

COMPARACIÓN DE CUATRO TÉCNICAS PARA SELECCIONAR LA DISTRIBUCIÓN DE PROBABILIDAD DE MEJOR AJUSTE PARA EL ANÁLISIS DE DATOS DE PRECIPITACIÓN MÁXIMA ANUAL EN MÉXICO

COMPARISON OF FOUR TECHNIQUES TO SELECT THE PROBABILITY DISTRIBUTION OF BEST FITTING FOR THE ANALYSIS OF DATA OF ANNUAL MAXIMUM PRECIPITATION IN MEXICO

Flowers-Cano R.S.^{1*}, Ortiz-Gómez R.², Burgos-Flores D.³, León-Jiménez J. E.¹,
Balladares-Sánchez M. Á.¹

^{1*} Universidad Juárez Autónoma de Tabasco, 86690, Tabasco, México.

² Universidad Autónoma de Zacatecas.

³ Universidad de Sonora.

floresrobert133@gmail.com*

RESUMEN

Se realizó un estudio acerca de la selección de la mejor distribución probabilística para el análisis de frecuencias de lluvias máximas anuales de 24 horas, utilizando registros de 21 estaciones meteorológicas de México. Se consideraron las distribuciones log-gamma, Gumbel, log-logística, log-Pearson tipo 3 (LP3), general de valores extremos (GVE) y logística-generalizada para ajustar los datos; y para encontrar cuál de estas era más apropiada para el análisis de frecuencias se utilizaron las pruebas de error estándar de ajuste (EEA) y Kolmogorov- Smirnov (KS), así como los criterios de información de Akaike (CIA) y Bayesiano (CIB). Los resultados muestran que, en general, los diferentes criterios eligen diferentes distribuciones, lo cual puede producir diferencias importantes en la estimación de los cuantiles. Las distribuciones que se ajustaron mejor en la mayoría de las ocasiones a las series analizadas fueron las de la familia log Pearson tipo 3, según CIA y CIB; y las de la familia GVE, según EEA y KS.

Palabras clave: Análisis de frecuencias; Criterios de selección; Distribuciones de probabilidad.

ABSTRACT

A study on the selection of the best probability distribution for the frequency analysis of annual maximum rainfall in 24 hours was performed, using records from 21 meteorological stations of Mexico. We considered the log-gamma distribution, Gumbel, log-logistic, log-Pearson type 3 (LP3), general extreme values (GEV) and generalized-logistic to adjust the data; and to find which of these was more appropriate for the analysis of frequencies tests of standard error of fitting (EEA) and Kolmogorov - Smirnov (KS) were used as well as the criteria of information (CIA) Akaike and Bayesian (CIB). The results show that the different criteria in general, choose different distributions, which can produce significant differences in the estimation of quantiles. The distributions that were better adjusted in the majority of cases to the analyzed series were those of the family log Pearson type 3, according to CIA and CIB; and those of the family GEV, according to EEA and KS.

Key words: Frequency analysis; Selection criteria; Probability distributions.

INTRODUCCIÓN

Los eventos hidrológicos extremos tienen impactos económicos y sociales negativos sobre las poblaciones humanas. La protección de dichas poblaciones requiere el correcto dimensionamiento de estructuras hidráulicas, que se plantea en términos de un evento extremo hipotético conocido como caudal o precipitación de diseño, el cual está asociado a un determinado periodo de retorno [1]. Para estimar este evento con frecuencia se recurre al análisis de frecuencias, el cual relaciona la magnitud de los eventos extremos con la probabilidad de ocurrencia de los mismos, por medio de distribuciones de probabilidad [2].

Un problema importante en el análisis de frecuencias es la selección de la distribución de probabilidad apropiada para describir el comportamiento de los datos observados. Existen diversas funciones de distribución de probabilidad para el análisis de frecuencias, pero ninguna es universalmente aceptada como la mejor para el análisis de variables hidroclimatológicas [3], [4], [5]. Algunas de las distribuciones más utilizadas en hidrología son la normal, log-normal, Gumbel, Weibull, General de Valores Extremos (GVE), Pearson y Log-Pearson tipo 3 [6], [7], [8]. En México y Latinoamérica, la función de distribución de Gumbel ha sido ampliamente utilizada como una distribución para el análisis de eventos extremos [9], [10], [11].

Los valores estimados por medio de diferentes distribuciones de probabilidad difieren entre sí, por lo tanto, es necesario seleccionar entre ellas cuál se ajusta mejor a la serie de tiempo bajo análisis; dicha selección se basa generalmente en pruebas de bondad

de ajuste, las cuales representan la concordancia entre la distribución empírica de frecuencias y una distribución teórica. Las pruebas de bondad de ajuste incluyen métodos gráficos y estadísticos, siendo preferibles los métodos estadísticos por su objetividad [12]. Entre los métodos estadísticos con mayor difusión se encuentran las pruebas de chi- cuadrado (χ^2) y del error estándar de ajuste (EEA) [13]. Otros métodos usados con frecuencia son los de función de distribución empírica, que incluyen las pruebas de Kolmogorov- Smirnov (KS), Cramer- Von Mises y Anderson- Darling (AD) [14], [15], [16], [17].

