ESTIMACIÓN DE EXTREMOS DE PRECIPITACIÓN DIARIA A PARTIR DE UN PROCESO ESTOCÁSTICO DE CLÚSTER

José Antonio López Díaz[(1)]
[(1)] AEMET, jlopezd@aemet.es

En este trabajo se valora la utilidad de los procesos estocásticos de cluster de tipo Neyman-Scott Rectangular Pulses con dos tipos de célula para la estimación de la distribución de los extremos de precipitación máxima diaria. Este tipo de procesos estocásticos ofrecen la posibilidad de tener en cuenta los dos tipos principales de precipitación en nuestras latitudes: la convectiva y la frontal. El ajuste se ha efectuado utilizando la serie horaria de precipitación en el observatorio de Daroca por tener esta una longitud de más de 20 años.

1. DATOS

Los datos utilizados para este estudio han sido los de precipitación horaria en el observatorio de Daroca (Zaragoza) procedentes del banco de datos climatológico de AEMET. Este observatorio se ha seleccionado para este estudio por la longitud de su serie de precipitación horaria. Los datos horarios de precipitación comienzan en 1982 en Daroca. Se han seleccionado solo aquellos meses que no presentaban ninguna laguna en los datos horarios. Han resultado para los sucesivos meses del año un número de meses completos de 25, 27, 30, 26, 30, 29, 29, 29, 29, 29, 27 y 24.

2. METODOLOGÍA

2.1 El proceso estocástico *GNSRP(2)*

Se define a continuación el proceso puntual de clúster generalizado de Neyman-Scott de pulsos rectangulares con dos tipos de célula *GNSRP(2)* (Cowpertwait, 1994). Suponemos que los orígenes de las tormentas ocurren según un proceso de Poisson de tasa (por unidad de tiempo) λ y que un número aleatorio C de orígenes de células se asocia con cada origen de tormenta. En este trabajo la distribución de C se ha supuesto geométrica con parámetro v, por tanto con media v^{-1}. Recordemos que para un proceso de Poisson de tasa λ en cada intervalo temporal infinitesimal δt la probabilidad de que caiga un origen de tormenta es $\lambda\ \delta t$, y los sucesos correspondientes a intervalos distintos son independientes. Los orígenes de células están retrasados de sus orígenes de tormenta por distancias que son variables aleatorias exponenciales con parámetro β. Cada célula independientemente es clasificada como de tipo *I* o *II*, con probabilidades respectivas α_1 y $\alpha_2 = 1 - \alpha_1$. Un pulso rectangular de lluvia es asociado independientemente con cada origen de célula. La duración del pulso es una variable exponencial independiente de parámetro η_i para el tipo i de célula, $i = 1, 2$. La intensidad del pulso, notada X_i, es también una variable exponencial independiente de parámetro ι_i según el tipo de célula, $i = 1, 2$.

En la referencia citada antes de Cowpertwait (1994) se derivan diversos estadísticos para este tipo de procesos en función de los parámetros del mismo. El procedimiento para el ajuste a los datos consiste en minimizar una función cuadrática que mide la desviación entre los valores teóricos de distintos estadísticos (fundamentalmente momentos de primer y segundo orden de varias magnitudes) y los valores empíricos de esos estadísticos derivados a partir de los datos. Esto se describe en el siguiente apartado.

2.2 Función de coste para el ajuste a los datos del proceso *GNSRP(2)*

Para cada mes del año se ha procedido a ajustar un proceso *GNSRP(2)* a los datos de precipitación horaria de Daroca. Para especificar el proceso *GNSRP(2)* son necesarios 8 parámetros: λ, v, β, α_1, η_1, ι_1, η_2, ι_2. Para medir el grado de disparidad entre el proceso estimado *GNSRP(2)* y los datos de precipitación se ha utilizado una suma de cuadrados de errores relativos de estadísticos del proceso agregado, de la forma:

$$SS = \sum_{h\in H} \sum_{g_i\in G} \left\{ \left(1-\frac{g_i(h)}{\hat{g}_i(h)} \right)^2 + \left(1-\frac{\hat{g}_i(h)}{g_i(h)} \right)^2 \right\} \quad (1)$$

donde G es un conjunto de propiedades estadísticas agregadas para el modelo *GNSRP(2)*, $\hat{g}_i$ denota el estimador muestral (a partir de los datos) del estadístico teórico del proceso g_i y H es un conjunto de niveles de agregación temporal.

