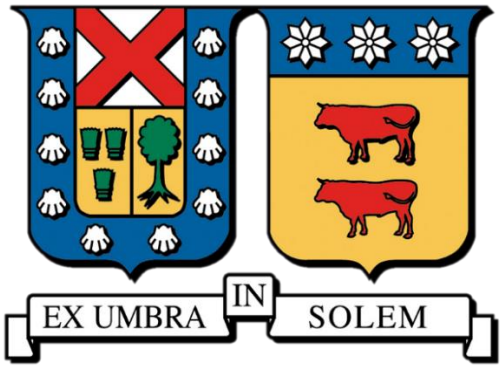


UNIVERSIDAD TECNICA FEDERICO SANTA MARÍA

DEPARTAMENTO DE OBRAS CIVILES

NUEVA METODOLOGIA PARA LA ESTIMACION DE LA PRECIPITACION MAXIMA PROBABLE, UTILIZANDO LA DISTRIBUCION GENERALIZADA DE VALORES EXTREMOS Y ANALISIS REGIONAL

Memoria de título y tesis de grado presentado por:

IGNACIO ANDRES VILLAVICENCIO ARRIAZA

Como requisito parcial para optar al título de

Ingeniero civil

y al grado de

Magister en Ciencias de la Ingeniería Civil

Profesores Guía:

Dr. ALVARO OSSANDON

Dr. JOAQUIN MEZA

Valparaiso, mayo de 2025

CONSTANCIA DE VALIDACIÓN Y CONFIDENCIALIDAD DE MONOGRAFÍA A REPOSITORIO ACADÉMICO

1.- IDENTIFICACIÓN DEL TRABAJO ACADÉMICO

Tipo de monografía (marcar una opción): ☐ Memoria o trabajo de título; ☒ Tesis de Postgrado;

Título del trabajo: Nueva Metodologia para la Estimacion de la Precipitacion Maxima Probable, Utilizando Distribucion Generalizada de Valores Extremos y Analisis Regional

Nombre del candidato(a): Ignacio Andres Villavicencio Arriaza

Carrera / Grado: Magister en Ciencias de la Ingeneiria Civil

Campus: Casa Central Valparaiso **;** **Departamento**: Obras Civiles

2.- VALIDACIÓN DEL PROFESOR GUÍA/DIRECTOR DE TESIS

Yo, Alvaro Ossandón, en mi calidad de profesor(a) guía/director(a) del trabajo académico mencionado anteriormente **DEJO CONSTANCIA** que:

- He revisado esta versión del documento y corresponde a la versión final aprobada del trabajo.
- El trabajo cumple con los requisitos académicos y de formato establecidos por la institución

3.- EVALUACIÓN DE CONFIDENCIALIDAD POR PROPIEDAD INDUSTRIAL

☒ El trabajo **NO contiene información que amerite confidencialidad** y puede ser publicado de inmediato en repositorio con acceso abierto.

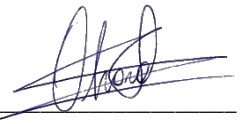
☐ El trabajo **CONTIENE** información con potenciales implicancias de propiedad industrial o intelectual y requiere un periodo de confidencialidad (embargo) por:

☐ 6 meses; ☐ 12 meses; ☐ 2 años; ☐ 3 años; ☐ 5 años; ☐ 10 años

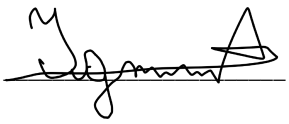
Fundamentación de la necesidad de confidencialidad (obligatorio si se solicita embargo):

4.- FIRMAS

Profesor(a) guía o director(a) de memoria o tesis:

Fecha: 17-06-2025 ; **Firma:** 

Estudiante o Candidato(a):

Fecha: 17-06-2025 ; **Firma:** 

Este formulario debe ser insertado como página 2 de la memoria o tesis, completado y firmado por estudiante y profesor(a) antes de la entrega en portal PRISMA de Biblioteca USM.

