Trabajo Fin de Máster

Estudio de la precipitación a partir de la regresión por cuantiles

Gleisis Alvarez Socorro

Máster en Técnicas Estadísticas

Curso 2023-2024

Propuesta de Trabajo Fin de Máster

Título en galego: Estudio da precipitación a partir da regresión por cantos.

Título en español: Estudio de la precipitación a partir de la regresión por cuantiles.

English title: Precipitation study using the quantile regression.

Modalidad: Modalidad A

Autor/a: Gleisis Alvarez Socorro, Universidad de Vigo

Director/a: Javier Roca Pardiñas, Universidad de Vigo

Tutor/a: Luis Gimeno Presa, Universidad de Vigo

Breve resumen del trabajo:
El mecanismo conocido como chorro de los bajos niveles de las Grandes Llanuras Americanas consiste en vientos muy fuertes en la tropósfera inferior que transportan una gran cantidad de humedad desde el Golfo de México y está activo principalmente durante el verano. El estudio de la precipitación asociada a este mecanismo es de gran importancia para dicha región. Variables meteorológicas como la humedad transportada, la inestabilidad atmosférica y el agua total en la columna, juegan un importante papel en la ocurrencia de la precipitación. Por tanto, el objetivo de este trabajo consiste en estudiar la relación entre estas variables y la precipitación en la zona de interés a partir de modelos de regresión cuantil. Para ello se cuenta con una serie de valores diarios de estas variables, de los meses junio-julio-agosto, entre 1980 y 2017. Se empleó el criterio BIC para la selección de las variables predictoras más significativas. Se utilizaron diferentes modelos de regresión, desde los más simples a más complejos, siendo el modelo de localización y escala el que presentó los mejores resultados. Este modelo representó correctamente la distribución de la precipitación para los diferentes cuantiles y permitió identificar las variables predictoras que más influyeron en su variabilidad.

Recomendaciones: Se recomienda utilizar datos meteorológicos de mayor resolución espacial para comprobar los resultados y en otras regiones de interés. Además, aplicar otras técnicas estadísticas para modelar la relación entre estas variables.

Otras observaciones:

Don Javier Roca Pardiñas, Profesor y estadístico de la Universidad de Vigo, y don Luis Gimeno Presa, Profesor catedrático, de Universidad de Vigo, informan que el Trabajo Fin de Máster titulado

Estudio de la precipitación a partir de la regresión por cuantiles

fue realizado bajo su dirección por don/doña Gleisis Alvarez Socorro para el Máster en Técnicas Estadísticas. Estimando que el trabajo está terminado, dan su conformidad para su presentación y defensa ante un tribunal.

En Universidad de Vigo, a 3 de junio de 2024.

ROCA PARDIÑAS JAVIER - 77592181W
Firmado digitalmente por ROCA PARDIÑAS JAVIER - 77592181W
Fecha: 2024.06.03 16:29:13 +02'00'

El director/a:
Don Javier Roca Pardiñas

GIMENO PRESA LUIS - 03838319X
Firmado digitalmente por GIMENO PRESA LUIS - 03838319X
Fecha: 2024.06.03 11:28:45 +02'00'

El tutor/a:
Don Luis Gimeno Presa

Firmado por ALVAREZ SOCORRO GLEISIS - ****9471* el día 03/06/2024 con un certificado emitido por AC FNMT Usuarios

La autora:
Doña Gleisis Alvarez Socorro

Declaración responsable. Para dar cumplimiento a la Ley 3/2022, de 24 de febrero, de convivencia universitaria, referente al plagio en el Trabajo Fin de Máster (Artículo 11, Disposición 2978 del BOE núm. 48 de 2022), **el/la autor/a declara** que el Trabajo Fin de Máster presentado es un documento original en el que se han tenido en cuenta las siguientes consideraciones relativas al uso de material de apoyo desarrollado por otros/as autores/as:

- Todas las fuentes usadas para la elaboración de este trabajo han sido citadas convenientemente (libros, artículos, apuntes de profesorado, páginas web, programas,...)
- Cualquier contenido copiado o traducido textualmente se ha puesto entre comillas, citando su procedencia.
- Se ha hecho constar explícitamente cuando un capítulo, sección, demostración,...sea una adaptación casi literal de alguna fuente existente.

Y, acepta que, si se demostrara lo contrario, se le apliquen las medidas disciplinarias que correspondan.

Agradecimientos

Durante estos dos años de etapa de formación profesional han sido muchas las personas que de una forma u otra me han ayudado a culminar mis estudios de Máster. En especial quiero agradecer a:

- Toda mi familia, que desde la distancia han estado siempre al pendiente de mí, de cada uno de mis exámenes y trabajos, apoyándome en todo momento para cumplir este importante objetivo. En especial a mis padres, mi mima y mis hermanos.
- Mi esposo José Carlos, sin él esto no hubiera sido posible. Gracias por todo el tiempo que has dedicado en ayudarme y por siempre estar a mi lado en todo momento.
- Los profesores Raquel y Luis por su guía y apoyo cada vez que lo he necesitado durante estos años.
- El profesor Javier Roca por sus acertadas orientaciones y correcciones durante el desarrollo de este trabajo.
- Olamar, porque ha sido un gran descubrimiento en estos años, por todas las horas de estudio juntas que fueron mucho más fáciles al poder compartirlas con ella.
- Mis amistades, tanto los que están cerca como los que están lejos, por siempre darme ánimos para seguir y compartir conmigo lindos momentos.
- Los profesores del Máster por su tiempo y preparación en cada clase o tutoría, gracias por transmitirme sus amplios conocimientos estadísticos.
- Grupo Ephyslab por permitirme formar parte de este grupo como una nueva integrante y crecer como profesional. Por facilitarme los datos para realizar este trabajo.

A todos los que de una forma me han apoyado durante estos dos importantes años de mi vida, MUCHAS GRACIAS.

Índice general

Resumen

Resumen en español

El mecanismo conocido como chorro de los bajos niveles de las Grandes llanuras americanas consiste en vientos muy fuertes en la tropósfera inferior que transportan una gran cantidad de humedad desde el Golfo de México y está activo principalmente durante el verano. El estudio de la precipitación asociada a este mecanismo es de gran importancia para dicha región. Variables meteorológicas como la humedad transportada, la inestabilidad atmosférica y el agua total en la columna, juegan un importante papel en la ocurrencia de la precipitación. Por tanto, el objetivo de este trabajo consiste en estudiar la relación entre estas variables y la precipitación en la zona de interés a partir de modelos de regresión cuantil. Para ello se cuenta con una serie de valores diarios de humedad transportada, inestabilidad atmosférica, agua total en la columna y precipitación, de los meses junio-julio-agosto, entre 1980 y 2017. Se empleó el criterio BIC para la selección de las variables predictoras más significativas. Los cuantiles utilizados en el análisis fueron 0.25, 0.50, 0.75, 0.95, 0.99. Se utilizaron diferentes modelos de regresión, desde los más simples a los más complejos, siendo el modelo de localización y escala el que presentó los mejores resultados. Este modelo representó correctamente la distribución de la precipitación para los diferentes cuantiles y permitió identificar las variables predictoras que más influyeron en la variabilidad de dicha precipitación.

English abstract

The Great Plains low-level jet mechanism consists of very strong winds in the lower troposphere that transport a large amount of moisture from the Gulf of Mexico and is mainly active during the summer. Studying the precipitation associated with this mechanism is of great importance for this US region. Meteorological variables such as transported moisture, atmospheric instability, and total column water play an important role in the precipitation occurrence. Therefore, the research objective is to study the relationship between these variables and precipitation in the interest area using quantile regression models. For this purpose, daily value series of transported moisture, atmospheric instability, total column water, and precipitation from June-July-August between 1980 and 2017 were used. The BIC criterion was employed to choose the most significant predictor variables. The quantiles used in the analysis were 0.25, 0.50, 0.75, 0.95, and 0.99. Different regression models were used, from the simplest to the most complex, with the location and scale model showing the best results. This model correctly represented the precipitation distribution for the different quantiles and allowed the identification of the predictor variables that most influenced the precipitation variability.

Acrónimos

- ACF: función de autocorrelación.
- AIC: criterio de información Akaike.
- BIC: criterio de información bayesiano.
- CDF: función de distribución acumulativa condicional.
- ELF: pérdida extendida de log-f.
- FLEXPART: FLEXible PARTicle dispersion model.
- GAM: modelos aditivos generalizados.
- GCV: validación cruzada generalizada.
- GLM: modelos lineales generalizados.
- GPLLJ: chorro de bajos niveles de las Grandes Llanuras Americanas.
- Inest: inestabilidad atmosférica.
- Inest1: inestabilidad atmosférica con retardo igual a 1.
- LE: modelo de localización y escala.
- LLJ: chorro de bajos niveles.
- LM: modelos lineales.
- MSE: error cuadrático medio.
- Prec1: precipitación con retardo igual a 1.
- QGAM: regresión cuantil de modelos aditivos generalizados.
- RMSE: raíz del error cuadrático medio.
- TCW: agua total en la columna.
- TCW1: agua total en la columna con retardo igual a 1.
- Trans.Humed: humedad transportada.
- Trans.Humed1: humedad transportada con retardo igual a 1.

Introducción

Las técnicas estadísticas a día de hoy tienen una gran aplicación tanto para el trabajo científico como operativo. Un ejemplo de ello es la aplicación que tienen en las Ciencias Meteorológicas, permitiendo a los investigadores decodificar patrones, establecer relaciones y predecir futuros escenarios basados en datos observados. Dentro de las técnicas estadísticas, en particular los modelos de regresión, se han convertido en herramientas indispensables dentro de esta disciplina. Estos modelos ofrecen la capacidad de describir y cuantificar las relaciones entre variables meteorológicas, ya sea para identificar factores que contribuyen a un fenómeno específico, o para anticipar cambios en un sistema bajo diferentes condiciones.

Los modelos de regresión en las ciencias ambientales no son solo herramientas analíticas: son puentes que conectan la observación con la acción. Las ciencias meteorológicas y ambientales, por su naturaleza, a menudo se enfrentan a sistemas complejos y multifacéticos, por lo que es esencial comprender cómo las distintas variables interactúan y se influencian mutuamente. Los modelos de regresión proporcionan una estructura para desglosar esta complejidad, permitiendo a los científicos identificar causas, evaluar impactos y, crucialmente, proponer soluciones basadas en evidencia.

Uno de los aspectos que se pueden utilizar en un modelo de regresión para realizar un análisis de la relación entre las variables es a partir de los cuantiles de las mismas. De forma general, lo habitual es utilizar la media μ (o la mediana) como resumen de la posición, mientras que la desviación típica σ se utiliza como medida de dispersión. Sin embargo, una información más completa y detallada de las variables viene dada por sus cuantiles. La regresión por cuantiles es muy útil cuando la distribución de los datos es asimétrica o cuando hay presencia de valores atípicos. Además cada cuantil representa un porcentaje específico de la distribución y proporciona información detallada, así como una descripción completa sobre la variabilidad en diferentes partes de la distribución. Su aplicación se extiende al análisis de datos climáticos y meteorológicos, los cuales en su mayoría presentan distribución asimétrica y pueden contener valores extremos y atípicos.

Entre las variables meteorológicas más estudiadas, la precipitación es una de las más importantes. Su análisis es fundamental por diversas razones debido a que tienen implicaciones tanto en la escala local como global. Además, es crucial para una variedad de disciplinas y aplicaciones, desde la gestión de recursos y la conservación ecológica, hasta la planificación urbana y la predicción del tiempo. La comprensión de la precipitación es esencial para abordar muchos de los desafíos globales actuales y futuros relacionados con el agua, el clima, las sequías y el medio ambiente.

En este trabajo se pretende realizar un estudio de la precipitación asociada al chorro de bajos niveles en las Grandes LLanuras Americanas (GPLLJ, por sus siglas en inglés). Un chorro de bajos niveles (LLJ, por sus siglas en inglés) es un mecanismo de vientos muy fuertes en la tropósfera inferior, típicamente en los primeros 1000 metros de altura (Stensrud, 1996). Este GPLLJ, es uno de los más estudiados debido a sus efectos, tanto para la economía como para la sociedad en general, ya que transporta una enorme cantidad de humedad desde el Golfo de México hacia las Grandes Llanuras Americanas. Son considerados como uno de los principales mecanismos del transporte de humedad a escala planetaria, lo que favorece a la precipitación en determinadas regiones. La estrecha relación entre el transporte de humedad y los eventos extremos de precipitación se maximiza cuando se estudia en las áreas de influencia de los principales mecanismos globales de transporte de humedad atmosférica, entre ellos los LLJs (Gimeno et al., 2016). El GPLLJ es un mecanismo extremadamente localizado en

tiempo y espacio, y su papel en el equilibrio de humedad continental es difícil de estudiar únicamente a partir de observaciones (Algarra et al., 2019).

La presencia de humedad es una condición necesaria para la ocurrencia de precipitación, pero no es suficiente por sí sola. También se requieren mecanismos de levantamiento, como por ejemplo corrientes ascendentes, así como núcleos de condensación, como partículas de polvo o sal, para que el vapor de agua se condense y eventualmente caiga como precipitación. La inestabilidad atmosférica se relaciona con dichos mecanismos de levantamiento en la atmósfera. Esta se produce cuando una masa de aire caliente y húmedo se eleva y se enfría, lo que puede provocar la formación de nubes y precipitaciones. De ahí que otro factor importante, que juega un papel crucial en la formación y distribución de la precipitación que pueda caer en una región, es la inestabilidad atmosférica que esté presente en dicha región. Una atmósfera más inestable generalmente favorece la formación de precipitaciones más intensas y, a menudo, de corta duración, mientras que una atmósfera estable tiende a inhibir la formación de precipitación o a favorecer precipitaciones más ligeras y constantes.

El mecanismo subyacente en la relación entre el GPLLJ y la precipitación es un fuerte transporte de humedad y calor en niveles bajos desde el Golfo de México. Además, la convergencia del viento en niveles bajos implica inestabilidad atmosférica en el área de salida del GPLLJ, favoreciendo el movimiento ascendente. Por lo tanto, es evidente que la humedad transportada y la inestabilidad atmosférica son dos factores que juegan un papel importante en la precipitación. De ahí la importancia de estudiar la precipitación asociada al GPLLJ, en función de la humedad transportada y la inestabilidad atmosférica presente en dicha región.

Otra variable que se puede considerar relevante para estudiar la precipitación asociada a este fenómeno es el agua total en la columna (TCW, por sus siglas en inglés). Esta variable es una medida de la cantidad total de agua en una columna de aire, la cual en este caso se corresponde con el sumidero de humedad asociado a la región de ocurrencia del GPLLJ. Un valor alto de TCW indica una mayor cantidad de vapor de agua disponible, que potencialmente puede condensarse dando lugar a la precipitación. Las regiones con valores altos de TCW son más propensas a experimentar precipitaciones significativas, especialmente si se combinan con condiciones de inestabilidad atmosférica.

En resumen, para que ocurra la precipitación, entre otras condiciones, se requiere una combinación de inestabilidad atmosférica (que promueve el ascenso del aire húmedo), suficiente humedad transportada desde la región fuente a la región sumidero (que provee el vapor de agua necesario), y un TCW elevado (que indica la disponibilidad de vapor de agua en la atmósfera). La interacción entre estos elementos determinará la ocurrencia, el tipo y la intensidad de la precipitación en una región determinada.

Para analizar la relación existente entre estas variables meteorológicas y cómo influyen en la precipitación asociada al GPLLJ se aplicarán diferentes modelos de regresión, comenzando desde uno más sencillo, a más complejo. De esta manera se determinará cuál de ellos representa mejor la relación entre dichas variables. Específicamente, se utilizarán modelos de regresión cuantil, a partir de los cuales se podrá identificar cómo estas variables explicativas afectan, no solo la precipitación media, sino también las precipitaciones en el extremo inferior y superior de la distribución. Al centrarse la regresión cuantil en la estimación de varios cuantiles, como por ejemplo, el 25 %, el 50 % o el 95 %, permite una compresión más detallada de la influencia de las variables independientes sobre la variable respuesta, en lugar de solo su media.

Actualmente esta investigación forma parte de los trabajos realizados por el grupo Ephyslab, en el departamento de Física Aplicada, perteneciente al Campus de Ourense de la Universidad de Vigo.

Este trabajo estará dividido en 4 Capítulos, los cuales se resumen a continuación:

- Capítulo 1: Realiza una breve descripción de la región de estudio seleccionada, así como del mecanismo de humedad en el que se enfoca la investigación. Además, se explican las variables meteorológicas utilizadas en el análisis, presentado un resumen de la descripción de las mismas.
- Capítulo 2: Se realiza una presentación de los modelos de regresión que se aplicarán a dichos datos, comenzando desde los más simples a los más complejos. También, cuenta con un resumen del criterio aplicado para la selección de variables que se incorporarán en dicho modelos.

- Capítulo 3: Presenta una introducción a la regresión cuantil, la cual se utilizará para analizar la relación que existe entre las variables de estudio seleccionadas para este trabajo y su aplicación a los diferentes tipos de modelos empleados.
- Capítulo 4: Expone los resultados obtenidos aplicando las diferentes técnicas de regresión por cuantiles, empleando distintos modelos, así como un análisis detallado de los mismos.
- Conclusiones: Se presentan las conclusiones resumiendo los resultados más significativos obtenidos durante el trabajo realizado. Además, se mencionan los trabajos futuros que se pretenden realizar para ampliar y mejorar el estudio.

Capítulo 1

Región de estudio y datos utilizados

En este capítulo 1 se realizará primeramente un resumen de la región a la que pertenecen los datos analizados en este trabajo, así como una breve descripción de los mismos y de sus propiedades.

1.1. Chorro de bajos niveles de las Grandes Llanuras Americanas

El GPLLJ es un mecanismo confinado dentro de los primeros kilómetros de la tropósfera y está estrechamente relacionado con la estación cálida (Bonner, 1968). En términos generales, transporta un tercio de todo el vapor de agua que entra a los Estados Unidos continentales (Helfand y Schubert, 1995), y está asociado con el 10 - 45 % de la precipitación de verano de la región de las Grandes Llanuras Americanas (Hodges y Pu, 2019). El GPLLJ afecta la precipitación al aumentar su frecuencia, modificar su distribución espacial y aumentar su intensidad (Pitchford y London, 1962; Mo et al., 1995; Walters y Winkler, 2001; Schumacher y Johnson, 2009; Squitieri y Gallus, 2016; Squitieri y Gallus Jr, 2016).

La importancia económica del GPLLJ es enorme en el sentido que determina la precipitación promedio y extrema de una vasta región agrícola, cuya producción depende de la precipitación, ocasionando grandes pérdidas debido a inundaciones y sequías (Basara et al., 2013). También es importante en la determinación del recurso eólico y especialmente, en los daños generados por las condiciones meteorológicas severas, ya que el GPLLJ está estrechamente relacionado con el desarrollo de sistemas convectivos de mesoescala (Chen et al., 1998) y estos están asociados con fuertes precipitaciones, tormentas supercelulares y desarrollo de tornados. Además, Higgins et al. (1997), informaron diferencias significativas en el patrón de precipitación en coincidencia (o no) con eventos LLJ. Cuando ocurre un evento LLJ, las observaciones muestran una precipitación sobre el norte-central de Estados Unidos y la región de las Grandes Llanuras, junto con una disminución a lo largo del Golfo de México y el Atlántico occidental.

En la Figura 1.1 se muestra la climatología del GPLLJ para los meses de junio, julio y agosto, en los que este mecanismo presenta mayor intensidad, influyendo así en la precipitación en la región analizada.

1.2. Datos utilizados

Los datos utilizados corresponden con una serie temporal de los meses de verano (junio, julio y agosto) en el período comprendido entre 1980 y 2017. Se escogieron estos meses porque como se había mencionado anteriormente el GPLLJ es más activo en esta época del año, con una frecuencia de ocurrencia cercana al 70 % de los días (Gimeno-Sotelo, 2021). Se cuenta con un total de 3496 valores diarios de cada una de las variables analizadas (precipitación, humedad transportada, inestabilidad

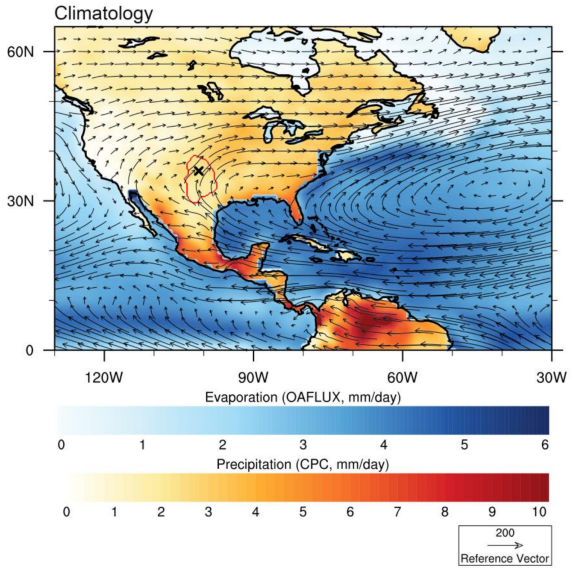

Figura 1.1: Climatología del sistema de chorros de bajos niveles de las Grandes LLanuras Americanas para los meses de junio, julio y agosto. La región con mayor ocurrencia del LLJ está dentro del contorno rojo, con la cruz indicando el punto en el cual la proporción de días en los que ocurre LLJ es la más alta. Los colores azulados representan la evaporación (mm/día), los colores rojizos indican la precipitación (mm/día) y las flechas simbolizan el flujo de humedad en cada punto de la cuadrícula considerada ($\mathrm{kgm^{-1}s^{-1}}$). Figura cortesía del Dr. Iago Algarra (Universidad de Vigo, España).

atmosférica, agua total en la columna). Destacar que en el caso de la variable humedad esta fue simulada utilizando el modelo lagrangiano de dispersión de partículas FLEXPART, forzado con los datos de reanálisis ERA-Interim. A continuación se detallan las variables mencionadas anteriormente, basadas en las áreas de fuentes y sumideros de humedad vinculadas al GPLLJ:

- **Precipitación** en la región del sumidero GPLLJ (mm/día): serie diaria de precipitación integrada en la toda la región del sumidero de humedad del GPLLJ tomada del conjunto de datos del Centro de Predicción del Clima (CPC) (Xie et al., 2010).

- **Humedad Transportada** desde la región fuente del GPLLJ al dominio del jet (mm/día) (teniendo en cuenta como se calculó en Algarra et al. (2019)). En este estudio, se utilizó un enfoque lagrangiano con el fin de identificar las principales fuentes y sumideros de humedad asociados al GPLLJ. La humedad transportada se calculó entonces sumando las ganancias de humedad de las partículas en la región fuente antes de llegar al dominio del jet.

- **Inestabilidad atmosférica** en la región sumidero del GPLLJ (omega, Pa/s): series diarias de velocidad vertical calculada como la media de omega a 850 hPa en la región sumidero, tomada del reanálisis ERA-5 (Hersbach et al., 2020). Omega se define como la componente vertical de la velocidad en coordenadas de presión (estas coordenadas tridimensionales se definen reemplazando la usual coordenada z por la presión atmosférica (p)). Es decir, $w = dp/dt$, por lo que los valores negativos de w representan movimientos ascendentes y los valores positivos corresponden a movimientos descendentes. El nivel de 850 hPa (alrededor de 1500 m de altura) se considera para w porque representa el movimiento vertical en la tropósfera inferior, donde ocurre el GPLLJ y donde la mayor parte de la humedad está confinada.

  Para una mejor interpretación de los resultados, el nombre de la variable que se tendrá en cuenta es Inestabilidad atmosférica, la cual será igual a $-Omega$. Destacar que Omega era negativo

cuando había inestabilidad y movimientos ascendentes. De esta manera, se hablará entonces de inestabilidad positiva cuando hay movimientos ascendentes y de inestibilidad negativa cuando los movimientos sean descendentes.

- **Agua total en la columna** en la región sumidero (kg/m^2): es la integral en la vertical desde la superficie hasta el tope de la atmósfera que expresa la cantidad total de agua en esa columna (vapor + agua de nube + hielo de nube), pero no incluye precipitación (Urban, 2006).

A continuación se muestra la Figura 1.2 con la región de mayor ocurrencia de LLJs. El área dentro de la curva roja en esta figura, corresponde con el dominio del LLJ, es decir, es la región con la mayor ocurrencia de este durante el período de mayo a octubre, siendo la cruz el punto geográfico en el que ocurren con mayor frecuencia (36°N, 101°W, 500m de altura); el área en azul identifica la principal región oceánica fuente de humedad que llega al dominio del LLJ. El área en verde corresponde al principal sumidero de esa humedad, una vez que ha sido transportada por el mismo. Por lo tanto, hay dos regiones de interés en nuestro análisis: las regiones fuente y sumidero de humedad, conectadas por la estructura del GPLLJ en un dominio temporal de varios días, desde la evaporación en la fuente hasta la precipitación en el sumidero.

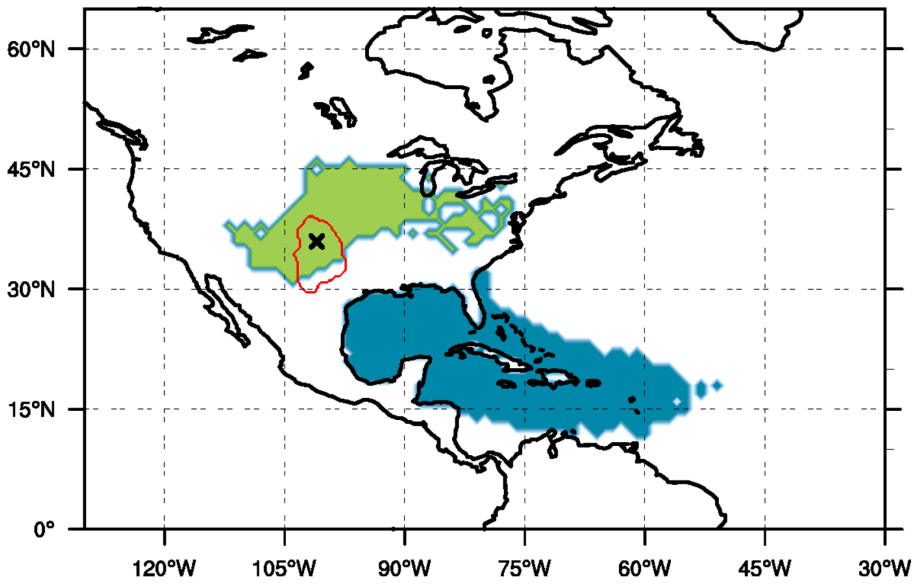

Figura 1.2: Región con la mayor ocurrencia de LLJs (dentro del contorno rojo, con la cruz indicando el punto en el que la proporción de días en que ocurre el LLJ es la más alta); la principal región de origen de humedad oceánica del GPLLJ (en azul) y su principal región sumidero de humedad (en verde). Figura cortesía del Dr. Iago Algarra (Universidad de Vigo, España).

Destacar que la base de datos empleada para la realización de este trabajo se utilizó anteriormente en un Trabajo de Máster en Estadística e Investigación Operacional en la Universidad de Lisboa (Gimeno-Sotelo, 2021) y en el artículo científico (Algarra et al., 2019).

1.2.1. Propiedades de las variables

Resumen numérico

A continuación se realiza un análisis, tanto desde el punto de vista numérico como gráfico, de las propiedades de las variables que se van a utilizar. Primeramente se muestra en la Tabla 1.1 un resumen

numérico de dichas variables, presentando sus valores mínimos, medios y máximos, así como la mediana y los cuantiles primero y tercero.

Tabla 1.1: Resumen estadístico de las variables utilizadas.

Variables	Mínimo	1er cuantil	Mediana	Media	3er cuantil	Máximo
Precipitación (mm/día)	0.026	1.473	2.556	2.814	3.785	11.609
Inestabilidad (Pa/s)	- 0.058	- 0.002	0.013	0.014	0.029	0.100
Humedad (mm/día)	0	0.411	0.836	0.917	1.321	4.086
Agua columna (kg/m^2)	14.410	23.780	27.030	26.920	29.850	39.920

Para el caso de la **precipitación**, sus valores mínimos y máximos son, 0.026 mm/día y 11.609 mm/día, respectivamente. Esto significa que hay días donde prácticamente no se reportó precipitación, a diferencia de otros donde la precipitación fue superior. Teniendo en cuenta, por ejemplo el valor máximo, este significa que en un período de 24 horas se registraron 11 l/m^2, es decir, 1 mm de lluvia equivale a 1 litro de agua por metro cuadrado (https://todosloshechos.es/cuantos-litros-por-metro-cuadrado-es-1-mm). Este valor se puede considerar una cantidad moderada de lluvia, la cual, teniendo en cuenta las condiciones locales como por ejemplo la capacidad de absorción del suelo, la urbanización y la preparación de drenaje en el área, podría provocar efectos y daños significativos en la región. Por otro lado, la mediana de estos datos es igual a 2.556 mm/día, mientras que la media fue ligeramente superior, 2.841 mm/día. Este último valor, comparado con el valor máximo, es relativamente pequeño, por lo que, a pesar de analizar el período estacional (de junio a agosto) donde el GPLLJ alcanzaba su máximo de intensificación, la precipitación en término medio no fue tan elevada. En cuanto a los cuantiles, su primer cuantil es igual a 1.473 mm/día, mientras que el tercero fue de 3.785 mm/día. Además, hay que destacar, que esta cantidad fue registrada en un día, pero pudo haber caído en un corto período tiempo, lo que puede conllevar a problemas mayores.

En cuanto a la variable **inestabilidad atmosférica**, esta puede tomar valores tanto negativos como positivos, ya que como se había mencionado anteriormente, indica la presencia de movimientos ascendentes y descendentes, los cuales en este caso serán considerados con valores positivos y negativos, respectivamente. Cabe destacar además, que con valores positivos existirán condiciones de inestabilidad, pero si son negativos, se estará en presencia de una atmósfera estable. Con respecto a los resultados obtenidos en el resumen numérico, su valor mínimo fue -0.058 Pa/s y su valor máximo 0.10 Pa/s. La mediana para la inestabilidad fue de 0.013 Pa/s, mientras que el valor medio fue de 0.014 Pa/s, lo que pudiera indicar un estado neutro, o sea ni muy estable ni muy inestable, por estar próximo a 0. Por último, en cuanto a los cuantiles primero y tercero, los valores son, -0.002 Pa/s y 0.029 Pa/s, respectivamente. Con todos estos resultados, se pudiera decir que de manera general, se contaba con una atmósfera relativamente inestable, la cual podría provocar en determinados momentos condiciones para la formación de tormentas severas o precipitaciones significativas.

Para la variable **humedad transportada**, el valor mínimo es 0, lo que significa ausencia de humedad, es decir en algunos de los días del período de estudio no hubo transporte de humedad desde la región fuente a la región sumidero, lo que podría estar asociado con condiciones secas, cuando no hay suficiente contenido de humedad en la atmósfera o cuando los patrones de circulación atmosférica no favorecen el transporte de humedad. Mientras que por otro lado, el valor máximo es de 4.086 mm/día, lo que significa un transporte considerable de humedad. La mediana de estos datos es 0.836 mm/día y la media, ligeramente superior, 0.917 mm/día. En relación con los cuantiles, el primero es igual a 0.411 mm/día, mientras que el tercero es igual a 1.321 mm/día.

