ANÁLISIS DE LAS PRUEBAS DE DETECCIÓN DE VALORES ANORMALMENTE EXTREMOS (OUTLIERS) EN SERIES HIDROLÓGICAS

Ricardo A. Smith y Claudia P. Campuzano
Posgrado en Aprovechamiento de los Recursos Hidráulicos
Facultad de Minas, Universidad Nacional de Colombia, Sede Medellín
cpcampuz@andromeda.unalmed.edu.co

RESUMEN

Se presentan un conjunto de pruebas para detección de valores anormalmente extremos (Outliers) en series hidrológicas. Se hace un análisis de la potencia de estas pruebas utilizando series sintéticamente generadas con las distribuciones Normal, Gumbel y Weibull, para diferentes números de puntos anormales, y tanto para puntos anormales por encima como por debajo de la media. Igualmente se presentan algunos resultados utilizando información histórica, tratando de analizar la relación de ocurrencia de puntos anormales con la ocurrencia del fenómeno de El Niño – Oscilación del Sur (ENSO) en sus dos fases (El Niño y La Niña). Se presentan finalmente algunas conclusiones y recomendaciones.

PALABRAS CLAVES: Puntos extremos, Puntos anormales, Potencia, Pruebas estadísticas, Generación sintética, El Fenómeno de El Niño

ABSTRACT

Several tests for dtectiion of puntos anormales in hidrological time series are presented. A power analysis of those tests using synthetic series is also presented. The Normal, Gumbel and Weibull distributions and puntos anormales of different charateristics (size, location and above or below the mean) were used for the power analysis. Some results using historical series, trying to analyze the influence of ENSO over puntos anormales, are also presented. Some conclusions and recomendations are finally presented.

KEY WORDS: Outliers, Power analysis, Statistical tests, Synthetic series, ENSO.

1. INTRODUCCION

Un punto anormalmente extremo (Outlier) es una observación alejada anormalmente del comportamiento general de las observaciones en una serie. Estos puntos pueden ser causados por errores de medida, errores de transcripción, averías de los instrumentos, y problemas de calibración. También pueden indicar mayor variabilidad espacial o temporal que la esperada. Los puntos anormalmente extremos (Outliers) dan la impresión de ser muestras de una distribución distinta al resto de las observaciones, parecen no ser representativos de la muestra. En los procesos de estimación estadística los puntos anormales a menudo llevan a resultados sesgados y escogencia de distribuciones inadecuadas. Se acostumbra entonces en hidrologia, antes de los procesos de estimación de las distribuciones, detectar si la muestra disponible tiene puntos anormales, y si así es, proceder a la remoción de los mismos.

En los análisis hidrológicos que normalmente se hacen en Colombia, no se usan pruebas de detección de puntos anormalmente extremos, ni siquiera pruebas de homogeneidad y consistencia de las series hidrológicas.

Estos análisis son de primera importancia, pues si las series tiene puntos anormalmente extremos o tienen no homogeneidades (cambios o tendencias) y se ignoran, los análisis hidrológicos pueden llevar a conclusiones equivocadas.

Las pruebas de detección de puntos anormalmente extremos comúnmente utilizadas en hidrología se aplican para rangos de tamaños de muestras diferentes, lo cual hace difícil su comparación. Se espera que estas pruebas tengan diferente potencia (en el sentido estadístico) y sería conveniente analizarla para concluir sobre su confiabilidad en aplicaciones hidrológicas. Se entiende por potencia de una prueba estadística la probabilidad de que la hipótesis nula sea rechazada cuando es verdadera (Mood, Graybill y Boes, 1974).

Se presenta a continuación un análisis experimental de la potencia de las pruebas de detección de puntos anormales comúnmente utilizadas en hidrología basada en generación de series sintéticas. Igualmente se presentan algunas conclusiones y recomendaciones basadas en estos experimentos.

Se hace también un análisis de la aplicabilidad de estas pruebas en nuestro medio utilizando información histórica. Se intenta encontrar si hay relación entre la ocurrencia de puntos anormalmente extremos y de el fenómeno El Niño – Oscilación del Sur (ENSO).

2. PRUEBAS DE DETECCION DE PUNTOS ANORMALMENTE EXTREMOS (OUTLIERS)

Diferentes pruebas han sido propuestas para detectar puntos anormalmente extremos (Outliers) y se encuentran referenciadas en algunos libros de estadística aplicada a la hidrología (Gilbert, 1987; Gibbons, 1994; Kottegoda y Rosso, 1997; Helsel y Hirsch, 1992) y otros en la literatura sobre el tema (Tietjen y Moore, 1972). Todas esas pruebas han sido desarrolladas básicamente para probar la hipótesis nula de que todas las observaciones fueron tomadas de poblaciones igualmente distribuidas. Si no se rechaza la hipótesis nula, entonces se asume que la muestra no tiene puntos anormales. Las pruebas propuestas en general se pueden clasificar en dos grupos:

- Basados en distribuciones
- Basados en cercanía

Las pruebas basadas en distribuciones prueban si la distribución de la muestra con y sin el punto o los puntos supuestamente puntos anormales es la misma. Si es la misma se acepta la hipótesis de que la muestra no tiene puntos anormales. Las pruebas basadas en cercanías prueban si el punto supuestamente punto anormal esta lo suficientemente alejado de la siguiente observación más cercana a ese punto. Si no lo esta se acepta la hipótesis de que la muestra no tiene puntos anormales.

Las diferentes pruebas disponibles se aplican para diferentes rangos de tamaño de la muestra. Algunas de ellas solo pueden detectar un punto anormal en la muestra y otras pueden detectar varios. A continuación se describen brevemente estas pruebas.

2.1. Pruebas basadas en Cercanías

- **Prueba de Dixon**

Esta es una prueba propuesta para un grupo de observaciones con un número pequeño de observaciones que se asumen fueron tomadas de una población normalmente distribuida con media y varianza desconocidas. Se recomienda que primero se identifique el grupo de observaciones que se sospecha pueden ser anormales. Si M representa el número de observaciones que se sospecha pueden ser anormales, entonces la prueba de Dixon podría ser aplicada para todas las M observaciones. Una vez se identifica un punto anormal usando la prueba este es removido del grupo de observaciones y se puede proceder a probar si hay otro punto anormal. Esta prueba puede usarse tanto para detectar puntos anormales en la dirección de los máximos como en la dirección de los mínimos.

