MODELOS ESTADÍSTICOS DE PREDICCIÓN ARIMA DE PRECIPITACIONES EN DOS ESTACIONES ESPAÑOLAS REPRESENTATIVAS DE DOS GRUPOS CON DIFERENTES CARACTERÍSTICAS CLIMÁTICAS

Esperanza AYUGA TÉLLEZ[1], Concepción GONZÁLEZ GARCÍA[1], Mª José MONTERO GARCÍA-ANDRADE[2], José Carlos ROBREDO SÁNCHEZ[3]
[1]*Departamento de Economía y Gestión Forestal. Escuela Técnica Superior de Ingenieros de Montes, Universidad Politécnica de Madrid. (UPM)*
[2]*Departamento de Construcción y Vías Rurales. Escuela Técnica Superior de Ingenieros Agrónomos. UPM*
[3]*Departamento de Ingeniería Forestal. Escuela Técnica Superior de Ingenieros de Montes. UPM*
esperanza.ayuga@upm.es, concepción.gonzalez@upm.es, mj.montero@upm.es, josecarlos.robredo@upm.es

RESUMEN:
Las series de datos pluviométricas, además de otras variables climáticas (temperaturas, insolación, etc.), desempeñan un papel muy importante en la caracterización del clima. Para su análisis con el fin de obtener modelos de predicción, las series de datos y, en este caso concreto, de precipitaciones mensuales deben haber sido observadas por un periodo de, al menos 20 a 30 años, lo que se llama un período largo de observación.
Mediante el empleo de técnicas estadísticas multivariantes se redujo la dimensionalidad del conjunto de datos de partida (distintos puntos de medición y un gran número de variables observadas en cada uno). Resultaron seleccionadas dos ciudades (Barcelona y Madrid) con estaciones de medición y, como variable más influyente, la precipitación media mensual. Estos datos han sido analizados mediante técnicas basadas en la teoría de los procesos estocásticos para la obtención de modelos tipo ARIMA de predicción.
Por otra parte, es bien conocida la importancia de estudiar y tratar de predecir precipitaciones máximas. El estudio de las precipitaciones pluviales extremas es fundamental para conocer posibles pérdidas de suelo por erosión y para un correcto dimensionamiento de estructuras de ingeniería civil que deben soportar grandes avenidas para garantizar la seguridad de construcciones y poblaciones afectadas. La precipitación máxima sigue modelos univariantes de probabilidad bien estudiados (distribución Gumbel, Square Root-Exponential Type Distribution of Maximum, por ejemplo), sin embargo en este trabajo se obtienen modelos de regresión en función de la variable "precipitación media mensual" que pueden contribuir a posibles predicciones de precipitaciones máximas en las poblaciones seleccionadas.

Palabras clave: Precipitaciones, Caracterización Climática, Procesos estocásticos, Modelos ARIMA.

ABSTRACT
Pluviometric data series and other climate variables (temperatures, hours of sunshine,…) play a very important role in the characterization of climate. To analyse them with the aim to obtain

forecast models, the data series and, in this case, monthly rainfall data, must have been measured during a long time period, at least 20 or 30 years (long record of data).
Using multivariate statistical techniques were reduced dimensionality of the starting data set (different measuring points and a large number of variables observed in each). Were selected two cities (Barcelona and Madrid) with measuring stations and, like most influential variable, the average monthly precipitation. These data series were analysed using techniques based on the theory of stochastic processes to obtain forecasting ARIMA models.
On the other hand, is well known the importance of studying and trying to predict maximum rainfall (Koutsoyiannis et al, 2000). Extreme rainfall analysis is basic to know feasible loss of soil by erosion and for a right design of the structures that have to support large rainfalls to guarantee the safety of the constructions and the affected population. It is well studied fitting of univariate probability models to maximum precipitation data (Gumbel distribution, Square Root-Exponential Type Distribution of Maximum for example). However in this paper are obtained regression models based on the variable "monthly average precipitation that may contribute to possible predictions of rainfall in selected populations.

