Revista Brasileira de Meteorologia, **v.22**, **n.2**, 273-285, 2007

ASPECTOS METODOLÓGICOS DEL USO DEL ANÁLISIS DE COMPONENTES PRINCIPALES EN CAMPOS DE ANOMALIAS DE ALTURA GEOPOTENCIAL EN EL SUR DE SUDAMERICA

MARIA LAURA BETTOLLI[1, 2, 3], WALTER MARIO VARGAS[1, 2, 4] y OLGA CLORINDA PENALBA[2, 5]

[1] Consejo Nacional de Investigaciones Científicas y Técnicas (CONICET)
[2] Departamento de Ciencias de la Atmósfera y los Océanos, FCEN, UBA.
Ciudad Universitaria, Pabellón 2 Piso 2. (1428) Buenos Aires, Argentina.
E-mails: [3] bettolli@at.fcen.uba.ar; [4] vargas@at.fcen.uba.ar; [5] penalba@at.fcen.uba.ar

Recibido Diciembre 2005 - Aceptado Noviembre 2006

RESUMEN

En este trabajo se discutieron aspectos metodológicos del uso del Análisis de Componentes Principales (ACP) a fin de verificar la eficacia del método para el conjunto de campos diarios de anomalías de altura geopotencial de 1000 y 500 hPa de los reanálisis 2 del NCEP en el sur de Sudamérica de los años 1979-2001 en el período comprendido entre octubre y mayo. Este análisis metodológico se realizó con el objetivo de obtener una base representativa de la información para clasificaciones sinópticas futuras. Se evaluó la necesidad de realizar rotaciones ortogonales y su efectividad, encontrando que éstas no mejoran la redistribución de la varianza explicada. A partir de una submuestra de 100 días elegidos al azar se analizó la influencia del efecto de la persistencia de los campos diarios consecutivos. Dicho efecto no modificó sustancialmente los resultados del ACP y no introdujo otras componentes principales significativas. El conjunto de datos de anomalías de altura geopotencial puede reducirse efectivamente realizando un ACP como técnica de síntesis. Este resultado se corroboró utilizando el test de esfericidad de Bartlett y una prueba empírica modificando la matriz de correlación. Finalmente, para representar un campo de anomalía real y, de esta manera, facilitarle un sentido físico es necesaria la combinación de dos o más CP.

Palabras claves: patrones sinópticos; altura geopotencial; sur de Sudamérica; análisis de componentes principales

ABSTRACT: METHODOLOGICAL ASPECTS OF THE USE OF THE PRINCIPAL COMPONENT ANALYSIS OF 1000 AND 500 hPa ANOMALIES OF GEOPOTENTIAL HEIGHT FIELDS IN SOUTHERN SOUTH AMERICA.

This paper presents a discussion of methodological aspects of the use of the principal component analysis (PCA) to verify its efficacy synthesis for the daily 1000 and 500 hPa anomalies of geopotential height fields of the October to May 1979-2001 NCEP reanalyses II in southern South America. This analysis was performed in order to obtain a representative data base for future synoptic classifications. The need of orthogonal rotations and their efficacy was analyzed showing that the explained variance redistribution was not improved. A random sub-sample of 100 days was used to study the influence of the persistence effect of consecutive daily fields. This effect did not significantly modify the PCA results and did not introduce other significant PCs. The set of geopotential height data may be considerably reduced using PCA as a synthesis technique. This result was corroborated using Bartlett's sphericity test and an empirical test on the modified correlation matrix. Finally, to represent a real field and thus facilitate a physical sense, it is found that the combination of two or more PCs is necessary.

Keywords: synoptic pattern; geopotential height; southern South America; principal component analysis.

1. INTRODUCCIÓN

Los patrones sinópticos han sido utilizados con el fin de sintetizar los modos típicos de circulación atmosférica en una región, asociados en general con algún fenómeno en superficie. Estas síntesis son herramientas importantes ya que permiten reducir el volumen de información para obtener una representación de la circulación atmosférica. Asimismo, la identificación de los patrones sinópticos, su frecuencia, distribución y variabilidad temporal son elementos importantes de diagnóstico y pronóstico, fundamentalmente en sistemas que hacen uso de ellos tales como el hidrológico, el agropecuario, etc. Más aún, el estudio de las características climatológicas de los patrones sinópticos constituye una referencia útil para definir los cambios en la circulación de la atmósfera.

Revisiones sobre los métodos de síntesis de información y aplicaciones a la climatología sinóptica pueden encontrarse, entre otros, en Barry y Perry (1973), Yarnal (1993) y más recientemente en Yarnal et al. (2001). Dentro de estos métodos, la técnica de Análisis de Componentes Principales (ACP) (Green, 1978) es una de las técnicas más utilizadas en forma exploratoria. Sus propósitos principales son reducir la dimensión del conjunto de datos e identificar nuevas variables o factores significativos subyacentes en el mismo. Esta metodología aplicada a campos de circulación reproduce con pocos patrones la mayor parte de la variación presente en la muestra de datos (Jolliffe, 1986). Como técnica para extractar y reproducir tipos de circulación en Argentina, Compagnucci y Vargas (1985) la utilizaron para estudiar los patrones de presión de superficie del mes de julio y Compagnucci y Salles (1997) a lo largo de todo el año en el período 1972-1983. Compagnucci y Vargas (1998) estudiaron los patrones de invierno asociados a caudales de los ríos de Cuyo, mientras que Müller et al. (2003) los asociaron con eventos de heladas en la Pampa Húmeda. Compagnucci et al. (2001) estudiaron la evolución de los sistemas atmosféricos sobre la base de patrones de secuencias principales para campos de altura geopotencial de 1000 hPa del año 1997 y Escobar et al. (2004) estudiaron los patrones de secuencias principales de campos de altura geopotencial en 1000 y 500 hPa asociados a entradas de aire frío en el centro de Argentina, como una extensión de la técnica de ACP. La misma técnica de secuencias principales pero en el modo S (extended empirical orthogonal functions - EEOF) fue utilizada por Vera y Vigliarolo (2000) y Vera et al. (2002) para estudiar las irrupciones de aire frío y las ondas de escala sinóptica en América del Sur para los inviernos de los años 1979-1993 utilizando la componente meridional del viento en 850 y 300 hPa. Asimismo, la técnica de ACP sirve como base para técnicas de clasificación. En este sentido, Chaves y Cavalcanti (2001) estudiaron las características de la circulación asociadas a la variabilidad de la precipitación en el sur del noreste de Brasil utilizando el ACP combinado con un método de clásificación. En el sur de Sudamérica, Solman y Menéndez (2003) clasificaron los campos diarios de alturas geopotenciales de 500 hPa en el período 1966-1999 utilizando EOF (modo-S en el ACP) como técnica inicial para el método de clasificación "K-means". Bejarán y Camilloni (2003) aplicaron un método de clasificación sinóptica para identificar masas de aire homogéneas que afectan la ciudad de Buenos Aires.

El ACP es una técnica muy utilizada en los estudios de la climatología sinóptica y en la construcción de series multivariadas para el pronóstico objetivo. Sin embargo, su utilización sin determinadas consideraciones conduce a inferencias sin justificación o a interpretaciones erróneas. Para su utilización es necesario analizar primeramente aspectos metodológicos a fin de verificar la eficacia y la factibilidad de empleo correcto del método para un conjunto de datos en particular. En este sentido, en este trabajo se propone analizar características del conjunto de datos que pueden tener relevancia en los resultados del ACP, y por lo tanto en su interpretación, bajo los siguientes objetivos específicos: a) evaluar cómo influye el efecto de la persistencia en los resultados de la aplicación del ACP, es decir, se analizará cuán robusto es el ACP frente a dicho efecto; b) evaluar la esfericidad del conjunto de datos analizando si el mismo puede reducirse efectivamente realizando un ACP como técnica de síntesis; c) analizar la representatividad de la base de componentes principales retenidas. Para llevar a cabo estos objetivos se utiliza un conjunto de datos en particular que se discutirá en la sección siguiente.

