Theor Appl Climatol (2009) 97:195–203
DOI 10.1007/s00704-008-0056-y

ORIGINAL PAPER

Factor analysis for El Niño signals in sea surface temperature and precipitation

Christine K. Lee · Samuel S. P. Shen · Barbara Bailey · Gerald R. North

Received: 24 February 2008 / Accepted: 12 August 2008 / Published online: 17 September 2008

Abstract Maximum likelihood factor analysis (MLFA) is applied to investigate the variables of monthly Tropical Pacific sea surface temperatures (SST) from Niño 1+2, Niño 3, Niño 3.4, and Niño 4 and precipitation over New South Wales and Queensland of eastern Australia, Kalimantan Island of Indonesia, and California and Oregon of the west coast of the United States. The monthly data used were from 1950 to 1999. The November-February SST with time leads of 0, 1, 2, and 3 months to precipitation are considered for both El Niño warm phases and non El Niño seasons. Interpretations of the factor loadings are made to diagnose relationships between the SST and precipitation variables. For El Niño signals, the rotated FA loadings can efficiently group the SST and precipitation variables with interpretable physical meanings. When the time lag is 0 or 1 month, the November–February El Niño SST explains much of the drought signals over eastern Australia and Kalimantan. However, when the time lag is 2 or 3 months, the same SST cannot adequately explain the precipitation during January–May over the two regions. Communality results of five factors for precipitation indicate nearly 100% explanation of variances for Queensland and California, but the percentages are reduced to only about 30% for Oregon and Kalimantan. Factor scores clearly identify the strongest El Niño relevant to precipitation variations. Principal component factor analysis (PCFA) is also investigated, and its results are compared with MLFA. The comparison indicates that MLFA can better group SST data relevant to precipitation. The residuals of MLFA are always smaller than the PCFA. Thus, MLFA may become a useful tool for improving potential predictability of precipitation from SST predictors.

C. K. Lee · S. S. P. Shen (✉) · B. Bailey
Department of Mathematics and Statistics,
San Diego State University, San Diego, CA 92182, USA
e-mail: shen@math.sdsu.edu

G. R. North
Department of Atmospheric Sciences, Texas A&M University,
College Station, Texas 77843, USA

1 Introduction

Factor analysis (FA) is a multivariate statistical analysis method that decomposes a covariance or correlation matrix and reduces the dimension of data (Gorsuch 1983). It has often been used in social sciences, marketing, and operations research. However, its use in climate research is still rare, although some successful applications have effectively demonstrated its value. Examples include identification of dynamical modes (Walters 1999) and correlation patterns (Bartzokas and Metaxas 1995; Dinpashoh et al. 2004).

The commonly used FA decomposition methods are the empirical orthogonal function (EOF) approach and the maximum likelihood (ML) approach (Johnson and Wichern 1992; Bartzokas and Metaxas 1995; Wilks 2006). The purpose of this paper is to explore the advantages of using ML factor analysis (MLFA) in climate data analysis. In the process, we further introduce the fundamentals of the method to climate research. As a computational example, MLFA is used to investigate the relationships between the Tropical Pacific sea surface temperature (SST) in Niño 1+2, Niño 3, Niño 3.4, and Niño 4 and precipitation data of five $5° \times 5°$ land grid boxes over Kalimantan of Indonesia, New South Wales and Queensland of Australia, and California and Oregon of the United States.

The EOF analysis, also known as principal component (PC) analysis, has been extensively used in climate data analysis for various purposes: explaining dynamical modes, exploring teleconnections of climate parameters, reducing data dimension, and identifying footprints of forced climate changes (North et al. 1982, 1995). The loadings of the PC factor analysis (PCFA) are the EOFs multiplied by their corresponding variances (i.e., eigenvalues), and thus the PCFA approach maximizes the variance of the climate component projected onto the EOF direction, and the PCFA computation is a well-defined eigenvalue problem. The ML approach, by name, maximizes the normally distributed multivariate likelihood function. The computational algorithm for ML factor analysis (MLFA) is not a simple eigenvalue approach like that of PCFA; instead, it is an iterative approach and the procedures are quite complicated. Fortunately there is standard statistical software such as SAS, to conduct MLFA analysis. The main advantage of MLFA is that it usually leads to smaller residuals compared to the PCFA. Thus, MLFA can identify patterns or modes that can better represent the covariance or correlation matrix, in particular, the off-diagonal elements. Such elements are also more closely associated with climate dynamics and atmospheric circulations. Thus, the loyalty implies more meaningful interpretation of the factor loadings. This is important for statistical prediction of monthly or seasonal climate since the skills of linear prediction such as canonical correlation analysis (CCA) method or direct multivariate regression method, come from the correlation between the predictor and predictand (Barnett and Preisendorfer 1987; Barnston and Smith 1996; Shen et al. 2001; Lau et al. 2002).