Existen claramente muchas posibilidades para escoger G y H. En este trabajo se eligieron las siguientes: media de 1 h, para las agregaciones de 1, 3, 6, 12 y 24 h las varianzas, la proporción de intervalos secos y la probabilidad de transición de húmedo a húmedo.

El cálculo del valor teórico de este último parámetro (la probabilidad de transición de húmedo a húmedo) requiere la evaluación numérica de una integral que

dio problemas de convergencia que conducían a resultados incoherentes. Esto se pudo solventar desarrollando una cota superior para el error de esta integral.

2.3 Algoritmo numérico de optimización

Para minimizar (1) respecto a los 8 parámetros del proceso *GNSRP(2)* desconocidos se ha utilizado el algoritmo de Nelder -Mead. Este método usa el concepto de simplex, que es una generalización de una línea en dos dimensiones, un triángulo en tres o un tetraedro en cuatro dimensiones, y es un politopo de N+1 vértices en N dimensiones. El algoritmo de Nelder-Mead genera una nueva posición de prueba a partir del comportamiento de la función objetivo en los vértices del simplex. Por ejemplo se puede reemplazar el peor punto del simplex por medio de una reflexión respecto al centroide de los puntos restantes del simplex. Si este punto es mejor que los anteriores se alarga exponencialmente a lo largo de esta línea. En cambio si el nuevo punto no mejora sustancialmente los anteriores, entonces probablemente estemos en una zona de valle, y lo que hacemos es encoger el simplex hacia un nuevo punto.

Realizando diversas pruebas de minimización de la función de coste (1) para distintos meses partiendo de distintas condiciones iniciales se vio que el mínimo obtenido variaba, por lo que se ha utilizado un procedimiento de búsqueda aleatoria previa de condiciones iniciales favorables antes de lanzar el algoritmo numérico de optimización. Este proceso se describe en el próximo apartado.

3. AJUSTE A LOS DATOS

El primer paso en el ajuste a los datos consistió en buscar un ajuste óptimo de un proceso *GNSRP(1)*, que tiene un solo tipo de célula, y de un *GNSRP(2)*. La idea es que, en teoría, si el ajuste a ambos modelos fuera el óptimo, el modelo con una sola célula, al ser una caso particular del modelo con dos tipos de célula, debería de dar un valor de la función cuadrática de ajuste de (1) mayor o igual que el modelo completo con dos células. De esta forma se puede comprobar si el ajuste óptimo es realizable. Además el ajuste al modelo más sencillo *GNSRP(1)* debe ser más estable al reducirse el número de parámetros que hay que ajustar de 8 a 5.

Para cada mes del año se ha procedido a ajustar los modelos *GNSRP(1)* y *GNSRP(2)* a los datos horarios con una búsqueda previa aleatoria de condiciones iniciales favorables y posterior búsqueda del óptimo de la función de coste con el algoritmo de Nelder-Mead. Los términos empíricos en la función de coste (1) se han calculado a partir de los datos horarios de cada mes concatenando años sucesivos. Para el proceso *GNSRP(2)* se ha evaluado la función de coste (1) 1000 veces variando los parámetros del proceso aleatoriamente con condiciones realistas de rangos de variación de cada parámetro. Para el *GNSRP(1)* se ha usado un procedimiento similar.

