Proceedings of Joint Conference of Korean Data And Information Science Society and The Korean Data Analysis Society,
April 28 - 29, 2006, pp. 335-343

Forecasting of Monthly Mean of Surface Air Temperature on Korean Region

Jeong Hyeong Lee[1], *Keon Tae Sohn*[2], *and Riyu Lu*[3]

1. Introduction

The problem of detecting and forecasting a climate change signal in the climatological record is of obvious importance in any strategy to understand regional and global change. Over the last few decades there has been growing interest in trying to quantify the magnitude of regional and global climatic trends. Trends are defined here by the mean slope of the change in amplitude of a given climatic variable with time over some prescribed period.

Since understanding climatic changes at the regional scale is one of the most important and uncertain issues within the climate change debate, the need to provide regional climate change information has increased for both impact assessment studies and policymaking(Mearns et al. 2001, Lee et al. 2005). A regional climate is determined by interactions between large, regional and local scales, but the available tools have directed research toward understanding the climate system as a whole.

On the other hand, the statistical analysis of observed temperature time series has not yet provided conclusive evidence about the climatic warming effect. In fact, much has been accomplished in the detection and attribution of climatic change on a global scale. All projections of future change indicate that the warming is likely to continue. This conclusion holds regardless of the computer model used or the 'emission scenario'-the particular set of data describing the future emissions of greenhouse gases and aerosols-applied in the model. The advantage is that no arbitrary assumptions about the variable's statistical properties are required, the catch being that we are assuming that the dynamic models are correctly reproducing these properties. However, the statistical analysis of observed temperature time series using statistical models is also important, at least as a complement to the physical models(Zwiers, 2002).

[1] Division of Management Information Science, Dong-A University, Busan 604-714, Korea.
Institute of Atmospheric Physics, Chinese Academy of Sciences, Beijing 100080, China.
E-mail : jeonglee@dau.ac.kr

[2] Department of Statistics, Pusan National University, Busan 609-735, Korea.
E-mail : ktsohn@pusan.ac.kr

[3] Institute of Atmospheric Physics, Chinese Academy of Sciences, Beijing 100080, China.

A summary of results for the previous global and regional climate change studies is presented in Table 2 of Casey and Cornillon(2001). For the globe, the 1960~1990 anomaly trends calculated using the Comprehensive Ocean-Atmosphere Data Set(COADS; Woodruff et al. 1987) are found to be 0.09±0.03°C and 0.10±0.03°C per decade for the 5° bin and temperature class averaging techniques, respectively.

Trends for other climatic variables have also been estimated, but normally for shorter periods than for surface air temperature. Wu and Straus(2004) have computed 55-year duration trends, for the period 1948~2002, for surface pressure, temperature at different heights etc.

Trends have also been estimated for individual countries for a range of climatic variables, usually over only multi-decadal periods. See for example, Jonsson and Fortuniak(1995), surface wind directions for Sweden; Brunetti et al.(2000), temperature trends for Italy; Kaiser(2000), cloudiness and other trends for China; and Osborn et al.(2000), precipitation trends for the UK. Lee et al.(2005) studied climate change detection, attribution, and prediction for the surface air temperature in the Northeast Asian region.

Our object is to predict surface air temperature(SAT) of Korean region using variance decomposition method in time series analysis.

2. Data and regions

The primary dataset used in this study is the National Centers for Environmental Prediction(NCEP) and the National Center for Atmospheric Research(NCAR) Collaborative Reanalyses(hereafter, NNR) obtained from the National Oceanic and Atmospheric Administration's Climate Diagnostics Center(NOAA) website http://www.cdc.noaa.gov/.

We used the monthly mean of SAT of the NNR data from January 1948 to December 2005. The resolution of data is 2.5° longitude by 2.5° latitude. The dataset analyzed in this paper is Korean region monthly SAT anomalies from January 1948 to February 2006. We defined that Korean region is (122.5°E~132.5°E, 32.5°N~42.5°N) (Fig. 1).

The plots of the data set is given in Figure 2. We have subtracted sample mean from the original data and thus the data plotted correspond to this mean deleted observation.