To demonstrate the advantages of MLFA, we chose to analyze sea surface temperatures (SST) and its relationship to the precipitation, because SST data are often used as predictors for monthly and seasonal temperature and precipitation predictions. We explore the El Niño MLFA loadings and specific variances imbedded in the SST data over the tropical Pacific (Niño 1+2, Niño 3, Niño 3.4, and Niño 4) and precipitation data of five 5°×5° land grid boxes over Kalimantan of Indonesia, New South Wales and Queensland of Australia, and California and Oregon of the United States. The monthly data used were from 1950 to 1999. The SST leads of 0, 1, 2, and 3 months to precipitation are considered for both El Niño warm phases and non El Niño seasons. We have found that, compared with PCFA, the MLFA loadings in the calculated factors can more effectively group the SST and precipitation variables with interpretable physical meanings. When the time lag is 0 or 1 month, the November-February El Niño explains much of the precipitation signals over eastern Australia and Kalimantan, but when the time lag is 2 or 3 months the same SST cannot adequately explain the precipitation during January-May over the two regions.

The rest of the paper is arranged as follows: section Method and data describes the data and analysis methods, Results are included in the third section, and Conclusions and discussion are presented in the fourth section.

2 Method and data

The FA's mathematical expression is that an N-dimensional covariance or correlation matrix is decomposed into M common factors of N-dimension and a diagonal matrix consisting of specific variances, i.e.,

$$\sum\nolimits_{ij} = \langle X_i X_j \rangle = \sum_{m=1}^{M} l_{im} l_{jm} + \psi_i \delta_{ij} + R_{ij}, \quad i,j = 1,2,\ldots,N \tag{1}$$

where $\sum_{ij}$ is the covariance matrix for climate anomaly data, $\langle \cdot \rangle$ stands for the ensemble mean operation, l_{im} is the ith variable X_i's loading of mth factor, ψ_i is the specific variance for X_i, δ_{ij} is the Kronecker delta, and R_{ij} is the decomposition residual—see chapter 9 of Johnson and Wichern (1992). For many covariance matrices, there exists a sufficiently large M_o, but still less or equal to N, such that R_{ij}=0, when $M = M_o$ However, the practically useful FA retains only a few factors, i.e., M is small compared to the number of variables N, and it is expected that these few factors and the specific variances can accurately approximate the covariance matrix $\sum_{ij}$ with a very small residual R_{ij}. Then, the orthogonal factor model can allow the variable decomposition into common factors in the following way

$$X_i = \sum_{m=1}^{M_o} F_m l_{im} + \varepsilon_i, \quad i = 1,2,\ldots,N, \tag{2}$$

where F_m is the mth common factor with the following orthonormal properties,

$$\langle F_m F_n \rangle = \delta_{mn}, \quad \langle \varepsilon_i \varepsilon_i \rangle = \left\langle \varepsilon_i^2 \right\rangle \delta_{ij}, \quad \langle F_m \varepsilon_i \rangle = 0. \tag{3}$$

Again, the practically useful FA requires a small number of factors M. Thus, the decomposition Eq. (2) becomes

$$X_i = \sum_{m=1}^{M} F_m l_{im} + \sum_{m=M+1}^{M_o} F_m l_{im} + \varepsilon_i, \quad i = 1,2,\ldots,N. \tag{4}$$

The diagonal elements of the decomposition residual are forced to be zero. Hence,

$$R_{ij} = \sum_{m=M+1}^{M_o} l_{im} l_{jm} - \sum_{m=M+1}^{M_o} l_{im}^2 \tag{5}$$

$$\psi_i = \left\langle \varepsilon_i^2 \right\rangle + \sum_{m=M+1}^{M_o} l_{im}^2 \tag{6}$$

The sum of the factor loadings across the *i*th variable is called the communality

$$h_i^2 = \sum_{m=1}^{M} l_{im}^2, \tag{7}$$

which represents the proportion of the total variance that the M retained factors can explain. The specific variances ψ_i represent the proportion of the total variance for which the M retained factors cannot explain. Thus, the relationship between the specific variances ψ_i and communalities $h_i^2 = \sum_{m=1}^{M} l_{im}^2$ is

$$\sum\nolimits_{ii} = h_i^2 + \psi_i. \tag{8}$$

In cases where the correlation matrix is used, Eq. (7) reduces to $h_i^2 + \psi_i = 1$, a property important for iterative methods of FA.

Using FA to explore the El Niño signals and relationship imbedded in SST and precipitation data requires data from locations with strong El Niño signatures: SSTs from tropical Pacific regions Niño 1+2, Niño 3, Niño 3.4, and Niño 4, and precipitation from Kalimantan of Indonesia, New South Wales and Queensland of Australia, and California and Oregon of the United States. The SST data are from the US Climate Prediction Center's (CPC) Monthly Atmospheric and SST Indices. The CPC monthly SST anomaly data over Niño 1+2 (0–10°S, 90–80°W), Niño 3 (5–5°S, 150–90°W), Niño 4 (5–5°S, 160–150°W), and Niño 3.4 (5–5°S, 170–120°W) are from January 1950 to current (see Fig. 1 for the locations of the Niño regions). Although the anomalies are based on the 1961–1990 climatology base period, the sum from 1961 to 1990 for each month over Niño regions is not exactly zero, because this dataset is the reconstructed SST (Climate Prediction Center 2007).