2. DATOS Y METODOLOGÍA

Para realizar este trabajo se utilizaron campos medios diarios de altura geopotencial en 1000 y 500 hPa en el período 1979-2001 correspondientes a la base de Reanálisis 2 (http://www.cpc.ncep.noaa.gov; Kanamitsu et al., 2002) del National Center for Environmental Prediction (NCEP) provistos por el NOAA-CIRES Climate Diagnostics Center. El dominio elegido se extiende desde 15° S a 60°S de latitud y desde 30°O a 90°O de longitud e incluye 475 puntos de enrejado (2.5° de latitud por 2.5° de longitud). Se trabajó con campos de anomalías diarias de altura geopotencial que se calcularon restando a cada campo diario el campo promedio correspondiente a ese día del año en el período de análisis, filtrando de esta forma la onda estacional. El período analizado corresponde al comprendido entre el 1 de octubre y el 31 de mayo del año siguiente (22 períodos de 243 días sin considerar los 29 de febrero). Se conformaron así las matrices finales con una dimensión de: 475 filas x 5346 columnas cada una. La selección de este período se debió a que en el mismo ocurren las mayores precipitaciones en la región centro-oriental de Argentina (Prohaska, 1976). Asimismo,

corresponde con el período de crecimiento de los principales cultivos de verano en la Pampa Húmeda argentina, los cuales dependen críticamente de la cantidad de precipitación y de su distribución temporal, ya que se desarrollan fundamentalmente en condiciones de secano (sin riego artificial). Los resultados de este trabajo serán la base para estudios futuros de aplicación relacionados con un cultivo en particular, que se desarrolla en este período, y con las lluvias extremas en esa época del año.

Se aplicó el método de Análisis de Componentes Principales (Richman, 1986; Jolliffe, 1986) en modo T con matriz de correlación, donde las variables son los 5346 campos de anomalías diarias y las observaciones corresponden a los 475 puntos de enrejado. La selección de la cantidad de componentes principales (CP) a retener se realizó en base a una combinación de distintos criterios como los diagramas de LEV (logaritmo del autovalor, Craddock y Flood,1969), la regla de Kaiser (1960), el 'scree graph' (Cattell, 1966) y la varianza total explicada por las componentes. Se exploró la redistribución de la varianza explicada por las CP con las rotaciones ortogonales Varimax (Kaiser, 1958) y Quartimax (Harman, 1967).

La influencia del efecto de la persistencia de los campos diarios consecutivos en los resultados del ACP se analizó en base a una submuestra de 100 días elegidos al azar. Se utilizó el test de esfericidad de Bartlett (1950) y una prueba empírica modificando la matriz de correlación, a fin de analizar si el conjunto de datos de anomalías de altura geopotencial bajo estudio puede reducirse efectivamente realizando un ACP como técnica de síntesis.

3. RESULTADOS

3.1. Patrones del ACP

La determinación de la cantidad de CP que representan la parte sustancial de la información contenida en el conjunto original de datos se realizó en base a distintos criterios. En la Figura 1 se muestran los gráficos de 'scree' para ambos niveles y en ellos se observa que la curva posee un quiebre o codo para la CP número 6 tanto para 1000 como para 500 hPa. En los diagramas de LEV, los puntos muestran una alineación a partir de este número de CP (Figura 1) y el porcentaje de varianza explicado por estas seis primeras CP para ambos niveles es superior al 80% (Tabla 1). Al aplicar la regla de Kaiser, en 1000 hPa deberían retenerse 55 CP y en 500 hPa 41 CP. Dado que, en cada nivel, con las 6 primeras CP se explica más del 80% de la varianza total y que los dos primeros criterios son consistentes entre sí se decidió trabajar con la base ortogonal de las 6 primeras CP.

La distribución espacial de las CP puede ser interpretada en su fase positiva o negativa (Figura 2, No Rotadas). Las componentes de peso positivas tienen el mismo signo que las anomalías de altura geopotencial. Por lo tanto, las componentes de puntaje positivas (negativas) representan anomalías positivas (negativas) de altura geopotencial, las que se denominarán modo directo (modo indirecto).

Para el nivel de 1000 hPa las primeras 6 componentes principales explican el 80.7% de la varianza mientras que en 500 hPa, las primeras 6 componentes principales explican el 83.5% de la varianza total (Tabla 1). Para cada uno de los niveles, se evaluó el porcentaje de varianza explicada por las CP en el modo directo e indirecto por separado observándose una similitud en los valores de varianzas para cada una de las CP (Tabla 1; columnas No Rotadas). Los gráficos de las componentes de peso para una CP en función de las componentes de peso de otra CP ('pairwises plots') permiten analizar si las matrices de componentes de peso presentan estructura simple (Thurstone, 1947). En estos diagramas, la estructura simple ideal debe presentar una gran concentración de puntos a lo largo de los ejes, un gran número de puntos cerca del origen y un pequeño número de puntos fuera de la línea de los ejes (Richman, 1986). El análisis de estos gráficos (no mostrados) reveló que las matrices de componentes de peso en ambos niveles no presentaban estructura simple. Por ello, se exploró la redistribución de la varianza explicada por las CP con las

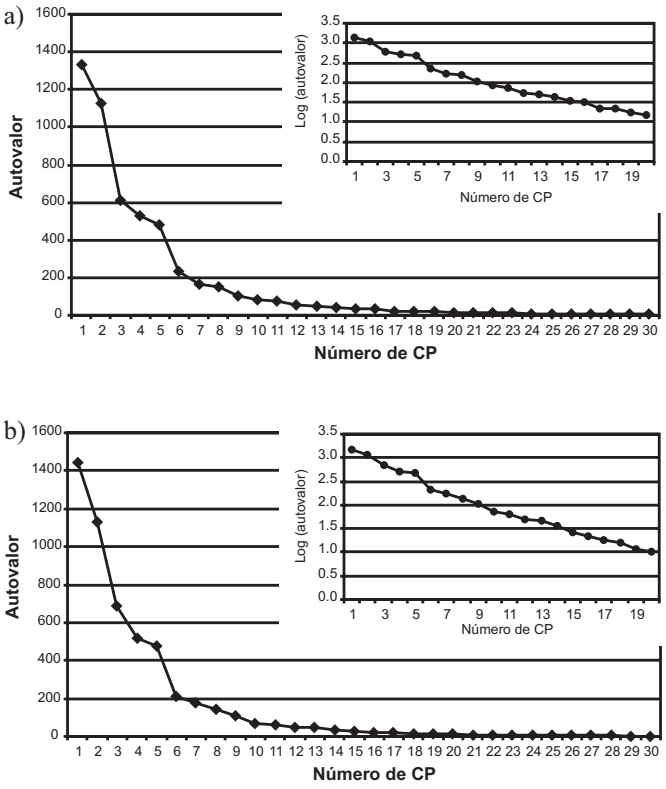

Figura 1 – Autovalores de las 30 primeras CP en función del número de CP para 1000 **(a)** y 500 **(b)** hPa (gráficos exteriores). Gráficos interiores: Diagramas de LEV (log-eigenvalue).

Tabla 1 – Porcentajes de varianza explicada y acumulada de las seis primeras componentes principales, no rotadas y rotadas Quartimax y Varimax, en el modo directo e indirecto para 1000 y 500 hPa. *Var: varianza explicada, Dir: modo directo, Ind: modo indirecto, Ac: varianza acumulada.*

	No Rotadas				Rotadas Quartimax				Rotadas Varimax			
1000 hPa												
	Var (%)	Dir (%)	Ind (%)	Ac (%)	Var (%)	Dir (%)	Ind (%)	Ac (%)	Var (%)	Dir (%)	Ind (%)	Ac (%)
CP1	24.9	12.5	12.4	24.9	23.1	12.0	11.1	23.1	23.1	11.9	11.1	23.1
CP2	21.1	11.0	10.1	46.0	22.8	11.8	11.1	46.0	22.5	11.6	10.9	45.6
CP3	11.4	5.7	5.7	57.5	11.4	5.8	5.7	57.4	11.8	5.9	5.9	57.4
CP4	9.9	4.7	5.2	67.4	9.5	4.5	5.1	66.9	9.6	4.5	5.1	67.0
CP5	9.0	4.5	4.5	76.4	9.4	4.9	4.5	76.4	9.4	4.9	4.5	76.4
CP6	4.3	2.3	2.0	80.7	4.3	2.3	2.0	80.7	4.4	2.3	2.1	80.7
500 hPa												
CP1	27.0	13.6	13.4	27.0	26.9	13.6	13.3	26.9	24.7	12.8	11.9	24.7
CP2	21.1	9.9	11.2	48.1	21.0	9.9	11.1	47.9	22.9	11.1	11.8	47.6
CP3	12.9	6.6	6.2	61.0	12.9	6.5	6.3	60.8	12.9	6.6	6.3	60.5
CP4	9.7	5.0	4.7	70.6	9.7	5.0	4.7	70.5	9.9	5.1	4.8	70.4
CP5	9.0	4.7	4.2	79.6	9.1	4.9	4.2	79.6	9.2	4.9	4.3	79.6
CP6	4.0	1.8	2.1	83.5	4.0	1.9	2.1	83.5	4.0	1.9	2.1	83.5

rotaciones ortogonales Varimax (Kaiser, 1958) y Quartimax (Harman, 1967). Luego de las rotaciones ortogonales, las varianzas explicadas por las CP rotadas no cambiaron sustancialmente, y tampoco lo hicieron en la distribución de varianza explicada los modos directo e indirecto (Tabla 1; columnas Rotadas Quartimax y Rotadas Varimax). Del mismo modo, las matrices de componentes de peso tampoco mostraron estructura simple en ambos niveles.

