Characterizing the Spatiotemporal Variation of Environmental Data

I. Principal Component Analysis
II. Time Series Analysis

Brian K. Eder

Air Resources Laboratory
National Oceanic and Atmospheric Administration

National Exposure Research Laboratory
Environmental Protection Agency
RTP, NC 27711

Principal Component Analysis

Objective

Identify, through data reduction, the characteristic, recurring and independent modes of variation (signals) of a large, noisy data set.

Approach

Sorts, initially correlated data, into a hierarchy of statistically independent modes of variation (mutually orthogonal linear combinations), which explain successively less and less of the total variation.

Utility

Facilitates identification, characterization and understanding of the spatiotemporal variation of the data set across a myriad of spatial and temporal scales.

Numerous applications

- The spatial and temporal analysis of the **Palmer Drought Severity Index** over the Southeastern US. (*J. of Climatology* - 7, pp 31-56)
- A principal component analysis of **$SO_4^=$ precipitation concentrations** over the eastern US. (*Atmospheric Environment* 23, No. 12, pp 2739-2750)
- A characterization of the spatiotemporal variation of **non-urban ozone** in the Eastern US. (*Atmospheric Environment*, 27A, pp. 2645-2668)
- A climatology of **total ozone mapping spectrometer** data using rotated principal component analysis. (*Journal of Geophysical Research*. 104, No. D3, pp 3691-3709)
- A climatology of **air concentration** data from the Clean Air Status and Trends Network (CASTNet). (*Atmospheric Environment*)

Methodology

Spatial

Calculate a square, symmetrical *correlation* matrix **R** having dimensions j x j, from the original data matrix having dimensions j (e.g. stations, grid cells) x i (e.g. days, weeks).

By using **R** and the Identity matrix **I**, of the same dimensions, j characteristics roots or eigenvalues (λ) can be derived that satisfy the following polynomial equation:

$$\det [{}_j\mathbf{R}_j - \lambda_j \mathbf{I}_j] = 0 \qquad (1)$$

Methodology

Spatial

For each root **l** of (1) which is called the characteristic equation, a nonzero vector **e** can be derived such that:

$$ {}_{j}\mathbf{R}_{j}\ \mathbf{e}_{1} = \lambda_{j}\ \mathbf{e}_{1} \qquad (2) $$

where **e** is the characteristic vector (eigenvector) of the correlation matrix **R,** associated with its corresponding eigenvalue **l**.

- The eigenvectors represent the mutually orthogonal linear combinations (modes of variation) of the matrix.

- The eigenvalues represent the amount of variation explained by each of the eigenvectors.

Spatial

When the elements of each eigenvector (**e**) are multiplied by the square root of the associated eigenvalue ($\boldsymbol{\lambda}^{0.5}$),

the principal component (pc) **Loading** (L) is obtained.

L : provides the correlation between the pc and the station (grid cell)

L^2: provides the proportion of variance at an individual station (grid cell) that can be attributed to a particular pc

The sum of the squared Loadings indicates the total variance accounted for by the pc, which stated earlier is called the eigenvalue

$$\lambda_k = \sum_j L_{kj}^2 \qquad (3)$$

For station j and pc k

The pc Loadings can then be spatially mapped onto their respective stations (grid cells) identifying homogeneity or "*influence regimes*".

Methodology

Spatial

By retaining the first few eigenvector-eigenvalue pairs or principal components, a substantial amount of the variation can be explained while ignoring higher-order pcs, which explain successively less of the variance.

How many principal components should be retained??

- Scree test
- $\lambda > 1$ criteria
- Overland-Priesendorfer “Rule N” test
- Common Sense

Methodology

Spatial

Rotation of Retained Principal Components

Facilitates spatial interpretation allowing better identification of areas that are homogeneous

Oblique Rotation
Orthogonal Rotation

An orthogonal rotation developed by Kaiser ('58) increases the segregation between principal component loadings which in turn better defines a distinct group or cluster of homogeneous stations.

Stations (grid cells) are then assigned to the pc ("*influence regime*") having the largest pc loading.

Methodology

Temporal

Having identified *influence regimes*, we can examine their temporal structure thru calculation of the pc **Score**

The pc **Score** for time period i on principal component k are weighted, summed values whose magnitudes depend upon the observation O_{ij} for time i at station j and L_{jk} is the loading of station j on component k as seen below:

$$(PCscore)_{ik} = \sum_{j} O_{ij} L_{jk} \qquad (4)$$

The pc **Scores** are standardized (mean: 0, std dev: 1)

Methodology

Temporal

When plotted as a time series, the pc **Scores** provide excellent insight into the spectrum of temporal variance experienced by each of the influence regimes.

This temporal variance can then be examined using:

Spectral Density Analysis

Correlograms

Filters

Red and White Noise tests

JOURNAL OF GEOPHYSICAL RESEARCH, VOL. 104, NO. D3, PAGES 3691–3709, FEBRUARY 20, 1999

A climatology of total ozone mapping spectrometer data using rotated principal component analysis

Brian K. Eder[1] and Sharon K. LeDuc[1]

Atmospheric Sciences Modeling Division, NOAA Air Resources Laboratory, Research Triangle Park North Carolina

Joseph E. Sickles II

Environmental Sciences Division, National Exposure Research Laboratory, U.S. Environmental Protection Agency, Research Triangle Park, North Carolina

Abstract. The spatial and temporal variability of total column ozone (Ω) obtained from the total ozone mapping spectrometer (TOMS version 7.0) during the period 1980–1992 was examined through the use of a multivariate statistical technique called rotated principal component analysis. Utilization of Kaiser's varimax orthogonal rotation led to the identification of 14, mostly contiguous subregions that together accounted for more than 70% of the total Ω variance. Each subregion displayed statistically unique Ω characteristics that were further examined through time series and spectral density analyses, revealing significant periodicities on semiannual, annual, quasi-biennial, and longer term time frames. This analysis facilitated identification of the probable mechanisms responsible for the variability of Ω within the 14 homogeneous subregions. The mechanisms were either dynamical in nature (i.e., advection associated with baroclinic waves, the quasi-biennial oscillation, or El Niño–Southern Oscillation) or photochemical in nature (i.e., production of odd oxygen (O or O_3) associated with the annual progression of the Sun). The analysis has also revealed that the influence of a data retrieval artifact, found in equatorial latitudes of version 6.0 of the TOMS data, has been reduced in version 7.0.

"Dynamical Forcing"

Related to transport and tropopause height.

Sharp late winter, early spring peak.

More broad, late summer, early autumn minimum.

Strong **annual** signal (Periodicity = 2B/*f*)

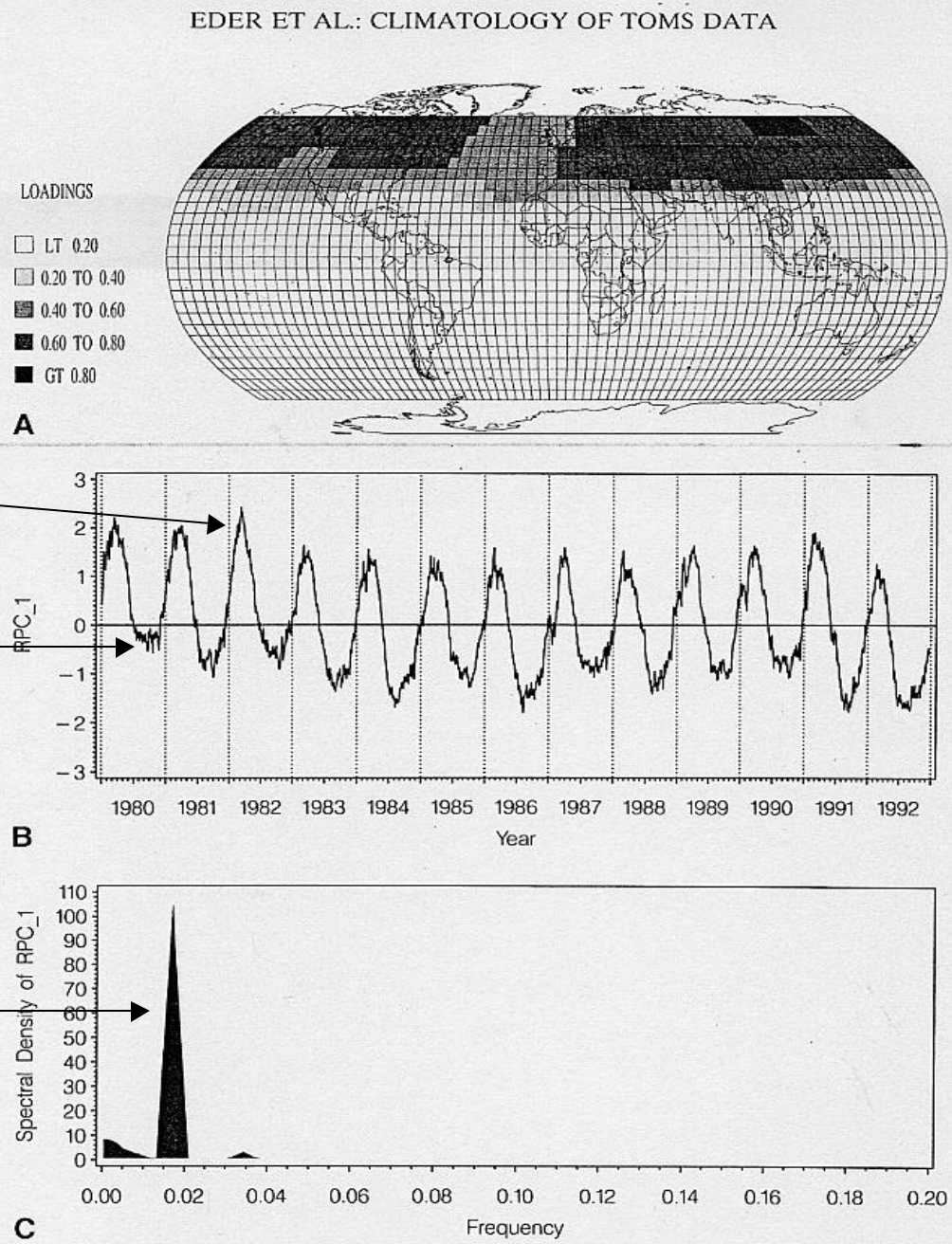


Figure 2. (a) Principal component loadings associated with RPC_1, which accounted for 31.9% of the total variance; (b) standardized principal component scores associated with RPC_1; and (c) spectral density analysis of the principal component scores associated with RPC_1.