Key words: Maximum Precipitations, Climatic Characterization, Stochastic Process, ARIMA Models

1. INTRODUCCIÓN

El agua es un bien indispensable para el hombre, siendo de gran importancia para la promoción del bienestar de la sociedad. El incremento de la población y el crecimiento tecnológico han potenciado la intensidad del empleo del agua y la variedad de sus usos. Esto ha hecho que aumente su importancia estratégica y ha puesto en evidencia la falta de medios para su gestión. Por esta razón, hay que potenciar las tareas de planificación que tiendan a su uso racional, preservando su cantidad y calidad en las distintas zonas de España.

Las precipitaciones son la fuente principal de entrada de agua en las cuencas. Cualquier proyecto de diseño de una estructura hidráulica (presa, balsa, encauzamiento, etc.) o análisis del territorio en relación con la hidráulica fluvial necesita de un conocimiento preciso y científico de las condiciones hidrológicas de la cuenca en que se realice ese estudio. Es necesario conocer las precipitaciones máximas que se producen en una cuenca para poder estimar las máximas avenidas asociadas a las mismas.

La información sobre precipitaciones máximas está asociada al tiempo de dos formas distintas, la primera hace referencia a la duración de esa precipitación máxima, mientras que la segunda se relaciona con el periodo de tiempo con que nos encontramos precipitaciones de esa magnitud máxima. Por esto la máxima lluvia probable en una cuenca es la mayor altura de agua que puede producirse durante un tiempo dado y en un periodo de tiempo prefijado denominado tiempo de retorno.

En una primera fase del estudio se emplean técnicas estadísticas multivariantes para seleccionar variables y reducir la dimensión del conjunto de variables y poblaciones a estudiar. Con las variables y ciudades seleccionadas, por el tipo de datos, tomados a lo largo del tiempo, se analizará si presentan dependencias entre ellos y por tanto se deberán aplicar técnicas derivadas de los procesos estocásticos. Un primer análisis descriptivo de las series consideradas, aporta información sobre la estructura de dependencias y, como suele ocurrir en

series de variables climáticas (AYUGA, GONZÁLEZ, ET AL. 1997, 2003, 2005, 2007), se estudia si son estacionarias, ya que pueden presentar tendencia (variaciones en la media) y estacionalidad (periodicidades cada cierto número de observaciones). Una vez analizada la estructura de dependencias a distintos intervalos de medida, mediante los correlogramas, se procederá a identificar algún modelo estocástico según la metodología BOX-JENKINS (1976). Por último, se ha considerado de interés estimar modelos de regresión que relacionen precipitaciones máximas y precipitaciones medias mensuales en las ciudades consideradas.

2. MATERIAL Y MÉTODO:

En este trabajo se ha contado con datos del Instituto Nacional de Meteorología (INM). Se ha partido de los datos disponibles de los últimos 30 años en las 52 ciudades españolas de 13 variables climáticas. Las variables a estudiar para el Análisis Multivariable han sido: altitud (A), temperatura media (T), temperatura media máxima (Tmax), temperatura media mínima (Tmin), precipitación media (R), humedad relativa (H), número medio de días con precipitación media mayor o igual a 1 mm (DR), número medio de días con nieve (DN), días de tormenta (DT), días de niebla (DF), días con heladas (DH), días despejados (DD) y media anual de horas de sol (I).

En las ciudades seleccionadas se analizan series de medias mensuales de precipitaciones. Como las series disponibles son datos tomados a intervalos regulares de tiempo y en un amplio periodo de observación, son susceptibles de ser tratadas mediante técnicas de series temporales: modelos ARIMA de Box-Jenkins (1976).

Los modelos ARIMA (de las iniciales en inglés, Autoregresive Integrated Moving Average) constan de una parte autorregresiva (AR, cada variable en un instante t de medición está relacionada con los valores hasta "p" instantes anteriores) y otra de media móvil (MA, el residuo o parte aleatoria de cada variable en un instante t de medición está relacionada con los residuos de los valores medidos hasta "q" instantes anteriores).

