

---

# ANÁLISIS Y VISUALIZACIÓN DE DATOS - MENTORÍA: CUÁNDO LLUEVE?

---

**Álvarez, Javier**

javieralvarez.ic@gmail.com

**Ferraro, María Eugenia**

ferraro.mariaeugenias@gmail.com

**Trógolo, Nair**

natrogolo2002@gmail.com

**Villarreal D'Angelo, Juan Manuel**

juanmav07@gmail.com

31 de mayo de 2019

## RESUMEN

En el presente trabajo se llevó a cabo un análisis estadístico sobre registros de precipitaciones en la localidad de Teodelina, provincia de Santa Fe. Se realizó un análisis exploratorio de los datos, encontrando principalmente que existe un periodo de precipitaciones que alterna con un periodo de sequía. Se analizaron las distribuciones mensuales y anuales acumuladas y se reagruparon los datos con el fin de buscar posibles comportamientos en las precipitaciones, así como correlaciones temporales.

## 1. Introducción

Teodelina es una comuna ubicada al sur de la provincia de Santa Fe, a 3 km del límite con la provincia de Buenos Aires. Dicha comuna está circundada por varios arroyos (ver figura 1) afluentes de la *Laguna El Chañar*, considerada como nacimiento del *Río Salado* cuyas aguas desembocan en la Bahía Samborombón.

Dada su ubicación hidrológica, la comuna de Teodelina ha sufrido los estragos de las precipitaciones en varias ocasiones, tanto en las regiones urbanas, como en las rurales donde se desempeña la agricultura. Por ello, el análisis estadístico de las precipitaciones en la región otorga una base que permitirá, por un lado, justificar la petición de fondos para obras que ayuden a evitar inundaciones y por otro, generar modelos predictivos de probabilidad de ocurrencia de lluvias y de su intensidad, con el fin de alertar a la comunidad y agricultores.

El presente trabajo lleva a cabo un análisis exploratorio del registro de precipitaciones en la comuna de Teodelina entre el 1 de Enero de 1978 hasta el 28 de febrero de 2019. Este proyecto se encuentra enmarcado en el primer trabajo práctico relacionado con la mentoría "¿Cuándo llueve?".



Figura 1: Imagen satelital de la localidad de Teodelina y los arroyos que confluyen a la Laguna El Chañar.

|            | rain |
|------------|------|
| date       |      |
| 1978-01-01 | 0    |
| 1978-01-02 | 0    |
| 1978-01-03 | 0    |
| 1978-01-04 | 0    |
| 1978-01-05 | 0    |
| 1978-01-06 | 0    |

Figura 2: Visualización parcial del *Dataset*.

## 2. Datos

El presente análisis utiliza un *Dataset* (ver datos aquí) de registros de lluvia que datan desde el primero de enero de 1978 a la fecha, generado por la familia Prone en la comuna de Teodelina. La estructura del mismo (ver Figura 2) consiste de dos columnas, la columna *date* de tipo índice, con formato *datetime64*, caracteriza a fechas con resolución temporal de un día, y la columna *rain*, con formato *int64*, contiene el registro de precipitaciones en milímetros, para cada fecha. El mismo presenta fechas faltantes, ni celdas con datos *NaN*. El *Dataset* abarca 15034 días de los cuales, por lo menos el 75 % no registra lluvia, lo que induce un valor medio de 3 mm con una dispersión de 12 mm (tabla 1). Si en su lugar se analizan solo los días en los que hubo precipitaciones (columna *SubDataset* de la Tabla 1), se obtiene una muestra de 2459 datos (~ 16 %) de los cuales, por lo menos el 75 % registra valores menores que 29 mm y el 50 % menores a 13 mm (mediana), aunque existen días que han alcanzado a obtener más de 200 mm de precipitaciones, lo cual se ve reflejado en el valor medio con un valor mayor al de la mediana.

A modo de exponer una primera visualización de los datos, la Figura 3 muestra el histórico de lluvias diarias registradas en el *Dataset* bajo estudio. Cada punto azul se corresponde con el valor

|        | <i>Dataset</i> | <i>SubDataset</i> |
|--------|----------------|-------------------|
| N días | 15034          | 2459              |
| media  | 3              | 20                |
| sigma  | 12             | 21                |
| min    | 0              | 1                 |
| 25 %   | 0              | 5                 |
| 50 %   | 0              | 13                |
| 75 %   | 0              | 28                |
| max    | 220            | 220               |

Tabla 1: Principales estadísticos que caracterizan el Dataset de estudio.

en milímetros, de precipitaciones en un día. De la figura se puede observar la existencia de una periodicidad o ciclo de lluvias. Estos ciclos han sido delimitados por líneas verticales que abarcan doce meses (año hidrológico de aquí en adelante). Se considera el inicio de dicho año como el primero de junio de cada año debido a que son los meses en que menos precipitaciones acumuladas mensuales se observan en el año (tal como veremos más adelante en la tabla 2).

