A New Forecasting Model For A Nonstationary Environmental Data

Shou Hsing Shih and Chris P. Tsokos

Department of Mathematics and Statistics
University of South Florida, USA

Abstract

The object of the present study is to develop a new forecasting model for the atmospheric temperature of the continental United States. We shall analyze the pattern of the temperature time series, and illustrate the usefulness of the duplicated mean of the signal. In removing the duplicated mean time series from the original temperature recording series simplifies the forecasting process. The accuracy of this proposed methodology will be demonstrated in comparison with the classical multiplicative Autoregressive Integrated Moving Average, ARIMA model that is often used.

Introduction

There are two methods being used in recording atmospheric temperatures in the continental United States and we shall refer them as Version 1 and Version 2 data sets. Version 1 data was collected by the United States Climate Division, USCD, and Version 2 data by the United States Historical Climatology Network, USHCN. For additional information concerning Version 1 and Version 2 data sets, see (Alexandersson & Moberg, 1997; Baker, 1975; Easterling & Peterson, 1995; Easterling et al., 1996; Easterling et al., 1999; Hughes et al., 1992; Karl et al., 1986; Karl & Williams, 1987; Karl et al., 1988;

Karl et al., 1990; Karl et al., 1990; Karl et al., 1988; Karl et al., 1986; Karl & Williams, 1987; Lund & Reeves, 2002; Menne & Williams, 2005; Peterson & Easterling, 1994; Quayle et al., 1991; Quinlan et al., 1987; Vose et al., 2003; Wang, 2003). Although we found the two different sets of temperature data to be somewhat similar, we believe from a statistical perspective that the Version 2 data set is more appropriate to use. Therefore, we will use the Version 2 data to represent the temperature series of continental United States in this study.

In the present study, our object is to forecast the monthly average atmospheric temperature in degrees Fahrenheit in the Continental United States using historical monthly data from 1895-2007. A graphical presentation of the monthly average temperature of the continental United States is given by Figure 1.

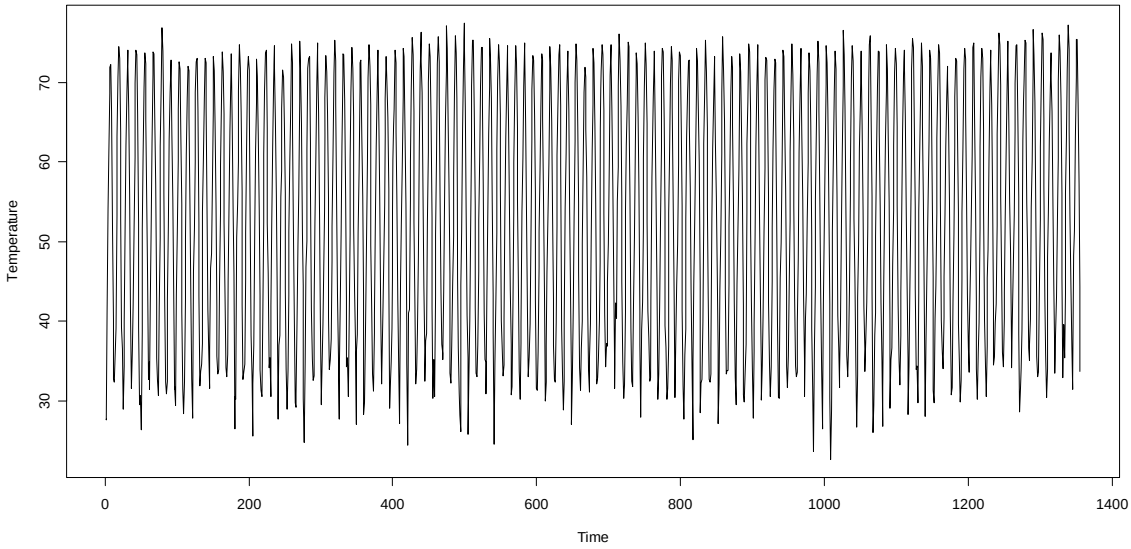

Figure 1 Time Series Plot for Monthly Atmospheric Temperature, 1895-2007

As can be visually seen the temperature series contains a seasonal pattern, and it repeats itself every 12 months. We shall discuss it further later in our study.