En la práctica las series de valores suelen presentar variaciones en la media (tendencia o estacionalidad) y en la varianza. Para conseguir una serie estacionaria y poder realizar el ajuste de un modelo ARMA se aplican diferencias entre valores de la serie de datos x_t ($x_t - x_{t-1}$, o estacional $x_t - x_{t-s}$, s = periodo estacional, según sea la variación observada) (MARTÍNEZ FALERO *et al.*, 1995). La inicial I del término "integrado" en la denominación ARIMA indica que, llamando z_t al proceso estacionario, x_t se obtiene como suma (integración) de z_t (PEÑA, 2005). Si se observan variaciones en la varianza conviene transformar los datos observados.

Para la realización del estudio se ha utilizado el programa informático STATGRAPHICS 5.1 que contiene sencillos análisis multivariantes y de series temporales (MARTÍN *et al*.2001).

3. RESULTADOS:

Dado el gran número de variables y estaciones de medición, se emplearon, en primer lugar, técnicas de Análisis Multivariante para reducir la dimensión del problema. Se realizó un Análisis de Componentes Principales (ACP) y un Análisis de Cluster con los que se determinaron dos variables y dos grupos de estaciones con características climatológicas

semejantes. Como estaciones representativas de cada uno de los grupos se escogieron Madrid y Barcelona por ser dos de las ciudades con mayor número de datos, más de 100 años de datos mensuales. Para el caso de las variables climáticas se han seleccionado la precipitación media (R) y la temperatura media (T).

Con el ACP se realizó la evaluación de las variables climáticas en las ciudades españolas, la valoración de la importancia de cada una de ellas y el análisis de la relación entre las mismas. El objetivo fue obtener un número más reducido de variables que resultan ser combinaciones lineales de las 13 variables iniciales. Las variables obtenidas explican la mayor parte de la variabilidad en los datos observados. En este caso resultaron las siguientes componentes:

Componente 1 = 0,26.A - 0,37.T - 0,37.Tmax - 0,33.Tmin + 0,21.R + 0,15.H + 0,30.DR +
+ 0,29.DN + 0,23.DT + 0,23.DF + 0,27.DH - 0,27.DD - 0,26.I

Componente 2 = - 0,37.A + 0,15.T + 0,03.Tmax + 0,24.Tmin + 0,38.R + 0,38.H + 0,27.DR -
- 0,26.DN - 0,04.DT + 0,20.DF - 0,36.DH - 0,27.DD - 0,34.I

Con estas dos variables queda explicada el 77,5% de la variabilidad en los datos originales.

Las ecuaciones de las componentes muestran la importancia del conjunto de las trece variables para resumir la información. No obstante, el mayor peso de las variables que miden las temperaturas en la *Componente 1* y de la humedad (H) y precipitación (R) en la *Componente 2*, permiten suponer que con una variable de cada grupo, se puedan agrupar las diferentes estaciones de ciudades españolas. Para verificar esto, se realiza un análisis de cluster por variables.

Con el análisis de cluster se obtuvieron los dendogramas correspondientes a los grupos representativos en el caso de variables climáticas (figura 1) y con la variable T y R se agruparon las ciudades españolas, empleando el método del vecino más lejano y la distancia euclídea (figura 2).

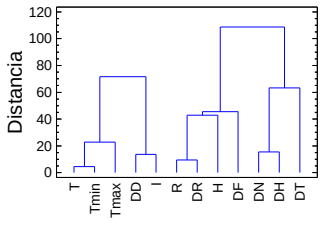

Fig. 1. Variables Climáticas

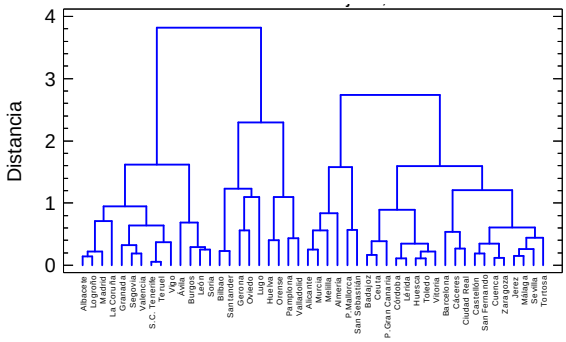

Figura 2. Ciudades Españolas

Para la validación de los grupos del cluster se aplicó un análisis de la varianza (ANOVA), con el objeto de comprobar que la diferenciación de grupos es aceptable. Para ello se compararon

los valores medios de T y R en los dos grupos resultantes del análisis de cluster. La diferencia entre ambos niveles era estadísticamente significativa para un nivel de confianza de un 95% (Figura 3).