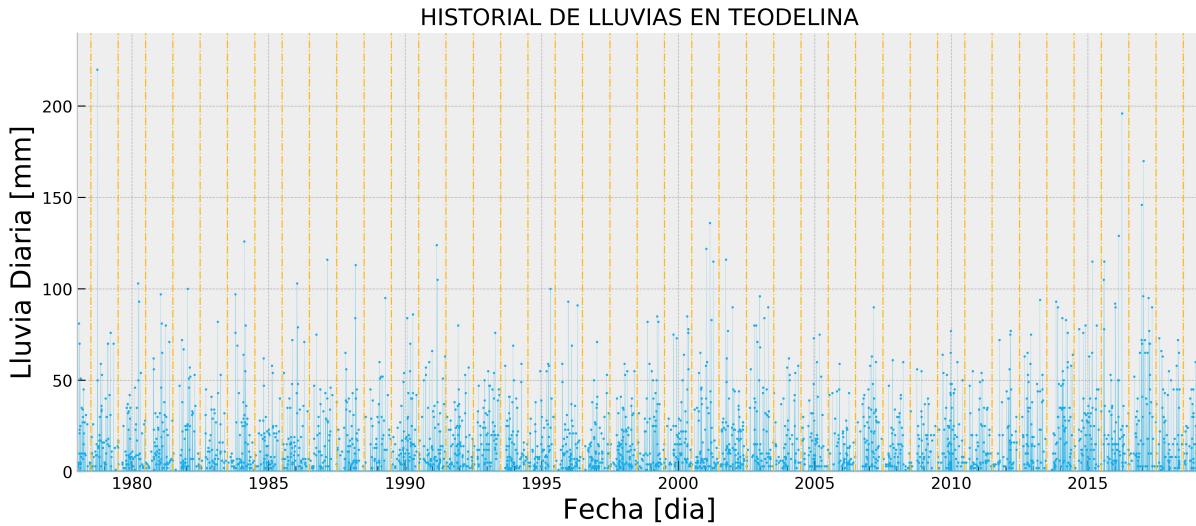


Figura 3: Registro histórico de lluvias en Teodelina. Puntos celestes: registro diario de lluvia en milímetros de agua almacenada. Líneas verticales: demarcación visual de ciclo aparente en el comportamiento de las precipitaciones.

### 3. Procesamiento

El gráfico de la sección anterior, que muestra el histórico de lluvias (Figura 3) almacena mucha información. La visualización del mismo no permite extraer grandes conclusiones más allá del ciclo que se puede observar, por lo tanto, es necesario procesar y agrupar los registros con el fin de poder exponer, si existen, relaciones o información con mayor fuerza.

A continuación se llevan a cabo distintas instancias de tratamiento de los datos con el fin de encontrar tendencias que permitan inducir conclusiones. El procesamiento básico consiste en: analizar cómo se distribuyen los datos, caracterizar los valores poco frecuentes denominados *outliers* y definir

qué hacer con los mismos. Además, se estudiará posibles distribuciones que siguen los datos y la existencia de correlaciones.

### 3.1. Registro mensuales acumulados

| Mes    | 1      | 2      | 3      | 4      | 5      | 6      | 7      | 8      | 9      | 10     | 11     | 12     |
|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|
| count  | 41.00  | 41.00  | 41.00  | 41.00  | 41.00  | 41.00  | 41.00  | 41.00  | 41.00  | 41.00  | 41.00  | 41.00  |
| mean   | 156.93 | 140.66 | 147.39 | 134.10 | 67.15  | 28.95  | 31.80  | 37.02  | 66.32  | 136.22 | 132.71 | 135.90 |
| std    | 81.75  | 82.81  | 90.11  | 84.69  | 58.81  | 26.82  | 29.55  | 47.99  | 58.96  | 70.03  | 60.13  | 84.91  |
| min    | 21.00  | 25.00  | 19.00  | 4.00   | 0.00   | 0.00   | 0.00   | 0.00   | 31.00  | 18.00  | 3.00   |        |
| 25 %   | 93.00  | 79.00  | 87.00  | 77.00  | 24.00  | 10.00  | 7.00   | 5.00   | 24.00  | 89.00  | 101.00 | 81.00  |
| 50 %   | 141.00 | 134.00 | 128.00 | 111.00 | 55.00  | 23.00  | 22.00  | 24.00  | 56.00  | 128.00 | 127.00 | 129.00 |
| 75 %   | 210.00 | 177.00 | 196.00 | 174.00 | 96.00  | 38.00  | 47.00  | 52.00  | 79.00  | 174.00 | 167.00 | 167.00 |
| max    | 331.00 | 365.00 | 367.00 | 368.00 | 298.00 | 118.00 | 120.00 | 255.00 | 303.00 | 377.00 | 266.00 | 360.00 |
| bigInf | -82.50 | -68.00 | -76.50 | -68.50 | -84.00 | -32.00 | -53.00 | -65.50 | -58.50 | -38.50 | 2.00   | -48.00 |
| bigSup | 385.50 | 324.00 | 359.50 | 319.50 | 204.00 | 80.00  | 107.00 | 122.50 | 161.50 | 301.50 | 266.00 | 296.00 |

Tabla 2: Principales estadísticos para las lluvias acumuladas mensuales entre los años 1978 y 2019.