3. Method and results

The fundamental statistical model used in this study is

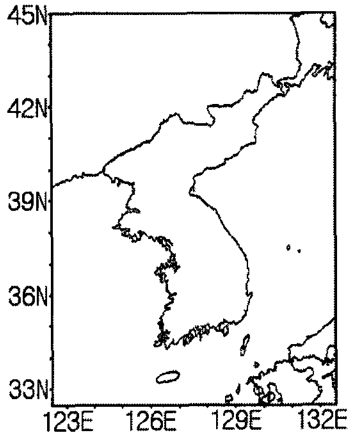

Fig. 1. Regional map showing the region used to compute regional anomaly SAT.

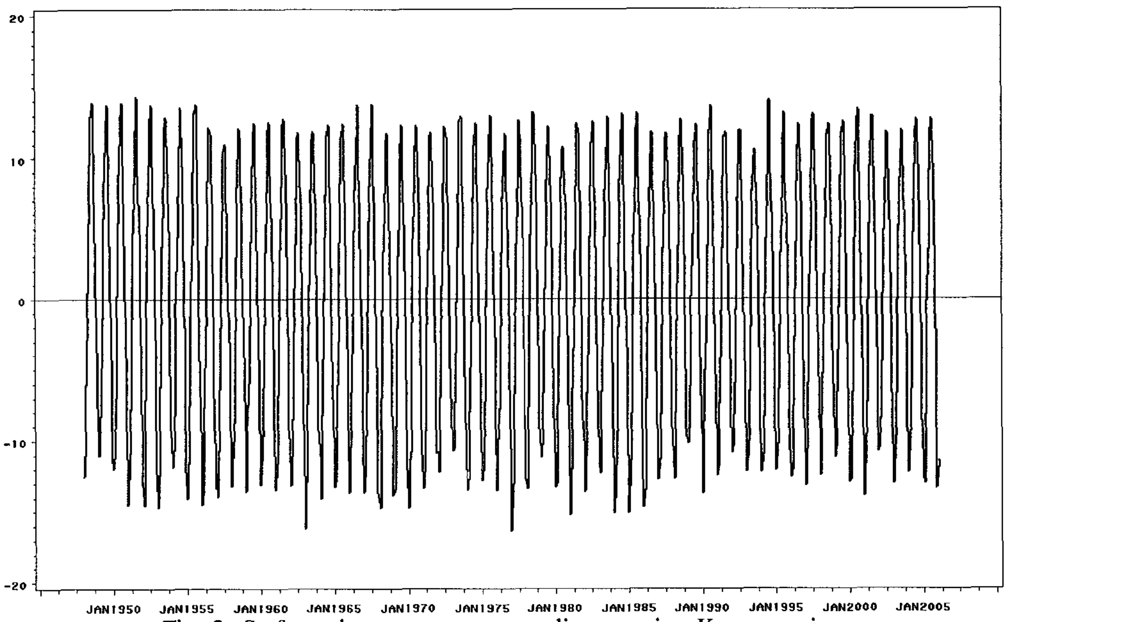

Fig. 2. Surface air temperature anomalies covering Korean region.

$$Y_t = T_t + S_t + I_t, \quad t = 1, \cdots, N,$$

where T_t is the long-time trend, S_t is the seasonal component and I_t is the random noise. For example, in the study of global warming, the SAT data consists of the long-term warming trend, seasonal variations within years and random variation.

We consider additive model given below for the SAT data where $\{Y_t\}$ are the observed time series.

$$Y_t = c_0 + c_1 t + \sum_{i=1}^{12} r_i I_{it} + f(t, \theta) X_t, \qquad t = 1, \cdots, N, \tag{1}$$

where c_0, c_1, and $r_i(i=1,\cdots,12)$are parameters to be estimated, 12 is the number of months a year,

$$I_{it}=\begin{cases}1, & t(\text{mod } 12)=i \\ 0, & \text{otherwise}\end{cases}$$

is index function of month as we are dealing with monthly SAT data and $f(t,\theta)>0$ $(1\le t\le N)$ is bounded parametric deterministic function of t. Also, we assume that the stationary time series $\{X_t\}$ can be represented by an $ARMA(p,q)$ model of the form