Una herramienta para evaluar el ajuste entre los campos espaciales de las CP y los campos reales es el coeficiente de congruencia (Richman, 1986). En la Figura 3 se muestran las medianas de los coeficientes de congruencia en función de la cantidad de CP no rotadas y rotadas Varimax y Quartimax para ambos niveles. De esta figura se observa que, tanto para las CP no rotadas como las rotadas, el ajuste entre los patrones espaciales y los campos de anomalías de altura geopotencial es muy bueno para las primeras CP (superior a 0.9). Las diferencias entre las curvas no son relevantes como para concluir que las rotaciones ortogonales mejoran el ajuste entre los campos matemáticos y los campos reales. Por otro lado, para ambos niveles y rotaciones los patrones espaciales rotados resultantes fueron muy similares a los no rotados (Figura 2). Por lo tanto, al no lograrse una redistribución efectiva de la varianza luego de la rotación ni un mejor ajuste entre los patrones espaciales matemáticos y los campos de anomalías reales, para los análisis posteriores se utilizará la base ortogonal de las seis componentes no rotadas en ambos niveles. Por otro lado, existen estudios que muestran que las soluciones rotadas estarían menos afectadas por la dependencia del dominio (Richman, 1986). Sin embargo, este resultado muestra que luego de la rotación las estructuras espaciales se mantienen, de forma tal que el tamaño del dominio actúa del mismo modo en todas las soluciones (no rotadas y rotadas).

3.2. Efecto de la Persistencia

En general, en la escala sinóptica, se asume que los campos de altura geopotencial de días consecutivos presentan una persistencia de al menos una semana de duración, dado que las ondas que la definen se encuentran en este rango de frecuencia. Con el fin de analizar el efecto de la persistencia en los resultados del ACP aplicado a estos campos diarios de anomalías de altura geopotencial se procedió a analizar en primera instancia cuál es la memoria de dichos campos. Para

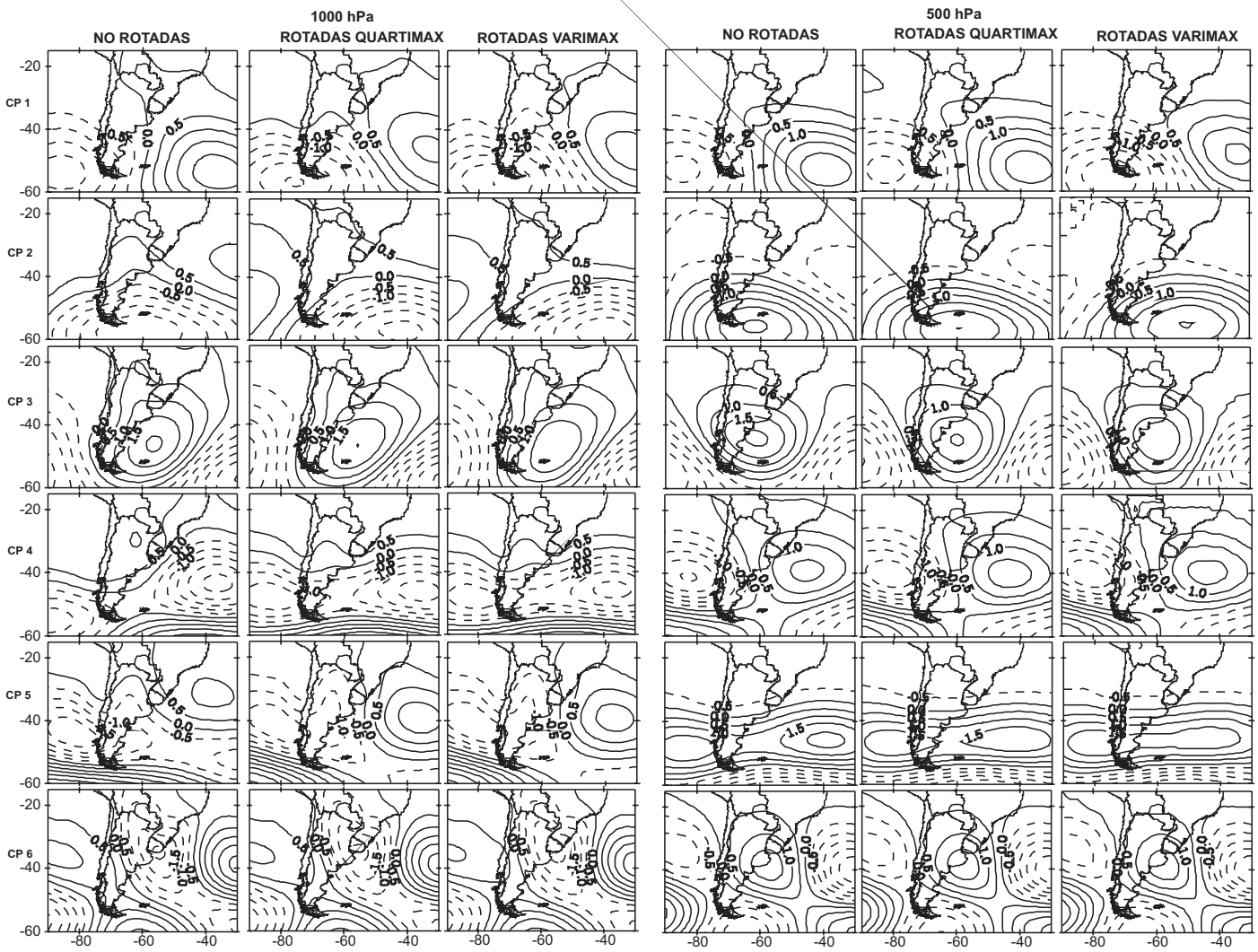

Figura 2 – Patrones espaciales de las primeras seis componentes principales no rotadas y rotadas Quartimax y Varimax para 1000 y 500 hPa. Línea de puntos: valores negativos.

ello se seleccionó una submuestra de la muestra total del análisis, el período comprendido entre el 1 de octubre de 1980 al 31 de mayo de 1981 (243 días seguidos). Este período se eligió debido a que no es un período anómalo ya que no posee ocurrencia de eventos muy extremos en precipitación (Bettolli et al., 2005) y temperatura (Barrucand y Rusticucci, 2001) y tampoco coincide con eventos El Niño o La Niña (Trenberth, 1997). Para cada uno de los niveles, se calcularon los distintos coeficientes de correlación entre campos diarios de anomalías de altura geopotencial para distintos desfasajes en días. En la Figura 4 se presentan las distribuciones de los coeficientes de correlación según el desfasaje en días. Para cada desfasaje se muestran los 'box plots' de los coeficientes de correlación con la mediana como medida central, el intervalo intercuartil y el rango como medidas de dispersión. En ambos niveles la distribución de los coeficientes correspondientes al desfasaje de un día es marcádamente asimétrica con medianas que alcanzan valores de 0.71 y 0.78 para 1000 y 500 hPa respectivamente e intervalos intercuartiles reducidos (0.26 y 0.22). En ambos niveles, casi la totalidad de la distribución de coeficientes de correlación para este desfasaje es positiva. A medida que se aumenta el desfasaje los valores de las medianas van disminuyendo estabilizándose alrededor del cero, de manera que a partir de un desfasaje de cuatro días las medianas se encuentran concentradas en un rango pequeño de variación. Los valores de las medianas de los coeficientes de correlación presentan su primer mínimo relativo para un desfasaje de siete días en ambos niveles y a partir de allí se mantienen estables cercanos a cero. Del mismo modo, a medida que aumenta el desfasaje los intervalos intercuartiles van aumentando y se

