The Korean Journal of Applied Statistics (2009)
22(6), 1143–1152

Climate Prediction by a Hybrid Method with Emphasizing Future Precipitation Change of East Asia

Yaeji Lim[1] · Seongil Jo[2] · Jaeyong Lee[3] · Hee-Seok Oh[4] · Hyun-Suk Kang[5]

[1]Department of Statistics, Seoul National University
[2]Department of Statistics, Seoul National University
[3]Department of Statistics, Seoul National University
[4]Department of Statistics, Seoul National University
[5]National Institute of Meteorological Research Korea Meteorological Administration

(Received July 2009; accepted September 2009)

Abstract

A canonical correlation analysis(CCA)-based method is proposed for prediction of future climate change which combines information from ensembles of atmosphere-ocean general circulation models(AOGCMs) and observed climate values. This paper focuses on predictions of future climate on a regional scale which are of potential economic values. The proposed method is obtained by coupling the classical CCA with empirical orthogonal functions(EOF) for dimension reduction. Furthermore, we generate a distribution of climate responses, so that extreme events as well as a general feature such as long tails and unimodality can be revealed through the distribution. Results from real data examples demonstrate the promising empirical properties of the proposed approaches.

Keywords: Canonical correlation analysis, empirical orthogonal function, climate change, precipitation, prediction.

1. Introduction

Predictions of future climate change are issues of global importance affecting the whole ecosystem and human community. Especially, regional climate change projections are critical in many areas such as food production, planning of industrial investment, energy demand and water resources. This paper mainly emphasizes the future precipitation behavior in summer season near east Asia region including Korea. A better projection of future precipitation changes during summer season in Korea is crucial for controlling water resources and various industries such as insurance company.

The goal of this study is to develop a statistical method for reliable prediction of future climate values and to provide a system for probabilistic prediction of climate.

[1]Corresponding author: Department of Statistics, Seoul National University, Seoul 151-747, Korea.
E-mail: yaeji@snu.ac.kr

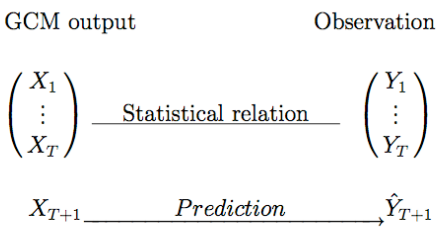

Figure 1.1. The prediction procedure of MOS.

There are three conventional approaches for climate prediction, namely a purely statistical approach, a dynamical approach using the GCMs and model output statistics(MOS). Pure statistical approach is the method that predicts unobserved climate value through a parametric statistical equation based on past observed values. A popular choice is regression analysis. A purely statistical approach has been used from the beginning of quantitative climate researches. However, it may not be appropriate to perform a long-term prediction, because a statistical equation is valid only in the range of the past climate which has been used for building the equation.

Numerical experiments based on coupled AOGCMs are typically used for projections of future climate change. These models are based on the integration of a variety of fluid dynamical, chemical, and sometimes biological equations. Although AOGCMs do an excellent job of representing global climate change, it has been known that numerical realizations from AOGCMs are very limited to represent regional climates (Greene *et al.*, 2006).

The MOS approach more recently developed combines the above two approaches (Wilks, 2006; Storch and Zwiers, 1999). Suppose that we have GCM output variables X_t and observations of interests Y_t $(t = 1, \ldots, T)$, where t denotes a time point. The prediction procedure by MOS is two-fold: 1) make a statistical equation that reflects the relationship between Y_t and X_t, and 2) predict future climate value Y_{T+1} at $T + 1$ time point using the statistical equation and X_{T+1} which is available from numerical experiments. The procedure is clearly displayed in Figure 1.1. The prediction accuracy of MOS is generally far better than either a pure statistical model or a AOGCM prediction.