‘Photochemical Forcing”

Related to solar insolation.

Broad, mid-summer maximum.

Sharp, mid-winter minimum.

Strong **annual** signal

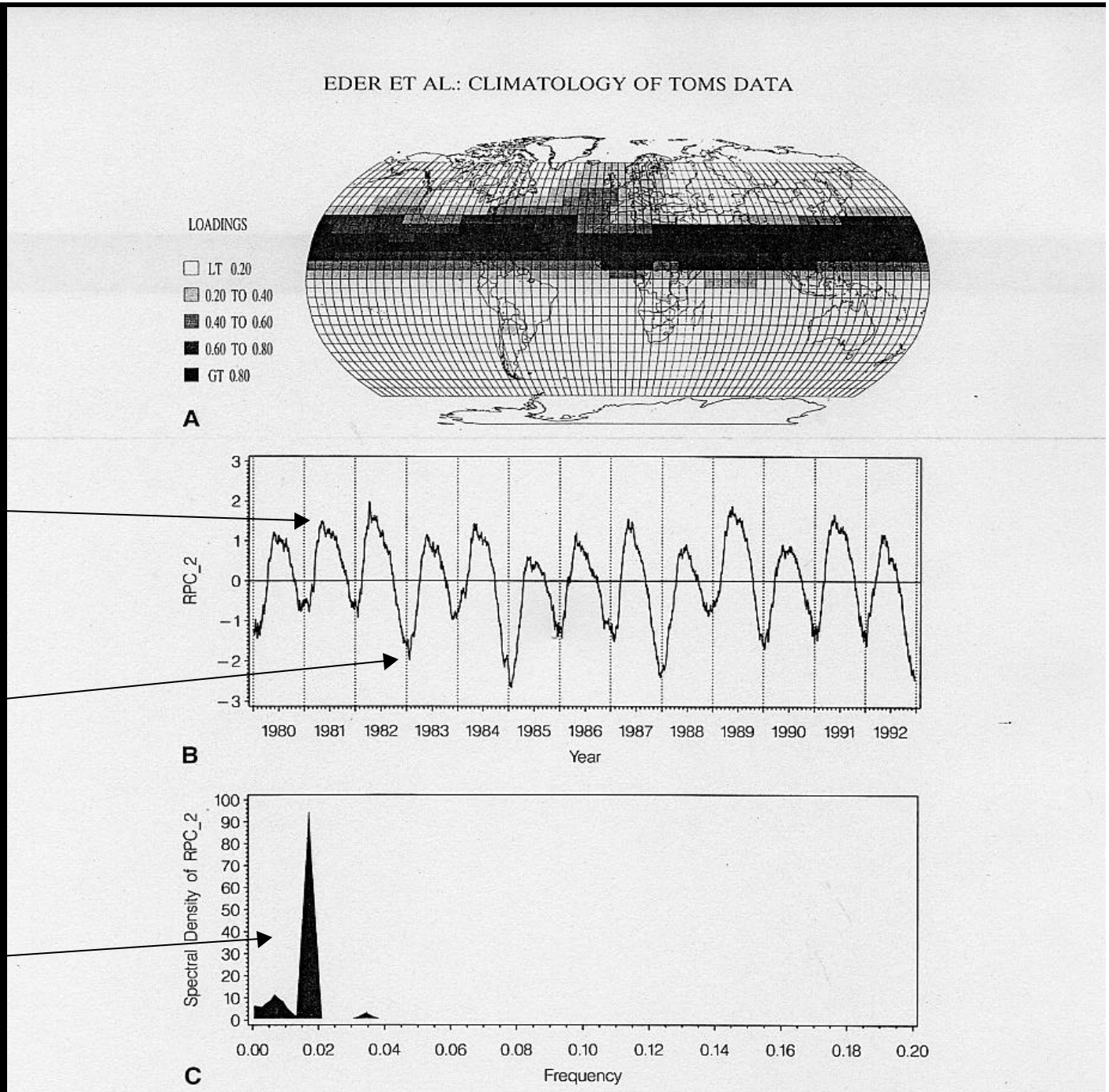


Figure 3. (a) Principal component loadings associated with RPC_2, which accounted for 17.7% of the total variance; (b) standardized principal component scores associated with RPC_2; and (c) spectral density analysis of the principal component scores associated with RPC_2.

“Dynamical Forcing”

Related to annual transport and SAO in wind field (peaks at equinoxes)

Strong **annual, semi-annual** and a **long term** signal.

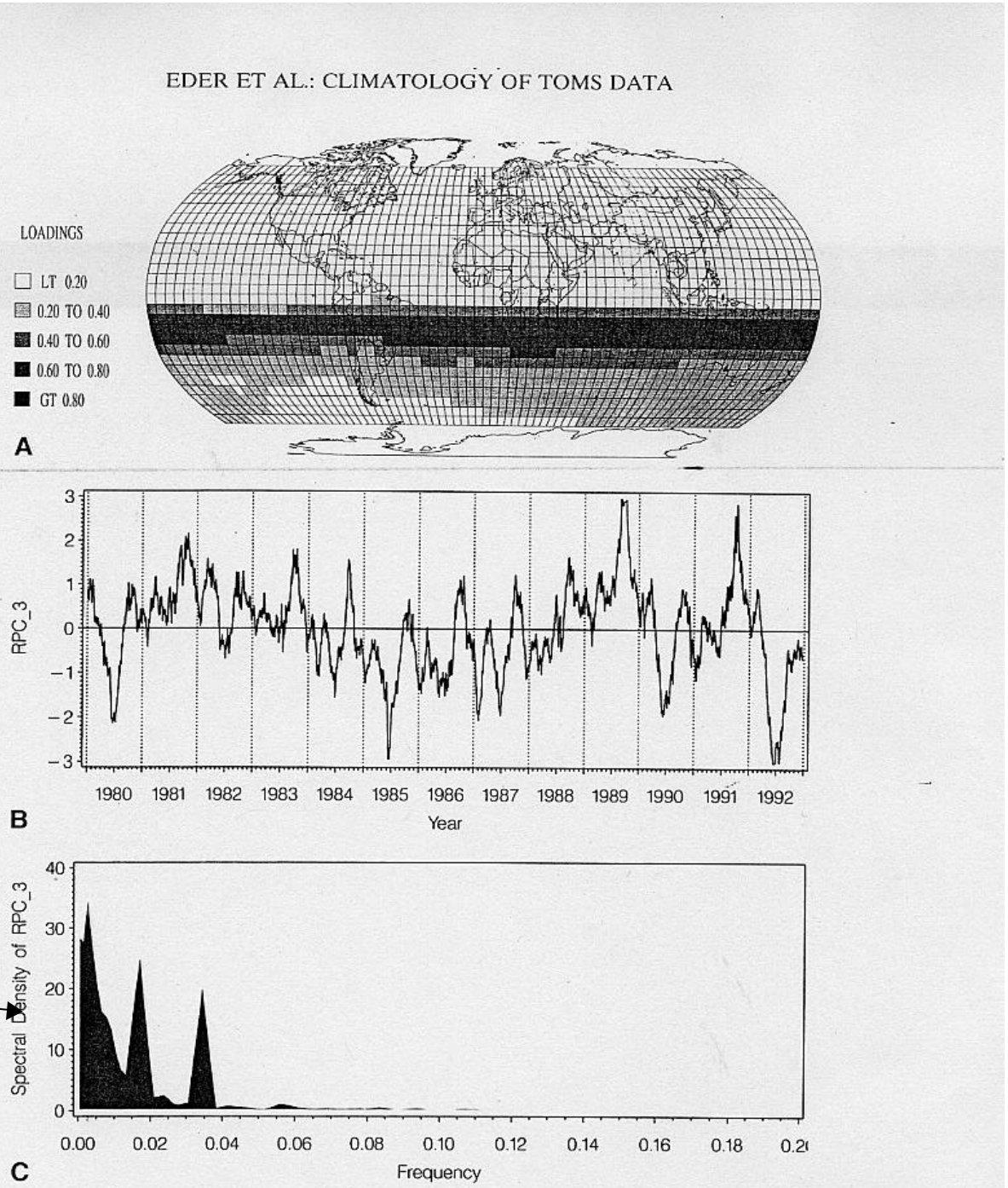


Figure 4. (a) Principal component loadings associated with RPC_3, which accounted for 5.8% of the total variance; (b) standardized principal component scores associated with RPC_3; and (c) spectral density analysis of the principal component scores associated with RPC_3.

‘Quasi-Biennial Forcing”

Related to QBO of tropical winds in the stratosphere.

Note peaks in ‘80, ‘82, 85, ‘87, ‘90 and ’92.