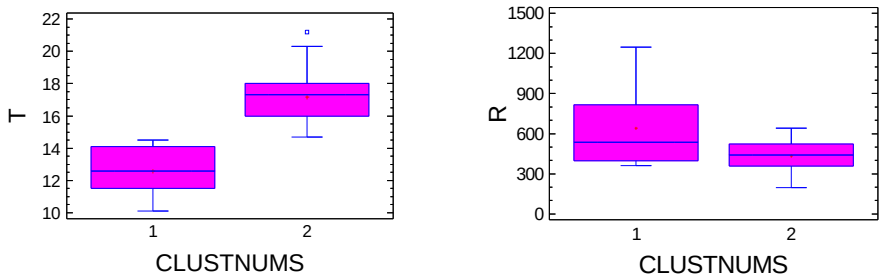

Fig. 3 y 4. Cajas para diferencia de medias (Análisis ANOVA)

Una vez seleccionadas las ciudades tipo (Madrid y Barcelona) y las series más representativas, se procedió a la aplicación de técnicas de series de tiempo a los datos de precipitaciones medias mensuales en dichas ciudades. Se observaron datos faltantes en ambas series de datos (2 en la de Madrid y 37 en la de Barcelona), para lo cual se realizaron estimaciones utilizando datos vecinos.

Se estimaron las autocorrelaciones, obteniéndose los Figuras correlogramas (función de autocorrelación simple, fas y función de autocorrelación parcial, fap), Figura 5 para la serie PME. En ambas series (Madrid y Barcelona) los correlogramas no indican tendencia (correlación en primeros retardos 1 y 2), pero si estacionalidad (correlaciones significativas en retardos 12, 24, 36). Por ello, se realizó una diferencia de orden estacional.

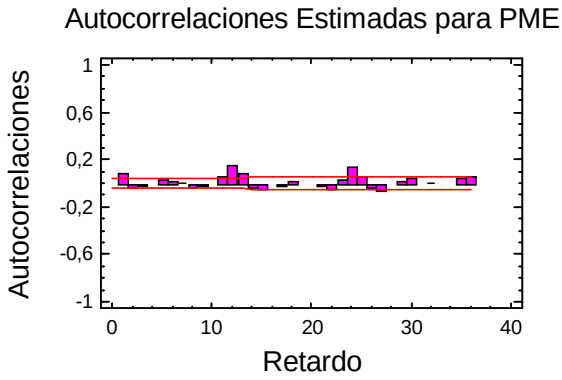

Fig. 5

Las dos series diferenciadas mostraron correlogramas similares (Figuras 6 y 7, sólo la de Madrid), con la diferencia de identificar un ARIMA (0,,0,1) $x(0,1,1)_{12}$ para la de Madrid y sólo uno estacional del tipo SARIMA $(0,1,1)_{12}$ para la de Barcelona.

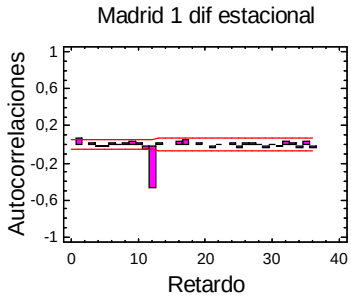

Fig. 6

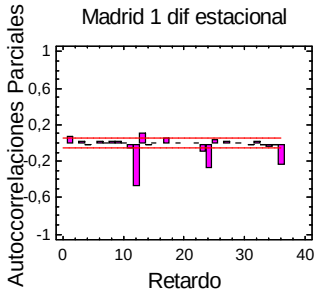

Fig. 7

Una vez estimados los modelos se comprobaron con una serie de contrastes para los residuos, que debian aparecer con valores no significativos ($p \geq 0,05$).

- Para excesivas ejecuciones arriba y abajo.
- Para excesivas ejecuciones por encima y por debajo de la mediana.
- De Box-Pierce para excesivas correlaciones.
- Para la diferencia en la media de la 1ª mitad a la 2ª mitad.
- Para la diferencia en la varianza en la 1ª mitad a la 2ª mitad.