However, typically X_t is a large dimensional vector of model output values including temperature, precipitation and pressure evaluated on a large spatial scale. The vector of past climate observations Y_t has relatively small dimension. That is, X_t and Y_t are p and q dimensional vectors with $p, q \geq T$. In the cases where the number of variables $p$and q are larger than the number of observation T, the classical statistical methods such as regression analysis cannot be directly applicable due to singularity problems.

In this paper, a statistical method based on CCA coupled with EOF is applied to predict monthly precipitation anomalies with lead times of up to 3 months. Previously, Landman and Goddard (2002) used a similar approach for regional rainfall forecasts of southern Africa. Here, by refinement and elaboration of the hybrid method by coupling CCA and EOF, we provide a clear explanation of the procedure on statistical theoretic ground and provide an exact formulation of prediction mechanism, so that anyone can use this method for one's climate prediction. Furthermore, we propose a resampling-based method to estimate probability distribution function(PDF) of the climate values. Extreme events as well as a general trend of climate values will be revealed through the information of the PDF, so that it will be useful tools for probabilistic prediction.

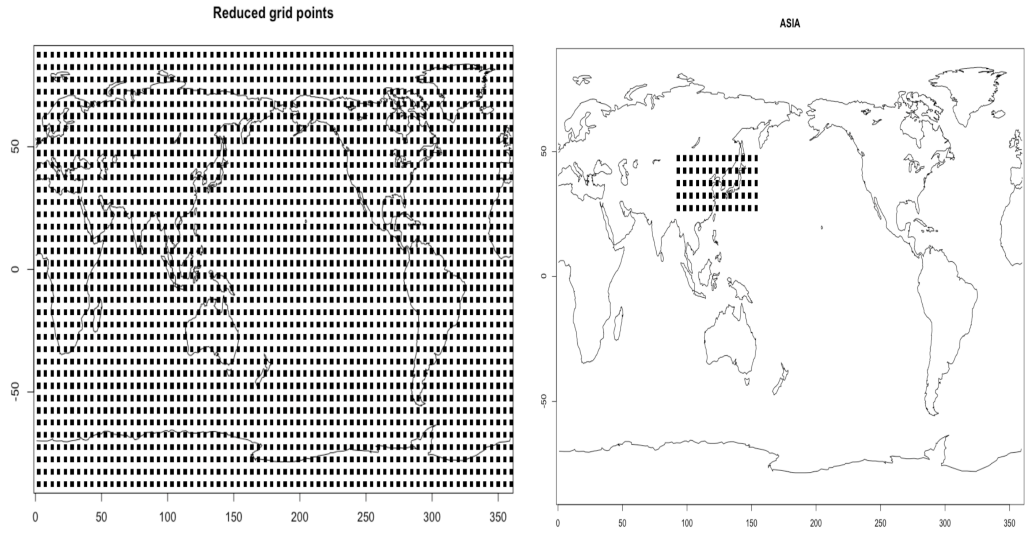

Figure 2.1. The left panel indicates reduced grid points by merging adjacent 9 grid points from 144×73 grid points which covers the whole global at regular interval by 2.5, and the right one shows the region of the data used for the regional precipitation prediction.

The rest of the paper is organized as follows. Section 2 contains a general information of data used in this study. In Section 3, we present the statistical methodology. We apply the methodology to data and discuss the results in Section 4. The conclusions are offered in Section 5.

2. Data

2.1. GCM output data

We use GCM outputs of Global Data Assimilation and Prediction System(GDAPS) from Korea Meteorological Administration(KMA). It is monthly data from 1979 to 2008 that covers the whole globe by 2.5×2.5 grid. There are 2-dimensional data such as sea level pressure(SLP), surface u wind(Us) and ground temperature(Ts) and also 3-dimensional data such as height(HGT), u wind(U) and surface t-td(DEPR) ranged 15 layers. As the initial time changes, there are 20 ensemble members in GCM outputs. For seasonal predictions, we average 3 consecutive months such as representing summer by June, July and August(JJA) and winter by December, January and February(DJF). In this study, we merge adjacent 9 grid points as one new point with equal weight as in the left panel of Figure 2.1 for effective computations.