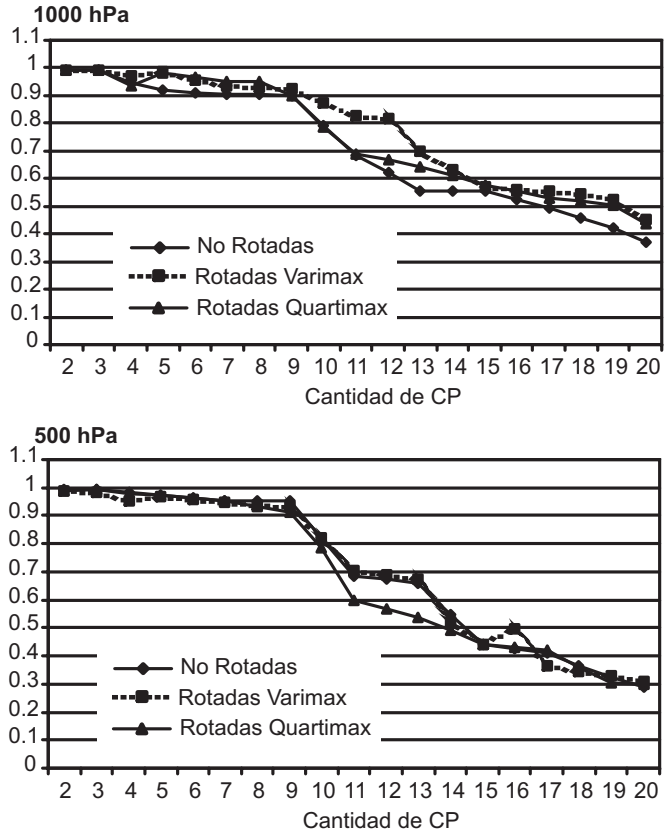

Figura 3 – Medianas de los coeficientes de congruencia en función de la cantidad de CP no rotadas y rotadas Varimax y Quartimax.

estabilizan en un valor promedio de 0.62 en ambos niveles. Si bien sólo se muestran los primeros 50 desfasajes, para mayores desfasajes la figura no cambia sustancialmente. Estos resultados evidencian, por un lado, la persistencia de los campos diarios en los primeros días, fundamentalmente hasta 4, y por el otro, que los campos atmosféricos de anomalías de altura geopotencial presentan disposiciones espaciales que varían dentro de un rango de estructuras espaciales, característica que se refleja en la dispersión de los coeficientes de correlación alrededor del cero.

Lo analizado anteriormente pone en evidencia que las muestras de campos diarios consecutivos de anomalías de altura geopotencial analizadas en este trabajo presentan un factor de persistencia. En este punto se desea analizar si dicha persistencia influye en los resultados del ACP. Para ello, en cada nivel se tomó una submuestra de 100 días al azar de manera de eliminar la persistencia y el orden cronológico de los campos de anomalías de geopotencial. Se realizó un ACP a cada submuestra. Utilizando para las submuestras los mismos criterios empleados para determinar la cantidad de componentes a retener en las muestras totales, se encontró que para ambos niveles las seis primeras componentes principales satisfacían dichos criterios (no mostrado). Los patrones resultantes en 1000 y 500 hPa se muestran en la Figura 5 y las varianzas explicadas en la Tabla 2. Al comparar estos patrones con los hallados para la muestra total se observa que son similares aunque no conservan estrictamente el mismo orden de varianza explicada pero sí el mismo orden de magnitud. En 1000 hPa, los patrones de las CP 1 y 2 se corresponden con los respectivos patrones de las CP 1 y 2 de la muestra total, tanto en la estructura de los campos como en la similitud de los porcentajes de varianza explicada. El patrón de la CP3 de la submuestra se puede vincular con el de la CP5 de la muestra total, del mismo modo que los patrones de las CP4 y CP5 de la submuestra se pueden asociar a los patrones de las CP3 y CP4 de la muestra total. En la muestra total los patrones de las CP3, CP4 y CP5 explican valores similares de varianza, por lo tanto se puede esperar que en la submuestra los patrones de las CP3, CP4 y CP5 no conserven el mismo orden aunque sí valores de varianza explicada similares a los de la muestra total. La CP6 de la submuestra presenta un patrón comparable al de la CP6 de la muestra total, con

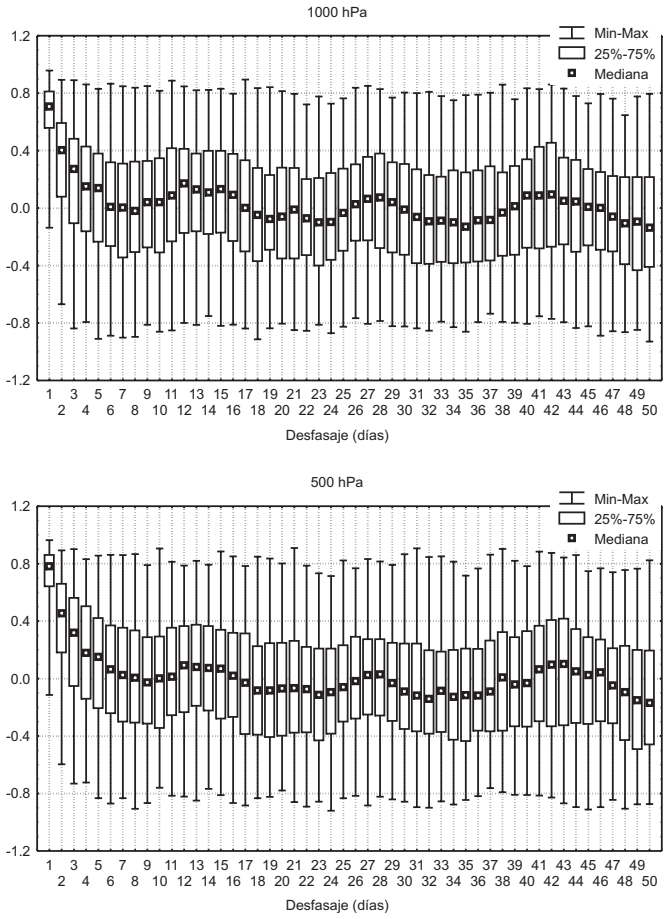

Figura 4 – Distribución de los coeficientes de correlación entre campos de anomalías altura geopotencial en función del desfasaje en días para 1000 y 500 hPa en el período 1980/81 y hasta un desfasaje de 50 días.

un cambio en el eje de inclinación de las anomalías pero sin cambios en la disposición espacial de las mismas. El análisis de la similitud entre los patrones espaciales de la submuestra y la muestra se realizó, adicionalmente, mediante el cálculo de los coeficientes de correlación entre patrones, dándole un enfoque más objetivo. Los valores de las correlaciones corroboran la correspondencia hallada entre los patrones de la submuestra y la muestra (no mostrado). Asimismo, el porcentaje total de varianza explicada por las seis CP de la submuestra (81.1% - Tabla 2) es muy similar al explicado por las primeras seis CP de la muestra total (80.7% - Tabla 1). De la Tabla 2 también se observa que los modos directo e indirecto son posibles en la submuestra y que los porcentajes de varianza explicada en ambos modos son más similares en la muestra total, situación que puede deberse a la mayor cantidad de casos y, por lo tanto, estabilidad de la muestra total.