Seasonal Multiplicative ARIMA Forecasting Model

In time series analysis, seasonal variations usually dominate the variations of the original nonstationary time series and make it very difficult to analyze. It occurs often on environmental data along with some type of periodic trend that we must address in developing a forecasting model. In our study we will treat the seasonal time series as a nonstationary time series that varies along some sort of seasonal periodic trend. Hence, addressing the seasonal variations for the forecasting model becomes very useful when we deal with these types of difficulties.

(Box & Jenkins 1994) first introduced the seasonal multiplicative autoregressive integrated moving average, ARIMA, model that is capable of developing a forecasting model of a given time series with seasonal variation. This forecasting process addresses the issue of the incapability of predicting a time series with seasonal trends for the classical ARIMA methodology. The seasonal multiplicative ARIMA model is defined by

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

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 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:

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

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

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

and

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

The order of the multiplicative ARIMA model determines the structure of the model and it is essential to have a good procedural approach in terms of developing the forecasting model.

(Shih & Tsokos 2008) summarized the model identifying procedure as follows:

- 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 $p + q + P + Q \leq m$.
- 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).

The forecasting model that we identified using the above procedure is

ARIMA$(2,1,1)\times(1,1,1)_{12}$ process, analytical given by

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

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 obtain

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

Thus, the one-step ahead forecasting model for the atmospheric temperature data is calculated to be

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

where $\phi_1 = .0899$, $\phi_2 = .04$, $\theta_1 = .9853$, $\Phi_1 = -.0058$, and $\Gamma_1 = .9761$.

We shall compare the result of the classical multiplicative ARIMA with our proposed methodology in later section.

The Proposed Forecasting Model

It is clear that the average monthly atmospheric temperature of the Continental United States contains a seasonal pattern without any upward or downward trend present, see Figure 1. The idea of our proposed forecasting model is assuming that the overall

mean of all January, February, …, December data equals to 12 constants, hence, we can treat the atmospheric temperature series as a nonstationary time series that varies along those 12 constants that we can estimate.

Let $x_1, x_2, ..., x_{1356}$ denotes the average monthly atmospheric temperature from year 1895 to 2007, $m_1, m_2, ..., m_{12}$ denotes the mean of Januarys, Februarys, … , Decembers, and the year 1901 to 2000 be the base period. It is obvious that $x_{73}, x_{74}, ..., x_{1272}$ of the atmospheric temperature series represents the monthly temperature from 1901 to 2000, and then we can calculate $m_1, m_2, ..., m_{12}$ by using the following transformation.

$$m_1 = \frac{x_{73} + x_{85} + ... + x_{1261}}{100}$$

$$m_2 = \frac{x_{74} + x_{86} + ... + x_{1262}}{100}$$

$$\vdots$$

$$m_{12} = \frac{x_{84} + x_{96} + ... + x_{1272}}{100} \tag{4}$$

We proceed to create a new time series $\{\gamma_t\}$ by simply repeating the series $m_1, m_2, ..., m_{12}$, that is,

$$\{\gamma_1, \gamma_2, ..., \gamma_{12}\} = \{m_1, m_2, ..., m_{12}\}$$

$$\{\gamma_{13}, \gamma_{14}, ..., \gamma_{24}\} = \{m_1, m_2, ..., m_{12}\}$$

$$\vdots$$

$$\{\gamma_{1345}, \gamma_{1346}, ..., \gamma_{1356}\} = \{m_1, m_2, ..., m_{12}\} \tag{5}$$

The time series plot of the new time series $\{\gamma_t\}$ is illustrated below by Figure 2.