We chose to use 50 years of data for our analysis from January 1950 to December 1999, which corresponds to the period of good precipitation data with missing data of only 5 months over Kalimantan. The five missing months were August 1991, April 1992, May 1994, February 1995, and August 1995, and the missing data are filled by the data from the nearest grid box (5–10°S, 105–110°E) of Jakarta, Indonesia for the corresponding months.

Monthly precipitation data are from the US National Climatic Data Center's Global Historical Climatology Network (GHCN) 5°×5° gridded dataset. Again, the anomaly data were used and the climatology period is 1961–1990. Although the GHCN gridded data spans from 1900 to current, we choose to use the high quality data of few missing values over the five selected locations in the period between 1950 and 1999. Precipitation data are chosen from Indonesia, Australia, and the west coast of the United States because these areas are known to exhibit weather conditions related to El Niño that signifies tropical Pacific SST fluctuations. For Indonesia, the Kalimantan grid box (5°N–0, 110–115°E) of the island of Borneo is chosen because this area is known for drought, late rain months, and exacerbated bush and forest fires during an El Niño event (National Drought Mitigation Center 1997). For Australia, grid boxes in the southeastern state of New South Wales (35–40°S, 145–150°E) and the northeastern state of Queensland (20–25°S, 140–145°E) are selected. Both states are affected by El Niño-induced late autumn rainy seasons. The late rains exacerbate brush fires in New South Wales and fires from Southeast Asia blanket Queensland with smoky air. For the west coast of the United States, grid boxes over California (40–35°N, 125–120°W) and Oregon (45–40°N, 125–120°W) were chosen. Typical El Niño effects for these areas include earlier, heavier rainy seasons (in the winter months) for California. However, Oregon responds in an uncertain way to the El Niño variations of SST: often dry winters, but occasionally wet winters as well. Figure 1 shows a global picture of the locations of the chosen 5°×5° grid boxes and Niño regions. All together, nine variables (four SST variables and five precipitation variables) are chosen. On average, each El Niño episode has 4 months showing strong SST signals (National Weather Service Forecast Office 2004). Thus, the 10 episodes of El Niño during the 1950–1999 period have 40 months of data (Null 2004). Thus the data matrix for the FA analysis has 9 columns representing the 9 random variables and 40 rows representing 40 months of data. The covariance matrix is of order 9×9, i.e., N=9 in eq. (1).

Both SST and precipitation data require pre-processing before being used for FA. Data preprocessing include three steps: filling in missing values, standardizing anomalies and selecting years of strong El Niño effects, known as warm episodes. All data preprocessing and subsequent analysis are performed using SAS. Due to scale inhomogeneity in different months, SST and precipitation anomaly data are standardized before being used for FA. Since the SST data are already anomalies, we only need to standardize the data by dividing the anomalies standard derivation computed for the period of 1961–1990. In order to calculate the standardized precipitation anomalies, the precipitation data need to subtract 1961–1990 climatology before being divided by the 1961–1990 anomalies' standard deviation. The standardized data are closer to being stationary so that the temporal mean calculation of the covariance matrix makes sense approximately. The standard deviations for the SST over the four Niño zones range from 0.48°C (Niño 4,