Agradecimientos

Quiero agradecer infinitamente a mi tata Juan, mi mami Lola, mi madre, a mi papá Juanito y mi tío José, los cuales fueron las personas que están desde el principio de cualquier proceso que haya experimentado y me han apoyado en cada situación complicada que se ha presentado en el proceso. También agradecer a mis hermanos y mi tía, que en los últimos años fueron un cable a tierra en mi proceso final de Valparaíso.

Agradecer también a mis amigos de toda la vida y los que hice en la universidad, los cuales sin lugar a duda aportaron de una u otra forma para lograr el objetivo de completar mi proceso universitario. Finalmente agradecer a mis profesores guía Álvaro Ossandón y Joaquín Meza, los cuales me han guiado con mucha paciencia y entusiasmo en este último proceso universitario, ayudándome a encontrar mi siguiente objetivo como profesional.

Resumen

Una estimación precisa de la Precipitación Máxima Probable (PMP) es fundamental para diseñar infraestructuras hidráulicas resilientes y minimizar los riesgos hidrológicos. Los enfoques estadísticos tradicionales, como lo es el ampliamente utilizado método de Hershfield, se basan en supuestos simplificados que no necesariamente caracterizan correctamente los eventos extremos de precipitación, las cuales presentan una naturaleza asimétrica de los datos. Este tipo de simplificaciones frecuentemente conducen a sesgos sistemáticos, lo cual lleva a una sobreestimación en zonas húmedas el valor de PMP y subestimaciones en zonas áridas. Este estudio propone un método robusto para la estimación de la PMP, que integra un modelo de distribución Generalizada de Valores Extremos (GEV por sus siglas en inglés) y métodos de agrupamientos, con el fin de abordar las limitaciones de los métodos tradicionales. Esto consiste en estimar los parámetros de la distribución GEV mediante método de "L-moments", corregir el parámetro de forma en función de la longitud de registro para finalmente en estimar el resto de los parámetros. Posterior a esto, el algoritmo "K-means" (KMA por sus siglas en inglés) divide el área de estudio en regiones de precipitaciones homogéneas, basadas en atributos geográficos (Latitud y Altura) y estadísticos, lo cual reduce la influencia de los valores atípicos y proporciona estimaciones de PMP con mayor sentido físico. El caso de estudio de este análisis es Chile continental, el cual tiene el desafío de ser un país con una variabilidad climática latitudinal y longitudinal relevante, lo que constituye contexto idóneo para verificar la aplicabilidad y robustez del método propuesto. Referente a la corrección de sesgo, se redujo sustancialmente la variabilidad del parámetro de forma de la distribución GEV en estaciones que presentan longitudes de registro cortos, lo cual mejora la estabilidad en las estimaciones de valores extremos usando el modelo estadístico. El KMA logra delimitar seis regiones coherentes, cada una con patrones de precipitación distintos. En cuanto a la comparación con el enfoque tradicional de Hershfield, el método propuesto produjo estimaciones de PMP más consistentes espacialmente, particularmente en áreas áridas o semiáridas con alta asimetría. En general, el nuevo método evitó las sobreestimaciones extremas que se observan frecuentemente con los enfoques estadísticos sin corrección. Al combinar el modelado GEV corregido de sesgos con el agrupamiento regional y distribuciones con límites superiores, este estudio ofrece una metodología más robusta para la estimación de la PMP. La alineación verificable con patrones climáticos establecidos, junto con una reducción sustancial de sobreestimaciones extremas, subraya su aplicabilidad en diversos contextos hidrológicos.

Abstract

Accurate estimation of Probable Maximum Precipitation (PMP) is paramount for designing resilient hydraulic infrastructure and minimizing hydrological risks. Traditional statistical approaches—such as the widely used Hershfield method—rely on simplified assumptions that can misrepresent heavy-tailed extreme precipitation events and their asymmetric nature. Such simplifications frequently lead to systematic biases, overestimating in humid zones and underestimating in arid areas. This study proposes a robust PMP estimation framework integrating bias-corrected Generalized Extreme Value (GEV) modeling with a data-driven regionalization approach to address these shortcomings. Specifically, GEV parameters are derived via L-moments and adjusted for short-record biases using a shape-parameter correction. The K-means algorithm (KMA) then partitions the study area into homogeneous precipitation regions based on geographic and statistical attributes, reducing outliers' influence and providing more physically coherent PMP estimates. This methodology was tested in Chile, a country spanning marked latitudinal and climatic variability. The bias correction substantially reduced the variability of the GEV shape parameter for short-record length stations, improving stability in extreme-value estimates. KMA delineated six coherent regions, each exhibiting distinct precipitation patterns. Compared to the traditional Hershfield approach, the proposed framework produced more consistent PMP estimates, particularly in arid or semi-arid areas with high skewness. Overall, the new method avoided the extreme overestimates frequently observed with uncorrected statistical approaches. By coupling bias-corrected GEV modeling with regional clustering and upper-bounded distributions, this study offers a robust framework for PMP estimation. The verifiable alignment with established climatic patterns, together with substantially fewer extreme overestimates, underscores its applicability for diverse hydrological contexts.