Al analizar los resultados de 500 hPa se observa una situación similar: los patrones de las CP de la submuestra son comparables a los de la muestra total aunque no conservan el mismo orden en cuanto a la varianza explicada pero sí el mismo orden de magnitud. Este cambio en el orden de la varianza explicada se produce en los patrones de las CP4 y CP5 de la submuestra que se asocian respectivamente a los patrones de las CP5 y CP4 de la muestra total. Sin embargo, los órdenes de varianza explicada son equivalentes. Del mismo modo que en 1000 hPa, los valores de las correlaciones confirman la correspondencia entre los patrones espaciales de la submuestra y la muestra (no mostrado). Asimismo, el porcentaje total de varianza explicada por las seis CP de la submuestra (85.2% - Tabla 2) es muy similar al explicado por las primeras seis CP de la muestra total (83.5% - Tabla 1). Al considerar los modos directo e indirecto se observa que ambos modos existen en la

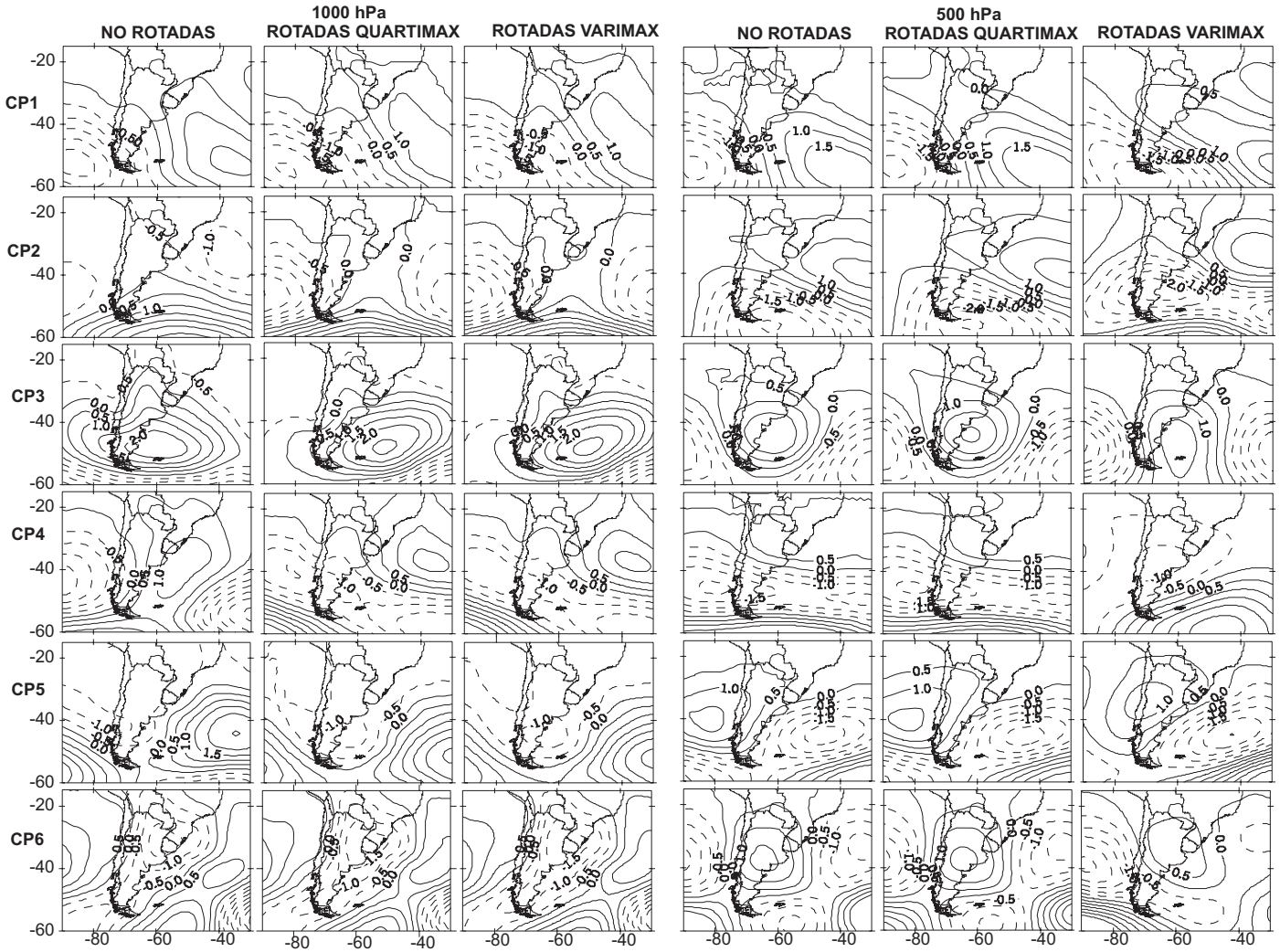

Figura 5 – Patrones espaciales de las primeras seis componentes principales no rotadas y rotadas Quartimax y Varimax de las submuestras de 100 días al azar de 1000 y 500 hPa. Línea de puntos: valores negativos.

Tabla 2 – Porcentajes de varianza explicada y acumulada de las seis primeras componentes principales, no rotadas y rotadas Quartimax y Varimax, en el modo directo e indirecto para las submuestras de 100 días al azar de 1000 y 500 hPa.

	1000 hPa											
	No Rotadas				**Rotadas Quartimax**				**Rotadas Varimax**			
	Var (%)	**Dir (%)**	**Ind (%)**	**Ac (%)**	**Var (%)**	**Dir (%)**	**Ind (%)**	**Ac (%)**	**Var (%)**	**Dir (%)**	**Ind (%)**	**Ac (%)**
CP1	26.3	16.7	9.6	26.3	25.4	17.0	8.4	25.4	25.1	16.8	8.3	25.1
CP2	19.9	6.5	13.4	46.1	16.6	6.2	10.4	42.0	15.9	6.0	9.9	41.0
CP3	12.2	6.5	5.7	58.3	10.9	5.7	5.2	52.9	11.0	5.8	5.2	51.9
CP4	10.3	5.4	4.8	68.6	12.3	7.6	4.7	65.2	12.4	7.7	4.7	64.3
CP5	8.1	4.9	3.2	76.7	11.1	5.1	5.9	76.3	11.8	5.4	6.3	76.1
CP6	4.4	2.9	1.5	81.1	4.8	3.2	1.7	81.1	5.0	3.2	1.8	81.1
	500 hPa											
CP1	27.5	17.9	9.7	27.5	24.5	17.8	6.7	24.5	23.3	17.0	6.2	23.3
CP2	20.3	13.0	7.3	47.8	20.3	11.3	8.9	44.8	10.1	5.0	5.2	33.4
CP3	15.1	7.9	7.2	62.9	16.9	9.2	7.7	61.7	16.8	9.3	7.5	50.2
CP4	10.6	4.4	6.2	73.5	9.3	4.4	4.9	71.0	19.9	8.9	11.0	70.1
CP5	8.0	3.4	4.6	81.5	7.7	2.8	4.9	78.7	7.9	2.8	5.1	78.0
CP6	3.8	1.4	2.4	85.2	6.5	2.7	3.9	85.2	7.2	2.8	4.4	85.2

submuestra aunque en los porcentajes de varianza explicada por ambos modos en la muestra total, la diferencia es menor.

A fin de explorar la redistribución de la varianza explicada, se realizaron las rotaciones ortogonales Quartimax y Varimax en las bases de CP de las submuestras de 100 días al azar. Los patrones resultantes de la rotación se muestran en la Figura 5. Como puede observarse no hay cambios importantes entre los patrones no rotados y rotados ni tampoco en los porcentajes de varianza explicados (Tabla 2). En el nivel de 1000 hPa se encuentran las mayores diferencias en las estructuras de las CP, principalmente en la CP4.

3.3. Esfericidad de la Muestra

Uno de los aspectos a considerar cuando se desea utilizar la técnica de ACP en un conjunto de datos es si efectivamente el conjunto de datos puede reducirse en espacio a través de esta técnica. El interés en este punto del trabajo es analizar si la técnica de ACP se puede utilizar como técnica de síntesis del conjunto de campos diarios de anomalías de altura geopotencial de interés. Este aspecto debería evaluarse de manera preliminar antes de realizar el ACP, sin embargo, en este trabajo se plantea en forma posterior a fin de poder comparar los resultados con los de la muestra total. En un ACP los autovectores de la matriz de similitud con la que se trabaja (en este caso matriz de correlación) son las nuevas direcciones donde se proyecta la matriz de datos originales. Si la matriz de correlación fuera una matriz diagonal estaría indicando que entre los campos de entrada no hay correlación excepto para el caso del campo consigo mismo (donde la correlación es 1). En este caso, la utilización del ACP como técnica para identificar patrones que representen al conjunto de datos resultaría inadecuada ya que los campos en sí representan patrones (o nuevas direcciones) ortogonales. Este caso representa una situación hipotética teórica que en el conjunto de datos meteorológicos no ocurre ya que los campos diarios están vinculados unos a otros por causas físicas como puede observarse en la Figura 4. Sin embargo, con una estructura dada de la matriz de correlación podría ocurrir que no se encuentren direcciones preferenciales donde se vaya maximizando progresivamente la varianza explicada. Por lo tanto, un conjunto cualquiera de ejes, variables o factores (incluyendo las variables originales) resultaría tan bueno como cualquier otro para preservar la información original. Cuando ocurre esta situación la configuración geométrica de la nube

de puntos alrededor de los ejes principales es una esfera (para el caso de tres dimensiones). Por esta razón debe determinarse si el conjunto de datos permite descomponerse en patrones ortogonales representativos del mismo a través de un ACP.

