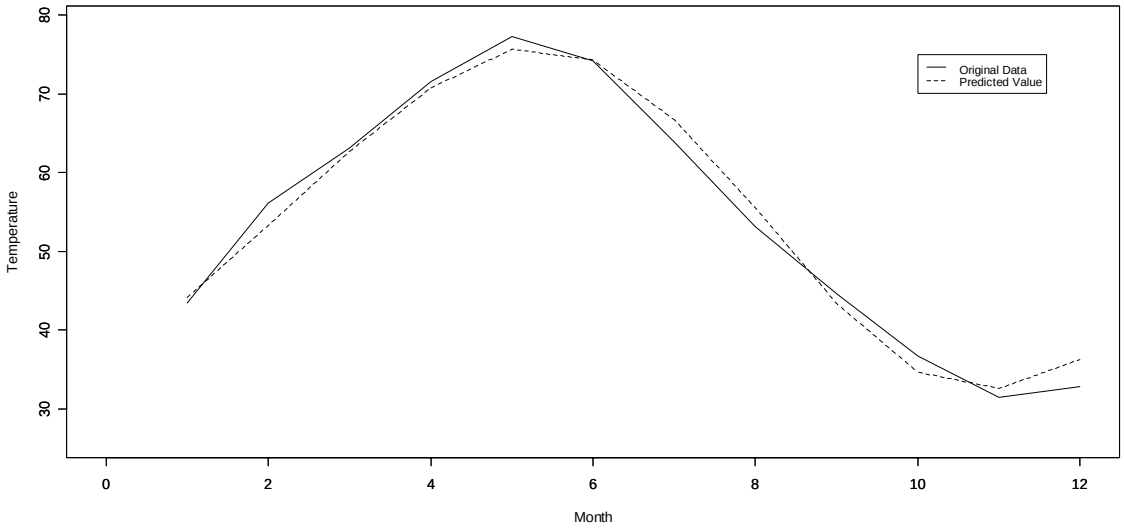

Figure 4.6. Monthly Temperature VS. Our Predicted Values for the Last 12 Observations (Version 2 Dataset)

We remark the similarity of the results of both models and the good forecast values.

5. CONCLUSION

We have developed two seasonal autoregressive integrated moving average models to forecast the monthly average temperature in the Continental united States using historical data for 1895-2007. The two models are based on two different methods, USCD and USHCN, that are been used to create the two temperature basis. The two developed models were evaluated and it was shown that the processes give good forecast values. In addition we can conclude that both Version 1 and 2 give really similar results and thus, both methods are not necessary.

6. ACKNOWLEDGEMENT

The authors wish to thank Dr. G. Ladde for his fruitful discussions and assistance during the present study.

REFERENCES

1. Akaike, H. (1974). A New Look at the Statistical Model Identification, IEEE Transactions on Automatic Control, AC-19, 716-723.

2. Alexandersson, H. and A. Moberg, (1997). Homogenization of Swedish temperature data. Part I: Homogeneity test for linear trends. Int. J. Climatol., 17, 25-34.

3. Baker, D. G., (1975). Effect of observation time on mean temperature estimation. J. Appl. Meteor., 14, 471-476.

4. Easterling, D.R., and T.C. Peterson, (1995). A new method of detecting undocumented discontinuities in climatological time series, Int. J. of Climatol., 15, 369-377.

5. Easterling, D. R., T. R. Karl, E.H. Mason, P. Y. Hughes, and D. P. Bowman. (1996). United States Historical Climatology Network (U.S. HCN) Monthly Temperature and Precipitation Data. ORNL/CDIAC-87, NDP-019/R3. Carbon Dioxide Information Analysis Center, Oak Ridge National Laboratory, U.S. Department of Energy, Oak Ridge, Tennessee.

6. Easterling, D. R., T. R. Karl, J. H. Lawrimore, and S. A. Del Greco. (1999). United States Historical Climatology Network Daily Temperature, Precipitation, and Snow Data for 1871-1997. ORNL/CDIAC-118, NDP-070. Carbon Dioxide Information Analysis Center, Oak Ridge National Laboratory, U.S. Department of Energy, Oak Ridge, Tennessee.

7. Hughes, P. Y., E. H. Mason, T. R. Karl, and W. A. Brower. (1992). United States Historical Climatology Network Daily Temperature and Precipitation Data. ORNL/CDIAC-50, NDP-042. Carbon Dioxide Information Analysis Center, Oak Ridge National Laboratory, U.S. Department of Energy, Oak Ridge, Tennessee.

8. Karl, T. R., C. N. Williams, Jr., and F. T. Quinlan. (1990). United States Historical Climatology Network (HCN) Serial Temperature and Precipitation Data. ORNL/CDIAC-30, NDP-019/R1. Carbon Dioxide Information Analysis Center, Oak Ridge National Laboratory, U.S. Department of Energy, Oak Ridge, Tennessee.

9. Karl, T.R., C.N. Williams, Jr., P.J. Young, and W.M. Wendland, (1986). A model to estimate the time of observation bias associated with monthly mean maximum, minimum, and mean temperature for the United States, J. Climate Appl. Meteor., 25, 145-160.

10. Karl, T.R., and C.W. Williams, Jr., (1987). An approach to adjusting climatological time series for discontinuous inhomogeneities, J. Climate Appl. Meteor., 26, 1744-1763.

11. Karl, T.R., H.F. Diaz, and G. Kukla, (1988). Urbanization: its detection and effect in the United States climate record, J. Climate, 1, 1099-1123.

12. Karl, T.R., C.N. Williams, Jr., F.T. Quinlan, and T.A. Boden, (1990). United States Historical Climatology Network (HCN) Serial Temperature and Precipitation Data, Environmental Science Division, Publication No. 3404, Carbon Dioxide Information and Analysis Center, Oak Ridge National Laboratory, Oak Ridge, TN, 389 pp.

13. Kwiatkowski, D., Phillips, P. C. B., Schmidt, P., & Shin, Y. (1992). Testing the Null Hypothesis of Stationarity against the Alternative of a Unit Root., Journal of Econometrics, 54, 159-178.

14. Lund, R., and J. Reeves, (2002). Detection of undocumented changepoints: a revision of the two-phase regression model. J. Climate, 15, 2547-2554.

15. Menne, M.J., and C.N. Williams, Jr., (2005). Detection of undocumented changepoints using multiple test statistics and composite reference series. J. Climate, 18, 4271-4286.

16. Peterson, T.C., and D.R. Easterling, (1994). Creation of homogeneous composite climatological reference series, Int. J. Climatol., 14, 671-680.

17. Quayle, R.G., D.R. Easterling, T.R. Karl, and P.Y. Hughes, (1991). Effects of recent thermometer changes in the cooperative station network, Bull. Am. Meteorol. Soc., 72, 1718-1724.

18. Quinlan, F. T., T. R. Karl, and C. N. Williams, Jr. (1987). United States Historical Climatology Network (HCN) Serial Temperature and Precipitation Data. NDP-019. Carbon Dioxide Information Analysis Center. Oak Ridge National Laboratory, U.S. Department of Energy, Oak Ridge, Tennessee.

19. Vose, R.S., C.N. Williams, T.C. Peterson, T.R. Karl, and D.R. Easterling, (2003). An evaluation of the time of observation bias adjustment in the US Historical Climatology Network. Geophysical research letters, 30 (20), 2046, clim3-1--3-4 doi:10.1029/2003GL018111.

20. Wang, X.L., (2003). Comments on "Detection of undocumented changepoints: A revision of the two-phase model". J. Climate, 16, 3383-3385.