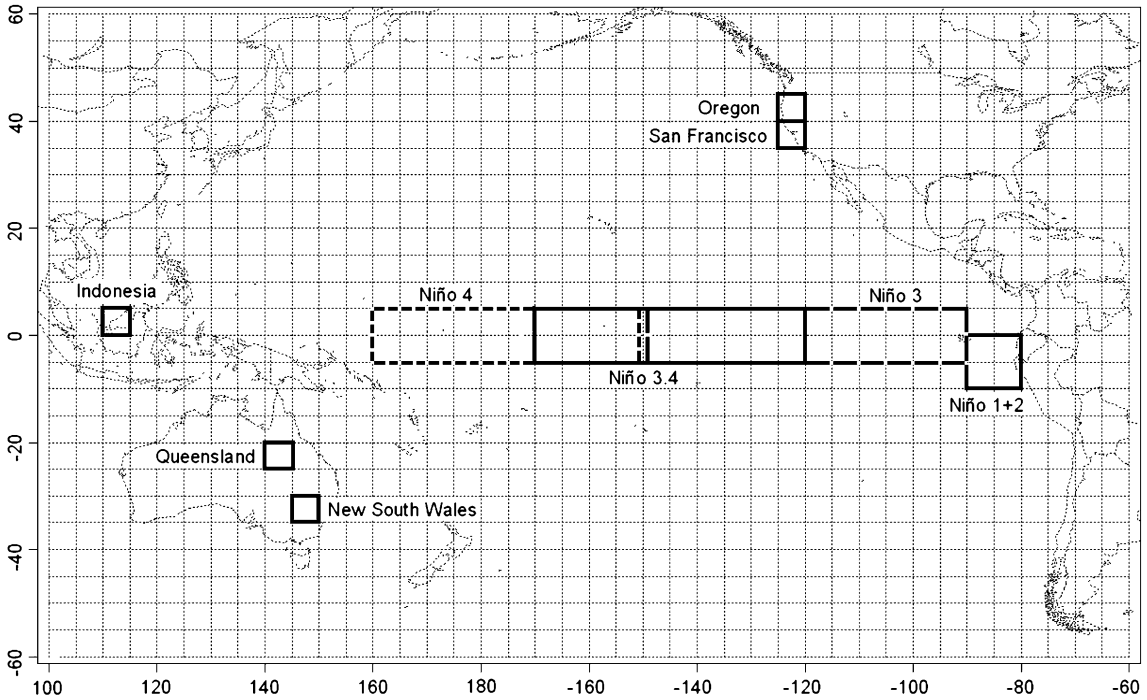

Fig. 1 Location of the five 5° × 5° grid boxes for precipitation data and the four Niño regions for SST data

April) to 1.36°C (Niño 1+2, June). The standard deviations for the precipitations over the five grid boxes range from 12.6 mm (California, July) to 183.4 mm (Kalimantan, January).

Our FA's focus is to explore El Niño effects on land precipitation in a few selected areas of strong El Niño signals. Thus, between 1950 and 1999, only 19 years (10 El Niño events) classified as Tropical Pacific warm phases are included in the analysis, which are 1957–58, 1965–66, 1972–73 (strong El Niño), 1977–1978, 1982–1983 (strong El Niño), 1987–1988, 1991–1992 (strong El Niño), 1992–1993, 1994–1995, and 1997-1998 (strong El Niño). For each event, the SST data from warm months are pooled for analysis: November, December, January, and February. The corresponding 4-month precipitation data for each event are pooled according to 0, 1, 2, and 3-month lags.

The FACTOR procedure in SAS is used for both MLFA and PCFA. Our FA's goal is to reduce the SST and precipitation variables over the spatial regions into a few FA factors, which may be physically interpretable. It is often that these factors tend to have one-sign loadings or even close-to-uniform loading for the first factors, which may be interpreted as the weighted spatial average that explains the most variance. However, physical interpretations often require interpretable observations of variables. Thus, a rigid rotation of the factors often needs to be done to make the factors readily interpretable. Our rotation is done by the varimax method, which rigidly rotates the M factors to maximize the variances of the square of the ratio between the factor loadings and the square root of the corresponding communalities (Johnson and Wichern 1992). After the rotation, the first factor does not usually correspond to the largest variance. Instead, rotated factors tend to group a few related factor loadings into each factor. The grouping shows large loadings for one or several variables and very small loadings of other variables. Therefore, the rotated factors become more readily interpretable physically. The varimax rotation method has been commonly used in climate research (Smith et al. 1998).

The SAS FACTOR procedures for MLFA and PCFA both yield ordered eigenvalues $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_N \geq 0$ and corresponding factors. The eigenvalues for the correlation matrix are the same for both methods, but factors are different. Usually MLFA factors lead to smaller elements in the residual matrix and are more accurate representations of off-diagonal elements of a correlation matrix. Thus, MLFA may yield improved potential predictability in a linear statistical climate forecasting that is based on a regression analysis and correlations between predictors and predictands.

Several criteria exist for selecting a sufficient number of eigenvalues for both MLFA and PCFA. For PCFA, the SAS default MINEIGEN=1 is used. This is also known as the Kaiser criterion. An eigenvalue and its corresponding factor are retained if the corresponding eigenvalue is larger than or equal to 1. The reasoning is that an eigenvalue should be able to extract at least as much variability as one of the original variables. For MLFA, the SAS default PROPOR-

TION=1 is used. Eigenvalues are retained until their cumulative percentage of variance explanation exceeds 100%. Due to its maximization of the likelihood function, sometimes negative values for the diagonal entries of ψ are obtained. These values are unacceptable for the ML iterations and are known as Heywood cases. When this occurs, the SAS HEYWOOD option is employed to reset out of bounds communalities to 1. Both MLFA and PCFA are conducted to see whether the factor loadings would reveal any underlying climate patterns.

For the PCFA, the explicit relationship between EOFs and factor loadings is as follows:

$$l_{im} = \sqrt{\lambda_m} e_{im}, \tag{9}$$

where e_{im} is the ith component (i.e., location i) of the mth normalized EOF (i.e., mode m) with $\sum_{i=1}^{N} e_{im}^2 = 1$. Because the MLFA may yield different factors from the PCFA, the MLFA factor loadings usually do not have the same relationship with EOFs shown in eq. (9). The differences between the MLFA and PCFA factors will be numerically demonstrated in the next section.

3 Results

In this section, we present and interpret the results of factor loadings for the SST over the four Niño regions and the precipitation over five land areas: eastern Australia, Kalimantan of Indonesia, and the western United States. It is generally regarded that the above SST and precipitation are strongly correlated with precipitation lagged behind the SST signals. Four time lags are considered here: 0, 1, 2, and 3 months. Our analysis shows that the factor loadings of our analysis for the El Niño warm phases clearly group the SST and precipitation parameters. This information can be useful for climate predictions.