Por último, la variable **agua total en la columna**, presentó unos valores mínimos y máximos iguales a 14.41 kg/m^2 y 39.92 kg/m^2, los cuales están relacionados con condiciones más secas y más húmedas, respectivamente. Destacar que este valor máximo de TCW, puede ser indicativo de condiciones relativamente potenciales para la formación de nubes densas y de precipitación significativa. Por otro lado, la mediana de esta variable es 27.03 kg/m^2, mientras que el valor medio fue ligeramente inferior, igual a 26.92 kg/m^2. Ambos valores relativamente altos, pudieran indicar que por lo general los valores de TCW para esta región reflejan contenido alto de agua total en la columna. Los cuantiles primero y tercero fueron iguales a 23.78 kg/m^2 y 29.85 kg/m^2, respectivamente, los cuales indican la dispersión y la variabilidad de los datos.

Se analizó también la correlación lineal que presentaban cada una de las variables, arrojando los siguientes resultados (Figura 1.3):

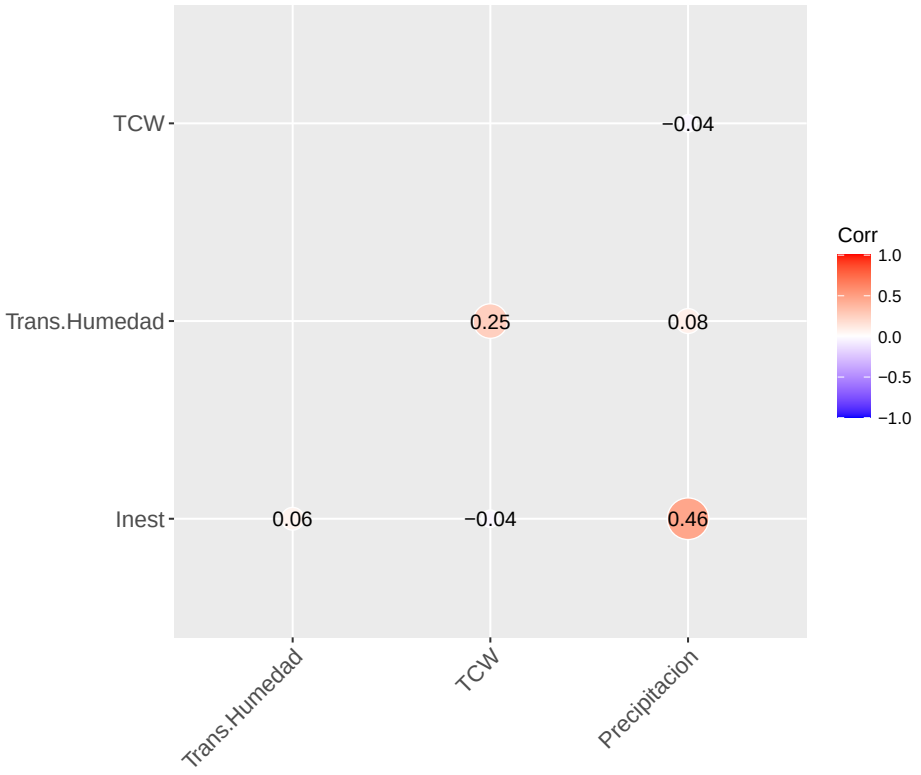

Figura 1.3: Correlación entre las variables.

Si se analizan cada uno de los valores de correlación lineal obtenidos, se pueden extraer las siguientes conclusiones, aunque de forma general no se observa alta correlación entre las variables. En primer lugar la correlación entre la inestabilidad atmosférica y la precipitación es moderada y positiva. Esto significa que si la inestabilidad aumenta la precipitación también tiende a aumentar, teniendo en cuenta que la correlación lineal entre ellas no es del todo fuerte. A diferencia de este caso, para las otras dos variables predictoras los valores de correlación lineal son muy bajos, prácticamente 0, lo que indica una relación muy débil, o prácticamente ninguna relación lineal entre la cantidad de humedad transportada y el agua total en la columna con respecto a la precipitación. Si se tiene en cuenta el signo para cada correlación, para el transporte de humedad es positiva y para el agua total en la columna es negativa, lo que sinifica que si ambas variables aumentan, la precipitación en un caso aumentará y en el otro disminuirá. Por otra parte, la correlación entre la inestabilidad con el agua total en la columna y la humedad transportada es muy baja, prácticamente 0. Mientras que, entre esta última y el agua total en

la columna, es un poco mayor, llegando hasta 0.25, aunque se puede seguir considerando baja. Desde el punto de vista meteorológico esta relación, aunque sea pequeña, tiene sentido debido que cuando hay un fuerte transporte de humedad hacia una región (sumidero), típicamente hay un incremento en el agua total en la columna en esa región. Esto es porque el vapor de agua transportado por los vientos se acumula en la atmósfera de la región sumidero.

En resumen estos resultados demuestran la complejidad de las relaciones entre dichas variables meteorológicas y que pudieran influir otros factores para la ocurrencia de precipitación, como factores locales o temporales. De ahí la necesidad de analizar como influyen estas variables en la precipitación, no solo de forma individual, sino también de manera conjunta y aplicando no solo modelos de regresión sencillos, sino también complejos.

Resumen gráfico

Desde el punto de vista gráfico, se representaron las variables en un histograma para analizar su distribución (Figura 1.4). Cada histograma muestra la frecuencia de los valores en el eje vertical y el rango de valores de cada una de las variables analizadas en el eje horizontal, así como su variabilidad. Con respecto a la precipitación, la mayor cantidad de datos se concentran entre 1 y 4 mm/día. Para valores mayores que 8 mm/día se registraron muy pocos días de precipitación, los cuales se podrían considerar como extremos alejándose ligeramente del resto de los datos.

Por otra parte, los valores de inestabilidad atmosférica se centran entre -0.01 y 0.03 Pa/s, fundamentalmente. Para aquellos que resultaron ser menores que -0.04 Pa/s y mayores que 0.07 Pa/s, la frecuencia de días es muy baja. En cuanto a la variable humedad transportada, a medida que aumentaban los valores de humedad registrados la frecuencia de días disminuía. Es decir, para valores de humedad mayores que 2.5 mm/día, la cantidad de días es menor, mientras que, la mayor frecuencia se observa entre los 0 y 1.5 mm/día. Por último, con respecto al agua total en la columna, hasta aproximadamente los 30 kg/m^2, la cantidad de días va aumentando a medida que aumentan los valores de la variable. Mientras que para valores mayores que 30 kg/m^2, la frecuencia de días tiende a la disminución.

En ninguno de los casos se puede decir que estamos en presencia de una distribución normal. Para comprobarlo se aplicó el test de Shapiro-Wilk para analizar la normalidad de los datos utilizando la función **shapiro.test** del paquete **stats**. Los resultados arrojaron, en los 4 casos, p-valores menores que cualquier nivel de significación, por lo tanto se concluye que existen evidencias significativas para rechazar la hipótesis nula de normalidad a favor de la hipótesis alternativa.

Estos resultados también se pueden observar en los siguientes gráficos secuenciales que muestran la variabilidad temporal para cada una de las variables analizadas (Figura 1.5). Las series muestran una componente estacional ya que presentan un patrón repetitivo. Para cada una de ellas se observa un rango de valores relativamente bien definido, destacándose muchos valores que sobresalen de este rango.

Además, en la Figura 1.6, se muestran los gráficos de dispersión, para analizar visualmente, como influyen cada una de las variables predictoras en la precipitación.

Estos gráficos reafirman los valores obtenidos al analizar la correlación lineal entre las variables. Se observa, que en cierta medida, los valores de precipitación tienden a aumentar a medida que aumenta la inestabilidad. Sin embargo, en los otros dos gráficos no se observa el mismo comportamiento, existiendo una mayor dispersión de los datos lo que ratifica los valores tan bajos de correlación lineal entre estas variables predictoras y la precipitación.

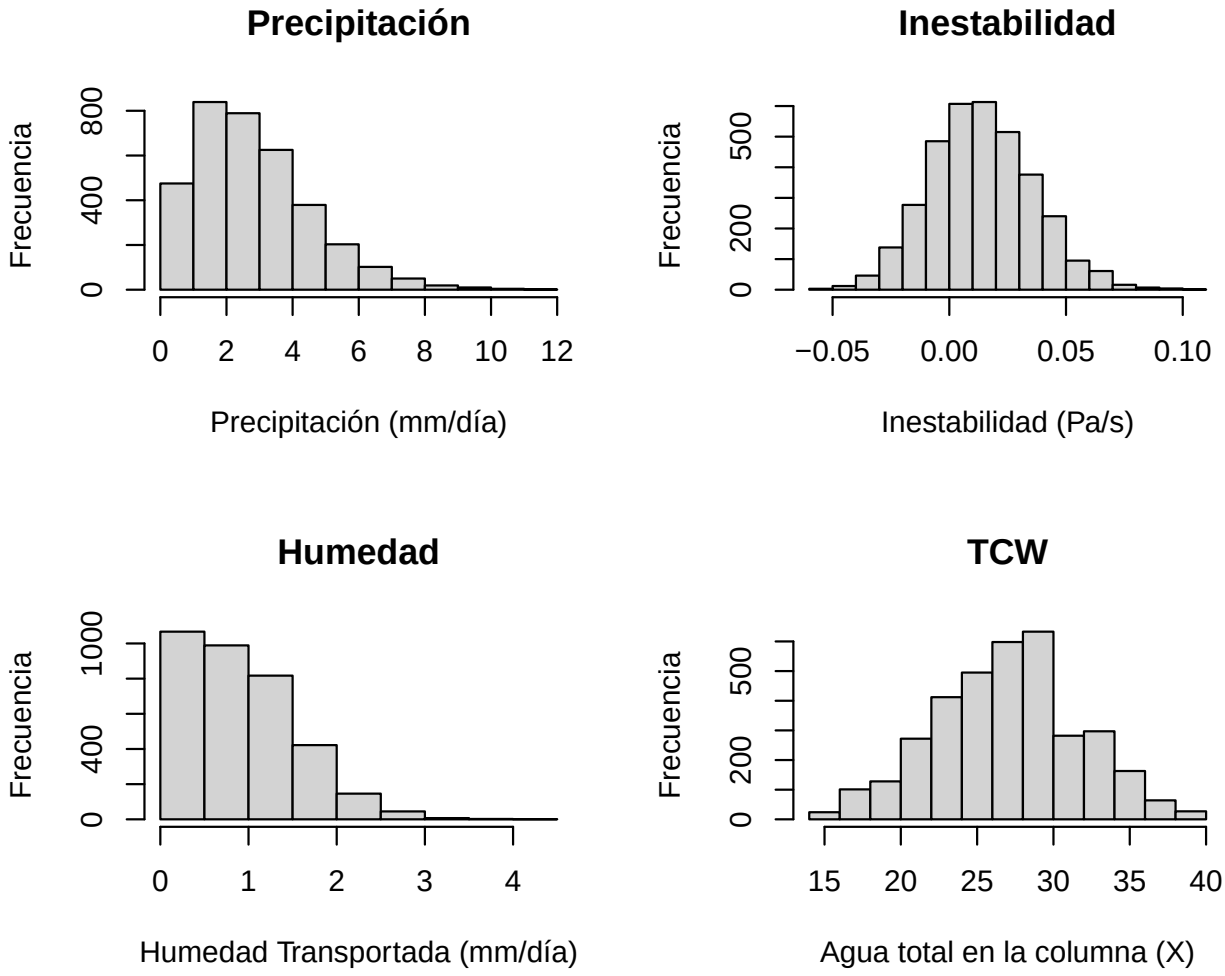

Figura 1.4: Representación de las variables precipitación, inestabilidad atmosférica, humedad transportada y agua total en la columna.

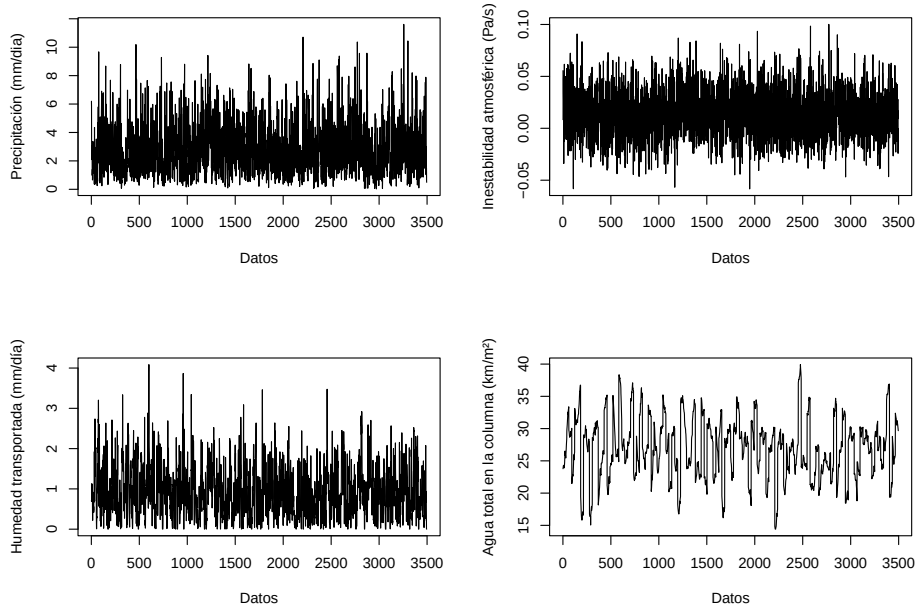

Figura 1.5: Gráficos secuenciales de las variables precipitación, inestabilidad atmosférica, humedad transportada y agua total en la columna.

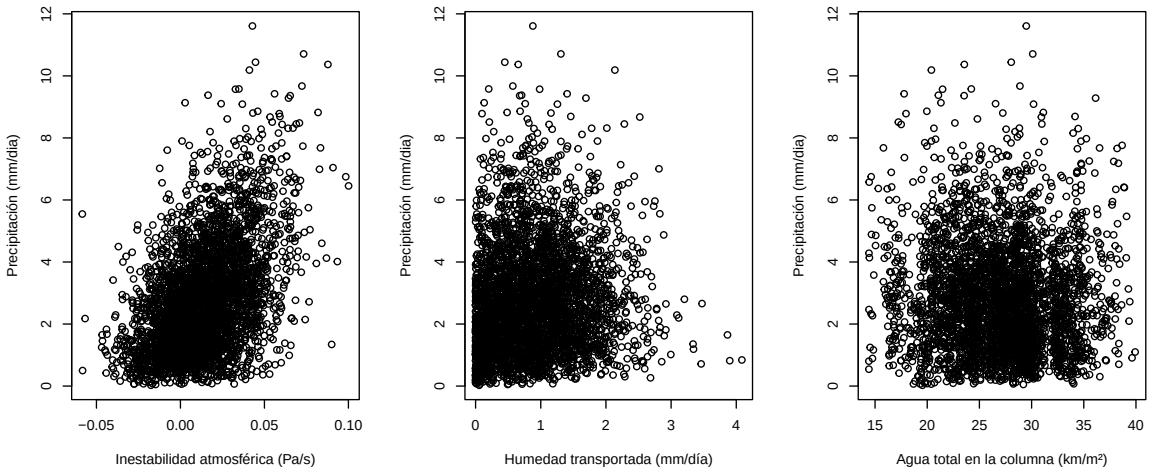

Figura 1.6: Diagramas de dispersión de la variable precipitación frente a cada una de las variables predictoras: inestabilidad atmosférica, humedad transportada y agua total en la columna.

Capítulo 2

Modelos de regresión

Este capítulo expone las técnicas estadísticas basadas en modelos de regresión utilizadas durante el desarrollo del trabajo. Primeramente se explican los modelos más simples, los modelos de regresión lineal. Debido a ciertas restricciones que presentan estos modelos es necesario aplicar otros más generales que se adapten a las características de los datos utilizados. De ahí que se extiende el análisis a los modelos aditivos generalizados y los modelos de localización y escala.

2.1. Introducción a la regresión multivariante

Cuando estamos en presencia de más de una variable predictora se dice que la regresión es multivariante. Estos modelos de regresión multivariante son conocidos también como modelos de regresión múltiple y estudian la relación entre una variable de interés o variable dependiente (respuesta) y un conjunto de variables independientes o explicativas (predictoras) (Rodríguez, 2023; Izenman, 2013).

La principal diferencia que presenta la regresión multivariante con la regresión simple, es que esta última, estudia la relación entre solamente 2 variables, es decir entre una única variable explicativa y una única variable respuesta. Es por ello que en este trabajo se empleará la regresión multivariante porque hay 3 variables explicativas (humedad transportada, inestabilidad atmosférica y agua total en la columna) y una variable respuesta (precipitación).

2.2. Modelos de regresión

Uno de los principales propósitos de las técnicas estadísticas consiste en examinar la influencia de un conjunto de variables explicativas (predictoras), conocidas como covariables, sobre una variable respuesta. Para llevar a cabo este análisis y comprender la relación entre estas variables, se recurre a los modelos de regresión, los cuales pueden ir desde lo más simple a lo más complejo dependiendo de las variables que se desean analizar y de la naturaleza de las mismas (Fox, 2019). La propuesta del primer modelo de regresión en el siglo XIX marcó un hito, convirtiendo a estos modelos en una de las aplicaciones más fundamentales tanto en el campo de la estadística, como en muchas otras ciencias, entre ellas las Ciencias Meteorológicas.

Entre los principales objetivos por los que fueron concebidos los modelos de regresión, se descatan los siguientes:

- Conocer de qué modo la variable respuesta Y depende de la(s) variable(s) explicativa(s) X, ya sea el caso de una regresión univariante o multivariante.
- Una vez construido el modelo de regresión realizar predicciones, es decir, predecir el valor de la variable Y, cuando se conoce el valor de X.

En aplicaciones reales, la variable respuesta, no puede predecirse exactamente de las variables predictoras. Es por ello, que a menudo se resume el comportamiento de la respuesta para valores fijos de los predictores utilizando medidas de tendencia central, como son: la media, la moda y la mediana (Hao y Naiman, 2007). Son varios los aspectos que se deben tener en cuenta a la hora de construir un modelo de regresión adecuado teniendo en cuenta el análisis que se desee realizar. Entre ellos se destacan el tipo de variables, tanto las explicativas como la respuesta, ya sean discretas, continuas, categóricas, así como la forma de la función de regresión, que puede ser lineal, polinómica, entre otras. Además, otras cuestiones importantes son la forma de obtener los datos muestrales con los que se están trabajando y el tipo de distribución del error.

2.2.1. Modelos de regresión lineal

La regresión, en términos generales, se suele formalizar como la media condicionada de la variable respuesta en función del valor que toma la variable explicativa. Para ello se supone un conjunto de observaciones $\{Y_i, i = 1, ..., n\}$ de una variable respuesta, junto con el valor de las covariables $\{X_1, ..., X_i\}$ que conforman toda la información disponible sobre el individuo i. La expresión queda planteada como sigue:

$$\mu(x) = E(Y \mid X). \tag{2.1}$$

De ahí que se pueda descomponer la variable respuesta en función del resultado de X, a través de la media condicionada, más un error, como se muestra a continuación:

$$Y = \mu(X) + \epsilon, \tag{2.2}$$

donde ϵ es el error, verificando que $E(\epsilon) = 0$. La estructura elegida para $\mu(X)$ determina el modelo de regresión considerado. El más simple es aquel que supone que el efecto de las covariables es lineal. Este modelo lineal considera que (Crujeiras y Conde, 2019):

$$\mu = \beta_0 + \beta_1 X_1 + ... + \beta_i X_i, \tag{2.3}$$

donde β_i, con $i = 1, ...n$ son los coeficientes de regresión del modelo (desconocido) a estimar.

El modelo de regresión queda expresado en notación matricial:

$$Y = X\beta + \epsilon, \tag{2.4}$$

siendo X la matiz de diseño.

Existen hipótesis básicas sobre las que se sustenta este modelo, las cuales se deben cumplir para llevar a cabo una buena interpretación de los resultados. Estas hipótesis se resumen a continuación (Balaguer y Ruiz, 2021):

- **Linealidad**: La función de regresión es una línea recta. Esto significa que la media de la variable respuesta Y crecerá una cantidad fija β_1 cada vez que X incrementa en una unidad. La variable respuesta tomará un valor inicial β_0 cuando la variable explicativa X sea igual a 0.
- **Homocedasticidad**: La varianza del error es la misma cualquiera que sea el valor de la variable explicativa, es decir, la dispersión de cada ϵ_i en torno a su valor esperado es siempre la misma, esto es, $Var(\epsilon / X = x) = \sigma^2$ para todo x.
- **Normalidad**: El error tiene distribución Normal, es decir, $\epsilon \in N(0, \sigma^2)$.
- **Independencia**: Las variables aleatorias que representan los errores $\epsilon_1, ..., \epsilon_n$ son mutuamente independientes, entendiendo que se obtiene una muestra de n observaciones bajo el modelo de regresión. Esta suposición dice que los n errores (no observados) serían mutuamente independientes. Esto implica que el valor del error para cada observación muestral no está influenciado por los valores de los errores correspondientes a otras observaciones muestrales.

Estimación de los parámetros por mínimos cuadrados

Con el modelo de regresión lineal planteado anteriormente, se pasa entonces a estimar los parámetros β, utilizando el procedimiento de mínimos cuadrados. Este método permite escoger los estimadores que den lugar a residuos más pequeños. Se escoge como estimador, aquel $\hat{\beta}$ que verifique que (Lema, 2022):

$$\hat{\beta} = \arg\min_{\beta} \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2 = \arg\min \sum_{i=1}^{n} (Y_i - x_i\beta)^2. \tag{2.5}$$

En notación matricial se puede expresar este problema de minimización de forma equivalente:

$$\hat{\beta} = \arg\min_{\beta} \sum_{i=1}^{n} (Y - X\beta)'(Y - X\beta). \tag{2.6}$$

Al derivar la expresión anterior se obtiene que la solución para estimar los parámetros del modelo por mínimos cuadrados es:

$$\hat{\beta} = (X'X)^{-1}X'Y. \tag{2.7}$$

Es necesario que la matriz $X'X$ sea invertible para que el estimador esté bien definido, de ahí que el número de observaciones debe ser mayor o igual que el número de covariables a considerar en el modelo.

2.2.2. Modelos Aditivos Generalizados

Con los modelos lineales exiten limitaciones con respecto a las variables explicativas ya que no nos permiten introducir variables que tengan un efecto no lineal sobre la variable respuesta, es decir, siguen considerando que la relación entre la media de la variable respuesta y las variables explicativas debe ser lineal y constante. Es por este motivo que surgen entonces, los modelos aditivos, los cuales son una extensión de los modelos lineales, con la diferencia de que la relación entre la media de la variable respuesta con cada predictor se hace a través de funciones $f_i(X_i)$. Al igual que los modelos lineales generalizados (GLM), que son una generalización de los modelos lineales, estos modelos aditivos cuentan con una generalización, apareciendo así, los Modelos Aditivos Generalizados (GAM) (Wood, 2017). Dichos modelos GAM permiten una mayor variedad de efectos de covariables que son continuas, incorporar relaciones no lineales entre las diferentes variables explicativas y la media de la variable respuesta, así como interacciones entre dichas variables. En estos modelos el predictor lineal se reemplaza por un predictor aditivo.

Un GAM, según Hastie y Tibshirani, 1990), es una extensón de un modelo lineal generalizado con un predictor lineal que implica una suma de funciones suaves de las covariables. Por tanto, el efecto de las variables explicativas sobre la variable respuesta ya no tiene que ser lineal, quedando representado como se muestra a continuación:

$$\eta_i = g(\mu_i) = X_i^*\theta + f_1(X_{1i}) + f_2(X_{2i}) + ..., \tag{2.8}$$

donde

- $\mu_i = E(Y_i)$,
- Y_i representa la variable respuesta y sigue alguna distribución de la familia exponencial,
- X_i^* son los valores de la fila i-ésima de la matriz del modelo, para cualquier componente del modelo estrictamente paramétrico,
- θ es el vector de parámetros correspondiente,
- $f_i, i = 1, ...m$ son las funciones de suavizado desconocidas que deben estimarse, que pueden ser tanto lineal como no lineal, que toma como argumentos los valores de las variables explicativas.

Estos modelos permiten una especificación bastante flexible de la dependencia de la respuesta de las covariables ya que pueden incluir tanto efectos lineales, como estimaciones no paramétricas de efectos no lineales de dichas covariables y de sus interacciones. Al especificar el modelo sólo en términos de *"funciones suaves"*, en lugar de relaciones paramétricas detalladas, se evita el tipo de modelos engorrosos y difíciles de manejar e interpretar. Esta flexibilidad y conveniencia tiene el costo de dos nuevos problemas teóricos. Las funciones suaves que se aplicarán a cada variable explicativa es necesario representarlas y elegir qué tan suaves deben ser, para evitar modelos demasiados flexibles o rígidos (Wood, 2017). El grado de suavidad de los términos del modelo se estima como parte del ajuste.

Librería de R

Para la implementación de estos modelos GAM a los datos que se están analizando, se utiliza la librería "mgcv"de R. Dicha librería representa las funciones suaves mediante splines de regresión penalizados y utiliza funciones base para estos splines, diseñadas para ser óptimas, dada la cantidad de funciones base utilizadas. Los términos suaves pueden ser funciones de cualquier número de covariables y el usuario tiene cierto control sobre cómo se mide la suavidad de las funciones. Por defecto emplea "plate regression splines penalizados multivariantes", con la opción de selección automática de los parámetros de suavizado mediante distintos criterios.

2.2.3. Funciones suaves

Como bien se mencionaba anteriormente al incorporar funciones suaves en los modelos GAM para representar la relación entre las variables explicativas y la respuesta, resulta necesario representar dichas funciones y decidir su grado de suavidad. Para la representación de los modelos GAM se pueden utilizar los splines de regresión penalizados, cuya estimación resulta de aplicar métodos de regresión penalizados. Por otra parte, el grado óptimo de suavidad para las funciones, se puede estimar utilizando diferentes métodos, como el de validación cruzada generalizada (GCV, por sus siglas en inglés). En resumen, los términos suavizados se representan mediante splines de regresión penalizados (o suavizadores similares) con parámetros de suavizado seleccionados por GCV o mediante splines de regresión con grados de libertad fijos (se permiten mezclas de los dos).

Si se permitieran absolutamente cualquier tipo de funciones suaves en el ajuste del modelo, la estimación de máxima verosimilitud de dichos modelos inevitablemente resultaría en estimaciones complejas de sobreajuste de f_i. Por esta razón, los modelos generalmente se ajustan mediante la maximización de la verosimilitud penalizada, en la que la verosimilitud del modelo se modifica mediante la adición de una penalización para cada función suave, penalizando su rugosidad. Para controlar el equilibrio entre penalizar la rugosidad y penalizar el mal ajuste, cada penalización se multiplica por un parámetro de suavizado asociado: cómo estimar estos parámetros y cómo representar las funciones suaves son las principales cuestiones estadísticas introducidas al pasar de GLM a GAM. Dichas cuestiones, de forma más detallada se pueden encontrar en Wood (2017).

2.2.4. Modelos de localización y escala

En muchas ocasiones, los supuestos de linealidad de muchos métodos parámetricos son demasiados restrictivos, y en consencuencia, se obtienen predicciones incorrectas. De ahí, que con el objetivo de contar con un modelo de regresión más flexible, que tenga una mejor capacidad de captar la posible relación que existe entre las variables predictoras a tener en cuenta y la variable respuesta, se plantea el uso de modelos de localización y escala. Dichos modelos se pueden considerar más flexibles ya que incluyen la varianza como parte de la modelación. Es por ello, que son particularmente útiles cuando se tiene sospecha que la varianza de la variable respuesta no es homógenea, sino que dependen del nivel de uno o más predictores. Los modelos de localización y escala expresan la media condicional y la varianza condicional como funciones no paramétricas aditivas de las covariables (Silva et al., 2016).

Es decir, se utilizarán los modelos aditivos generalizados mencionados anteriormente para expresar la media y varianza condicionales.

Se supone un modelo de localización y escala:

$$Y = m(X) + \sigma(X) \cdot \epsilon, \tag{2.9}$$

siendo m y σ funciones desconocidas y ϵ la variable de error cuya distribución no depende de X.

En resumen, cuando se combinan ambos apectos en este tipo de modelos, da la posibilidad de describir, no solo como cambia la variable respuesta en función de las variables predictoras, sino también como cambia la variabilidad de dicha variable respuesta. Esto proporciona una comprensión más completa de la relación entre las variables predictoras y la respuesta, en comparación con un modelo que solo predice la media. Tanto la magnitud como la variabilidad de los resultados constituyen información relevante para la toma de decisiones, lo que justifica la implementación de los modelos de localización y escala (Viechtbauer y López-López, 2022).

2.3. Series de tiempo

En este trabajo se utilizan un conjunto de observaciones diarias, en un período de 3 meses, para diferentes variables meteorológicas. Es por ello que resulta necesario realizar un breve recordatorio de las series de tiempo.

Una serie de tiempo es el resultado de una colección de observaciones de una variable, X, tomadas secuencialmente a lo largo del tiempo en intervalos regulares (cada día, hora, mes, ...) (Cowpertwait y Metcalfe, 2009). El objetivo fundamental del estudio de las series temporales es el conocimiento del comportamiento de una variable a través del tiempo para, a partir de dicho conocimiento, y bajo el supuesto de que no van a producirse cambios estructurales, poder realizar predicciones, es decir, determinar qué valor tomará la variable objeto de estudio en uno o más períodos de tiempo situados en el futuro, mediante la aplicación de un determinado modelo calculado previamente.

Un proceso estocástico se define como un conjunto de variables aleatorias asociadas a distintos instantes de tiempo. Así, en cada período o momento temporal se dispone de una variable que tendrá su correspondiente distribución de probabilidad. La relación existente entre una serie temporal y un proceso estocástico que la genera es análoga a la que existe entre una muestra y la población de la que procede, de tal forma que podemos considerar una serie temporal como una muestra o realización de un proceso estocástico, formada por una sola observación de cada una de las variables que componen el proceso (Parra, 2019).

En este trabajo, resultó importante determinar la dependencia de los errores de los modelos ajustados. Entre los criterios que se puede utilizar para analizar las correlaciones entre los residuos del modelo para describir la serie temporal son los gráficos de la función de autocorrelación (ACF). Dichos gráficos muestran si hay alguna correlación lineal entre los residuos en diferentes retardos. Si los residuos son independientes, la ACF debería caer rápidamente hacia cero.

2.4. Criterio de selección de variables

Es importante resaltar que en situaciones que involucran una gran cantidad de variables, que pueden ser o no relevantes para hacer predicciones sobre la variable respuesta, es útil reducir la complejidad del modelo. De esta manera se garantiza que contenga sólo las variables que brindan información importante sobre la variable de respuesta. Como regla general, un aumento en el número de variables incluidas en un modelo proporciona un ajuste aparentemente mejor de los datos observados. Sin embargo, estas estimaciones no siempre son satisfactorias por diferentes razones. Por un lado, la inclusión de variables irrelevantes aumentaría la varianza de las estimaciones, resultando en una pérdida parcial de la capacidad predictiva del modelo. Además, la inclusión de muchas variables suele producir un mo-

delo difícil de interpretar. Por tanto, es necesario proporcionar herramientas de análisis que permitan determinar el mejor modelo a utilizar.

2.4.1. Criterio de información bayesiano

En este contexto han sugerido muchos criterios diferentes para comparar diferentes modelos, asignando una puntuación a cada uno de ellos, basado en algún principio estadístico subyacente: criterio del coeficiente de determinación (R^2), criterio de información de Akaike (AIC), criterio de información bayesiano (BIC), entre otros.