$$X_t+a_1X_{t-1}+\cdots+a_pX_{t-p}=e_t+b_1e_{t-1}+\cdots+b_qe_{t-q}, \tag{2}$$

where $\{e_t,\ t=1,\cdots,N\}$ is a sequence of independent and identically distributed(i.i.d) random variables with mean zero and variance σ_e^2. We assume that the roots of the polynomial $a(B)=1+a_1B+\cdots+a_pB^p$ and $b(B)=1+b_1B+\cdots+b_qB^q$ lie outside the unit circle. We now consider the estimation of the parameters of nonstationary model (1), the parameters of the deterministic function $f(t,\theta)$ as well as orders and the parameters of the stationary $ARMA(p,q)$ model (2). $ARMA(p,q)$ processes are stationary, shortrange correlated normal processes. White noise residuals($ARMA(0,0)$) and red noise residuals($ARMA(1,0)$) are the most commonly used.

We first consider the estimation of the parameters c_0 and c_1 by the method of ordinary least square(OLS). Let

$$\zeta_t=\sum_{i=1}^{12}r_iI_{it}+f(t,\theta)X_t,\qquad t=1,\cdots,N.$$

Then

$$Y_t=c_0+c_1t+\zeta_t,\qquad t=1,\cdots,N.$$

Let

$$c=\begin{bmatrix}c_0\\ c_1\end{bmatrix},\ Y=\begin{bmatrix}Y_1\\ \vdots\\ Y_N\end{bmatrix},\ \zeta=\begin{bmatrix}\zeta_1\\ \vdots\\ \zeta_N\end{bmatrix},\ S_1=\begin{bmatrix}1 & 1\\ \vdots & \vdots\\ 1 & N\end{bmatrix},$$

where ' denotes the transpose of a matrix.

We estimate the parameter c_0 and c_1 by the method of OLS. Toyooka(1977, 1980) showed that under some conditions on the function $f(t,\theta)$ and on the regression variables, OLS can give consistent estimates of the parameters even in the case where the errors are nonstationary and uniformly modulated. We have the OLS estimate $\hat{c}$ of c as $\hat{c}=(S_1'S_1)^{-1}S_1'Y$.

Having estimated the parameters c_0 and c_1 we can obtain the residuals' $\{Z_t\}$ from

$$Z_t=Y_t-\widehat{c_0}+\widehat{c_1}t, \qquad t=1,\cdots,N.$$

We now estimate r_i parameter 12 using the residual $\{Z_t\}$ is given by

$$Z_t=\sum_{i=1}^{12}r_iI_{it}+\zeta_t, \qquad t=1,\cdots,N, \tag{3}$$

where

$$\zeta_t\approx f(t,\theta)X_t, \qquad 1,\cdots,N.$$

The estimation of r_i is quite simple. Let

$$Z=\begin{bmatrix}Z_1\\ \vdots\\ Z_N\end{bmatrix},\ R=\begin{bmatrix}r_1\\ \vdots\\ r_{12}\end{bmatrix},\ \xi=\begin{bmatrix}\xi_1\\ \vdots\\ \xi_N\end{bmatrix},\ D_{N\times 12}=\begin{bmatrix}1&0&0&0&\cdots&0\\ 0&1&0&0&\cdots&0\\ \vdots&\vdots&\vdots&\vdots&\vdots&\vdots\\ 0&1&0&0&\cdots&0\end{bmatrix}.$$

Estimates of $\{r_i\}$ can then be obtained by OLS. The least squares estimates $\{\widehat{r_i}\}$ of the unknown parameters $\{r_i\}$ are given by $\hat{R}=(D'D)^{-1}D'Z$.