INDEX

FIGURE INDEX

TABLE INDEX

1. Introduction

The design and safety of critical infrastructure —such as dams, spillways, tailings dams, flood defenses, and nuclear facilities —rely heavily on accurate assessments of extreme precipitation events. Failures in this infrastructure can lead to catastrophic consequences, including loss of life, environmental damage, and substantial economic costs (Salas et al., 2020; Panday et al., 2023). The increased frequency and intensity of extreme weather events projected under climate change scenarios (Planton et al., 2008; Kunkel et al., 2013; O'Gorman, 2015; Stott, 2016; Papalexiou & Montanari, 2019) are expected to heighten infrastructure vulnerability. In this context, the Probable Maximum Precipitation (PMP) approach —defined by the World Meteorological Organization (WMO) as the "greatest depth of precipitation for a given duration meteorologically possible for a particular location and time of year" (WMO, 2009) —is widely used to quantify extreme precipitation for infrastructure design (Hansen, 1987; Papalexiou & Koutsoyiannis, 2006; Kulkarni et al., 2010; Stratz & Hossain, 2014). PMP serves as a critical parameter for designing hydraulic structures. However, estimating PMP remains challenging, particularly in regions with limited data availability or important hydroclimatic variability.

PMP estimation methods can be broadly classified into hydrometeorological and statistical approaches. Hydrometeorological methods are mainly based on the physical mechanisms of precipitation and include the local, transposition, combination, inferential, and generalized methods (WMO, 2009; Stöwhas, 2016; Salas et al., 2020). While hydrometeorological methods are grounded in physical principles, their implementation can be challenging due to the need for extensive meteorological data and specialized expertise. This is particularly problematic in countries where data and resources may be limited (Mishra & Coulibaly, 2009). In contrast, statistical methods have gained prominence due to their practicality and reliance on historical precipitation records. The Hershfield method (1961) is the most widely used statistical approach, employing a frequency factor to standardize extreme precipitation data. This method serves as the foundation of the WMO manual on PMP estimation and has been applied extensively worldwide (Rezacova et al., 2005; Desa M. & Rakhecha, 2007; Chavan & Srinivas, 2015; Casas-Castillo et al., 2018; Wangwongwiroj & Khemngoen, 2019).

Despite its widespread use, the Hershfield method has strong limitations. Its reliance on two statistical parameters—the mean and standard deviation—fails to adequately capture the asymmetry and heavy tails inherent in extreme precipitation data (Papalexiou & Koutsoyiannis, 2013). A probabilistic analysis by Koutsoyiannis (1999) revealed that the frequency factor used in Hershfield's method (Hershfield, 1961) corresponds to an exceptionally long return period (approximately 60,000 years) when modeled with a Generalized Extreme Value (GEV) distribution, suggesting that Hershfield's provides an extreme statistical extrapolation, rather than reasonable physical upper limit of precipitation. Moreover, the method assumes spatial homogeneity across the study area, which can result in substantial inaccuracies in regions with diverse hydroclimatic and topographic characteristics (Koutsoyiannis, 1999; Chavan & Srinivas, 2017). Such assumptions often lead to overestimating PMP in humid regions and underestimation in arid areas (Desa M et al., 2001; Desa M. & Rakhecha, 2007). Additionally, the Hershfield method's empirical adjustments for outliers and record length biases lack theoretical robustness, raising concerns about its reliability in extreme scenarios.