Sea Y_j $j=1,\ldots,N$ un grupo de N observaciones y W_j $j=1,\ldots,N$ son las mismas observaciones pero organizadas en orden creciente de magnitud $(W_1 < W_2 < \ldots < W_N)$. La prueba estadística de Dixon está definida de acuerdo al rango del tamaño de la muestra N y probando si el valor más alto o más bajo en el grupo de observaciones es un punto anormal de acuerdo con la siguiente Tabla 1 (Gibbons, 1994).

TABLA 1. Estadístico para la Prueba de Dixon

N	Highest Value	Lowest Value
3-7	$D = \frac{W_N - W_{N-1}}{W_N - W_1}$	$D = \frac{W_2 - W_1}{W_N - W_1}$
8-10	$D = \frac{W_N - W_{N-1}}{W_N - W_2}$	$D = \frac{W_2 - W_1}{W_{N-1} - W_1}$
11-13	$D = \frac{W_N - W_{N-2}}{W_N - W_2}$	$D = \frac{W_3 - W_1}{W_{N-1} - W_1}$
14-25	$D = \frac{W_N - W_{N-2}}{W_N - W_3}$	$D = \frac{W_3 - W_1}{W_{N-2} - W_1}$

La hipótesis nula de que la observación más alta o más baja es un punto anormal se rechaza si

$$d \leq C(N, \alpha) \qquad (1)$$

en donde d es una estimado de la prueba estadística D dado en la Tabla anterior y $C(N, \alpha)$ es el valor crítico de la prueba para una muestra de tamaño N y un nivel de significancia a. Tablas para determinar $C(N, \alpha)$ están disponibles en Gibbons (1994).

- **Pruebas Estadísticas Em y Lm**

Las pruebas de los estadísticos Em y Lm pueden usarse para probar la existencia de M puntos anormales en un grupo de observaciones que se asume fue tomado de una población normalmente distribuida con media y varianza desconocidas. Se recomienda primero identificar el grupo de M observaciones sospechosas de ser puntos anormales, pudiendo aplicarse la prueba para diferentes valores de M. La prueba del estadístico Lm es una prueba de hipótesis de dos lados y se puede usar para identificar puntos anormales por encima o por debajo de la media. La prueba del estadístico Em puede usarse para detectar si las M observaciones más extremas (por encima o por debajo de la media) en una muestra son puntos anormales.

Sea Y_j $j=1,\ldots,N$ un grupo de N observaciones. Para calcular el estadístico para la prueba L_M, inicialmente la serie se organiza en orden creciente de magnitud (si se están detectando puntos anormales por encima de la media) o en orden decreciente de magnitud (si se están detectando puntos anormales por debajo de la media). W_j $j=1,\ldots,N$ representa la serie ordenada.

El estadístico para la prueba L_M, para puntos anormales por encima de la media, está definido como (Tiejten y Moore, 1972):

$$L_M = \frac{\sum_{j=1}^{N-M} \left(Y_j \; \mu_Y^{(N-M)}\right)^2}{\sum_{j=1}^{N} \left(Y_j \; \mu_Y\right)^2} \qquad (2a)$$

En donde μ_Y es la media muestral y $\mu_Y^{(N-M)}$ es la media sin las M observaciones más cercanas.

Y para puntos anormales por debajo de la media:

$$L_M = \frac{\sum_{j=M+1}^{N} \left(Y_j \; \mu_Y^{(N-M)}\right)^2}{\sum_{j=1}^{N} \left(Y_j \; \mu_Y\right)^2} \qquad (2b)$$

La hipótesis nula de que el grupo de M observaciones en el grupo de observaciones son puntos anormales se rechaza al nivel de significancia a si (Tietjen y Moore, 1972):

$$l_M \; £ \; L(N, M, \alpha) \qquad (3)$$

en donde l_M es una estimado del estadístico L_M de la ecuación (4), $L(N, M, \alpha)$ es el valor crítico de la prueba dado en función del tamaño de la muestra N, del número M de puntos anormales que se esta probando, y nivel de significancia a. Las Tablas para determinar $L(N, M, \alpha)$ se encuentran en Tiejten y Moore (1972).

Para determinar el estadístico para la prueba E_M inicialmente se determina la serie $\tilde{Y}_j$ de las desviaciones con respecto a la media en valor absoluto y luego se define la serie W_j dada por la serie $\tilde{Y}_j$ ordenada en orden creciente de magnitud $(W_1 < W_2 < \ldots < W_N)$. Una nueva serie Z_j se define, en donde Z_j es igual a la *Y-ava*

observación cuyo $\tilde{Y}_i$ representa el i-ésimo valor más grande. Usando la serie Z_j se calcula la media de la muestra usando todas las observaciones μ_Z y usando las primeras N- M observaciones, $\mu_Z^{(N-M)}$. El estadístico para la prueba E_M se define entonces como (Tietjen y Moore, 1972):

$$E_M = \frac{\sum_{j=1}^{N-M} \left(Z_j - \mu_Z^{(N-M)}\right)^2}{\sum_{j=1}^{N} \left(Z_j - \mu_Z\right)^2} \qquad (4)$$

La hipótesis nula de que el grupo de M observaciones en el grupo de observaciones son puntos anormales se rechaza al nivel de significancia a si (Tietjen y Moore, 1972):

$$e_M \leq E\,(N, M, \alpha) \qquad (5)$$

en donde e_M es un estimado del estadístico E_M de la ecuación (10) y $E\,(N, M, \alpha)$ es el valor crítico de la prueba en función del tamaño de la muestra N, del número M de puntos anormales y del nivel de significancia a. Las Tablas para determinar $E\,(N, M, \alpha)$ se encuentran igualmente en Tietjen y Moore (1972).