En este trabajo se utilizó el criterio BIC, el cual fue introducido por Gideon E. Schwarz en 1978 y está estrechamente relacionado con el criterio AIC, publicado en 1974. El BIC se deriva para servir como una aproximación asintótica a una transformación de la probabilidad bayesiana a posteriori de un modelo candidato. Cuando se cuenta con muestras grandes, el modelo seleccionado por el criterio BIC, corresponde idealmente al modelo candidato que es el más probable a posteriori, es decir, el modelo más plausible teniendo en cuenta los datos disponibles. El cálculo del BIC se basa en la probabilidad logarítmica empírica y no requiere la especificación a priori, de ahí que sea de gran utilidad para problemas de modelados donde los antecedentes son difíciles de establacer con precisión (Neath y Cavanaugh, 2012).

La aproximación en la que basa el BIC se obtiene a partir de la siguiente expresión:

$$-2\log(L(\hat{\theta}) + k\log(N), \tag{2.10}$$

donde

- N es el número de observaciones,
- k es el número de parámetros que estima el modelo,
- θ es el conjunto de todos los parámetros y,
- $L(\hat{\theta})$ representa la probabilidad del modelo probado, dado sus datos, cuando se evalúa con valores de máxima verosimilitud de θ.

Por otro lado, en ciertos contextos, la selección de modelos según el criterio BIC es equivalente a la selección de modelos basada en factores de Bayes. De esta manera, se puede realizar comparaciones entre modelos cuantificadas por el BIC diferencial, con respecto al modelo que tiene el menor valor de BIC. En la Tabla 2.1 se muestra un resumen de la comparación entre dos modelos M_1 y M_2, siendo este último el que presenta menor valor de BIC (Neath y Cavanaugh, 2012):

Tabla 2.1: Criterio BIC para la comparación de modelos.

Diferencias	Evidencia a favor de M_2 sobre M_1
0-2	No vale más que una simple mención
2-6	Positiva
6-10	Fuerte
> 10	Muy fuerte

En este trabajo, para la selección de las variables en los modelos se aplicó un algoritmo que probaba todas las posibles combinaciones entre dichas variables a incorporar. La combinación de las variables en cada uno de los modelos se evalúa, tanto de una, dos o la cantitad total posible y se calcula su valor de BIC correspondiente. Seguidamente, se ordenan los modelos según el valor de BIC obtenido en forma ascendente y se selecciona como mejor modelo aquel que presente el menor valor de BIC. En

resumen, con este criterio de BIC, se proporciona un equilibrio entre la complejidad del modelo y el ajuste de los datos.

2.4.2. Coeficientes de determinación

Los valores de los coeficientes de determinación (R^2) y R^2 ajustado (R^2_{ajust}) fue otro criterio tenido en cuenta a la hora de seleccionar el mejor modelo. El primero de ellos es un indicador de la bondad de ajuste del modelo, muestra lo bien que se ajusta un modelo de regresión a un conjunto de datos y hace referencia al porcentaje explicado de dicho modelo. No tiene unidades de medidas y toma valores entre 0 y 1, será más cercano a 1 a medida que aumente la cantidad de variables incorporadas en el modelo, a pesar de que no todas sean significativas. Es por ello, que para resolver este problema aparece el R^2_{ajust}. Este coeficiente es mejor que el R^2, ya que mide la bondad de ajuste penalizando por cada variable añadida (Novales, 2010). Además el R^2_{ajust} es mejor para comparar dos modelos de regresión que incluyan diferente número de variables explicativas.

El valor de R^2_{ajust} se expresa como sigue (Rodríguez, 2023):

$$R^2_{ajust} = (1 - R^2)\frac{n-1}{n-k}, \tag{2.11}$$

donde n es el número de observaciones y k el número de variables explicativas. De esta manera, al incorporar variables explicativas al modelo, el término (1-R^2) disminuirá y el término $\frac{n-1}{n-k}$ aumentará. Mientras que el R^2 se calcula mediante la siguiente expresión:

$$R^2 = 1 - \frac{\sigma_r^2}{\sigma^2} = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}, \tag{2.12}$$

donde

- σ_r^2 es la varianza residual,
- σ^2 es la varianza de la variable dependiente Y,
- y_i es el valor de la variable dependiente de la observación i,
- $\hat{y}_i$ es el valor aproximado por el modelo de regresión para la observación i,
- $\bar{y}$ es la media de la variable dependiente de todas las observaciones.

Capítulo 3

Regresión cuantil

3.1. Introducción a la regresión cuantil

En los modelos de regresión, los errores se asumen como una sucesión e_n de variables aleatorias independientes e indénticamente distribuidas con media cero ($E(e_n = 0)$) y una distribución normal. Sin embargo, esta última suposición no siempre se cumple ya que la distribución puede ser asimétrica. Como solución a dichos problemas, Koenker y Bassett (1978), introdujeron el concepto de regresión cuantílica y demostraron que los estimadores por cuantiles son más eficientes que el estimador de máxima verosimilitud de muchos modelos paramétricos convencionales. En los métodos de regresión clásicos el objetivo es minimizar la suma de los residuos al cuadrado y utilizar la media como estimador, mientras que la regresión cuantil busca minimizar una suma de errores absolutos ponderados con pesos asimétricos, utilizando los cuantiles como estimadores (López y Mora, 2007). La localización de los cuantiles asegura un tipo de robustez, por lo que debido a su importancia, en muchos casos se adaptan una gran cantidad de técnicas de inferencia relacionadas con la tradicional regresión en media a los modelos de regresión cuantil. Esta se utiliza cuando se pretende obtener una estimación de las diferentes posiciones, en este caso denomidadas cuantiles, de una variable de interés, o variable respuesta, en función de ciertas variables explicativas.

La regresión cuantil es considerada una técnica estadística de gran utilidad en diversas ramas de las ciencias, ya que realiza un análisis más completo y detallado de los datos con los que se pretenda trabajar. De forma general, tiene el mismo objetivo que la regresión con la media condicional, que es representar la relación entre la variable respuesta del modelo y un conjunto de covariables. Sin embargo, cuando la distribución es altamente asimétrica, la media puede resultar díficil de interpretar, mientras que la mediana permanece altamente informativa. Como consecuencia, el modelado de la mediana condicional tiene el potencial de ser más útil. Además permiten abordar problemas de regresión con datos más complejos, una mejor descripción del comportamiento de la variable respuesta y se pueden utilizar en situaciones bajo condiciones más generales de la distribución del error. Además las estimaciones de regresión cuantil son más robustas frente a los valores extremos. De ahí también su ventaja con respecto a una regresión en media.

La mediana es un cuantil especial, el cual describe la localización central de la distribución, mientras que otros cuantiles pueden utilizarse para describir posiciones no centrales de dicha distribución. La noción del cuantil generaliza términos específicos como cuartil, decil y percentil. El $p-$ésimo cuantil denota el valor de la respuesta por debajo del cual se encuentra una proporción p de la población, por ello los cuantiles pueden especificar cualquier posición de la distribución (Hao y Naiman, 2007).

3.1.1. Definición de cuantil

Para una mejor compresión, se definen a continuación una serie de criterios importantes de la regresión por cuantiles (Conde, 2013).

Dada cualquier variable aleatoria $X : \omega \to R$ definida en un espacio muestral ω asociado a un experimento aleatorio, estará caracterizada por su **función de distribución** que viene determinada por la siguiente expresión:

$$F(x) = P(X \leq x). \tag{3.1}$$

Si la variable aleatoria X es discreta, es decir, su recorrido es un conjunto discreto, entonces su función de distribución viene dada por:

$$F(x) = P(X \leq x) = \sum_{x_i \leq x} P(X = x_i), \tag{3.2}$$

donde $x_i \leq x$ se corresponde con los valores que toma la variable X inferiores o iguales al valor de x.

Si la variable aleatoria X es continua, es decir, si su recorrido no es un conjunto numerable, entonces su función de dstribución viene determinada por la siguiente expresión:

$$F(x) = P(X \leq x) = \int_{-\infty}^{x} f(x)dx, \tag{3.3}$$

para cualquier valor de x donde $f : R \longrightarrow R$ se conoce como la función de densidad, la cual es una función no negativa e integrable.

A continuación se pasa a definir entonces el cuantil de orden τ, para cada $0 < \tau < 1$, dada cualquier variable aleatoria X. Dicho cuantil se denota como c_τ y es el valor que verifica que:

$$P(X \leq c_\tau) \geq \tau, \tag{3.4}$$

$$P(X \geq c_\tau) \geq 1 - \tau. \tag{3.5}$$

Si la variable aleatoria X es una variable continua, pues el cuantil de orden τ se verifica como:

$$P(X \leq c_\tau) = P(X < c_\tau) = \tau. \tag{3.6}$$

Cualquier variable aleatoria X puede ser caracterizada por su función de distribución, ya sea discreta o continua para un $0 < \tau < 1$. De esta manera aparece la función cuantil de una distribución de probabilidad, la cual se corresponde con la inversa de la función de distribución. Dicha función inversa, en algunos casos no está bien definida, por lo que se establece como alternativa la siguiente expresión:

$$F^{-1}(\tau) = inf\{x : F(x) \geq \tau\}. \tag{3.7}$$

La mediana, $F^{-1}(1/2)$ juega un rol central (Koenker y Bassett, 1978). Para una probabilidad de $0 < \tau < 1$, la función cuantil devuelve el valor mínimo de x para el cual se mantiene la probabilidad anterior.

Específicamente se pueden distinguir los siguientes casos:

1. La función cuantil, F^{-1} dada una función de distribución $F : R \to (0, 1)$ continua y estrictamente monótona, devuelve un valor de x tal que (ejemplo Figura 3.1):

$$P(X \leq x) = \tau. \tag{3.8}$$

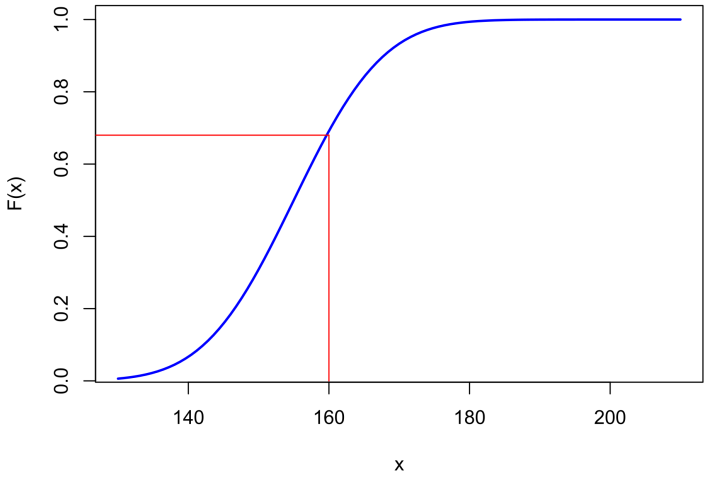

Figura 3.1: Ejemplo de función de distribución continua.

2. En el caso de que la función de distribución fuese discreta, pueden haber saltos entre los valores del dominio de dicha función (ejemplo Figura 3.2) .

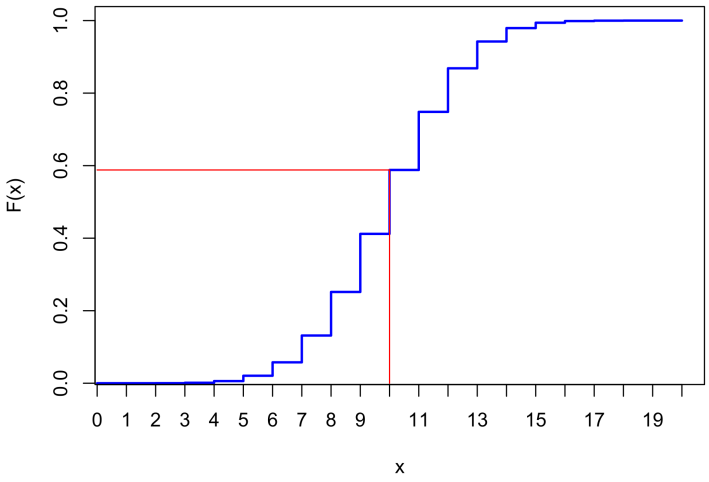

Figura 3.2: Ejemplo de función de distribución discreta.

3. Por útlimo, si la función de distribución fuera monótona no estricta, pueden haber intervalos en los que el valor de la función se mantiene constante (ejemplo Figura 3.3).

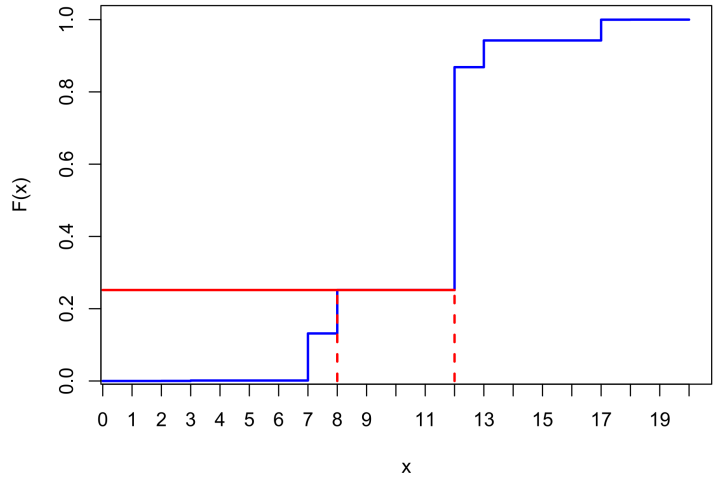

Figura 3.3: Ejemplo de función de distribución monótona no estricta.

3.1.2. Función de pérdida cuantílica

Los cuantiles surgen de un simple problema de optimización. El concepto de cuantil no se aplica directamente a la regresión, sino que se aplica a través de la función de pérdida cuantílica. Para una variable dada U con $u \in U$ y un cuantil de orden $\tau \in (0, 1)$, la función de pérdida cuantílica viene determinada por la siguiente función lineal definida a trozos (Figura 3.4):

$$\rho_\tau(u) = u(\tau - I(u < 0)). \tag{3.9}$$

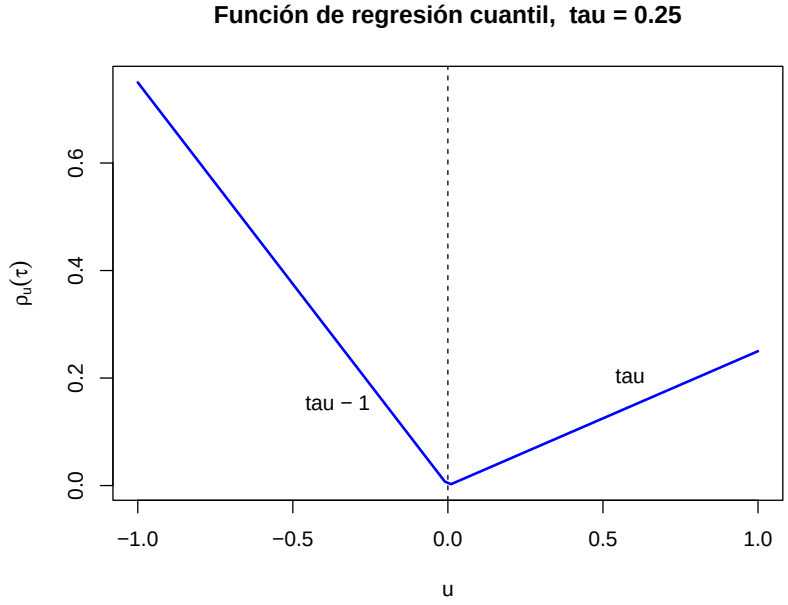

Figura 3.4: Función de regresión cuantil ρ.

Se pueden dar dos casos:

- si $u \geq 0$ entonces $\rho_\tau(u) = u\tau$,
- si $u < 0$ entonces $\rho_\tau(u) = u(\tau - 1)$.

Dependiendo del valor del cuantil se penalizarán más las observaciones superiores o inferiores a dicho valor. Por ejemplo, si se consideran valores entre -1 y 1 y se utilizan diferentes valores de cuantiles las penalizaciones de las observaciones quedan como sigue (Figura 3.5):

- $0 < \tau < 0.5$: los valores negativos serán penalizados más que los positivos. Se desplazan las estimaciones hacia posiciones inferiores,
- $\tau = 0.5$: los valores serán igualmente penalizados,
- $0.5 < \tau < 1$: los valores positivos serán penalizados más que los negativos. Se desplazan las estimaciones hacia posiciones superiores.

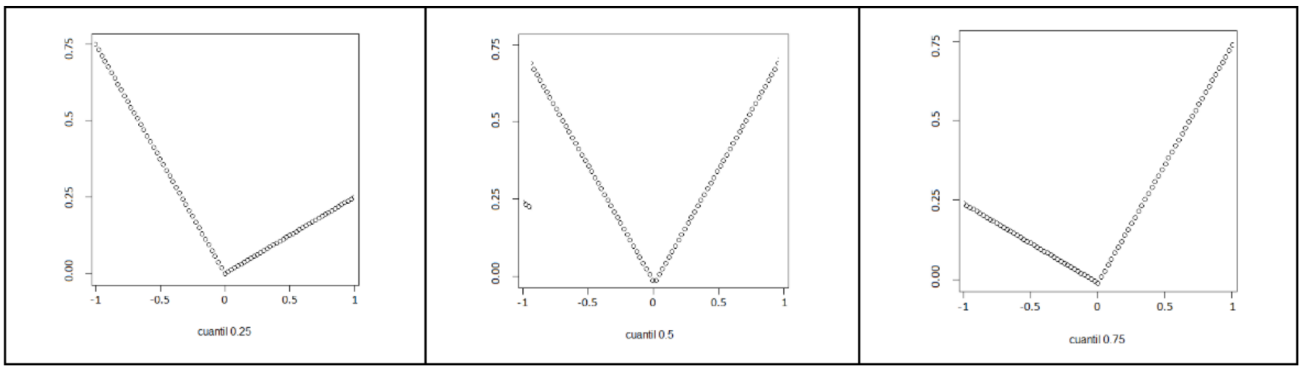

Figura 3.5: Representación de funciones de pérdida cuantílica para diferentes valores del cuantil τ. Tomado de (Díaz, 2020).

El objetivo es encontrar $\hat{x}$ para minimizar la pérdida esperada, es decir, encontrar un elemento c_τ, para cada $\tau \in (0, 1)$, que minimice la pérdida, lo que queda planteado de la siguiente manera:

$$E_{\rho_\tau}(X - \hat{x}) = (\tau - 1) \int_{-\infty}^{\hat{x}} (x - \hat{x}) dF(x) + \tau \int_{\hat{x}}^{\infty} (x - \hat{x}) dF(x). \tag{3.10}$$

Diferenciando la expresión anterior respecto a x se tiene:

$$0 = (1 - \tau) \int_{-\infty}^{\hat{x}} dF(x) - \tau \int_{\hat{x}}^{\infty} dF(x) = F(\hat{x}) - \tau. \tag{3.11}$$

Dado que F es monótona, cualquier elemento de $\{x : F(x) = \tau\}$ minimiza la pérdida esperada. Cuando la solución es única, $\hat{x} = F^{-1}(\tau)$, de lo contrario se tiene un intervalo de cuantiles de τ, del cual el elemento más pequeño debe ser elegido para adherirse a la convención de que la función cuantílica empírica sea continua a la izquierda (Koenker, 2005).

3.2. Regresión cuantil basada en modelos de regresión lineal

3.2.1. Función rq

Librería quantreg de R

El software estadístico R, utilizado para la obtención de los resultados de este trabajo, presenta rutinas basadas en la regresión cuantil, las cuales tienen como objetivo ajustar una recta de regresión. Específicamente, la librería *quantreg*, incluye la función *rq*, que una vez introducidos los valores de las variables explicativas y de la variable respuesta, obtiene una recta de regresión teniendo en cuenta el

orden del cuantil, a partir del parámetro *tau*. A menos que se indique lo contrario, por defecto, este cuantil será la mediana (Koenker, 2005).

En resumen, la función calcula una estimación de la función cuantil condicional **tau** de la variable respuesta, dadas las covariables, suponiendo una especificación lineal para el modelo de regresión cuantil. Dicha función minimiza una suma ponderada de residuos absolutos que pueden formularse como un problema de programación lineal. El método utilizado se describe en Koenker y dOrey (1987, 1994). Este método es eficiente, y más para problemas con miles de observaciones, que es el caso de esta trabajo, y puede usarse para calcular el proceso de regresión cuantil completo.

3.2.2. Regresión cuantil flexible

Los modelos de regresión lineales pueden ser demasiados rígidos y en ocasiones no ajustan ni se adaptan bien a la naturaleza de los datos, por lo que surge la necesidad de utilizar otros modelos más flexibles. Con la función *rq*, mencionada anteriormente, se pueden obtener estimaciones más flexibles incluyendo bases de B-Splines.

Las B-Splines son una serie de funciones base que se utilizan para construir curvas o superficies suavizadas. Además pueden ser usadas para modelar componentes suavizados de un modelo permitiendo una mayor flexibilidad y suavidad en la forma funcional de la relación modelada. Específicamente son muy útiles cuando la relación entre las variables dependientes e independientes no es estrictamente lineal, o también cuando dicha relación puede cambiar de forma para diferentes niveles de las variables independientes (De Boor y De Boor, 1978).

3.3. Regresión cuantil con modelos GAM

Como bien se ha mencionado, los modelos GAM, son modelos de regresión no lineal flexibles, que se pueden ajustar de manera eficiente utilizando los métodos bayesianos aproximados. La función **bf** de R, proporciona métodos bayesianos calibrados rápidos para ajustar los cuantiles de dichos modelos GAM. Los QGAM se basan en una versión suavizada de la pérdida de pinball de (Koenker y Bassett 1978), en lugar de la probabilidad. La ventaja del uso de esta función para determinar los cuantiles de modelos GAM, es precisamente que proporcionan más flexibilidad al modelar los cuantiles de la distribución de la respuesta condicional individualmente, evitando así cualquier supuesto paramétrico sobre la distribución de la variable respuesta. La suavidad de la nueva pérdida se determina minimizando el error cuadrático medio (MSE) asintótico de los coeficientes de regresión estimados, los cuales son aproximados utilizando un modelo GAM de localización y escala.

Para una información más detallada de esta metodología aplicada en los modelos QGAM, se puede consultar (Koenker y Bassett 1978).

3.4. Regresión cuantil basada en modelos de localización y escala

La regresión cuantil basada en modelos de localización y escala expresa la media condicional y la varianza condicional como funciones no paramétricas aditivas de las covariables. A continuación se supone un modelo de localización y escala:

$$Y = m(X) + \sigma(X) \cdot \epsilon, \tag{3.12}$$

con m y σ funciones desconocidas y ϵ la variable de error cuya distribución no depende de X. Bajo este modelo, el cuantil de orden p de Y dado x viene dado directamente por

$$y_p(x) = m(x) + \sigma(x) \cdot \epsilon_p, \tag{3.13}$$

siendo ϵ_p el cuantil de orden p de la distribución de la variable de error ϵ.

Para obtener la estimación de $y_p(x)$ hay que seguir los siguientes pasos que se mencionan a continuación:

- Obtener la estimación de $\hat{m}$ ajustando un modelo de regresión no paramétrico de Y sobre X.
- Obtener la estimación de σ^2 ajustando un modelo de regresión de respuesta $(Y - \hat{m}(X))^2$ sobre X.
- Calcular e_p definido como el cuantil de orden p de los errores estandarizados.

$$\hat{\epsilon} = \frac{Y - \hat{m}(X)}{\sigma(X)}. \tag{3.14}$$

- Finalmente se obtiene la estimación

$$\hat{y}_p(x) = \hat{m}(X) + \hat{\sigma}(X) \cdot \hat{\epsilon_p}, \tag{3.15}$$

donde ϵ_p es el cuantil empírico de orden p de los errores obtenidos en el paso anterior.

Para estos modelos se utiliza la librería **mgcv** y la función **gam**, explicada en la sección 2.2.2.

Capítulo 4

Análisis y discusión de los resultados

En este capítulo se exponen los resultados obtenidos utilizando las variables de interés. Se aplicaron diferentes modelos de regresión, desde modelos más sencillos a modelos más complejos para analizar la relación entre la precipitación (Prec), como variable respuesta, y las variables predictoras inestabilidad atmosférica (Inest), humedad transportada (Trans.Humed) y agua total en la columna (TCW).

4.1. Regresión lineal

Como bien se comentó en capítulos anteriores, el modelo de regresión más sencillo, es el modelo lineal, de ahí que para analizar la relación entre las diferentes variables de interés se comenzó aplicando un modelo de regresión lineal, el cual queda planteado como sigue:

```
modelo <- lm(Prec ~ Inest+Trans.Humed+TCW)
```

Al realizar el summary se obtienen los siguientes resultados:

```
Call:
lm(formula = Prec ~ Inest+Trans.Humed+TCW)

Residuals:
    Min      1Q  Median      3Q     Max
-4.2116 -1.1047 -0.1841  0.8777  7.8096

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  2.571829   0.153424  16.763  < 2e-16 ***
Inest       35.381490   1.181311  29.951  < 2e-16 ***
Trans.Humed  0.165245   0.042868   3.855 0.000118 ***
TCW         -0.014744   0.005752  -2.563 0.010409 *
---


Residual standard error: 1.542 on 3492 degrees of freedom
Multiple R-squared:  0.2126, Adjusted R-squared:  0.2119
F-statistic: 314.3 on 3 and 3492 DF,  p-value: < 2.2e-16
```

El R^2_{ajust} tiene un valor muy pequeño, solo de 0.2119, ligeramente menor que el R^2, lo que significa que la relación entre las variables no es capaz de representarse correctamente con un modelo lineal. Este valor indica que la inestabilidad, la humedad transportada y el agua total en la columna, explican aproximadamente, el 21 % de la variabilidad de la precipitación en este modelo lineal. Las dos primeras variables resultan ser significativas para cualquier nivel de significancia, sin embargo el agua total en la columna, solo es significativa para un 90 %. Dicho resultado, da una primera idea de cuales condiciones pudieran haber influido más, en este caso, en la precipitación en la región de interés.

Se realizó la diagnosis del modelo ajustado, para comprobar las hipótesis del modelo lineal, aplicando los test correspondientes para analizar la normalidad, homocedasticidad e independencia. En ninguno de los casos se aceptaron dichas suposiciones, obteniéndose p-valores menores que cualquier nivel de significancia. Además se analizó la autocorrelación de los residuos del modelo lineal anterior, los cuales no son más que las desviaciones de los puntos de los valores reales a la recta de regresión ajustada. De esta manera se puede comprobar si el modelo planteado ha captado toda la información relevante de los datos analizados. El gráfico ACF se muestra en la Figura 4.1.

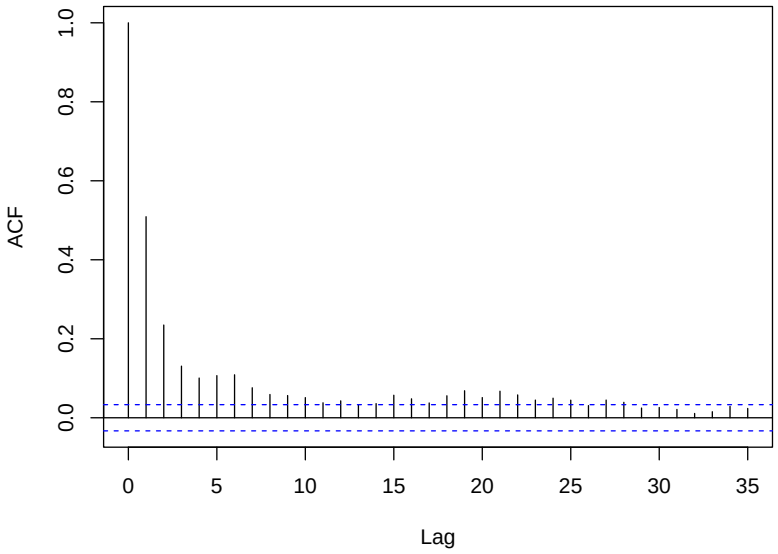

Figura 4.1: Autocorrelación de los residuos del modelo lineal original.

Teniendo en cuenta las líneas azules que limitan los umbrales de significancia, hay barras, las cuales se corresponden cada una con un lag (retardo) específico, que se encuentran por encima de este intervalo. Esto significa la presencia de una alta autocorrelación, sobre todo para los primeros retardos de las variables, lo que podría indicar un problema en el modelo lineal propuesto. Se pueden aplicar diferentes métodos para solucionar este problema de autocorrelación, por ejemplo, aplicar transformaciones a los datos, ajustar un modelo diferente o más complejo, o introducir términos de retardos adicionales. Teniendo en cuenta que se está trabajando con variables meteorológicas, las cuales en muchas ocasiones están influencias por condiciones que ocurren en días anteriores, la última cuestión mencionada puede ser una solución.

Para corregir esta autocorrelación en los residuos del modelo, se incorporaron como nuevas variables predictoras, las mismas variables independientes pero con retardos igual a 1, incluso de la propia variable precipitación. Se pasa de contar con 3 variables para analizar la precipitación a tener 7 variables. De esta manera se espera corregir la autocorrelación, en los primeros retardos, que se observaron en

el gráfico de ACF mostrado anteriormente. Se debe tener en cuenta, que en el momento de incorporar estas nuevas variables, la base de datos utilizada se modifica ligeramente, ya que para incorporar el primer retardo, la serie tendrá que comenzar el 2 de junio de 1980. En resumen, el modelo de regresión lineal sería la precipitación de un día, en función de la inestabilidad, el transporte de humedad y el agua total en la columna de ese mismo día, así como de la inestabilidad, el transporte de humedad, el agua total en la columna y la precipitación del día anterior.

Separación de la base de datos

La nueva base de datos modificada ligeramente con la nueva incorporación de las variables con retardo igual a 1, es considerada entonces, como la base de datos final para realizar el análisis. Es por ello que, teniendo todas las variables a considerar en el modelo, se decidió separarla en dos muestras. Una de ellas será la muestra de entrenamiento para ajustar los modelos de regresión que se van a utilizar y una muestra test, para evaluar la precisión de las predicciones. Para ello se consideró como muestra de entrenamiento el 80 % de los datos y como muestra test el 20 % restante.

Selección de variables

Al incorporar nuevas variables predictoras en el modelo, surge la necesidad de aplicar un criterio de selección para reducir la complejidad del mismo y a su vez tener en cuenta solo las variables más significativas. El criterio aplicado para la selección fue el BIC, el cual busca un equilibrio entre la bondad de ajuste del modelo y su complejidad, imponiendo una penalización más fuerte por el número de parámetros estimados, en comparación con el criterio AIC. Esto lo hace menos propenso a sobreajustar, especialmente en muestras grandes. Además la penalización con el BIC, se basa directamente en el número de datos, lo que hace que su interpretación y comparación entre modelos sea más intuitiva.

Se analizaron todas las combinaciones posibles de variables explicativas, incluyendo las que tienen lag 1. En la Tabla 4.1 se presentan las variables a incluir de los primeros 10 mejores modelos según este criterio. En ella, Inest e Inest1, se corresponden con la variable inestabilidad atmosférica sin retardo y con retardo igual a 1, respectivamente. Trans.Humed y Trans.Humed1 son la humedad transportada desde la fuente, sin lag y con lag 1, respectivamente. El agua total en la columna sin retardo y con retardo igual a 1, se corresponden con TCW y TCW1, respectivamente. Y por último Prec1 es la precipitación con retardo igual a 1.