To overcome the limitations of the Hershfield method and improve PMP estimates, statistical and regionalization techniques could be considered. The Generalized Extreme Value (GEV) distribution offers a flexible probabilistic framework for modeling extreme precipitation, as it unifies the three

limiting distribution laws of extremes (Coles, 2001; Carney, 2016). Such flexibility enables a more accurate representation of asymmetry and heavy tails in extreme rainfall data, addressing key weaknesses of the Hershfield method. Additionally, correction techniques for GEV parameters, such as those proposed by Papalexiou & Koutsoyiannis (2013), could help to mitigate biases caused by short record lengths, enhancing the reliability of PMP estimates. Beyond purely statistical approaches, spatial clustering methods, such as KMA, could be considered to account for regional variability by delineating regions with similar hydroclimatic and geographic characteristics (Chavan & Srinivas, 2017; Tuel & Martius, 2022). These methodological advancements highlight potential pathways to account for regional variability and ensure more accurate PMP estimates, particularly in areas with complex topography and diverse hydroclimatic conditions.

In this paper, we propose a refined PMP estimation methodology that integrates GEV-based standardization and clustering techniques to overcome the limitations of traditional statistical approaches. We select continental Chile as the case study due to its pronounced hydroclimatic and physiographic variability, providing a suitable context to assess the applicability and robustness of the proposed methodology. By combining advanced statistical modeling with spatial clustering, this approach seeks to enhance the accuracy and adaptability of PMP estimation, offering a framework with implications for hydrological risk assessments across diverse geographical settings.

2. Objectives

The goal of this study is to develop a new methodology for estimating the Probable Maximum Precipitation (PMP) across continental Chile, incorporating the asymmetric behavior and spatial variability of extreme precipitation. The proposed approach is based on the Generalized Extreme Value (GEV) distribution and regional clustering techniques, aiming to overcome key limitations of traditional statistical methods.

The specific objectives of the study are:

1. Propose a new frequency factor derived from the GEV distribution for PMP estimation, explicitly accounting for the asymmetry of extreme precipitation events.
2. Implement a regionalization framework for continental Chile using the k-means clustering algorithm to represent the spatial heterogeneity of climatic conditions effectively.
3. Evaluate and compare the results of the proposed methodology against the traditional Hershfield method in terms of robustness and spatial consistency.

3. Methodology

We propose a new framework for Probable Maximum Precipitation (PMP) estimation, whose schematic representation is provided in Figure 3-1. The framework comprises three main components: (i) fitting the Generalized Extreme Value (GEV) distribution to annual maximum daily precipitation (AMDP) time series in order to estimate the maximum modified frequency factor at individual stations; (ii) applying spatial clustering techniques to delineate homogeneous regions; and (iii) estimating the Regional Modified Frequency Factor by fitting an upper-bounded distribution to the series of maximum modified frequency factors from all stations within each homogeneous region. The following sections outline the methods corresponding to each component of the PMP framework and the steps involved in its implementation.

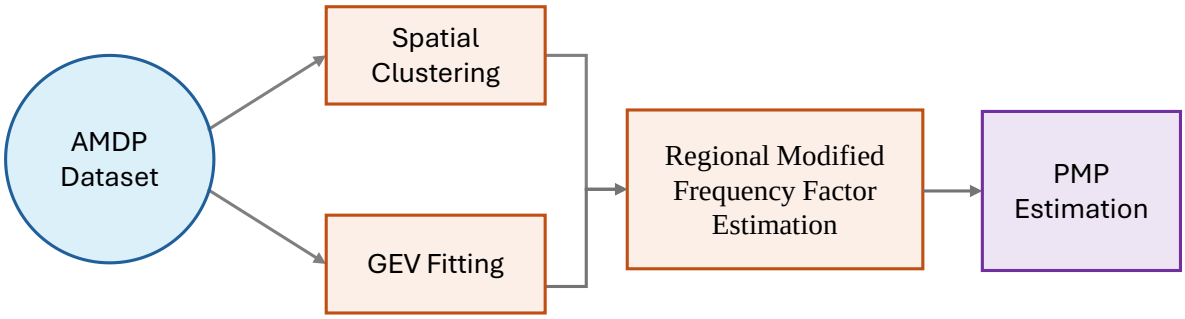


Figure 3-1. Workflow of the proposed framework for estimating PMP.