Para evaluar este aspecto se realizaron dos procedimientos utilizando la matriz de correlación de la submuestra de 100 días elegidos al azar:

a) aplicar el test de esfericidad de Bartlett (1950), el cual compara la matriz de correlación con la matriz identidad;
b) modificar la matriz de correlación correspondiente a la submuestra de 100 días al azar, reemplazando los coeficientes de correlación no significativamente distintos de cero por ceros (independencia total entre campos) y buscar sus autovectores.

Para el primer procedimiento se planteó la hipótesis nula del test de Bartlett la cual propone que la matriz de correlación pertenece a una población de matrices de correlación identidad. Por lo tanto, si se rechaza la hipótesis nula, la matriz analizada se aleja de la matriz identidad. El estadístico chi-cuadrado propuesto por Bartlett tiene la forma

$$\chi^2_{\left[0.5(n^2-n)\right]} = -\left[m - 1 - \frac{1}{6}(2n+5)\right]\ln|R|$$

donde R es la matriz de correlación, *n* es el número de variables en R (columnas de las matrices de las submuestras, 100), *m* el tamaño de la muestra (filas de las matrices de las submuestras, 475) y ln|R| es el logaritmo natural del determinante de R. De manera que si la matriz R es igual a la matriz identidad su determinante sería 1 y el estadístico χ^2 sería 0. Los estadísticos obtenidos fueron 186466.78 y 218862.60 para 1000 y 500 hPa respectivamente, los cuales son altamente significativos rechazando la hipótesis nula. Dado que el test de Bartlett se emplea para muestras con $n >= 10$ y $m \approx 200$, se volvió a aplicar dicho test a las matrices de correlación de las submuestras de 100 días al azar reducidas. En cada matriz de las submuestras se redujo la cantidad de filas de las mismas quitando puntos de enrejado, es decir, se trabajó con un enrejado de 5° x 5° (238 puntos o filas). Esta reducción de puntos de enrejado no altera las estructuras de los campos de anomalías de altura geopotencial y por lo tanto, no altera los resultados del ACP (no mostrado). Con esta modificación en las matrices de entrada, los estadísticos χ^2 para 1000 y 500 hPa resultaron 41272.02 y 48593.34 respectivamente, los cuales nuevamente son altamente significativos. Esto significa que existen correlaciones significativas entre las variables y que por lo tanto las matrices son 'apropiadas' para el ACP.

Para el segundo procedimiento, en las matrices de correlación de las submuestras de 100 días al azar se mantuvieron sólo los coeficientes que resultaron estadísticamente distintos de cero con un nivel de significancia superior al 99% (Panofsky y Brier, 1965) y los demás se reemplazaron por cero. A partir de esta nueva matriz de correlación modificada se buscaron los autovalores y autovectores. Estos últimos son las direcciones sobre las cuales se proyecta la matriz de datos originales generando la matriz de nuevas variables. En ambos niveles se calcularon las primeras 6 nuevas variables resultantes de la proyección de las variables originales sobre los autovectores de la matriz de correlación modificada (Figura 6). Estas nuevas variables, que se denominarán CP por analogía, se compararon con los resultados encontrados previamente para evaluar las diferencias. De esta figura se observa que los patrones obtenidos son muy similares a los hallados sin modificar la matriz de correlación (Figura 5). El mismo resultado se encontró al comparar las series de componentes de peso entre ambos casos (resultados no mostrados). Como prueba empírica adicional, se utilizó un coeficiente de correlación crítico de 0.3 para modificar las matrices de correlación en base a los resultados de la Figura 4 donde se observó que las medianas de los coeficientes de correlación tendían a cero pero el intervalo intercuartil se estabilizaba con una amplitud de 0.3. Es decir, se retuvieron los coeficientes superiores en valor absoluto a 0.3 y los demás se consideraron 0. Los resultados, con este nuevo coeficiente crítico, no presentaron modificaciones con respecto al caso de estudio anterior (no mostrados).

Estos resultados indican que la técnica de ACP resulta válida en términos de sintetizar la muestra en base a patrones representativos de los campos diarios de anomalías de altura geopotencial analizados.

El estudio de la esfericidad se planteó a través de un método sencillo que se basa en el análisis de una submuestra de 100 días tomados al azar. Los resultados obtenidos para la submuestra pueden generalizarse a la muestra total en el caso en que se pruebe que la matriz de correlación de la submuestra no sea una matriz identidad. Esto es, si la matriz de correlación de la submuestra no es una matriz diagonal entonces tampoco lo será la de la muestra total. Otras ventajas de utilizar la matriz de la sumbuestra de 100 días al azar es que el efecto de persistencia no está presente en los campos de anomalías de altura geopotencial de manera que pueda influir sobre las correlaciones y que el costo computacional es menor al disminuir las dimensiones de la matriz de correlación.

3.4. Representatividad de los patrones espaciales del ACP

El valor de las componentes de peso permite evaluar el porcentaje de varianza explicada por los patrones teóricos (componentes principales) de las estructuras de anomalías reales ya que son el coeficiente de correlación entre cada campo de anomalía diaria y cada patrón teórico. Asimismo, cada día de

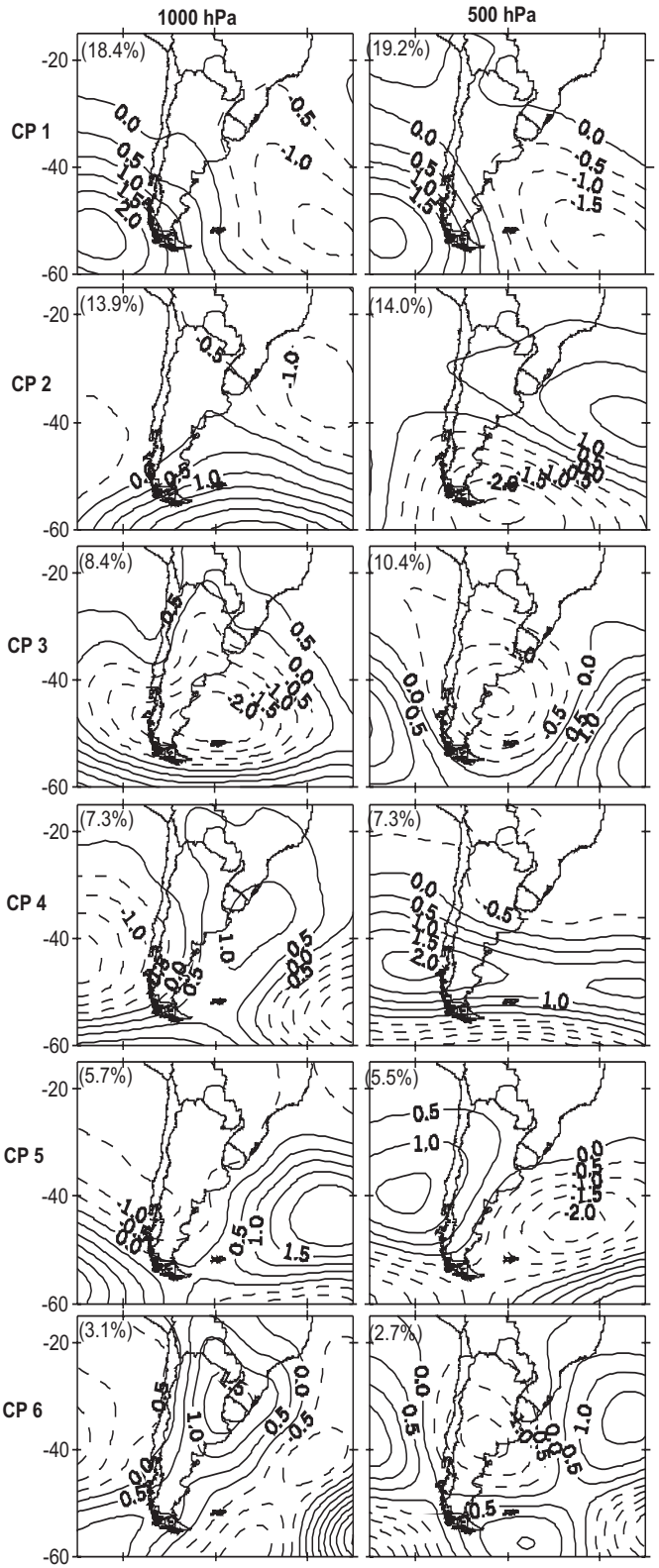

Figura 6 – Patrones espaciales de las primeras seis componentes principales de las sumbuestras de 100 días al azar para la matriz de correlación modificada para 1000 y 500 hPa. Línea de puntos: valores negativos. Entre paréntesis: porcentajes de varianza explicada.