La tabla 2, presenta los principales valores estadísticos asociados a la lluvia acumulada en cada mes para el total de los datos analizados. Es posible observar que existen dos épocas bien diferenciadas: una en la que las cantidades de precipitaciones son mayores (época humeda) y otra donde la cantidad de precipitaciones decrece de forma considerable (época seca). De la tabla se observa que en los meses de menores precipitaciones, la desviación estándar ronda en torno al valor de la media, por lo que indicaría que la precipitación en estos períodos puede alcanzar valores nulos, hasta valores de magnitud considerable.

En la figura 4, se presenta un gráfico de barras para una mejor visualización de los patrones medios de las lluvias acumuladas mensuales. A partir de la misma podemos distinguir que el periodo seco abarca los meses junio, julio y agosto, que se caracterizan por tener precipitaciones  $P < 50$  mm mensuales; dos meses de transición entre la época seca a húmeda correspondiente a los meses mayo y septiembre, donde  $50 \text{ mm} < P < 75 \text{ mm}$ ; y por último la época húmeda caracterizada por  $P > 75 \text{ mm}$  mensuales, correspondientes a los meses enero, febrero, marzo, octubre, noviembre y diciembre.

Para tener una visualización más detallada de las lluvias acumuladas de forma mensual, debemos observar la figura 5. Los boxplots son útiles a la hora de identificar outliers. En la figura podemos ver que existen ciertos años en los que -para determinado mes- las precipitaciones fueron excesivamente más altas de lo usual (outliers arriba de los boxplots), sin embargo, no existen años en los que -para un determinado mes- haya llovido una cantidad significativa menor que lo previsto, es por eso que no se observan valores atípicos en la parte inferior del gráfico.

La figura 6 muestra en la primer columna la distribución mensual acumulada del registro histórico, la segunda contiene un qqplot donde se la compara con una distribución normal, y la tercer columna la distribución para cada mes de precipitaciones entre 1978 y 2019. En la parte superior derecha de cada gráfico de la columna 1, se observan los valores resultantes de aplicar un test de prueba de hipótesis, teniendo como hipótesis nula que las precipitaciones acumuladas mensuales siguen una distribución normal. Podemos observar que en general las distribuciones por mes poseen asimetría

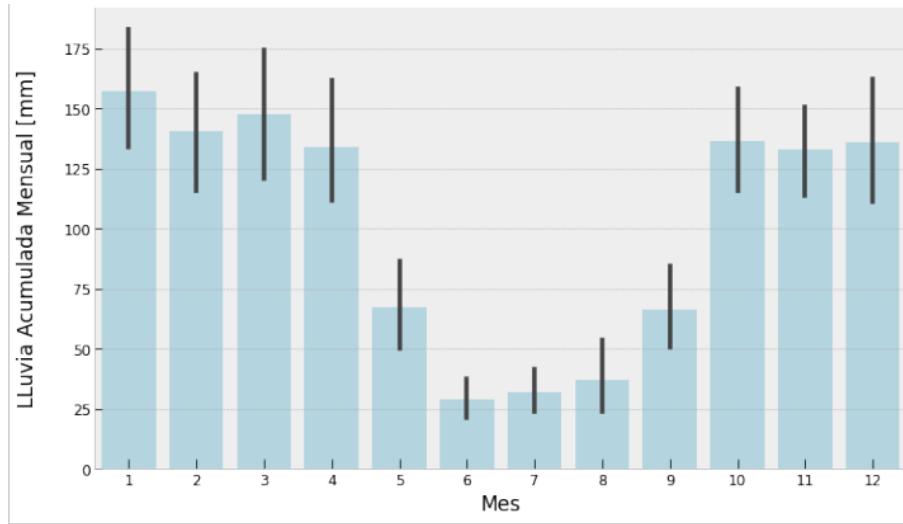


Figura 4: Promedio e intervalo de confianza de la Precipitación Acumulada Mensual.

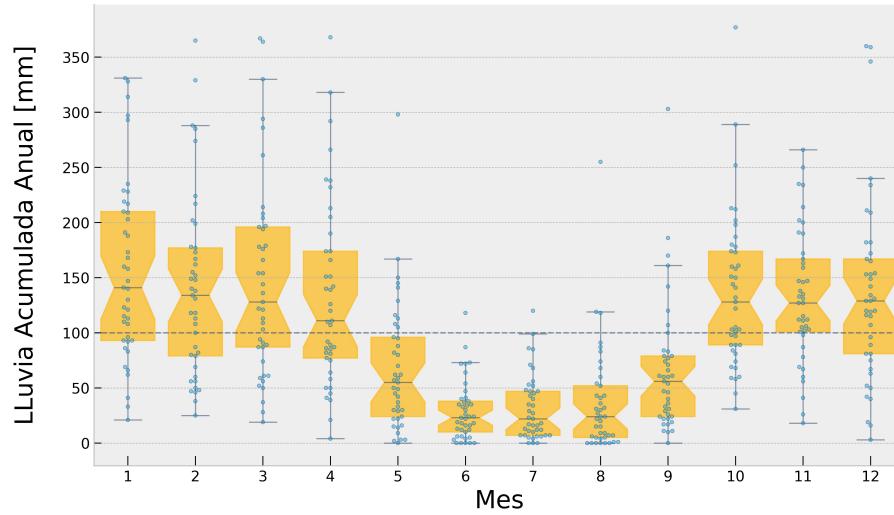


Figura 5: Lluvia acumulada mensual entre los años 1978 y 2019.