3.1. Generalized Extreme Value Distribution

The generalized extreme value (GEV) distribution is a fundamental tool in extreme value analysis, particularly well-suited for modeling block maxima, such as annual maximum daily precipitation (AMDP). This approach analyzes the maximum values within fixed time intervals (typically one year). According to the extreme value theory, there exist only three possible limiting distributions for block maxima: type I or Gumbel, type II or Fréchet, and type III or reversed Weibull (Papalexiou & Koutsoyiannis, 2013; Salas et al., 2020). These three distributions can be unified under the GEV distribution, whose cumulative distribution function (CDF) is given by:

$$F(x) = \exp\left\{-\left[1+\xi\frac{(x-\mu)}{\sigma}\right]^{-\frac{1}{\xi}}\right\}, \quad 1+\xi\frac{(x-\mu)}{\sigma} > 0, \sigma > 0 \tag{1}$$

where x corresponds to the random variable analyzed (e.g., AMDP), and μ, σ, and ξ are the location, scale, and shape parameters. The values of the shape parameter define the type of GEV distribution with $\xi > 0$, $\xi = 0$, and $\xi < 0$ corresponding to Fréchet, Gumbel, and reversed Weibull distributions, respectively (Coles, 2001). Here, we propose a modified frequency factor (K_{EV}) to estimate the PMP, which corresponds to the GEV transformed variable and is expressed as:

$$K_{EV} = \frac{1}{\xi}\ln\left(1+\xi\frac{(x-\mu)}{\sigma}\right) \tag{2}$$

3.2. Parameter Estimation and Correction Methods

Several methods are available for estimating the GEV distribution parameters, including the maximum likelihood (ML), moments (M), and L-moments (LM) methods (Gu et al., 2022). Among these, the L-Moments method has been shown to outperform ML and M methods in terms of bias and variance, particularly for small sample sizes (Hosking et al., 1985; Šimková & Picek, 2017; Tosunoğlu, 2018). The L-moments method is derived from the probability weights moments (PWM; Hosking, 1990). The expressions for estimating the GEV parameters using L-moments are the following (Carney, 2016; Hosking & Wallis, 1997a):

$$\xi \approx -7.8590c - 2.9554c^2 \tag{3}$$

$$\sigma = \frac{-l_2\xi}{\Gamma(1-\xi)(1-2^{\xi})} \tag{4}$$

$$\mu = l_1 - \frac{\sigma}{\xi}[\Gamma(1-\xi) - 1] \tag{5}$$

where c is defined as

$$c = \frac{2l_2}{l_3 + 3l_2} - \frac{\ln(2)}{\ln(3)} = \frac{2}{t_3 + 3} - \frac{\ln(2)}{\ln(3)} \tag{6}$$

l_1, l_2, and l_3 are the sample L-moments, and $\Gamma(\cdot)$ denotes the gamma function (Hosking & Wallis, 1997b). In addition, L-variation (L-CV), L-skewness, and L-kurtosis coefficients can be estimated as the sample L-moments ratios $t = l_2/l_1$, $t_4 = l_4/l_2$, and $t_3 = l_3/l_2$, respectively (Hosking & Wallis, 1997)

However, several authors have reported that GEV shape parameter (ξ) estimates are sensitive to the sample size regardless of the estimation method (e.g., Carney, 2016; Hosking et al., 1985; Hossain et al., 2022; Martins & Stedinger, 2000). Papalexiou & Koutsoyiannis (2013) proposed a bias correction technique for the GEV shape parameter obtained with the L-moments based on the sample size. Their study analyzed a dataset of daily maximum precipitation records from about 15,000 stations worldwide, each with at least 40 years of data. They assumed that shape parameters are normally distributed for any record length and that an unbiased distribution exists as the length approaches infinity.