2.2. Observed data

For observations of climate values, we analysis precipitation data from CPC Merged Analysis of Precipitation(CMAP) anomalies analyzed the Climate Research Unit, UK during the period 1979–2007. These are monthly data in 144×73 grid points that covers the whole global at regular interval 2.5×2.5. Similarly, we merge adjacent 9 grid points as one new point with equal weight, so that the number of resulting grid points are 2592.

2.3. Regional precipitation data

For regional climate prediction, we collect a portion of the CMAP data. The range of the data is 27.5°–47.5°N and 93.15°–153.57°W which covers only east Asia region including Korea in the right panel of Figure 2.1.

3. Methods

3.1. Prediction method

The future climate value Y_{T+1} can be predicted by

$$\hat{Y}_{T+1} = f(X_{T+1}),$$

where X_{T+1} is a GCM output at a future time point $T+1$ and f represents the statistical relation between predictors $\{X_t\}_{t=1}^T$ and responses $\{Y_t\}_{t=1}^T$.

Given the data $\{X_t\}_{t=1}^T$ and $\{Y_t\}_{t=1}^T$, building the system f is the most important task for reliable predictions. A popular approach is the classical regression analysis, where the f is a linear function. However, the classical regression cannot be directly applied to this problem because X and Y are high-dimensional data. For instance, for global precipitation prediction, we need to consider millions of variables such as GCM precipitation field, pressure level, wind, etc, while the number (year or month) of observation are so limited. Moreover, there are spatial correlations between two variables, X and Y. We cannot obtain well-performed models without considering this correlation.

To solve the dimensionality problem and enhance the prediction power simultaneously, we propose to use a hybrid method based on EOF and CCA. As two main ingredients of the hybrid method, EOF is employed for dimensionality reduction of climate data, and CCA plays an important role in improving the prediction ability. To be specific, the hybrid method consists of three steps: (1) dimension reduction by EOF, (2) constructing prediction equation based on CCA, and (3) synthesizing predicted values by regression.

Suppose that X be a $p \times T$ dimensional matrix of GCM outputs and Y be a $q \times T$ dimensional matrix of a climate response variable. In the first step, by using EOF, X and Y can be represented by

$$X = \sum_{i=1}^{p} \alpha_i \boldsymbol{\xi}_i \quad \text{and} \quad Y = \sum_{j=1}^{q} \beta_j \boldsymbol{\gamma}_j,$$

where $\boldsymbol{\xi}_1, \ldots, \boldsymbol{\xi}_p$ and $\boldsymbol{\gamma}_1, \ldots, \boldsymbol{\gamma}_q$ are orthonormal basis functions which are the eigenvectors of $\mathrm{Cov}(X, X)$ and $\mathrm{Cov}(Y, Y)$, respectively. Since the bases are orthonormal, the coefficients can be expressed as

$$\boldsymbol{\alpha} = \Xi^T X \quad \text{and} \quad \boldsymbol{\beta} = \Gamma^T Y,$$

where $\Xi = [\boldsymbol{\xi}_1 \cdots \boldsymbol{\xi}_p]$ and $\Gamma = [\boldsymbol{\gamma}_1 \cdots \boldsymbol{\gamma}_q]$. Thus, the individual coefficients are $\alpha_i = X^T \boldsymbol{\xi}_i$ $(i = 1, \ldots, p)$ and $\beta_j = Y^T \boldsymbol{\gamma}_j$ $(j = 1, \ldots, q)$. For a subsequent analysis, we select the first few bases, $p_1 (\leq p)$ and $q_1 (\leq q)$, so that X and Y can be approximated by

$$X \approx \sum_{i=1}^{p_1} \alpha_i \boldsymbol{\xi}_i \quad \text{and} \quad Y \approx \sum_{j=1}^{q_1} \beta_j \boldsymbol{\gamma}_j.$$

Therefore, after applying EOF to X and Y respectively, we obtain new variables $\boldsymbol{\alpha} = \{\alpha_1, \ldots, \alpha_{p_1}\}$ and $\boldsymbol{\beta} = \{\beta_1, \ldots, \beta_{q_1}\}$ which are linear combinations of the original data chosen to represent the

maximum possible fraction of the variability contained in the original data. To perform CCA in the next step, the data X and Y should have less variables than the sample size so that the corresponding inverse matrices can be computable. In general, climate data is high dimensional, which has more variables than the sample size, so that the CCA method cannot be directly employed. For this reason, EOF analysis is performed on the both GCM outputs and observed variables.