la muestra puede reproducirse a partir de la combinación lineal de las componentes principales retenidas más un término de error, donde los coeficientes de la combinación lineal vienen dados por los factores de peso. De esta manera, sumando los factores de peso al cuadrado se puede determinar la varianza explicada por una componente en sí o por la combinación de varias componentes según sea el caso. En base a esta propiedad de linealidad, se analizó la representatividad de la base de CP elegidas para 1000 y 500 hPa fijando distintos umbrales (49%, 64% y 81%) de varianza explicada por la combinación lineal. Dado que ese umbral de varianza puede ser explicado por una única componente o por la combinación de varias, en este análisis se consideró que hasta tres CP podían explicar esa varianza. La Figura 7 muestra los porcentajes del total de días (5346) para los cuales una única CP, la combinación de a lo sumo 2CP y la combinación de a lo sumo 3 CP explican un porcentaje de varianza superior al umbral prefijado. En ambos niveles y en todos los casos (1CP, a lo sumo 2CP y a lo sumo 3CP) se observa que a medida que mayor es la exigencia del porcentaje de varianza explicado de cada día, menor es la cantidad de días que pueden ser representados. El porcentaje de días que se pueden representar con una sola componente principal va del 2% al 37 % en 1000 hPa y del 3% al 42% en 500 hPa a medida que se disminuye la exigencia en el umbral de varianza explicada. Sin embargo, si se considera cada umbral en particular, se observa que para el 49%, el porcentaje de días representados por la combinación de a lo sumo 2CP es más del doble del porcentaje de días representados por 1CP. Mientras que, para el umbral 81%, la cantidad de días representados por la combinación de a lo sumo 2CP es más del triple del porcentaje de días representados por 1 CP. Para este último umbral, el aumento de la representatividad de la combinación de 2CP a 3CP es notable (más del doble). En todos los umbrales, el aumento en la cantidad de días representados es más abrupto cuando se pasa de 1CP a la combinación de a lo sumo 2CP que cuando se pasa de esta última a la combinación de a lo sumo 3CP o a la combinación de la base total. Estos resultados están indicando que si bien los patrones de las CP en sí pueden ser parecidos a campos físicos reales, es necesaria la combinación de dos o más patrones para poder representar un campo de anomalía real y, de esta manera, facilitarle un sentido físico. Visto de otro modo, cada patrón de CP por sí solo no es representativo de la física de un campo de anomalía real, sólo es representativo de un determinado porcentaje de varianza del mismo.

Cuando se comparan ambos niveles se encuentra que en todos los casos en el nivel de 500 hPa se explica un mayor porcentaje de campos diarios de anomalías de altura geopotencial. Este resultado refleja el hecho de que en este nivel los campos de anomalías diarias son menos perturbados que en niveles inferiores.

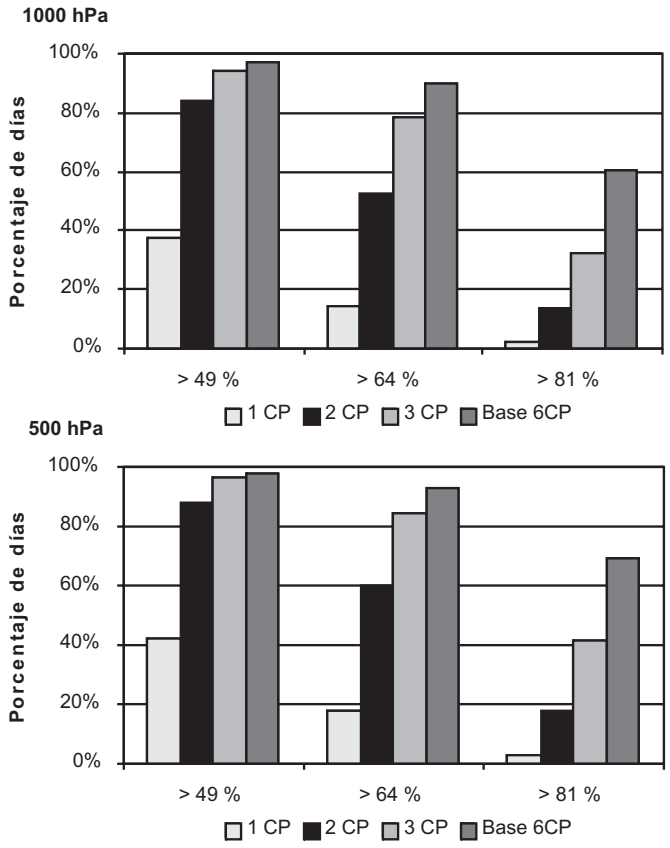

Figura 7 – Porcentaje de días (de los 5346 días) en los cuales con una sola componente principal (1CP), la combinación de a lo sumo dos (2CP), la combinación de a lo sumo tres (3CP) o la combinación de toda la base de CP retenidas se explica más del porcentaje de varianza indicado en el eje de las abscisas.

4. CONCLUSIONES

En este trabajo se discutieron aspectos metodológicos del uso del Análisis de Componentes Principales a fin de verificar la eficacia del método y la factibilidad de empleo correcto para el conjunto de campos diarios de anomalías de altura geopotencial de 1000 y 500 hPa de los reanálisis 2 del NCEP en el sur de Sudamérica de los años 1979-2001 en el período de octubre a mayo. Este análisis se plantea como base para una futura clasificación sinóptica y estudios de aplicación relacionados con un cultivo en particular, que se desarrolla en este período, y con las lluvias extremas en esa época del año

La herramienta del ACP es utilizada por diversos autores para el tratamiento de campos espaciales de distintas variables. Es una herramienta que se utiliza con el fin de simplificar el sistema que se pretende estudiar. Sin embargo, como toda herramienta estadística su aplicación sin determinadas consideraciones puede llevar a resultados y/o conclusiones erróneas. La herramienta provee una representación de la muestra total a través de un conjunto de elementos linealmente independientes entre sí forzando, en este caso, al conjunto de campos atmosféricos a poseer la propiedad de ortogonalidad. Dicha propiedad surge de la matemática de la herramienta y no de la naturaleza del sistema físico que se estudia, que es altamente no lineal. Este trabajo aborda esta problemática, proponiendo un análisis, en forma simple, de las características del conjunto de datos que pueden influenciar los resultados del ACP. Este análisis debería plantearse como un paso inicial básico necesario en la aplicación de ACP para luego realizar el análisis sinóptico/climatológico correspondiente.

El análisis de las rotaciones indica que, previo a decidir con qué base de CP se va a trabajar, se debe evaluar la necesidad de realizar una rotación y su efectividad. Si bien, las soluciones no rotadas del ACP no presentan estructura simple, las rotaciones ortogonales Quartimax y Varimax, comúnmente utilizadas en la literatura, no logran una redistribución efectiva de la varianza explicada por la base de CP ni modificaron las estructuras espaciales principales de los campos diarios de anomalías de altura geopotencial analizados.