Given the initial estimates of $\{r_i\}$ we can proceed to estimate the parameters of the function $f(t,\theta)$. We assume $f(t,\theta)>0$, for all t which is satisfied by the chosen functions. By replacing the parameters of the $f(t,\theta)$ with their estimates $\hat{\theta}$ we obtain the model

$$\frac{Z_t}{f(t,\hat{\theta})}=\frac{1}{f(t,\hat{\theta})}\sum_{i=1}^{12}r_iI_{it}+X_t, \qquad t=1,\cdots,N.$$

The form of $\{X_t\}$ is unknown. Therefore we proceed assuming that $\{X_t\}$ are i.i.d random variables and re-estimate $\{X_t\}$ following the above procedure; the only difference being that the new estimates of $\{r_i\}$ will be given by $\hat{R}=(D_{(2)}'D_{(2)})^{-1}D_{(2)}'Z_{(2)}$ where $Z_{(2)}$ and $D_{(2)}$ are obtained from Z and D where each element is divided by $f(\cdot,\theta)$.

Our objective here is to find a suitable functional form of $f(t,\theta)$ to describe our SAT data. On the basis of the simulation study we concluded that the choice of $f(t,\theta)=t^{c_2}$ $(t>0)$ with $c_2<0$. From the model (1) we have

$$Y_t - c_0 - c_1 t - \sum_{i=1}^{12} r_i I_{it} = f(t, \theta) X_t, \qquad t = 1, \cdots, N.$$

Let

$$W_t = Y_t - \widehat{c_0} - \widehat{c_1} t - \sum_{i=1}^{12} \widehat{r_i} I_{it}, \qquad t = 1, \cdots, N,$$

where $\widehat{r_i}$ is the estimates of r_i. Hence

$$\ln|W_t| = c_2 \ln t + \mu + \eta_t, \qquad t = 1, \cdots, N$$

where $\mu = E(\ln|X_t|)$ is a constant to be estimated and $\eta_t = \ln|X_t| - \mu$ is a random variable with mean zero and variance σ_η^2. Let

$$\theta = \begin{bmatrix} \mu \\ c_2 \end{bmatrix}, \quad W = \begin{bmatrix} \ln|W_1| \\ \vdots \\ \ln|W_N| \end{bmatrix}, \quad E = \begin{bmatrix} \eta_1 \\ \vdots \\ \eta_N \end{bmatrix}, \quad S = \begin{bmatrix} 1 & \ln 1 \\ \vdots & \vdots \\ 1 & \ln N \end{bmatrix}.$$

OLS estimates of μ and c_2 are given by $\widehat{\theta} = (S'S)^{-1} S'W$. Since $c_2 < 0$, the function of $f(t, \theta) = t^{c_2}$ will tend to 0 as $t \to \infty$. This means that as $t \to \infty$ the model for Y_t will reduce to a deterministic function.

The final stage of the estimation of the parameters of model (1) is to fit an *ARMA* model to $\{W_t / f(t, \widehat{\theta})\}$ where $f(t, \widehat{\theta}) = t^{\widehat{c_2}}$. The function of $\{W_t / f(t, \widehat{\theta})\}$ is approximately an estimate of $\{X_t\}$ and thus we can assume that the resulting series is stationary and satisfies an *ARMA* model. We used Akaike Information Criteria(AIC, Akaike 1977) to determine the orders estimate the parameters of *ARMA* model and then estimate the parameters.

Following the method described above we fitted models of the equation (1) to the dataset of SAT data for

$$f(t, \theta) = t^{c_2}, \quad c_2 < 0.$$

We summarize the obtained models for dataset of the Korean region.

$$\begin{aligned} Y_t = & -0.216342 + 0.000619 \cdot t - 12.9975 \cdot I_{1t} - 11.5972 \cdot I_{2t} - 7.3665 \cdot I_{3t} \\ & - 1.1975 \cdot I_{4t} + 4.0129 \cdot I_{5t} + 8.0316 \cdot I_{6t} + 11.7127 \cdot I_{7t} + 12.5099 \cdot I_{8t} \\ & + 8.6322 \cdot I_{9t} + 2.8431 \cdot I_{10t} - 4.0173 \cdot I_{11t} - 10.1412 \cdot I_{12t} + t^{-0.096528} X_t \end{aligned}$$

for $t = 1, \cdots, 698$ where $\{X_t\}$ is given by

$$(1-0.6622B)X_t=(1-0.3438B)e_t, \quad t=1,\cdots,698$$

and $\{e_t,\ t=1,\cdots,N\}$ are i.i.d with mean zero and variance σ_e^2. An estimate of variance of is 1.7090.