For the zero time lag, the correlation matrix is shown in Table 1. The MLFA for the correlation matrix retains five factors and the loadings of the first four rotated factors are in Table 2. The loadings of factor 5 are all very small and have no apparent physical meaning.

Table 2 clearly indicates that the SSTs of the eastern tropical Pacific are grouped together to reflect the strong El Niño signal (see factor 1), while the Niño 4's SST is influenced by the relatively stable western Pacific warm pool (WPWP) and has shown little El Niño signal, which is depicted in factor 4. Eastern Australia's drought signal in an El Niño warm phase is clear and is well reflected in factor 2's loadings of 0.71 for the New South Wales grid box and 0.95 for the Queensland grid box, respectively. However, the El Niño drought signal over the Kalimantan grid box is not as strong (factor loading 0.44) because the precipitation over this region is more influenced by monsoon signals, which in turn is predominantly influenced by WPWP. The opposite signs of SST loadings with precipitation loadings in factor 2 reflect the drought occurrence of eastern Australia and eastern Indonesia. The same sign of SST loadings with the precipitation loadings reflect the United States west coast's wet phase locking with the eastern tropical Pacific's warming phase. The zero loading of Niño 4's SST in factor 3 implies that Niño 4's warm phase has little to do with the wet anomalies of the west coast of the United States. The large loading of Niño 4's SST in factor 4 indicates that in the El Niño and precipitation analysis, Niño 4 stands out as an independent predictor. The above analyses imply that the SST signal of the eastern tropical Pacific has precipitation teleconnections to the north reaching the west coast of the United States and to the southwest reaching Australia. However, the western tropical Pacific SST signal during the El Niño warm phase has little or no connection to the precipitation over the west coast of North America. The main reason is likely that the strength of the El Niño signal is in the eastern tropical Pacific, which is associated with a slightly inclined east-west oscillation of atmospheric pressure over the southern tropical Pacific.

Communalities shown in Table 2 are based on five factors and are invariant under the varimax rotation. The communalities indicate that the variances of the SSTs of Niño 1+2, Niño 3, and Niño 3.4, and the variances of Queensland and California precipitations are 100% explained by the five factors, while the variance of Oregon's precipitation is only explained to 28%. This corresponds to 33% explanation for Kalimantan and 65% for New South Wales. It is surprising that the variance of Niño 4 is only explained to 60%. These results provide very important information for monthly precipitation predictions by a statistical method: the necessity of using multiple predictors for a statistical prediction method (Shen et al. 2001; Lau et al. 2002). The current MLFA results strongly indicate that the forecasting of Oregon's monthly precipitation should also consider other predictors besides SST, and that Kalimantan's precipitation is not only influenced by the SSTs of the Niño regions, but also by the monsoon dynamics powered by the SSTs of the Indian Ocean and WPWP.

The MLFA's residuals are almost zero except for Oregon's precipitation which has non-zero residuals corresponding to Niño 4's SST, New South Wales' precipitation, and Kalimantan's precipitation. The possible reason is that Oregon's precipitation is heavily influenced by local microclimate and is sometimes de-phased from El Niño signals. In particular, Oregon's rugged and diverse terrain yields large differences in temperature (from –48°C to +48°C) and precipitation (average annual rainfall

Table 1 Correlation matrix of zero lag SST and precipitation during the El Niño warm phases

Correlation matrix	SST Niño 1+2	SST Niño 3	SST Niño 3.4	SST Niño 4	PCPN N S Wales	PCPN Queensland	PCPN Kalimantan	PCPN California	PCPN Oregon
Niño 1+2	1	.94	.84	.28	–.35	–.26	–.35	.33	.16
Niño 3	.94	1	.96	.35	–.37	–.32	–.33	.27	.16
Niño 3.4	.84	.96	1	.49	–.35	–.38	–.29	.22	.14
Niño 4	.28	.35	.49	1	–.17	–.32	–.01	.09	.06
N S Wales	–.35	–.37	–.35	–.17	1	.73	.27	.30	–.06
Queensland	–.26	–.32	–.38	–.32	.73	1	.41	.18	.07
Kalimantan	–.35	–.33	–.29	–.01	.27	.41	1	–.26	.04
California	.33	.27	.22	.09	.30	.18	–.26	1	.51
Oregon	.16	.16	.14	.06	–.06	.07	.04	.51	1

Status of El Niño year was based on consensus of four different climate centers: *WRCC*, *CDC*, *CPC*, and *MEI*

between 203 mm and 5,080 mm; National Oceanic and Atmospheric Administration 1986).

The PCFA retains only three factors according to the Kaiser criterion and it groups Niño 4's SST together with the SST signals of the other three eastern tropical Pacific regions in factor 1, which is different from the MLFA result that makes Niño 4's SST factor loading stand alone in factor 4. The groupings for precipitation are the same as MLFA. The residuals of PCFA are clearly larger than that those of MLFA even if five factors are retained. Thus, if one uses FA for climate predictions, it is likely that MLFA will render higher potential predictability.