Las correlaciones entre los campos de anomalías diarias de altura geopotencial, calculadas para un año en particular, pone de manifiesto la existencia de una persistencia de hasta 4 días. Se analizó si dicha persistencia influye en los resultados del ACP considerando en ambos niveles submuestras de 100 días elegidos a azar. El ACP realizado a dichas submuestras arroja resultados similares a los de la muestra total, tanto en la estructura espacial de los patrones como en los porcentajes de varianza explicados por los mismos. De este análisis, el cual es planteado como un método exploratorio sencillo, se concluye que la persistencia natural entre los campos diarios consecutivos de anomalías de altura geopotencial analizados no modifica sustancialmente los resultados y no introduce otras componentes principales significativas. De manera que si lo que se desea es buscar las estructuras espaciales principales de este conjunto de campos de altura geopotencial sin interesarse en la evolución temporal de la asociación con dichos patrones, con una muestra pequeña tomada al azar es suficiente y el costo computacional involucrado es bajo.

Adicionalmente, se analizó si el espacio original permite ser reducido a través de una base ortogonal de pocas variables. Tanto para las muestras totales como para las submuestras, los resultados del test de Bartlett, significativos al 99%, indican que las matrices de correlación son distintas de la matriz identidad y que por lo tanto las variables están relacionadas. Reforzando este resultado, las descomposiciones en factores de las matrices de correlación modificadas arrojan CP con estructuras altamente similares a las originales. De esta manera las bases de datos analizadas justifican un análisis de factores y permiten ser reducidas en espacios de

menores dimensiones a través de un ACP. El método simple de análisis propuesto en este trabajo para evaluar esta premisa en forma preliminar consiste en trabajar con una submuestra de días (o variables del ACP en modo T) elegidos al azar. De esta manera se garantiza mejor la independencia entre variables, se reduce el costo computacional y los resultados pueden generalizarse a la muestra total en el caso en el que se rechace la hipótesis de que la submuestra posee una matriz de correlación identidad.

La representatividad de las bases de CP retenidas se analizó en base al porcentaje de varianza de cada día explicado por los patrones teóricos (CP) a partir de distintos umbrales de varianza explicada y de las combinaciones lineales considerando una, dos y tres CP. Los resultados indican que es necesaria la combinación de dos o más patrones para poder representar un campo de anomalías real y, de esta manera, facilitarle un sentido físico, ya que cada patrón de CP por sí solo no es representativo de la física de un campo real, sólo es representativo de un determinado porcentaje de varianza del mismo. En este punto es importante destacar que, al utilizar la matriz de correlación como matriz de similitud, la comparación de las estructuras espaciales de las CP con las estructuras de anomalías reales es una función de forma y no de intensidad.

5. AGRADECIMIENTOS

Los autores agradecen al revisor anónimo por sus comentarios y sugerencias que mejoraron la claridad del trabajo. Este trabajo fue financiado por los proyectos UBA X135 y X234, CONICET PIP N°5139 y GOCE-CT-2003-001454 CLARIS.

6. REFERENCIAS BIBLIOGRAFICAS

BARRUCAND, M.; RUSTICUCCI, M. Climatología de temperaturas extremas en la Argentina. Variabilidad temporal y regional. **Meteorologica**, v. 26, p. 85-102, 2001.

BARTLETT, M. S. Tests of significance in factor analysis. Brit. **J. Psychol. Statist**. Section, v. , p. 77-85, 1950.

BARRY, R. G.; PERRY, A. H. **Synoptic Climatology: Methods and applications.** Methuen, London, 1973. 555 p.

BEJARAN, R. A.; CAMILLONI, I. A. Objective method for classifying air masses: an application to the analysis of Buenos Aires' (Argentina) urban heat island intensity. **Theor. Appl. Climatol.**, v. 74, p. 93-103, 2003.

BETTOLLI, M L.; PENALBA, O. C.; VARGAS, W. M. Características de la precipitación diaria en la región núcleo sojera argentina. In: IX CONGRESO ARGENTINO DE METEOROLOGIA, 10, 2005, Buenos Aires. Anales… Buenos Aires: CAM, 2005.1CD-ROM.

CATTELL, R. B. The scree test for the number of factors. **J. Multiv. Behav**. Res., v. 1, p. 245-276, 1966.

CHAVES, R. R.; CAVALCANTI, I.F.A. Atmospheric circulation features associated with rainfall variability over Southern Northeast Brazil. **Mon Wea Rev**, v. 129, n. 10, p. 2614-2626, 2001.

COMPAGNUCCI, R. H.; VARGAS, W. M. Tipificación de los campos báricos de superficie para julio 1972-1977. Análisis por componentes principales no-rotadas. **Geoacta,** v. 13, p. 57-70, 1985.

COMPAGNUCCI, R. H.; SALLES, M. A. Surface pressure patterns during the year over southern South America. **Int. J. Climatol.**, v. 17, p. 635-653, 1997.

COMPAGNUCCI, R. H.; VARGAS,W. M. Inter-annual variability of the cuyo rivers' streamflow in the Argentinean Andean mountains and ENSO events. **Int. J. Climatol.**, v. 18, n. 14, p. 1593-1609, 1998.

COMPAGNUCCI, R. H., ARANEO, D.; CANZIANI, P. O. Principal sequence pattern analysis: a new approach to classifying the evolution of atmospheric systems. **Int. J. Climatol.**, v.21, n. 2, p. 197-217, 2001.

CRADDOCK, J. M.; FLOOD, C. R.. Eigenvectors for representing the 500 mb geopotential surface over the Northern Hemisphere. **Q. J. R. Met. Soc.**, v. 95, p. 576-593, 1969.

ESCOBAR, G.; COMPAGNUCCI, R.; BISCHOFF, S. Sequence patterns of 1000 hPa and 500 hPa geopotential height fields associated with cold surges over Central Argentina. **Atmósfera**, v. 17, n. 2, 69-89, 2004.

GREEN, P. E. **Analysing Multivariate Data**. The Dryden Press: Illinois, USA, 1978. 519 p.

HARMAN, H. **Modern Factor Analysis**. The University of Chicago Press, 1967. 474 p.

JOLLIFFE, I. T. **Principal Component Analysis**. Springer-Verlag. 1986. 271 p.

KAISER, H. F. The Varimax criterion for analytic rotation in factor analysis. **Psycometrika**, v. 23, p. 187-200, 1958.

KAISER, H. F. The application of electronic computers to factor analysis. **Educ. Psychol. Meas**., v. 20, p. 141-151, 1960.

KANAMITSU, M.; EBISUZAKI, W.; WOOLLEN, J.; YANG, S-K; HNILO, J.J.; FIORINO, M.; POTTER, G. L. **Bulletin of the American Meteorological Society**, p. 1631-1643, 2002.

MÜLLER, G. V.; COMPAGNUCCI; R. H., NUÑEZ, M. N.; SALLES, M. A. Surface circulation associated with frost in the wet Pampas. **Int. J. Climatol**., v. 23, n. 8, p. 943-961, 2003.

PANOFSKY, H.; BRIER, G. **Some applications of Statistics to Meteorology**. College of Mineral Industries, The Pennsylvania State University, 1965. 223 p.

PROHASKA, F. J. **Climates of Central and South America World Survey of Climatology**, Elsevier Cientific Publishing Company, Amsterdam, p. 57-69, 1976.

RICHMAN, M. Rotation of Principal Components. **J. Climatol**., v. 6, p. 293-335, 1986.

SOLMAN, S. A.; MENENDEZ, C. G. Weather regimes in the South American sector and neighbouring oceans during winter. **Clim Dyn**, v. 21, n. 1, p. 91-104, 2003.

THURSTONE, L. L. **Multiple Factor Analysis**. Chicago: University of Chicago Press, 1947.

TRENBERTH, K. The definition of El Niño. **Bulletin of the American Meteorological Society**, v. 78, p. 2771-2777, 1997.

VERA, C. S.; VIGLIAROLO, P. K. A diagnostic study of cold-air outbreaks over South America. **Mon Wea Rev**, v.128, p. 3-24, 2000.

VERA, C. S.; VIGLIAROLO, P. K.; BERBERY, E. H.. Cold Season Synoptic-Scale Waves over Subtropical South America. **Mon Wea Rev**, v. 130, p. 684-699, 2002.

YARNAL, B. **Synoptic Climatology in Environmental Analysis**, Belhaven Press, London, 1993. 195 p.

YARNAL, B.; COMRIE, A. C.; FRAKES, B.; BROWN, D. P. Developments and prospects in synoptic climatology. **Int. J. Climatol**., v. 21, n. 15, p. 1923-1950, 2001.