Strong QBO signal (~2.5 years)

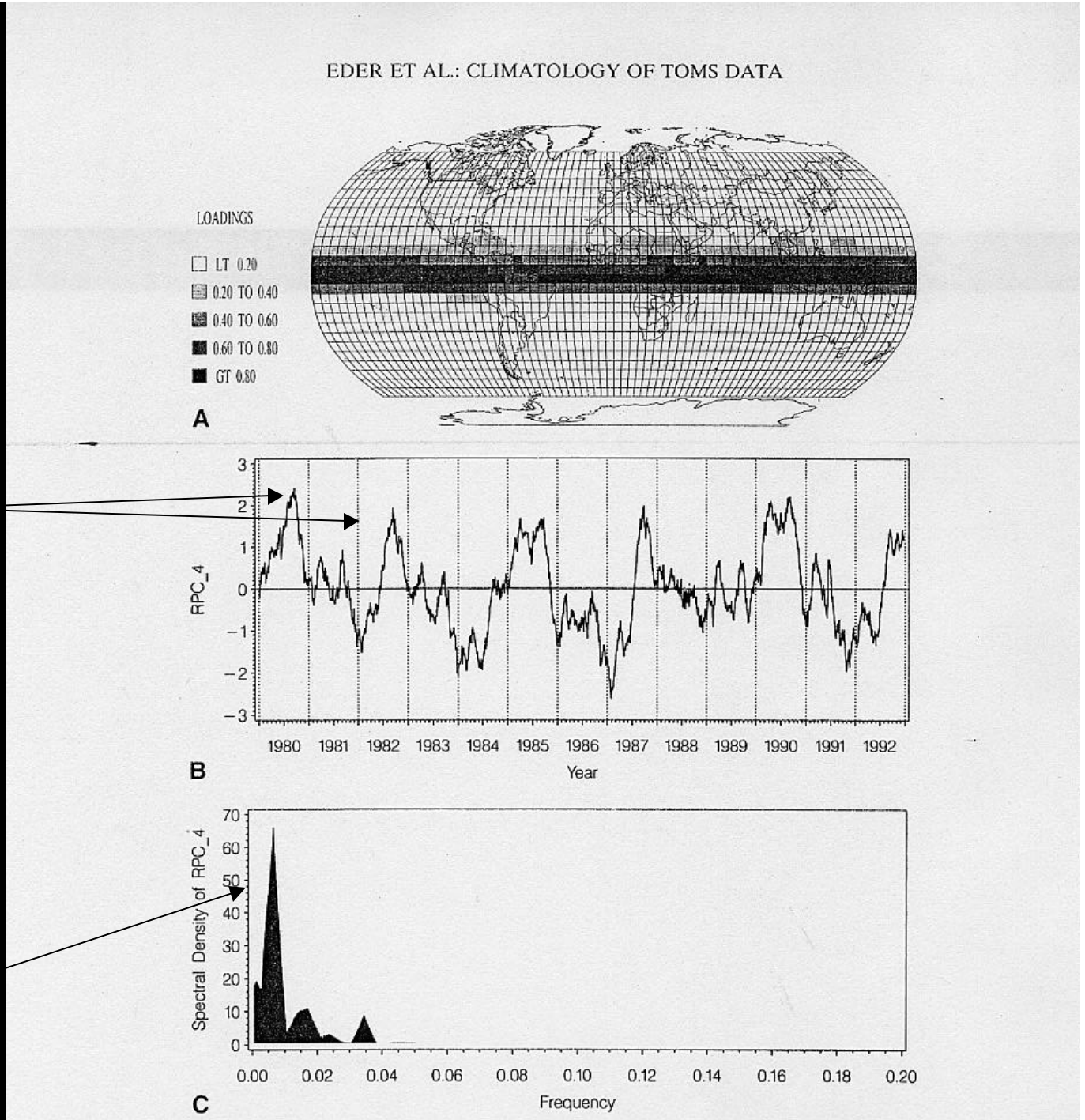


Figure 5. (a) Principal component loadings associated with RPC_4, which accounted for 4.2% of the total variance; (b) standardized principal component scores associated with RPC_4; and (c) spectral density analysis of the principal component scores associated with RPC_4.

"Wave Number 5"

One of 5 similar patterns found between 45^0 - 65^0 S.

Due to medium scale baroclinic waves associated with Antarctic Polar Jet stream.

Note variability.

Note **trend** and **semi-annual** periodicity.

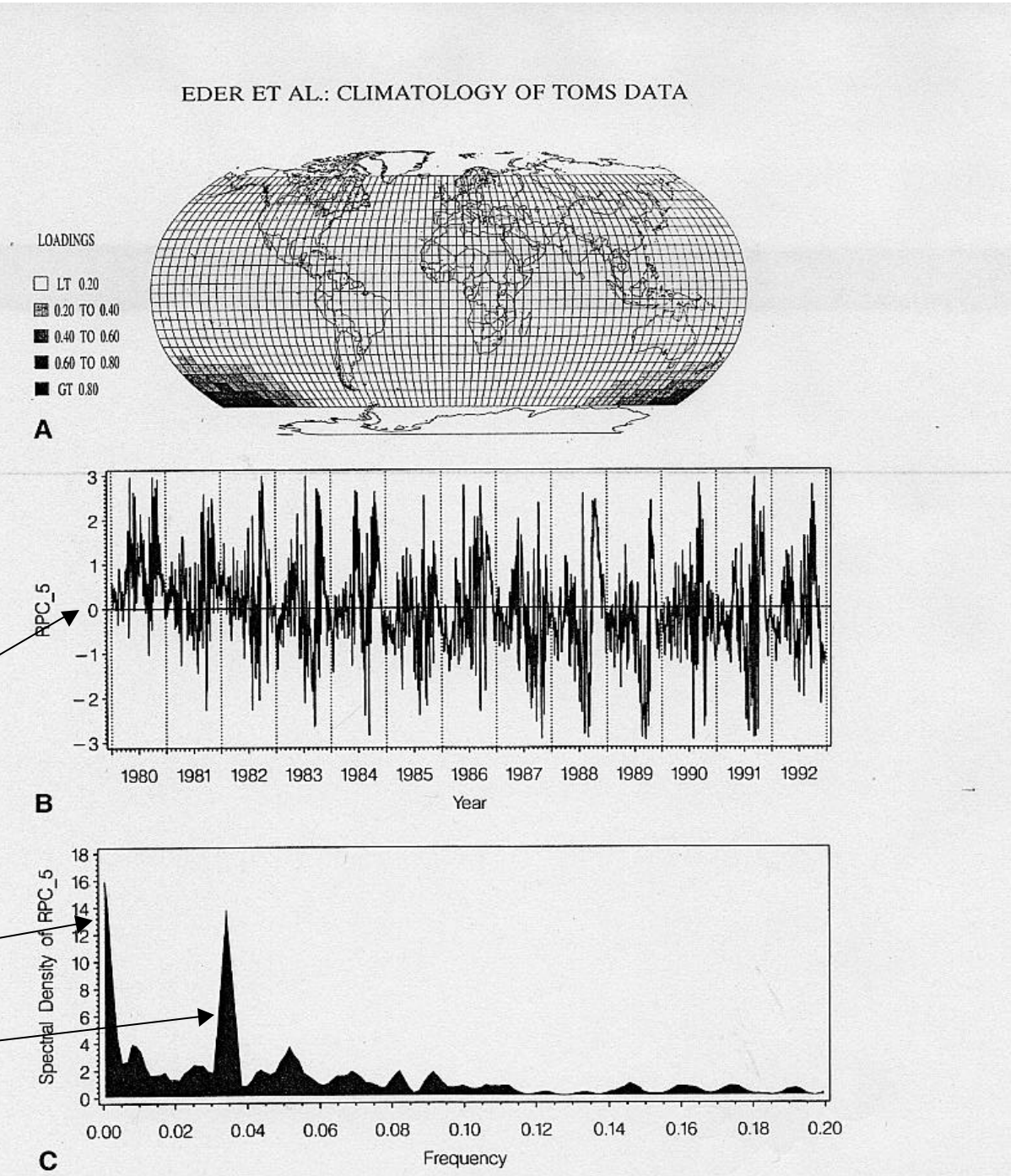


Figure 6. (a) Principal component loadings associated with RPC_5, which accounted for 2.0% of the total variance; (b) standardized principal component scores associated with RPC_5; and (c) spectral density analysis of the principal component scores associated with RPC_5.

‘El-Nino-Southern Oscillation”

During **ENSO** years of 82-83, ’87 and ’91-’92, ozone values are very low, while in none ENSO years ozone values are high.

Note strong periodicity of ~ 4 years.

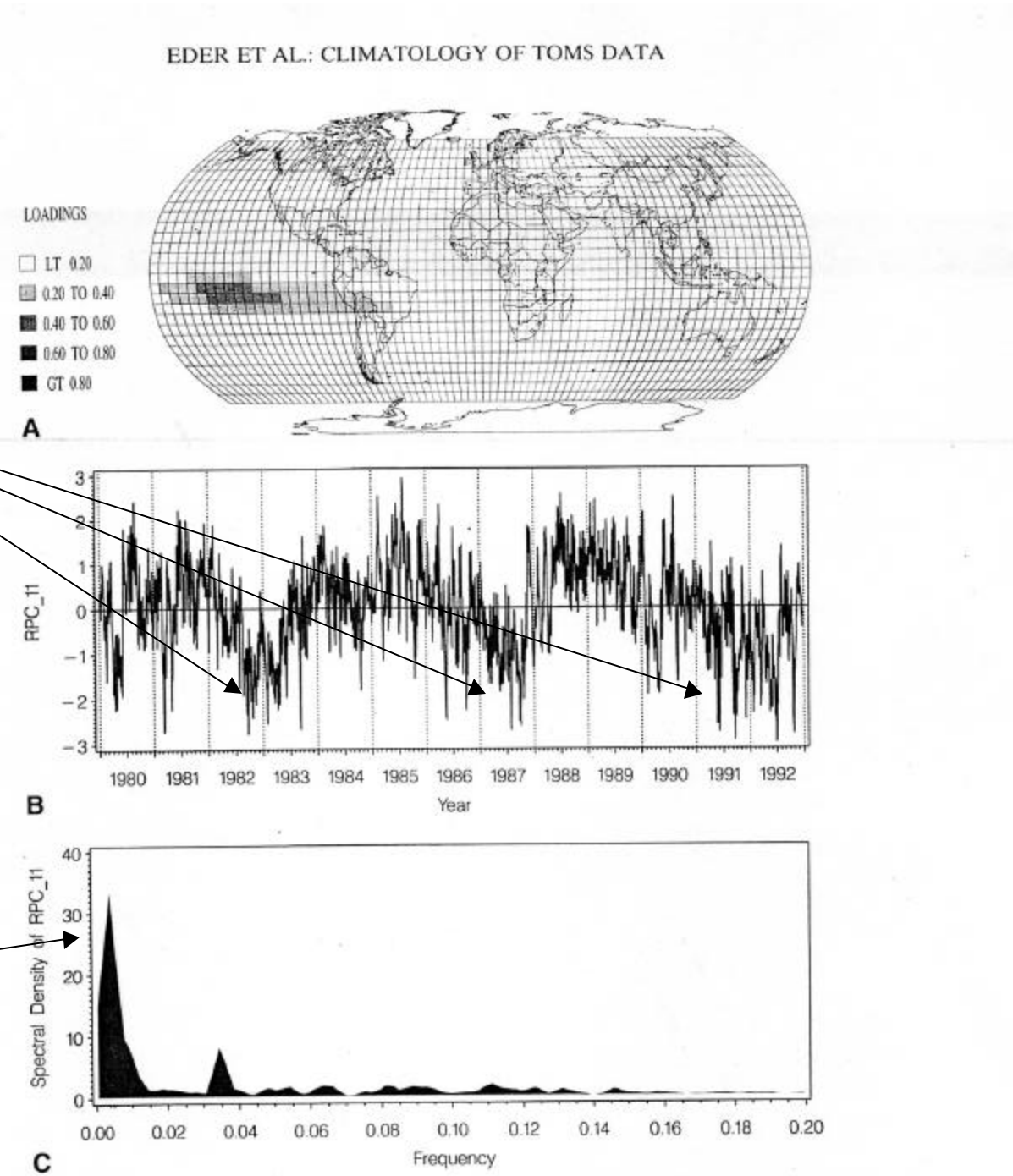


Figure 12. (a) Principal component loadings associated with RPC_11, which accounted for 0.9% of the total variance; (b) standardized principal component scores associated with RPC_11; and (c) spectral density analysis of the principal component scores associated with RPC_11.