With a 1-month time lag, the MLFA results are similar to those of the 0-month time lag, and physical implications are the same as those concluded for the 0-month time lag analysis. Niño 4's SST loading again stands out alone in factor 4. The 1-month lag results have practical values in statistical prediction of precipitation from SST by using various kinds of methods. In particular, the MLFA results may be used for the recently developed CEC (canonical ensemble correlation) method for monthly and seasonal predictions, since the CEC method can divide the predictor into several critical sub-regions and thus multiple predictors are possible (Shen et al. 2001; Lau et al. 2002). Consequently, due to the consideration of atmospheric circulation and climate dynamics, the CEC prediction is not a linear method, rather it is a quasi-linear method, i.e., linear only for an ensemble member.

The 2-month lag results have minor differences from those of 0-month and 1-month lags. The MLFA retains five factors and their loadings are shown in Table 3.

Table 3 indicates that the SST and precipitation factor groupings are more obvious than those shown in Table 2. The Niño 1+2, Niño 3 and Niño 3.4 SST loadings are very big in factor 1 (over .92), and the Niño 4's SST loading is also big (.96) in factor 4. The precipitation loadings are also large for Queensland (.98) in factor 2 and California (.95) in factor 3. These results indicate high potential predictability of precipitation of the two regions when using SST as a predictor. Kalimantan's precipitation loading is now in factor 3, rather than factor 2 as in the 0- and 1-month time lags. This seemingly surprising result from the point of

Table 2 MLFA rotated factor loadings and communality for 0-month time lag

	Factor 1	Factor 2	Factor 3	Factor 4	Communality
Niño 1+2	.92[a]	–.21	.20	.07	1.00[b]
Niño3	.95[a]	–.21	.15	.15	1.00[b]
Niño 3.4	.87[a]	–.21	.12	.37	1.00[b]
Niño 4	.22	–.12	.00	.74	.61[b]
New S. Wales	–.26.27	.71[a]	.26	.00	.65[b]
Queensland	.09	.95[a]	.10	–.26	1.00[b]
Kalimantan	–.20	.44[a]	–.29	.01	.33
California	.14	.12	.97[a]	.10	1.00[b]
Oregon	.09	.02	.51[a]	–.04	.28

[a] Numbers in the same column of factors are classified as a group according to relative importance of the factor loadings
[b] Indicates a significant explanation of variances by the factors

Table 3 MLFA rotated factor loadings and communality for 2-month time lag

	Factor 1	Factor 2	Factor 3	Factor 4	Factor 5	Communality
Niño 1+2	.92[a]	–.01	.13	.06	.38[a]	1.00[b]
Niño 3	.98[a]	.02	.13	.10	.05	1.00[b]
Niño 3.4	.94[a]	.00	.15	.25	–.15	.97[b]
Niño 4	.26	–.06	–.02	.96[a]	–.06	1.00[b]
New S. Wales	–.10	.82[a]	.11	–.08	–.08	.70[b]
Queensland	.12	.98[a]	.00	.02	.01	.97[b]
Kalimantan	–.07	–.03	–.54[a]	.03	.03	.31
California	.13	.13	.95[a]	.10	.20	1.00[b]
Oregon	.15	–.13	.39[a]	–.13	.38[a]	.35

[a] Numbers in the same column of factors are classified as a group according to relative importance of the factor loadings
[b] Indicates a significant explanation of variances by the factors

view of El Niño dynamics can be explained by WPWP and Australian-Indian monsoon dynamics. The ENSO signal shows clearly in the Indonesia's precipitation in October and November, but diminishes from December to February (Giannini et al. 2007). The January–April precipitation over Indonesia is predominantly controlled by the Australian-Indian monsoon (Haylock and McBride 2001; Aldrian et al. 2003). The 2-month lags include precipitation in the months of March and April. The negative sign of Kalimantan's loading in factor 3 is the reflection of the out-phase locking between ENSO and Australian-Indian monsoon signals.

Oregon's precipitation loading is distributed between factors 3 and 5. This means that the precipitation is not strongly locked with El Niño signal at a 2-month lag. As a matter of fact, with the 2-month time lag the Oregon's precipitation has a negative correlation with Niño 4 SST and has very weak correlations with Niño 3 and Niño 3.4's SSTs. However, it has a relatively strong correlation with Niño 1+2's SST. This may explain why Niño 1+2 and Oregon's loadings appear together in factor 5. Despite this correlation, the tropical SST may not be a good predictor for the Oregon precipitation during the El Niño warm phases when the 2-month time lag is considered.

The communality in Table 3 indicate that the variances of all the SSTs are completely explained by the five factors, and the variances of Queensland and California precipitations are also completely explained by the five factors. However, the variance of Kalimantan's precipitation is only explained to 31%, which corresponds to 35% for Oregon and 70% for New South Wales. These results indicate that the forecasting of Kalimantan and Oregon's monthly precipitation should also consider other predictors besides the SSTs of the Niño regions. Microclimate such as wind, influences Oregon's precipitation in a significant way. Kalimantan's precipitation in March and April is predominantly influenced by Australian-Indian monsoons.

The PCFA this time does not tightly group Niño 4's SST together with other SSTs, compared to the cases of 0-month and 1-month time lags; rather, it distributes Niño 4's SST loadings into two factors: factors 1 and 4. It groups the Kalimantan and California precipitation more tightly than the MLFA.