positiva, y aunque ninguna tiene un p-value igual a cero, valor que rechaza la hipótesis nula. En los gráficos QQ de la segunda columna podemos observar que ningún cuantil graficado se aproxima a la linea identidad por lo que las distribuciones si se pueden descartar de ser distribuciones normales. Por ultimo también podemos observar que no hay correlaciones entre los meses y los milímetros de lluvia caídos en ese mes, ya que la <https://www.overleaf.com/project/5cd9e6f669b9343cf7e03589s> dispersiones son grandes (excepto en los meses secos Junio, Julio, Agosto).

### 3.2. Registros anuales acumulados

En esta sección, haremos un análisis similar al anterior, pero teniendo en cuenta la cantidad de precipitaciones acumuladas anuales. Se quiere analizar si existe alguna correlación entre el avance

de los años y la lluvia total anual. En la figura 7, se muestra una gráfica de puntos de lluvia anual versus tiempo y le ajustamos una recta.

La línea de ajuste aparenta una tendencia en su media, lo cual podría interpretarse como una tendencia a que en el pasar de los años, la lluvia aumente. Sin embargo, también es posible analizar el intervalo de confianza de la recta mencionada, donde se puede observar que al inicio del período de estudio el valor mayor casi alcanzaba 1250 mm, en cambio al final del período el valor mínimo ronda en 1100 mm. Es decir, dentro del intervalo de posibles rectas sería posible ajustar una línea recta horizontal tal que la correlación entre lluvia acumulada anual y años sea nula. Por último, se concluye que no es posible afirmar que exista correlación entre la precipitación acumulada anual y el transcurso de los años.

En la figura 8, se observa que la curva de densidad de probabilidades empírica se asemeja a una distribución normal aunque su cola izquierda no se comporta como una distribución normal. Por este motivo, proponemos como hipótesis nula que las precipitaciones acumuladas anuales podrían representarse con una distribución normal sesgada hacia la izquierda. Para analizar esto, a continuación se realiza el test de Man Kendal.

Como podemos observar en la figura 9, los puntos (cuantiles) formados por los pares ordenados ( $x,y$ ) siendo  $x$  los valores correspondientes a los percentiles de una distribución normal estimada y siendo  $y$  los valores correspondientes a los percentiles de la distribución empírica (datos medidos), no forman una recta identidad por lo que esta ultima no se asemeja a una distribución normal.

### 3.3. Registros por décadas

Con el objetivo de analizar el comportamiento de las precipitaciones en distintas épocas, decidimos reagrupar los datos para cada mes en un periodo temporal de 10 años consecutivos. La figura 10a muestra boxplots construídos para cada mes entre el año 1978 y 1988, la (b) entre 1988 y 1998, la (c) entre 1998 y 2008 y un último grupo (d) que abarca los años 2008 hasta 2018. A partir de estos esquemas puede observarse con claridad la existencia de una temporada seca o de bajas precipitaciones y una temporada húmeda o de precipitaciones más intensas, tal como se mencionó con anterioridad. Los meses secos comprenden: Mayo, Junio, Julio, Agosto y Septiembre; los meses húmedos: Enero, Febrero, Marzo, Abril, Octubre, Noviembre y Diciembre; mientras que Abril-Mayo y Agosto-Septiembre pueden considerarse meses de transición de una época a la otra. En general, los meses correspondientes a la época de sequía poseen una dispersión menor en los datos que los meses donde las lluvias son más frecuentes. Solo en la primer década se observa una caída en las lluvias entre los meses Noviembre-Diciembre. Por otro lado, a medida que nos acercamos al periodo 2008-2018 podemos observar cómo aumentan las precipitaciones en la época seca y cómo meses como Mayo y Septiembre pasan a tener precipitaciones comparables con la de meses lluviosos. Otra característica sobresaliente de la última década es la cantidad de años donde las precipitaciones superan los 300 mm mensuales, comportamiento que no es posible observar en las tres décadas anteriores. En definitiva, este gráfico expone que en la última década en Teodelina, las precipitaciones mensuales se han ido incrementando y el periodo temporal de escazas lluvias se redujo.

### 3.4. Análisis del SubDataset

La Figura 11 muestra, en la parte superior, la distribución de lluvias del *SubDataset*, es decir, la distribución de los registros cuyos valores son mayores a cero y, en la parte inferior, se expone un diagrama de caja con un intervalo de confidencia del 95 %. Este diagrama muestra que la mediana corresponde al valor de 13 mm y que todos los *outliers* de la muestra se ubican a la derecha del segundo bigote, es decir, corresponden a días en los que las precipitaciones fueron inusualmente más intensas comparadas con el valor medio. La distribución observada tiene un máximo en valores de precipitaciones bajos, con una media en 20 mm y luego disminuye hasta alcanzar un valor máximo de 220 mm diarios.