Data Retrieval Artifact"

An earlier analysis of TOMS Version 6.0 included a "**cross-track" bias** related to successive orbital scans of the surface.

Note the tremendous "**pulse**" in the spectral plot.

NASA was unaware of this artifact.

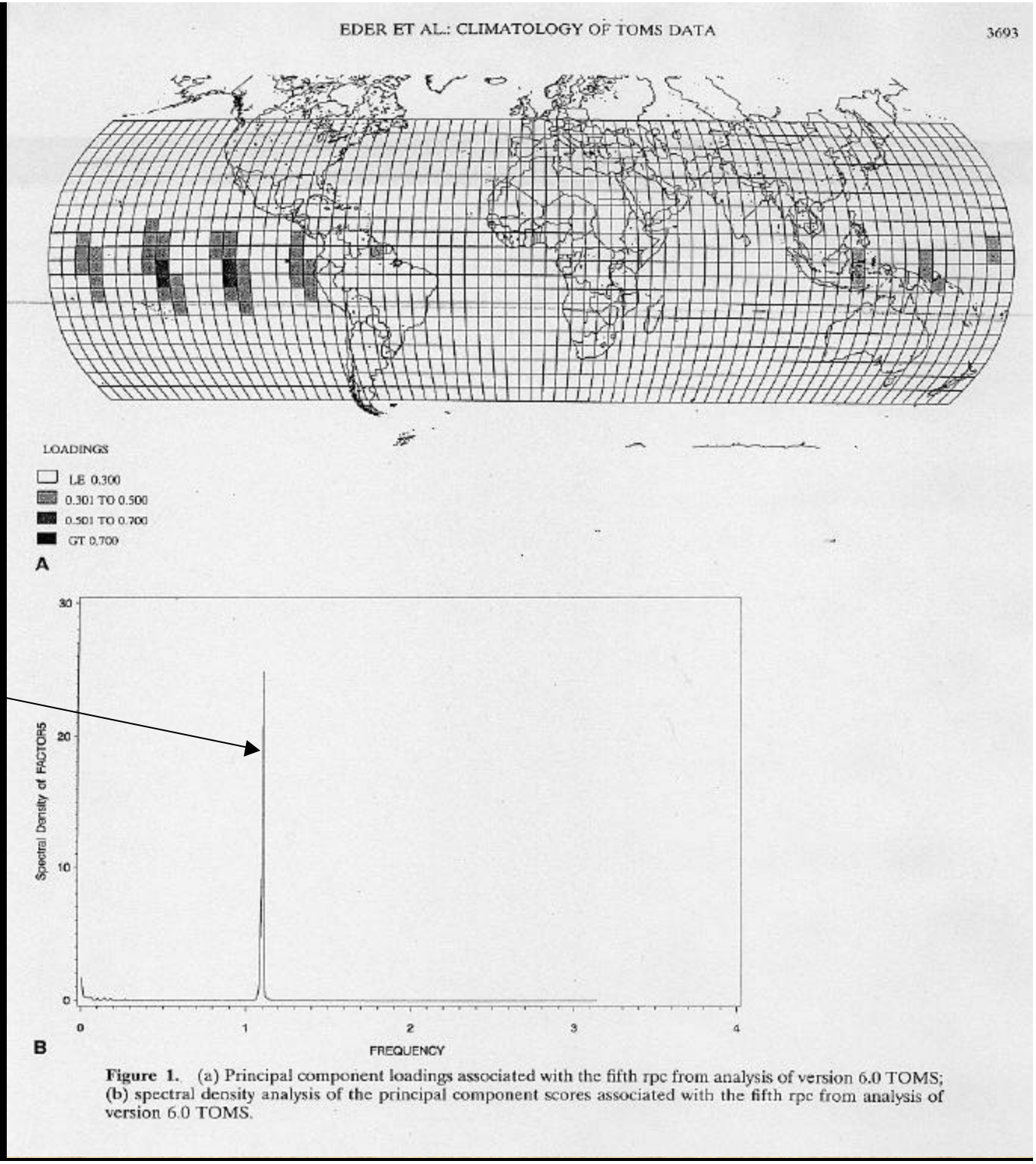


Figure 1. (a) Principal component loadings associated with the fifth rpc from analysis of version 6.0 TOMS; (b) spectral density analysis of the principal component scores associated with the fifth rpc from analysis of version 6.0 TOMS.

Printed in Great Britain.

Pergamon Press Ltd

A CHARACTERIZATION OF THE SPATIOTEMPORAL VARIABILITY OF NON-URBAN OZONE CONCENTRATIONS OVER THE EASTERN UNITED STATES

BRIAN K. EDER*

Atmospheric Sciences Modeling Division, Air Resources Laboratory, National Oceanic and Atmospheric Administration, Research Triangle Park, NC 27711, U.S.A.

JERRY M. DAVIS

Department of Marine, Earth and Atmospheric Sciences, North Carolina State University, Raleigh, NC 27695, U.S.A.

and

PETER BLOOMFIELD

Department of Statistics, North Carolina State University, Raleigh, NC 27695, U.S.A.

(*First received* 6 *October* 1992 *and in final form* 12 *April* 1993)

Abstract—The spatial and temporal variability of the daily 1-h maximum O_3 concentrations over non-urban areas of the eastern United States of America was examined for the period 1985–1990 using principal component analysis. Utilization of Kaiser's Varimax orthogonal rotation led to the delineation of six contiguous subregions or "influence regimes" which together accounted for 64.02% of the total variance. Each subregion displayed statistically unique O_3 characteristics and corresponded well with the path and frequency of anticyclones. When compared to the entire domain, the mid-Atlantic and south subregions observe higher mean daily 1-h maximum concentrations. Concentrations are near the domain average for the northeast and southwest subregions and are lowest in the Great Lakes and Florida subregions. The percentage of observations exceeding 120 ppb were greatest in the mid-Atlantic and southwest subregions, near the domain average in the northeast and south subregions, and lowest in the Great Lakes and Florida subregions.

Examination of the time series of the principal component scores associated with the subregions indicated that Great Lakes and mid-Atlantic subregions tend to observe a stronger seasonal cycle, with maximum concentrations occurring during the last week in June and first week in July, respectively. The strength of this seasonality is weakened for the northeast and south subregions and its timing delayed, until the end of July and the first of August, respectively. The southwest subregion experiences a greatly diminished seasonality, with maximum concentrations delayed until the middle of August. The seasonality found in the Florida subregion is unique in both its strength and timing, as the highest concentrations consistently occur during the months of April and May. The time series were then deseasonalized and autocorrelations and spectral density estimates calculated, revealing that persistence is much more prevalent in the Florida (autocorrelation significant to a lag of 4 days), south (3 days) and southwest (3 days) subregions. Conversely, autocorrelations are only significant to a lag of one day in the northeast and two days for the Great Lakes and mid-Atlantic subregions.

Key word index: Ozone, principal component analysis, influence regimes, time series analysis, persistence, seasonality, anticyclones.