Una vez calculado el valor óptimo de las 1000 simulaciones previas aleatorias, se ha usado esa solución como condición inicial para lanzar el algoritmo numérico de Nelder-Mead.

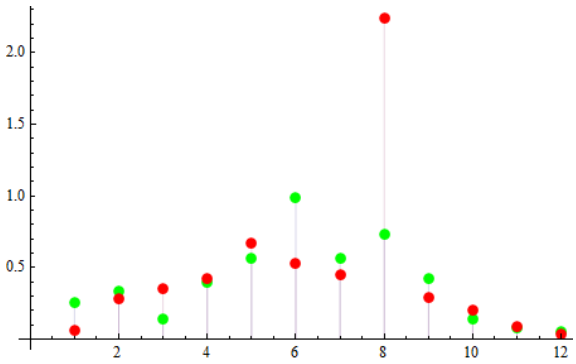

Fig. 1: función de coste (1) para cada mes con el ajuste inicial de *GNSRP(1)* (verde) y *GNSRP(2)* (rojo).

Los valores de la función de ajuste (1) para estos ajustes iniciales se muestran en la fig. 1. Vemos que para algunos meses se viola el orden de errores entre los dos tipos de proceso, con una y dos células, notoriamente en agosto.

A la vista de estos problemas se ha procedido a ajustar un *GNSRP(2)* partiendo de las condiciones iniciales dadas por el ajuste a *GNSRP(1)* anterior para los meses del 2 al 10. Con esto se ha mejorado notablemente el ajuste final a *GNSRP(2)*, obteniéndose los errores cuadráticos de la figura 2.

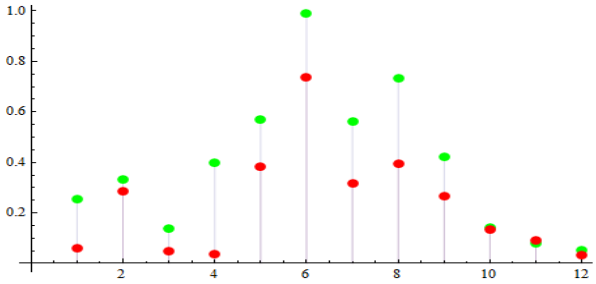

Fig.2 : Función de coste (1) para ajuste de *GNSRP(1)* (verde) y *GNSRP(2)* (rojo) en la segundo fase.

Nótese en la figura 2 la considerable disminución del error para ajuste a *GNSRP(2)* en agosto, abril o marzo. Ahora ya todos los errores del ajuste para el modelo más complejo son inferiores que para el modelo más simple.

Para decidir en cada mes con qué modelo quedarnos se ha tenido en cuenta que el modelo más complejo debe disminuir de forma aceptable la función de

coste para ser preferible al modelo más simple. Se ha usado el criterio de considerar aceptable para un mes el modelo *GNSRP(1)* si el valor de la función de coste (7) es inferior a 0.1, y si el valor de (7) es inferior a 1.2 veces el valor de (7) para dos células. En caso contrario se toma el *GNSRP(2)*. Con este criterio la elección para cada mes queda: ene (2), feb(1), mar(2), abr(2), may(2), jun(2), jul(2), ago(2), sep(2), oct(1), nov(1), dic(1).

Pero falta un último paso, porque analizando las probabilidades de los dos tipos de células en los meses con modelo *GNSRP(2)* en la lista anterior se ve que marzo y abril dan una probabilidad muy pequeña a unos de los tipos de célula, del orden de un 0.1 %. Por tanto es natural probar si reduciendo el modelo a *GNSRP(1)*, eliminando el tipo de célula de baja probabilidad, el valor de la función de coste (1) no se incrementa. Esto se comprobó que era cierto, y por tanto estos dos meses pasaron a *GNSRP(1)*, con lo que la asignación final de modelos a cada mes queda como se recoge en la figura 3.

