# MODELOS LINEALES DINÁMICOS PARA VALORES EXTREMOS

Elvis Rafael Arrazola Acosta[a], Izhar Asael Alonzo Matamoros[b], Cristian Andrés Cruz Torres[c]

[a]Departamento de Estadística Matemática, Universidad Nacional Autónoma de Honduras, elvis.arrazola@unah.edu.hn, ORCID: https://orcid.org/0009-0008-6807-0896

[b]Departamento de Estadística Matemática, Universidad Nacional Autónoma de Honduras, izhar.alonzo@unah.edu.hn

[c]Departamento de Estadística Matemática, Universidad Nacional Autónoma de Honduras, cristian.cruz@unah.edu.hn, ORCID: https://orcid.org/0000-0002-2185-5783

**DOI:** https://doi.org/10.5377/pc.v1i19.18700

***Recepción:*** 19/08/2023
***Aceptación:*** 15/05/2024

## *Resumen*

Actualmente, el cambio climático es uno de los fenómenos que preocupa a la población mundial, por ello, proponemos un enfoque para modelar valores extremos medidos de lluvia, sequía, etc. Primero, las observaciones siguen una distribución de valor extremo generalizado (GEV) para la cual los parámetros de ubicación, escala o forma definen la estructura espacio-temporal. La distribución generalizada de valores extremos se amplía para incorporar la dependencia del tiempo, utilizando una representación de espacio de estado, donde las variables de estado se miden a través de un modelo lineal dinámico (DLM). El elemento espacial se impone a través de la matriz de evolución del DLM, en la que adoptamos una forma de proceso de convolución. Mostramos cómo producir estimaciones temporales y espaciales de nuestro modelo a través de una simulación personalizada de Markov Chain Monte Carlo (MCMC). La metodología se ilustra utilizando rendimientos extremos de datos mediante mediciones diarias de los niveles de precipitación producidos diariamente en el estado de Washington, EE. UU.

**Palabras clave:** valores extremos, modelos lineales dinámicos, Markov Chain Monte Carlo, procesos de convolución, precipitación

# DYNAMIC LINEAR MODELS FOR EXTREME VALUES

Elvis Rafael Arrazola Acosta[a], Izhar Asael Alonzo Matamoros[b], Cristian Andrés Cruz Torres[c]

[a]Departamento de Estadística Matemática, Universidad Nacional Autónoma de Honduras, elvis.arrazola@unah.edu.hn, ORCID: https://orcid.org/0009-0008-6807-0896

[b]Departamento de Estadística Matemática, Universidad Nacional Autónoma de Honduras, izhar.alonzo@unah.edu.hn

[c]Departamento de Estadística Matemática, Universidad Nacional Autónoma de Honduras, cristian.cruz@unah.edu.hn, ORCID: https://orcid.org/0000-0002-2185-5783

**DOI:** https://doi.org/10.5377/pc.v1i19.18700

***Recepción:*** 19/08/2023
***Aceptación:*** 15/05/2024

## *Abstract*

Currently, climate change is one of the phenomena that worries the world's population, which is why we propose an approach to model measured extreme values of rainfall, drought, etc. First, observations follow a Generalized Extreme Value (GEV) distribution for which location, scale, or shape parameters define the spatiotemporal structure. The generalized distribution of extreme values is extended to incorporate time dependence using a state space representation where state variables are measured through a Dynamic Linear Model (DLM). The spatial element is imposed through the evolution matrix of the DLM where we adopt a form of convolution process. We show how to produce temporal and spatial estimates of our model through a custom Markov Chain Monte Carlo (MCMC) simulation. The methodology is illustrated using extreme data yields through daily measurements of precipitation levels produced daily in Washington State, USA.

**Keywords:** extreme values, dynamic linear models, Markov Chain Monte Carlo, convolution processes, precipitation

## *Introducción*

En el campo de las ciencias ambientales y la meteorología, comprender y predecir patrones de precipitación es de vital importancia, debido a sus amplias implicaciones en la agricultura, la hidrología y la gestión de desastres naturales. A lo largo de los años, los avances en el análisis de datos de series temporales han brindado herramientas poderosas para abordar la complejidad de los registros de precipitación.

En esta introducción, exploraremos la combinación de tres enfoques clave: modelos de espacios de estado, modelos lineales dinámicos y la teoría de valores extremos generalizada, que juntos forman un marco sólido para el estudio de precipitaciones extremas. En este contexto nos centramos en analizar las precipitaciones diarias de 18 estaciones meteorológicas del estado de Washington mediante un modelo lineal dinámico (DLM, por sus siglas en inglés) para la parte espacio-temporal, el cual se presenta como un caso especial de los modelos en espacio de estados (SSM, por sus siglas en inglés), siendo estos SSM lineales y gaussianos. Para un DLM, la estimación y el pronóstico se pueden obtener recursivamente mediante el conocido filtro de Kalman (Triantafyllopoulos, 2021). Modelos en espacio de estados han emergido como una potente metodología para analizar series temporales complejas. Estos modelos permiten descomponer la serie en dos componentes fundamentales: la componente observada y la componente no observada o latente, que captura la dinámica subyacente del fenómeno. En el contexto de las precipitaciones, los modelos de espacios de estado pueden ayudar a identificar patrones de comportamiento y tendencias ocultas, lo que es esencial para comprender los cambios climáticos a lo largo del tiempo. Consideremos una serie de tiempo $Y_t, t = 1,2 \ldots\}$, donde $\boldsymbol{Y_t}$ es un vector aleatorio observable $(s \times 1)$; por ejemplo, $Y_t = (Y_{1,t}, \ldots, Y_{s,t})$ son $\boldsymbol{S}$ estaciones de alguna región en el momento $\boldsymbol{t}$. Para hacer inferencias sobre la serie de tiempo, en particular para predecir el siguiente valor $Y_{t+1}$ dadas las observaciones $(Y_1, \ldots, Y_t)$ necesitamos especificar la ley de probabilidad del proceso $(Y_t)$, lo que significa dar la estructura de dependencia entre las variables de $Y_t$. Para una mejor comprensión de los SSM es pertinente notar que estos se basan en la idea de que la serie de tiempo $(Y_t)$ es una función incompleta y ruidosa de algún proceso subyacente no observable $\boldsymbol{\theta_t}$, denominado como proceso de estados. En aplicaciones de ingeniería, $\theta_t$ generalmente describe el estado de un sistema físico que produce la salida $\boldsymbol{Y_t}$ con perturbaciones aleatorias.

De forma general, podríamos pensar en $\{\boldsymbol{\theta_t}\}$ como un proceso aleatorio auxiliar que facilita la tarea de especificar la ley de probabilidad de la serie de tiempo. El proceso observable $(\boldsymbol{Y_t})$ depende del proceso de estado latente $(\boldsymbol{\theta_t})$, que tiene una dinámica Markoviana más simple. En consecuencia, podemos suponer razonablemente que la observación $\boldsymbol{Y_t}$ solo depende del estado del sistema y en el momento de la medición se toma $\boldsymbol{\theta_t}$.

Los supuestos de un SSM son:

1. $\{\boldsymbol{\theta_t}\}$ es una cadena de Markov, es decir, $\theta_{t-1}$ depende únicamente del valor pasado $\theta_{t-1}$. Así, la ley de probabilidad del proceso $\{\boldsymbol{\theta_t}\}$ se especifica asignando la densidad inicial $\pi_t(\boldsymbol{\theta_t})$ de $\boldsymbol{\theta_0}$ y las densidades de transición $\pi(\theta_t|\theta_{t-1})$ de $\boldsymbol{\theta_t}$ condicionada en $\theta_{t-1}$.