In the second step, we apply the CCA to the new variables $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$. CCA is used as one of the MOS methods that relate GCM outputs and observed climate variables (Glahn, 1963). It seeks vectors $\boldsymbol{\phi}_1$ and $\boldsymbol{\psi}_1$ such that the variables $U_1 = \boldsymbol{\alpha}^T\boldsymbol{\phi}_1$ and $V_1 = \boldsymbol{\beta}^T\boldsymbol{\psi}_1$ maximize the correlation $\rho = \mathrm{Corr}(U_1, V_1)$, termed the first canonical correlation. Here, the variables U_1 and V_1 are the first pair of canonical variables, and the vectors $\boldsymbol{\phi}_1$ and $\boldsymbol{\psi}_1$ are the first pair of canonical coefficients. Subsequently, we try to find vectors maximizing the same correlation subject to the constraint that they are to be uncorrelated with the first pair of canonical variables; this gives the second pair of canonical variables. After the above procedure is repeated at p_2 times, we obtain two canonical variables $U = (U_1, \ldots, U_{p_2})^T = \Phi^T\boldsymbol{\alpha}$ and $V = (V_1, \ldots, V_{p_2})^T = \Psi^T\boldsymbol{\beta}$, where $\Phi = \{\phi_1, \ldots, \phi_{p_2}\}$ and $\Psi = \{\psi_1, \ldots, \psi_{p_2}\}$ with $p_2 \leq \min\{p_1, q_1\}$. Then, by using best linear unbiased prediction(BLUP) theory, we obtain predicted values $\hat{V}_i$ $(i = 1, \ldots, p_2)$ as

$$\hat{V}_i = \frac{\mathrm{Cov}(V_i, U_i)}{\mathrm{Cov}(U_i, U_i)} U_i = \frac{\mathrm{Cov}(\boldsymbol{\beta}^T\boldsymbol{\psi}_i, \boldsymbol{\alpha}^T\boldsymbol{\phi}_i)}{\mathrm{Cov}(\boldsymbol{\alpha}^T\boldsymbol{\phi}_i, \boldsymbol{\alpha}^T\boldsymbol{\phi}_i)} \boldsymbol{\alpha}^T\boldsymbol{\phi}_i = \frac{\boldsymbol{\psi}_i^T \mathrm{Cov}(\boldsymbol{\beta}, \boldsymbol{\alpha})\boldsymbol{\phi}_i}{\boldsymbol{\phi}_i^T \mathrm{Cov}(\boldsymbol{\alpha}, \boldsymbol{\alpha})\boldsymbol{\phi}_i} \boldsymbol{\alpha}^T\boldsymbol{\phi}_i.$$

As one can see, the whole procedure depends on the estimators of $\mathrm{Cov}(\boldsymbol{\alpha}, \boldsymbol{\alpha})$ and $\mathrm{Cov}(\boldsymbol{\alpha}, \boldsymbol{\beta})$. A more accurate estimator of covariances will improve the quality of prediction. Although the above two steps produce the most highly related canonical patterns of X and Y with a simple calculation for prediction, it is performed on new canonical variables, but we are interested in the prediction of the response variable Y.

To accomplish the prediction of the response variable, it is necessary to synthesize $\hat{Y}$ from the predicted canonical variables V. To that end, by using BULP theory again, we obtain predictions of Y as $\mathrm{Cov}(Y, V)\mathrm{Cov}(V, V)^{-1}V$. By simply plugging an estimate $\hat{V}$ into V, we finally obtain the predicted value $\hat{Y} = \mathrm{Cov}(Y, V)\mathrm{Cov}(V, V)^{-1}\hat{V}$, where $\hat{V} = (\hat{V}_1, \ldots, \hat{V}_{p_2})$.