2.2. Pruebas basadas en Distribuciones

• Prueba de la Distribución Normal

Pruebas de normalidad realizadas utilizando grupos de observaciones con puntos anormales en general rechazan la hipótesis de que los datos fueron tomados de una población normalmente distribuida. Una prueba para detectar puntos anormales puede basarse entonces en probar repetidamente la hipótesis de normalidad cuando las observaciones que se sospechan son puntos anormales se remueven de manera secuencial del grupo de observaciones. Cualquiera de las pruebas de normalidad existentes en la literatura se puede usar con este propósito.

Sea Y_j $j = 1, N$ un grupo de N observaciones. Se puede entonces realizar una prueba de normalidad utilizando esta muestra y si no se rechaza la hipótesis de normalidad, entonces se rechaza la hipótesis nula de puntos anormales en el grupo de observaciones. Si se rechaza la hipótesis de normalidad, entonces se acepta la hipótesis de que hay puntos anormales en el grupo de observaciones. Las pruebas de normalidad se aplican comenzando con $i=1$. Si no se rechaza la hipótesis de normalidad sobre las N -1 observaciones restantes, entonces hay solo un punto anormal en el grupo de observaciones. Si se rechaza, hay más de un punto anormal en el grupo de observaciones y en este caso las pruebas de normalidad se aplican considerando $i=2$. Si no se rechaza la hipótesis de normalidad sobre las N-2 observaciones restantes, entonces hay dos puntos anormales en el grupo de datos. Este procedimiento continúa hasta que no se rechaza la hipótesis de normalidad. Por lo tanto, si no se rechaza la prueba de normalidad para $i = k$ significa que hay k puntos anormales en el grupo de observaciones.

• Prueba de las Desviaciones Normalizadas

La prueba de las Desviaciones Normalizadas es una prueba de detección de múltiples puntos extremos. Se asume que el grupo de observaciones se tomó de una población normalmente distribuida. La hipótesis nula es que la distribución de la población de donde se tomaron las observaciones que se sospecha son anormales es la misma distribución normal de la población de donde se obtuvo el resto de las observaciones de la muestra. La prueba requiere que se especifique, antes de realizar la misma, un número superior M de puntos anormales potenciales en el grupo de observaciones.

Sea Y_j $j = 1, \ldots, N$ un grupo de N observaciones y se sospecha que hay M puntos anormales dentro del grupo de datos. Primero el grupo de observaciones se organiza en orden creciente de magnitud y se define una nueva serie W_j tal que $W_1 < W_2 < \ldots < W_N$. $\mu_W^{(N-i+1)}$ y $\sigma_W^{(N-i+1)}$ representan la media y la desviación estándar del grupo W estimadas usando las primeras N-$i+1$ observaciones.

El estadístico para la prueba se define como (Kottegoda y Rosso, 1997, p .322)

$$S_i = \frac{W_{N-i+1} - \mu_W^{(N-i+1)}}{\sigma_W^{(N-i+1)}} \qquad (6)$$

La hipótesis nula de que hay i puntos anormales en el grupo de M observaciones que se sospechan que son puntos anormales del grupo de observaciones se rechaza si:

$$|s_i| \leq B_{Mi}\,(N, \alpha) \qquad (7)$$

en donde s_i es un estimado de la prueba estadística S_i de la ecuación (14) y $B_{Mi}\,(N, \alpha)$ es el valor crítico para la prueba del i-ésimo punto anormal en el grupo de M observaciones que se sospechan que son puntos

anormales, dado en función del tamaño de la muestra N, y del nivel de significancia a para la prueba. Tablas para encontrar $B_{M,i}(N, a)$ se encuentran en Kottegoda y Rosso (1997, p .699). La prueba de las Desviaciones Normalizadas descrita es para puntos anormales por encima de la media. En este caso $\mu_W^{(N-i+1)}$ y $\sigma_W^{(N-i+1)}$ se calculan usando los $N-i+1$ valores de la serie W_i. La prueba estadística S_i de la ecuación (6) se define usando W_i en lugar de W_{N-i+1}, y la prueba es desarrollada usando la ecuación (7) en la misma manera como se expone arriba.

- **La Prueba de Roesner**

Esta es una prueba de las llamadas múltiples puntos extremos ya que permite detectar hasta diez puntos anormales en un grupo de datos. La prueba es válida para tamaños de muestra mayores o iguales a 25 observaciones. El procedimiento asume que las observaciones se tomaron de una población normalmente distribuida. La prueba trata de evitar el enmascaramiento de un punto anormal por otro cuando están relativamente cercanos.

Antes de realizar la prueba se necesita especificar un número M de puntos anormales potenciales presentes en el grupo de observaciones. La hipótesis nula es que todo el grupo de observaciones representa una muestra tomada de una distribución normal y la hipótesis alternativa es que hay M puntos anormales, o M - 1 puntos anormales, o....., o 1 punto extremo (Gilbert, 1987). La prueba de Roesner es una prueba de hipótesis de dos lados e identifica puntos anormales por encima o por debajo de la media.

Sea Y_j, $j = 1,\ldots, N$ un grupo de observaciones donde N es el número total de observaciones. Se sospecha que hay M puntos anormales en el grupo. Sean $\mu^{(i)}$ y $\sigma^{(i)}$ la media estimada y la desviación estándar estimada de las $N - i$ observaciones del grupo de observaciones después de que las i observaciones más extremas (en la dirección que interesa hacer la prueba) han sido removidas (Gilbert, 1987; Gibbon, 1994).

$Y^{(i)}$ representa la observación más extrema (la más alejada de la media $\mu^{(i)}$) en el grupo de observaciones después de que las i observación más extrema han sido removidas de ese grupo. El estadístico para la prueba Rosner se define entonces como:

$$R_{i+1} = \frac{Y^{(i)} - \mu^{(i)}}{\sigma^{(i)}} \tag{8}$$

La hipótesis nula de que hay i + 1 puntos anormales en el grupo de observaciones se rechaza sí:

$$|r_{i+1}| \leq \lambda_{i+1}(N, \alpha) \tag{9}$$

en donde r_{i+1} es un estimado de la prueba estadística R_{i+1} de la ecuación (18) y $l_{i+1}(N, \alpha)$ es el valor crítico para la prueba en función del número de puntos anormales $i+1$, el tamaño de la muestra N, y el nivel de significancia a. Las Tablas para determinar $l_{i+1}(N, \alpha)$ están disponibles en Gilbert (1987) o Gibbon (1994).