2. Condicionalmente a $\{\boldsymbol{\theta_t}\}$, los $\boldsymbol{Y_t}$ son independientes entre sí y $\boldsymbol{Y_t}$ depende únicamente de $\boldsymbol{\theta_t}$. Se sigue que, para cualquier n ≥ 1, $(Y_t, \ldots, Y_t|_{\theta_1,\ldots,\theta_n})$, tienen densidad condicional conjunta $\prod_{t=1}^{n} f(y_t|\theta_t)$.

De lo anterior, se nota que el modelo queda completamente especificado por la distribución inicial o distribución *a priori,* $\pi(\theta_t|\theta_{t-1})$ y las densidades condicionales $\pi(\theta_t|\theta_{t-1})$ y $\pi(y_t|\theta_t)$. En particular, vemos que el proceso $\{(\theta_t|\ Y_t)\}$ es de Markov. La densidad de $(Y_1,\ldots,Y_n)$ se puede obtener integrando todas las variables $\boldsymbol{\theta}$ de la densidad conjunta. Las siguientes secciones muestran que los cálculos son bastante simples en el SSM lineal gaussiano; pero, en general, la densidad de $(Y_1,\ldots,Y_n)$ no está disponible en forma cerrada y el proceso observable $(Y_t)$ no es Markoviano. Sin embargo, podemos ver una propiedad importante: $\boldsymbol{Y_t}$ es condicionalmente independiente de las observaciones pasadas $(Y_1,\ldots;$ $Y_{t-1})$ dado el valor de $\theta_t$. Esto nos da una interpretación atractiva del estado $\theta_t$: representa alguna información cuantitativa que resume la historia pasada del proceso observable y es suficiente para predecir su comportamiento futuro.

En este estudio, nos enfocamos en explorar el uso de la distribución de valores extremos generalizada

(GEV) y su capacidad para permitir la variación tanto en el tiempo como en el espacio de sus parámetros. En la sección 2 del artículo, proponemos diferentes enfoques para analizar series de tiempo de eventos extremos registrados en diversas ubicaciones espaciales. Mediante la aplicación de modelos lineales dinámicos, inspirados en el trabajo de West y Harrison (1997), investigamos el comportamiento cambiante en el tiempo de los parámetros que gobiernan la distribución de estos datos extremos. De este modo, logramos capturar tendencias temporales variables, lo que nos permite detectar cambios de corto y largo plazo en los eventos extremos.

Además, nos enfrentamos al desafío de relacionar estas tendencias temporales con covariables que también varían en el tiempo. De esta manera, podemos describir cómo la fuerza de la relación lineal se modifica con el transcurso del tiempo. La utilización de un enfoque dinámico en nuestro modelo nos brinda la capacidad de adaptarnos a las fluctuaciones y evoluciones en los datos extremos a lo largo del tiempo. Para lograr un mejor análisis, basamos nuestro enfoque en convoluciones de procesos, siguiendo métodos presentados por Cressie (1992) y otros investigadores. Al hacerlo, logramos una eficiente reducción del espacio de parámetros, lo cual es fundamental cuando trabajamos con un gran número de ubicaciones espaciales, esto nos permite examinar y comparar patrones de extremos en diferentes lugares geográficos. Trabajamos usando el modelo presentado por Huerta, Sansó y Stroud (2004) y un modelo de distribución generalizada de valores extremos con dependencia del tiempo, empleando los modelos lineales dinámicos en forma de espacio de estado. Aplicamos métodos computacionales para la estimación de parámetros vía HMC (Strawderman & Friel, 2000) y Metropolis-Hasting (Metropolis, Rosenbluth, Rosenbluth, Teller & Teller, 1953), además, del algoritmo FFBS (Frühwirth-Schnatter, 1994) mencionado anteriormente.

## *Modelos lineales dinámicos*

Los modelos lineales dinámicos son una clase especial de modelos de espacios de estado que asumen linealidad en las relaciones entre las variables y en la evolución temporal. En el estudio de precipitaciones estos modelos pueden ser especialmente útiles para analizar la variabilidad de los patrones de precipitación a lo largo de diferentes estaciones del año o en diferentes regiones geográficas. Además, los modelos lineales dinámicos permiten estimar la incertidumbre asociada con las predicciones, lo que es esencial para la toma de decisiones informadas en el manejo de recursos hídricos y prevención de inundaciones.

Uno de los casos particulares y más importantes de los SSM son los DLM. Un DLM queda especificado por una distribución *a priori* normal p-dimensional, para el estado del vector al tiempo t = 0, es decir,

$$(1)\ \theta_0 \sim N_p(c_0, m_0)$$

junto con las ecuaciones para cada tiempo,

$$(2)\ Y_t = F_t\,\theta_t + v_t\,, \qquad v_t \sim N(0\,, V_t)$$
$$(3)\ \theta_t = G_t\,\theta_{t-1} + w_t\,, \quad w_t \sim N(0\,, W_t)$$

donde $Y_t$ es un vector de orden, $m, \theta_t$ es un vector de orden, p, $G_t$ y $F_t$ son matrices conocidas de orden $p \times p$ y $m \times p$, respectivamente, y $\{v_t\}$ y $\{w_t\}$ son dos sucesiones de variables aleatorias independientes con distribución gaussiana, con media cero y varianzas $\{v_t\}$ y $\{w_t\}$, respectivamente (Pole, West, & Harrison, 1994). Además, se supone que $\theta_0$ es independiente de $\{v_t\}$ y $\{w_t\}$. Las ecuaciones (2) y (3) son conocidas como las ecuaciones de observación y estados respectivamente. Para el resto del trabajo adoptaremos la notación propuesta por West y Harrison (1997). La cuádrupla DLM $\{F_t, G_t, V_t, W_t\}$ representa el DLM propuesto en las ecuaciones (2) y (3), y cuando las matrices en la cuádrupla son constantes, se dice que el DLM es invariante con respecto al tiempo.

Los DLM pueden ser considerados como una generalización de los modelos de regresión lineal porque permiten la variación de los coeficientes de regresión en el tiempo (Pole, West, & Harrison, 1994). De igual forma, los DLM son un caso particular de los SSM, dado que estos permiten especificar cualquier otra distribución *a priori* distinta a la de Gauss, junto con la especificación de ecuaciones no lineales:

Los DLM pueden ser considerados como una generalización de los modelos de regresión lineal porque permiten la variación de los coeficientes de regresión en el tiempo (Pole, West & Harrison, 1994). De igual forma, los DLM son un caso particular de los SSM, dado que estos permiten especificar cualquier otra distribución *a priori* distinta a la de Gauss, junto con la especificación de ecuaciones no lineales:

$$Y_t = h_t(\theta_t, v_t)$$

$$\theta_t = g_t(\theta_{t-1}, w_t)$$

para cualquier par de funciones continuas

$$g_t \text{ y } h_t : R \rightarrow \text{R}$$

## *Teoría de valores extremos generalizada*

El análisis de valores extremos es crucial para comprender eventos climáticos excepcionales como sequías prolongadas o precipitaciones extremadamente intensas. La teoría de valores extremos generalizada proporciona una herramienta matemática poderosa para modelar y estimar eventos extremos en una distribución de probabilidades. Al combinar esta teoría con los modelos de espacios de estado y los modelos lineales dinámicos, podemos explorar cómo los eventos extremos de precipitación evolucionan en el tiempo y cómo se relacionan con el comportamiento general de las series temporales. El enfoque clásico para el estudio de los extremos se basa principalmente en un resultado asintótico que puede considerarse como un análogo del teorema del límite central. Antes de exponerlo, se necesitan algunas especificaciones del modelo. Sean $X_1, X_2, \ldots, X_n$ variables aleatorias independientes e idénticamente distribuidas (I.I.D.), tal que $n \in Z^+$, $X_i \in R$, para todo $i = 1,2,3, \ldots, n$; $R$ y $Z^+$denotan el conjunto de los números reales y enteros positivos, respectivamente. El objetivo principal es estudiar el comportamiento máximo de la muestra $M_n = \max\{X_1, X_2, \ldots, X_n\}$; en nuestro enfoque tomaremos la metodología de superación de umbral descrita en Coles (2001).

En este artículo, abordamos una metodología de estimación novedosa para un modelo de valor extremo con dependencia del tiempo, que es inducido por una variable latente que varía en el tiempo de un modelo de espacio de estado no gaussiano. Comenzamos con la distribución de valor extremo generalizado (GEV) dada por:

$$G(z) = exp\left\{-\left[1 + \xi\left(\frac{z-\mu}{\sigma}\right)^{\left(\frac{-1}{\xi}\right)}\right]\right\} \quad (4)$$

definido en el conjunto: $z: \left\{1 + \xi\left(\frac{z-\mu}{\sigma}\right) > 0\right\}$, donde los parámetros $\mu, \xi \in R$, $\sigma > 0$.

El modelo de la ecuación (4) se denomina distribución generalizada de valores extremos (GEV):

- μ un parámetro de ubicación,
- σ > 0 parámetro de escala,
- ξ parámetro de forma.

Coles (2001) presenta una descripción clara del modelado estadístico para valores extremos utilizando la distribución GEV. Bajo el supuesto de independencia, condicional a μ, σ y ξ. Por lo tanto, se puede realizar un análisis bayesiano imponiendo una distribución previa en μ, σ y ξ como en Coles y Tawn, (1996), «A Bayesian Analysis of Extreme Rainfall Data», y en Gaetan y Grigoletto (2004), quienes presentan un modelo con parámetros dinámicamente variables para el análisis de extremos de una serie temporal univariada. Su modelo incluye cambios dinámicos para los parámetros de escala/forma y actualización secuencial con filtros de partículas, sin una estructura de espacio o espacio-tiempo. Estudiamos el comportamiento variable en el tiempo de los parámetros que gobiernan la distribución de los datos usando modelos lineales dinámicos, como se presenta en West y Harrison (1997). Ilustramos las posibilidades de capturar tendencias variables en el tiempo. El uso de un modelo dinámico permite la detección de cambios de corto y largo alcance, también consideramos el problema de relacionar las tendencias con las covariables que varían en el tiempo. Al hacerlo, podemos describir como cambia la fuerza de la relación lineal con el tiempo.

## *Modelos GEV temporales*

Coles (2001) presenta un enfoque para modelar cambios a lo largo del tiempo en máximos utilizando la distribución GEV. Estos cambios se definen a través de funciones deterministas y pueden referirse a la tendencia o estacionalidad de los datos. Considere una secuencia de observaciones condicionalmente independientes $y_1, y_2, \dots, y_m$ tal que cada observación sigue una distribución GEV con un parámetro de ubicación variable en el tiempo, es decir, $y_t \sim GEV(\mu_t\,,\ \sigma\,, \xi)$. Para obtener $y_t$ fijamos un periodo de tiempo o bloque de datos y calculamos el máximo de las observaciones obtenidas durante el i-ésimo periodo de tiempo o bloque. Para modelar los cambios en el tiempo, Coles (2001) menciona varias posibilidades. Por ejemplo, $\mu_t = \beta_0 + \beta_1 t, \mu_t = \beta_0 + \beta_1 t + \beta_2 t^2$, los cuales son polinomios y estos pueden ser también de algún otro orden superior sobre *t*.

Además, sugiere el uso de covariables con una expresión de la forma $\mu_t = \beta_0 + \beta_1 t$.

También se pueden formular modelos no estacionarios para los parámetros de forma y/o escala. Por ejemplo,

$$\sigma_t = \exp(\beta_0 + \beta_1 t);\ \xi_t = \beta_0 + \beta_1 t\ o\ \xi_t = \beta_0 + \beta_1 t + \beta_2 t^2$$

Alternativamente, proponemos el uso de los DLM, una clase muy general de modelos de series de tiempo (West & Harrison, 1997) para modelar cambios de parámetros a lo largo del tiempo para ubicación, escala o forma. Para un modelo de ubicación variable en el tiempo, asumimos que los datos $z_1, z_2, \dots, z_m$ definen una sucesión condicionalmente independiente que supera algún umbral y $z_t \sim GEV\ (\mu_t\,, \sigma\,, \xi)$ ,por lo que la función de distribución acumulativa de cada $z_t$ es:

$$G(z) = \exp\left\{-\left[1 + \xi\left(\frac{z_t - \mu_t}{\sigma}\right)^{\left(\frac{-1}{\xi}\right)}\right]\right\} \qquad (5)$$

Esta suposición se basa en resultados asintóticos, por lo que puede considerarse una aproximación. En lugar de una función determinista en μt, suponemos que:

$$\mu_t = F_t\,\theta_t + v_t\,, v_t \sim N(0\,, V_t) \quad (6)$$

$$\theta_t = G_t\,\theta_{t-1} + w_t\,,\ \ w_t \sim N(0\,, W_t) \quad (7)$$

Donde $\theta_t$ es un vector de estado de dimensiones, $k \times 1$; $F_t$ es un «regresor» de dimensiones, $k \times 1$; $G_t$ es una matriz de evolución, $k \times k$; $V$ es la varianza Observacional y $W$ es una matriz de covarianza de evolución $k \times k$. Esto define un DLM ($F_t$, $G_t$, V, W) como se definió anteriormente. Al usar un DLM obtenemos más flexibilidad, en el sentido de que las tendencias no están restringidas a tener una forma paramétrica específica. Podemos evaluar la importancia de los cambios a corto plazo junto con los de largo plazo. Además, es posible cuantificar como los efectos debidos a las covariables cambian con el tiempo.

## *Especificación del modelo*

Sea $Z = z_1, z_2, \dots, z_n$ una sucesión de valores extremos. Definimos el modelo GEV mediante DLM, es decir, un modelo GEV-DLM, para cada $t$ sea $\mu_t = \mu_1, \mu_2, \dots, \mu_t$ mediante las ecuaciones (6) y (7).

Se supone que la variable de estado $\mu_t$ las tomaremos del FFBS partiendo de la ecuación (6). Además, tomamos que F es un vector de 1x1 y G es la matriz de convolución de 1x1. Suponemos que el error de observación $v_t \sim N\,(0\,, \sigma_v^2)$ y el error de evolución $w_t \sim N\,(0\,, \sigma_w^2)$ para $t = 1, \dots, m$. Los parámetros del modelo podemos estimarlos mediante la inferencia y la estructura posteriores bajo nuestra nueva distribución de GEV, siguiendo los métodos estándar de la cadena Markov Chain Monte Carlo (MCMC) (ver, por ejemplo, Lindsten, 2013). Describimos brevemente las distribuciones condicionales completas relevantes.