Six Homogenous Regions

Great Lakes
Northeast
Mid-Atlantic
Southwest
South
Florida

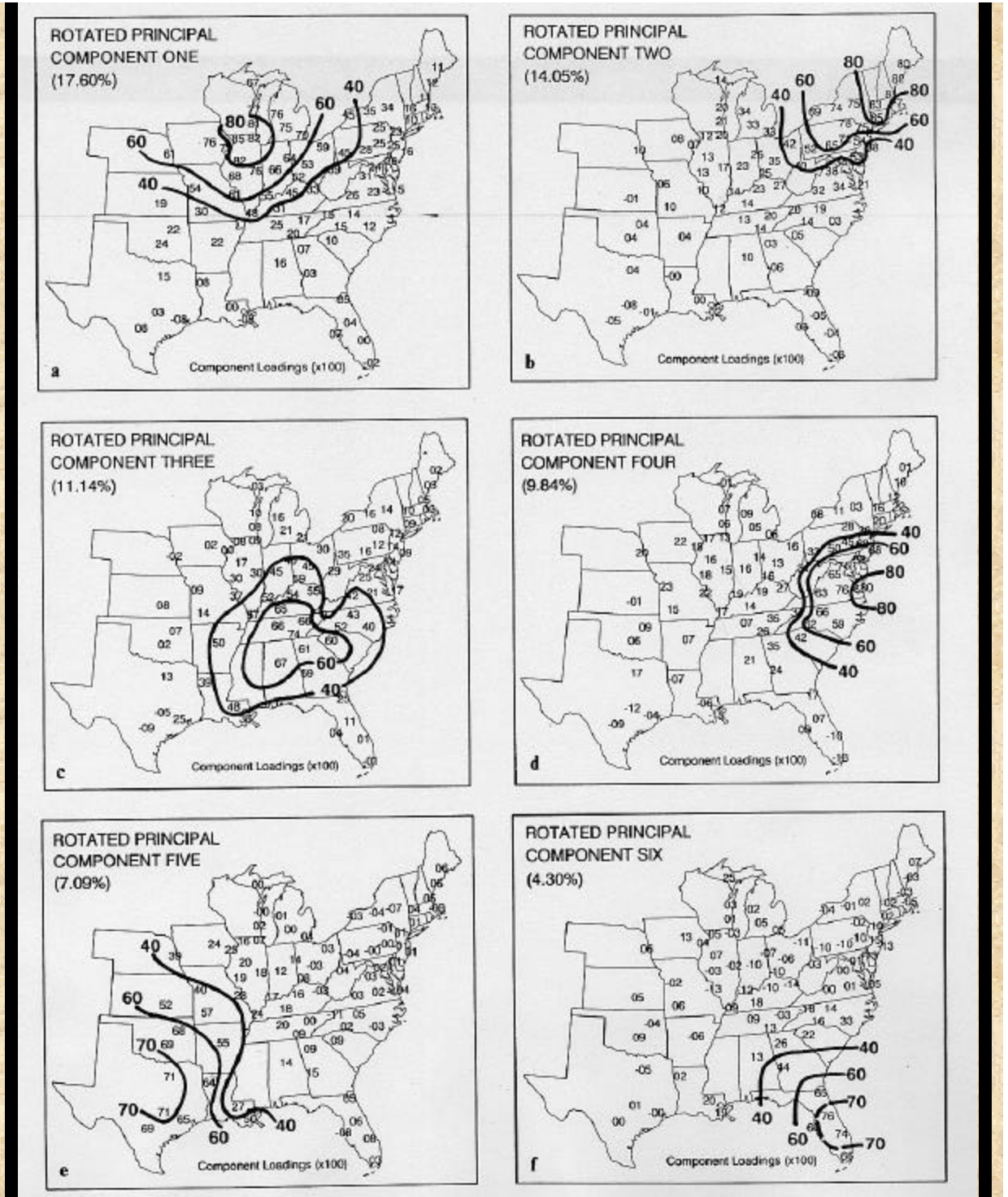


Fig. 3. Component loadings associated with the six rotated principal components.

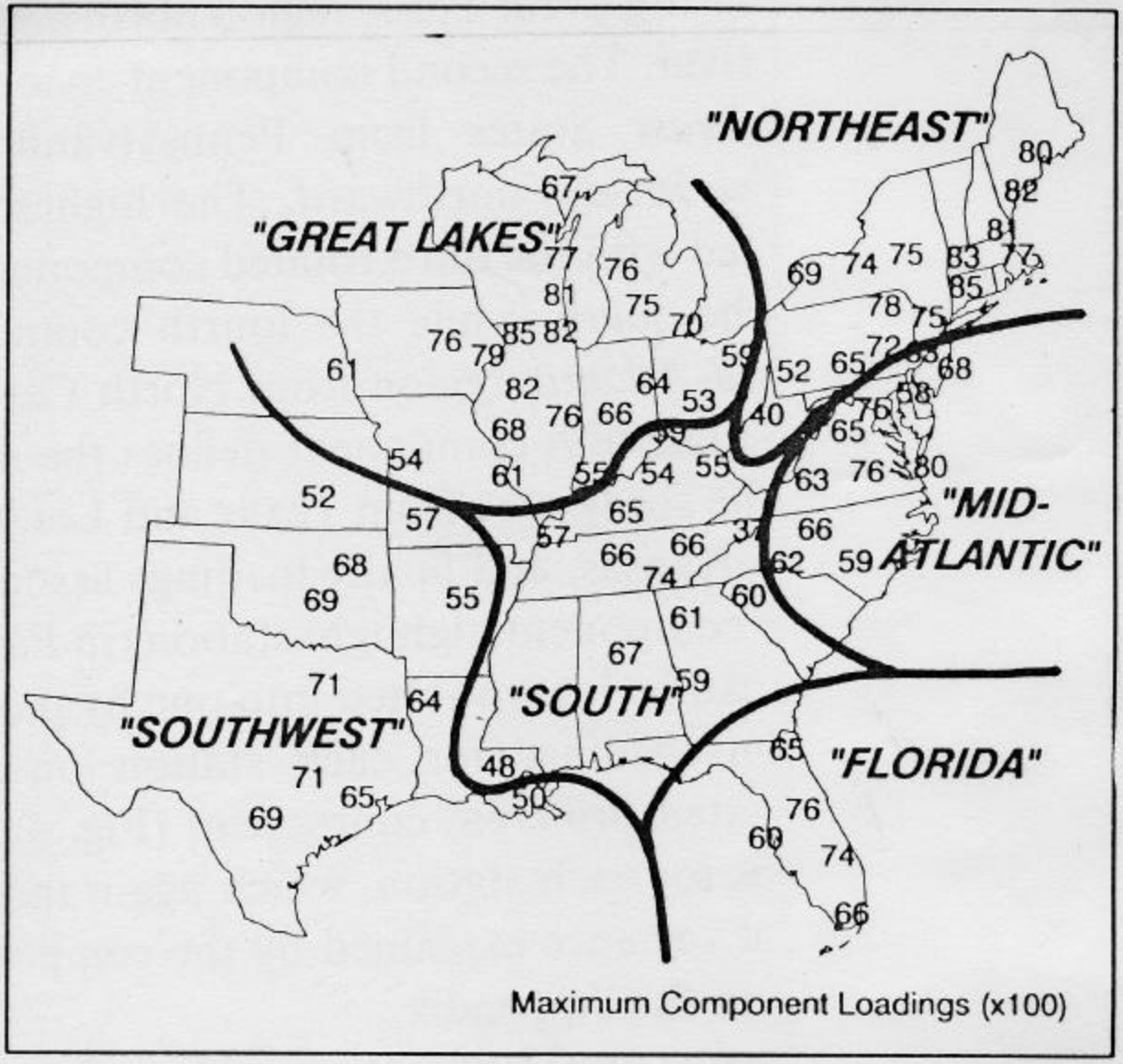


Fig. 4. Six homogeneous O_3 concentration regions as depicted by the maximum component loadings.

Daily Time Series

Standardized PC scores

Summer Peak

No pronounced Peak

Spring Peak

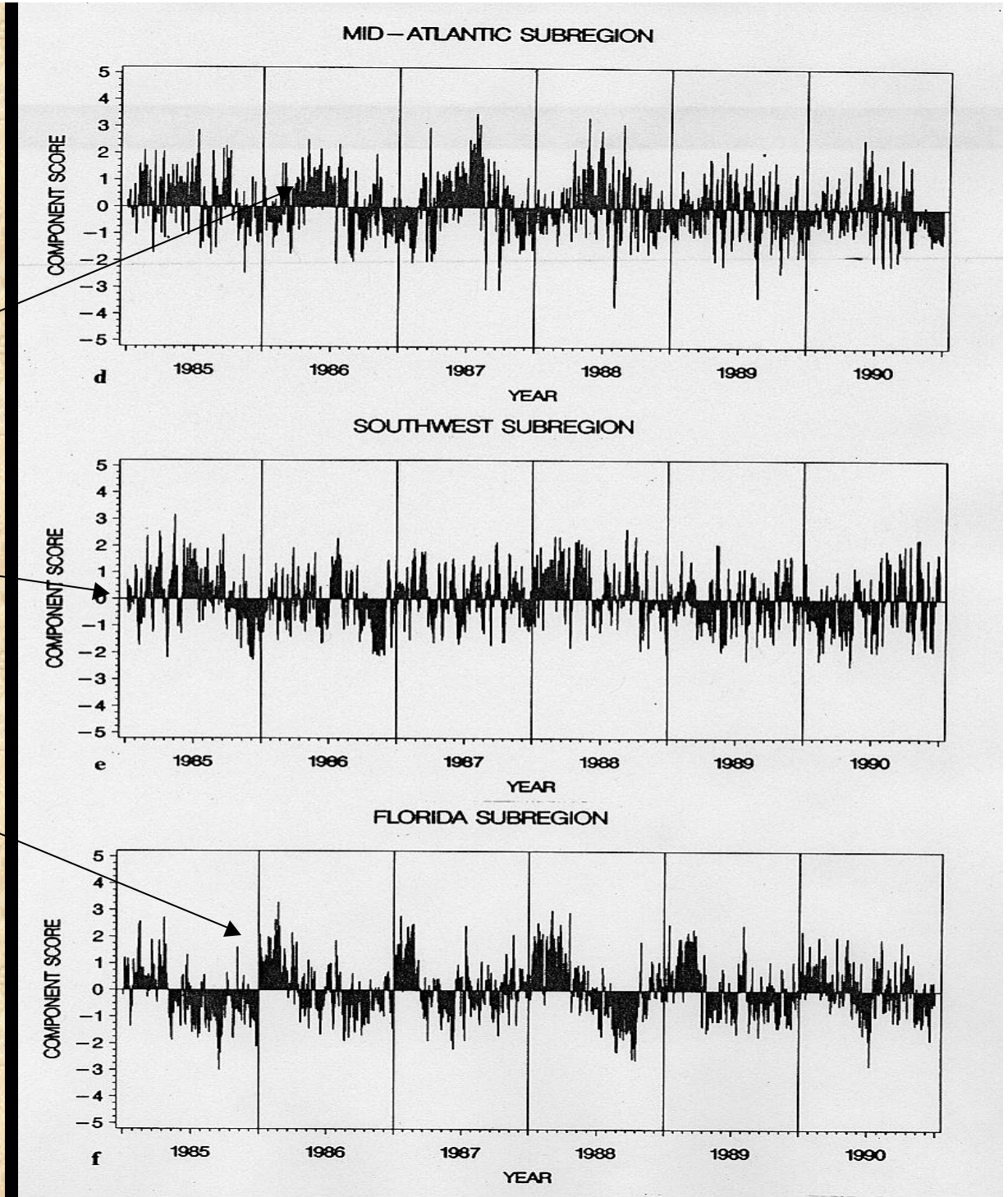


Fig. 5. Daily time series of the standardized principal component scores associated with the six homogeneous subregions.