Se destaca que en todos los modelos se incorporan las variables Prec1, Inest e Inest1. Seguido de estas 3 variables se encuentra Trans.Humed1, incorporada en 8 de los 10 modelos presentados, mientras que sin retardo se incorporó solo en 6. Esto tiene mucho sentido, porque la humedad transportada desde el Golfo de México hasta las Grandes Llanuras Americanas, se corresponde con días atrás, teniendo en cuenta toda el área que ocupa esta fuente de humedad y el tiempo que demora en llegar todo este contenido de humedad hasta la región de interés. Y por último, como era de esperar, según la significancia mostrada en el primer ajuste lineal realizado, las variables TCW y TCW1, se incorporaron en menos modelos.

Teniendo en cuenta estos resultados, se puede concluir que el mejor modelo, con un valor de BIC igual a 9150.219, es el que incorpora 4 de las 7 posibles variables predictoras: Inest, Inest1, Trans.Humed1 y Prec1. Según el orden el octavo mejor modelo, con un valor de BIC igual a 9163.796, incorpora solo 3 variables, de las mismas anteriores, elimina Trans.Humed1. Esto significa que el mejor modelo no se corresponde con el que menos variables tiene, entre estos 10 mejores, aspecto a tener en cuenta, debido a que la incorporación de más variables en el modelo, hace que este sea más complejo y difícil de interpretar. Es por ello que se decidió analizar estos dos modelos, el primero por ser el que menor valor de BIC presenta y el otro, porque entre los mejores, fue el que menos variables incorporó.

A la hora de seleccionar uno de estos modelos, uno de los criterios a tener en cuenta, además de la cantidad de variables incorporadas, fue la diferencia de BIC que hay entre ellos. En este caso, dicha diferencia es de aproximadamente 13.5, siendo un valor superior a 10, lo que significa que la diferencia

Tabla 4.1: Mejores modelos según el criterio de selección de variables BIC.

Orden	Variables	Inest	Trans.Humed	TCW	Inest1	Trans.Humed1	TCW1	Prec1	BIC
1	**4**	✓			✓	✓		✓	**9150.219**
2	5	✓		✓	✓	✓		✓	9152.418
3	5	✓	✓		✓	✓		✓	9153.742
4	5	✓			✓	✓	✓	✓	9154.268
5	6	✓	✓	✓	✓	✓		✓	9156.644
6	6	✓	✓		✓	✓	✓	✓	9158.096
7	6	✓	✓		✓	✓	✓	✓	9159.489
8	**3**	✓			✓			✓	**9163.796**
9	7	✓	✓	✓	✓	✓	✓	✓	9164.039
10	4	✓	✓		✓			✓	9164.592

es muy fuerte entre ambos modelos. Otro aspecto fue la determinación de su R^2_{ajust}, obteniéndose prácticamente el mismo valor para ambos modelos.

- Modelo 1: 0.4965.
- Modelo 8: 0.4927.

Por último, en la Figura 4.2, se muestran las comparaciones de las predicciones realizadas por ambos modelos. Al contar en la muestra test con 699 días, para poder observar mejor el comportamiento de las predicciones, se decidió dividirla en 3 intervalos de 15 días cada uno, en el inicio, medio y final de las predicciones.

Se comprueba con este gráfico que las predicciones para ambos modelos son muy similares, por lo que en este sentido tampoco presentan diferencias. A su vez, de manera general, siguen el patrón de los valores reales de la precipitación, sin embargo, ninguno es capaz de predecir correctamente los valores más extremos, sobreestimando en muchos de los días. Esto se relaciona, con los bajos valores de R^2, que expresan la baja variabilidad de la variable respuesta que son capaces de explicar las variables predictoras que incorporan.

Por todas las cuestiones planteadas anteriormente, se propone seleccionar para continuar con la representación de la relación entre la precipitación y el resto de las variables predictoras, el modelo 1, que se corresponde con el mejor modelo propuesto por el criterio de selección BIC. Teniendo en cuenta que los valores de R^2_{ajust} y las predicciones son muy similares, se consideró importante la diferencia entre los valores de BIC entre ambos modelos, sugiriendo una diferencia muy fuerte entre ellos.

4.1.1. Regresión cuantil basada en modelos de regresión lineal

Teniendo en cuenta el modelo lineal seleccionado anteriormente con las variables predictoras incluidas, se puede pasar a aplicar la regresión cuantil. Con el uso de los cuantiles se podrá realizar un análisis proporcionando una visión más completa y detallada del efecto de los predictores en toda la

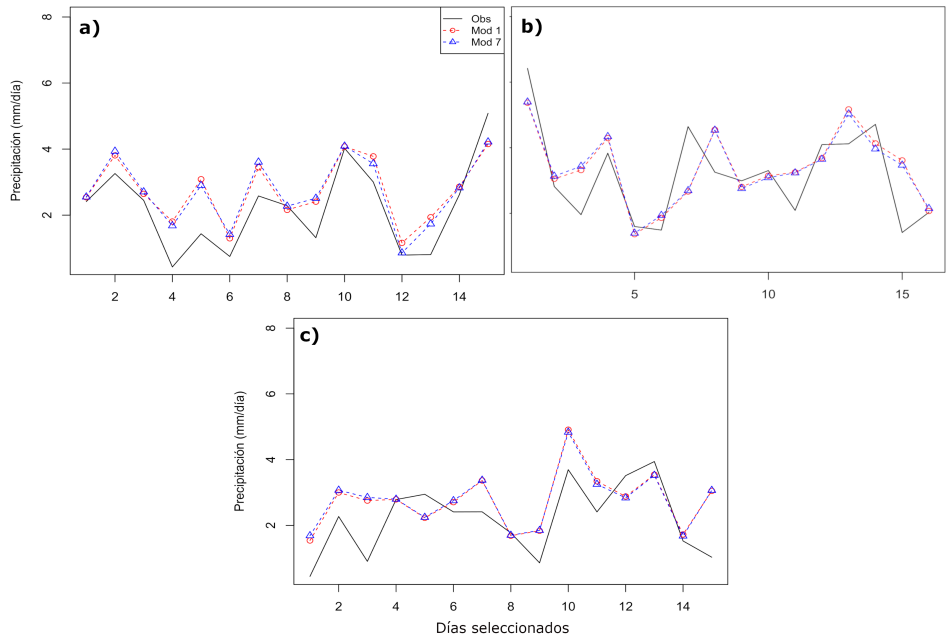

Figura 4.2: Comparación entre los valores observados (línea continua negra) y las predicciones realizadas por ambos modelos de regresión lineal. Las líneas discontinuas roja y azul son las predicciones de los modelos 1 y 7, respectivamente. a) inicios, b) mediados y c) final.

distribución de la variable respuesta. Además, como la variable respuesta es precipitación puede ser interesante para encontrar valores atípicos o extremos.

El summary para el modelo lineal seleccionado, con las 4 variables predictoras incluidas, se muestra a continuación:

```
Call:
lm(formula = P_resp ~ Inest + Prec1 + Trans.Humed1 + Inest1, data = train)

Residuals:
    Min      1Q  Median      3Q     Max
-4.6397 -0.8275 -0.1077  0.6697  6.4605

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)      0.69441    0.05653  12.284  < 2e-16 ***
Inest           31.90515    1.13749  28.049  < 2e-16 ***
Prec1            0.50816    0.01546  32.873  < 2e-16 ***
Trans.Humed1     0.17123    0.03688   4.643 3.59e-06 ***
Inest1           6.05945    1.27347   4.758 2.05e-06 ***
---

Residual standard error: 1.233 on 2791 degrees of freedom
Multiple R-squared:  0.4972, Adjusted R-squared:  0.4965
F-statistic: 689.9 on 4 and 2791 DF,  p-value: < 2.2e-16
```

De esta manera se comprueba que, los valores de R^2 y R^2_{ajust} para este modelo, son mayores que para el modelo lineal inicial (el que no presentaba las variables predictoras con retardos), ambos

aproximadamente igual a 0.5. Este resultado significa que cerca del 50 % de la variabilidad de la precipitación, puede ser explicada por la inestabilidad atmosférica, del mismo día y del día anterior, así como por la precipitación y el transporte de humedad, ambos con retardo igual a 1. Además se destaca que, en este modelo, todas las variables incluidas son significativas para cualquier nivel de significancia. A continuación se interpretan los valores estimados de las variables independientes, teniendo en cuenta que en este primer caso la regresión es en media y no tiene en cuenta los cuantiles:

- El valor esperado de la precipitación cuando todas las variables predictoras son 0, es ~ 0.69 mm/día. Dicho valor es bajo, teniendo en cuenta que en este caso se considera que no hubo ningún contenido de humedad transportado desde el Golfo de México, que no precipitó el día anterior y que hay baja inestabilidad. Es decir, no existían condiciones atmosféricas en la región de interés para que precipitara, lo que se puede traducir a un ambiente de estabilidad.

- Una de las contribuciones mayores es cuando aumenta el valor de la inestabilidad atmosférica, ya que manteniendo constantes el resto de las variables predictoras, por cada unidad que aumenta la inestabilidad en esa región, se espera que aumente la precipitación aproximadamente 32 unidades. Esto demuestra la importancia de la existencia de inestabilidad en la atmósfera para que precipite.

- Un incremento en una unidad de la Inest1, significa un incremento de aproximadamente 6 unidades en la precipitación. Este incremento de la precipitación es menor, lo que puede estar relacionado con que la influencia de la inestabilidad del día anterior no es de forma directa.

- En menor medida, aunque también importante, es cuando hay un aumento de una unidad de Trans.Humed1 y Prec1, el valor de la precipitación se espera que aumente, 0.17 y 0.5 unidades, respectivamente. Estos valores bajos de precipitación demuestran que estas variables influyen en menor medida en la variable respuesta y confirman la necesidad de presencia de movimientos verticales en la atmosféra para que se genere inestabilidad.

Uno de los aspectos importantes es comprobar si al incorporar las variables predictoras con lag 1 se corrige la autocorrelación presente en los residuos del modelo. El gráfico ACF para este modelo se muestra en la Figura 4.3:

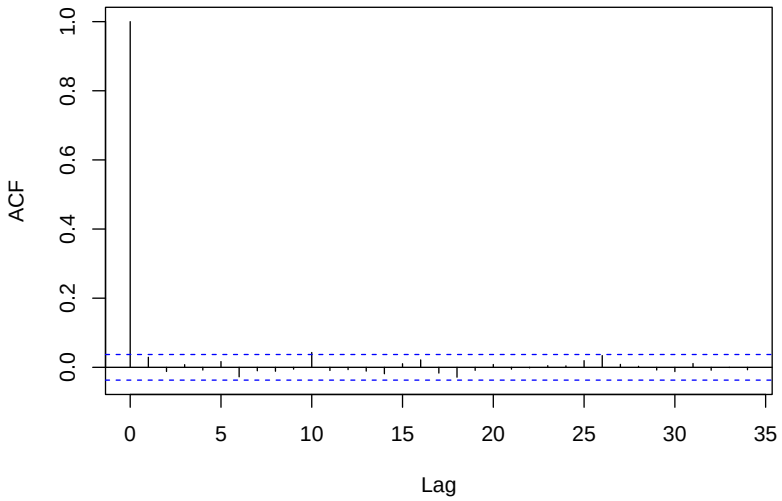

Figura 4.3: Autocorrelación del modelo de regresión lineal seleccionado.

Según los límites establecidos por las líneas de significancia, todas las barras se encuentran dentro de ellos, sobrepasando ligeramente en un retardo puntual. En comparación con el gráfico ACF presentado anteriormente en la Figura 4.1, se puede apreciar la corrección de la autocorrelación en los primeros

retardos. Esto demuestra que la incorporación de las variables predictoras con retardos adicionales solucionó el problema de la autocorrelación en los residuos del modelo. En este sentido, se comprueba que contamos con un modelo más adecuado que captura mayor cantidad de información de los datos analizados.

Además, se realizó la diagnosis del modelo, comprobando las hipótesis de normalidad de los residuos, homecedasticidad y la no autocorrelación, obteniéndose los resultados que se muestran en la Tabla 4.2.

Tabla 4.2: Diagnosis del modelo de regresión lineal.

Test	p-valor
Shapiro-Wilk	2.2e-16
Breusch-Pagan	2.2e-16
Durbin-Watson	0.1212

Estos resultados demuestran que para un nivel de significancia al 95 %, no se cumplen las hipótesis de normalidad ni de homocedasticidad, sin embargo si se cumple la de autocorrelación ya que el p-valor obtenido es mayor que 0.05. Dichos resultados se pueden corroborar gráficamente en la Figura 4.4.

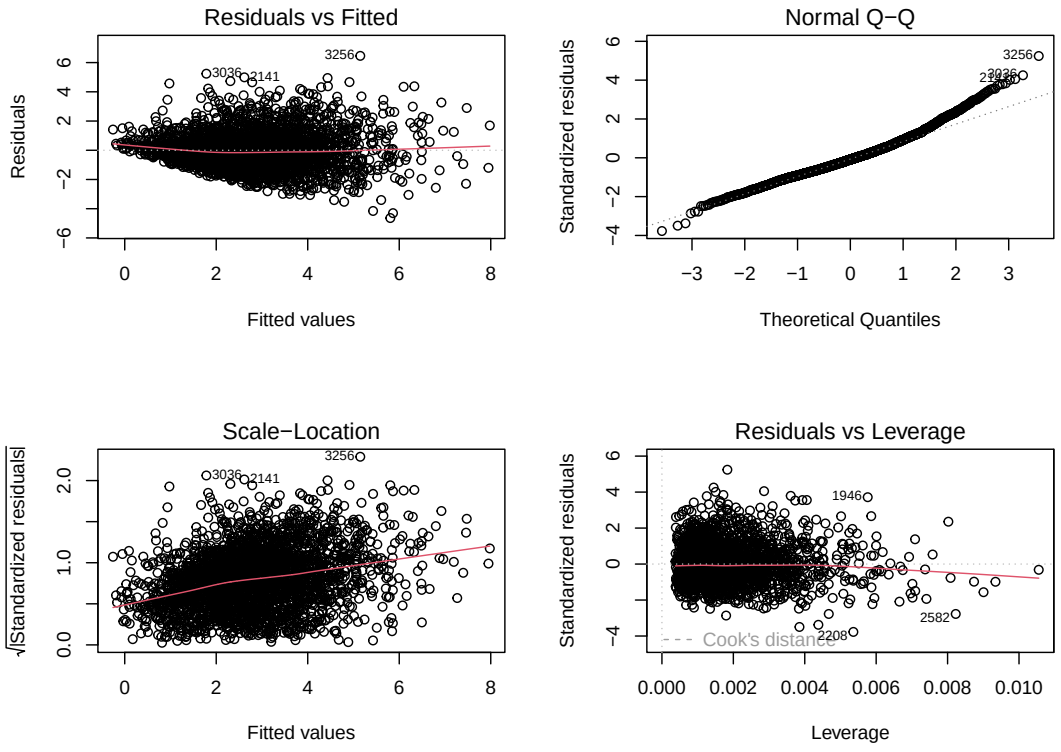

Figura 4.4: Diagnosis del modelo de regresión lineal.

El hecho de no cumplirse todas las hipótesis del modelo demuestra la complejidad de la relación entre

las variables meteorológicas estudiadas, ya que dicha relación no es lineal, sino que es más compleja. A pesar de esto, se decidió continuar con el análisis de los cuantiles para este modelo también, con el objetivo de comparar con el resto de los resultados que se irán obteniendo, teniendo en cuenta que los resultados no serán muy confiables al no poder validarse.

De esta manera se puede pasar a un análisis más profundo, para diferentes cuantiles con este modelo de regresión lineal. Para el cálculo de las nuevas predicciones se utilizó la muestra test mencionada anteriormente. Los cuantiles utilizados fueron: 0.25, 0.50, 0.75, 0.95 y 0.99, los cuales nos permitirán encontrar valores extremos que pueden estar presentes en los datos analizados. Se realizaron varias predicciones, las cuales representan condiciones atmosféricas diferentes, con el objetivo de analizar como las variaciones de las variables predictoras pueden influir en el comportamiento de la precipitación.

A continuación se presenta la Tabla 4.3 con los valores de los cuantiles para la precipitación, para 3 condiciones diferentes, considerando en primer lugar el valor mínimo para cada variable predictora, luego el valor medio y por último el valor máximo.

Tabla 4.3: Cuantiles de la precipitación en función de los valores mínimos, medios y máximos de las variables explicativas. Modelo lineal.

Valores	25 %	50 %	75 %	95 %	99 %
Mínimos	-0.580	-0.141	0.445	1.443	2.202
Medios	3.218	3.655	4.240	5.236	5.994
Máximos	10.833	11.274	11.864	12.868	13.631

Se puede apreciar como varía la precipitación según las diferentes condiciones meteorológicas (extremas, medias y mínimas) de las variables explicativas. Si se consideran los valores mínimos de las variables se obtienen valores muy bajos de precipitación, incluso, para los cuantiles del 25 y 50 %, los valores son negativos, cuestión que no tiene sentido cuando se está tratando con la variable precipitación, lo que se pudiera traducir a la ausencia de precipitación. Estos valores negativos, además de estar condiciondados por los valores mínimos de las variables predictoras, pueden relacionarse con errores propios del modelo lineal ajustado, unido además a no cumplirse con todas las hipótesis para su validación. Según estos resultados, el 99 % de los valores de precipitación serán iguales o inferiores a 2.2 mm/día, lo que significa que cuando hay condiciones de estabilidad en la región de interés y no hay suficiente humedad transportada desde la fuente del Golfo de México de días anteriores, entre otras condiciones, se espera que la precipitación asociada al GPLLJ, sea muy baja. Esto es un aspecto a tener en cuenta para evitar posibles eventos de sequía o problemas agrícolas en esta amplia zona de las Grandes Llanuras, por falta de precipitación.

Cuando se utilizan los valores medios de cada una de las variables predictoras, la precipitación aumenta en cada uno de los cuantiles analizados, presentando valores razonables y similares con los valores reales. Para este caso, el 99 % de los valores de precipitación, son inferiores o iguales a aproximadamente 6 mm/día, lo que significa que a pesar de que las condiciones no sean completamente favorables, puede llegar a precipitar, aunque no cantidades considerables. Por último, se utilizaron los valores máximos de las variables predictoras, aumentando considerablemente los valores de la precipitación en cada uno de los cuantiles, llegando incluso a alcanzar diferencias de aproximadamente 8 y 11 mm/día con respecto a cuando se consideraron los valores medios y mínimos, respectivamente, de las variables predictoras. Esto demuestra que la presencia de inestabilidad atmosférica en la región para ese día, así como, la cantidad de humedad transportada, la inestabilidad y la precipitación del día anterior, pueden influir en la precipitación registrada en la región de las Grandes Llanuras, asociado con la presencia del chorro de bajos niveles.

Estos cuantiles pueden proporcionar una herramienta para la planificación y la gestión de recursos hídricos, así como para evaluar los posibles riesgos que pueden existir, ya sea por problemas de sequías o por precipitaciones intensas. Se podría pensar que esta cantidad de precipitación no es significativamente alta, pero sigue siendo igual de importante su estudio debido a la amplia región agrícola que abarca y a las consecuencias que la variación de la precipitación puede traer en dicha región.

Análisis individual de cada variable

Para un análisis más específico y con el objetivo de ver cuánto puede influir la diferencia de las condiciones de las variables explicativas, se decidió variar solo una variable predictora, desde el valor mínimo hasta el valor máximo, manteniendo fijas, con su valor medio, el resto de las variables explicativas. Esto permitiría además, representar el comportamiento de la precipitación en función de cada una de esas variables predictoras, así como los cuantiles utilizados. Para el análisis se emplearon los mismos cuantiles anteriores: 0.25, 0.50, 0.75, 0.95 y 0.99. A continuación se mostrarán los resultados para cada una de las variables analizadas.

Inestabilidad atmosférica

Los valores de la precipitación, para cada uno de los cuantiles, para diferentes valores de inestabilidad atmosférica se muestran en la Tabla 4.4.

Tabla 4.4: Cuantiles de la precipitación en función de la inestabilidad atmosférica.

Inest (Pa/s)	25 %	50 %	75 %	95 %	99 %
-0.044	1.038	1.469	2.046	3.030	3.781
-0.030	1.572	2.003	2.578	3.560	4.309
-0.016	2.107	2.537	3.111	4.091	4.838
-0.002	2.642	3.071	3.645	4.624	5.370
0.0123	3.178	3.606	4.180	5.158	5.904
0.026	3.713	4.142	4.716	5.694	6.441
0.041	4.249	4.678	5.253	6.232	6.979
0.055	4.785	5.215	5.791	6.771	7.520
0.069	5.321	5.753	6.329	7.312	8.063
0.083	5.858	6.290	6.869	7.855	8.607

En la tabla se observa como al aumentar los valores de la inestabilidad atmosférica en la región de estudio, la precipitación tiende a aumentar en cada uno de los cuantiles. Esto demuestra una vez más que, la existencia de inestabilidad en la atmósfera es una condición necesaria para la ocurrencia de precipitaciones. Para una inestabilidad de -0.044 Pa/s, se obtiene que para el 25 % de los días, el valor de precipitación es muy bajo, a pesar de tener valores medios del resto de las variables predictoras. Para el valor máximo de inestabilidad, se obtuvo que, solo en el 1 % de los casos, la precipitación

estará por encima de aproximadamente 8.6 mm/día, sin embargo, si la inestabilidad fuera mínima, no superaría los 3.8 mm/día.

La representación gráfica para estos cuantiles de la precipitación en función de la inestabilidad atmosférica se puede obsevar en la Figura 4.5, donde el cuantil 0.5 se representa de color azul para diferencialo del resto.

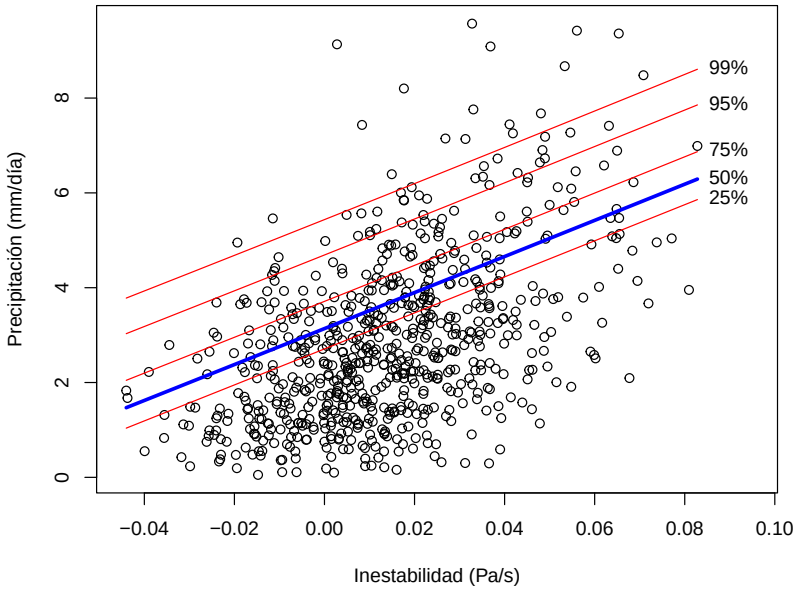

Figura 4.5: Regresión por cuantiles de la precipitación en función de la inestabilidad atmosférica.

De forma general, se ratifica lo obtenido en la tabla anterior. Como el modelo de regresión utilizado en este caso es un modelo lineal, las líneas de los cuantiles, no se ajustan del todo correcto a los datos, ya que no es lo suficientemente flexible para captar la relación que existe entre las dos variables y por tanto la distribución de los puntos. Además, con esta representación, no es posible ver la variabilidad de los valores de la precipitación.

Humedad transportada (lag 1)

Los resultados de los valores de la precipitación para cada uno de los cuantiles en función de la humedad transportada desde el Golfo de México se muestran en la Tabla 4.5.

En este caso el aumento de la humedad transportada desde el Golfo de México, produce también un incremento de la precipitación para cada uno de los cuantiles, pero mucho más ligero en comparación con la inestabilidad atmosférica. Sin embargo, para los valores más bajo del transporte de humedad, con los valores medios del resto de las variables, los cuantiles de la precipitación son mayores en relación con la inestabilidad. Mientras que para los valores máximos de humedad transportada, los valores de la precipitación son menores que los obtenidos cuando la inestabilidad es máxima. Estos resultados demuestran que puede haber gran contenido de humedad en la atmósfera, transportada desde una región fuente, pero es necesario la existencia de condiciones inestables en la región de interés para que se generen movimientos verticales, el vapor de agua ascienda y se condense para que se genere la precipitación.

La representación gráfica de la precipitación en función de la humedad transportada para los diferentes cuantiles se muestra en la Figura 4.6:

Tabla 4.5: Cuantiles de la precipitación en función de diferentes valores del transporte de humedad.

Trans.Humed1 (mm/día)	25 %	50 %	75 %	95 %	99 %
0	3.107	3.536	4.111	5.090	5.838
0.454	3.178	3.607	4.181	5.160	5.906
0.908	3.250	3.679	4.252	5.230	5.977
1.362	3.322	3.751	4.325	5.303	6.050
1.816	3.394	3.824	4.398	5.378	6.126
2.270	3.467	3.897	4.473	5.455	6.204
2.724	3.539	3.971	4.549	5.533	6.284
3.178	3.613	4.046	4.626	5.614	6.367
3.632	3.686	4.121	4.703	5.696	6.453
4.086	3.760	4.197	4.782	5.780	6.541

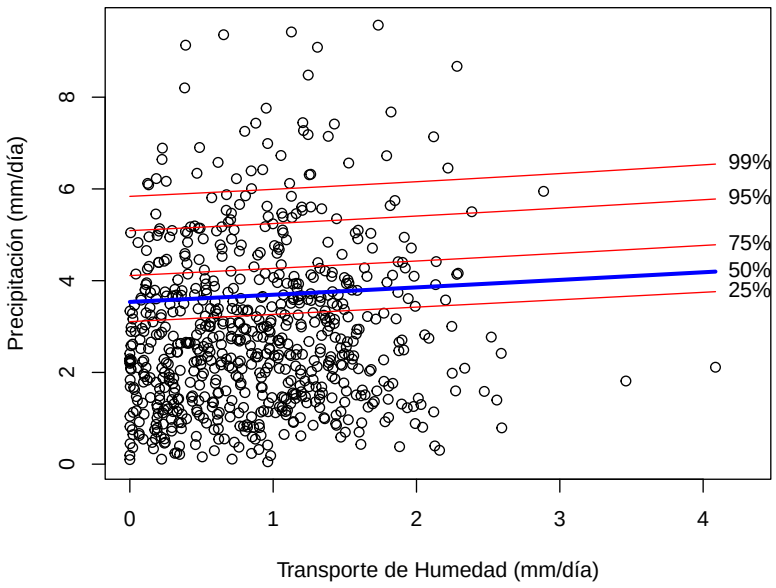

Figura 4.6: Regresión por cuantiles de la precipitación en función de la humedad transportada.

En este caso se puede observar, que las líneas que representan los cuantiles tampoco se adaptan a la distribución de los datos. La gran dispersión de los datos no es posible captarla con este modelo lineal. La inclinación de los cuantiles es muy ligera, a diferencia de la figura anterior con la inestabilidad, lo que significa que la precipitación tiende aumentar a medida que aumenta la humedad trasportada desde las regiones fuentes, pero en menor medida. Estos resultados corroboran los obtenidos en la tabla anterior.

Inestabilidad atmosférica (lag 1)

En la Tabla 4.6 se muestran los resultados de los cuantiles de la precipitación en función de la inestabilidad atmosférica, esta vez con retardo igual a 1.

Tabla 4.6: Cuantiles de la precipitación en función de diferentes valores de inestabilidad atmosférica con retardo igual a 1.

Inest1 (Pa/s)	25 %	50 %	75 %	95 %	99 %
-0.058	3.325	3.759	4.340	5.330	6.085
-0.040	3.305	3.737	4.315	5.300	6.051
-0.023	3.286	3.716	4.292	5.273	6.022
-0.006	3.267	3.697	4.271	5.250	5.997
0.011	3.249	3.678	4.252	5.230	5.976
0.029	3.232	3.661	4.235	5.213	5.960
0.046	3.214	3.644	4.219	5.199	5.947
0.064	3.198	3.629	4.206	5.189	5.939
0.081	3.182	3.615	4.194	5.182	5.935
0.098	3.166	3.602	4.184	5.178	5.936

Los valores de la precipitación para los diferentes cuantiles analizados, demuestran que la inestabilidad con lag 1 (inestabilidad del día anterior) en la región de interés, tiene menos influencia en la precipitación que la inestabilidad del mismo día. Esto tiene sentido, si se considera que esta última tiene una relación más directa e inmediata con la precipitación, ya que si el aire es lo suficientemente inestable y hay suficiente humedad, las corrientes ascendentes pueden favorecer la formación de nubes de tormenta que llegan a precipitar. Mientras que la inestabilidad del día anterior puede estar relacionada con la presencia de algún sistema meteorológico que presenta evento de precipitación y que perdure en el tiempo. En algunos casos, la inestabilidad puede no disiparse completamente de un día para otro, especialmente en situaciones donde sistemas de baja presión de lento movimiento, afectan un área determinada.

Además, teniendo en cuenta los resultados de la Tabla 4.6, los valores de precipitación, en cada uno de los cuantiles, son muy similares, con una diferencia de sólo aproximadamente 0.15 mm/día y con una ligera disminución a medida que aumenta la inestabilidad.

En la figura 4.7 se puede observar el comportamiento lineal de estos cuantiles, los cuales presentan menos pendiente que la inestabilidad del propio día y negativa. Esta representación confirma que los valores obtenidos anteriormente son muy similares y que no se ajustan correctamente a la distribución de la nube de puntos.

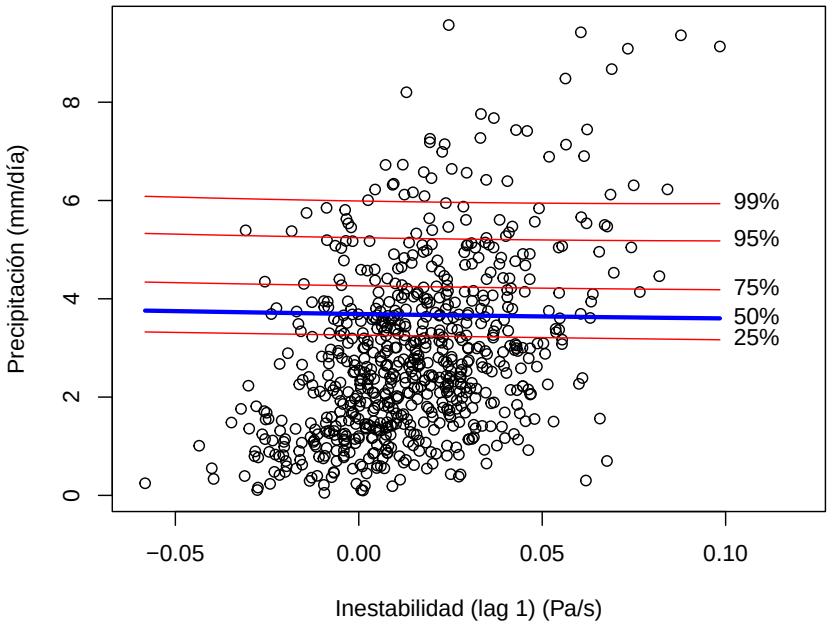

Figura 4.7: Regresión por cuantiles de la precipitación en función de la inestabilidad atmosférica con retardo igual a 1.

Precipitación (lag 1)

Por último, se presentan los resultados para cada uno de los cuantiles de la precipitación en función de la precipitación del día anterior, en la Tabla 4.7.