Una forma simple de ajustar el modelo es extraer una muestra para la posterior marginal de $v_t\,\mu_1, \mu_2, \dots, \mu_t |\, Z\,, \sigma, \xi$ mediante el algoritmo de Forward Filtering Backward Sampling (FFBS), asumiendo una aproximación normal y ajustando la media *a posteriori*. Los parámetros de escala y forma se simulan a partir de $p(\sigma|Z, \mu, \xi)$ y $p(\xi|Z, \mu, \xi)$ con pasos individuales de Metropolis-Hastings. Para generar valores iniciales utiliza-

remos una densidad log-normal para el parámetro de forma y una normal estándar para el parámetro de escala $\sigma \sim Halt - t(h), \xi \sim N(0, \sigma_\xi^2)$. Es importante resaltar que la student-t debe solo tomar valores positivos, ya que sigma es positiva, y estas densidades definidas en la mitad de su dominio se les conoce como densidades medias (half–student–t).

Los parámetros de escala y forma de la distribución GEV se suponen constantes en el tiempo.

La varianza observacional V se puede muestrear a partir de una distribución gamma inversa.

La matriz de covarianza de evolución W se puede modelar con factores de descuento o como funciones de la varianza observacional V. En el segundo caso, la evolución se puede muestrear con una distribución Inversa-Gamma o Inversa-Wishart.

Previo al ajuste del Metrópolis es necesario definir las prioris y los valores iniciales con las que se iniciaran las cadenas de Markov. En este caso, se utilizará una densidad t–student con h grados de libertad para el parámetro de escala del modelo y una densidad normal estándar para el parámetro de forma.

## *Modelos GEV espaciales*

Podemos extender el modelo presentado en la sección anterior a datos en el espacio y el tiempo. Basamos nuestro enfoque en convoluciones de proceso como en Higdon (2004). De hecho, las convoluciones de procesos proporcionan una representación lineal conveniente de los procesos gaussianos. Considere que en el tiempo $t$ tenemos un vector $z_t = z_{1t}, \dots, z_{1n}$, en el cual se registraron observaciones en los sitios $s_{1t}, \dots, s_{nt}$.

Una formulación de espacio-tiempo para $z_t$ se define como:

$$z_t = K^t x_t + \epsilon_t \quad (8)$$

$$x_t = x_{t-1} + v_t \quad (9)$$

Donde $K^t$ es una matriz $n_t \times k$ dada por $K_t = (s_i - w_j) \;\; t = 1, \dots, k$, donde k es un núcleo dado. Este modelo se incluye dentro de la clase de DLM utilizados en apartados anteriores para el modelado temporal. Suponemos que el error de observación $\epsilon_t \sim N(0, \sigma_\epsilon^2 I_K)$ y el error de evolución $v_t \sim N(0, \sigma_v^2 I_k)$ para $t = 1, \dots m$. Para el vector de estado $x_0 \sim N(0, \sigma_x^2 I_k)$.

Como se mencionó anteriormente, los elementos $k(-\omega_j)$ están definidos por un núcleo de suavizado, donde $\omega_1, \omega_2, \dots, \omega_k$ son los sitios espaciales o nudos donde se centran los núcleos (pesos espaciales). Dado que los términos de error no están correlacionados espacialmente, la dependencia espacial de este modelo está completamente definida por $k$. Algunos núcleos recomendados son:

- Gaussiano: $k(s) \propto \exp\{-||s||^2/2\eta^2\}$
- Exponencial: $k(s) \propto \exp\left\{-\frac{||s||}{\eta^2}\right\}$
- Esférica: $k(s) \propto \left(1 - \frac{||s||^3}{r^3}\right) I[s \geq r]$

Hacer una convolución en $\omega_1, \omega_2, \dots, \omega_k$ proporciona una aproximación finita a la convolución continua. Para la convolución integral es posible encontrar una equivalencia entre los núcleos y la función de covarianza del proceso. Ver Higdon (2004) para más detalles.

Para fusionar el DLM espacial con el análisis de los valores extremos, suponga que $Z_{s,t} \sim GEV(\mu_{s,t}, \sigma, \xi)$ para $s = 1, \dots S$ y $t = 1, \dots, m$. La función de distribución para $z_{s,t}$ es:

$$G_{s,t}(z_{s,t}) = \exp\left\{-\left[1 + \xi\left(\frac{z_t - \mu_t}{\sigma}\right)^{\left(\frac{-1}{\xi}\right)}\right]\right\} \quad (10)$$

Para cada $t$ sea $\mu_t = \mu_{1,t}, \mu_{2,t}, \dots, \mu_{s,t}$. Ahora definimos un DLM en μt utilizando el enfoque de convolución del proceso descrito anteriormente.

$$\mu_t = K^t \theta_t + \epsilon_t \quad (11)$$

$$\theta_t = \theta_{t-1} + v_t \quad (12)$$

donde el vector de estado $\theta_t = \theta_{t,1}, \dots, \theta_{t,k}$, $\epsilon_t = \epsilon_{t,1}, \dots, \epsilon_{t,k}$, $\theta_t \; v_t = v_{t,1}, \dots, v_{t,k}$ Además, $\epsilon_t \sim N(0, \sigma_\epsilon^2 I_s)$ y $v_t \sim N(0, \sigma_v^2 I_k)$. Como ejemplo si se usa un núcleo gaussiano tenemos que la matriz $s \times k$ donde $k$ viene dada por: $K_{ij} = \exp\{-d \left||s_i - s_j|\right|^2/2\}$, donde $s_i$ es la posición de la estación $i = 1, \dots, k$; $s_i$ es la posición de la estación $j = 1, \dots, k$ y $\forall$ i, j. Si se usa un núcleo gaussiano y pesos espaciales tenemos que la matriz $s \times k$ , donde $k$ viene dada por: $K_{ij} = \exp\{-d ||\omega_i - s_i||^2/2\}$, donde

$\omega_i$ es la posición del peso más cercano a la estación $s_i$, $i = 1, \dots, k$.

Para los parámetros en la primera etapa del modelo especificamos las distribuciones *prioris* $p(\sigma) \sim half - t(5)$ y $p(\xi) \sim N(\mu_\xi, s_\xi)$. Para los parámetros del DLM $\theta_0 \sim N(0, \sigma_\theta^2 I_k)$; $\sigma_\epsilon^2 \sim IG(\alpha_\epsilon, \beta_\epsilon)$; $\sigma_v^2 \sim IG(\alpha_v, \beta_v)$; $\sigma_\theta^2 \sim IG(\alpha_\theta, \beta_\theta)$; $\sigma_\eta^2 \sim IG(\alpha_\eta, \beta_\eta)$.

La inferencia posterior y la simulación para este modelo de espacio-tiempo siguen una estructura similar a la del modelo variable en el tiempo GEV. Las distribuciones para $\mu_t, \sigma$ y $\xi$ se muestrean cada uno con un paso Metrópolis-Hastings. Generamos el vector $\mu_t$ con simulación hacia atrás de filtrado directo FFBS. Para $\mu_t$ $\sigma_\epsilon^2$; $\sigma_v^2$; $\sigma_\theta^2$; $\sigma_\eta^2$ se muestrean con distribuciones de Gamma-Inversa (*IG*). Normalmente, para las convoluciones de procesos, los núcleos se centran alrededor de puntos en una cuadrícula regular y el parámetro de rango se fija como $d = c\phi$, donde $\phi$ es la distancia entre los puntos de la cuadrícula y c es una constante, generalmente entre 1/2 y 2.