Sin embargo, las pruebas estadísticas de bondad de ajuste tienen poco poder para rechazar distribuciones equivocadas [18], por lo que en muchos casos, más de una distribución puede ser aceptada por una prueba específica [19]. Por lo cual, el concepto de criterio de selección de modelos representa una alternativa a las pruebas de bondad de ajuste. Diversos criterios de selección pueden definirse en función de los estadísticos de bondad de ajuste antes mencionados. Por ejemplo, el criterio de selección de AD consiste en calcular el estadístico del mismo nombre para cada una de las distribuciones de interés, y seleccionar aquella que tenga el valor más pequeño. Una vez seleccionada la distribución, puede hacerse una prueba de bondad de ajuste, con la ventaja de aplicarla a una sola distribución. Otros criterios de selección se basan en la función de verosimilitud, como el criterio de información de Akaike (CIA) y el criterio de información Bayesiano (CIB) [19]. Los criterios de selección de modelos probabilísticos han recibido poca atención en la literatura hidrológica. En México, el EEA es usado frecuentemente [20], [21], [22], [23].

El objetivo de este trabajo es comparar el desempeño de las pruebas de error estándar de ajuste y Kolmogorov- Smirnov, así como los criterios de información de Akaike y Bayesiano, en la selección de diferentes distribuciones de probabilidad en el análisis de frecuencia de series de precipitación máxima anual, con el fin de observar con qué frecuencia coinciden en la selección de las distribuciones, y en caso de que no coincidan, si los valores estimados difieren mucho entre sí.

METODOLOGÍA

Información climatológica. Los datos utilizados en este trabajo fueron los registros de precipitación máxima anual de 24 horas obtenidos de la base de datos CLICOM [24], de 21 estaciones meteorológicas, las cuales están ubicadas en cinco estados de la república mexicana con tres climas diferentes. La longitud de los registros varía desde 32 hasta 89 años para el periodo de 1923 a 2012 (ver Tabla 1).

Tabla 1. Localización geográfica e información relacionada a las estaciones climatológicas de estudio.

Clave	Nombre	Tipo de clima	Latitud (N)	Longitud (E)	Altitud (msnm)	Periodo de	No. de
6017	Madrid	Cálida subhúmeda	19.1122	-103.8839	195	1970-2012	43
6054	M. Ávila Camacho	Cálida subhúmeda	19.2850	-104.2450	376	1978-2012	35
6058	Tecomán	Cálida subhúmeda	18.9083	-103.8744	30	1954-2012	59
11003	Agua Tibia	Templada subhúmeda	20.5103	-101.6294	1720	1949-2012	64
11014	Cuerámaro	Templada subhúmeda	20.6256	-101.6758	1732	1967-2012	46
11028	Irapuato	Templada subhúmeda	20.6689	-101.3372	1729	1923-2011	89
11035	La Sandía	Templada subhúmeda	20.9211	-101.6974	1771	1965-2012	48
11036	Adjuntas	Tempalda subhúmeda	20.6753	-101.8442	1727	1944-2012	69
11134	El Conejo	Templada subhúmeda	20.7158	-101.3697	1740	1978-2012	35
14011	Apazulco	Cálida subhúmeda	19.3064	-104.8875	5	1961-2011	51
14036	Cuautitlán	Cálida	19.4506	-104.3592	600	1958-	54

		subhúmeda				2011	
14038	Cuixtla	Templada subhúmeda	21.0519	-103.4389	1000	1954- 2011	58
14067	Higuera Blanca	Cálida subhúmeda	19.9942	-105.1625	140	1956- 2002	47
14148	Tecomates	Cálida subhúmeda	19.5583	-104.5000	286	1961- 2006	46
16100	Puente San Isidro	Templada subhúmeda	19.8658	-101.5189	2022	1947- 1992	45
27019	Jalapa	Cálida húmeda	17.7233	-92.8117	14	1971- 2012	42
27024	La Huasteca	Cálida húmeda	17.5200	-92.9267	80	1970- 2012	32
27037	P. Nuevo	Cálida húmeda	17.8542	-92.8792	21	1949- 2012	64
27042	Tapijulapa	Cálida húmeda	17.4611	-92.7775	44	1962- 2012	50
27044	Teapa	Cálida húmeda	17.5489	-92.9533	51	1960- 2012	52
27061	Puyacatengo	Cálida húmeda	17.5133	-92.9200	86	1972- 2012	39

Distribuciones de probabilidad consideradas. Las funciones de distribución de probabilidad usadas en el análisis de frecuencias de las series de precipitaciones máximas anuales fueron la log-gamma, Gumbel, log-logística, log-Pearson tipo 3 (LP3), general de valores extremos (GVE) y logística-generalizada. Estas distribuciones fueron seleccionadas por las razones siguientes: las distribuciones LP3, GVE y logística generalizada son distribuciones de tres parámetros usadas comúnmente en hidrología. La distribución LP3 fue adoptada por las agencias federales de los Estados Unidos y Australia para el análisis de frecuencia de caudales [25], [26], [27], mientras que en el Reino Unido la distribución logística generalizada fue seleccionada para este mismo fin

[12], [28]. La distribución GVE, así como sus casos especiales, son utilizadas ampliamente en el ámbito hidrológico [25], [26], [29], [30]. Las otras tres distribuciones son distribuciones de dos parámetros, las cuales están relacionadas con alguna de las distribuciones de tres parámetros seleccionadas. La distribución log-gamma pertenece a la misma familia que la LP3; la Gumbel pertenece a la familia de la GVE; y la log-logística a la familia de la logística generalizada.