As estimate of the constant c_1 are small values, we calculate the standard error. Thus we have $SE(c_1)=1.6754\times10^3$.

Statistically significant trends did not have been detected in dataset of regional monthly anomalies SAT series. Also from the slop of the linear trend fitted to data set we notice that there is an increase in the SAT which is calculated from

$$\text{Increase} = \text{Slope} \times \text{number of interval}$$

where the number of interval is number of months. The SAT trend per decadal estimated for Jan1948 to Feb2006 was not statistically significant at 0.4321°C.

The trends in the Korean region SAT show interesting features that appear to have relevance to global change issues. These result agree with Zwiers(2002) and Stott and Kettleborough(2002). Zwiers(2002) estimate that the global mean SAT has risen by 0.6±0.2°C during the past century. Stott and Kettleborough(2002) estimate that the global mean SAT in the decade 2020~2030 will be 0.3~1.3°C greater than in 1990~2000(5~95% likelihood range). This result is unaffected by the choice of emission scenario used to make the projection.

Our final objective is to forecast the future values of the SAT data. A good forecast must be close to the original values. Thus, given a series $\{X_t\}$ which is zero mean, second stationary, a sensible criterion would be to minimize the mean square error given by

$$M(m)=E(X_{N+m}-\widetilde{X}_{N(m)})^2$$

where m is the step in the future at which we want to predict our series and $\widetilde{X}_{N(m)}$ is the prediction of X_{N+m}. Wei(1990) showed that the optimum predictor is the conditional expectation of the future values given the past observations, i.e.

$$\widehat{X}_{N(m)}=E(X_{N+m}|X_l,\ l<N).$$

The forecast error $e_{N(m)}$ is given by $e_{N(m)}=(X_{N+m}-\widehat{X}_{N(m)})$.

To forecast values from a fitted model, one has to assume that its parameters are known. By replacing the parameters c_0, c_1 and r_i for $i=1,\cdots,12$ as well as the parameter of the

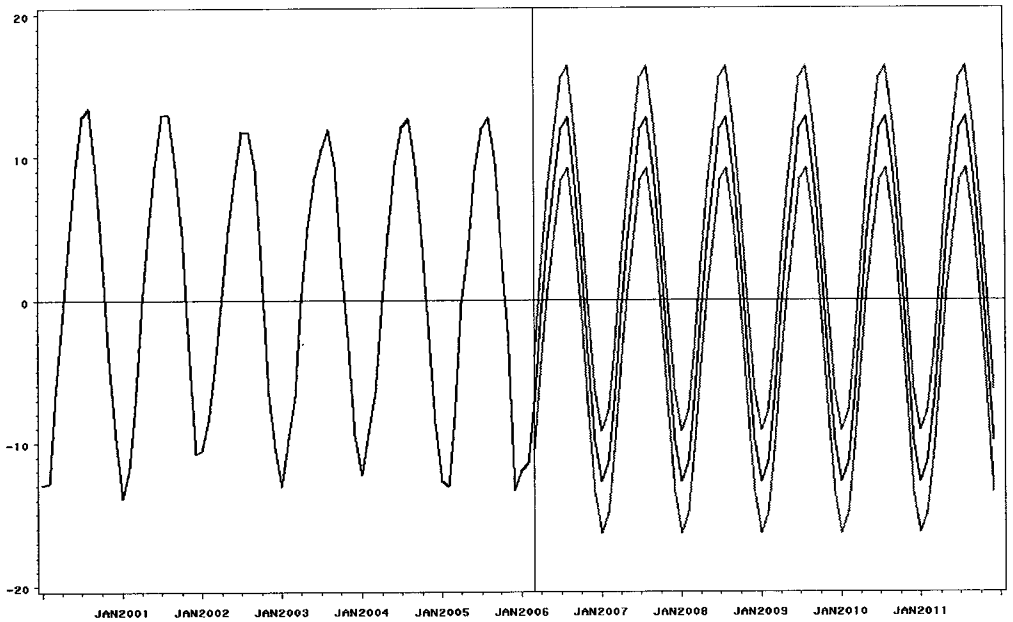

Fig. 3. Prediction for Korean region surface air temperatures

function $f(t, \theta)$ in model (1) by its estimates we can write

$$\widehat{X_t} = (Y_t - \widehat{c_0} - \widehat{c_1}t - \sum_{t=1}^{12} \widehat{r_i} I_{it}) / f(t, \widehat{\theta}), \qquad t = 1, \cdots, N,$$