Cabe resaltar que la estadística espacial con procesos de convolución proporciona una metodología para analizar patrones espaciales complejos en datos geoespaciales. Al combinar esta técnica con los otros enfoques mencionados, podemos identificar áreas geográficas con patrones de precipitación extremos, evaluar su evolución en el tiempo y estimar las distribuciones de probabilidades para eventos extremos en ubicaciones específicas. Ampliamos el modelo para permitir la variación espacial de los parámetros. Nuestro modelo espacio-temporal para extremos se basa en convoluciones de procesos utilizando métodos presentados, por ejemplo, en Higdon (2004). Esto proporciona una reducción del espacio de parámetros que es eficaz en el manejo de un gran número de ubicaciones. En la sección final presentamos una discusión de nuestros resultados.

## *Estudio de simulación ilustrativo*

Para la primera etapa de nuestro análisis no estamos tomando en cuenta la parte espacial, solamente la estimación de estados mediante el DLM usando FBBS para el parámetro μ.

En este caso, ajusta un modelo que sigue una distribución GEV con una estructura lineal temporal para medir las dependencias entre los datos a través del tiempo. Para eso modificaremos el modelo original agregando una ecuación adicional al parámetro de locación para que esta se adapte a la estructura temporal $y_t \sim GEV\ (\mu_t, \sigma, \xi)$ donde $\mu_t$ parte del modelo lineal dinámico $\xi \sim N(0,1)$ y haremos un pequeño cambio para $\sigma$ tomando $\sigma \sim half - t(5)$ para mantener $\sigma$ positivo, esto por definición. Para simular los datos, primero simulamos las locaciones en cada tiempo usando un DLM $(F = 1, G = 1, V = 1, W = 0.1)$ constante. Luego, extraemos los valores de $\mu_t$ para simular las observaciones mediante la ecuación anterior. Los parámetros de escala y forma se simularon con variables log normal y normal estándar, respectivamente.

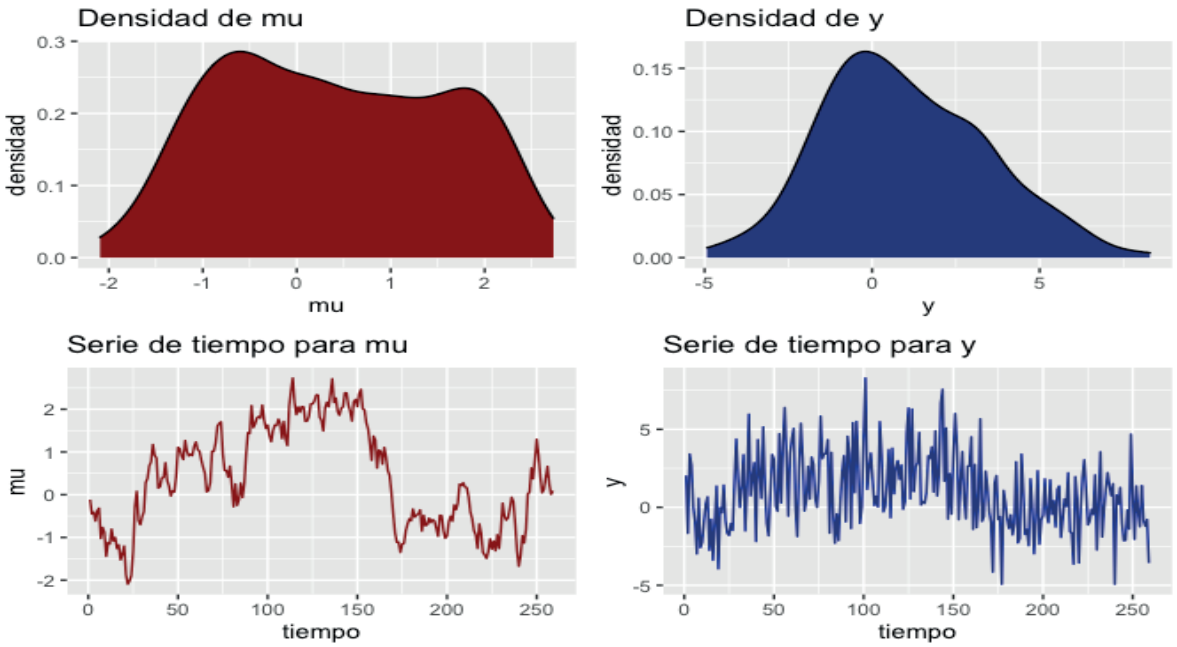
En la Figura 1 vemos el muestreo para el parámetro mediante FFBS partiendo del DLM, en esta parte se observa que el modelo no presenta estacionalidad ni tendencia, además mostramos los gráficos de los datos correspondientes.

Partiendo de los datos obtenidos de este muestreo se presentan los resultados mediante el DLM, tomamos como valor la media de estos datos para llevarlo al Metropolis-Hasting y así tener un valor de estimación más preciso.

Para ajustar el modelo se extrae una muestra para la posterior marginal de $\mu_1, \mu_2, \dots, \mu_n | y$ mediante el algoritmo de Forward Filtering Backward Sampling, asumiendo una aproximación normal, ajustar la media *a posteriori* y ajustar un algoritmo de Metrópolis-Hastings para los otros dos parámetros restantes. Previo al ajuste del metrópolis es necesario definir las prioris y los valores iniciales con los que se iniciaran las cadenas de Markov. En este caso, se utilizará una densidad half–student–t con 3 grados de libertad para la escala del modelo y una densidad normal estándar. Para generar valores iniciales utilizaremos una densidad log-normal para la escala y una normal estándar para la forma. Ahora corremos el metrópolis-hastings para los parámetros faltantes, pero con los datos filtrados por la estimación posterior de $\mu_1, \mu_2, \dots, \mu_n | y$. En este caso se corren 4 cadenas de 5000 iteraciones por cadena, donde las primeras 2500 iteraciones se descartan por warm-up.

En la Figura 2 observamos que las densidades y las trazas de las cadenas en ambos casos no se distinguen comportamientos anómalos, por lo tanto, aceptamos el modelo.

Figura 1. Muestreo parámetro de ubicación para la estructura temporal del modelo




Seguidamente, realizamos posterior *predictive checks* para revisar el ajuste de nuestro modelo, el siguiente algoritmo genera muestras de nuestro modelo usando las posterioris de nuestros parámetros y la media a posterior para las locaciones. Por costos computaciones, solo utilizaremos una muestra de 500, que extraemos del MCMC para hacer revisar el ajuste (Lynch & Western, 2004).