Para Madrid, los residuos presentaban con alta significación el contraste último para la diferencia de varianzas entre la primera mitad de la serie y la segunda mitad. Esta anomalía sólo se puede eliminar transformando los datos iniciales. Por ello se eligió una transformación "raíz cuadrada", la logarítmica no fue posible por tener valores nulos en ambas series.

Las series transformadas mostraron un comportamiento similar a las originales en cuanto a la estructura de dependencias analizada mediante los correlogramas. Tanto las series transformadas iniciales como las series con una diferencia estacional permitieron identificar los mismos modelos ARIMA.

El ajuste para la serie transformada de la PME en Madrid resultó con todos los contrastes para los residuos no significativos, por lo que se admitió como válido el modelo. Se verificó su comportamiento en predicciones realizando la transformación inversa (elevando al cuadrado las predicciones proporcionadas por el modelo ajustado). De 12 meses pronosticados y comparados con los datos reales observados, resultaron dos meses con el valor real fuera de los límites de confianza obtenidos para las predicciones.

En el caso de la PME de Barcelona no fue posible ajustar un modelo sencillo que permitiera aceptar todos los contrastes para los residuos. Aún para la serie transformada, el SARIMA $(0,1,1)_{12}$, sólo daba tres contrastes aceptables, el de las diferencias de varianza resultaba ligeramente significativo (p-valor, próximo a 0,05 pero manteniéndose inferior) y el de "excesivas ejecuciones arriba y debajo de la mediana" indicaba falta de aleatoriedad.
El único modelo que se aceptaría sería el ajuste de una tendencia lineal y a los valores residuales resultantes ajustarles un ARIMA (0,0,1).

En resumen los modelos ARIMA ajustados a los valores de la raiz cuadrada de PME fueron:

Para la estación de Barcelona SARIMA $(0,1,1)_{12}$, con problemas en residuos.
Para la estación de Madrid: ARIMA(0,0,1)x$(0,1,1)_{12}$, válido para predicción.

A continuación se presentan algunos de los Figuras más representativos para la serie de Madrid transformada

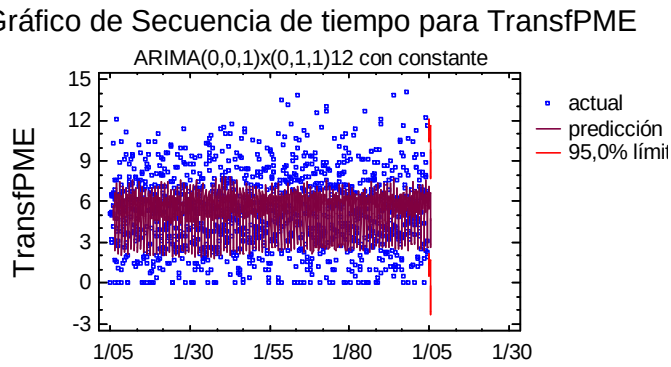

Fig. 8. Secuencias de tiempo ajustadas

En el caso de Madrid se verificó el comportamiento de las predicciones para la serie original (Tabla 1)

Las predicciones obtenidas para ese periodo, octubre/2003 a noviembre/2004, y sus límites, así como, la transformación de esos valores, elevando al cuadrado y su comparación con los datos reales indican que éstos se encuentran dentro de los límites de confianza de predicción al 95%, salvo en observaciones más extremas como se observa en la tabla.

LimlSupTransfP	LimInfTransf	PredicTransfPME	*PredPME Sup*	*PredPME Inf*	*Observados*	*Fecha*
11.5431	1.66403	6.60355	133.2431576	2.76899584	148.5	*Oct/2003
12.0084	2.08412	7.04625	144.2016706	4.34355617	79.7	
11.4066	1.4823	6.44443	130.1105236	2.19721329	44	
10.6627	0.738412	5.70054	113.6931713	0.54525228	5.9	
10.4195	0.495205	5.45734	108.5659803	0.24522799	64	
10.202	0.277687	5.23982	104.080804	0.07711007	60.5	
11.3302	1.4059	6.36803	128.373432	1.97655481	45.9	
11.4184	1.49416	6.45629	130.3798586	2.23251411	135.3	*May/2004
9.487	-0.437268	4.52486	90.003169	0.1912033	5	
7.8041	-2.12017	2.84196	60.90397681	4.49512083	9.4	
7.49053	-2.43374	2.52839	56.10803968	5.92309039	38.6	
9.7879	-0.136363	4.82577	95.80298641	0.01859487	3.7	