Estimación de parámetros. Para cada serie de precipitación máxima anual, se ajustaron las seis distribuciones consideradas. Es decir, se estimaron los valores de los parámetros de estas distribuciones a partir de las observaciones de cada serie. Los parámetros de las distribuciones Gumbel, GVE, log-logística y logística generalizada se estimaron por el método de los momentos L, mientras que los parámetros de la distribución log-gamma y de la LP3 se estimaron usando momentos convencionales.

Criterios de selección de modelos Para determinar cuál de las distribuciones consideradas era la más apropiada para modelar las observaciones de precipitación máxima anual de cada estación, se utilizaron las pruebas de error estándar de ajuste (EEA) y Kolmogorov-Smirnov (KS), así como el Criterio de Información Bayesiano (CIB) y el Criterio de Información de Akaike (CIA).

El EEA se define como [31]:

$$EEA = \left[\frac{\sum_{i=1}^n (x_i - \hat{x}_i)^2}{n-p} \right] \quad (1)$$

Donde: x_i son las observaciones ordenadas de menor a mayor; i es el número de orden de la observación x_i ; $\hat{x}_i = F^{-1}\left(\frac{i}{n+1}\right)$; $F(x_i)$ es la función de distribución teórica; n es el número de observaciones; y p el número de parámetros de la distribución ajustada.

El estadístico de KS se define de la siguiente manera:

$$KS = \max_{1 \leq i \leq n} \left[F(x_i) - \frac{i-1}{n}, \frac{i}{n} - F(x_i) \right] \quad (2)$$

Donde: x_i son las observaciones; i es el número de orden de la observación x_i ; $F(x_i)$ es la función de distribución teórica; y n es el número de observaciones.

El CIA se define mediante la ecuación (3) [32]:

$$CIA_j = -2 \ln(L_j(\hat{\phi})) + 2p_j \quad (3)$$

Donde: $L_j(\hat{\phi}) = \prod_{i=1}^n f_j(x_i, \hat{\phi})$ es la función de verosimilitud, evaluada en $\phi = \hat{\phi}$, que corresponde al estimador del vector de parámetros ϕ ; f_j es la j -ésima función de distribución teórica considerada; y p_j es el número de parámetros de dicha función.

Para el caso particular de este trabajo, $j=1,2,\dots,6$, ya que las distribuciones teóricas fueron seis.

El CIB, por su parte se define como [33]:

$$CIB_j = -2 \ln(L_j(\hat{\phi})) + p_j \ln(n) \quad (4)$$

En ambos casos, se selecciona la función de distribución f_j para la cual el correspondiente criterio de información haya tenido el valor más pequeño.

Tanto el CIA y como el CIB toman en cuenta el principio de parsimonia, es decir, entre dos modelos con un nivel comparable de ajuste, seleccionan el modelo más sencillo. La forma en que toman en cuenta este principio es por medio del segundo término de las ecuaciones (3) y (4), que penaliza el número de parámetros de una distribución.

Estimación de cuantiles. Para las distribuciones de mejor ajuste según los criterios aplicados, se estimaron los cuantiles correspondientes a los periodos de retorno de 2, 5, 10, 20, 50, 100, 200, 500 y 1000 años. Como las formas inversas de las distribuciones Gumbel, GVE, log-logística y logística generalizada se pueden definir analíticamente, los estimadores \hat{x}_T de los cuantiles resultan de invertir las respectivas funciones de distribución de probabilidad, ecuaciones (5) a (8), respectivamente. Las formas inversas de las distribuciones log-gamma y LP3 no se pueden definir analíticamente. Para estos casos se utilizó el software Scilab que incluye una función inversa de la distribución gamma, mediante la cual se obtienen los cuantiles de dicha distribución. Scilab calcula esta función inversa usando el algoritmo descrito por [34]: Dicha función inversa sirvió como base para calcular los cuantiles de las distribuciones log-gamma y LP3.

$$\hat{x}_T = \hat{\theta} - \hat{\beta} \ln \left[\ln \left(\frac{T}{T-1} \right) \right] \quad (5)$$

$$\hat{x}_T = \hat{\theta} - \frac{\hat{\beta}}{\hat{\alpha}} \left\{ 1 - \left[\ln \left(\frac{T}{T-1} \right) \right]^{-\hat{\alpha}} \right\} \quad (6)$$

$$\hat{x}_T = \exp \{ \hat{\theta} + \hat{\beta} \ln(T - 1) \} \quad (7)$$

$$\hat{x}_T = \hat{\theta} + \frac{\hat{\beta}}{\hat{\alpha}} [1 - (T - 1)^{-\hat{\alpha}}] \quad (8)$$

RESULTADOS

En la tabla 2 se muestran las distribuciones seleccionadas en primer lugar por los criterios KS, EEA, CIA y CIB. En ella se observa que los cuatros criterios empleados sólo coincidieron una vez en la selección de la mejor distribución (la distribución GVE para la serie Tecomates), y que al menos tres de los criterios coincidieron en la selección de la mejor distribución en siete ocasiones. También se observa que los criterios CIA y CIB casi siempre coinciden, excepto para las series Pueblo Nuevo y Puyacatengo. Los criterios CIA y CIB coinciden con el criterio KS cuatro veces, mientras que el criterio EEA coincide en cinco, y estos últimos dos (KS y EEA) coinciden entre si un total de siete veces.