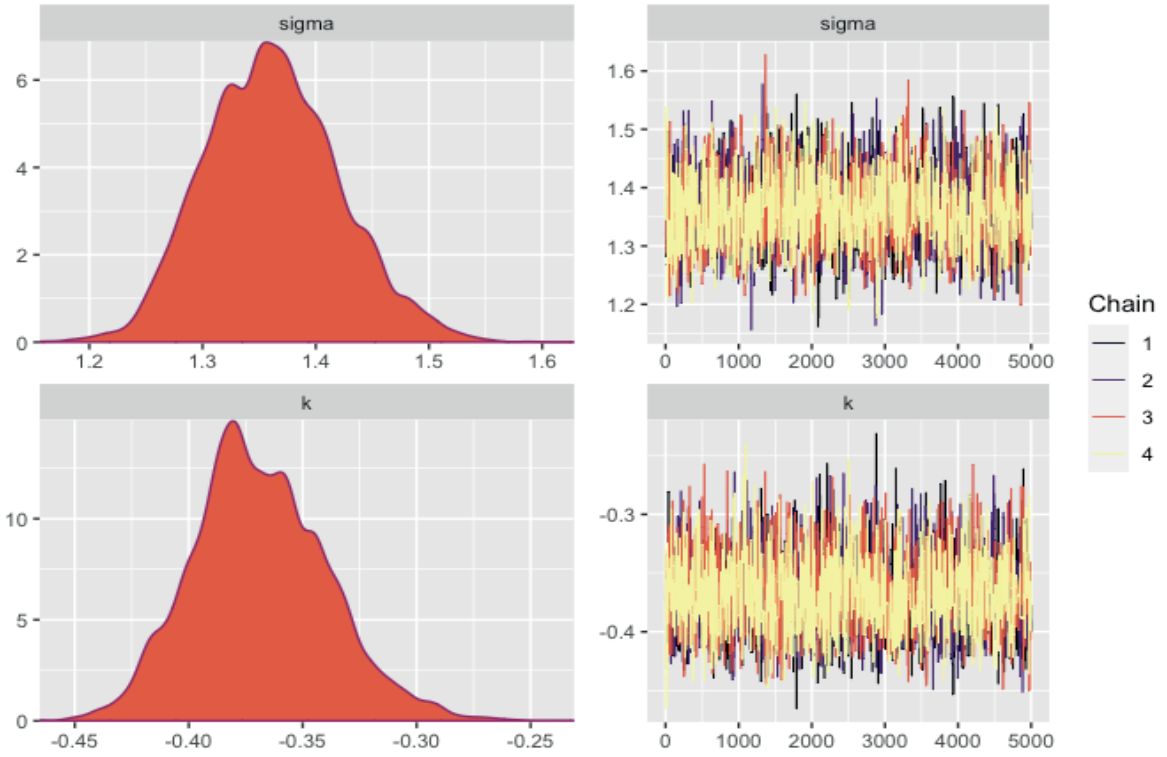
En la Figura 3 vemos que el modelo muestra una recuperación algo débil de los datos, pero se pudo recuperar el comportamiento temporal de los datos, es decir, que recuperamos bien el parámetro en su parte temporal mediante FFBS.

## *Aplicación del modelo en valores de precipitaciones en el estado de Washington, EE. UU.*

Considere el problema de modelar los valores que superen el umbral, para ello, en este trabajo se tomaron los datos del estado de Washington (Estados Unidos), tomando valores de las precipitaciones de 18 estaciones meteorológicas. Los datos de las estaciones fueron proporcionados por la Administración Nacional Oceánica y Atmosférica (NOAA, por sus siglas en inglés), agradecemos su valiosa colaboración por proporcionar los datos necesarios para este trabajo. De los datos completos de las precipitaciones diarias por hora, seleccionamos las estaciones que tienen datos desde agosto de 2002 a agosto de 2022. Para cada estación seleccionada se tomaron registros diarios en un periodo de 20 años.

El estado de Washington es nuestro punto de enfoque. Contiene los datos de 50 estaciones meteorológicas, sin embargo, buscamos la misma cantidad de datos en un mismo periodo de tiempo y tenemos que 32 estaciones no proporcionan los datos de manera correspondiente, es decir, muchas de ellas sus periodos de tiempos no eran uniformes entre sí. Por ello, se tomaron solamente 18 estaciones en las cuales el periodo de tiempo de observación de datos entre estaciones era correspondiente, es decir, mismo periodo de tiempo para todas. Cada una de las estaciones proporciona un total de 414 mediciones por año, es decir, 8208 mediciones para cada una de las 18 estaciones mencionadas. Es importante recordar en nuestro análisis que no en todos los días del año hay precipitaciones, por lo tanto, resulta relevante analizar los días en donde las mismas son más elevadas.

Figura 2. Densidad de los parámetros de forma y escala



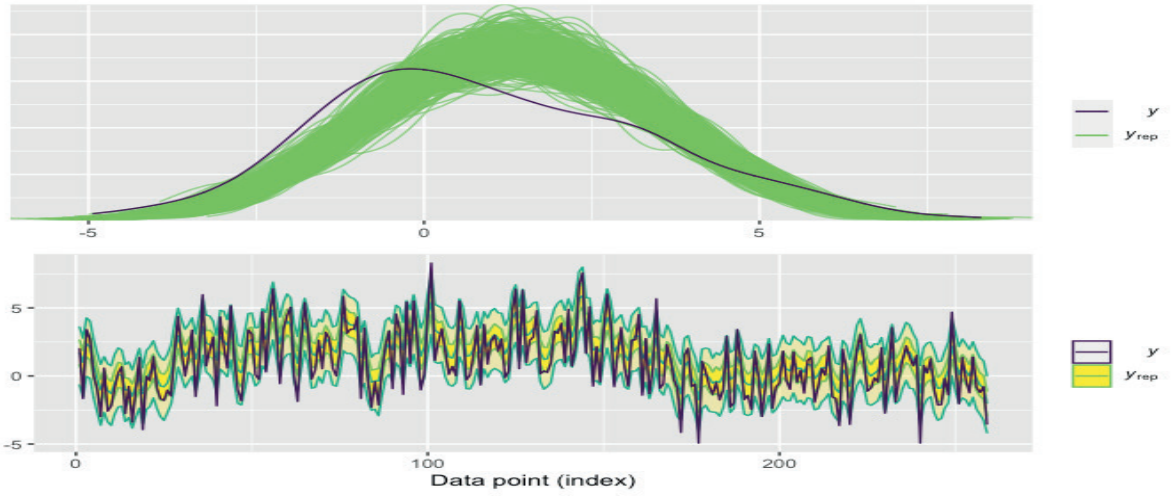

Para ello, hacemos uso del cálculo del umbral adecuado, este se tomó haciendo uso del gráfico de vida media residual, ese umbral proporcionará los valores de excedencia para cada estación, ya que los que se toman son los que superen el umbral seleccionado. En este punto cada estación tiene un número diferente de registros, ahora nos interesan únicamente aquellos valores que exceden el umbral, recordemos que los datos varían según la estación que estamos estudiando, lo que significa que también el umbral cambiará a medida escogemos otra estación.

Naturalmente, podemos extender el modelo presentado en la sección anterior a datos en el espacio y el tiempo, esto podemos hacerlo usando procesos de convolución como se muestra en Higdon (2004). De hecho, las convoluciones de procesos proporcionan una representación lineal conveniente de los procesos gaussianos. Considere que en el tiempo t tenemos un vector $y_t = y_{1,t}, \ldots, y_{n_t,t}$para el cual se registraron observaciones en los sitios $y_t\, s_1, \ldots, s_{n_t}$. Definimos el modelo espacio-tiempo para $y_t$ mediante (11) y (12). En geoestadística, los pesos espaciales son una herramienta fundamental para modelar la dependencia espacial entre observaciones en un conjunto de datos georreferenciados. Uno de los enfoques es usar pesos espaciales que representan la influencia relativa que tiene cada ubicación sobre las ubicaciones circundantes, y se utilizan para realizar interpolación espacial, estimar valores en ubicaciones no muestreadas y analizar patrones espaciales.