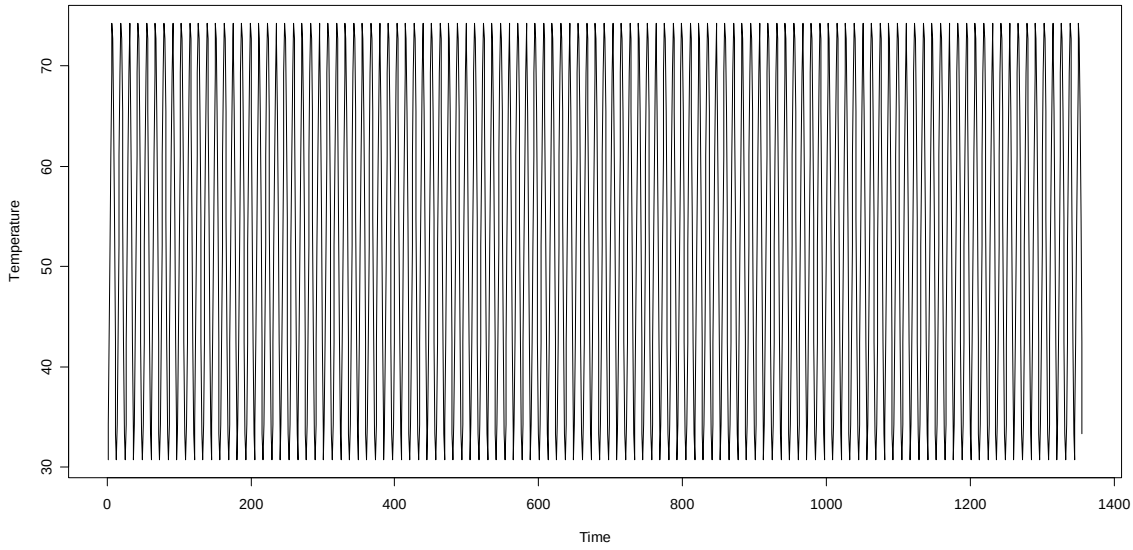

Figure 2 Time Series Plot of the Series $\{\gamma_t\}$

We have assumed that the original atmospheric temperature series varies along the series $\{\gamma_t\}$. Let $\{\lambda_t\}$ be the difference between the atmospheric temperature series $\{x_t\}$ and the new series $\{\gamma_t\}$, that is,

$$\lambda_t = x_t - \gamma_t \tag{6}$$

and the resulting nonstationary time series $\{\lambda_t\}$ will be used to develop the forecasting process. The new nonstationary time series is shown below by Figure 3.

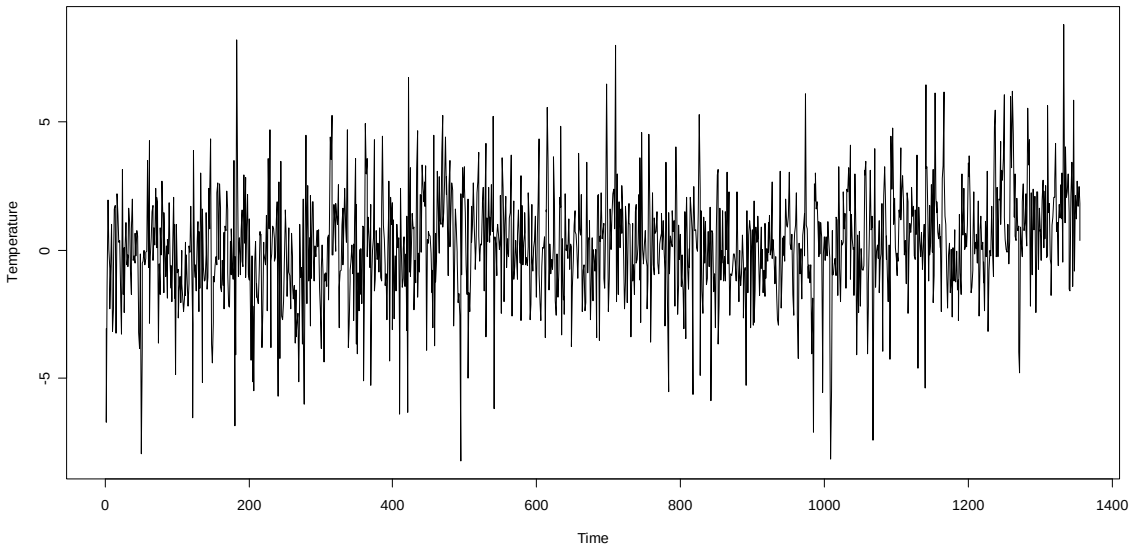

Figure 3 Time Series Plot of the Series $\{\lambda_t\}$

Since the above process is invertible, we proceed to turn a seasonal nonstationary time series into a simple nonstationary time series without losing any information. Once we obtain the forecasts from the series $\{\lambda_t\}$, we can obtain the forecasts of the actual time series of the atmospheric temperature series $\{x_t\}$. By using the methodology that we discussed in (Shih & Tsokos 2008), we have found that the best ARIMA model on the series $\{\lambda_t\}$ is ARIMA(2,1,1), that is,