3. ANÁLISIS EXPERIMENTAL DE LA POTENCIA DE LAS PRUEBAS

Para realizar el análisis experimental de la potencia de las pruebas, se generaron 100 series sintéticas con las distribuciones Normal, Weibull y Gumbel de 5, 10, 15, 20, 25, 30, 35, 40, 45, 50, 100, 500, 1000 datos. Las series normales se generaron utilizando un método basado en la aproximación del teorema del límite central (Salas, et al, 1992). Las series Weibull se generaron utilizando la ecuación (Salas et al, 1992):

$$X = x_o + \alpha\left[-ln(1.0 - U)\right]^{1/\beta} \tag{10}$$

Donde x_o es el parámetro de ubicación, α es el parámetro de escala y β es el parámetro de forma de la distribución de Weibull y U es un número aleatorio uniforme (0,1). Las series Gumbel se generaron utilizando la inversa de la FAP (Función Acumulada de Probabilidades) de la distribución Gumbel con parámetros x_o y α (Salas, et al, 1992). Así, para un número aleatorio uniforme U(0,1) se tiene:

$$X = x_o - \alpha\, ln\left[-ln U\right] \tag{11}$$

el cual es un número aleatorio distribuido Gumbel.

Una vez generadas las series sintéticas se procedió a incluir los puntos anormales en las mismas. Con este objetivo se introdujeron puntos anormales en esas series de 1 hasta 10 puntos anormales dependiendo del tamaño de las series. Para determinar el tamaño de los puntos anormales se usó la expresión m + σ K, variando el valor de K hasta 10, y utilizando la media y la desviación típica estimadas con los datos de la serie.

De acuerdo con las pruebas realizadas se concluyó inicialmente que se acepta la presencia de puntos anormales cuando K≥ 4. En este análisis se hicieron extensivas pruebas con las series generadas variando el tamaño de la muestra N, el valor de K, el número de puntos anormales M y la distribución de la serie sintética.

A manera de ejemplo, a continuación se presentan solo algunos resultados, ya que no es posible presentarlos todos acá. Los resultados se presentan en forma de tablas en donde se indica el porcentaje de aciertos de las pruebas. En las tablas que se presentan, R corresponde a la prueba de Roesner, SD a la prueba de las Desviaciones Normalizadas, N a la prueba Normal, D a la prueba Dixon y Em y Lm a las pruebas de los estadísticos Em y Lm. Por ejemplo, en la Tabla 2 se muestra para K=5 en el cuadro de 0 puntos anormales detectados, la columna de Em y para 10 datos un valor de 21. Esto significa que con la prueba del estadístico Em para las series de 10 observaciones no se detectaron puntos anormales en el 21% de los casos y si se detectaron (en el cuadro de 1 punto anormal detectado) en el 79% de los casos.

En la Tabla 2 se presentan los resultados para series de una distribución Normal con M=1, analizando puntos anormales por encima de la media y K=5 y K=10. La Tabla 3 presenta los mismos resultados para la distribución Gumbel y la Tabla 4 para la distribución Weibull. La Tabla 5 es para el mismo caso pero para puntos anormales por debajo de la media. La Tabla 6 presenta los resultados para series sacadas de una distribución Gumbel con M=4, analizando puntos anormales por encima de la media y K=6, 7, 9 y 10 (tamaño de los cuatro puntos anormales). En este caso se presentan los resultados solo para las detecciones de 2, 3 y 4 puntos anormales. La Tabla 7 presenta los resultados para series tomadas de la distribución Weibull con M=10, analizando puntos anormales por encima de la media y K=5, 5, 6, 6, 6, 9, 9, 9, 10 y 10 (tamaño de los puntos anormales). En este caso se presentan los resultados solo para las detecciones de 0,1 y 10 puntos anormales.

TABLA 2. Porcentaje de aciertos para las series con distribución Normal, con M=1 y analizando para puntos anormales por encima de la media

K=5	0 puntos anormales detectados							1 punto anormal detectado					
Datos	N	D	Em	Lm	R	SD		N	D	Em	Lm	R	SD
5	0	43	64	41				100	57	36	59		
10	0	19	21	7				100	81	79	93		
15	0	6	5	1				100	94	95	99		
20	0	1	2	0				100	99	98	100		
25	0	2	1	0	1	0		100	98	99	100	99	100
30	0		0	0	0	0		100		100	100	100	100
35	0		0	0	0	0		100		100	100	100	100
40	0		0	0	0	0		100		100	100	100	100
45	0		0	0	0	0		100		100	100	100	100
50	0		0	0	0	0		100		100	100	100	100
100	0				0	0		100				100	100
500	0				0			100				100	
1000	0				0			100				100	

Continuación **Tabla 2**

K=10

	0 puntos anormales detectados						1 punto anormal detectado					
Datos	N	D	Em	Lm	R	SD	N	D	Em	Lm	R	SD
5	0	0	6	0			100	100	94	100		
10	0	0	0	0			100	100	100	100		
15	0	0	0	0			100	100	100	100		
20	0	0	0	0			100	100	100	100		
25	0	0	0	0	0	0	100	100	100	100	100	100
30	0		0	0	0	0	100		100	100	100	100
35	0		0	0	0	0	100		100	100	100	100
40	0		0	0	0	0	100		100	100	100	100
45	0		0	0	0	0	100		100	100	100	100
50	0		0	0	0	0	100		100	100	100	100
100	0				0	0	100				100	100
500	0				0		100				100	
1000	0				0		100				100	

TABLA 3. Porcentaje de aciertos para las series con distribución Gumbel, con M=1 y analizando para puntos anormales por encima de la media