where $\{X_t\}$ is an $ARMA(p, q)$ stationary process of the form

$$X_t + \widehat{a_1}X_{t-1} + \cdots + \widehat{a_p}X_{t-p} = e_t + \widehat{b_1}e_{t-1} + \cdots + \widehat{b_q}e_{t-q}$$

and $\{e_t,\ t = 1, \cdots, N\}$ is a Gaussian white noise process with mean zero and variance σ_e^2. The order of $ARMA$ model for each set of data has selected an $ARMA(1,1)$ model for $\{X_t\}$. As shown above the optimum predictor is given by (4). Thus we have that for $ARMA(1,1)$ the one-step ahead forecast is given by

$$X_N(1) = \widehat{a_1}X_N + \widehat{b_1}(X_N - \widehat{X}_{N-1}(1)).$$

We obtain monthly forecasts for the SAT dataset. In Fig. 3, we plot the real data of Jan2000~Feb2006 and the predicted values of Mar2006~Dec2011 for dataset.

4. Concluding remarks

In this paper we analyzed SAT data covering the Korean region SAT anomalies and propose models under realistic assumptions.

The conclusions of the analysis of this data lead us to propose uniformly modulated linear models for the SAT data. We proposed an additive model consisting of linear trend, seasonal terms and finally a uniformly modulated process. The parameters were estimated for the SAT dataset and used later for prediction for Mar2006~Dec2011.

References

[1] Akaike, H. (1977). On entropy maximization principle, *Application of Statistics*, P. R. Krishnaiah, Ed., North Holland, 27-41.

[2] Kistler, R., Collins, W., Saha, S., White, G., Woollen, j., Kalnay, E., Chelliah, M., Ebisuzaki, W., Kanamitsu, M., Kousky, V., van den Dool, H., Jenne, R., and Fiorino, M. (2001). The NCEP-NCAR 50-Year Reanalysis: Monthly Means CD-ROM and Documentation, *Bulletin of the American Meteorological Society,* 82, 247-267.

[3] Lee, J. H., Kim, B., Sohn, K. T., Kown W. T., and Mim, S. K. (2005). Climate Change Signal Analysis for Northeast Asian Surface Temperature, *Advances in Atmospheric Sciences,* 22, 159-171.

[4] Mearns, L. O., M. Hulme, T. R. Carter, R. Leemans, M. Lal, and P. Whetton (2001). Climate change 2001, the scientific basis, *Contribution of Working Group I to the Third Assessment Report of the Intergovernmental Panel on Climate Change IPCC,* Cambridge University Press, Cambridge, UK, pp739-768.

[5] Rao, T. S. and Tsolaki, E. P. (2004). *Nonstationary time series analysis of monthly global temperature anomalies, Time Series Analysis and Applications to Geophysical Systems,* Brilliner et al.(Eds), Springer-Verlag, New York.

[6] Stott P. A. and Kettleborough. J. A. (2002). Origins and estimates of uncertainty in predictions of twenty-first century temperature rise, *Nature,* 416, 719-723.

[7] Toyooka, Y. (1977). A note on weak consistency of simple least squares estimators in polynomial regression with a nonstationary error process, *Keio Engineering Report,* 30, 35-44.

[8] Toyooka, Y. (1980). An asymptotically efficient estimation procedure in time series regression model with uniformly modulated error process, *Mathematica Japonica,* 25, 525-532.

[9] Wei , W. (1990). *Time series analysis: Univariate and multivariate methods,* Addison Wesley.

[10] Woodruff, S., R. Slutz, R. Jenne, and P. Steurer (1987). A comprehensive ocean-atmosphere data set, *Bulletin of American Meteorological Society,* 68, 1239-1250.

[11] Zwiers, F. W. (2002). The 20-year forecast, *Nature,* 416, 690-691.