Tabla 2. Funciones de distribución de probabilidad seleccionadas por los diferentes criterios de selección.

Serie	Criterio de selección			
	KS	EEA	CIA	CIB
Madrid	L-gen	LP3	LP3	LP3
M. Ávila	GVE	LP3	LP3	LP3
Tecomán	L-gen	LP3	L-gam	L-gam
Agua Tibia	GVE	GVE	LP3	LP3
Cuerámara	L-gen	L-gen	GVE	GVE
Irapuato	GVE	L-log	L-log	L-log
La Sandía	L-gen	GVE	L-gen	L-gen
Adjuntas	L-gen	GVE	L-gam	L-gam
El Conejo	GVE	GVE	L-gam	L-gam
Apazulco	L-gam	LP3	L-gam	L-gam
Cuautitlán	GVE	GVE	LP3	LP3
Cuixtla	GVE	GVE	L-gam	L-gam
Higuera Blanca	L-gen	GVE	Gumbel	Gumbel
Tecomates	GVE	GVE	GVE	GVE

Puente San Isidro	LP3	Gumbel	Gumbel	Gumbel
Jalapa	GVE	L-gen	L-log	L-log
La Huasteca	GVE	LP3	L-gam	L-gam
Pueblo Nuevo	Gumbel	GVE	LP3	L-gam
Tapijulapa	L-gam	GVE	L-gam	L-gam
Teapa	GVE	GVE	Gumbel	Gumbel
Puyacatengo	L-gen	GVE	LP3	L-gam

En la Tabla 3 se muestra el número de veces que quedó en primer lugar cada distribución. La distribución más seleccionada por los criterios KS y EEA fue la GVE, y posteriormente la distribución logística generalizada y la LP3, respectivamente. La distribución seleccionada el mayor número de veces por los criterios CIA y CIB fue la log-gamma, seguida por la LP3.

Para 13 de las 21 series analizadas, los criterios CIA y CIB seleccionaron distribuciones de la familia LP3 como las de mejor ajuste. Las distribuciones de la familia GVE, que comúnmente se emplean en América Latina para el análisis de frecuencias de variables hidrológicas, sólo fueron seleccionadas en 5 ocasiones por estos criterios. En cambio, las distribuciones de la familia GVE fueron las más seleccionadas por las pruebas de KS y de EEA, mientras que las distribuciones de la familia LP3 fueron poco seleccionadas por estos criterios.

El CIA eligió en 12 casos una distribución de dos parámetros (2p) y en 9 casos una de tres parámetros (3p); el CIB escogió en 14 casos una distribución de 2p y 7 casos una de 3p; KS eligió en 3 casos una distribución de 2p y en 18 casos una de 3p; y por último, el EEA escogió en 2 casos una distribución de 2p y en 19 casos una de 3p. Es

decir, los criterios CIA y CIB mostraron una preferencia por las distribuciones de 2p, en tanto que el KS y el EEA prefirieron distribuciones de 3p.

Con las distribuciones seleccionadas se estimaron los cuantiles de interés (Tablas 4, 5 y 6).

Tabla 3. Número de veces que quedó en primer lugar cada distribución por criterio de selección.

Distribución	Criterios de selección			
	KS	EEA	CIA	CIB
Gumbel	1	1	3	3
Log Logística (L-log)	0	1	2	2
Log Gamma (L-gam)	2	0	7	9
Gumbel Valores Extremos (GVE)	10	12	2	2
Log Generalizada (L-gen)	7	2	1	1
Log Pearson Tipo 3 (LP3)	1	5	6	4

Tabla 4. Cuantiles estimados de las series de precipitación máxima anual. Clima cálido subhúmedo.

Serie	Fdp de ajuste	Cuantiles estimados (mm)								
		T (años)								
		2	5	10	20	50	100	200	500	1000
Madrid	LP3	93.2	142	184	233	312	384	469	607	733
	L-gen	93.5	137	176	225	311	399	514	722	936
M. Ávila Camacho	LP3	88.5	135	176	225	305	380	470	617	755
	GVE	87.4	131	171	221	308	396	508	706	905
Tecomán	L-gam	96.5	149	190	232	294	345	400	481	548
	L-gen	96.3	144	183	229	305	377	465	614	759
	LP3	95.7	149	191	235	302	358	420	512	590
Apazulco	L-gam	110	169	213	259	326	381	440	525	596
	LP3	114	171	209	248	298	338	378	433	475
Cuautitlán	LP3	101	141	174	211	267	317	374	462	541
	GVE	101	139	171	210	272	330	401	518	629

	Gumbel	99.6	158	197	235	283	319	355	403	439
Higuera Blanca	L-gen	95	149	190	238	315	385	470	610	740
	GVE	93.9	152	196	243	313	371	437	534	617
Tecomates	GVE	93.4	120	142	166	202	234	270	326	375

Tabla 5. Cuantiles estimados de las series de precipitación máxima anual. Clima templado suhúmedo.