Daily Time Series

Standardized PC scores

Early summer peak

Late summer peaks

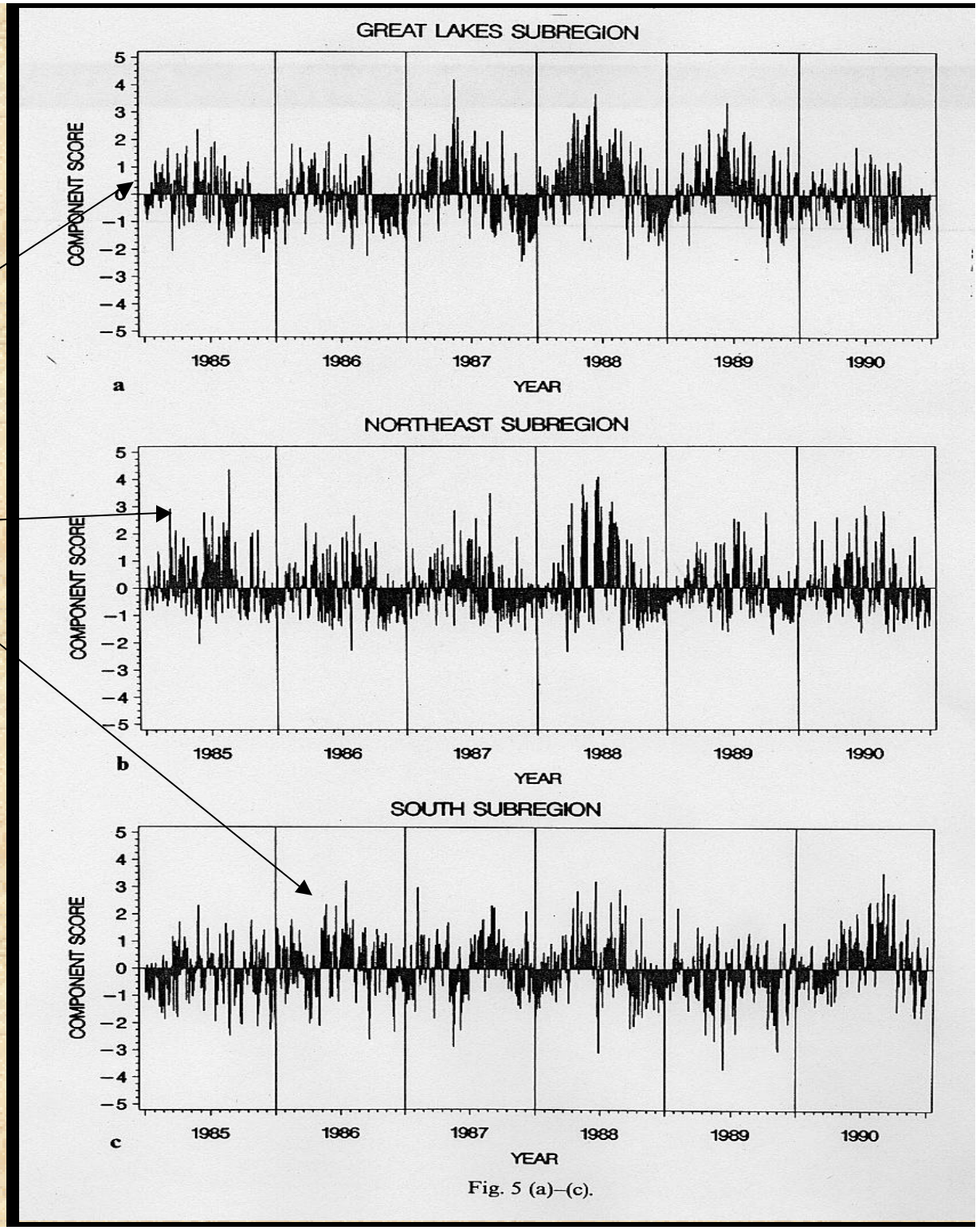


Fig. 5 (a)–(c).

Seasonal Time Series

Standardized PC scores

Medians over 6 years
Cubic Spline Smoother

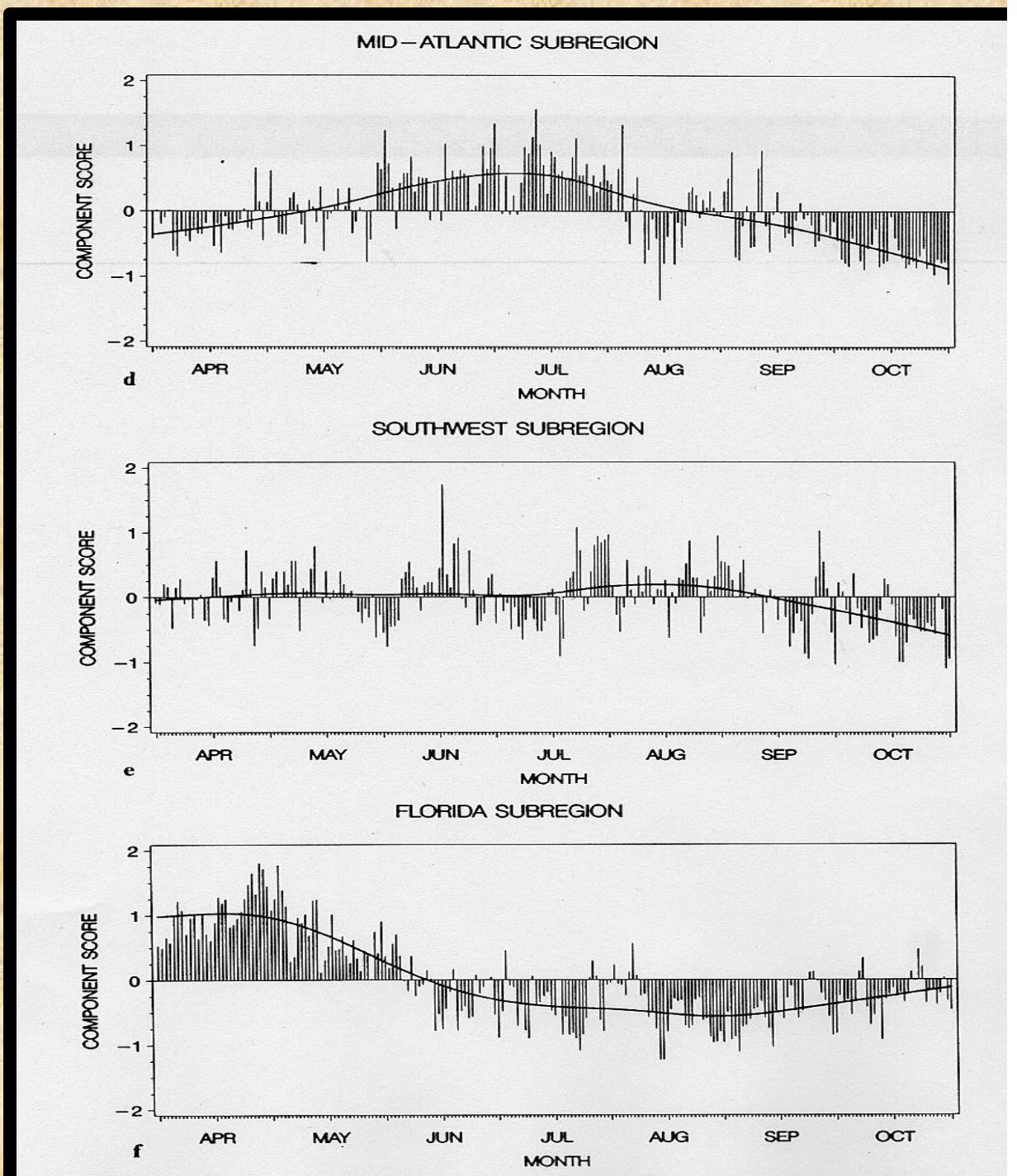


Fig. 6. Seasonal time series of the standardized principal component scores associated with the six homogeneous subregions as defined by the median scores for the six year period. A cubic spline function was used to smooth the data (thick line).