To develop the correction, they subsampled series with record lengths ≥ 80 years into subsets of varying lengths (from 10 to 115 years by a step of 5 years) and modeled the mean and variance of the shape parameter as:

$$\mu_\xi(n) = \mu_\xi + b_\mu n^{-c_\mu}; \quad \sigma_\xi(n) = \sigma_\xi + b_\sigma n^{-c_\sigma}; \tag{7}$$

where n represents the record length, and the parameters a_μ, b_μ, c_μ, a_σ, b_μ, and c_σ are estimated via the minimum least square error method for each curve. Note that $a_\mu = \mu_\xi$ and $a_\sigma = \sigma_\xi$ correspond to the limiting value of each curve when $n \to \infty$. Thus, assuming that ξ is normally distributed, the true distribution of ξ is $N(\mu_\xi, \sigma_\xi^2)$.

Then, considering that for a record length n $\xi(n) \sim N\left(\mu_\xi(n), \sigma_\xi(n)^2\right)$, the "unbiased" GEV shape parameter, $\tilde{\xi}(n)$, can be estimated as:

$$\hat{\xi}(n) = \frac{\sigma_\xi}{\sigma_\xi(n)}\left(\hat{\xi} - \mu_\xi(n)\right) + \mu_\xi \tag{8}$$

where n is sample size (number of years), $\hat{\xi}$ is the L-moments estimate of ξ, μ_ξ, σ_ξ, $\mu_\xi(n)$, and $\sigma_\xi(n)$ corresponds to the values estimated in Eq. (7). Once ξ is bias corrected, μ and σ must be updated using Eqs. (4) and (5). Values proposed by Papalexiou & Koutsoyiannis (2013) are as follows: $\mu_\xi = 0.114$, $\sigma_\xi = 0.045$, $\mu_\xi(n) = 0.114 - 0.69 \cdot n^{-0.98}$, and $\sigma_\xi(n) = 0.045 + 1.27 \cdot n^{-0.7}$. Although this is a useful approach to address the bias in the shape parameter, Carney (2016) shows that Eq. (7) parameters can vary when applied to regions with specific climate conditions.

3.3. Spatial Clustering

Estimating the PMP in large regions with substantial hydroclimatic variability requires identifying homogeneous areas or clusters with similar extreme precipitation characteristics. Hence, we delineate clusters with the K-means algorithm (KMA) because of its simplicity, efficiency, and effectiveness in partitioning high-dimensional datasets into distinct clusters with minimal within-group variance (Roushangar & Alizadeh, 2018; Tuel & Martius, 2022).

To obtain the optimal number of clusters, P, we apply the elbow method (Humaira & Rasyidah, 2020; Syakur et al., 2018). This technique involves plotting the within-cluster sum of squares (WCSS) as a function of P to identify the point at which increasing P results in diminishing returns in reducing WCSS (Hastie et al., 2009). The optimal P corresponds to the "elbow" of the curve, where the slope begins to level off. Appendix A provides a detailed description of KMA.

3.4. Upper Bounded Distribution for Regional Frequency Factor Estimation

Bounded distributions play a crucial role in hydrology, particularly in Probable Maximum Precipitation (PMP) and Probable Maximum Flood (PMF) estimation, by incorporating an upper limit that constrains extreme values (Botero & Francés, 2010). The upper bound (g) plays a fundamental role in PMP estimation and is closely related to the regional frequency factor, which defines the maximum expected precipitation intensity in a given area. The upper bound (g) can be determined using either a fixed approach, based on climatological principles (e.g, envelopes), or based on statistical approach, where g is treated as a parameter and inferred alongside other distribution parameters (Salas et al., 2020). We adopt the latter approach, as detailed in Section 3.4.4.

The most commonly applied bounded distributions in hydrology include:

3.4.1. Four-Parameter Extreme Value Distribution (EV4)

The EV4 distribution extends extreme value theory by incorporating upper and lower bounds (Kanda, 1981). It has been successfully applied in hydrological extreme frequency analysis (Takara & Tosa, 1999). The CDF of EV4 is given by:

$$F(x) = \exp\left(-\left[\frac{g - x}{v(x - a)}\right]^k\right), \qquad k > 0;\ v > 0;\ a \le x \le g \tag{11}$$

where a and g are the lower and upper bounds, v is the scale parameter, and k is the shape parameter.