3.2. Probabilistic prediction using distribution of climate values

Here we propose a procedure for probabilistic prediction. To that end, it is required to obtain PDF of climate values, which is useful tool for evaluating a normal event or an extreme one. For estimating PDF of climate values, we generate predicted values by cross-validation which is very popular statistical method to enlarge the sample. In this study, we used one-year-out cross-validation. More precisely, one year is removed from the observed N years. Then the prediction method described in Section 3.1 is applied to the remaining $N - 1$ year data, and evaluate the predicted value for the removed year. After the above procedure is repeated over all N years, we can generate one-year-out cross-validated data $\hat{y}_1, \ldots, \hat{y}_N$.

Given the data $\{\hat{y}_i\}_{i=1}^N$, we estimate PDF of Y based on kernel density estimation method,

$$\hat{f}_h(y) = \frac{1}{Nh} \sum_{i=1}^{N} K\left(\frac{y - \hat{y}_i}{h}\right),$$

where K is proper kernel function and h is the smoothing parameter. More detail descriptions of the kernel density estimation are in Parzen (1962).

Once the PDF of Y is obtained, we evaluate a lower and an upper quantile values which can be

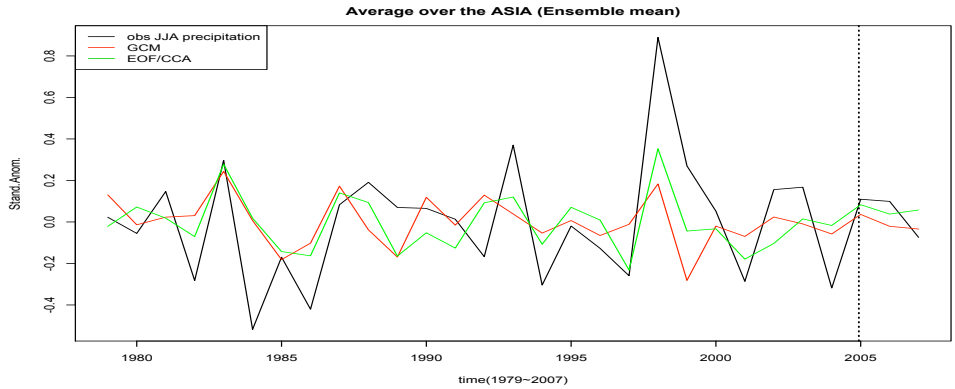

Figure 4.1. Time series of the comparing prediction for the Asia precipitation. The observed precipitation (black line), GCM output (red line), and the result of the hybrid method (green line) are displayed.

used for threshold values to separate abnormal events from normal climates. Finally, we perform probabilistic prediction using the values generated from the next year's GCM outputs.

4. Results

4.1. Performance of the prediction method

For prediction of precipitation, we select JJA season GCM simulation precipitation field as the best predictor X. In fact, the GCM simulation precipitation field produces the highest correlation coefficient with the JJA season observed precipitation field Y. Figures 4.1 and 4.2 show the JJA season prediction results of the hybrid method. To perform the prediction, we separate the whole years into training years(1979–2004) and test years(2005–2007). The average fitted values and average prediction values across years from 20 ensemble GCM members are displayed in Figures 4.1. As shown, the overall patterns of prediction by the hybrid method are much closer to the real observations than those of GCM results. Unfortunately, in the last test year, 2007, the GCM prediction is closer to observed precipitation than prediction results of the hybrid method. But this does not means that GCM prediction is better, it might be just due to the rapid change of observed precipitation. Through the image plot of the globe in Figure 4.2, we can compare the performance of the hybrid method with the GCM simulation of precipitation during JJA season of year 2007. As shown, the hybrid method can capture the overall trend as well as an important local feature such as La nina well. For an accurate comparison, we compute root mean squared error(RMSE) for each method. RMSE is defined as