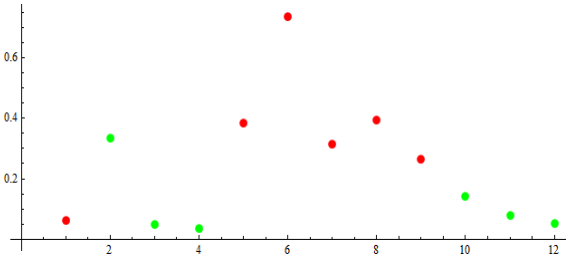

Fig. 3: Función de coste (1) para ajuste de *GNSRP(1)* (verde) y *GNSRP(2)* (rojo) en el ajuste final.

El patrón que muestra la figura 3 en cuanto a la selección final de un modelo con una o con dos tipos de célula tiene bastante coherencia, ya que son los meses de carácter veraniego (de mayo a septiembre) los que ajustan mejor con dos tipos de célula. Esto indicaría que las precipitaciones convectivas de esos meses requieren un segundo tipo de célula que no es estrictamente necesaria en los otros meses del año. El caso de enero, que en el ajuste final aparece con dos tipos de célula, según el criterio de elección entre modelos antes enunciado, no desentona en realidad tanto si tenemos en cuenta que para este mes el modelo con un solo tipo de célula ya da un error de ajuste pequeño, como muestra la figura 2.

4. COMPROBACIÓN DEL AJUSTE CON LA PRECIPITACIÓN TOTAL ANUAL

En la figura 4 se muestran las funciones de densidad del total anual de precipitación, estimadas a partir de un kernel gaussiano, para los datos observados en los últimos 30 años (rojo) y para los datos simulados a lo largo de 1000 años con el *GNSRP* con una o dos células en cada mes, según el ajuste óptimo descrito en la sección anterior, ajustado a los datos (azul). Se aprecia que la densidad simulada es más simétrica que la obtenida a partir de los datos, y tiene una cola izquierda bastante menos larga que la observada. Esta insuficiente representación de los años más secos en el proceso ajustado se podría explicar por el hecho de que el ajuste se ha hecho mes a mes, y por tanto el total anual simulado no puede capturar las correlaciones entre meses que sin duda están presentes en los años más secos (al combinarse los meses del año de forma aleatoria en la simulación). Esto también explicaría la mayor simetría. En cambio la cola derecha, correpondiente a años muy húmedos, sí que encaja muy bien con la observada.

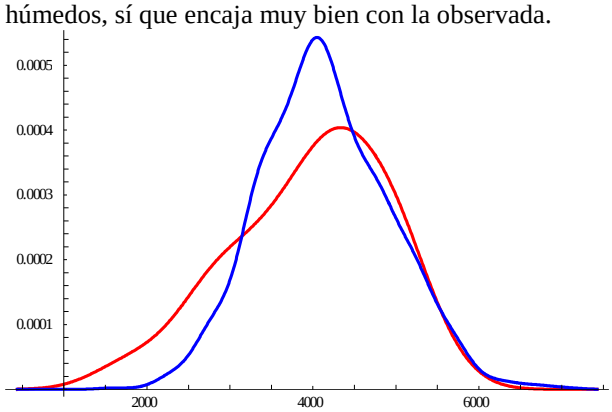

Fig.4.- Funciones de densidad de la precipitación total anual obtenidas aplicando un kernel gaussiano. En rojo función de densidad de los totales observados de los últimos 30 años en Daroca, en azul función de densidad a partir de 1000 años de simulación del GNSRP(2) ajustado.