K=5

	0 puntos anormales detectados						1 punto anormal detectado					
Datos	D	Em	Lm	N	R	SD	D	Em	Lm	N	R	SD
5	62	75	63	0			38	25	37	100		
10	65	58	42	0			35	42	58	100		
15	50	50	34	0			50	50	66	100		
20	49	46	30	0			51	54	70	100		
25	41	41	25	0	35	25	59	59	75	100	65	75
30		36	18	0	36	24		64	82	100	64	76
35		35	16	0	33	18		65	84	100	67	82
40		39	24	0	33	16		61	76	100	67	84
45		35	18	0	30	18		65	82	100	70	82
50		33	18	0	30	19		67	82	100	70	81
100				0	25	15				100	75	85
500				0	14					100	86	
1000				0	4					100	96	

K=10

	0 puntos anormales detectados						1 punto anormal detectado					
Datos	D	Em	Lm	N	R	SD	D	Em	Lm	N	R	SD
5	14	32	11	0			86	68	89	100		
10	2	1	0	0			98	99	100	100		
15	0	0	0	0			100	100	100	100		
20	0	0	0	0			100	100	100	100		
25	0	0	0	0	0	0	100	100	100	100	100	100
30		0	0	0	0	0		100	100	100	100	100
35		0	0	0	0	0		100	100	100	100	100
40		0	0	0	0	0		100	100	100	100	100
45		0	0	0	0	0		100	100	100	100	100
50		0	0	0	0	0		100	100	100	100	100
100				0	0	0				100	100	100
500				0	0					100	100	
1000				0	0					100	100	

TABLA 4. Porcentaje de aciertos para las series con distribución Weibull, con M=1 y analizando para puntos anormales por encima de la media

K=5

	0 puntos anormales detectados						1 punto anormal detectado					
Datos	N	D	Em	Lm	R	SD	N	D	Em	Lm	R	SD
5	0	4	13	2			100	96	87	98		
10	0	0	0	0			100	100	100	100		
15	0	0	0	0			100	100	100	100		
20	0	0	0	0			100	100	100	100		
25	0	0	0	0	0	0	100	100	100	100	100	100
30	0		0	0	0	0	100		100	100	100	100
35	0		0	0	0	0	100		100	100	100	100
40	0		0	0	0	0	100		100	100	100	100
45	0		0	0	0	0	100		100	100	100	100
50	0		0	0	0	19	100		100	100	100	81
100	0				0	0	100				100	100
500	0				0		100				100	
1000	0				0		100				100	

K=10

	0 puntos anormales detectados						1 punto anormal detectado					
Datos	N	D	Em	Lm	R	SD	N	D	Em	Lm	R	SD
5	0	0	0	0			100	100	100	100		
10	0	0	0	0			100	100	100	100		
15	0	0	0	0			100	100	100	100		
20	0	0	0	0			100	100	100	100		
25	0	0	0	0	0	0	100	100	100	100	100	100
30	0		0	0	0	0	100		100	100	100	100
35	0		0	0	0	0	100		100	100	100	100
40	0		0	0	0	0	100		100	100	100	100
45	0		0	0	0	0	100		100	100	100	100
50	0		0	0	0	0	100		100	100	100	100
100	0				0	0	100				100	100
500	0				0		100				100	
1000	0				0		100				100	

TABLA 5. Porcentaje de aciertos para las serie con distribución Normal, con M=1 y analizando para puntos anormales por debajo de la media

K=5

	0 puntos anormales detectados						1 punto anormal detectado					
Datos	N	D	Em	Lm	R	SD	N	D	Em	Lm	R	SD
5	0	43	64	41			100	57	36	59		
10	0	19	21	7			100	81	79	93		
15	0	6	5	1			100	94	95	99		
20	0	1	2	0			100	99	98	100		
25	0	2	1	0	1	0	100	98	99	100	99	100
30	0		0	0	0	0	100		100	100	100	100
35	0		0	0	0	0	100		100	100	100	100
40	0		0	0	0	0	100		100	100	100	100
45	0		0	0	0	0	100		100	100	100	100
50	0		0	0	0	0	100		100	100	100	100
100	0				0	0	100				100	100
500	0				0		100				100	
1000	0				0		100				100	

Continuación **Tabla 5.**

K=10

	0 puntos anormales detectados						1 punto anormal detectado					
Datos	N	D	Em	Lm	R	SD	N	D	Em	Lm	R	SD
5	0	0	6	0			100	100	94	100		
10	0	0	0	0			100	100	100	100		
15	0	0	0	0			100	100	100	100		
20	0	0	0	0			100	100	100	100		
25	0	0	0	0	0	0	100	100	100	100	100	100
30	0		0	0	0	0	100		100	100	100	100
35	0		0	0	0	0	100		100	100	100	100
40	0		0	0	0	0	100		100	100	100	100
45	0		0	0	0	0	100		100	100	100	100
50	0		0	0	0	0	100		100	100	100	100
100	0				0	0	100				100	100
500	0				0		100				100	
1000	0				0		100				100	

TABLA 6. Porcentaje de aciertos para las series Gumbel, con M=4 y analizando para puntos anormales por encima de la media

K = (6,7,9,10)

	2 puntos anormales detectados						3 puntos anormales detectados						4 puntos anormales detectados					
N	R	SD	N	D	Em	Lm	R	SD	N	D	Em	Lm	R	SD	N	D	Em	Lm
10			0	0	0	0			0	0	0	0			0	0	0	0
15			1	0	0	0			0	5	0	0			83	70	42	83
20	0	0	6	0	0	5	0	0	4	5	4	4	100	100	90	77	88	90
25	0	0	2	0	1	2	0	0	5	8	5	5	100	100	92	78	92	92
30	0	0	1		1	1	0	0	2		2	2	0	0	97		97	97
35	0	0	1		1	1	0	0	3		3	3	0	0	96		96	96
40	0	0	0		0	0	0	0	3		3	3	0	0	97		97	97
45	0	0	0		0	0	0	0	2		2	2	0	0	98		98	98
50	0	0	0		0	0	0	0	1		1	1	0	0	99		99	99
100	0	0	0				0	0	1				0	0	99			
500	0	0	0				0	0	0				0	0	100			
1000	0	0	0				0	0	0				100	0	100			

TABLA 7. Porcentaje de aciertos para las series Weibull, con M=10 y analizando para puntos anormales por encima de la media

K=(5,5,6,6,6,9,9,9,10,10)