Serie	Fdp de ajuste	Cuantiles estimados (mm)								
		T (años)								
		2	5	10	20	50	100	200	500	1000
Agua Tibia	LP3	52.3	66.5	74.1	80.4	87.5	92.1	96.2	101	104
	GVE	52.1	66.4	74.5	81.3	88.9	93.9	98.3	103	107
Cuerámara	GVE	49.3	63.8	71.2	76.8	82.7	86.1	88.9	91.9	93.6
	L-gen	49.4	62.4	70.1	77.1	86	92.6	99.1	108	114
Irapuato	L-log	51.5	64.5	73.5	83	96.8	108	121	141	158
	GVE	51.7	65.9	74.8	82.9	92.9	100	107	115	122
La Sandía	L-gen	54.2	66.1	72.5	78	84.6	89.1	93.3	98.5	102
	GVE	54.3	67.5	73.3	77.3	81	82.9	84.3	85.6	86.3
Adjuntas	L-gam	50.8	66.7	77.3	87.5	101	111	122	136	147
	L-gen	49.5	64.1	75.9	89.5	112	133	158	201	242
	GVE	49.1	64.9	77.3	90.9	112	129	150	181	208
El Conejo	L-gam	52.7	66.5	75.3	83.7	94.4	102	110	121	129
	GVE	53	66.9	75.6	83.5	93.1	99.9	106	114	120
Cuixtla	L-gam	55.6	68.2	76	83.3	92.4	99.2	106	115	121
	GVE	55.5	68.2	76.4	83.9	93.3	100	107	115	121
Pte. Sn. Isidro	Gumbel	44.7	54	60.2	66.1	73.8	79.6	85.3	92.8	98.5
	LP3	44.9	54	60	65.6	72.9	78.4	83.9	91.3	97

Tabla 6. Cuantiles estimados de las series de precipitación máxima anual. Clima cálido húmedo.

Serie	Fdp de ajuste	Cuantiles estimados (mm)								
		T (años)								
		2	5	10	20	50	100	200	500	1000

		2	5	10	20	50	100	200	500	1000
Jalapa	L-log	155	191	215	241	277	308	342	392	435
	GVE	153	192	219	245	279	305	332	368	396
	L-gen	154	189	215	242	283	318	358	419	472
La Huasteca	L-gam	142	187	217	246	283	312	341	380	411
	GVE	140	186	218	250	294	329	365	415	455
	LP3	140	186	218	251	297	333	371	425	469
Pueblo Nuevo	LP3	116	157	186	217	260	295	333	387	431
	L-gam	118	158	185	210	244	270	297	333	362
	Gumbel	119	160	187	213	247	272	297	330	355
	GVE	116	156	186	218	263	301	341	400	450
Tapijulapa	L-gam	188	252	295	337	391	434	477	535	581
	GVE	192	255	293	327	368	397	424	457	480
Teapa	Gumbel	171	210	237	262	295	319	344	376	400
	GVE	169	209	237	265	303	333	365	409	444
Puyacatengo	LP3	179	233	272	313	370	417	466	537	596
	L-gam	183	236	270	302	343	375	406	448	480
	L-gen	178	228	267	311	381	447	525	653	773
	GVE	177	231	272	315	380	434	494	583	659

En la Tabla 5 (clima templado subhúmedo) se observa que los criterios de selección no siempre coinciden en seleccionar la distribución más adecuada, aunque para la mayor parte de las series, los cuantiles estimados por medio de las distintas distribuciones seleccionadas no son tan diferentes entre sí. Sin embargo, al observar las Tablas 4 y 6 (climas cálido subhúmedo y cálido húmedo) se aprecian diferencias más importantes, especialmente para los cuantiles correspondientes a periodos de retorno mayores a 100 años.

En la Figura 1 se muestran gráficamente las diferencias entre los cuantiles estimados por medio de distintas distribuciones para las series de precipitación con las mayores diferencias. Se observa que para $T < 50$ años las diferencias son pequeñas; para $T > 50$ años las diferencias son mayores, y en algunos casos, considerables. Esto es razonable, ya que los registros históricos de las series analizadas son de alrededor de 50 años.