5. ESTIMACIÓN DE EXTREMOS

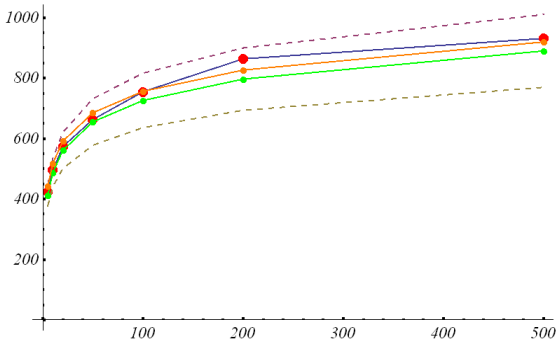

Fig. 5.- Valores de la precipitación máxima diaria anual para los tiempos de retorno en años en abscisas. La línea verde corresponde a ajuste de Gumbel con la serie de datos observados de los últimos 30 años, la línea naranja a ajuste Gumbel con los últimos 100 años de datos, la línea azul con puntos rojos estimación a partir de simulación de GNSRP(2) ajustado a cada mes.

Una de las posibles aplicaciones de este tipo de procesos ajustados a la precipitación estriba en obtener una estimación de los valores extremos de la precipitación directamente a partir de la simulación estocástica del GNSRP(2) ajustado, en lugar de por medio de la estimación paramétrica de