Distintas distribuciones teóricas fueron graficadas para la distribución observada, tal como se muestra en la primer imagen de la figura 12. La curva continua color naranja corresponde a una distribución *lognormal*, la curva de trazos azul a una *chi-cuadrado* y la roja a trazos a una *gumbel*. En el gráfico de la derecha se exponen las distribuciones de probabilidad acumuladas para las ya mencionadas distribuciones teóricas de prueba.

Si bien visualmente al menos dos de ellas parecen guardar similitud con los datos, decidimos someterlas a un test de prueba de hipótesis. El resultado del test muestra que no existe similitud entre la distribución observada con las tres distribuciones teóricas propuestas.

Una segunda prueba se realizó mediante *qqplots*, como se observa en la figura 13. A partir de los mismos se corroboran los resultados del test de prueba de hipótesis, donde se verifica que la distribución de lluvias observadas no coincide con las distribuciones propuestas. No obstante, entre estas tres, la distribución *chi-cuadrado* es la que mejor ajusta a los valores observados; la diferencia se genera en el último cuantil que posee un valor atípico.

## 4. Conclusiones

A partir de este primer análisis exploratorio del registro histórico de las precipitaciones en la localidad de Teodelina se encontró que:

- De los datos recolectados entre el 1 de enero de 1978 y el 2 de febrero de 2019, 15034 días en total, solo el 16 % de los días se registraron precipitaciones (2459 días), mientras que el 84 % restante no se registraron lluvias.
- En un primer análisis, puede decirse que en la localidad de Teodelina llueve en promedio 3 mm diarios, con una desviación estándar de 12 mm. Sin embargo, al considerar únicamente los días que sí llueve (*sub-dataset*), el promedio asciende a 20 mm, con una desviación estándar de 21 mm.
- Históricamente, el mes más lluvioso del año -en promedio- es enero y el menos lluvioso, junio. Del mismo modo, existen tres períodos al año bien diferenciados, la época de sequía en la que llueve  $P < 50$  mm mensuales (de junio a agosto), una época húmeda con precipitaciones  $P > 75$  mm mensuales (entre octubre y marzo) y dos meses de transición entre ambos periodos con precipitaciones promedio entre 50 y 75 mm mensuales, que se

corresponde con los meses mayo y septiembre. Se evidencia un período húmedo que resulta más extenso que el periodo seco.

- Al estudiar en mayor detenimiento el sub-dataset se observa una gran cantidad de días en los que llovió una cantidad atípica (outliers). Sin embargo, no fueron quitados de la serie porque se deben analizar con especial atención, ya que la mayoría de los valores se encuentra en torno a la nulidad o unidad. Del mismo análisis, se demostró que la distribución empírica de los datos de lluvia diaria no se corresponde ni con la normal, ni lognormal, ni chi-cuadrado, ni gumbel.
- Por otra parte, las precipitaciones acumuladas anuales poseen una distribución de probabilidades empírica con un marcado sesgo hacia la izquierda, presentando una cola más prolongada a la derecha del valor medio. Por lo tanto, no se comporta como una distribución normal teórica.
- En la última década se ha tenido un comportamiento en el ciclo anual de lluvias diferente a los registrados entre 1978 y 2008. Es entonces posible observar una leve tendencia alcista en las precipitaciones totales anuales que. Sin embargo, al existir una gran dispersión en los datos, el ajuste posee una pendiente baja y no resulta posible asegurar que esta tendencia sea significativa.