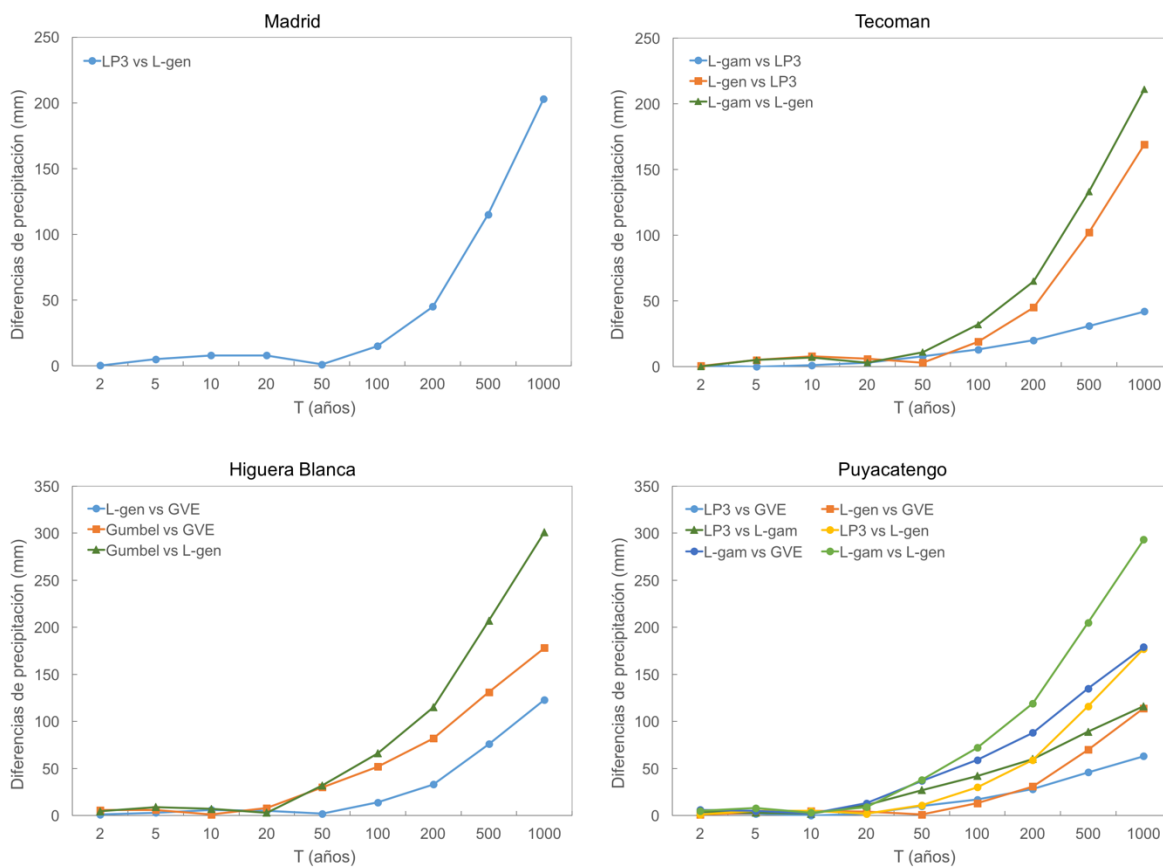


Fig. 1. Diferencias de los cuantiles de las distribuciones de probabilidad.

DISCUSIÓN

Los resultados muestran que, por lo general, los criterios CIA y CIB coincidieron en la selección de la mejor distribución, pero los criterios KS y EEA no coincidieron con el

CIA y el CIB. En general, los criterios KS y EEA prefirieron distribuciones de 3p sobre distribuciones de 2p, y distribuciones de la familia GVE sobre otras familias. En cambio, el CIA y el CIB prefirieron distribuciones de 2p sobre distribuciones de 3p, y distribuciones de la familia 3p sobre otras familias. Estas diferencias son importantes en la medida en que se traducen en diferencias en la estimación de los cuantiles, es decir, de los valores de precipitación asociados a determinados periodos de retorno.

Ahora bien, las diferencias entre los cuantiles de las distribuciones seleccionadas por medio de los diferentes criterios de selección no fueron iguales para todas las series analizadas. Se observó una relación entre las regiones climáticas a las que pertenecían las series y las diferencias entre los cuantiles. Así, para las series de la región templada subhúmeda (Tabla 5), no se observaron grandes diferencias entre los cuantiles de diferentes distribuciones. Esto significa que, para las series analizadas de dicha región, la discrepancia entre los criterios de selección no constituye un problema grave, pues no importa cuál de ellos se tenga por bueno, al final se obtienen estimaciones muy similares.

No sucede lo mismo para las series de las regiones cálida subhúmeda y cálida húmeda (Tablas 4 y 6). Para estas series, los cuantiles de las distribuciones seleccionadas por medio de los diferentes criterios pueden ser muy diferentes entre sí. En tales casos, la discrepancia entre los criterios de selección es importante, pues la selección de una distribución equivocada conduciría a graves errores de estimación, cuyas consecuencias podrían ser desastrosas.

Las diferencias más pequeñas entre los cuantiles se observan para $T < 50$ años, mientras que las diferencias más grandes se presentan para $T > 50$ años (Figura 1). Considerando que los registros de las series analizadas son de alrededor de 50 años, los periodos de retorno menores a 50 años corresponden más o menos a la región de interpolación, mientras que los periodos de retorno mayores a 50 años corresponden a la región de extrapolación. Pero es precisamente esta región de extrapolación la que interesa cuando se ajustan distribuciones de probabilidad, ya que para hacer interpolaciones basta la distribución empírica. Y las extrapolaciones pueden variar mucho según la distribución ajustada. Los criterios de selección tienen la finalidad de ayudar en la selección de una distribución de probabilidad adecuada, pero los diferentes criterios seleccionan diferentes distribuciones y no se sabe cuál de ellos es mejor.