Los pesos espaciales se pueden definir de diferentes maneras, dependiendo del enfoque y del tipo de dependencia espacial que se desea capturar. A continuación, se presentan algunas formas comunes de definir los pesos espaciales:

- *Pesos basados en la distancia.* Se asignan pesos inversamente proporcionales a la distancia entre las ubicaciones. Las ubicaciones más cercanas tienen mayores pesos, lo que indica una mayor influencia en la estimación o interpolación espacial.

Figura 3. Validación del parámetro de ubicación vía posterior *predictive check*




Ejemplos de métodos que utilizan pesos espaciales basados en la distancia incluyen la interpolación por Kriging y el método de los vecinos más cercanos (Cressie, 1992).

- *Pesos basados en la vecindad.* Se definen pesos según la relación de vecindad entre las ubicaciones. Las ubicaciones vecinas tienen mayores pesos, mientras que las ubicaciones no vecinas tienen pesos cercanos a cero. Los pesos espaciales basados en la vecindad se utilizan en métodos como la interpolación por polígonos de Thiessen y los modelos de auto-regresión espacial (SAR, por sus siglas en inglés) (Bivand, Pebesma & Gómez-Rubio, 2013).
- *Pesos basados en funciones de influencia.* Se utilizan funciones de influencia espacial para asignar pesos a las ubicaciones en función de su relación espacial. Estas funciones pueden ser basadas en la distancia, la dirección o cualquier otra medida de proximidad espacial. Los pesos espaciales basados en funciones de influencia son comúnmente utilizados en métodos como la regresión espacial y los modelos de tendencia espacial (Chou, 1992).

Es importante destacar que la elección de los pesos espaciales depende del objetivo y de las características específicas del conjunto de datos y del fenómeno estudiado. Diferentes métodos y enfoques pueden dar lugar a diferentes pesos espaciales y, por lo tanto, a diferentes resultados en el análisis espacial. Para nuestro análisis nos enfocaremos en desarrollar un proceso de convolución analizando la longitud y latitud por estación meteorológica, lo cual permitirá crear una matriz de autocorrelación para analizar de mejor manera la dependencia espacial del modelo.

En la Figura 4 representamos las 18 estaciones para el análisis de la parte espacial del modelo.

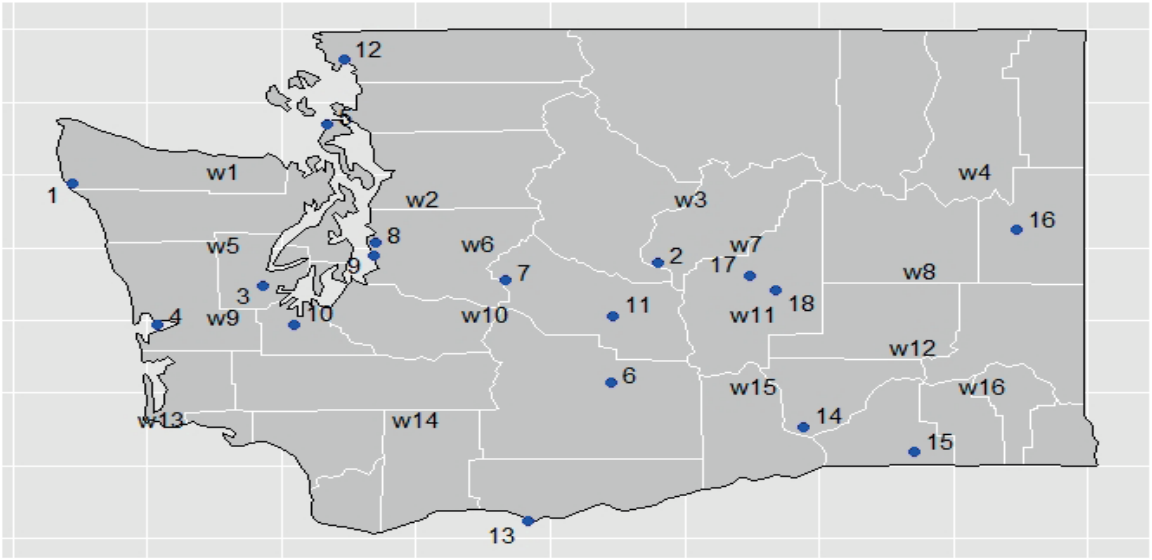
En la Figura 5 mostramos que el modelo genera una buena convergencia para la estimación de parámetros, lo cual hace que validemos el método.

La Figura 6 muestra cómo el modelo reproduce adecuadamente la estructura temporal de los datos correspondientes. Además, la validación cruzada indica que el modelo replica suficientemente bien los datos, lo que valida su efectividad.

## *Discusión*

En conjunto, el enfoque de modelos GEV temporales y espaciales con modelos lineales dinámicos brinda una herramienta poderosa y flexible para analizar, comprender y predecir comportamientos en datos extremos en diferentes contextos. La

Figura 4. Mapa del estado de Washington con sus respectivas estaciones meteorológicas




combinación de técnicas de modelado, inferencia y simulación proporciona una metodología sólida para investigar tendencias, patrones y riesgos en fenómenos extremos, contribuyendo así al avance del conocimiento en diversas áreas de estudio. El trabajo presentado es un marco general para modelar valores extremos que exhiben un comportamiento variable en el tiempo. Los modelos se basan en el uso de la distribución GEV y asumen que el parámetro de ubicación varía con el tiempo según un DLM. El enfoque se puede ampliar fácilmente para tener en cuenta la variabilidad espacial mediante el uso de convoluciones de proceso. Usamos validación cruzada y posterior *predictive check* para validar el modelo. El uso de la distribución GEV hace que la inferencia sea sencilla, dentro de un enfoque MCMC, todo lo que se necesita es la evaluación de la función del GEV para cada muestra conjunta de los parámetros. Por lo tanto, las estadísticas de verificación de modelos son fáciles de calcular. En el caso espacio-temporal, podemos considerar parámetros de escala y forma que varían con la ubicación. Nos gustaría mencionar que se hizo un análisis por estación dando resultados favorables para cada una de ellas, para ello, el FFBS fue fundamental para obtener la réplica en el parámetro de ubicación, que es el que corresponde a dependencia espacio-temporal.

## *Agradecimientos*

Agradezco a los autores en su asesoría para poder desarrollar este documento. La ayuda tanto a Asael Alonzo y Cristian Cruz por su colaboración en la comprensión y desarrollo de elementos importantes de la teoría de valores extremos, modelos lineales dinámicos y la estimación bayesiana del método.