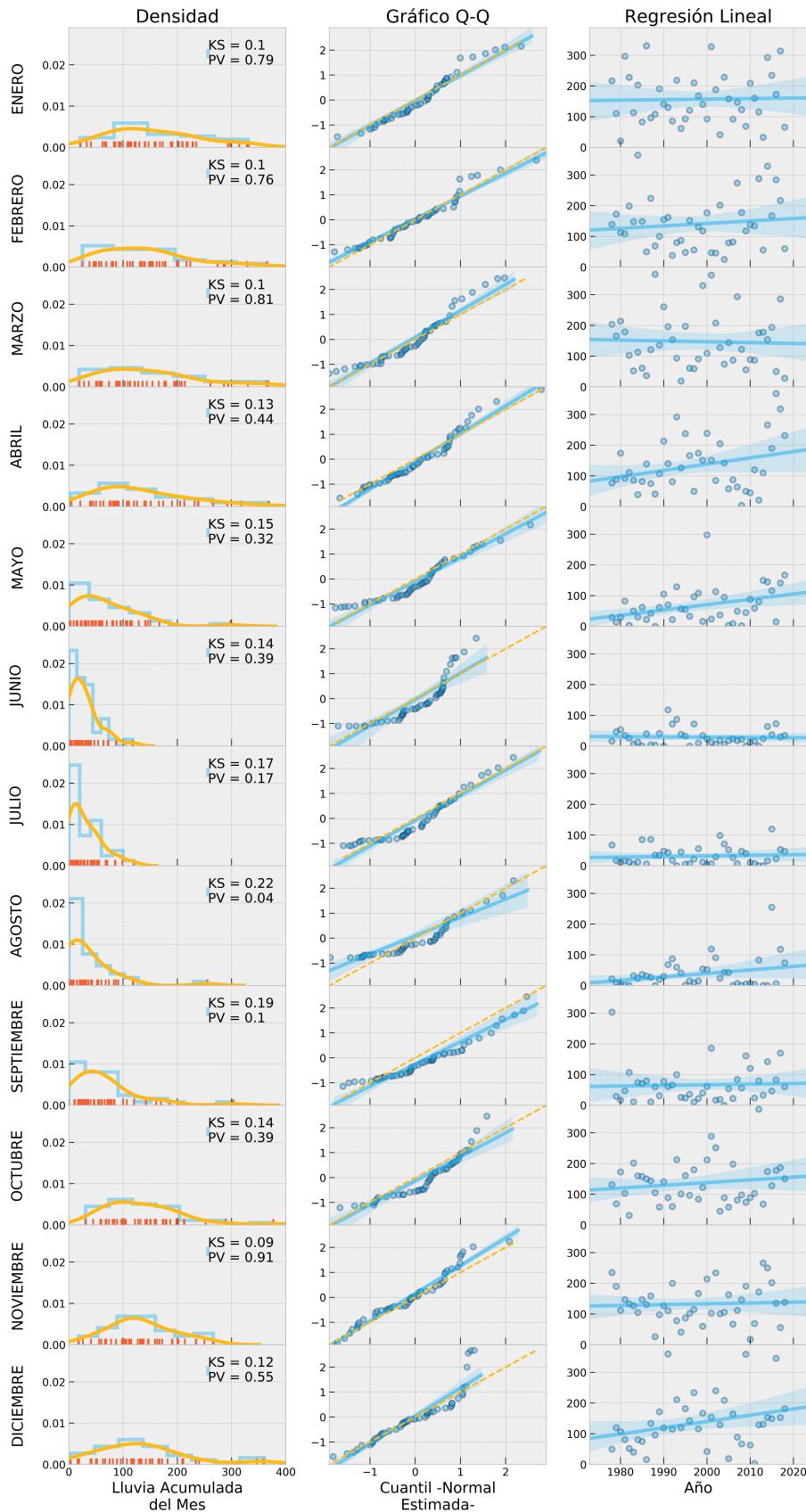


Figura 6: Análisis de las precipitaciones acumuladas para cada mes del año. Izq: Histograma y curva de densidad de probabilidades empírica asociada al histograma. Centro: Gráfico QQ para comparar la distribución de probabilidades empírica con una distribución normal teórica. Der: Precipitación acumulada mensual en función de los años. En azul, línea de ajuste de los valores graficados.

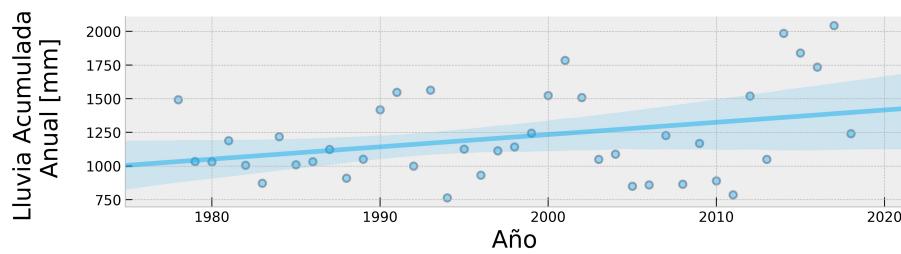


Figura 7: Precipitación acumulada anual en función de los años. En azul, línea de ajuste de los valores graficados.

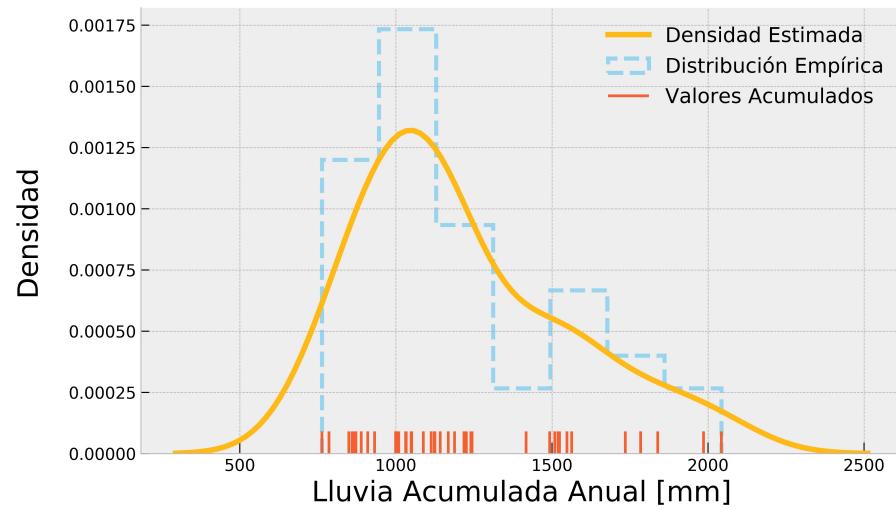


Figura 8: Histograma de precipitaciones acumuladas anuales y curva de densidad de probabilidades empírica asociada al histograma.