	0 puntos anormales detectados					1 punto anormal detectado					10 puntos anormales detectados			
N	R	N	D	Em	Lm	R	N	D	Em	Lm	R	N	D	Em	Lm
15		0	100	100	100		100	0	0	0		0	0	0	0
20		0	100	100	100		100	0	0	0		0	0	0	0
25	0	0	100	100	100	0	100	0	0	0		0	0	0	0
30	100	0		4	0	0	0		0	0	0	100		96	100
35	100	0		0	0	0	0		0	0	0	100		100	100
40	100	0		0	0	0	0		0	0	0	100		100	100
45	100	0		0	0	0	0		0	0	0	100		100	100
50	100	0		0	0	0	0		0	0	0	100		100	100
100	100	0				0	0				0	100			
500	100	0				0	0				0	100			
1000	0	0				0	0				100	100			

A partir de los experimentos realizados, y analizando todos los resultados obtenidos, se pueden hacer las siguientes conclusiones:

a. Para las series generadas con la Distribución Normal, y para las pruebas de puntos anormales por encima de la media se observó:

- Cuando el número de datos es muy pequeño (<10), los resultados de todas las pruebas no son buenos.
- La prueba de las Desviaciones Normalizadas es la que mejores resultados da para el rango de aplicación (25 £ N £ 100 y M £ 5).
- Para M=1, en el esquema de m + s K, K debe ser ³ 4 para que las pruebas den resultados satisfactorios.
- Con M=1, mientras menor sea el número de datos en el análisis, mayor debe ser K para que se obtengan resultados satisfactorios. Por ejemplo para:

N = 15	K ³ 5
N = 10	K ³ 6
N = 5	K ³ 7

- Cuando K ³10 y M =1 todas las pruebas detectan siempre los puntos anormales sin importar el valor de N
- La prueba de la distribución normal, cuando los números bases son normales y M =1, es una prueba muy acertiva, detectando siempre todos los puntos anormales.
- En los análisis para M =1, las pruebas que se van deteriorando a medida que K decrece son las siguientes:

Prueba	K	N a partir del cual se obtienen resultados satisfactorios
E_m	10	5
	9	5
	8	10
	7	10
	6	10
	5	15
	4	35
L_mS y L_mP	10	5
	9	5
	8	5
	7	10
	6	10
	5	10
	4	25
Dixon	10	5
	9	5
	8	5
	7	10
	6	10
	5	15
	4	----

- La prueba Dixon solo debe usarse para:

K ³ 5	N ³ 15
6 £ K £ 7	N ³ 10
K ³ 8	N ³ 5

- La prueba Roesner es muy acertiva para K ³ 5, identificando siempre el punto anormal.
- La prueba de las Desviaciones Normalizadas es muy acertiva para K ³ 5, identificando siempre los puntos anormales.
- En general, la ubicación del punto anormal en la serie, no afecta los resultados.

- Para valores de M³1, las pruebas de Roesner, Em y Lm, presentan mejores resultados cuando la combinación de valores de K ³ 5.
- En la prueba de Roesner, a medida que aumenta el número de puntos anormales, el número de datos de la serie debe aumentar para poder detectarlos.

b. Para las series generadas con la Distribución Normal, realizando las pruebas para los puntos anormales por debajo de la media se observó:

- Para este caso las pruebas de Roesner, Desviaciones Normalizadas, Dixon, y Lm no dan buenos resultados.
- Las pruebas Normal y Em dan buenos resultados, en todos los casos detectan puntos anormales, aunque a medida que aumenta el número de puntos anormales a detectar, aumenta la exigencia en el número de datos que debe tener la serie.

c. Para las series generadas con la Distribución Gumbel, realizando las pruebas para los puntos anormales por encima de la media se observó:

- La prueba Normal es muy eficiente en el caso de M =1, siempre detecta el punto anormal sin importar ni N, ni K. Para el caso de M ≥ 2, da buenos resultados para combinaciones de K ≥ 5 y N ≥ 40.
- La prueba de Roesner, en general, no presenta resultados satisfactorios, solo detecta puntos anormales para N ≥ 1000. Solamente para el caso de M =1 y K ≥ 6 presenta resultados satisfactorios.
- La prueba de las Desviaciones Normalizadas no presenta resultados satisfactorios. Al igual que la prueba de Roesner solo para el caso de M = 1 y K ≥ 6 presenta resultados aceptables.
- Las pruebas de Em y Lm presentan buenos resultados en el caso de combinaciones de K ≥ 6 y N ≥ 40.

d. Para las series generadas con la Distribución Gumbel, realizando las pruebas para los puntos anormales por debajo de la media se observó:

- Las pruebas de Roesner, Desviaciones Normalizadas, Dixon y Lm no presentan resultados satisfactorios en todos los casos.
- La prueba Em da buenos resultados para N ≥ 20.
- La prueba Normal da buenos resultados para 25 ≤ N ≤ 100.

e. Para las series generadas con la Distribución Weibull, realizando las pruebas para los puntos anormales por encima de la media se observó:

- Para M = 1, todas las pruebas detectan el punto anormal para K ≥ 6. La prueba Roesner detecta el punto anormal para K ≥ 5 y la prueba Normal lo detecta en todos los casos.
- Para M ≥ 2, las pruebas Roesner y Desviaciones Normalizadas no detectan puntos anormales en ninguno de los casos, Roesner tiene una excepción detectando puntos anormales solamente para N ≥ 1000.
- Las pruebas Em y Lm presentan resultados satisfactorios, sin embargo, a medida que aumenta el número de puntos anormales, disminuye su capacidad de detectarlos para valores de N pequeños.
- La prueba Normal es la que presenta los mejores resultados, detectando puntos anormales en todos los casos.

f. Para las series generadas con la Distribución Weibull, realizando las pruebas para los puntos anormales por debajo de la media se observó:

- Las pruebas Normal y Em dan buenos resultados para valores de N ≥ 25.
- Las otras pruebas presentan malos resultados en todos los casos.