Figura 5. Densidad para los parámetros de forma y escala de la estación número 4

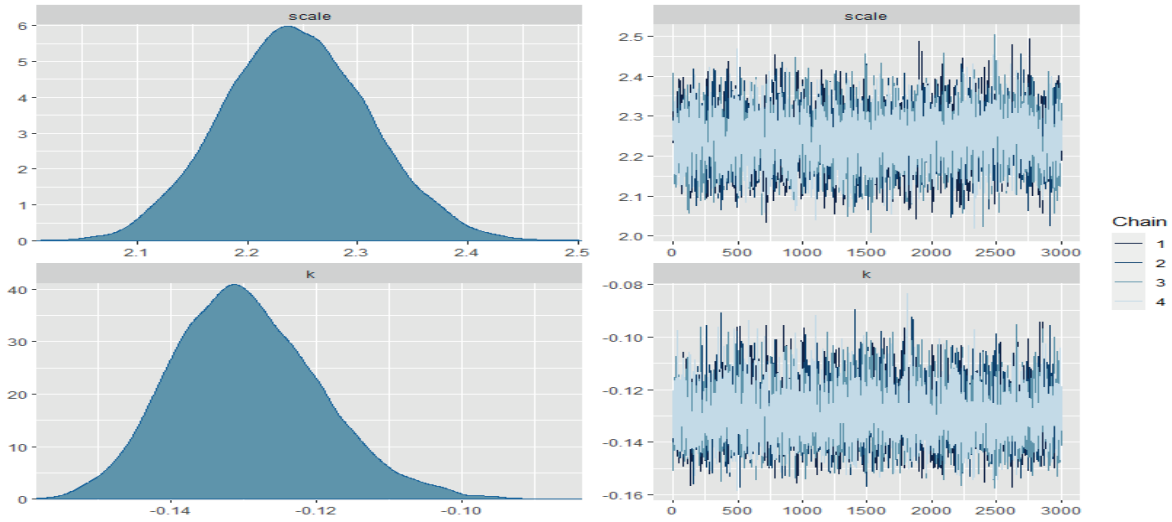
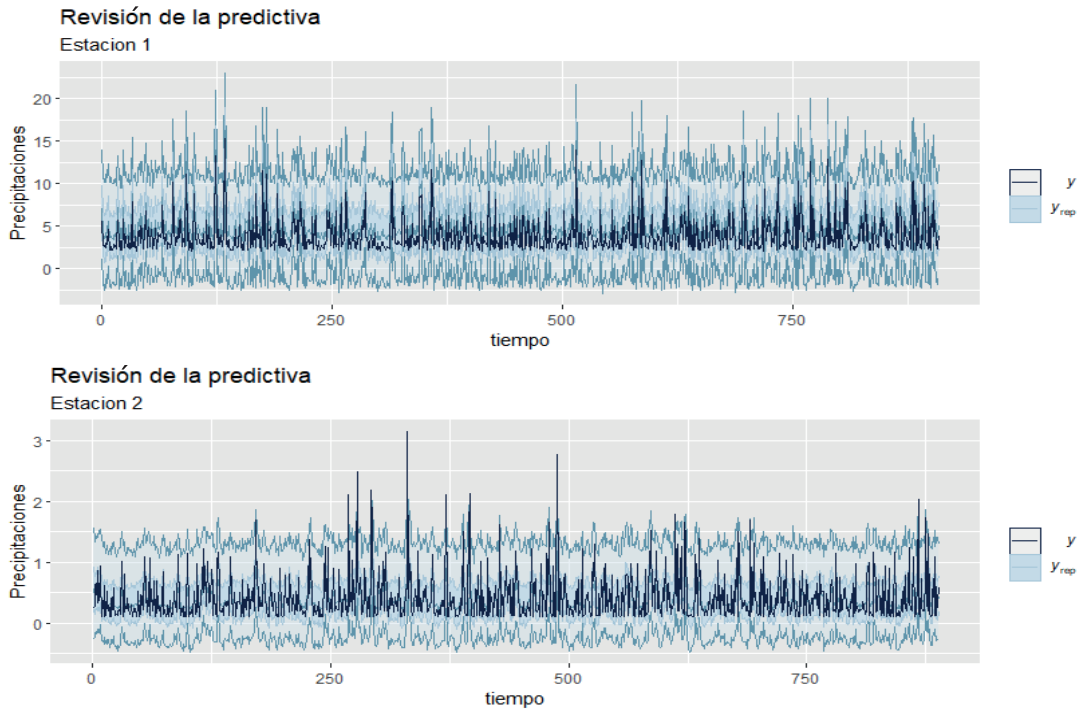



Figura 6. Validación mediante posterior *predictive check* para las estaciones 1 y 2



## *Referencias bibliográficas*


Bivand, R., Pebesma, E., & Gómez-Rubio, V. (2013). Applied Spatial Data Analysis with R. doi: 10.1007/978-1-4614-7618-4

Chou, Y. H. (1992). Spatial autocorrelation analysis and weighting functions in the distribution of wildland fires. *International Journal of Wildland* Fire. doi: 10.1071/wf9920169

Coles, S. (2001). *An introduction to statistical modeling of extreme values.* Springer. doi:10.1007/978-1-4471-3675-0

Coles, S., & Tawn, J. (1996). A Bayesian analysis of extreme rainfall data. *Applied statistics, 45*(4), 463. doi: 10.2307/2986068

Cressie, N. (1992). Statistics for spatial data. *Terra Nova, 4*(5), 613-617. doi: 10.1111/j.1365-3121.1992.tb00605.x

Frühwirth-Schnatter, S. (1994). Data augmentation and dynamic linear models. *Journal of Time Series Analysis, 15,* 183-202. doi: 10.1111/j.1467-9892.1994.tb00184.x

Gaetan, C., & Grigoletto, M. (2004). Smoothing Sample Extremes with Dynamic Models. *Extremes, 7*(3), 221-236. doi:10.1007/s10687-005-6474-7

Higdon, Dave. (2004). En Space and Space-Time Modeling using Process Convolutions, pp. 37-56. Springer. doi:10.1007/978-1-4471-0657-9\

Huerta, G., Sansó, B., & Stroud, J. (2004). A spatiotemporal model for Mexico City ozone levels. *Applied Statistics, 53*(2), 231-248. doi: 10.1046/j.1467-9876.2003.05100.x

Lindsten, F. (2013). *Backward simulation methods for Monte Carlo statistical inference.* Illustrated edn. Now Publishers Inc. doi:10.1561/9781601986993

Lynch, S., & Western, B. (2004). Bayesian posterior predictive checks for complex models. *Sociological Methods & Research, 32*(3), 301-335. doi: 10.1177/0049124103257303

Metropolis, N., Rosenbluth, A., Rosenbluth, M., Teller, A., & Teller, E. (1953). Equation of state calculations by fast computing machines. *Journal of Chemical Physics, 23*(6), 1087-1092. doi: 10.1063/1.1699114

Pole, A., West, M., & Harrison, J. (1994). Applied Bayesian Forecasting and time series analysis. Chadman and Hall. doi: 10.1007/978-1-4899-3432-1

Strawderman, R., & Friel, N. (2000). Markov Chain Monte Carlo: Stochastic simulation for Bayesian inference. *Journal of the American Statistical Association, 95*(449), 346. doi: 10.2307/2669581

Triantafyllopoulos, K. (2021). *Bayesian Inference of State Space Models.* Springer.

Vehtari, a., Gelman , A., & Gabry, J. (2016). Practical Bayesian model evaluation using leave-one-out cross-validation and **WAIC**. *Statistics and Computing, 27*(5), 1413-1432. doi: 10.1007/s11222-016-9696-4

West, M., & Harrison, J. (1997). Bayesian Forecasting and dynamic models. Springer. doi: 10.1007/b98971