Seasonal Time Series

Standardized PC scores

Medians over 6 years
Cubic Spline Smoother

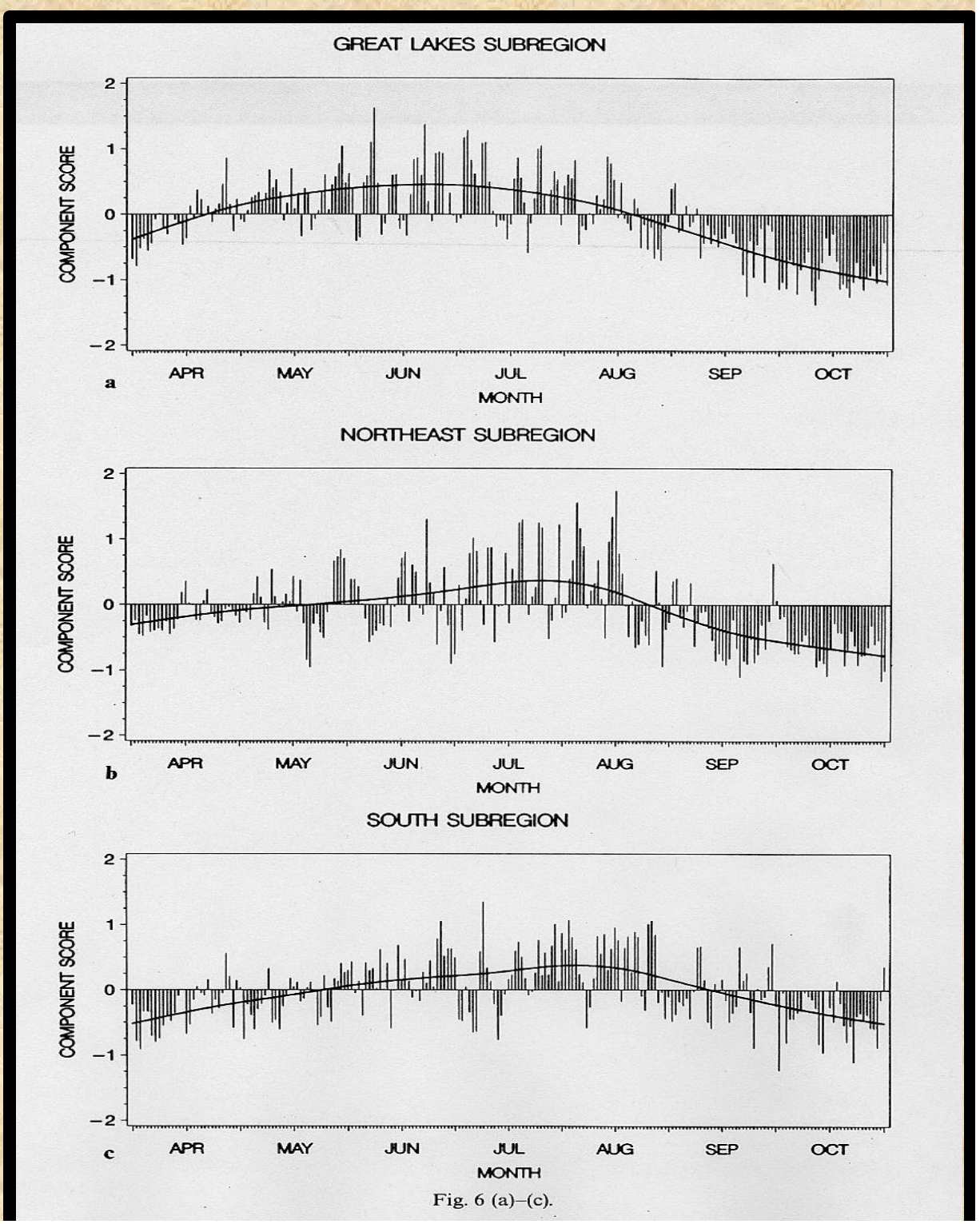


Fig. 6 (a)–(c).

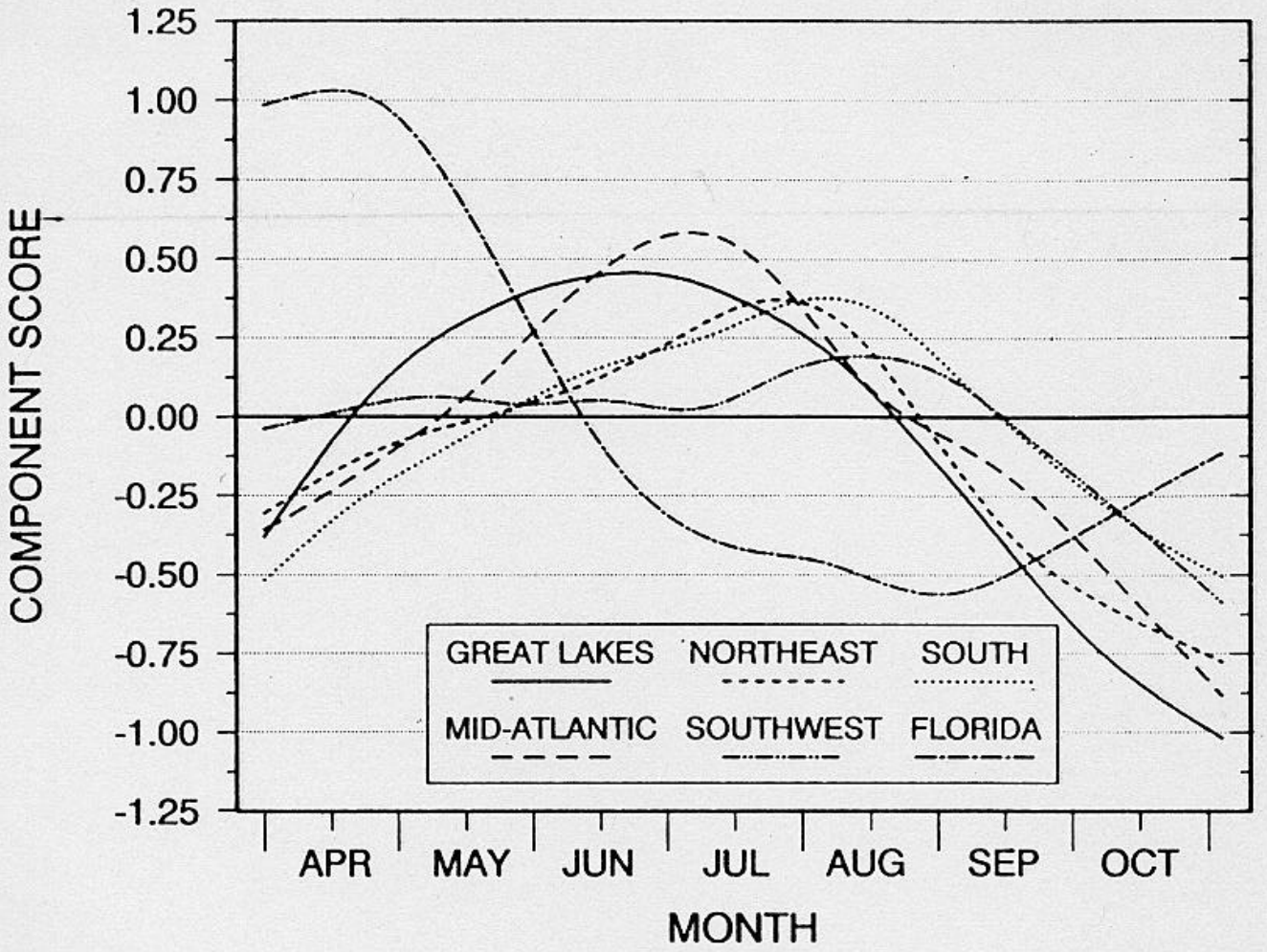


Fig. 7. Composite seasonal time series as defined by smoothed median principal component scores for each of the six homogeneous subregions.

Correlograms

Standardized PC scores

Deseasonalized

Weaker persistence

Stronger Persistence

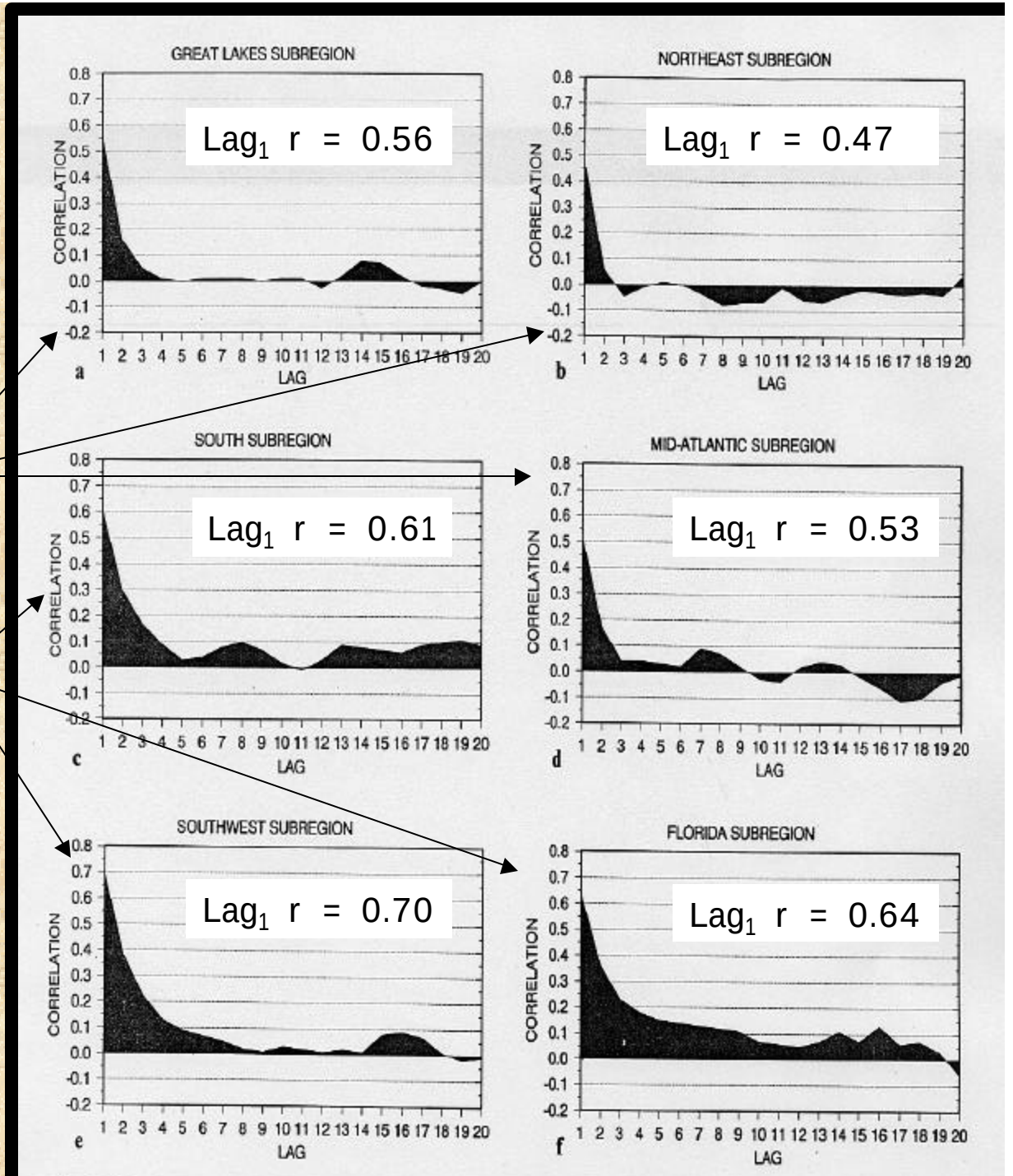


Fig. 8. Correlograms of the deseasonalized daily standardized principal component scores associated with the six homogeneous subregions.