Tabla 1

El análisis para valorar la relación de dependencia de las precipitaciones máximas (P_{MA24}) respecto de las medias (P_{ME}) se realizó mediante un análisis de regresión. El p-valor obtenido

en el análisis ANOVA fue inferior a 0,01, por lo que quedó demostrada una relación significativa entre estas dos variables para un nivel de confianza del 99%.

La ecuación del modelo ajustado es:

En Barcelona: $P_{MA24} = 3{,}87144 + 0{,}362903 * P_{ME}$

En Madrid: $P_{MA24} = -2{,}69406 + 2{,}98282 * \sqrt{P_{ME}}$

El valor de R-cuadrado que se obtuvo indica que el modelo explica un 73% y un 63% en cada caso (Barcelona y Madrid) de la variabilidad en P_{MA24} y el coeficiente de correlación de ambos modelos indica una relación moderadamente fuerte entre las variables.

Con el test de Durbin-Watson, se examinaron los residuos para determinar si existía correlación significativa basada en el orden en que se introdujeron los datos. El resultado que se obtuvo para p-valor fue inferior a 0,05 en las dos series, lo que indicó un posible modelo dependiente del tiempo.

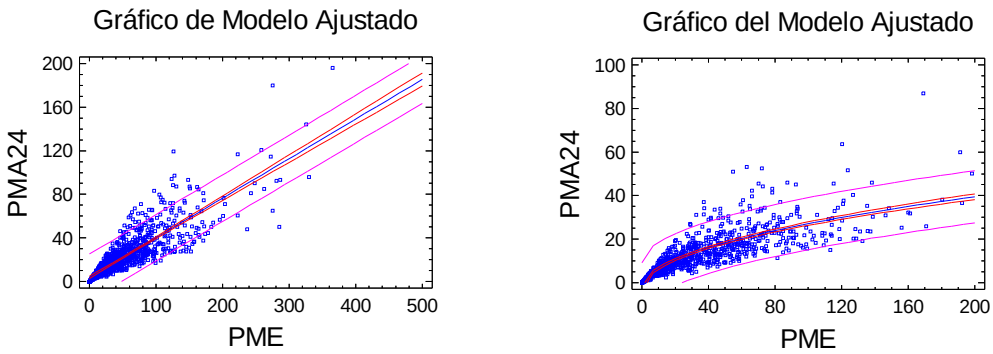

Fig. 9 y 10. Modelos Ajustados. Barcelona y Madrid.

4. CONCLUSIONES:

Las técnicas de análisis multivariante, ACP y el análisis de cluster han permitido seleccionar dos variables como representativas del grupo de variables climáticas iniciales, por su alta correlación (la R y la T) y, dos grupos de ciudades españolas por ciertas similitudes detectadas mediante el análisis de cluster: uno de los grupos reúne la mayoría de ciudades en la costa mediterránea, algunas del interior en la franja central (las dos de Extremadura y de Castilla – La Mancha), las Palmas de Gran Canaria, Ceuta y Melilla y próximas de la costa atlántica (Jerez, San Fernando), Vitoria y Lérida; en el otro grupo se recoge la mayoría de las ciudades en costa Atlántica y del Cantábrico, Sta. Cruz de Tenerife y de Castilla-León.

Con las series de precipitaciones medias mensuales) en las dos ciudades con periodo más amplio de datos (Madrid y Barcelona) se ha realizado un análisis de series de tiempo. Primero descriptivo mediante correlogramas para detectar regularidades frecuentes de encontrar en este tipo de datos (tendencias, estacionalidad). Se observó comportamiento estacional con periodo 12. Se tomó una diferencia estacional y, además fue necesario transformar los datos iniciales

para estabilizar la variación en la varianza de los datos a lo largo del tiempo. Los correlogramas de ambas permitieron identificar modelos ARIMA con parte estacional.