$$\mathrm{RMSE}(\boldsymbol{y}, \hat{\boldsymbol{y}}) = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2},$$

where y_i indicates observed precipitation at the i-th grid point, and $\hat{y}_i$ denotes the predicted values at the i-th grid point. The RMSE by the hybrid method is 0.619, while the RMSE from the GCM simulation is 0.806. Therefore, the values of RMSE support the results of the image plot in Figure 4.2. Furthermore, when we focus on the performance of regional prediction on Asia area in Figures 2.1, the RMSE of the hybrid method and GCM are 0.600 and 1.046, respectively.

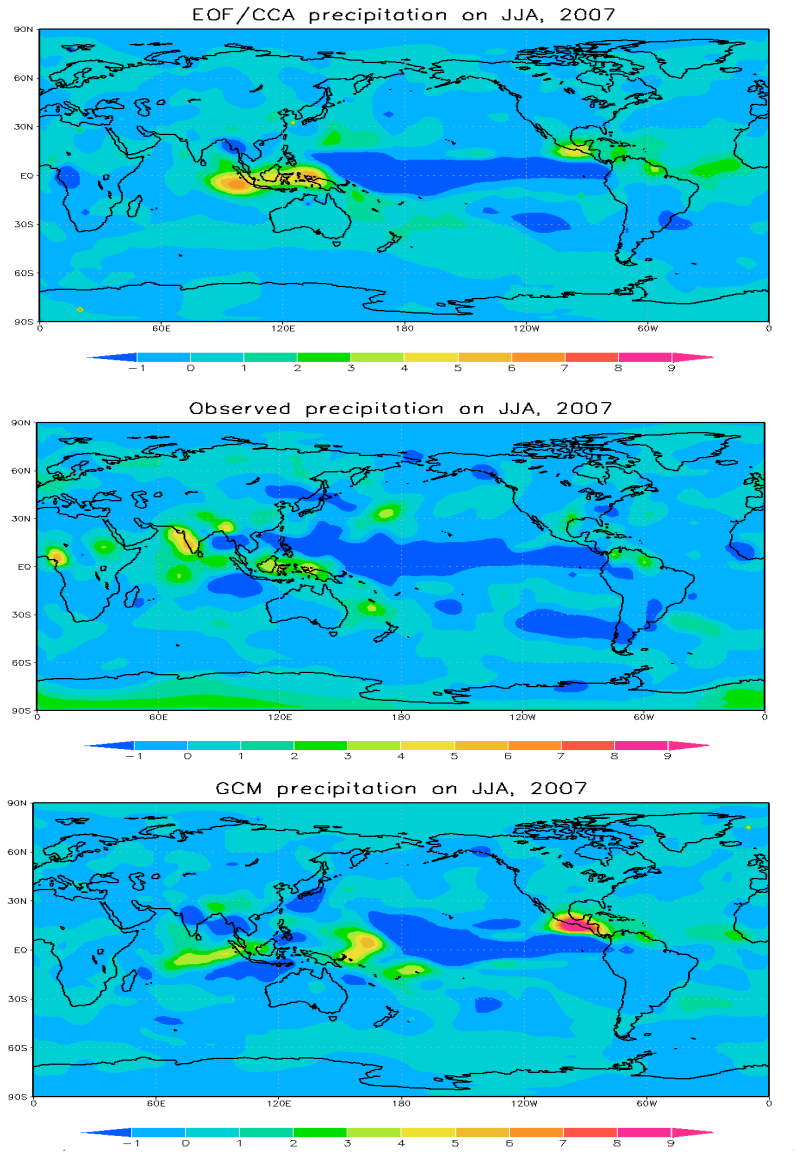

Figure 4.2. The globe image plot of the predicted precipitation in 2007. The first panel shows the result from the EOF/CCA method, the second one indicates the observed precipitation and the third one is for the GCM predicted value. All variables were standarized to remove its unnecessary trends.