Spectral Density

Standardized PC scores

Deseasonalized

"White Noise"

"Red Noise"

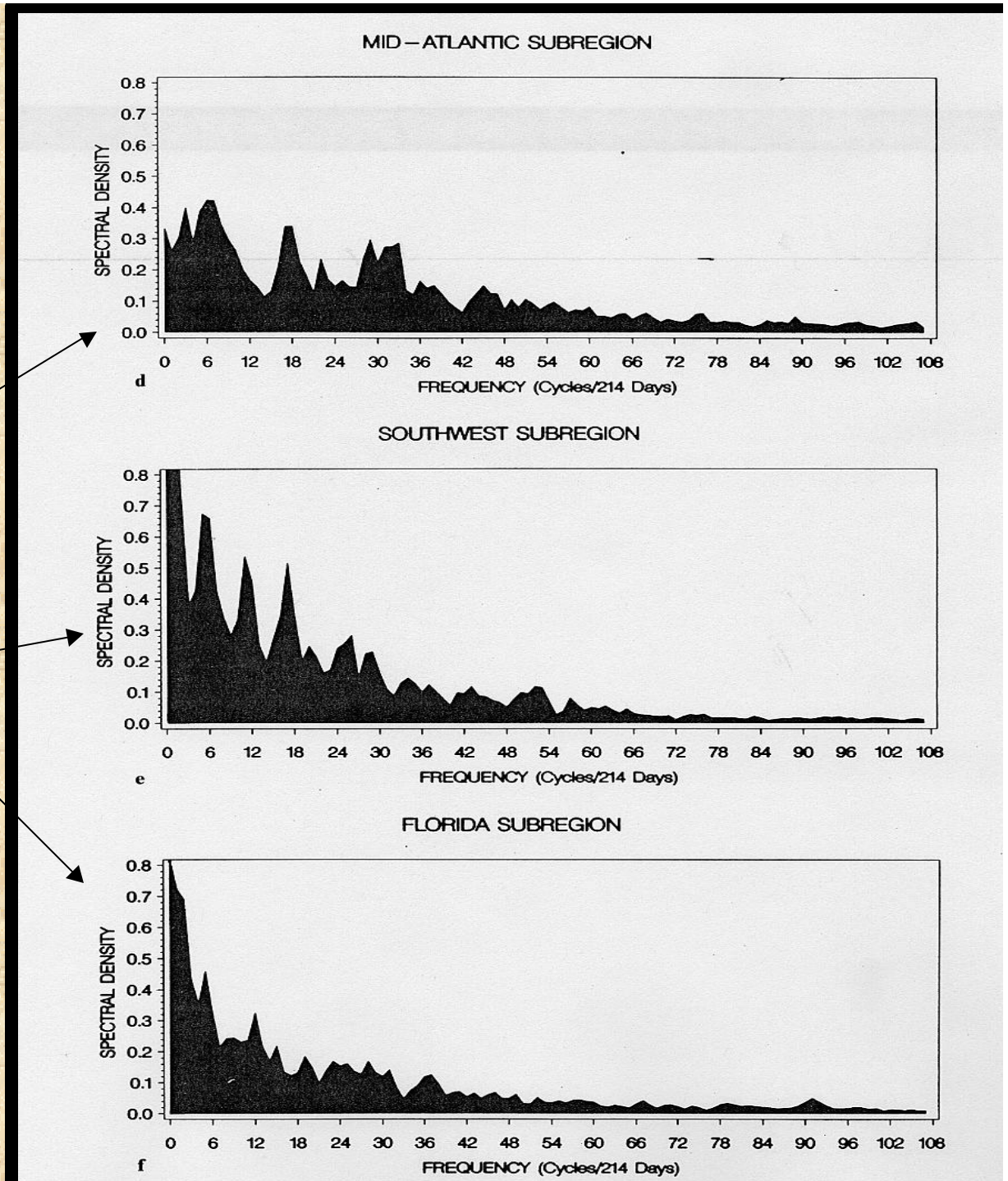


Fig. 9. Spectral density analysis of the deseasonalized daily standardized principal component scores associated with the six homogeneous subregions.

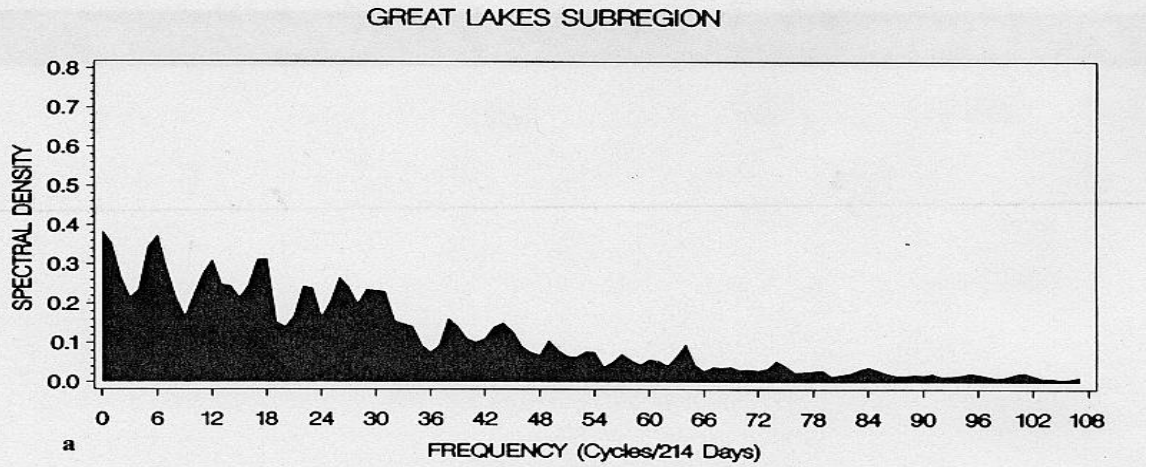
Spectral Density

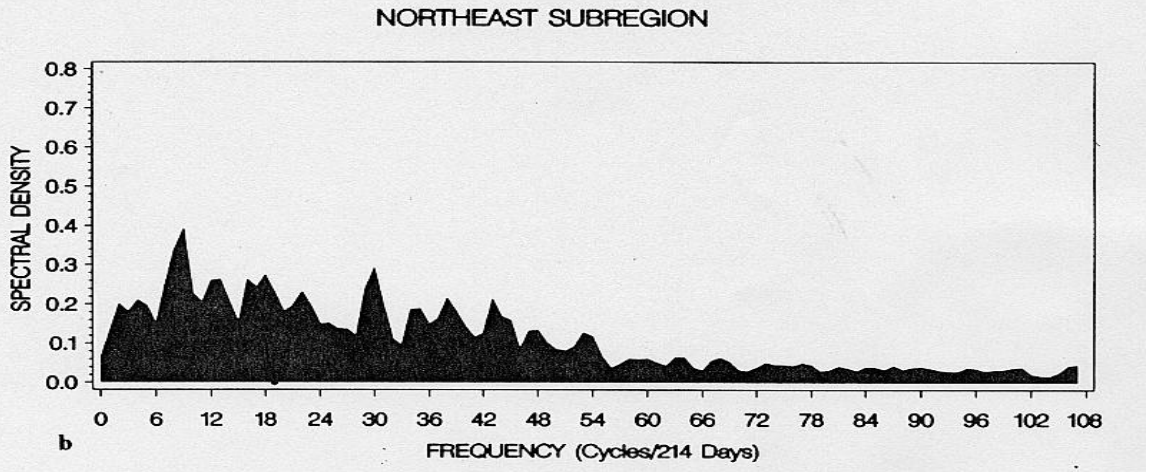
Standardized PC scores

Deseasonalized

"White Noise"

"Red Noise"





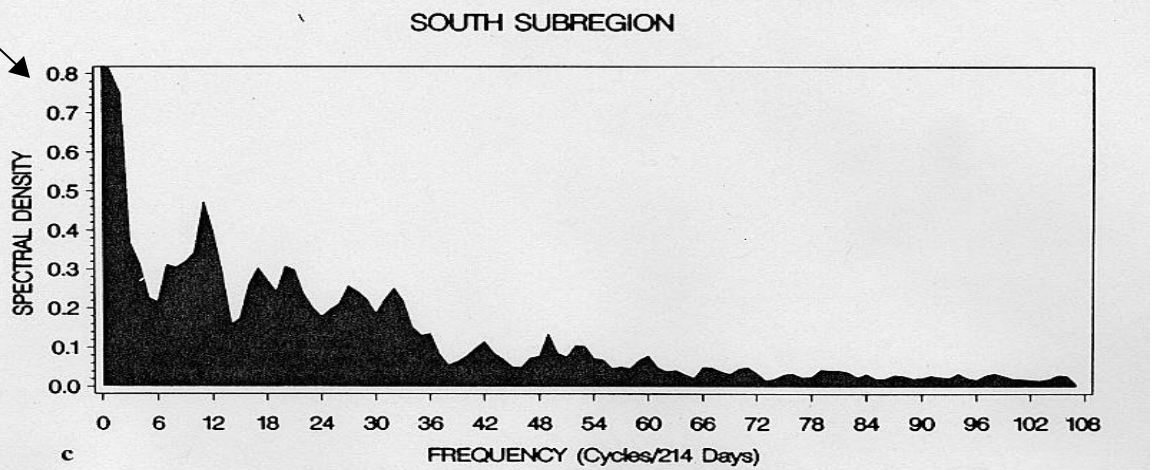


Fig. 9 (a)–(c).

Summary

Principal Component Analysis:

- allows one to examine the spatial and temporal variability of environmental data across a myriad of scales;

  - utilization of Kaiser's orthogonal rotation facilitates identification of "*influence regimes*" where concentrations exhibit statistically unique and homogenous characteristics.

  - utilization of time series analyses, including spectral density analysis, facilitates characterization of the "influence regimes"

Summary

Principal Component Analysis

- is useful in that it:

  - can provide “weight of evidence” of the regional-scale nature of environmental data,
  - facilitate understanding of the mechanisms responsible for the data’s unique variability among *influence regimes,*
  - designate stations (grid cells) that can be used as indicators for each *influence regime*,
  - identify erroneous data or data artifacts that are often missed with other analyses.