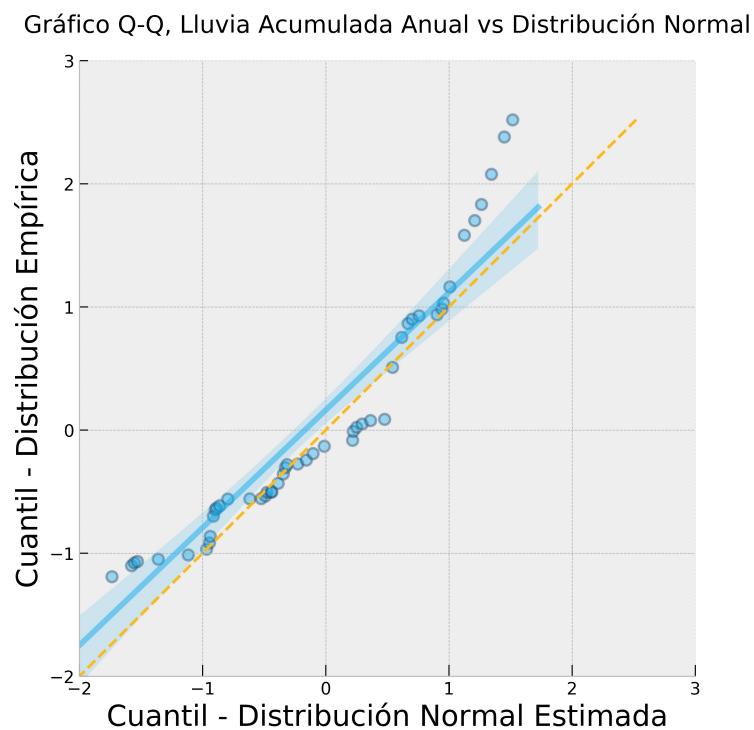


Figura 9: Gráfico QQ utilizado para comparar la distribución acumulada anual de lluvias con una distribución normal teórica.