4. ANÁLISIS DE SERIES HISTÓRICAS

Además de los experimentos anteriores con series sintéticas, las pruebas de detección de puntos anormalmente extremos se utilizaron para detectar puntos anormales en algunas series históricas de caudales máximos instantáneos en estaciones ubicadas en diferentes Departamentos, y en las cuales se sospecha la presencia de puntos anormales. En la Tabla 8 se presentan los resultados para los diferentes departamentos. En esta Tabla se usa la misma nomenclatura que en las Tablas anteriores para diferenciar las pruebas. Se añade la columna puntos anormales esperados para indicar el número de puntos anormales que visualmente identificaron los analistas.

TABLA 8. Puntos anormales detectados en las series de Caudales Máximos Instantáneos.

ANTIOQUIA

SERIE	Datos	R	SD	N	D	Em	Lm	Puntos anormales
		>=25	20-100	Todos	3-25	3-50	3-50	Esperados
1	12	-	-	1	1	1	1	1
2	21	-	0	1	1	1	1	2
3	16	-	-	1	0	0	0	2
4	11	-	-	8	1	0	0	1
5	18	-	-	1	1	1	1	1
6	28	1	0	1	-	1	1	2
7	15	-	-	1	0	0	1	3
8	11	-	-	1	1	1	1	1
9	12	-	-	1	0	0	1	1
10	11	-	-	1	0	0	0	1
11	26	1	0	1	-	1	1	2
12	17	-	-	1	0	0	0	3
13	16	-	-	7	0	0	0	2
14	24	-	0	1	0	0	0	1

CUNDINAMARCA

SERIE	Datos	R	SD	N	D	Em	Lm	Puntos anormales
		>=25	20-100	Todos	3-25	3-50	3-50	Esperados
1	37	0	0	1	-	0	0	5
2	41	0	0	3	-	3	3	3
3	48	0	0	1	-	0	0	2
4	34	0	0	10	-	10	10	10
5	37	1	0	1	-	1	1	3
6	30	2	0	2	-	2	2	2

RISARALDA

SERIE	Datos	R	SD	N	D	Em	Lm	Puntos anormales
		>=25	20-100	Todos	3-25	3-50	3-50	Esperados
1	23	-	0	1	0	0	0	1
2	7	-	-	1	1	0	1	1
3	49	0	0	1	-	0	0	3
4	15	-	-	1	0	0	0	3
5	24	-	0	1	0	0	0	2
6	29	1	0	1	-	1	1	1
7	20	-	-	1	0	0	0	1
8	23	-	0	1	1	1	1	1
9	23	-	0	1	1	1	1	2
10	25	0	0	1	-	0	0	1

Continuación **Tabla 8.**

VALLE

SERIE	Datos	R	SD	N	D	Em	Lm	Puntos anormales
		>=25	20-100	Todos	3-25	3-50	3-50	Esperados
1	31	0	0	1	-	0	0	1
2	46	0	0	1	-	0	0	4
3	26	0	0	1	-	0	0	0
4	47	0	0	1	-	0	0	1
5	27	1	1	1	-	1	1	2
6	42	0	0	1	-	0	0	2
7	18	-	-	1	0	0	0	2
8	35	0	0	1	-	0	0	1
9	15	-	-	1	0	0	0	2
10	34	0	4	2	-	2	2	2
11	21	-	0	1	0	0	0	3
12	39	1	1	1	-	1	1	2
13	21	-	0	1	0	0	0	1
14	16	-	-	1	0	0	0	2
15	21	-	1	1	1	1	1	2
16	47	0	2	2	-	2	2	1
17	16	-	-	1	0	0	0	1
18	23	-	3	1	0	0	0	3
19	41	2	3	2	-	2	2	3
20	15	-	1	1	0	0	0	1
21	47	0	1	1	-	0	0	1
22	31	0	1	1	-	0	0	1

Analizando estos resultados no se encontraron diferencias regionales. Aunque las pruebas detectaron puntos anormales, no fueron muy asertivas en la detección del número de puntos anormales esperados.

Un último análisis que se realizó en este estudio es la posible relación entre la ocurrencia del fenómeno de El Niño Oscilación del Sur (ENSO) y la ocurrencia de puntos anormalmente extremos. Debido a que la ocurrencia del ENSO genera una fase cálida de escasez hidrológica, y su fase fría (La Niña) de abundancia, se esperaría que los puntos anormalmente bajos ocurran con El Niño y los anormalmente altos con La Niña. Se trata de hacer un análisis preliminar de esta posible relación.

También se utilizaron las pruebas para detectar puntos anormales en series históricas de caudales, con el fin de determinar si los puntos anormales detectados correspondían a períodos pertenecientes a El Niño o a La Niña. En las Tablas 9 y 10 se presentan los resultados obtenidos para este análisis.

En estas Tablas se usaron series de caudales medios diarios con un número significativo de registros (N°3000) y los análisis se hicieron para los 100 valores más extremos. En la Tablas 9 y 10, la columna puntos anormales en año Niño o Niña indica el número de puntos anormales detectados en años Niño o Niña, en la siguiente columna se presenta el mismo resultado en porcentaje, la columna puntos anormales en otro, indica el número de puntos anormales en año no-Niño o no-Niña, en la siguiente columna se presenta el mismo resultado en porcentaje, y la columna posición valor máximo año Niño o Niña, indica la posición del máximo punto anormal detectado en año Niño o Niña en toda la serie de puntos anormales. Las series que se utilizaron corresponden a estaciones ubicadas en todo el país, las cuales tenían registros de 10 o más años de caudales medios diarios completos (sin información faltante).

Las pruebas de detección de puntos anormales asumen que la serie hidrológica es independiente, lo cual es una suposición razonable cuando se trata de series extremas anuales. Las series de caudales medios anuales no cumplen

esta condición, presentando una alta dependencia serial. Se decidió en este estudio aplicar estas pruebas a estas series a pesar de que no cumplen con la condición de independencia, tratando de analizar su aplicabilidad en estos casos, y debido a que sería más notorio el posible efecto del ENSO sobre estas series.

TABLA 9. Puntos anormales detectados en las series de caudales históricos para períodos de El Niño.