De forma general, se puede observar como al aumentar los valores de la precipitación con lag 1, aumenta también la precipitación en cada uno de los cuantiles. Con respecto, a los resultados obtenidos con las variables predictoras anteriores, en este caso, los valores de precipitación fueron mayores, sobre todo para los valores más altos de precipitación con lag 1 y para los cuantiles más altos también. Se destaca que para un valor de precipitación del día anterior, de 11.609 mm/día, con valores medios de inestabilidad atmosférica y transporte de humedad, se espera una cantidad de precipitación en la región de estudio, igual o inferior a aproximadamente 10 mm/día, en el 99 % de los días. Este valor de precipitación es el mayor de todos los cuantiles analizados en todas las variables predictoras. Esta relación entre la precipitación de ambos días está condicionado también con la presencia de inestabilidad atmosférica el día anterior y de sistemas que generen varios días de precipitación.

Dichos resultados se pueden corroborar en la representación de los cuantiles en la Figura 4.8.

Tabla 4.7: Cuantiles de la precipitación en función de diferentes valores de precipitación con retardo igual a 1.

Prec1 (mm/día)	25 %	50 %	75 %	95 %	99 %
0.062	1.909	2.339	2.914	3.894	4.642
1.345	2.531	2.960	3.534	4.513	5.260
2.628	3.154	3.583	4.157	5.135	5.881
3.911	3.777	4.206	4.780	5.758	6.505
5.194	4.400	4.830	5.405	6.384	7.132
6.477	5.024	5.455	6.031	7.013	7.762
7.760	5.648	6.080	6.658	7.643	8.395
9.043	6.273	6.706	7.287	8.276	9.030
10.326	6.898	7.334	7.917	8.911	9.669
11.609	7.523	7.961	8.548	9.548	10.311

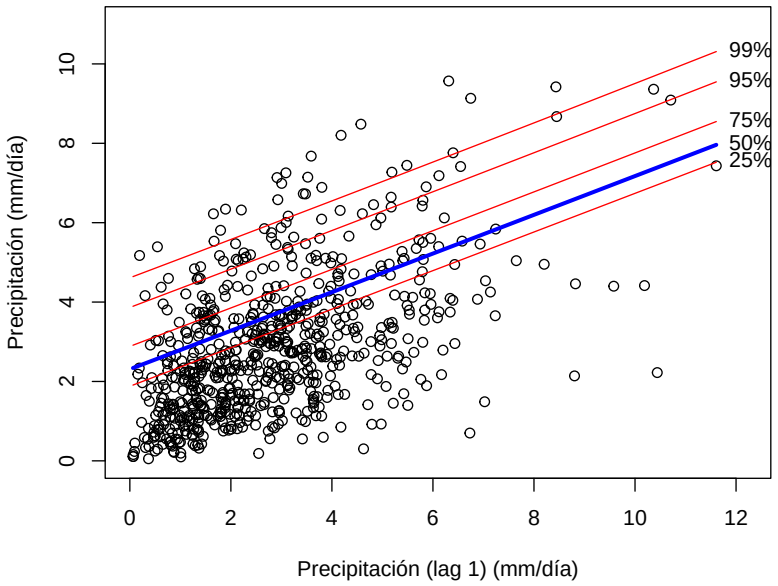

Figura 4.8: Regresión por cuantiles de la precipitación en función de la inestabilidad atmosférica con retardo igual a 1.

4.2. Función rq

Para continuar con el análisis de los cuantiles se decidió utilizar la función **rq**, mencionada en los capítulos anteriores. Esta función supone una especificación lineal del modelo de regresión cuantil y minimiza una suma ponderada de residuos absolutos que pueden formularse como un problema de programación lineal. Teniendo en cuenta que también es presentando como un problema lineal, se podrá comparar con los resultados obtenidos anteriormente.

Para continuar con el mismo procedimiento, se muestra la siguiente Tabla 4.8, con los valores de los cuantiles de la precipitación para diferentes condiciones de las variables predictoras.

Tabla 4.8: Cuantiles de la precipitación en función de los valores mínimos, medios y máximos de las variables explicativas.

Valores	25 %	50 %	75 %	95 %	99 %
Mínimos	-1.069	-1.047	-0.811	-0.082	0.624
Medios	1.962	2.697	3.532	4.984	6.262
Máximos	8.215	10.385	12.448	15.028	16.308

Un resultado interesante al utilizar esta función, fue que se obtuvieron valores negativos para todos los cuantiles de la precipitación cuando no estaban las condiciones meteorológicas creadas, excepto para el cuantil 0.99, obteniéndose un valor muy bajo, igual a 0.6 mm/día . Estos valores negativos no tienen sentido, ya que el valor mínimo de precipitación es cero, por lo que se puede traducir a que no precipite en aquellos días donde las condiciones no sean favorables. Por otro lado, el modelo utilizado pudo haber estimado valores que no tienen sentido físico, debido a errores propios del modelo, a la variabilidad física de los datos o a la extrapolación más allá del rango de los datos observados, como mismo ocurrió con el modelo lineal anterior. Sin embargo para los otros dos casos, los resultados sí son positivos. Con los valores máximos de las variables explicativas se obtienen las mayores precipitaciones, demostrando la necesidad de condiciones específicas para que precipite. En general los resultados son muy similares a los obtenidos con el modelo lineal, sin embargo en este caso para el 99 % de los días se espera una precipitación igual a aproximadamente 16 mm/día, valor superior que con el modelo lineal.

Para un análsis de la influencia de cada una de las variables predictoras en la precipitación, se realizó el mismo procedimiento anterior, mostrándose los resultados a continuación.

Inestabilidad atmosférica

En primer lugar se realizó el análisis con la variable inestabilidad atmosférica, obteniéndose los resultados, para cada uno de los cuantiles, que se muestran a continuación en la Tabla 4.9:

Con esta función se mantiene la conclusión de que a medida que aumenta la inestabilidad atmosférica aumenta también la precipitación, con valores similares de manera en general. Sin embargo, para una inestabilidad igual a 0.08 Pa/s, se obtuvo que para el 99 % de los datos, el valor de la precipitación es de aproximadamente 10 mm/día. Este valor es mayor que el obtenido para el modelo lineal con esta misma variable. Mientras que, en el 25 % de los datos, para una inestabilidad baja, se obtuvo un valor de precipitación de 0.684 mm/día, valor muy pequeño, incluso menor que el anterior con el modelo lineal. Se obtuvo una diferencia de aproximadamente 5 mm/día de precipitación entre el cuantil 0.5 y el 0.99 para la mayor inestabilidad.

Para una mejor visualización de dichos cuantiles, se presenta la Figura 4.9.

Con esta función **rq** las líneas de cada uno de los cuantiles siguen mejor la distribución de la nube de puntos. Además, ya estas dejan de ser paralelas, representando de mejor manera la variabilidad de

Tabla 4.9: Cuantiles de la precipitación en función de diferentes valores de inestabilidad atmosférica.

Inest (Pa/s)	25 %	50 %	75 %	95 %	99 %
-0.044	0.684	1.010	1.323	2.221	3.348
-0.030	0.994	1.419	1.858	2.890	4.054
-0.016	1.304	1.827	2.394	3.560	4.760
-0.002	1.613	2.236	2.929	4.229	5.466
0.012	1.923	2.645	3.464	4.898	6.172
0.026	2.233	3.053	3.999	5.568	6.878
0.041	2.542	3.462	4.535	6.237	7.584
0.05463	2.852	3.870	5.070	6.907	8.290
0.069	3.162	4.279	5.605	7.576	8.996
0.083	3.471	4.688	6.140	8.245	9.701

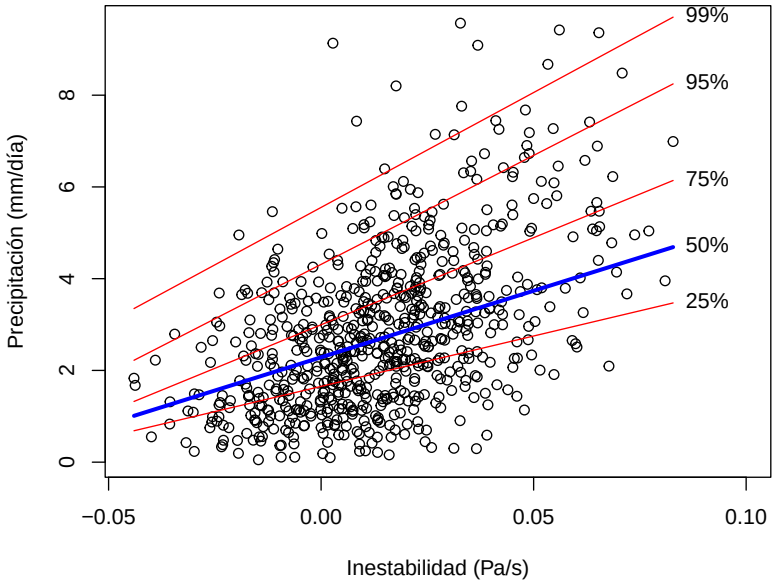

Figura 4.9: Regresión por cuantiles de la precipitación en función de la inestabilidad atmosférica.

los datos. Dicha variabilidad es menor para valores mas bajos de inestabilidad, mientras que a medida que esta variable va aumentado, la variabilidad también aumenta.

Humedad transportada (lag 1)

Seguidamente se aplicó la misma metodología pero esta vez con la variable relacionada con el transporte de humedad de días anteriores, proveniente de la fuente de humedad oceánica del Golfo de México. Los resultados para los diferentes cuantiles se muestran en la siguiente Tabla 4.10.

Tabla 4.10: Cuantiles de la precipitación en función de diferentes valores de humedad transportada (lag 1).

Trans.Humed1 (mm/día)	25 %	50 %	75 %	95 %	99 %
0	1.846	2.528	3.375	4.898	6.248
0.454	1.906	2.615	3.456	4.942	6.255
0.908	1.966	2.702	3.537	4.986	6.262
1.362	2.025	2.789	3.618	5.030	6.269
1.816	2.085	2.876	3.699	5.074	6.277
2.270	2.145	2.962	3.780	5.119	6.284
2.724	2.205	3.049	3.861	5.163	6.291
3.178	2.264	3.136	3.942	5.207	6.298
3.632	2.324	3.223	4.023	5.251	6.305
4.086	2.384	3.310	4.104	5.295	6.313

Los valores obtenidos para cada uno de estos cuantiles son muy similares, aumentando en todos los casos muy ligeramente, sobre todo, para el cuantil 0.99, ya que en este caso, el valor es prácticamente el mismo, con una diferencia de sólo 0.065 entre cuando hay mínima y máxima humedad transportada. La precipitación para el 99 % de los datos es aproximadamente igual a 6 mm/día para el mayor valor de humedad transportada.

Este comportamiento se aprecia mejor en la siguiente representación gráfica (Figura 4.10), como para el cuantil 0.99 se mantiene prácticamente constante. Otro aspecto importante, es que en esta variable la variabilidad es muy similar, tanto para los valores más bajos como los más altos de humedad transportada. A diferencia del modelo anterior, esta representación permitió ver las diferencias que pueden existir en diferentes partes de la distribución de la precipitación, debido que para el cuantil más extremo el comportamiento no fue igual que el resto.

Inestabilidad atmosférica (lag 1)

Seguidamente se pasó al análisis de la variable inestabilidad atmosférica, esta vez con un retardo igual a 1. La Tabla 4.11 muestra los resultados para cada uno de los cuantiles analizados.

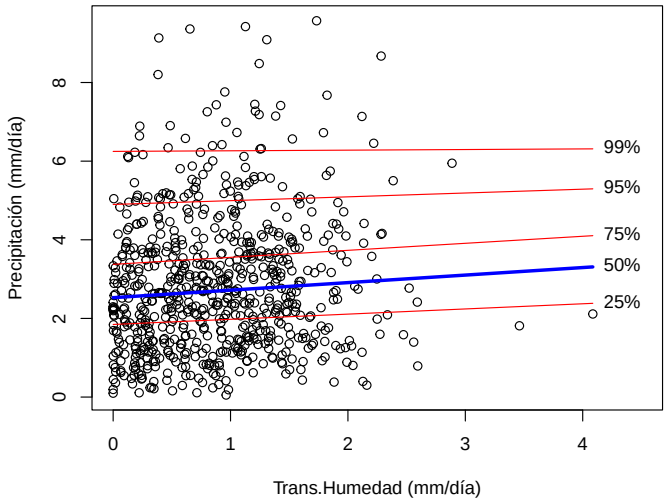

Figura 4.10: Regresión por cuantiles de la precipitación en función de la humedad transportada (lag 1).

Tabla 4.11: Cuantiles de la precipitación en función de diferentes valores de inestabilidad atmosférica (lag 1).

Inest1 (Pa/s)	25 %	50 %	75 %	95 %	99 %
-0.058	1.521	2.236	3.252	4.689	5.240
-0.041	1.626	2.346	3.319	4.759	5.483
-0.023	1.731	2.455	3.385	4.829	5.727
-0.006	1.836	2.565	3.452	4.899	5.970
0.011	1.941	2.674	3.519	4.969	6.213
0.029	2.046	2.784	3.585	5.040	6.456
0.046	2.151	2.894	3.652	5.110	6.699
0.064	2.256	3.003	3.719	5.180	6.942
0.081	2.361	3.113	3.785	5.250	7.185
0.098	2.467	3.222	3.852	5.320	7.428

En este caso hay diferencias con respecto a lo obtenido con la inestabilidad con lag 1 en el modelo lineal. La precipitación en cada uno de los cuantiles va aumentando a medida que aumenta dicha inestabilidad ligeramente. Los mayores cambios se apreciaron en el cuantil 0.99, con una diferencia de aproximadamente 2 mm/día. Lo que ratifica lo obtenido anteriormente de la mayor influencia en la precipitación de la inestabilidad del propio día a la inestabilidad del día anterior. A pesar de ello, se puede ver que para el 99 % de los datos se espera una precipitación de aproximadamente 7 mm/día, para los valores mayores de esta inestabilidad con lag 1, manteniendo los valores medios del resto de las variables. Este valor resultó ser superior, que el obtenido con la función **lm**.

En la Figura 4.11 se muestra la representación de estos cuantiles, la cual permite observar con mayor claridad, como para el cuantil 0.99, se observa un mayor cambio en la pendiente de la línea recta, en comparación con el resto de los cuantiles. La variabilidad de los datos, según esta función es un poco mayor para valores mayores de inestabilidad con lag 1.

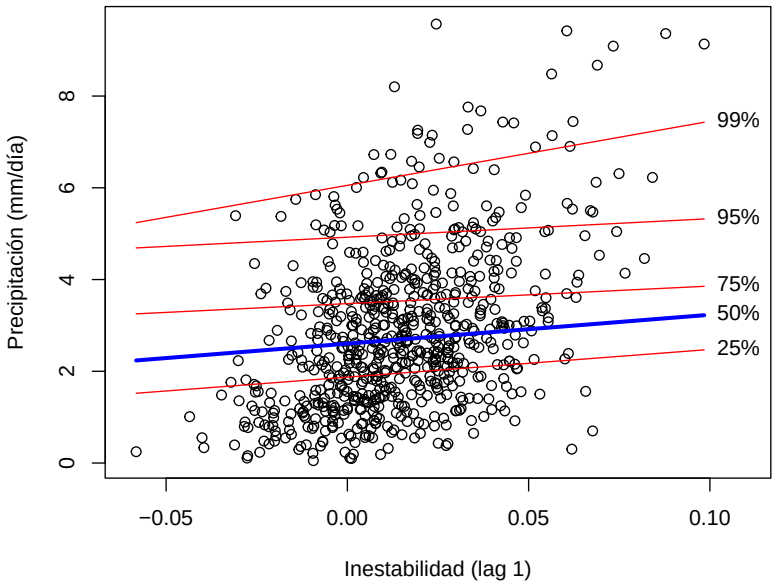

Figura 4.11: Regresión por cuantiles de la precipitación en función de la inestabilidad atmosférica (lag 1).

Precipitación (lag 1)

Por último con la variable precipitación con lag 1, se muestran los resultados para cada cuantil en la Tabla 4.12.

Se puede observar que en este caso los resultados son similares de manera general en comparación con el **lm**. Mientras mayor sea la precipitación registrada el día anterior, mayor se espera que sea la del día analizado, en cada uno de los cuantiles. Específicamente para el 99 % de los datos, para un valor casi nulo de precipitación con lag 1, se espera una precipitación de aproximadamente 4.6 mm/día. Mientras que para este mismo valor de cuantil, la precipitación esperada es prácticamente igual a la del día anterior, para su valor máximo, igual a 12 mm/día aproximadamente.

Para una mejor interpretación se muetra la representación de los cuantiles en la Figura 4.12.

En este caso los cuantiles se ajustan mejor a los distribución de los datos. Muestran una menor

Tabla 4.12: Cuantiles de la precipitación en función de diferentes valores de la precipitación (lag 1).

Prec1 (mm/día)	25 %	50 %	75 %	95 %	99 %
0.062	0.766	1.268	1.835	3.061	4.573
1.345	1.323	1.933	2.626	3.957	5.360
2.628	1.880	2.599	3.416	4.852	6.146
3.91118	2.437	3.264	4.206	5.747	6.932
5.19408	2.995	3.929	4.996	6.642	7.719
6.477	3.552	4.594	5.787	7.537	8.505
7.760	4.109	5.259	6.577	8.432	9.291
9.043	4.666	5.924	7.367	9.327	10.078
10.326	5.223	6.589	8.157	10.223	10.864
11.609	5.780	7.254	8.948	11.118	11.650

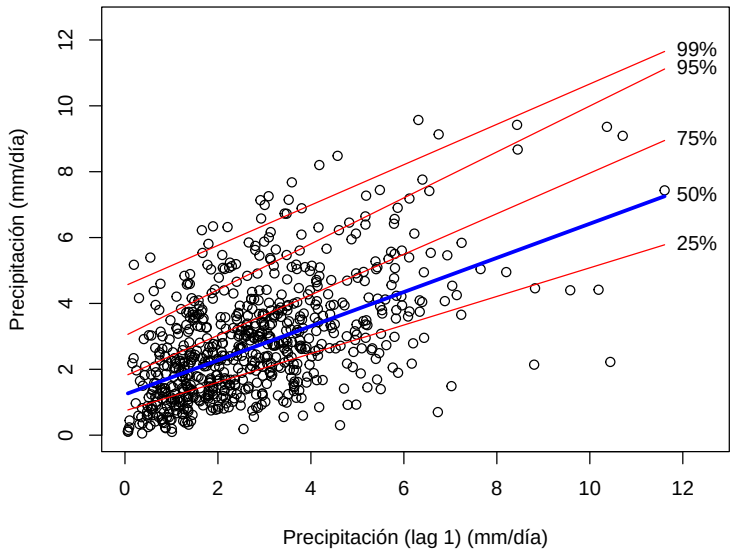

Figura 4.12: Regresión por cuantiles de la precipitación en función de la precipitación (lag 1).

variabilidad los datos para valores bajos de precipitación con lag 1, aumentando la variabilidad a medida que aumenta dicha precipitación. La mayoría de los datos se concentran por debajo de 6 mm/día de precipitación con lag 1, con una mayor dispersión por encima de este valor.

4.2.1. Regresión cuantil flexible

De forma general, con esta nueva función se obtuvieron mejores resultados, sobre todo en la representación de los cuantiles de la precipitación en esta región de las Grandes Llanuras Americanas, con respecto al modelo lineal con la función **lm**. Es por ello que, en busca de mejores soluciones se continuó el uso de esta función para otros modelos de regresión. Y es que, con la propia función **rq** en R es posible ajustar modelos de regresión más flexibles, incluyendo bases de B-splines para las variables predictoras. En este caso se utilizaron diferentes grados de libertad (4 y 5) para comparar y analizar las diferencias que pueden provocar en la representación de los cuantiles de la precipitación.

Se realizará el mismo procedimiento para cada una de las variables independientes, para analizar la posible influencia de cada una de ellas. Se mostrará solamente el análisis gráfico ya que el objetivo es comparar las diferencias en el uso de los grados de libertad y ver qué tan bien pueden representar la distribución de la nube de puntos.

Inestabilidad atmosférica

Se comenzó el análisis con la variable inestabilidad atmosférica, cuyos resultados se muestran en la Figura 4.13.

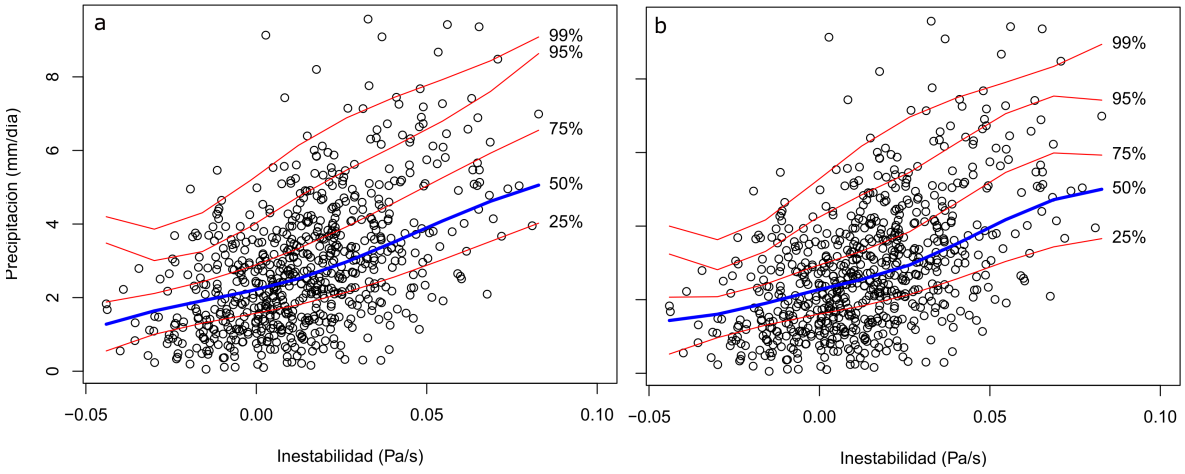

Figura 4.13: Regresión por cuantiles de la precipitación en función de la inestabilidad atmosférica. a) 4 grados de libertad y, b) 5 grados de libertad.

Se puede observar como la representación de los cuantiles en el modelo va cambiando al ajustar diferentes grados de libertad para las variables predictoras. Con el aumento de dichos grados, las líneas de los cuantiles son más flexibles, capturando mejor la variabilidad y los extremos de la precipitación. Sin embargo, hay que tener en cuenta que, tratando de representar todos los valores extremos presentes en la nube de puntos, podría obtenerse un sobreajuste. Cada una de las representaciones tiene sus particularidades, explicadas a continuación:

- De manera general, la representación de los cuantiles en ambos casos es muy similar, tendiendo al aumento de la precipitación a medida que aumenta la inestabilidad atmosférica. Representan mejor la distribución de la nube de puntos de la precipitación que los modelos anteriores, demostrando la flexibilidad que se obtiene al incluir bases B-Splines en las variables predictoras.

- Se observa una mayor variabilidad en los datos para los valores más altos de inestabilidad atmosférica.

- En el caso de considerar 4 grados de libertad, para los cuantiles 0.25, 0.50 y 0.75 siempre hay un aumento en la variable respuesta. Sin embargo para los cuantiles de precipitación más altos, 0.95 y 0.99, los cuales representan las precipitaciones más extremas, hay una ligera disminución entre los -0.05 y -0.04 Pa/s aproximadamente, para luego aumentar con el aumento de la inestabilidad. Pero para el cuantil 0.99, para los valores más altos de esta variable, la precipitación vuelve a tener una ligera disminución.

- En el caso de considerar 5 grados de libertad, el cuantil 0.75 presenta mayor similitud esta vez, con los cuantiles mayores que con los menores. Se mantiene el aumento en los cuantiles 0.25 y 0.50 desde los valores más bajos de inestabilidad. Luego para los cuantiles 0.75, 0.95 y 0.99, dicho aumento comienza en torno los -0.04 Pa/s. Mientras quue para los mayores valores de inestabilidad, ocurre una ligera disminución de la precipitación en los cuantiles 0.75 y 0.95.

Humedad transportada (lag 1)

Posteriormente se hizo el mismo procedimiento para la variable humedad transportada con retardo igual a 1, cuya representación aparece en la Figura 4.14.

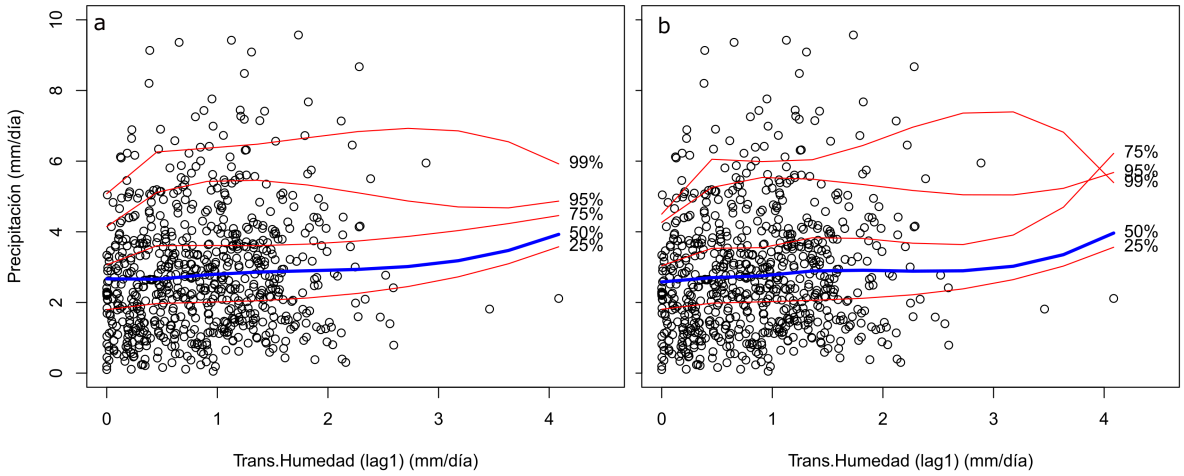

Figura 4.14: Regresión por cuantiles de la precipitación en función de la humedad transportada (lag 1). a) 4 grados de libertad y b) 5 grados de libertad.

Para esta variable las diferencias son menos significativas para los cuantiles 0.25 y 0.50, tendiendo a aumentar ligeramente, sobre todo para los mayores valores de humedad transportada. Mientras que para los cuantiles 0.75, 0.95 y 0.99 se observaron algunas diferencias en el comportamiento de la precipitación en función del transporte de humedad.

- En el caso de utilizar 4 grados de libertad, el comportamiento de la precipitación en el cuantil 0.75 es prácticamente constante hasta los 3 mm/día de humedad transportada aproximadamente, para luego presentar un ligero aumento. Dicho aumento con 5 grados de libertad es mucho más fuerte, llegando incluso a cruzar los cuantiles 0.95 y 0.99.

- En el cuantil 0.95, las diferencias son menos notables, destacándose una pequeña disminución de la precipitación a partir de 3 mm/día de humedad transportada, al utilizar 4 grados de libertad.

- Con respecto al cuantil 0.99, al considerar 5 grados de libertad, ocurre una disminución de la precipitación para los valores más altos de humedad, mientras que para 4 grados de libertad la disminución es menos notable.

- Estos resultados demuestran que para la ocurrencia de precipitaciones más altas o extremas, no es suficiente tener humedad transportada desde la región fuente, sino que deben existir otras condiciones

para que se genere la precipitación. Como por ejemplo, la inestabilidad atmosférica, que en este caso presenta un valor medio.

- Por último, para 5 grados de libertad se observa que los cuantiles más altos se cruzan para los mayores valores de humedad. En teoría esto no debiera suceder así ya que por definición los valores en los cuantiles más extremos siempre deben ser mayores que los cuantiles menores, sin embargo, esto pudiera estar relacionado con la propia inestabilidad en el método de estimación, sobre todo para los valores más extremos de las variables.

Inestabilidad atmosférica (lag 1)

La próxima variable analizada es la inestabilidad atmosférica con lag 1, cuyos resultados se muestran en la Figura 4.15.

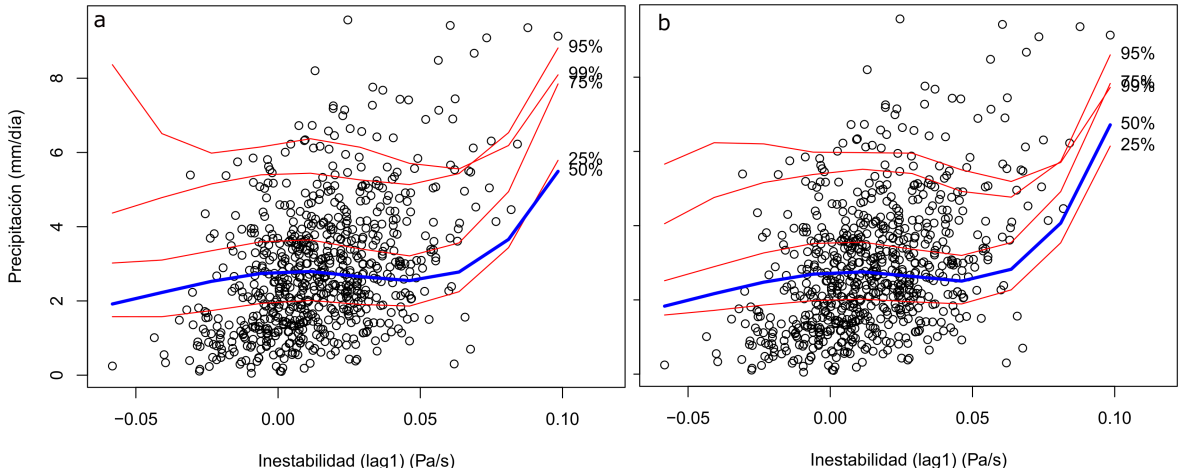

Figura 4.15: Regresión por cuantiles de la precipitación en función de la inestabilidad atmosférica (lag 1). a) 4 grados de libertad y b) 5 grados de libertad.

En general, para esta variable, el comportamiento de la precipitación para ambos grados de libertad utilizados es muy similar. La mayor diferencia se observa en el cuantil 0.99, ya que de forma general, para los valores más bajos de inestabilidad se observa un ligero aumento en la precipitación. Sin embargo, para 4 grados de libertad para estos primeros valores, la precipitación presenta una disminución. Luego, en todos los cuantiles, se mantiene sin mucho cambio hasta aproximadamente después de los 0.05 Pa/s, que nuevamente continúan con un aumento mucho mayor. Se observa una mayor variabilidad en los valores más bajos de la variable predictora. Para los valores más altos de inestabilidad, las líneas de los cuantiles también se cruzan lo que puede estar relacionado con las mismas causas ya mencionadas. Este resultado corrobora que la inestabilidad del día anterior influye en menos medida con la precipitación de un día y que los mayores valores de inestabilidad están muy relacionados con las precipitaciones más fuertes.

Precipitación (lag 1)

Por último se analizó la precipitación con lag 1 con los diferentes grados de libertad, cuyos resultados se muestran en la Figura 4.16.