The 3-month time lag MLFA results are similar to those of the 2-month time lag. The exceptions are that Niño 4's loading (0.96) is now more dominantly in factor 4, California's precipitation loading (0.97) dominates factor 3, and that Kalimantan's precipitation loading is never strong (–0.42) and only appears in factor 3. These results further imply that the tropical Pacific SST is a good predictor for California precipitation during the El Niño warm phases, but that it is not a good predictor for Kalimantan and Oregon precipitation when 2- or 3-month time lags are considered, if MLFA factors are used.

The 3-month time lag PCFA, however, has given quite different loadings for Kalimantan and Oregon precipitations: Kalimantan's loading is 0.97 in factor 4 and Oregon's loading is 0.91 in factor 3. Despite these large loadings, we feel that there is no physical base to support the predictability of the SST to precipitations over the two regions with a 3-month time lag. We thus tend to conclude that the large loadings are due to statistical artifacts rather than physical reality. Our checking with the primitive correlation matrix between the tropical Pacific SST anomalies and the precipitation confirms weak correlation. Therefore, the FA results need to be analyzed in conjunction with physical models and atmospheric circulation patterns. Simple statistical analysis without considering climate dynamics may be misleading.

We also conducted an FA for the non El Niño months: June–September SST and July–October precipitation. The SST and precipitation anomaly data of these 4 months from 1950 to 1999 are used. The SSTs of Niño 1+2 and Niño 3 have high correlations (0.85) because these two regions are

next to each other. The other SST and precipitation correlations are inconsistent and small. Despite these weak correlations, the SSTs still have considerable potential predictability for the precipitations over the investigated regions, since June–September is after the spring-barrier period of March–May. The rotated MLFA groups Niño 4 and Niño 3.4 in factor 1, Niño 1+2 and Niño 3 in factor 2, New South Wales and Queensland in factor 3, and Oregon and California in factor 4. Kalimantan's precipitation is not obviously grouped with any other variables, although it is weakly associated with Niño 4 and Niño 3.4's SST. These imply that the SSTs of these tropical Pacific regions have potential predictability for precipitations over the east coast of Australia and the west coast of United States. Kalimantan's precipitation is not controlled by these SSTs. As a matter of fact, it is predominantly controlled by the WPWP and the Indian Ocean's SST, i.e., the monsoon dynamics.

The un-rotated MLFA groups the factor loadings in the similar way, because the rotation transformation matrix is approximately an identity matrix after a re-arrangement of the column vectors. Namely, each column of the rotation matrix has a dominant element close to 1, and the other elements close to 0.

Both rotated and un-rotated PCFA loadings group all the four SSTs together in factor 1, Australia's east coast precipitations in factor 2, and the United States' west coast precipitations in factor 3. Kalimantan's precipitation does not stand out. The rotation matrix is approximately an identity matrix without rearrangement of columns. From a prediction point of view, the MLFA provides improved potential predictability since it distinguishes between SSTs from the eastern tropical Pacific and the western tropical Pacific.

We have also examined factor scores, which can be obtained for each factor for each month in the data matrix. Factor scores represent the relative importance of the month with respect to their overall factor's interpretation. Larger positive scores indicate a strong correspondence to the factor's interpretation, while larger negative scores indicate a weak correspondence. Our factor scores for month t are calculated by using the weighted least-square method, $\widehat{f}_t = \left(L'\Psi^{-1}L\right)^{-1}L'\Psi^{-1}X_t$, where L is the $N\times M$ factor loading matrix, Ψ is the $N\times N$ diagonal specific variance matrix, X_t the $N\times 1$ anomaly data vector for month t, and the superscript prime indicates the matrix transpose operation.

As discussed earlier, factor 1 represents a strong El Niño signal in the eastern tropical Pacific; the top three highest scores for factor 1 of the zero-lag MLFA analysis are December 1997, November 1997, and January 1998 (2.41, 2.20, and 1.97 respectively). These 3 months correspond to the strong 1997–1998 El Niño event. The 1982–1983 El Niño also corresponds to higher scores of factor 1.

4 Conclusions and discussion

Factor analysis is an effective statistical procedure that provides a way to group variables into factors, and hence can represent the many original variables with a few factors. We have introduced the MLFA to analyze the El Niño signals in precipitation. The El Niño signals are signified by the SSTs in the four Niño regions over the Tropical Pacific: Niño 1+2, Niño 3, Niño 3.4, and Niño 4. The precipitation data are analyzed for the east coast of Australia, Kalimantan island of Indonesia, and the west coast of the United States. Our MLFA and PCFA are performed by SAS programs. We have shown that FA is indeed an effective tool for climate data analysis, in particular for the analysis of covariance or correlation matrices. FA can subtly group meteorologically connected variables together and help raise the potential predictability of predictand variables. Our analysis of the SST and precipitation data concludes the following: for El Niño signals, the rotated MLFA factor loadings group the SST and precipitation variables with interpretable physical meanings. When the time lag is 0 or 1 month, the November–February El Niño explains much of the precipitation signals over east Australia and Kalimantan, but when the time lag is 2 or 3 months, the same SST cannot adequately explain the precipitation over the two regions, because Australia-Indian monsoon dynamics plays dominant role for this season. The tropical Pacific's SSTs always have high potential predictability for the precipitation over the west coast of the United States. PCFA is also investigated, and its results are compared with MLFA. The comparison indicates that MLFA can better group SST predictors for precipitation forecasting. The residuals of MLFA are always smaller than the PCFA. Thus, MLFA may have the potential to become a useful tool for raising the forecasting skill of precipitation from SST predictors.