Existen dos factores importantes que se deben considerar en la selección de una distribución de probabilidad: el nivel de ajuste y la incertidumbre de las estimaciones. Ambos factores están relacionados con el número de parámetros de la distribución. Al aumentar el número de parámetros se obtiene un mejor de ajuste, pero a costa de un incremento en la incertidumbre de estimación de cada parámetro adicional. Los criterios KS y EEA toman en cuenta únicamente el nivel de ajuste, y por eso no es de extrañar que prefieran distribuciones de 3p sobre distribuciones de 2p. En cambio, los criterios CIA y CIB penalizan el número de parámetros por medio del segundo término de las ecuaciones (3) y (4). Como resultado de esta penalización, seleccionaron distribuciones de 2p con mayor frecuencia que distribuciones de 3p. Sin embargo, no es evidente que

esta tendencia a seleccionar modelos más sencillos implique una superioridad de los criterios CIA y CIB, ya que podrían estar sacrificando el nivel de ajuste.

Así pues, los criterios de selección no resuelven el problema de saber cuál es la distribución más adecuada, ya que con frecuencia seleccionan distribuciones de probabilidad diferentes, y no se sabe cuál es el mejor criterio de selección. Más aún, los cuantiles estimados por medio de las diferentes distribuciones pueden ser muy diferentes entre sí. Si al menos tres de los criterios coinciden, se puede adoptar la regla de seleccionar la distribución con el mayor número de coincidencias, ya que los criterios coincidentes se respaldan entre sí. Pero es de esperarse que la mayoría de las veces esto no suceda. Por lo tanto, se requieren estudios adicionales que demuestren la superioridad de alguno de los métodos sobre los demás.

CONCLUSIONES

Los criterios CIA y CIB coincidieron en gran medida en cuanto a la selección de la distribución más adecuada para el análisis de frecuencias. Sin embargo, estos criterios generalmente no coincidieron con los resultados de los criterios KS y EEA, que tampoco coincidieron entre sí con mucha frecuencia. En la mayoría de los casos, los criterios CIA y CIB seleccionaron distribuciones de dos parámetros, lo cual es consistente con el hecho de que estos criterios penalizan el número de parámetros de una distribución. En cambio, los criterios KS y EEA, que sólo toman en cuenta el nivel de ajuste, seleccionaron principalmente distribuciones de tres parámetros.

De acuerdo con CIA y CIB, las distribuciones de la familia LP3 fueron las que mejor se ajustaron a las series analizadas en la mayoría de las ocasiones. En cambio, según las

pruebas de KS y EEA, las distribuciones de la familia GVE fueron las de mejor ajuste en la mayoría de los casos.

Se observaron algunos casos en que los cuantiles estimados por medio de distribuciones diferentes, seleccionadas por criterios diferentes, fueron muy similares. Sin embargo, hubo más casos en que los cuantiles estimados por medio de distribuciones diferentes presentaron diferencias importantes para los periodos de retorno más grandes.

En los casos en que al menos tres de los criterios coinciden, se puede adoptar la regla de seleccionar como distribución de mejor ajuste a aquella que presente el mayor número de coincidencias. Cuando esto no sucede, permanece la duda acerca de cuál es la mejor distribución para el análisis de frecuencias. La elección de una distribución incorrecta puede conducir a graves errores de estimación, y por lo tanto, a sub o sobre diseñar una obra hidráulica. Por lo tanto, se requieren estudios adicionales que verifiquen cuál de estos criterios tiene el mayor poder de selección, esto es, la mayor capacidad para seleccionar la distribución correcta.

REFERENCIAS

- [1] V. T Chow, D. R. Maidment and L. W. Mays. Hidrología aplicada. McGraw-Hill Interamericana, p. 584. 1994.
- [2] R. Flowers Cano, R. Flowers y J. Rivera Trejo, "Evaluación de criterios de selección de modelos probabilísticos: validación con series de valores máximos simulados," *Tecnol Cienc Agua*, V (5), 189-197. 2014.
- [3] B. W. Silverman. Density estimation for statistics and data analysis. Chapman & Hall. London, UK, 1986.
- [4] Y. Mon and U. Lall, "Kernel quantile function estimator for flood frequency analysis," *Water Resour Res*, Volumen 30, 3095-3103. 1994.
- [5] I. Cruz y F. García. Modelación de los escurrimientos del Río Yaqui. Instituto Tecnológico de Sonora, p. 145–152. 2008.
- [6] R. Maidment. Handbook of hydrology. David R Maidment, editor in chief. McGraw-Hill, p. 1424. 1992.
- [7] H. Aksoy, "Use of Gamma Distribution in Hydrological Analysis," *Turkish J Eng Env Sci*, XXIV (6), 419-428. 2000.
- [8] M. F. Aparicio. Fundamentos de Hidrología de Superficie. México D.F., México: Limusa Noriega Editores, p. 304. 2005.
- [9] J. Heo, D. Boes and J. Salas, "Regional flood analysis based on a Weibull model: Part 1. Estimation and asymptotic variances," *J Hydrol*, 242(3-4), 157-170. 2001a.
- [10] J. Heo, J. Salas, and D. Boes, "Regional flood frequency analysis based on a Weibull model: Part 2. Simulations and applications," *J Hydrol*, 242(3-4), 171-182. 2001b.
- [11] A. Gutiérrez y R. A. "Predicción hidrológica mediante el Método de la Avenida Índice para dos poblaciones," *Ing Hidraul Mex*, XX (2), 37-4. 1993.