Para PME en Madrid el modelo de predicción resultó

$$z_t = z_{t-12} + a_t - \vartheta a_{t-1} - \Theta a_{t-12} \qquad \text{ARIMA (0, 0,1) x (0,1,1)}_{12}$$

Aceptable para su empleo en predicción.

Para Barcelona se obtuvo como modelo ARIMA más simplificado y aproximado:

$$z_t = z_{t-12} + a_t - \Theta a_{t-12} \qquad \text{SARIMA (0,1,1)}_{12}$$

No aceptable por problemas en residuos.

Los modelos de regresión entre PME y PMA24 son aceptables por su elevado R^2 pero también presentan problemas de significación las hipótesis para los residuos. Por ello tampoco resultan adecuados para la predicción.

5. REFERENCIAS:

AYUGA, E.; GONZÁLEZ, C.; MARTINEZ, J.E.; PARDO, M.; SOLANA, J. (1997). *Análisis estadístico de algunas series pluviométricas de la provincia de Huelva*. Comunicación en actas Congreso, 298 – 308. 5th Meeting of the International Biometric Society Network for Central America, the Caribbean, Mexico, Colombia and Venezuela. Facultad de Estadística e Informática. Universidad Veracruzana, Mexico; Instituto Nacional de Estadística, Geografía e Informática. México.

AYUGA, E.; GONZÁLEZ, C. (2003). *Modelos estadísticos para las series pluviométricas de Huelva. Validación y actualización.* Comunicación en actas del III Congreso de Agroingeniería. Universidad de Córdoba.

AYUGA TÉLLEZ, E.; GONZÁLEZ GARCÍA, C.; ROBREDO SÁNCHEZ, J.C.; MARTÍN FERNÁNDEZ, A.J. Y GRANDE ORTIZ, M.A. (2005). *Modelos estocásticos de pluviometría y temperaturas medias mensuales en España*. Póster, en actas del IV Congreso Forestal Español. Zaragoza (España).

AYUGA TÉLLEZ, E.; GONZÁLEZ GARCÍA, C.; MONTERO GARCÍA-ANDRADE, M.J. ; RAMIREZ GÓMEZ, D.A.; GARCÍA GARCÍA, A.I.; GARCÍA MONTERO, L. G. (2007). *Estimación de la Evolución de Temperaturas Medias para Detectar Cambios en el Clima Español*. IV Congreso Nacional de Evaluación de Impacto Ambiental (IV CONEIA). Libro actas (en imprenta). Madrid.

BOX, G.E.P. y JENKINS G.M. (1976). *Time Series Analysis, Forecasting and Control*, Holden-Day, San Francisco.

KOUTSOYIANNIS, D; BALOUTSOS, G. (2000). *Analysis of a Long Record of Annual Maximum Rainfall in Athens, Greece, and Design Rainfall Inferences.* Natural Hazards, Volume 22, Number 1, July 2000 , pp. 29-48(20). Springer.

MARTÍNEZ FALERO, E.; GONZALEZ, C. y AYUGA, E. (1995). Techniques for Spatial and Temporal Analysis, en *Quantitative Techniques in Landscape Planning*. 137-189; CRC Press, Inc.., Florida**.**

PEÑA, D. (2005). *Análisis de Series Temporales.* Alianza Editorial, Madrid.

JONES, D.A.; SVENSSON, C.; STEWART, E. J.. (2007). *Use of values of maximum rainfall for incomplete records.* Project note to Defra, project WS194/2/39 (Reservoir Safety - Long

Return Period Rainfall. Wallingford, Centre for Ecology and Hydrology, 14pp. (CEH Project Number: C02760)

MARTÍN FERNÁNDEZ, S., AYUGA TÉLLEZ, E., GONZÁLEZ GARCÍA, C., MARTÍN FERNÁNDEZ, A. (2001). *Guía completa de Statgraphics. Desde MS-DOS a Statgraphics Plus.* Díaz de Santos, Madrid.

http://www.aemet.es/es/servicios/publicaciones/novedades/Estudio_sobre_precipitaciones