Para esta variable tampoco se observaron grandes diferencias con los diferentes grados de libertad utilizados. Para los cuantiles 0.25, 0.50, 0.75 y 0.95, el comportamiento de la precipitación en los 4 casos es muy similar, con tendencia al aumento. Destacando la disminución en el cuantil 0.5, para 5 grados de libertad, llegando a ser menor que para el 0.25, cuestión que ya se había comentado anteriormente. Mientras que con el cuantil 0.99 hay ligeras diferencias entre ambos casos. Para 4 grados de libertad,

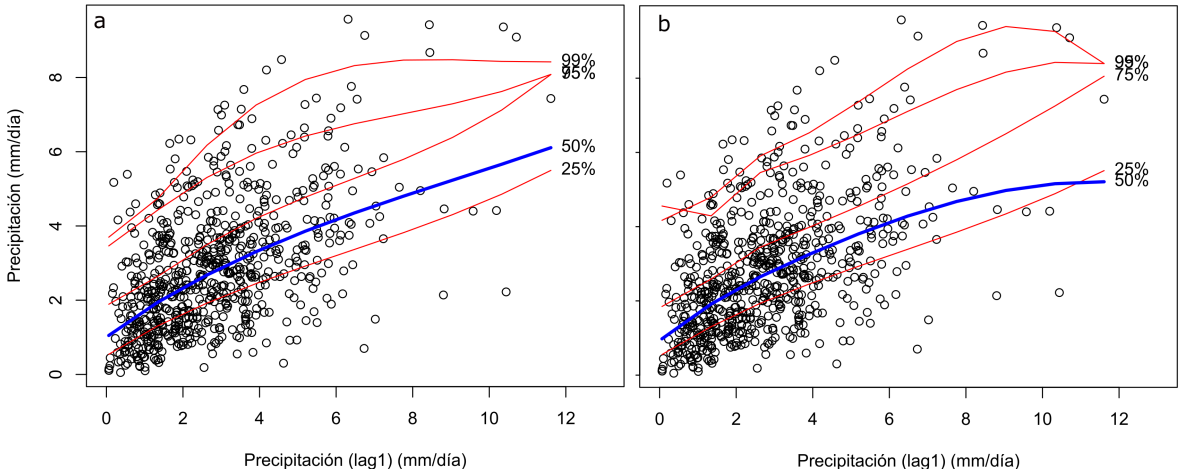

Figura 4.16: Regresión por cuantiles de la precipitación en función de la precipitación (lag 1).

el aumento de la precipitación ocurre hasta aproximadamente 6 mm/día de precipitación registrada el día anterior, para luego permanecer prácticamente constante. Mientras que con 5 grados de libertad, el aumento es hasta los 9 mm/día para posteriormente disminuir para valores mayores.

Con todos los resultados mostrados se puede concluir que al representar las variables predictoras con bases de B-splines, hace que estemos en presencia de modelos más flexibles, que son capaces de representar con mejor claridad la variabilidad de los datos, sin embargo han presentado ciertas dificultades. Los resultados observados con 4 y 5 grados de libertad, fueron muy similares de manera general. Esto demuestra que la relación entre las variables predictoras y la variable respuesta no es del todo lineal, por lo que se hace necesario la utilización de modelos de regresión más complejos y flexibles que sean capaces de representar con más claridad la relación entre estas variables.

4.3. Modelos aditivos generalizados

Para realizar un análisis aún más complejo y lograr representar relaciones más flexibles entre las variables predictoras y la respuesta se utilizaron los modelos de regresión aditivos generalizados. Teniendo en cuenta que es otro modelo, que tiene en cuenta otras características, se aplicó el mismo procedimiento que para el modelo lineal. Es decir, primeramente se ajustó el modelo para las variables originales sin ningún retardo, quedando de la siguiente manera:

```
modelo <- gam(Prec~s(Inest)+s(Trans.Humed)+s(TCW))
```

Al realizar el summary para este modelo GAM se obtuvieron los siguientes resultados:

```
Family: gaussian
Link function: identity

Formula:
Prec ~ s(Inest) + s(Trans.Humed) + s(TCW)

Parametric coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  2.81377    0.02584   108.9   <2e-16 ***
---
```

```
Approximate significance of smooth terms:
                edf Ref.df       F  p-value
s(Inest)      2.895  3.705 236.460  < 2e-16 ***
s(Trans.Humed) 3.865  4.849   7.370 1.95e-06 ***
s(TCW)        5.914  7.095   5.057 1.18e-05 ***
---

R-sq.(adj) =  0.226   Deviance explained = 22.9%
GCV = 2.3438  Scale est. = 2.3346    n = 3496
```

Se puede ver que el R^2_{ajust} es bajo, como ocurrió con el primer modelo de regresión lineal. Este es capaz de explicar aproximadamente el 23 % de la variabilidad de la precipitación en función de estas variables predictoras. A pesar de ser un modelo más flexible, no mejora la capacidad de representar la relación entre las variables analizadas. En este caso las 3 variables incluidas son significativas para cualquier nivel de significancia. En la diagnosis del modelo, tampoco se cumplieron los criterios de validación.

Destacar que se comprobaron diferentes combinaciones teniendo en cuenta la interacción entre las variables predictoras y el resultado era prácticamente el mismo. De ahí que, al considerar la interacción entre las variables, el modelo se convierte más complejo, pues tampoco se consideró una buena solución.

Por otro lado, se analizó también la autocorrelación de los residuos del modelo GAM planteado anteriormente, donde el gráfico ACF se muestra en la Figura 4.17.

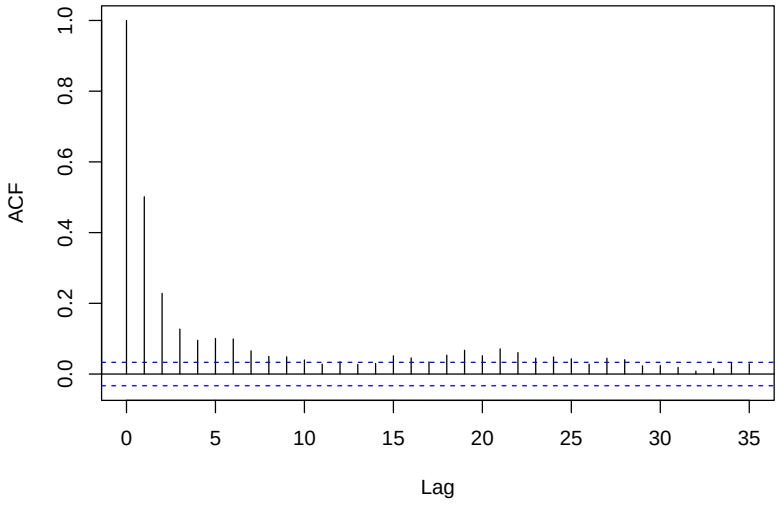

Figura 4.17: Autocorrelación del modelo GAM original.

Como mismo ocurría con el modelo lineal, para el modelo GAM, con los variables originales, sin incluir retardos, se aprecia una fuerte autocorrelación de los residuos, sobre todo para los primeros lags. Es por ello que resulta necesario, incluir como nuevas variables explicativas, los retardos igual a 1 de las variables originales, ya que hay que tener en cuenta la influencia que pueden tener las condiciones meteorológicas del día anterior en la precipitación. Con esto se espera corregir la autocorrelación de los residuos que hay en los primeros retardos, como mismo ocurrió en el modelo lineal.

Selección de variables

Al igual que en el caso anterior, se pasa de tener 3 variables predictoras a tener 7, por lo que resulta necesario aplicar nuevamente el criterio de selección de variables para este modelo GAM. Siguiendo el mismo procedimiento anterior, se muestra en la Tabla 4.13, las variables seleccionadas utilizando el criterio BIC para los primeros 10 mejores modelos.

Tabla 4.13: Mejores modelos según el criterio de selección de variables BIC.

Orden	Variables	Inest	Trans.Humed	TCW	Inest1	Trans.Humed1	TCW1	Prec1	BIC
1	**4**	✓			✓	✓		✓	**9155.420**
2	3	✓			✓			✓	9157.842
3	4	✓	✓		✓			✓	9159.444
4	**2**	✓						✓	**9160.843**
5	3	✓	✓					✓	9165.309
6	3	✓				✓		✓	9165.716
7	5	✓	✓		✓	✓		✓	9168.009
8	4	✓			✓		✓	✓	9172.422
9	3	✓					✓	✓	9172.562
10	4	✓	✓			✓		✓	9173.742

De forma general, estos modelos aditivos incorporan menos variables para realizar el ajuste en comparación con el modelo lineal. Sin embargo si fue similar en ambos modelos que se incluyeran en todos los casos las variables Inest y Prec1. Luego se destaca Inest1, que se incluyó en 5 de los 10 modelos presentados. Mientras que Trans.Humed y Trans.Humed1, se incorporaron en 4 modelos. Por último, TCW y TCW1 se incluyen en 2 y ningún modelo respectivamente.

El mejor modelo, seguún este criterio, resultó ser el que presentó un valor de BIC igual a 9155.420, incluyendo las variables predictoras: Inest, Inest1, Trans.Humed1 y Prec1. La incorporación de estas variables coincidió también con el mejor modelo lineal, lo que demuestra la importancia de estas variables en el estudio de la precipitación.

El cuarto mejor modelo, con un BIC igual a 9160.843, incluyó dos variables menos, Inest1 y Trans.Humed1. Teniendo en cuenta que de estos 10 mejores modelos, el mejor no es el que menos variables tiene, se realizó una comparación entre los dos modelos, a partir de varios criterios, siguiendo con la misma metodología llevada a cabo anteriormente. Primeramente se analizó la diferencia de los valores de BIC para estos dos modelos GAM, la cual es de 5.423, más baja, con respecto a los modelos lineales. Sin embargo, según los criterios planteados en el Capítulo 3, la diferencia sigue siendo positiva, considerando como mejor modelo el que incluye las 4 variables predictoras. La diferencia entre los valores de R^2_{ajust} y desviación explicada entre ambos modelos, no es muy considerable, siendo ligeramente superiores para el primero modelo, como se muestra en la Tabla 4.14.

Por otro lado, se compararon las predicciones de ambos modelos con respecto a los valores reales, para analizar si existían diferencias significativas entre ellas, cuyos resultados se pueden apreciar en la

Tabla 4.14: Comparación de los modelos GAM.

Modelos	R^2 ajustado	Desviación explicada
1	0.513	51.6 %
4	0.501	50.3 %

Figura 4.18, en diferentes períodos para tener una idea del comportamiento en general de las predicciones (inicio, mediados y final).

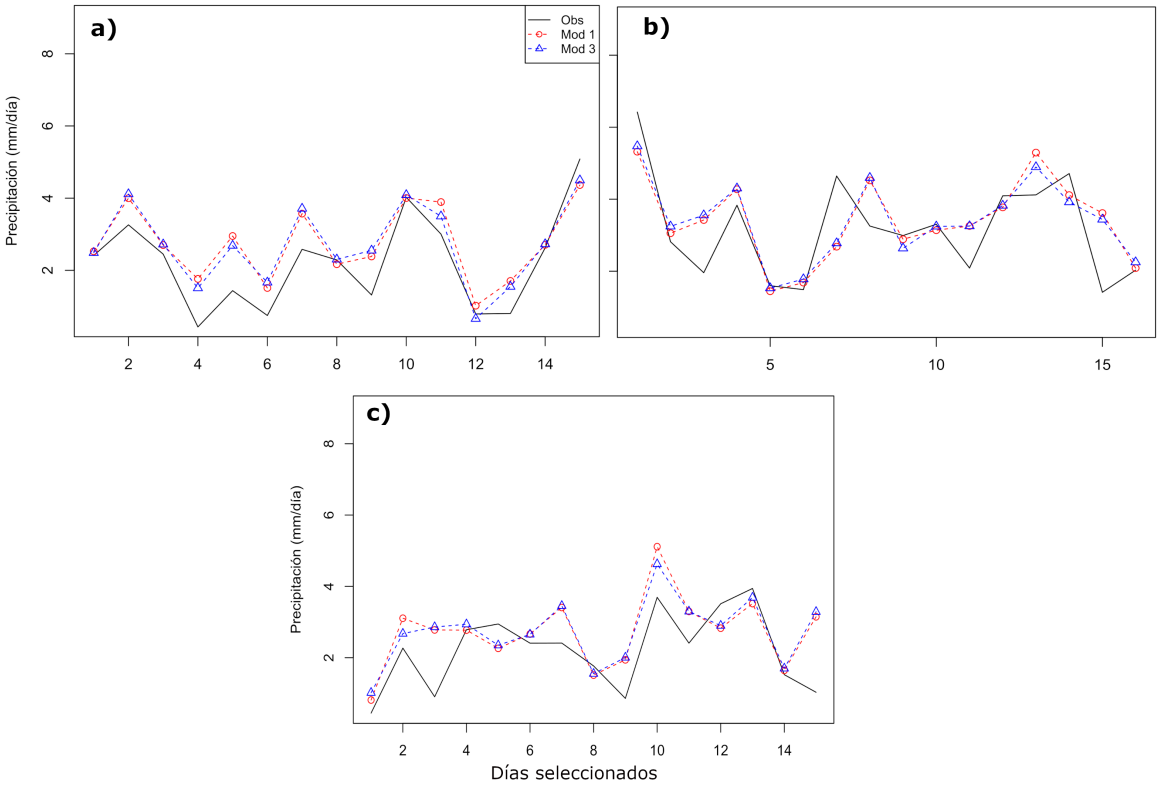

Figura 4.18: Comparación entre los valores observados (línea continua negra) y las predicciones realizadas por ambos modelos GAM. Las líneas discontinuas roja y azul son las predicciones de los modelos 1 y 4, respectivamente. a) inicios, b) mediados y c)final.

De forma general, ambos modelos siguen el patrón de la precipitación, sin observar grandes diferencias entre las predicciones de cada uno. A pesar de que las predicciones son ligeramente mejores que con el modelo lineal, hay momentos en que estos modelos GAM, no son capaces de predecir correctamente los valores de la precipitación, sobreestimando en algunos de los casos.

Con todos las comparaciones realizadas, se decide utilizar el modelo GAM con el menor valor de BIC, que cuenta con 4 variables predictoras, que a pesar de ser un poco más complejo que el modelo que incluye 2 variables predictoras, según la diferencia de BIC, es el mejor modelo para representar la relación entre las variables respuesta y predictoras, ya que en el resto de los aspectos tenidos en cuenta en las comparaciones, no se apreciaron diferencias significativas entre ambos modelos. Además coincide

con las variables incorporadas en los modelos lineales y así se pueden realizar mejor las comparaciones entre ellos.

4.3.1. Regresión cuantil con modelos GAM

El modelo GAM, teniendo en cuenta la selección de variables explicada en la sección anterior, quedaría planteado como se muestra a continuación:

```
modelo <- gam(y~s(Inest)+s(Ines1)+s(Trans.Humed1)+s(Prec1), data=train)
```

Con el summary realizado para este modelo se puede apreciar que el valor del R^2_{ajust}, igual a 0.513, no es tan alto, como se esperaba, pero si ligeramente mejor, en comparación con el resto de modelos analizados anteriormente. Además, el modelo explica cerca del 52 % de la variabilidad de la precipitación en función de las variables independientes, proporcionando una medida de qué tan bien el modelo se ajusta a los datos. Si bien este valor no es muy alto, teniendo en cuenta la complejidad de trabajar con datos climáticos y meteorológicos, se considera un buen resultado. A pesar de ello, hay que tener en cuenta para análisis futuros que, casi la mitad de la variabilidad de la variable respuesta no es capaz de explicarse. Por otra parte, todas las variables incluidas son estadísticamente significativas para cualquier nivel de significancia dado. Cada término de suavizado en el modelo tiene un número asociado de grados de libertad efectivos, que muestran la complejidad del ajuste del modelo en cada una de las variables.

```
Family: gaussian
Link function: identity

Formula:
P_resp ~ s(Inest)+s(Ines1)+s(Trans.Humed1)+s(Prec1), data=train

Parametric coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  2.80059    0.02294   122.1   <2e-16 ***
---

Approximate significance of smooth terms:
                        edf Ref.df       F  p-value
s(Inest)              2.797  3.590 227.906  < 2e-16 ***
s(Inest1)             5.719  6.951   7.369  < 2e-16 ***
s(Trans.Humed1)       3.390  4.276   5.097 0.000336 ***
s(Prec1)              6.550  7.703 149.553  < 2e-16 ***

---

R-sq.(adj) =  0.513   Deviance explained = 51.6%
GCV =  1.482  Scale est. = 1.4716    n = 2796
```

En cuanto a la interpretación de los resultados, se pueden extraer las siguientes conclusiones:

- Cuando todas las variables predictoras son 0, el valor esperado de la precipitación es aproximadamente 2.8 mm/día, con un error estándar muy bajo de 0.02.

- Los grados de libertad para la variable predictora Inest son 2.797, lo que indica una relación no lineal relativamente compleja entre dicha variable y la precipitación.

- Las variables predictoras Inest1 y Prec1, presentaron grados de libertad iguales a 5.719 y 6.550, respectivamente, sugiriendo una complejidad aún mayor en la relación entre estas variables y la precipitación.

- La variable Trans.Humed1 presenta 3.390 grados de libertad, indicando igualmente una relación compleja entre esta variable y la precipitación.

Dichos resultados se pueden corroborar en la representación de los efectos parciales estimados para cada predictor en la Figura 4.19.

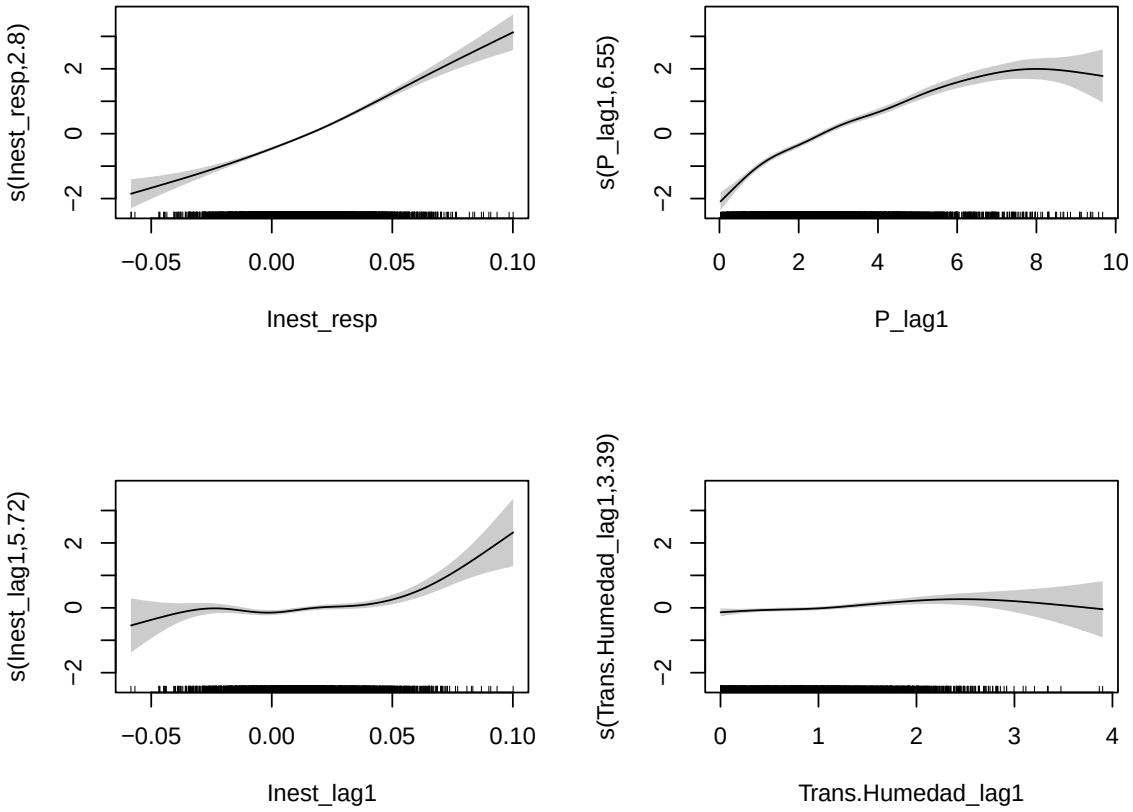

Figura 4.19: Estimaciones de los efectos parciales de las variables predictoras.

Se realizó, además, una diagnosis descriptiva y gráfica del modelo, la cual se muestra en la Figura 4.20. Esta diagnosis no arrojó tampoco los mejores resultados, ya que no se observa normalidad en los gráficos de la izquierda. Además, se observa un patrón en el gráfico de los residuos en la esquina superior derecha a medida que aumenta el predictor lineal y no se ajustan a una recta los valores ajustados y la respuesta, en el gráfico inferior derecho.

Una cuestión importante a destacar, es si la incorporación de estas variables predictoras fue suficiente para corregir la autocorrelación en los primeros retardos. Para ello, se realizó un gráfico ACF, que se muestra en la Figura 4.21, donde se puede ver que se corrigió la fuerte autocorrelación que había en los primeros retardos. Prácticamente todas las barras que representan cada uno de los lag se encuentran dentro de las líneas de significancia, exceptuando algún retardo puntual.

En resumen, este modelo GAM, a pesar de no cumplir todas las condiciones necesarias para la validación del mismo, teniendo en cuenta la complejidad de los datos, se puede decir que, muestra un mejor ajuste con relaciones no lineales significativas entre las variables predictoras y la respuesta. La significatividad de todas las variables y el porcentaje de desviación explicada sugieren que el modelo

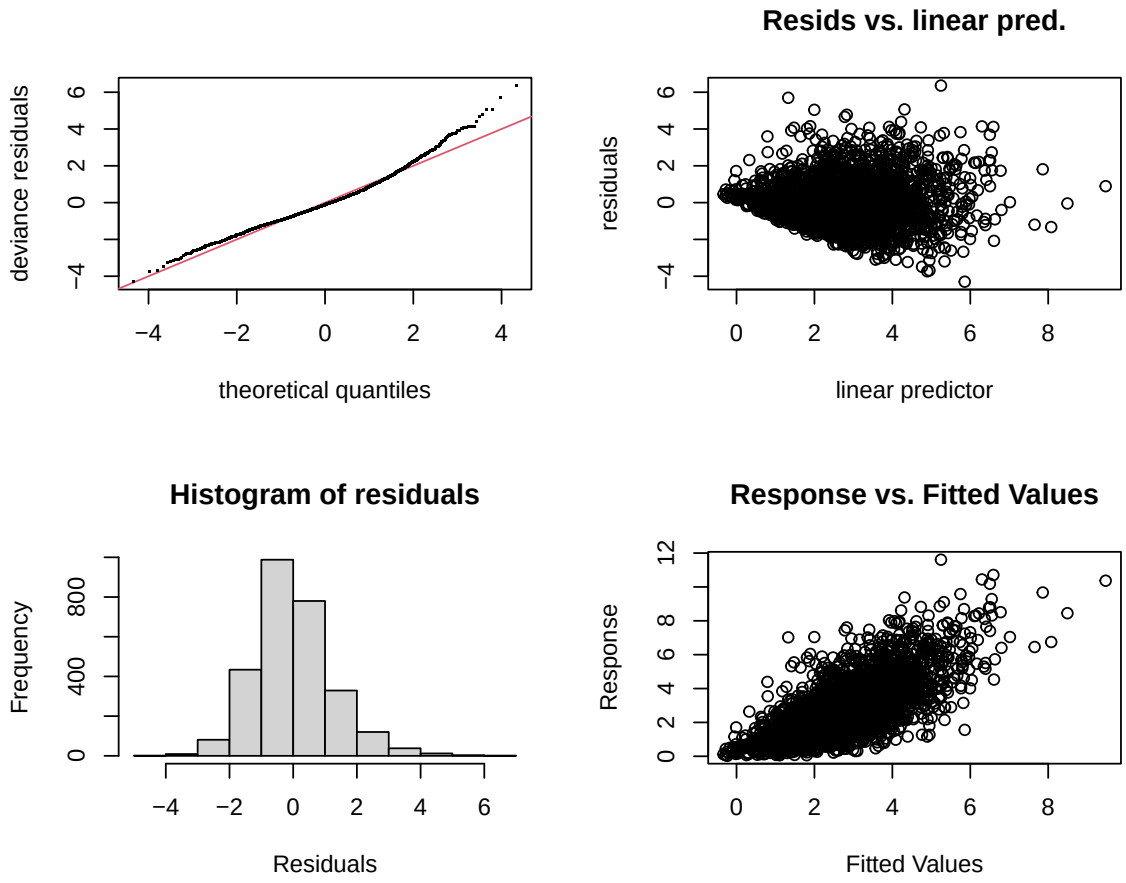

Figura 4.20: Gráfica de diagnóstico del modelo GAM seleccionado.

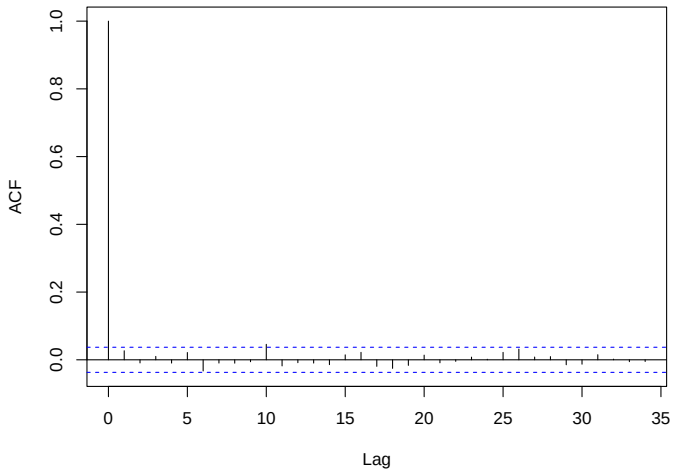

Figura 4.21: Autocorrelación del modelo GAM seleccionado.

puede ser adecuado brindando una representación de las relaciones que se establecen entre las variables analizadas. Sin embargo, el no poder validarse, es una cuestión a tener en cuenta a la hora de extraer conclusiones de los resultados.

Después de haber realizado todo este análisis con el modelo GAM, se decide continuar con el estudio y se considera que puede ser utilizado en modelos de regresión más complejos.

4.3.2. Función QGAM

La regresión cuantil en los modelos GAM se puede aplicar utilizando la función **qgam**, para ello se aplicó el mismo procedimiento mostrado anteriormente. Los primeros resultados para las diferentes condiciones de las variables predictoras se muestran en la Tabla 4.15.

Tabla 4.15: Cuantiles de la precipitación en función de los valores mínimos, medios y máximos de las variables explicativas. Modelo qgam.

Valores	25 %	50 %	75 %	95 %	99 %
Mínimos	-1.154	-0.992	-0.628	0.146	1.66
Medios	1.912	2.706	3.504	5.026	6.455
Máximos	7.897	9.647	11.727	12.046	12.569

Con estos valores se confirman los resultados obtenidos anteriormente, cómo al cambiar las condiciones meteorológicas la precipitación cambia también. Se obtuvieron igualmente valores negativos para los primeros cuantiles, lo cual se puede deber a las propios errores mencionados en los modelos anteriores. Para los valores máximos de las variables predictoras, puede llegar a registrarse hasta aproximadamente 13 mm/día de precipitación en el 99 % de los casos.

Para continuar con la metodología empleada en los modelos anteriores, se realizará el análisis de las predicciones para cada una de las variables predictoras incorporadas en los modelos.

Inestabilidad atmosférica

Se comienza el análisis para la variable inestabilidad atmosférica, el resultado para cada uno de los cuantiles se muestra en la Tabla 4.16.

Los resultados son muy similares a los obtenidos con los modelos utilizados anteriormente. Se comprueba como con el aumento de la inestabilidad atmosférica, la precipitación en cada uno de los cuantiles va aumentando también, sobre todo para los 4 primeros cuantiles. En el caso del cuantil 0.99, hay unas ligeras diferencias, ya que para los valores más bajos de inestabilidad se observa primero una ligera disminución. Se pasa de tener en el 99 % de los datos una precipitación igual o inferior a 5.32 mm/día, a tener 4.36 mm/día, para luego ir aumentado. Para el valor máximo de inestabilidad atmosférica, la cantidad de precipitación en el 99 % de los datos, es similar a la obtenida utilizando la función **rq**, con un valor igual o menor a aproximadamente 10 mm/día. Estos resultados se pueden corroborar gráficamente en la Figura 4.22.

Se puede observar como de manera general, las líneas de los cuantiles de la precipitación siguen la dstribución de la nube puntos correctamente. Se destaca la diferencia en el cuantil 0.99 con respecto al resto, para los valores más abajos de inestabilidad atmosférica. La presencia de un dato de precipitación de aproximadamente 5 mm/día, para un valor bajo de inestabilidad, produce dicho cambio en el cuantil 0.99, el cual teniendo en cuenta la distribución del resto de los valores, se puede considerar atípico. Lo mismo sucedió con la función **rq** al considerar los grados de libertad de las bases de las variables predictoras.

Tabla 4.16: Cuantiles de la precipitación en función de diferentes valores de la inestabilidad atmosférica.

Inest (Pa/s)	25 %	50 %	75 %	95 %	99 %
-0.058	0.446	1.138	1.620	2.749	5.317
-0.041	0.838	1.488	1.975	3.051	4.525
-0.023	1.226	1.852	2.367	3.464	4.360
-0.006	1.571	2.232	2.843	4.148	5.196
0.012	1.899	2.672	3.452	5.035	6.181
0.030	2.330	3.245	4.213	5.981	7.176
0.047	2.912	3.954	5.065	6.852	7.793
0.065	3.517	4.701	5.951	7.596	8.314
0.083	4.125	5.394	6.825	8.343	9.114
0.100	4.730	6.044	7.683	9.103	10.044

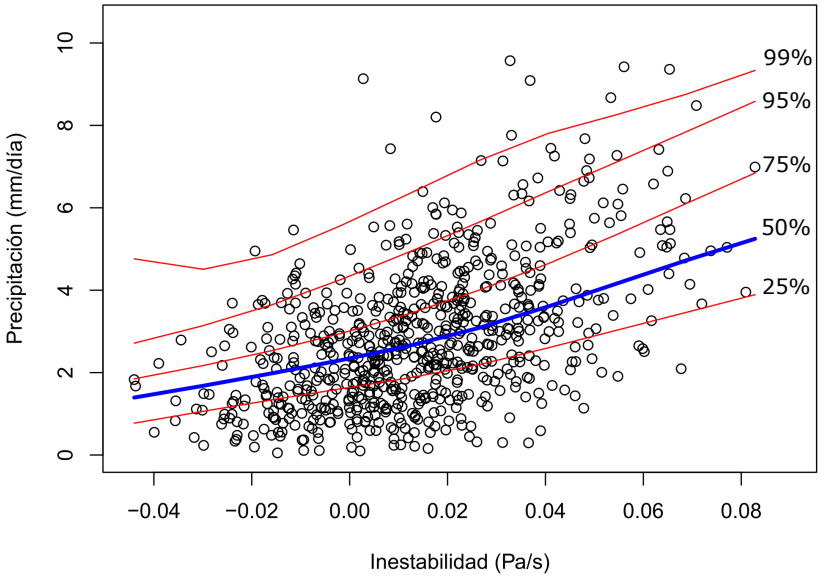

Figura 4.22: Cuantiles de la precipitación en función de la inestabilidad atmosférica.

Humedad transportada (lag 1)

La próxima variable predictora a analizar es la humedad transportada desde la región fuente del Golfo de México con retardo igual a 1. Los resultados para cada uno de los cuantiles de la precipitación, se muestran en la Tabla 4.17.

Tabla 4.17: Cuantiles de la precipitación en función de diferentes valores de la humedad transportada (lag 1).