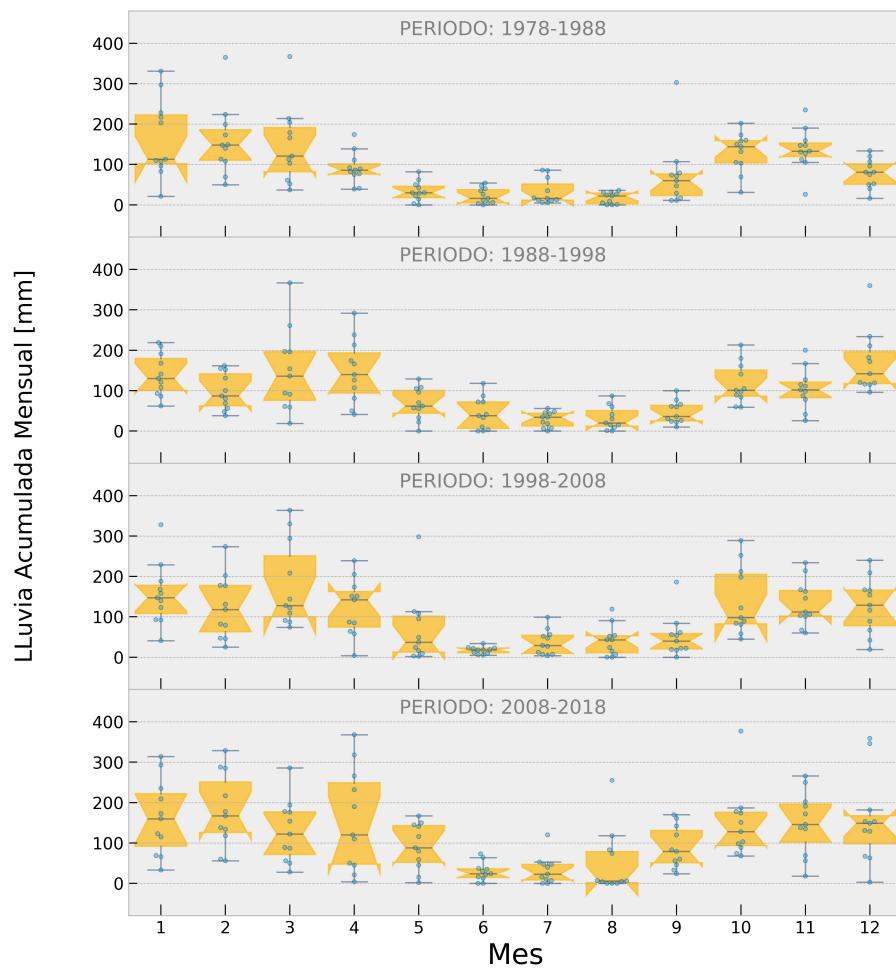


Figura 10: Cada boxplot representa un mes entre Enero-Diciembre,

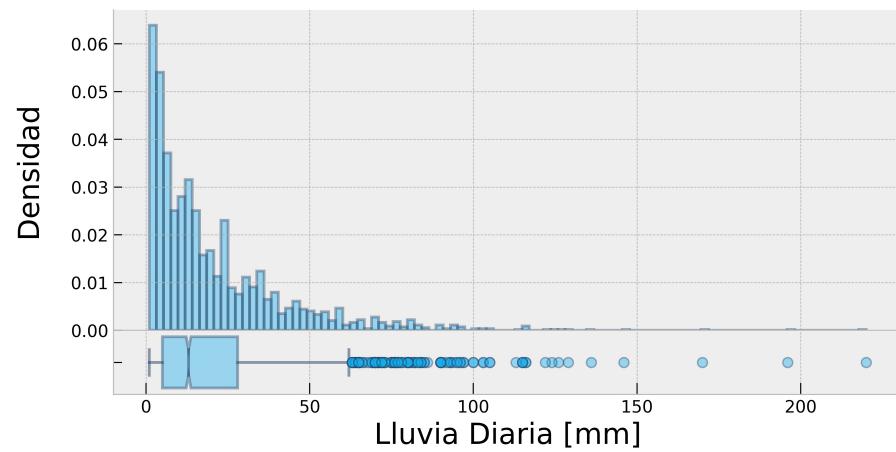


Figura 11: Arriba: distribución histórica de registros de lluvia. Abajo: diagrama de caja con intervalos de 95 % de confidencia

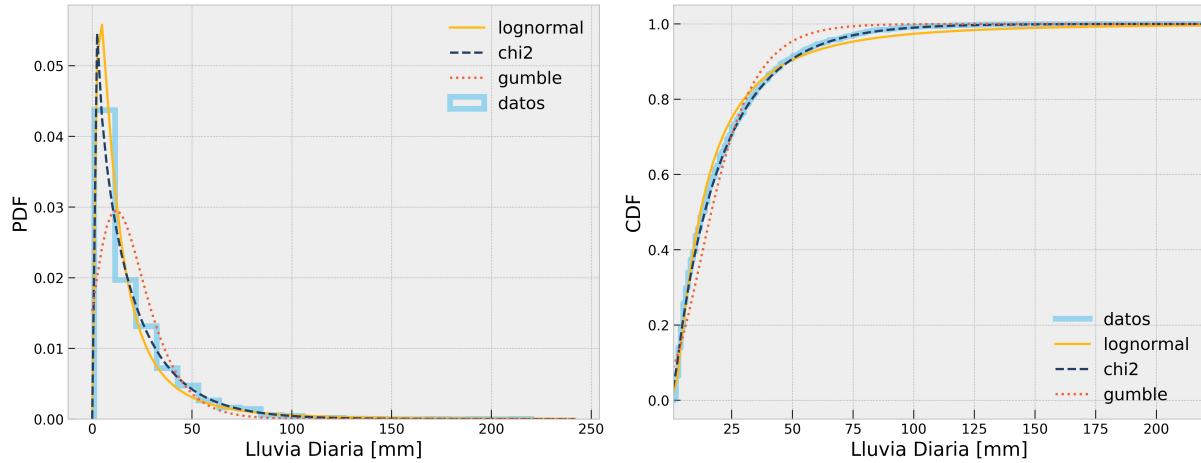


Figura 12: Izq.: distribución de densidad de probabilidad (línea continua celeste) junto a tres distribuciones propuestas. Der.: distribución de Probabilidad Acumulada observada, junto con las de las distribuciones propuestas.

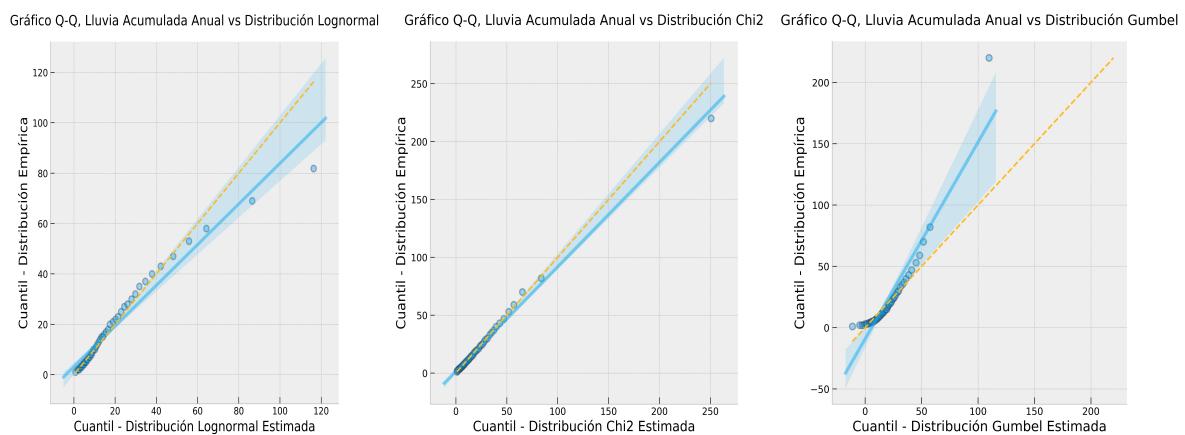


Figura 13: Los qqplots verifican que la distribución de lluvias observadas se apartan de una distribución lognormal (izquierda), de una distribución chi-cuadrado (medio) y de una distribución gumbel teórica (derecha).