SERIE	PUNTOS ANORMALES DETECTADOS	PUNTOS ANORMALES EN AÑO NIÑO	PUNTOS ANORMALES EN AÑO NIÑO (%)	PUNTOS ANORMALES EN OTRO (%)	POSICIÓN VALOR MAX AÑO NIÑO
1	100	70	70	30	1
2	100	37	37	63	4
3	100	28	28	72	10
4	100	15	15	85	29
5	100	9	9	91	41
6	100	33	33	67	5
7	100	71	71	29	1
8	100	62	62	38	1
9	100	54	54	46	22
10	100	49	49	51	1
11	100	99	99	1	1
12	100	54	54	46	1
13	100	3	3	97	38
14	100	57	57	43	1
15	100	12	12	88	20
16	100	4	4	96	82

Tabla 10. Puntos anormales detectados en las series de caudales históricos para períodos de La Niña.

SERIE	PUNTOS ANORMALES DETECTADOS	PUNTOS ANORMALES EN AÑO NIÑA	PUNTOS ANORMALES EN AÑO NIÑA (%)	PUNTOS ANORMALES EN OTRO (%)	POSICIÓN VALOR MAX AÑO NIÑA
1	100	13	13	87	29
2	100	42	42	58	1
3	100	55	55	45	9
4	100	2	2	98	40
5	100	36	36	64	2
6	100	24	24	76	1
7	100	79	79	21	1
8	100	87	87	13	1
9	100	47	47	53	2
10	100	24	24	76	6
11	100	28	28	72	9
12	100	26	26	74	6
13	100	14	14	86	4
14	100	5	5	95	7
15	100	15	15	85	3
16	100	8	8	92	19

En el análisis de las series de caudales medios diarios históricos se encontró que las series 7 y 8, correspondientes a la estación La Virginia en el departamento de Risaralda y La Pintada en el departamento de Caldas, poseen los porcentajes más altos de determinación de puntos anormales en año Niña, cuando se realizan los análisis por encima de

la media, como se mostró en la Tabla 10. Mientras que las series 1, 7 y 11, correspondientes a las estaciones Puente Santander en el departamento del Huila, La Virginia en el departamento de Risaralda y Piedras Blancas en el departamento de Antioquia, poseen los porcentajes más altos de determinación de puntos anormales en año Niño, cuando se realizan los análisis por debajo de la media, como se presentó en la Tabla 9.

Con información encontrada en estudios recientes sobre hidrología Colombiana como el de Rendón (2001), se determinó que estas estaciones pueden responder a la influencia que ejerce la corriente del chorro del Chocó sobre estas zonas, correspondiendo al debilitamiento de la corriente de chorro del occidente Colombiano o corriente de chorro del Chocó, como respuesta a la ocurrencia de la fase cálida del Niño y al aumento de la corriente de chorro del Chocó, como respuesta a la ocurrencia de la fase fría de La Niña. Estos resultados son un nuevo elemento que muestra nuevamente la ya comprobada relación entre la ocurrencia del fenómeno ENSO en sus dos fases y la hidrología colombiana.

5. CONCLUSIONES Y RECOMENDACIONES

De acuerdo con las características de las pruebas presentadas y con los resultados obtenidos se pueden hacer las siguientes observaciones y conclusiones:

a. Existen varias pruebas para detección de puntos anormalmente extremos con diferentes exigencias de información para su aplicación, y diferentes suposiciones sobre la distribución de donde se asume se tomó la muestra (paramétricas y no paramétricas).

b. En el caso de Colombia las series son de relativa poca longitud y por lo tanto esta situación tiende a favorecer las pruebas que son poco exigentes en el tamaño de la muestra, que generalmente son las pruebas de menor potencia estadística.

c. Las pruebas que mejores resultados arrojaron fueron las de la Distribución Normal y la de las Desviaciones Normalizadas, seguidas por las pruebas de Roesner y del estadístico Em.

d. En general, las pruebas tienden a dar resultados satisfactorios para K=6 (tamaño del punto anormal), N=30 (número de observaciones) y M (número de puntos anormales) pequeño.

e. Se confirmó la relación entre la hidrología colombiana y la ocurrencia del fenómeno ENSO en sus dos fases. Las estaciones en donde más claramente se detectaron los puntos anormales están influenciadas por la corriente del chorro del Chocó, el cual se ve claramente afectado por la ocurrencia del fenómeno ENSO.

f. Las pruebas detectaron puntos anormales, pero no fueron muy asertivas en la detección del número de puntos anormales esperados cuando se usaron las series de caudales medios diarios históricos. Las pruebas que mejores resultados dieron fueron las mismas que en el caso de series sintéticas.

g. Como recomendación se propone el uso de pruebas de detección de puntos anormales en análisis hidrológicos, de tal manera que se puedan identificar y ser removidos de los análisis posteriores que se hagan con las series. La no remoción de estos puntos significa trabajar con muestras que vienen de poblaciones distintas, lo cual puede tener consecuencias no apropiadas sobre los análisis que se hagan.

6. REFERENCIAS

Gibbon, R.D. 1994. Statistical Methods for Groundwater Monitoring. John Wiley and Sons Inc., New York.

Gilbert, R.O. 1987. Statistical Methods for Environmental Pollution Monitoring. Van Nostrand Reinhold Company, New York.

Helsel, D.R. and Hirsch, R.M., 1992. Statistical Methods in Water Resources. Elsevier, Amsterdam.

Kottegoda, N.T. and Rosso, R.1997. Probability, Statistics and Reliability for Civil and Environmental Engineers. McGraw Hill Book Co., New York.

Mood, A.M; Garybill, F.A and Boes, D.C.1974. Introduction to the Theory of Statistics. McGraw Hill Book Co., New York.

Rendón, A. 2001. Influencia de tres corrientes de chorro sobre la hidrología colombiana. Tesis de pregrado. Facultad de Minas. Universidad Nacional de Colombia. Medellín.

Salas, J; Smith, R; Tabios, G and Heo, J. 1992. Statistical computer techniques in water resources and environmental engineering. Department of civil engineering. Colorado State University. January.

Tietjen, G.L and Moore, R.H. 1972. Some Grubbs-Type statistics for the detection of several puntos anormales. Technometrics, 14(3): 583-597.