MLFA has application limitations. One of them is that the MLFA factors are not necessarily orthogonal. This gives the MLFA flexibility and makes it very powerful in diagnostic analysis, but also limits its applications when orthogonality is needed for mathematics operations. Of course, PCFA factors are always orthogonal as long as the rotation is rigid, although an oblique rotation can also be made for factors obtained by any FA calculation method (Johnson and Wichern 1992). The factors shown in Tables 2 and 3 are not orthogonal.

Acknowledgements This study was supported in part by the San Diego State University's start-up research fund for Shen. The authors thank Jay Lawrimore of the US National Climatic Data Center for extracting the GHCN data, Robert Field of the University of Toronto for comments related to Indonesia precipitation. GRN wishes to acknowledge support from CGM-NSF 480921–01001.

References

Aldrian E, Susanto RD (2003) Identification of three dominant rainfall regions within Indonesia and their relationship to sea surface temperature. Int J Climatol 23:1435–1452

Barnett TP, Preisendorfer R (1987) Origins and levels of monthly and seasonal forecast skill for United States surface air temperatures determined by canonical correlation analysis. Mon Weather Rev 115:1825–1850

Barnston AG, Smith TM (1996) Specification and prediction of global surface temperature and precipitation from global SST using CCA. J Clim 9:2660–2697

Bartzokas A, Metaxas DA (1995) Factor analysis of some climatological elements in Athens, 1931–1992: covariability and climatic change. Theor Appl Climatol 52:195–205

Climate Prediction Center. Monthly atmospheric and SST Indices. http://www.cpc.noaa.gov/data/indices/.

Dinpashoh Y, Fakheri-Fard A, Moghaddam M, Jahanbakhshc S, Mirnia M (2004) Selection of variables for the purpose of regionalization of Iran's precipitation climate using multivariate methods. J Hydrol 297:109–123

Giannini A, Robertson AW, Qian JH (2007) A role for tropical tropospheric temperature adjustment to ENSO in the seasonality of monsoonal Indonesia precipitation predictability. J Geophys Res (Atmosphere) 112, D16110. doi:10.1029/2007JD008519

Gorsuch RL (1983) Factor analysis, 2nd edn. Erlbaum, Philadelphia, PA, 452 pp

Haylock M, McBride J (2001) Spatial coherence and predictability of Indonesian wet season rainfall. J Clim 14:3882–3887

Johnson RA, Wichern DW (1992) Applied multivariate statistical analysis, 3rd edn. Prentice Hall, Englewood Cliffs, NJ, 642 pp

Lau WKM, Kim KM, Shen SSP (2002) Potential predictability of seasonal precipitation over the United States from canonical ensemble correlation predictions. Geophys Res Lett 29:10: 1029

National Drought Mitigation Center (2007) Reported effects of the 1997 El Niño through October 30: an NDMC analysis of media reports. University of Nebraska, Lincoln, NE. http://www.drought.unl.edu/risk/world/table2.pdf. Cited 29 August 2008

National Oceanic and Atmospheric Administration (1986) Climates of the states: narrative summaries, tables, and maps for each state with overview of state climatologist programs, 3rd edn, vol 2. NOAA, Washington, DC

National Weather Service Forecast Office (2004) El Niño impact map. National Weather Service Forecast Office, Washington, DC. http://www.cpc.ncep.noaa.gov/products/analysis_monitoring/impacts/warm.gif. Cited 29 August 2008

North GR, Bell TL, Cahalan RF, Moeng FJ (1982) Sampling errors in the estimation of empirical orthogonal functions. Mon Weather Rev 110:699–706

North GR, Kim KY, Shen SSP, Hardin JW (1995) Detection of forced climate signals, part I: theory. J Climate 8:401–408

Null J (2004) El Niño and La Niña Years: a consensus list. Golden Gate Weather Sevices, Saratoga, CA. http://ggweather.com/enso/years.htm. Cited 27 August 2008

Shen SSP, Lau KM, Kim KM, Li G (2001) A canonical ensemble correlation prediction model for seasonal precipitation anomaly. Technical Memorandum NASA-TM-2001–209989, NASA, Washington, DC, 53 pp

Smith TM, Livezey RE, Shen SSP (1998) An improved method for interpolating sparse and irregularly distributed data onto a regular grid. J Clim 11:1717–1729

Walters W (1999) Climate indices for use in social and behavior research. IASSIST Q 23(1):1–14

Wilks DS (2006) Statistical methods in the atmospheric sciences, 2nd edn. Elsevier, New York, 627 pp