- [12] H. Shin, Y. Jung, C. Jeong and J. Heo, "Assessment of modified Anderson-Darling test statistics for the generalized extreme value and generalized logistic distributions," *Stoch Environ Res Risk Assess*, Volumen 26, 105-114. 2011.
- [13] F. Ganancias. Cátedra de Hidrología y Procesos Hidráulicos. Clase de Estadística Hidrológica, Córdoba, Argentina: Universidad Autónoma de Córdoba, Facultad de Ciencias Exactas, Físicas y Naturales. 2009.
- [14] F. Laio, "Cramer-von Mises and Anderson-Darling goodness of fit tests for extreme value distributions with unknown parameters," *Water Resour Res*, Volumen 40, 10. 2004.
- [15] J. Suhaila and A. Jemain, "Fitting daily rainfall amount in Malaysia using the normal transform distribution," *J Appl Sci*, 7(14), 1880-1886. 2007.
- [16] S. Dan´Azumi, S. Shamsudin and A. Aris, "Modeling the Distribution of Rainfall Intensity using Hourly Data," *Am J Environ Sci*, 6(3), 238-243. 2010.
- [17] K. Atroosh and A. Moustafa, "An Estimation of the Probability Distribution of Wadi Bana Flow in the Abyan Delta of Yemen," *J Agr Sci*, 4(6), 80-89. 2012.
- [18] H. Mitosek, W. Strupczewski, and V.Singh. Toward an objective choice of an annual flood peak distribution. *Memorias de la 5th ICHE, CD Proceedings, Track C*, PDF 158, Warsaw, Poland, p. 16. 2002.
- [19] F. Laio, G. Di Baldassarre and A. Montanari, "Model selection techniques for the frequency analysis of hydrological extremes," *Water Resour Res*, Volumen 45, p. 11. 2009.
- [20] J. Raynal y L. García, "Análisis de caudales máximos anuales usando la distribución GVE para tres poblaciones," *Inf Technol*, 16(4), 69-75. 2005.
- [21] A. Breña, "Modelos Probabilísticos de Valores Máximos en Hidrología. Un Nuevo enfoque," *Contactos*, Volumen 66, 54-60. 2007.
- [22] D.F. Campos-Aranda, "Ajuste de las distribuciones GVE, LOG y PAG con momentos L depurados (1,0)," *Tecnol Cienc Agua*, VI (4), 153-167. 2015.

- [23] A. Ferriño y D. Burgos, "Delimitación de zonas federales y áreas de amortiguamiento en ríos afectados por el crecimiento urbano como estrategia para prevenir inundaciones," *Epistemus*, Volumen 19, 24-33. 2015.
- [24] Conagua. Base de datos climatológica nacional (Sistema CLICOM). [En línea] Disponible en: <http://clicom-mex.cicese.mx>. 2014.
- [25] R. Linsley, M. Kohler and J. Paulus. *Hidrología para ingenieros*. 2 ed. México: McGraw-Hill Interamericana S.A. 1988.
- [26] N. Gordon. *Stream Hydrology: An Introduction for Ecologists*. Second ed. Chichester, West Sussex, England: Wiley & Sons Ltd. 2004.
- [27] S. El Adlouni, B. Bobée and T. Ouarda, "On the tails of extreme event distributions," *J Hydrol*, Volumen 355, 16-33. 2008.
- [28] S. El Adlouni, and T. Ouarda, "Comparaison des méthodes d'estimation des paramètres du modèle GEV non-stationnaire," *Rev Sci Eau*, 21(1), 35-50. 2008
- [29] V. M. Ponce. *Engineering Hydrology, Principles and Practices*. Prentice Hall, p. 640, 1989.
- [30] A. R. Rao and K. H. Hamed. *Flood Frequency Analysis*. Chapter 7: Extreme value distributions. CRC Pres. Boca Raton, Fla, 207–223. 2000.
- [31] G. Kite. Chapter 12: Comparison of Frequency Distributions. In: *Frequency and Risk Analyses in Hydrology*. Colorado, USA: Water Resources Publications, 156-168. 1977.
- [32] H. Akaike. Information theory and an extension of the maximum likelihood principle. En: B. Petrov & F. Csaki, edits. *Second International Symposium on Information Theory*. Budapest: Acad. Kiadó, 267-281. 1973.
- [33] G. Schwarz. Estimating the dimension of a model. *Ann. Stat.*, Volumen 6, doi:10.1093/biomet/72.1.67, 461-464. 1978.
- [34] A. DiDonato and A. Morris, "Computation of the incomplete gamma function ratios and their inverse. ACM," *Trans Math Softw*, Volumen 12, 377-393. 1986.