$$(1 - \phi_1 B - \phi_2 B^2)(1 - B)\lambda_t = (1 - \theta_1 B)\varepsilon_t . \qquad (7)$$

Expanding the autoregressive operator and the first difference filter, we have

$$[1-(1+\phi_1)B+(\phi_1-\phi_2)B^2+\phi_2 B^3]\lambda_t = (1-\theta_1 B)\varepsilon_t$$

and the model can be written as

$$\lambda_t - (1+\phi_1)\lambda_{t-1} + (\phi_1-\phi_2)\lambda_{t-2} + \phi_2\lambda_{t-3} = \varepsilon_t - \theta_1\varepsilon_{t-1}.$$

The final analytical form of the above forecasting model can be written as

$$\hat{\lambda}_t = (1+\phi_1)\lambda_{t-1} - (\phi_1 - \phi_2)\lambda_{t-2} - \phi_2\lambda_{t-3} - \theta_1\varepsilon_{t-1} + \varepsilon_t . \quad (8)$$

where $\phi_1 = .0903$, $\phi_2 = .0414$, and $\theta_1 = .9837$.

We can transform back to the original atmospheric temperature series by combining (5) and (6). Thus we have

$$\hat{x}_t = 1.0903\lambda_{t-1} - .0489\lambda_{t-2} - 0.414\lambda_{t-3} + .09837\varepsilon_{t-1} + \gamma_t + \varepsilon_t \quad (9)$$

This forecasting model as we will show below gives a more accurate forecast then the classical model of Box & Jenkins.

Comparison of the Classical and Proposed Models

The basic statistics that shall we use to compare the classical and proposed models are the mean $\bar{r}$, variance S_r^2, standard deviation S_r and standard error $S_r/\sqrt{n}$ of the residuals. Table 1 and 2 are the basic results of the classical multiplicative ARIMA model and our proposed forecasting model.

Table1 Basic Evaluation Statistics for Classical Multiplicative ARIMA Model

$\bar{r}$	S_r^2	S_r	$S_r/\sqrt{n}$
-0.008972434	4.314059	2.077031	0.05640443

The mean residual, $\bar{r}$ = -0.008972434, of the classical multiplicative ARIMA model consistently overestimates the atmospheric temperature forecast by a factor of 0.008972434. Thus, we subtract 0.008972434 from (3) to correct for the overestimating the actual forecast.

Table2 Basic Evaluation Statistics for Our Proposed Model

$\bar{r}$	S_r^2	S_r	$S_r/\sqrt{n}$
0.09095724	4.253849	2.062486	0.05600944

In our proposed model, the mean residual, $\bar{r} = 0.09095724$, which consistently underestimates the atmospheric temperature forecast by a 0.09095724. Thus, we shall add 0.09095724 to (9) to correct the underestimating of the forecasting values.

From Table 1 and 2, it is clear that our proposed forecasting model given better forecast than the classical multiplicative ARIMA model with respect to variance, standard deviation and standard error. It speaks out the stability of our proposed methodology. Also the proposed forecasting model reduces the computational complexity that the classical forecasting model requires.