Trans.Humed1 (mm/día)	25 %	50 %	75 %	95 %	99 %
0	1.900	2.569	3.365	4.707	5.633
0.454	1.893	2.639	3.437	5.021	6.779
0.908	1.917	2.710	3.508	5.019	6.396
1.362	2.037	2.781	3.579	4.999	6.203
1.816	2.136	2.853	3.651	5.145	6.523
2.270	2.198	2.924	3.722	5.130	6.537
2.724	2.202	2.995	3.794	4.954	6.086
3.178	2.154	3.066	3.865	4.701	5.552
3.632	2.076	3.136	3.936	4.416	5.125
4.086	1.988	3.207	4.008	4.121	4.758

En este caso, no se cumple para todos los cuantiles, que a medida que aumenta la humedad transportada desde el Golfo de México, aumenta también la precipitación. Este aumento, solo ocurre, y de manera muy leve en los cuantiles 0.50 y 0.75 de la precipitación. Dicho aumento es de solo aproximadamente 0.6 mm/día para ambos cuantiles. Para los últimos cuantiles, los valores de precipitación varían, sobre todo para el cuantil 0.99, disminuyendo para los valores más altos de humedad trasnportada. Para el cuantil 0.25 los valores de la precipitación son muy similares, en torno a los 2 mm/día, destacándose igualmente una ligera disminución para los valores más altos de la variable predictora.

Desde el punto de vista gráfico se puede observar la representación de los cuantiles de la precipitación en función del transporte de humedad en la Figura 4.23. Con la representación se corroboran los resultados obtenidos y, hasta el momento, se puede considerar un buen ajuste. De esta manera se puede observar que la precipitación se comporta de manera diferente, en distintas partes de la distribución. Por ejemplo para el cuantil 0.50 la precipitación tiende siempre a aumentar ligeramente, a diferencia del cuantil 0.99, que tiene variaciones. Esto demuestra una vez más la importancia de la regresión cuantílica porque permite conocer con más profundidad el comportamiento de la variable de interés.

Inestabilidad atmosférica (lag 1)

Se continuó con el análisis de la variable explicativa inestabilidad atmosférica con retardo igual a 1, mostrando los resultados en la Tabla 4.18 para cada cuantil de la precipitación.

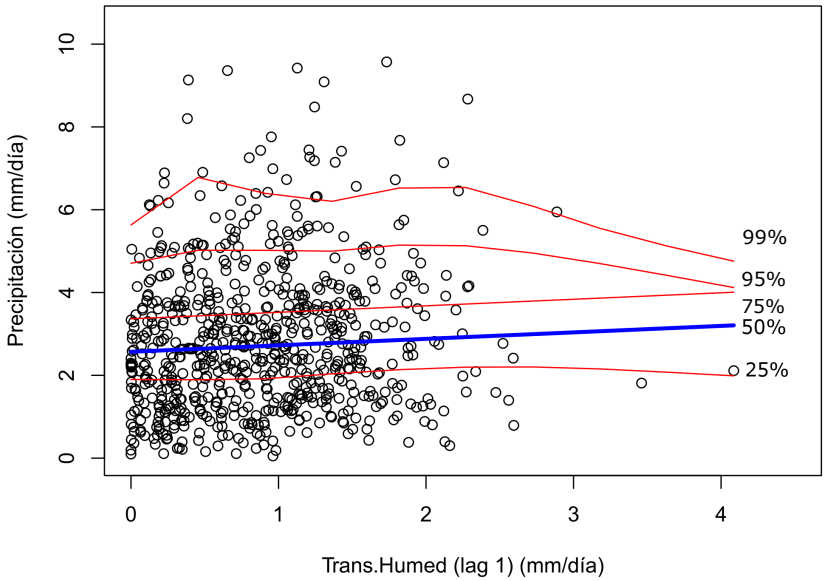

Figura 4.23: Cuantiles de la precipitación en función de la humedad transportada con la 1.

Tabla 4.18: Cuantiles de la precipitación en función de diferentes valores de la inestabilidad atmosférica (lag 1).

Inest1 (Pa/s)	25 %	50 %	75 %	95 %	99 %
-0.058	1.668	2.468	3.459	4.882	5.856
-0.041	1.824	2.625	3.534	4.955	5.999
-0.023	1.890	2.689	3.562	5.006	6.142
-0.006	1.809	2.605	3.501	5.008	6.284
0.011	1.882	2.675	3.493	5.020	6.427
0.029	2.024	2.809	3.576	5.066	6.569
0.046	2.245	2.980	3.746	5.243	6.712
0.064	2.656	3.344	4.032	5.548	6.855
0.081	3.303	3.912	4.395	5.899	6.997
0.098	4.070	4.571	4.785	6.266	7.140

En este caso, como mismo en los modelos anteriores, excepto en el modelo lineal, se mantiene la tendencia al aumento de la precipitación a medida que aumenta la inestabilidad del día anterior, en cada uno de los cuantiles. Los valores de forma general, son muy similares a los obtenidos con los modelos utilizados anteriormente. Lo que confirma que la inestabilidad del día anterior influye en la precipitación pero de una forma menos directa.

En la Figura 4.24 se muestra la rerpesentación de dichos cuantiles. Con ella se puede observar, como para los mayores valores de inestabilidad, disminuye ligeramente la variabilidad en la precipitación, sobre todo en los cuantiles 0.25, 0.50 y 0.75 por un lado, y por otro, los cuantiles 0.95 y 0.99. Para los valores más bajos de inestabilidad, el ajuste de los cuantiles no es del todo correcto, ya que muestra un poco más de variabilidad de la que hay a simple vista, como mismo ocurría con los modelos anteriores.

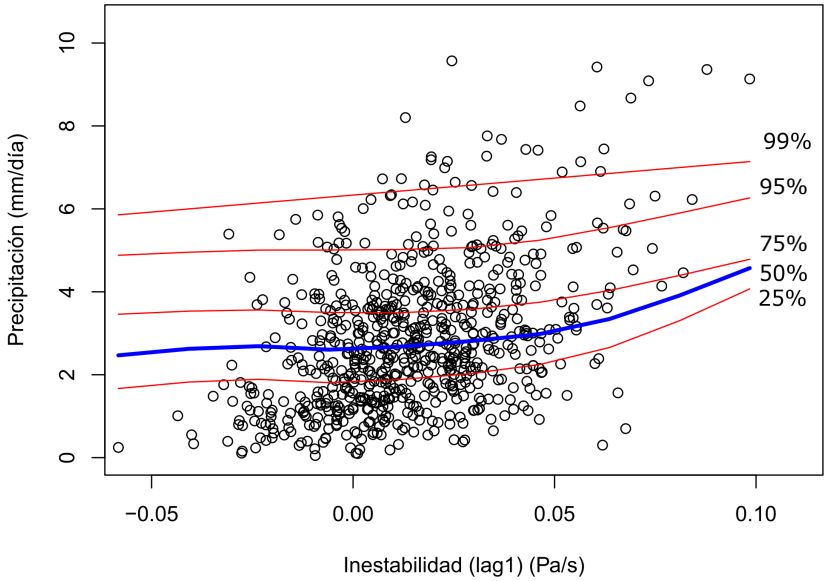

Figura 4.24: Cuantiles de la precipitación en función de la inestabilidad atmosférica con lag 1.

Precipitación (lag 1)

Por último se analizó la variable independiente que se corresponde con la precipitación con retardo igual a 1. En la Tabla 4.19 se presentan los resultados para cada uno de los cuantiles en este caso.

Estos valores confirman, como va aumentando la precipitación, en cada uno de los cuantiles, a medida que aumenta la precipitación registrada el día anterior. Comparado con los resultados anteriores, los valores obtenidos son ligeramente menores, donde el 99 % de los días presentan una precipitación igual o inferior a 10.7 mm/día. Un aspecto importante a destacar, que ya ocurrió también anteriormente, es que a pesar de que la variable respuesta va aumentado, para los valores más altos de precipitación con lag 1, siguen siendo menores que la precipitación del día anterior. Por ejemplo, solo en el 1 % de los días, la precipitación será mayor que 10.7 mm/día, cuya cantidad es menor que 11.6 mm/día, que se corresponde con la precipitación registrada el día anterior.

Desde el punto de vista gráfico, en la Figura 4.25, se muestran las líneas para cada uno de los cuantiles de la precipitación. Se corrobora que a medida que aumenta la precipitación con lag 1, la variable respuesta aumenta también. El ajuste es mejor que los obtenidos anteriormente con los otros modelos de regresión, lo que demuestra la importancia del uso de modelos más complejos y flexibles. Hay una mayor variabilidad de la precipitación para valores mayores de la variable independiente, observada en la separación de los cuantiles.

Tabla 4.19: Cuantiles de la precipitación en función de diferentes valores de la precipitación (lag 1).

Prec1(mm/día)	25 %	50 %	75 %	95 %	99 %
0.062	0.241	0.694	1.210	2.922	4.776
1.345	1.163	1.827	2.518	3.998	5.532
2.628	1.826	2.601	3.382	4.902	6.334
3.911	2.355	3.208	4.130	5.763	7.167
5.194	2.785	3.763	4.837	6.654	7.984
6.477	3.159	4.260	5.450	7.265	8.647
7.760	3.379	4.515	5.896	7.607	9.189
9.043	3.494	4.615	6.173	7.813	9.696
10.326	3.589	4.678	6.388	7.984	10.199
11.609	3.683	4.739	6.602	8.155	10.702

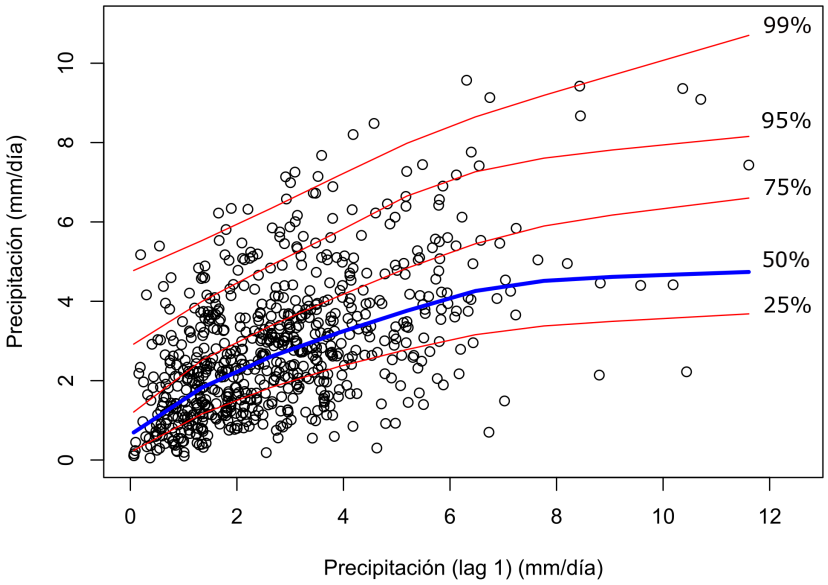

Figura 4.25: Cuantiles de la precipitación en función de la precipitación con lag 1.

4.4. Modelos de localización y escala

Para implementar un análisis más completo de la relación entre las variables predictoras y la precipitación se decidió utilizar los modelos de localización y escala, ya que ellos incluyen la varianza para la modelación. Ambos términos, el de localización y escala, en la modelación se representan como un modelo aditivo generalizado incluyendo las variables predictoras seleccionadas según el criterio BIC anterior. Dichos modelos quedarían planteados como se muestra a continuación:

```
modelo1 <- gam(P_resp~s(Inest)+s(Prec1)+s(Trans.Humed1)+s(Inest1), data=train)
muhat=predict(modelo1,type='response', data=train)
error2=(P_resp-muhat)**2

modelo2=gam(log(error2)~s(Inest)+s(Prec1)+s(Trans.Humed1)+s(Inest1), data=train)
sigma2=predict(modelo2,type='response', data=train)
sigma=sqrt(exp(sigma2))
error=(P_resp-muhat)/sigma
```

Con el objetivo de realizar predicciones en la muestra test se empleó el mismo procedimiento anterior para comparar los resultados. Los resultados obtenidos para los diferentes valores de cuantiles teniendo en cuenta condiciones atmosféricas distintas, se muestran en la Tabla 4.20.

Tabla 4.20: Cuantiles de la precipitación en función de los valores mínimos, medios y máximos de las variables explicativas. Modelo de localización y escala.

Valores	25 %	50 %	75 %	95 %	99 %
Mínimos	-3.447	-1.536	0.559	4.282	7.665
Medios	1.709	2.790	3.975	6.080	7.993
Máximos	4.731	8.359	15.712	22.214	35.385

Con los resultados obtenidos en la tabla anterior se pueden extraer las siguientes conclusiones:

- Para valores mínimos de las variables se obtuvieron valores negativos de precipitación en los cuantiles 0.25 y 0.50, lo que teniendo en cuenta que la variable analizada es precipitación no tienen sentido estos valores. Esto se puede traducir a la ausencia de precipitación asociada al GPLLJ, ya que no hay condiciones en la atmósfera suficientes para que precipite. Sin embargo para el 99 % de los días, en estas condiciones mínimas si puede llegar a precipitar hasta 7.665 mm/día.

- Por otra parte, al considerar los valores medios de las variables predictoras se obtuvieron valores de precipitación, para cada uno de los cuantiles, muy similares a los modelos anteriores, ligeramente superiores en los dos cuantiles más altos, destacando que en el 99 % de los días, la precipitación es igual o inferior a aproximadamente 8 mm/día.

- Resulta importante destacar, los altos valores de precipitación en todos los cuantiles, cuando se consideraron la inestabilidad, el transporte de humedad y la precipitación del día anterior máximos. En el 99 % de los días se obtuvo un valor de precipitación igual o inferior a 35 mm/día, resultado que se considera elevado, teniendo en cuenta los resultados obtenidos anteriormente con el resto de los modelos analizados.

De forma general, si hay presencia de valores altos de inestabilidad atmosférica, humedad transportada desde el Golfo de México y ha precipitado el día anterior, en la región de las Grandes Llanuras Americanas, se podría esperar valores cerca de los 35 mm/día. Puede ser que aún este valor no se considere significativo, pero hay que recordar que estos resultados se corresponden con la precipitación

asociada al chorro de los bajos niveles en esta región. Además, es una aproximación de esta región a un punto, en el cual se registró que este mecanismo alcanza su máximo en los meses de verano. Aún así este es el mayor valor de precipitación obtenido, que teniendo en cuenta los valores medios en estas fechas, es una cantidad superior. Esto pudiera indicar la importancia de conocer el comportamiento de estas variables en este período ya que se pudieran registrar, si están las condiciones necesarias, valores de precipitación superior a los habituales. También es importante destacar que esa cantidad es la registrada en un día, pero que puede caer toda esa cantidad en solo 1 hora y considerarse una lluvia muy fuerte.

Para un análisis más detallado de la influencia de cada una de las variables se continuó con el mismo procedimiento.

Inestabilidad atmosférica

Primeramente se realizó para la inestabilidad atmosférica para el mismo día, obteniéndose los resultados mostrados en la Tabla 4.21. Además se presenta en la Figura 4.26, el comportamiento de cada uno de los cuantiles, corroborando los resultados obtenidos numéricamente.

Tabla 4.21: Cuantiles de la precipitación en función de diferentes valores de la inestabilidad atmosférica.

Inest (Pa/s)	25 %	50 %	75 %	95 %	99 %
-0.044	0.144	1.292	2.551	4.788	6.821
-0.030	0.487	1.610	2.842	5.030	7.018
-0.016	0.863	1.954	3.151	5.277	7.209
-0.002	1.261	2.328	3.498	5.576	7.464
0.012	1.660	2.735	3.914	6.009	7.913
0.026	2.049	3.193	4.448	6.677	8.702
0.041	2.434	3.712	5.113	7.602	9.863
0.055	2.774	4.250	5.867	8.742	11.353
0.069	3.059	4.782	6.670	10.026	13.075
0.083	3.272	5.295	7.513	11.454	15.034

La precipitación aumenta, con el incremento de la inestabilidad atmosférica en esa región para ese mismo día, cuando la precipitación, la humedad transportada y la inestabilidad del día anterior, se mantienen constantes con su valor medio. Esto sugiere que días con mayor inestabilidad atmosférica tienden a estar asociados con mayores precipitaciones. La mayoría de los valores de precipitación se encuentran por debajo del cuantil del 99 %, exceptuando dos que están por encima. Este cuantil indica que en el 99 % de los días con estas condiciones la precipitación puede ser menor o igual a 15 mm/día aproximadamente, cuyo valor es superior a los obtenidos con las funciones anteriores. Los valores que están por encima del cuantil del 99 % pueden ser considerados extremos para este fenómeno específico que se está analizando y bajo estas condiciones. Si bien es cierto que estos valores no son elevados,

es un aspecto a tener en cuenta a la hora de planificar y entender las condiciones bajo las cuales se podría esperar una precipitación significativa.

Además, desde el punto de vista gráfico se puede apreciar que las líneas de cuantiles no son paralelas entre sí, indicando como varía la dispersión de la precipitación a medida que aumenta la inestabilidad. Para valores más pequeños de inestabilidad la diferencian entre los cuantiles es ligeramente menor, mientras que para los mayores valores estos están más separados.

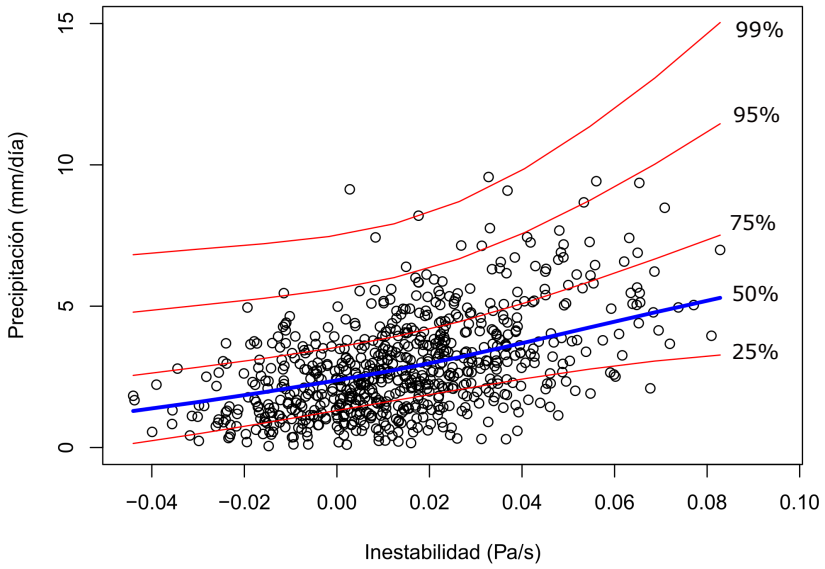

Figura 4.26: Cuantiles de la precipitación en función de la Inestabilidad atmosférica.

Humedad transportada (lag 1)

Los resultados de los cuantiles de la precipitación en función de la variación de la humedad trasnportada desde la región fuente del Golfo de México, se muestran en la Tabla 4.22.

Teniendo en cuenta estos resultados el aumento de la precipitación es muy poco a medida que aumenta la cantidad de humedad transportada, siendo ligeramente más notables en los últimos dos cuantiles. Con respecto a los valores obtenidos en los modelos anteriores, sobre todo para el valor más extremo, en este caso el resultado es superior. Y es que se obtiene que para el 99 % de los días, bajo estas condiciones, la precipitación es igual o inferior a casi 8.5 mm/día.

Desde el punto de vista gráfico, Figura 4.27 se puede apreciar un comportamiento similar a los resultados obtenidos con los modelos anteriores. La diferencia con respecto al modelo lineal, es que en este caso, las líneas que representan a cada uno de los cuantiles, se ajustan mejor a la distribución de los datos, ya que no son del todo paralelas. A pesar de ser un mejor modelo, se mantiene la conclusión de que el aumento de la humedad transportada, con presencia de valores medios de inestabilidad atmosférica y precipitación del día anterior, no influye considerablemente en el aumento de la precipitación para ese día. Esto demuestra que un gran contenido de humedad en la atmósfera no es suficiente por sí solo para que ocurra la precipitación, ya que se necesitan otras condiciones como la presencia de núcles de condensación y de movimientos verticales ascendentes. En este sentido al contar con valores de inestabilidad medios, tanto para ese mismo día como para el día anterior, pueden no ser suficientes para que la humedad se condense y precipite.

Por otro lado, para este caso de la humedad, la variabilidad entre los cuantiles de la precipitación es bastante similar, ligeramente superior para los mayores valores de humedad. Se observan igualmente valores muy alejados del cuantil 0.50, pir encima del cuantil 0.99, los cuales pueden ser considerados como valores extremos teniendo en cuenta estas condiciones.

Tabla 4.22: Cuantiles de la precipitación en función de diferentes valores de la humedad transportada (lag 1).

Trans.Humed1 (mm/día)	25 %	50 %	75 %	95 %	99 %
0	1.636	2.686	3.838	5.884	7.744
0.454	1.690	2.756	3.925	6.001	7.888
0.908	1.711	2.793	3.979	6.086	8.001
1.362	1.790	2.888	4.092	6.230	8.174
1.816	1.890	3.004	4.225	6.395	8.367
2.270	1.949	3.080	4.319	6.522	8.523
2.724	1.923	3.071	4.329	6.564	8.595
3.178	1.820	2.985	4.261	6.529	8.590
3.632	1.676	2.858	4.153	6.455	8.547
4.086	1.519	2.718	4.033	6.369	8.492

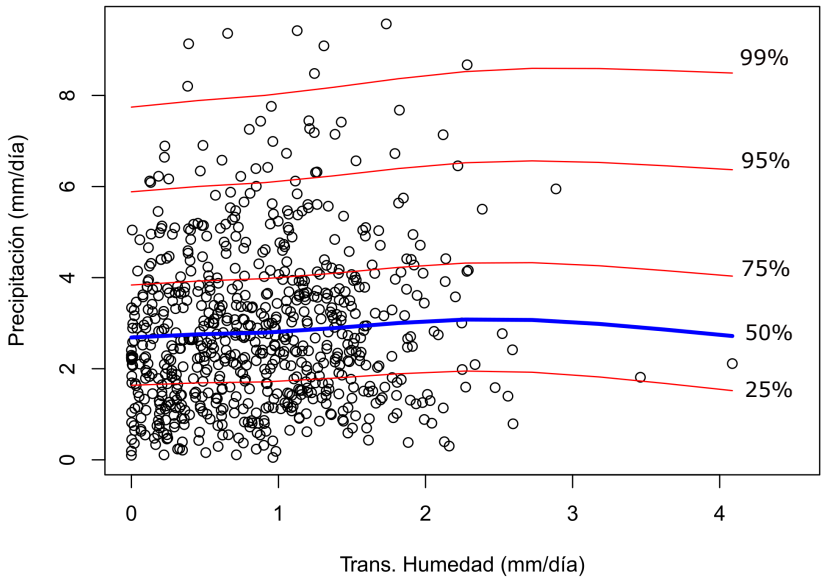

Figura 4.27: Cuantiles de la precipitación en función de la humedad transportada.

Inestabilidad atmosférica (lag 1)

Anteriormente se analizó como las variaciones de la inestabilidad atmosférica del mismo día, presentes en la región de interés, influye en la precipitación, por lo que ahora se analizará la inestabilidad del día anterior. Los resultados para cada uno de los cuantiles se muestran en la Tabla 4.23, donde se obtiene una mayor influencia de esta inestabilidad en la precipitación que la obtenida con los otros modelos. Incluso, en varios casos llega a ser mayor la cantidad de precipitación bajo estas condiciones que la obtenida al variar la inestabilidad sin ningún retardo. LLegando a alcanzar hasta valores menores o iguales a los 18 mm/día aproximadamente en el 99 % de los casos. Estos resultados demuestran la importancia de las condiciones meteorológicas de días anteriores para predecir variables como la precipitación.

Tabla 4.23: Cuantiles de la precipitación en función de diferentes valores de la inestabilidad atmosférica (lag 1).

Inest1 (Pa/s)	25 %	50 %	75 %	95 %	99 %
-0.058	0.866	2.261	3.790	6.508	8.977
-0.041	1.316	2.611	4.030	6.552	8.843
-0.023	1.588	2.793	4.114	6.461	8.593
-0.006	1.548	2.675	3.911	6.107	8.102
0.011	1.675	2.756	3.941	6.047	7.960
0.029	1.733	2.855	4.085	6.271	8.257
0.046	1.705	2.994	4.408	6.919	9.200
0.064	1.801	3.414	5.182	8.324	11.178
0.081	2.056	4.132	6.409	10.454	14.130
0.098	2.305	4.991	7.936	13.169	17.923

La representación de los cuantiles de la precipitación en función de la inestabilidad con lag 1 se muestra en la Figura 4.28.

Con este modelo de localización y escala, la representación de los cuantiles de la precipitación se ajustan mejor a la distribución de los datos. Como mismo ocurría con la inestabilidad del propio día, la precipitación tiende a aumentar a medida que aumenta la inestabilidad del día anterior. Hay que destacar que hasta aproximadamente una inestabilidad de 0.05 Pa/s, el aumento de la precipitación no era muy significativo, mientras que para valores más altos de inestabilidad, si se aprecia un mayor aumento de dicha variable. Incluso en los cuantiles 0.95 y 0.99, la precipitación tiene una pequeña disminución para luego volver a aumentar.

En este caso la variabilidad es mucho mayor para los valores más altos de inestabilidad. Se destacan también dos valores más extremos por encima del cuantil 0.99.

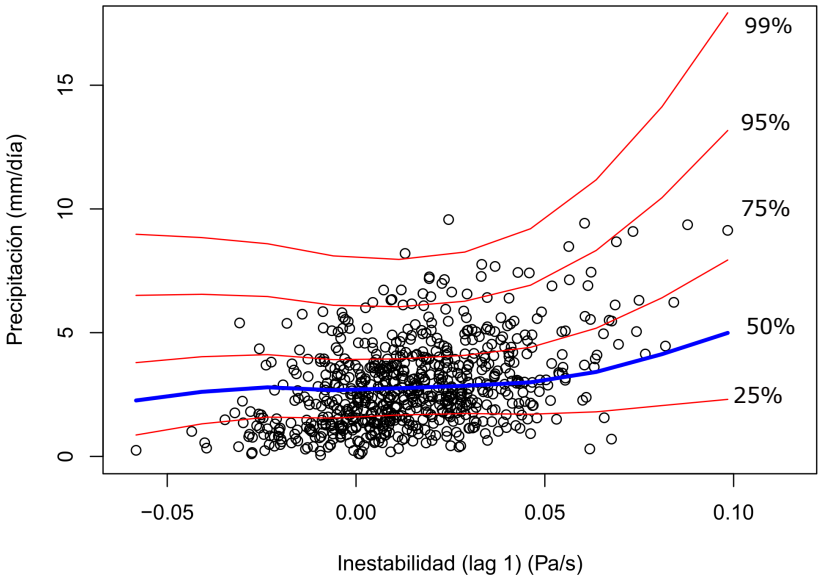

Figura 4.28: Cuantiles de la precipitación en función de la inestabilidad atmosférica con lag 1.

Precipitación (lag 1)

Por último se analizó la influencia de la precipitación del día anterior, como mismo en los modelos anteriores, cuyos resultados aparecen en la Tabla 4.24. Bajo estas condiciones los valores obtenidos para la precipitación, por lo general, tienden a ser superiores que el resto. Se resalta que, a medida que aumenta la variable independiente, la variable respuesta también aumenta, exceptuando para los valores más altos, que tiende a disminuir, en los cuantiles 0.25 y 0.50. Sin embargo para los cuantiles 0.95 y 0.99, es donde único se cumple que la precipitación predicha para el día de interés es superior a la del día anterior. Mientras que para el cuantil 0.25, solo es superior para el valor de precipitación con lag 1 igual a 0.062 mm/día, lo que significa que es prácticamente nula.

En la Figura 4.29 se representaron los cuantiles de la precipitación para este caso. Se confirma que con este modelo, se obtiene una mejor representación de los cuantiles, siguiendo de manera correcta la distribución de los datos. Se puede apreciar que mientras mayor sea la precipitación registrada en el día anterior, tenderá a ser mayor también la precipitación del día que se esté analizando. Como bien se mencionaba anteriormente, esta relación puede deberse a la presencia de algún sistema que esté afectando la región, lo que puede provocar la presencia de precipitaciones varios días consecutivos.

Para los valores más bajos de precipitación con lag 1, hay menos variabilidad, ya que las líneas de los cuantiles de la precipitación están más próximas. Sin embargo, para los valores mayores ocurre lo contrario, la diferencia es mayor, con una mayor variabilidad. En este caso, por encima de la línea que representa el cuantil 0.99 no se observa ningún valor.

4.5. Resumen de los resultados

Durante todo este capítulo se han aplicado diferentes modelos de regresión, desde algunos más simples hasta más complejos, para estudiar la precipitación asociada al chorro de bajos de niveles de las Grandes Llanuras Americanas. La regresión tuvo en cuenta diferentes cuantiles: 0.25, 0.50, 0.75, 0.95 y 0.99, lo que permitió el análisis más profundo de dicha variable. Para ello se utilizaron como variables predictoras la inestabilidad atmosférica y el agua total en la columna, presentes en la región de interés, así como la humedad transportada desde el Golfo de México desde días anteriores. Fue necesario incorporar nuevas variables predictoras debido a la alta autocorrelación presente en los residuos de

Tabla 4.24: Cuantiles de la precipitación en función de diferentes valores de la precipitación (lag 1).

Prec1 (mm/día)	25 %	50 %	75 %	95 %	99 %
0.062	0.235	0.598	2.171	4.964	7.502
1.345	0.770	1.935	3.212	5.480	7.542
2.628	1.600	2.678	3.860	5.961	7.870
3.911	2.113	3.278	4.556	6.826	8.889
5.194	2.532	3.898	5.397	8.059	10.478
6.477	2.780	4.376	6.126	9.235	12.061
7.760	2.816	4.620	6.597	10.111	13.304
9.043	2.495	4.516	6.731	10.666	14.243
10.326	2.005	4.269	6.751	11.161	15.169
11.609	1.481	4.018	6.799	11.742	16.232

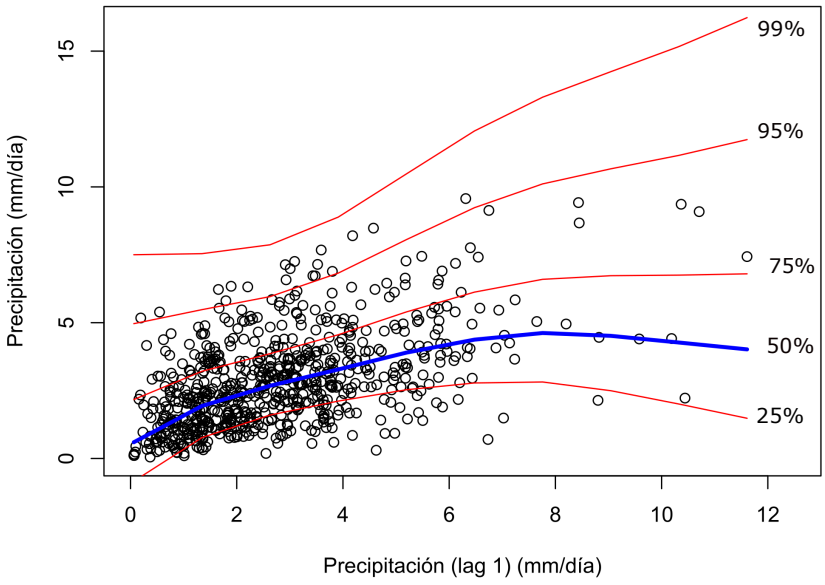

Figura 4.29: Cuantiles de la precipitación en función de la precipitación con lag 1.

los modelos utilizados. Dichas variables fueron las propias ya utilizadas pero con retardo igual a 1. La incorporación de estas nuevas variables corrigió en los modelos dicha autocorrelación, demostrando que las condiciones meteorológicas de días anteriores influyen en la cantidad de precipitación. Al incorporar nuevas variables fue necesario aplicar el criterio de selección de BIC para utilizar solo las variables más significativas en los modelos de regresión que se ajustaron. El ajuste de los modelos se realizó en una muestra de entrenamiento y las predicciones en una muestra test.