4.2. Probability distribution function of precipitation

By following the method described in Section 3.2, we compute cross-validated predicted values $\hat{y}$'s for 20 ensemble members and 29 years. Subsequently, the PDF of precipitation in Asia region is estimated by applying kernel density estimation to $\hat{y}$'s. For comparison, we also generate a PDF of GCM precipitation outputs. Figure 4.3 shows two estimated PDF with two thresholds(30 and

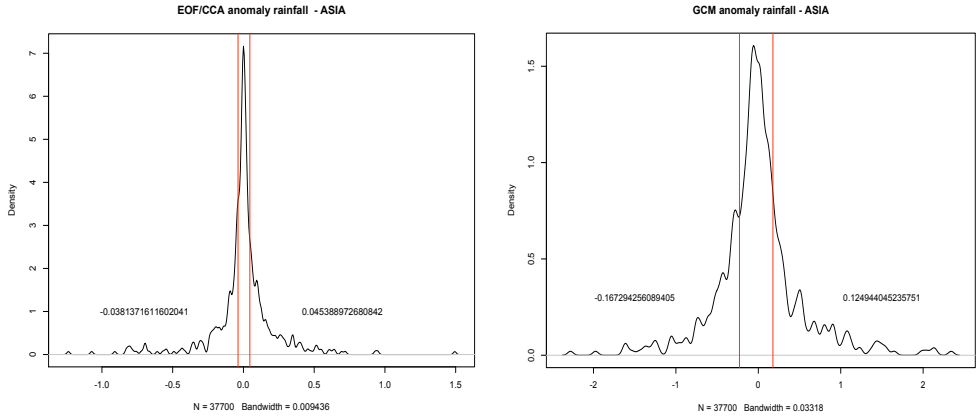

Figure 4.3. Estimated probability distribution function by applying the proposed method to 1979–2007 JJA season, Asia regional precipitation data. The left panel shows the estimated PDF of hybrid method, and the right one is for the GCM output.

70 percentiles). The left panel is PDF based on the predicted precipitations of the hybrid method, and the right panel is PDF from GCM outputs. Since the PDF of EOF/CCA method is longer and narrower, it is considerably less uncertain than GCM based PDF.

4.3. Validations

For assessing the quality of prediction, we employ two methods such as linear error in probability space(LEPS) and receiver operating characteristic(ROC) curve. LEPS score is defined as

$$\text{LEPS} = \frac{1}{N}\sum_{i=1}^{N} |cdf(F_i) - cdf(O_i)|,$$

where *cdf* denotes cumulative PDF, F_i denote the predicted value of the ith grid point, and O_i denotes the observation at the ith grid point. The value of LEPS is located in between 0 and 1. The perfect predicted case has 0 as LEPS value. Figure 4.4 displays LEPS scores from the predicted values of Asia region precipitation during years 2005–2007. As shown, the LEPS scores of the hybrid method are closer to 0 than those of GCM simulation.

As another validation measure, we consider ROC curve which has been widely used for visualizing and analyzing the behavior of diagnostic systems. ROC curves are two-dimensional graphs where true positive rate is plotted on the y-axis and false positive rate is plotted on the x-axis. An ROC curve depicts relative trade-off between benefits and costs. Thus, one point in ROC curve is better than another if it is located on upper-left area. Figure 4.5 shows ROC curves of the hybrid method and GCM simulation. As shown, the ROC curve of the hybrid method is more biased to left up side, which means that it provides more prediction accuracy than GCM prediction. Furthermore, for more accurate comparison, we compute the ROC area which is the area under the curves. The ROC area for the hybrid method is 0.5287 which is bigger than the GCM prediction's 0.5071.