Efficiency on Forecasting Models

To evaluate the efficiency of both forecasting models, we proceed to hide the atmospheric temperature observations of the last 12 months, and try to predict them only use the previous information. That is, we use $x_1, x_2, ..., x_{1344}$ to structure the model and we will predict $\hat{x}_{1345}$, $x_1, x_2, ..., x_{1345}$ to predict $\hat{x}_{1346}$, … , $x_1, x_2, ..., x_{1355}$ to predict $\hat{x}_{1356}$. Table 3 below provides a month by month comparison in predicting the average monthly atmospheric temperature of the year 2007 between the classical multiplicative ARIMA model and our proposed forecasting model. The “Actual” column represents the actual average monthly atmospheric temperature of the year 2007. The “Classical Forecasts” column represents the predictions generated by the classical multiplicative ARIMA

model. The “Proposed Forecasts” column represents the predictions generated by our proposed model. The residuals are calculated by using expression (10).

$$r_i = x_i - \hat{x}_i \tag{10}$$

Table 3 Month to Month Comparison

Month(2007)	Actual	Classical Forecasts	Classical Residuals	Proposed Forecasts	Proposed Residuals
January	31.45	32.68155	-1.23155	32.18227	-0.73227
February	32.87	36.29613	-3.42613	35.73977	-2.86977
March	48.23	43.54954	4.68046	43.84947	4.38053
April	51.3	53.73843	-2.43843	53.57107	-2.27107
May	63.2	62.37962	0.82038	62.51967	0.68033
June	70.66	70.53462	0.12538	70.74577	-0.08577
July	75.46	75.62914	-0.16914	75.70057	-0.24057
August	75.4	74.06826	1.33174	74.18117	1.21883
September	67.13	66.76519	0.36481	66.86947	0.26053
October	57.21	55.95558	1.25442	56.18677	1.02323
November	44.28	44.02936	0.25064	43.91927	0.36073
December	33.73	34.96153	-1.23153	34.81297	-1.08297

From Table 3, we can see that the residuals generated by our proposed forecasting model are better than the classical multiplicative ARIMA model almost for all months.

According to the evaluation demonstrated in Table 2, we notice that the proposed model consistently underestimate the original temperature series by the residual mean equal to 0.09095724. We can improve our forecasts further if we add the residual mean back to the model. Therefore, the final analytical form of the proposed model becomes

$$\hat{x}_t = 0.09095724 + 1.0903\lambda_{t-1} - .0489\lambda_{t-2} - 0.414\lambda_{t-3} + .09837\varepsilon_{t-1} + \gamma_t + \varepsilon_t . \tag{11}$$

Conclusion

The classical multiplicative ARIMA forecasting process is a useful tool in predicting seasonal time series. However, its complexity increases tremendously once the seasonal order (P, D, Q) increases. Expression (3) shows how complicated it becomes to actually obtain a workable final form of the forecasting model. We propose a new methodology that recognizes the presence of a periodic seasonal effect in the time series is actually a nonstationary time series varies along some periodic constants. Our results show that the proposed forecasting model is not only more accurate but also more effective than the classical multiplicative ARIMA forecasting model. In addition, the proposed model is far simpler in terms of its computational complexity than the classical model. Thus, we recommend the use of the proposed model over the classical multiplicative ARIMA process.

References

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

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

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

Easterling, D. R., T. R. Karl, E.H. Mason, P. Y. Hughes, & 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.

Easterling, D. R., T. R. Karl, J. H. Lawrimore, & 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.

Hughes, P. Y., E. H. Mason, T. R. Karl, & 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.

Karl, T. R., C. N. Williams, Jr., & 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.

Karl, T.R., C.N. Williams, Jr., P.J. Young, & W.M. Wendl&, (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.

Karl, T.R., & C.W. Williams, Jr., (1987). An approach to adjusting climatological time series for discontinuous inhomogeneities, J. Climate Appl. Meteor., 26, 1744-1763.
Karl, T.R., H.F. Diaz, & G. Kukla, (1988). Urbanization: its detection and effect in the United States climate record, J. Climate, 1, 1099-1123.

Karl, T.R., C.N. Williams, Jr., F.T. Quinlan, & 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.

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

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

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

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

Quinlan, F. T., T. R. Karl, & 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.

Shih, S. H., and C. P. Tsokoos (2008), A Temperature Forecasting Model for the Continental United States, The International Journal Neural, Parallel & Scientific Computations, Volume 16, Number 1, pp. 59-72.

Shih, S. H., and C. P. Tsokoos (2008), New Nonstationary Time Series Models with Economic Applications, Proceedings of Dynamic Systems and Applications, Volume 5, 2008, pp. 453-460

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