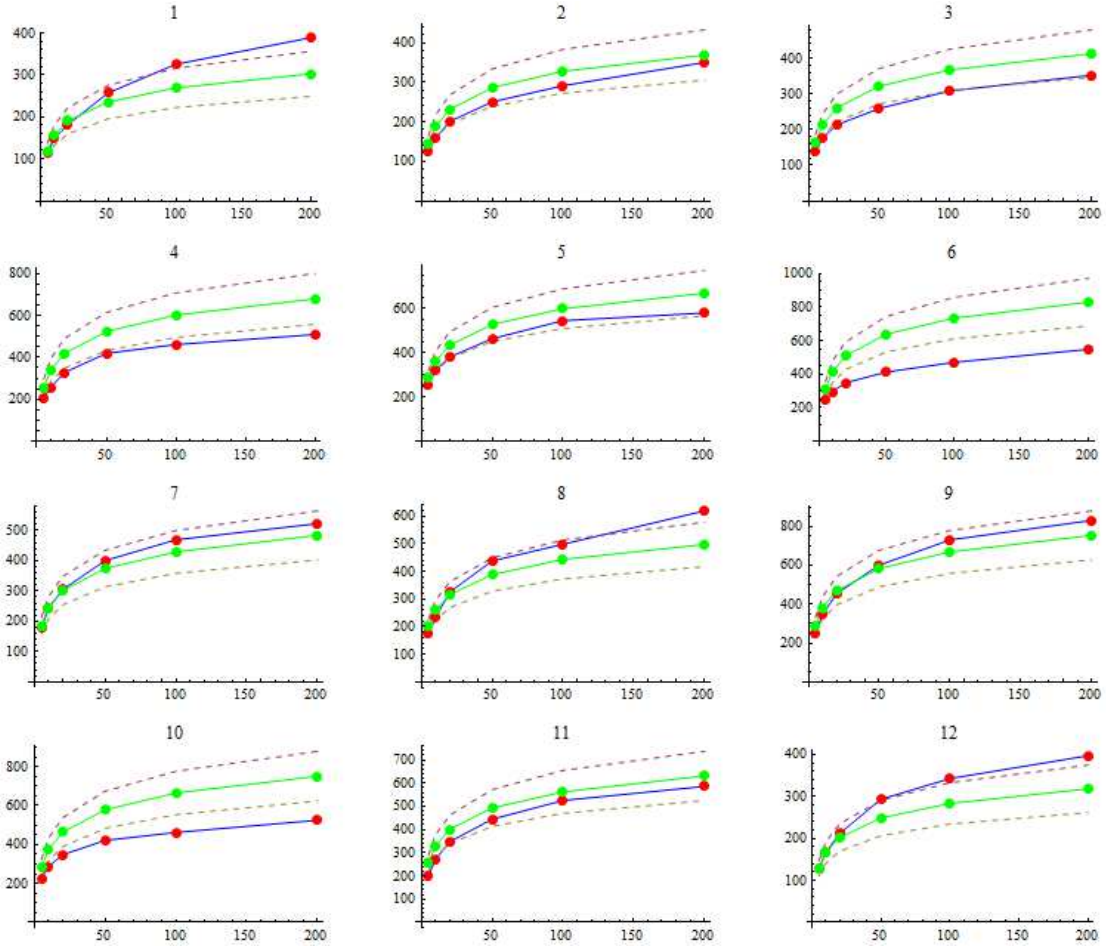

Fig. 6: Valores de la precipitación máxima diaria anual para los tiempos de retorno en años en abscisas. La línea verde corresponde a ajuste de Gumbel con la serie de datos observados de los últimos 30 años, la línea azul con puntos rojos estimación a partir de simulación de GNSRP(2) ajustado a cada mes.

los extremos postulando una función de densidad para los extremos como la función general de extremos. Además el proceso simulado puede proporcionar estadísticas de la distribución de muchas variables extremas, como totales diarios en el año o en meses concretos, o máximos en periodos superiores a un día, o en periodos inferiores como 6 horas. Con el ajuste del GNSRP(2) a los datos los estadísticos extremales deseados se estiman simulando el proceso ajustado un número grande de veces y computando los estadísticos empíricos de la simulación.

Para valorar estas posibilidades en la figura 5 se comparan las distribuciones extremas de la precipitación diaria máxima anual obtenidas de tres formas. Las líneas verde y naranja corresponden a ajuste de la distribución extrema Gumbel a las series de datos observados de precipitación máxima diaria anual en los últimos 30 años y con la serie larga de Daroca, desde 1910, respectivamente. Ambos ajustes muestran un grado de concordancia elevado según la figura. Además se ha representado en azul con puntos rojos el ajuste empírico a partir de la aaaaa simulación de 1000 años de precipitación horaria antes mencionado usando el proceso GNSRP(2). Queda patente el excelente acuerdo de las tres estiamaciones incluso para periodos de retorno de 500 años, el máximo representado en la figura 2, que exceden ampliamente el rango de datos observado. Incluso para este periodo de retorno tan elevado parece que el GNSRP(2) se acerca más al valor estimado con la serie larga que al estimado con la serie corta, lo cual podría indicar habilidad del modelo estocástico para producir características de la distribución extrema que la serie de máximos anuales corta, a lo largo del periodo de ajuste, desdibuja.

En la figura 6 se representan los mismos ajustes para cada mes del año. Vemos que hay en general una buena concordancia entre los valores estimados para los diferentes periodos de retorno tanto a partir de los datos con ajuste Gumbel como mediante simulación del proceso estocástico. En los meses de junio y octubre la concordancia es algo peor, con el proceso estocástico dando valores menores para los valores de retorno que el ajuste Gumbel. Como

muestra la figura 3 junio fue el mes con mayor error en el ajuste.

6. REFERENCIAS

O'Neill, R. (1971). “Algorithm AS 47: Function Minimization Using a Simplex Procedure”. Journal of the Royal Stat. Soc. Series C (Applied Statistics), Vol. 20, No. 3 , pp. 338-34

Cowpertwait, P.S.P. (1994). “A Generalised Point Process Model for Rainfall”. Proc. R. Soc. Lond. A 447, 23-27

Cox, D. R., y Isham, V. (1980). *Point Processes*. Chapman and Hall, 188 pp.

Morrissey, M. L. (2009). “Superposition of the Neyman-Scott Rectangular Pulses Model and the Poisson White Noise Model for the Representation of Tropical Rain Rates”. J. Of Hydrometeorology, vol. 10, pp. 395- 411.

Rodríguez-Iturbe, I.; Cox, D. y Isham, V. (1987). “Some models for rainfall based on stochastic point processes”. Proc. Roy. Soc. London, A410, 269–288.