5. Conclusions

In this paper, we have predicted the future precipitation values by a hybrid method which is based on CCA and EOF method. Since the dimension of climate data is very high, it is hard to

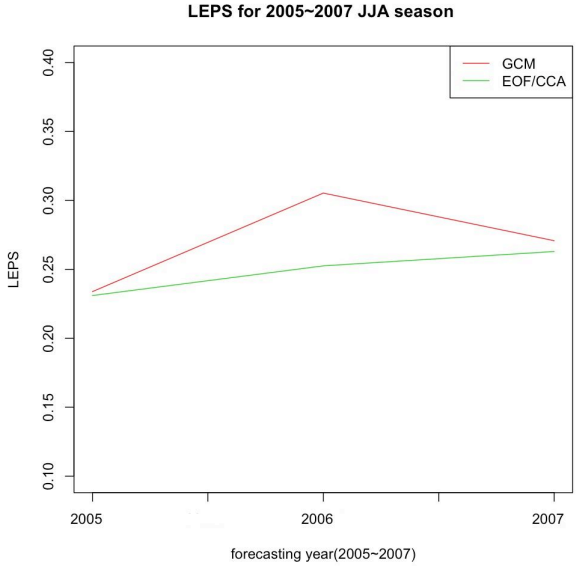

Figure 4.4. LEPS for the 2005–2007, JJA season for the each method. The green line is for the hybrid method and the red one is for GCM output.

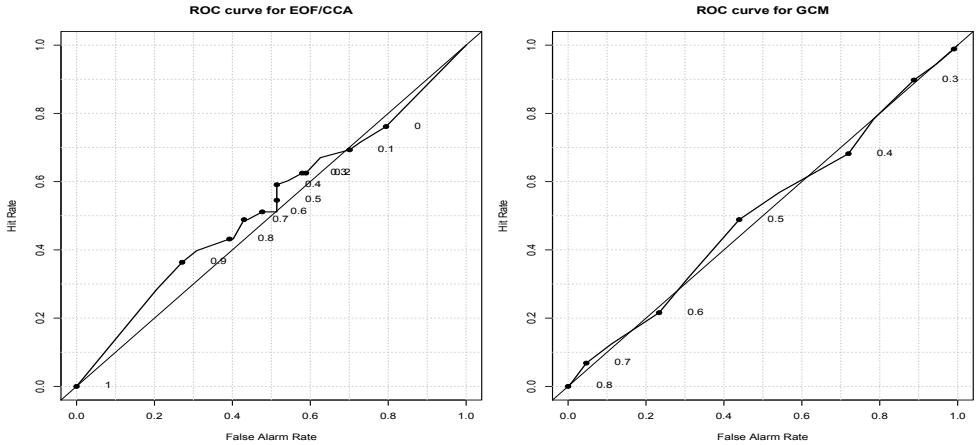

Figure 4.5. The ROC curves by the hybrid method(left) and GCM predicted precipitation(right) in 2005–2007. The numbers are indicating threshold values.

apply a conventional statistical method directly. The key idea of the hybrid method is reducing dimension using EOF analysis and then apply CCA to the reduced data. The method consists of three steps. The first step is reducing the dimension of GCM output, X and observed data, Y using EOF analysis. The second step is applying CCA to the reduced data. Then we obtain canonical variables U and V and using BLUP theory, construct a prediction equation. The last step is completing prediction by regression analysis. We show that prediction results by this method outperform GCM-simulated prediction. In addition, we have provided a statistical system for probabilistic prediction which is coupled of kernel density estimation with cross-validation method.

References

Glahn, H. (1963). Canonical correlation and its relationship to discriminate analysis and multiple regression, *Journal of the Atmospheric Sciences*, **25**, 23–31.

Greene, A. M., Goddard, L. and Lall, U. (2006). Probabilistic multimodel regional temperature change projections, *Journal of Climate*, **19**, 4326–4343.

Landman, W. A. and Goddard, L. (2002). Statistical recalibration of GCM forecasts over southern Africa using model output statistics, *Journal of Climate*, **15**, 2038–2055.

Parzen, E. (1962). On estimation of a probability density function and mode, *The Annals of Mathematical Statistics*, **33**, 1065–1076.

Storch, H. V. and Zwiers, F. W. (1999). *Statistical Analysis in Climate Research*, Cambridge.

Wilks, D. S. (2006). *Statistical Methods in the Atmospheric Sciences*, Academic Press.