Se comenzó el análisis con el modelo más sencillo, el modelo de regresión lineal. Para este modelo, de forma general, los resultados no fueron muy satisfactorios. No cumplió con las hipótesis para la validación del modelo y presentó algunas limitaciones en la representación de los cuantiles, porque no se ajustaron de forma correcta a la distribución de la nube de puntos de la precipitación. Se utilizó también la función **rq**, teniendo en cuenta, tanto el ajuste de un modelo lineal como incorporando bases de B-splines a las variables predictoras con diferentes grados de libertad. Esto permitió obtener una mejor representación de los cuantiles de la precipitación, obteniendo mejores resultados, aunque también con algunas limitaciones. El hecho de demostrar que las relaciones entre las variables no era lineal, posibilitó el uso de modelos de regresión aún más flexibles, los modelos aditivos generalizados. Al obtener el ajuste de este modelo, a pesar de no haber cumplido correctamente la diagnosis, fue posible aplicar un modelo más robusto y complejo: el modelo de localización y escala. Este modelo de regresión no solo tiene en cuenta la magnitud de las variables sino también su variabilidad, lo que permitió obtener los mejores resultados. Con este modelo, los cuantiles representaron correctamente la distribución de la precipitación, así como la variabilidad. Además se pudo determinar las variables meteorológicas que más y menos influyen en la precipitación en la región de interés. Un resumen más detallado para lograr una mejor comparación entre los modelos se expone a continuación.

Para llegar a una comparación final con los mejores modelos, se compararon primeramente los modelos ajustados a partir de la función **rq**. Con esta función se ajustaron 3 modelos, el primero como un problema lineal y los otros 2, incorporando bases de B-Splines a las variables predictoras con 4 y 5 grados de libertad. Hay que destacar que en este caso se muestran los resultados para el cuantil 0.50, que fue el que presentó los mejores resultados. En la Figura 4.30 se muestra un resumen del comportamiento de las predicciones para estos modelos, comparados con los valores reales de precipitación. De forma general los resultados son muy similares, todos siguen el comportamiento real de la precipitación, destacándose unas ligeras sobrestimaciones entre los días 4 y 7.

Para una comparación más profunda, se calcularon el error absoluto medio (MAE), la raíz del error cuadrático medio (RMSE) y el pseudo-R^2 (equivalente al R^2_{ajust}) para cada uno de los modelos, cuyos resultados se muestran en la Tabla 4.25. Los resultados numéricos son muy similares para los 3 modelos, ligeramente mejores para el modelo con las bases con 5 grados de libertad, con los errores más pequeños, 1.230 y 0.929, y un valor de pseudo-R^2 igual a 0.493, lo que significa que el modelo es capaz de explicar, aproximadamente el 50 % de la variabilidad de la precipitación. Esto tiene mucho sentido, ya que como se vio en las figuras mostradas en las secciones anteriores, para este caso se obtenía una ligera mejora en la representación de los cuantiles de la precipitación.

Tabla 4.25: Comparación de los modelos de regresión cuantil con rq para diferentes grados de libertad.

Modelos	RMSE	MAE	Pseudo-R^2
rq	1.233	0.930	0.491
rq (df=4)	1.235	0.931	0.489
rq (df=5)	1.230	0.929	0.493

Por otro lado, se ajustaron modelos **qgam**, para los cuantiles 0.25, 0.50, 0.75, 0.95 y 0.99. Los

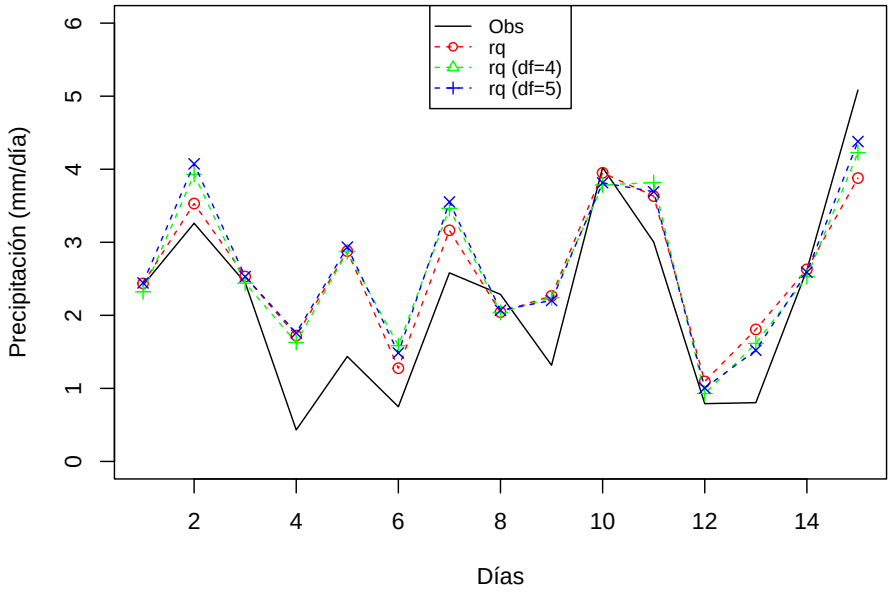

Figura 4.30: Comparación de los modelos de regresión cuantil rq para diferentes grados de libertad.

resultados de las predicciones para cada uno de ellos se muestran en la Figura 4.31.

En esta se puede apreciar como, a medida que va aumentando el cuantil, se va alejando más de los valores reales, lo que significa que, los cuantiles más altos representan los valores más extremos de la precipitación. Es por ello que, teniendo en cuenta esta representación, las mejores predicciones, con respecto a los valores reales, se corresponden con los cuantiles 0.25 y 0.50.

Para una correcta interpretación y comparación de los resultados, se calculó igualmente para estos modelos el RMSE, el MAE y el pseudo-R^2 (Tabla 4.26). En este caso, como era de esperar las diferencias son más significativas, incluso para los cuantiles superiores se obtiene valores de pseudo-R^2 negativos, lo cual no tiene sentido y demuestra el mal ajuste de estos modelos, ya que sus valores se alejan del resto. Además, presentaron los mayores errores, igual es 3.7 y 3.5. Para los modelos con los cuantiles 0.25, 0.50 y 0.75, los resultados son relativamente mejores, sobre todo para el cuantil 0.5. Este es capaz de explicar el 50 % de la variabilidad de la precipitación, en función de las variables predictoras, con valores de RMSE y MAE iguales a 1.2 y 0.9, respectivamente. De esta manera se concluye que de estos modelos el que se corresponde con el cuantil 0.5 es el que presentó los mejores resultados.

Con el análisis realizado se puede pasar a realizar la comparación final entre los diferentes modelos seleccionados. Dichos modelos son:

- el modelo lineal (LM),
- el modelo con la función rq con 5 grados de libertad (RQ),
- el modelo aditivo generalizado (GAM),
- el modelo aditivo generalizado con la función qgam para el cuantil 0.5 (QGAM) y,
- el modelo de localización y escala (LE).

En la Figura 4.32 se realiza una comparación de las predicciones de cada uno de estos modelos con respecto a los valores reales. Las predicciones realizadas por todos los modelos, excepto el de localización y escala, son muy similares entre sí. Dichas predicciones, de forma general, siguen el comportamiento de la precipitación real, aunque presentan algunas limitaciones. Además, no son capaces de predecir

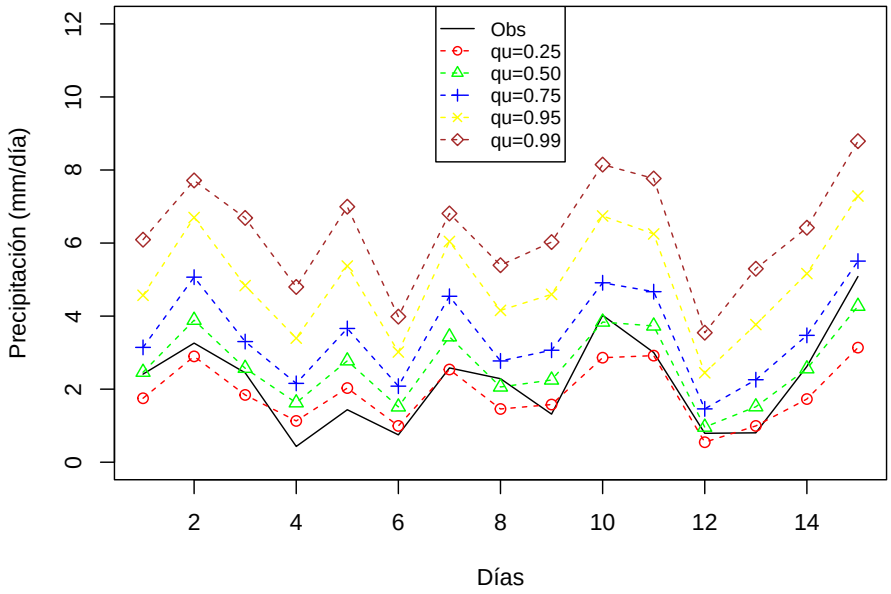

Figura 4.31: Comparación de los modelos de regresión cuantil qgam para diferentes cauntiles.

Tabla 4.26: Comparación de los modelos de regresión cuantil qgam para diferentes cauntiles.

Modelos	RMSE	MAE	Pseudo-R^2
qu=0.25	1.527	1.142	0.220
qu=0.50	1.220	0.921	0.502
qu=0.75	1.403	1.136	0.341
qu=0.95	2.512	2.212	-1.112
qu=0.99	3.727	3.473	-3.650

correctamente ciertos valores, sobreestimando los verdaderos valores de la variable respuesta. Mientras que, a diferencia del resto, el modelo de localización y escala, fue capaz de representar mejor los valores más extremos de la precipitación. Esto corrobora los resultados mencionados en la sección anterior, donde los mejores resultados se obtuvieron con este modelo. Y es que era de esperar este resultado, ya que este modelo incorpora en el ajuste, no solo la magnitud de los datos sino también la variabilidad. Esto hace que sea más robusto y más preciso que el resto de los modelos.

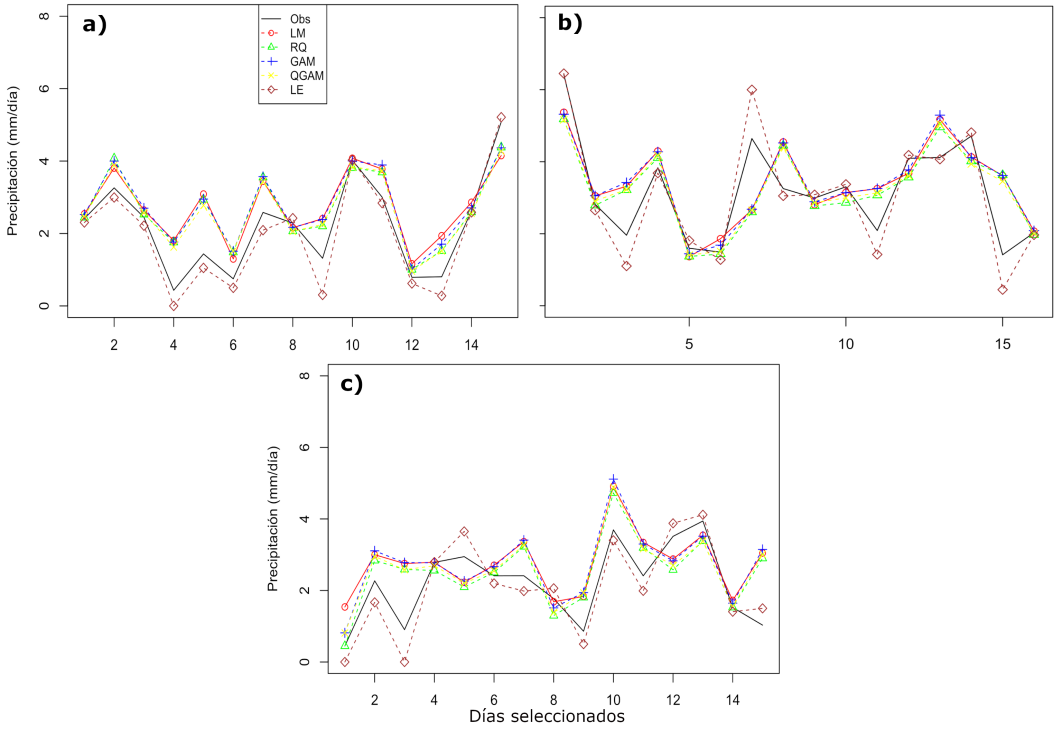

Figura 4.32: Comparación de los modelos de regresión finales. a) inicio, b) mediados y c) finales.

Para reafirmar estos resultados, se calcularon los valores de RMSE, MAE y pseudo-R^2, igualmente para cada uno de estos modelos, los cuales se muestran en la Tabla 4.27. Como se observó en la figura anterior, los resultados para los modelos LM, RQ, GAM y QGAM son muy similares, con un valor de pseudo-R^2 en torno a los 0.5, y RMSE y MAE, iguales a 1.2 y 0.9, respectivamente. Por otro lado, los resultados del modelo de localización y escala, confirman una vez que es el mejor modelo para representar la relación entre la variable respuesta precipitación y las variables explicativas: inestabilidad tanto con lag 1 como sin lag, y la humedad transportada y la precipitación, ambas con retardo igual a 1. Dicho modelo es capaz de explicar aproximadamente el 73 % de la variabilidad de la precipitación en función de estas variables predictoras. Este valor se considera un buen resultado, teniendo en cuenta la complejidad de trabajar con variables meteorológicas, en las cuales influyen condiciones tanto directas como indirectas. Los errores son más pequeños en comparación con el resto, iguales a 0.89 y 0.64. La obtención de estos resultados permite la utilización de este modelo de regresión en análisis más profundos, no solo de estas variables meteorológicas, sino también de muchas otras.

Además, otro método que se aplicó para evaluar la precisión de las predicciones son los gráficos de dispersión de las observaciones frente a las predicciones. Para considerar un buen ajuste los puntos debieran colocarse sobre la línea recta $y = x$. En la Figura 4.33, se muestran dichos gráficos para los modelos LM, RQ, GAM y QGAM. Para estos 4 modelos los resultados son muy similares, los puntos están dispersos alrededor de la línea, lo que confirma los resultados obtenidos anteriormente para estos modelos. A pesar de ello, de forma general, se observan unas ligeras infraestimaciones en los valores

Tabla 4.27: Comparación de los modelos de regresión finales.

Modelos	RMSE	MAE	Pseudo-R^2
LM	1.214	0.928	0.506
RQ	1.230	0.929	0.493
GAM	1.213	0.929	0.507
QGAM	1.220	0.921	0.502
LE	0.891	0.643	0.734

más bajos y sobreestimaciones para valores mayores.

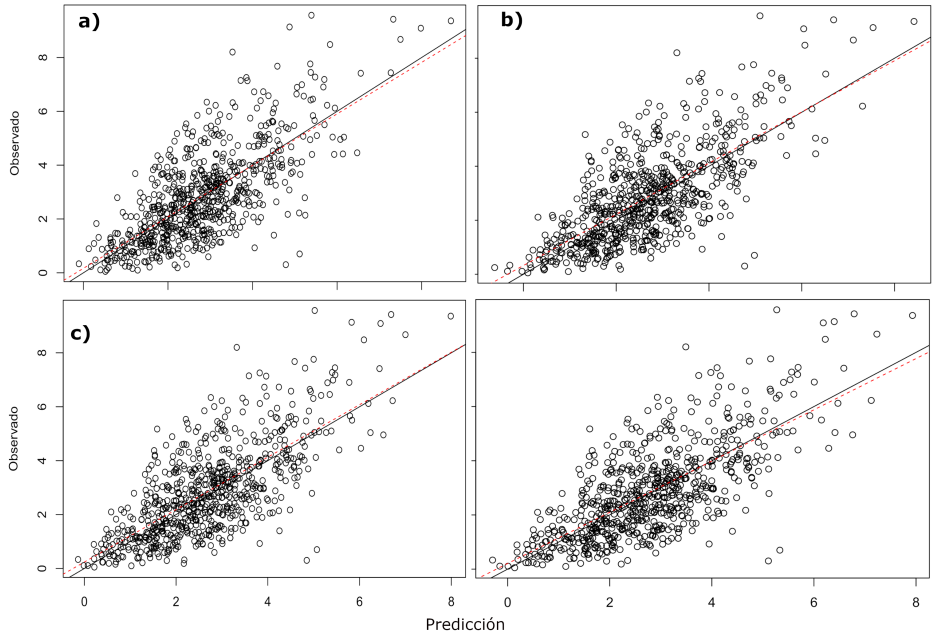

Figura 4.33: Gráfico de dispersión de observaciones frente a predicciones (línea continua roja es la identidad y línea discontinua es el ajuste). a) LM, b) RQ, c) GAM y d) QGAM.

Por otro lado, se muestra en la Figura 4.34, el gráfico de dispersión para el modelo de localización y escala. En este caso la dispersión de los datos es menor, es decir, se obtiene un mejor ajuste sobre la línea recta, lo que demuestra que este modelo es el más adecuado entre todos los ajustados para representar la relación entre las variables. Sin embargo, se puede ver que para valores más altos el modelo sobrestima los valores de la precipitación, lo que puede estar relacionado con los altos valores obtenidos en los cuantiles mostrados, mientras que para los valores más pequeños se observan unas ligeras infraestimaciones.

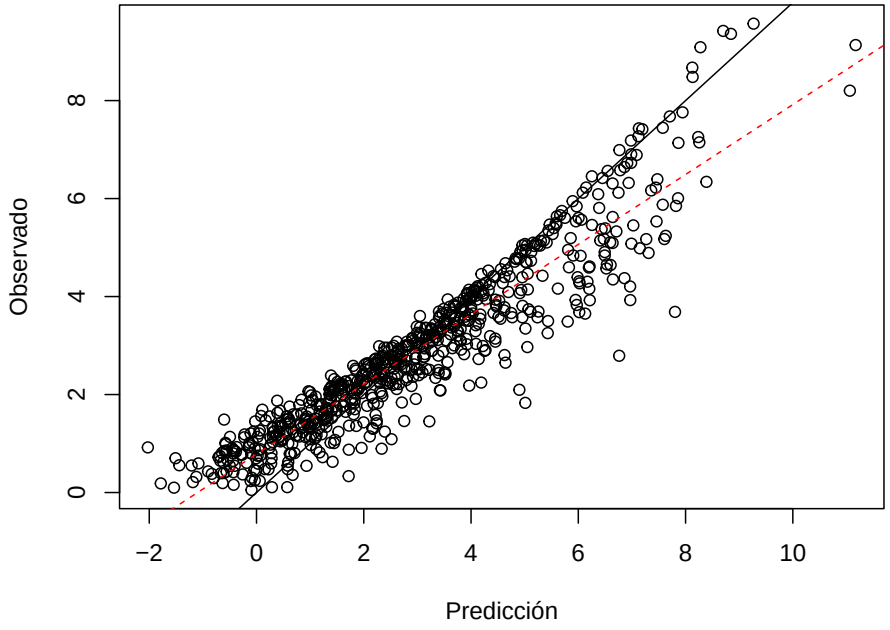

Figura 4.34: Gráfico de dispersión de observaciones frente a predicciones del modelo de localización y escala (línea continua roja es la identidad y línea discontinua es el ajuste).

Conclusiones

Durante la realización de este trabajo se aplicaron diferentes técnicas estadísticas de regresión en el estudio de la precipitación asociada al chorro de bajos niveles de las Grandes Llanuras Americanas en función de la inestabilidad atmosférica y el agua total en la columna en la región de estudio, así como, de la humedad transportada desde la región fuente del Golfo de México. Este análisis permitió extraer conclusiones, desde el punto de vista estadístico y también meteorológico, las cuales se destacan a continuación:

1. La incorporación de variables predictoras con retardos iguales a 1 corrigió la fuerte autocorrelación de los residuos de los modelos en los primeros lag, lo que demuestra la influencia de las condiciones meteorológicas en días anteriores sobre las condiciones de los días actuales.

2. El criterio de BIC seleccionó como variables más significativas para el ajuste de los modelos: la inestabilidad atmosférica sin lag y con lag 1, la humedad transportada con lag 1 y la precipitación con lag 1, lo que significa que en la precipitación que se registre hoy, influye en cierta medida, la inestabilidad para hoy, así como la inestabilidad, la humedad y la precipitación del día anterior.

3. La regresión por cuantiles permitió estudiar con profundidad el comportamiento de la precipitación teniendo en cuenta las variaciones de las variables predictoras en diferentes partes de la distribución. Las variaciones de la inestabilidad atmosférica del propio día y la precipitación del día anterior, resultaron ser las que más influyeron en las variaciones de la precipitación, aumentando, a medida que aumentan las variables predictoras.

4. Se comprobó que el comportamiento de la precipitación, en función de la variabilidad de cada una de las variables independientes, depende también de los cuantiles utilizados, ya que sobre todo para los modelos más complejos, la precipitación presentó diferentes comportamientos en determinados cuantiles.

5. Los modelos lineal y aditivos generalizados no cumplieron las hipótesis para su validación, sin embargo se realizó el procedimiento completo para analizar también el comportamiento de las predicciones, comparar los resultados y poder arrojar conclusiones.

6. Con el aumento de la complejidad de los modelos de regresión utilizados durante el desarrollo del trabajo, se fueron obteniendo mejores resultados y mejores ajustes en cuanto a la distribución de la precipitación, lo que demuestra la necesidad de modelos más robustos para el estudio de la relación entre las variables analizadas.

7. El modelo de localización y escala arrojó los mejores resultados, tanto desde el punto de vista gráfico como numérico. Presentó los menores valores de errores y el mayor pseudo-R^2, las predicciones eran muy similares a los valores reales, sobreestimando para los mayores valores, y las líneas que representaban a los diferentes cuantiles se ajustaban correctamente a la distribución de la precipitación.

8. Los resultados obtenidos con el ajuste del modelo de localización y escala son de gran importancia para comprender el comportamiento de la precipitación en la región de estudio y poder actuar

con mayor eficacia ante la ocurrencia de posibles riesgos, ya sea de precipitaciones extremas o sequías.

9. En trabajos futuros se pretende ampliar el estudio de la precipitación en otras regiones de interés, incorporando de ser necesario nuevas variables predictoras con el objetivo de obtener un mejor ajuste de los modelos. Además, aplicar otras técnicas estadísticas que cuenten con una mayor capacidad predictiva y que representen con más robustez la posible relación que existe entre las variables predictoras y la respuesta.

Bibliografía

[1] Algarra, I., Eiras-Barca, J., Nieto, R., y Gimeno, L. (2019). Global climatology of nocturnal low-level jets and associated moisture sources and sinks. Atmospheric Research, 229:39-59.

[2] Balaguer Beser, Á. A., y Ruiz Fernández, L. Á. (2021). Selección de un modelo de regresión lineal múltiple para el cálculo de la precipitación media en verano. Universidad Politécnica de Valencia.

[3] Basara, J. B., Maybourn, J. N., Peirano, C. M., Tate, J. E., Brown, P. J., Hoey, J. D., y Smith, B. R. (2013). Drought and associated impacts in the Great Plains of the United States-a review. International Journal of Geosciences, 4(6B):72-81.

[4] Bonner, W. D. (1968). Climatology of the low level jet. Monthly Weather Review, 96(12), 833-850.

[5] Chen, C., Tao, W.-K., Lin, P.-L., Lai, G. S., Tseng, S., y Wang, T.-C. C. (1998). The intensification of the low-level jet during the development of mesoscale convective systems on a mei-yu front. Monthly Weather Review, 126(2):349-371.

[6] Conde, M. (2013). Contraste de Bondad de Ajuste de Modelos de Regresión Cuantil. Proyecto de Fin de Máster. Universidad de Santiago de Compostela.

[7] Crujeiras, R.M. y Conde Amboage, M. (2019). El modelo de regresión lineal simple. Inferencia estadística-Grado en Matemáticas.

[8] De Boor, C., y De Boor, C. (1978). A practical guide to splines (Vol. 27, p. 325). New York: springer-verlag.

[9] Cowpertwait, P.S.P. y Metcalfe, A.V. (2009). Introductory Time Series with R. Springer.

[10] Díaz, S. (2020). Regresión cuantil con datos censurados. Proyecto de Fin de Máster en Técnicas Estadísticas.

[11] Fasiolo, M., Wood, S. N., Zaffran, M., Nedellec, R., y Goude, Y. (2020). qgam: Bayesian non-parametric quantile regression modelling in R. arXiv preprint arXiv:2007.03303.

[12] Fasiolo, M., Wood, S. N., Zaffran, M., Nedellec, R., y Goude, Y. (2021b). Fast calibrated additive quantile regression. Journal of the American Statistical Association, 116(535), 1402-1412.

[13] Fox, J. (2019). Regression diagnostics: An introduction. Sage publications.

[14] Gimeno, L., Dominguez, F., Nieto, R., Trigo, R., Drumond, A., Reason, C. J., Taschetto, A. S., Ramos, A. M., Kumar, R., y Marengo, J. (2016). Major mechanisms of atmospheric moisture transport and their role in extreme precipitation events. Annual Review of Environment and Resources, 41:117-141.

[15] Gimeno-Sotelo, L. (2021). Univariate and Bivariate Extremes in Meteorology: an application to the Great Plains Low-Level Jet System. Tesis de Maestrían en Estadística e Investigación Operativa. Universidad de Lisboa.

[16] Hampel, F. R., Ronchetti, E. M., Rousseeuw, P. J. y Stahel, W. A. (1986). Robust statistics: The approach based on influence functions. Wiley.

[17] Hao, L., y Naiman, D. Q. (2007). Quantile regression (No. 149). Sage.

[18] Hastie, T. y Tibshirani, R. (1990). Exploring the nature of covariate effects in the proportional hazards model. Biometrics, 1005-1016.

[19] Helfand, H. M. y Schubert, S. D. (1995). Climatology of the simulated Great Plains low-level jet and its contribution to the continental moisture budget of the United States. Journal of Climate, 8(4):784-806.

[20] Hersbach, H., Bell, B., Berrisford, P., Hirahara, S., Horányi, A., Muñoz-Sabater, J., Nicolas, J., Peubey, C., Radu, R. y Schepers, D. The ERA5 global reanalysis. Quarterly Journal of the Royal Meteorological Society, 146(730):1999-2049.

[21] Higgins, R. W., Yao, Y., Yarosh, E. S., Janowiak, J. E., y Mo, K. C. (1997). Influence of the great plains low-level jet on summertime precipitation and moisture transport over the central United States, J. Climate, 10: 481-507

[22] Hodges, D. y Pu, Z. (2019). Characteristics and variations of low-level jets in the contrasting warm season precipitation extremes of 2006 and 2007 over the Southern Great Plains. Theoretical and Applied Climatology, 136(1):753-771.

[23] Huber, P. J. (1981). Robust statistics. Wiley.

[24] Izenman, A. J. (2013). Multivariate regression. Modern multivariate statistical techniques (pp.159-194). Springer.

[25] Koenker, R. (2005). Quantile regression (Vol. 38). Cambridge university press.

[26] Koenker, R. y Bassett, G. W. (1978). Regression Quantiles. Econometrica, 46, 33-50.

[27] Koenker, R.W. and dOrey (1987, 1994). Computing regression quantiles. Applied Statistics, 36, 383-393, y 43, 410-414.

[28] Lema, M. (2022). Modelos de regresión de resposta multivariante e a súa aplicación a datos biomédicos. Tesis de Máster en Técnicas Estadísticas. Universidade da Coruña.

[29] Lin, X y Zhang, D. (1999) Inference in generalized additive mixed models by using smoothing splines. JRSSB. 55(2):381-400

[30] López, H. A., y Mora, H. (2007). Cálculo de los estimadores de regresión cuantílica lineal por medio del método ACCPM. Revista Colombiana de Estadística, 30(1), 53-68.

[31] Martínez Silva, I., Roca Pardiñas, J. y Ordóñez, C. (2016). Forecasting SO2 pollution incidents by means of quantile curves based on additive models. Environmetrics, 27(3), 147-1

[32] Conde, M. (2013). Contraste de bondad de ajuste de modelos de regresión cuantil. Máster en Técnicas Estadísticas. Universidad de Santiago de Compostela.

[33] Mo, K. C., Nogues-Paegle, J., y Paegle, J. (1995). Physical mechanisms of the 1993 summer floods. Journal of Atmospheric Sciences, 52(7):879-895.

[34] Neath, A. A., y Cavanaugh, J. E. (2012). The Bayesian information criterion: background, derivation, and applications. Wiley Interdisciplinary Reviews: Computational Statistics, 4(2), 199-203.

[35] Nelder, J. A., y Wedderburn, R. W. (1972). Generalized liner models. Journal of the Royal Statistical: Series A (General), 135(3): 370-384.

[36] Novales, A. (2010). Análisis de regresión. Departamento de Economía Cuantitativa de la Universidad Complutense de Madrid.

[37] Parra, F. (2019). Estadística y Machine Learning con R.

[38] Pitchford, K. L., y London, J. (1962). The low-level jet as related to nocturnal thunderstorms over Midwest United States. Journal of Applied Meteorology and Climatology, 1(1), 43-47.

[39] Rodríguez, S. (2023). Modelos de regresión multivariante y su aplicación en datos médicos. Tesis Máster en Técnicas Estadísticas. Universidad de Santiago de Compostela.

[40] Schumacher, R. S. y Johnson, R. H. (2009). Quasi-stationary, extreme-rain-producing convective systems associated with midlevel cyclonic circulations. Weather and Forecasting, 24(2):555-574.

[41] Martínez Silva, I., Roca Pardiñas, J., y Ordóñez, C. (2016). Forecasting SO2 pollution incidents by means of quantile curves based on additive models. Environmetrics, 27(3), 147-157.

[42] Squitieri, B. J. y Gallus, W. A. (2016). WRF forecasts of Great Plains nocturnal low-level jet-driven MCSs. Part I: Correlation between low-level jet forecast accuracy and MCS precipitation forecast skill. Weather and Forecasting, 31(4):1301-1323.

[43] Squitieri, B. J. y Gallus Jr,W. A. (2016). WRF forecasts of Great Plains nocturnal low-level jet-driven MCSs. Part II: Differences between strongly and weakly forced low-level jet environments. Weather and Forecasting, 31(5):1491-1510.

[44] Stensrud, D. J. (1996). Importance of low-level jets to climate: A review. Journal of Climate, 1698-1711.

[45] Viechtbauer, W. y López-López, J. A. (2022). Location-scale models for meta-analysis. Research Synthesis Methods, 13(6), 697-715.

[46] Walters, C. K. y Winkler, J. A. (2001). Airflow configurations of warm season southerly low-level wind maxima in the Great Plains. Part I: Spatial and temporal characteristics and relationship to convection. Weather and Forecasting, 16(5):513-530.

[47] Weakliem, D. L. (1999). A critique of the Bayesian information criterion for model selection. Sociological Methods & Research, 27(3), 359-397.

[48] Wood S.N. (2017). Generalized additive models: an introduction with R. Second Edition. Boco Raton: CRC Press.

[49] Wood, S.N., Pya, N y Saefken, B. (2016). Smoothing parameter and model selection for general smooth models (with discussion). Journal of the American Statistical Association 111, 1548-1575.

[50] Xie, P., Chen, M., y Shi, W. (2010). CPC unified gauge-based analysis of global daily precipitation. In Preprints, 24th Conf. on Hydrology, Atlanta, GA, Amer. Meteor. Soc, volume 2.

[51] Urban, J. (2006). Joint Meeting of CT-MTDCF/ET-DRC, Montreal. https://confluence.ecmwf.int/display/TIGGE/Total+column+water. Accedido el 20 de febrero de 2024.