3.4.2. Four-Parameter Lognormal Distribution (LN4)

The LN4 distribution is a bounded version of the lognormal distribution, helpful in modeling positively skewed hydrological variables (Slade, 1936; Takara & Loebis, 1996). The transformation defining LN4 is:

$$y = \ln\left(\frac{x-a}{g-x}\right), \qquad a \le x \le g \tag{12}$$

where y follows a lognormal distribution with location and scale parameters μ_y and σ_y, respectively, and a and g are the lower and upper bounds. The probability density function (PDF) is:

$$f(x) = \frac{g-a}{(x-a)(g-x)\sigma_y\sqrt{2\pi}} \exp\left(-\frac{\left(y-\mu_y\right)^2}{2\sigma_y^2}\right) \tag{13}$$

3.4.3. Transformed Extreme Value Distribution (TDF)

The TDF distribution is a modification of the Gumbel distribution, designed to incorporate an upper bound while retaining the simplicity of the Extreme Value Type I model (Elíasson, 1994, 1997). The CDF of TDF is defined as follows

$$F(x) = \exp\left(-\exp\left(-\frac{x}{a} + \frac{\alpha k^*}{g-x} - b\right)\right), \qquad \alpha > 0;\ k^* < 0; x \le g \tag{14}$$

where g is the upper bound, k^* is a negative constant, and α and b correspond to the scale and location parameters, respectively.

3.4.4. Upper Bound Estimation

To estimate the upper bound (g), we apply the Maximum Likelihood-Generic Equation (ML-GE) method proposed by Botero & Francés (2010). This method combines the Generic Equation proposed by Cooke (1979) to estimate g and the ML method to estimate the remaining distribution parameters. The estimation is performed by solving the following equation system:

$$\left.\begin{aligned} &\max L(\Theta) \\ g = \ & x_{max} + \int_{-\infty}^{g} [F_X(x;\Theta)]^n dx \end{aligned}\right\} \tag{15}$$

where x_{max} is the observed maximum, n is the observed sample size, and Θ corresponds to the set of distribution parameters excluding g. Notably, this approach provides a more conservative estimate than the ML estimate of g, ensuring that g is always greater than the maximum observed value.

3.5. Proposed Framework for Estimating PMP

The proposed framework to estimate PMP, considering a dataset of AMDP time series at M stations, is comprised of four steps, which are described below:

i. At station j, GEV parameters (μ_j, σ_j, ξ_j) are estimated via the L-moments method using the Eqs. (3)-(6). After that, GEV parameters are corrected based on the record length with Eq. (7). Then, the time series of the modified frequency factor (K_{EV}) are computed with Eq. (2), and the maximum value ($K_{EV,j,m}$) is selected at each station, which corresponds to

$$K_{EV,j,m} = \frac{1}{\xi_j} \ln\left(1 + \xi_j \frac{\left(x_{j,m} - \mu_j\right)}{\sigma_j}\right), \quad j = 1, 2, \dots, M \tag{16}$$

where $x_{j,m}$ is the maximum value of the AMDP at the station j. This procedure is carried out for all M stations.

ii. KMA is is applied to the dataset of AMDP (section 2.3) to obtain the regionalization of data, i.e., split the data into a subset of P homogeneous regions or clusters. As attributes for KMA implementation, geographical variables (longitude, latitude, and elevation), summary statistics —e.g., mean, variance, variation coefficient, and skewness coefficient —(Fırat et al., 2012; Kömüşcü et al., 2022) can be considered.

iii. At each cluster p, comprised of n_p points, an upper-bounded distribution is fitted to $\boldsymbol{K_{EV,p,m}} = \{K_{EV,1,m}, K_{EV,2,m}, \dots, K_{EV,n_p,m}\}$ for estimating the regional maximum modified frequency factor ($K_{EV,p}^{MAX}$), which corresponds to the upper bound of the distribution g.

iv. Finally, the PMP at each point can be estimated depending on the cluster to which it was assigned as follows:

$$PMP_{i,p} = \frac{\left(\sigma_{i,p} \exp\left(K_{EV,p}^{MAX} \xi_{i,p}\right) - \sigma_{i,p} + \mu_{i,p} \xi_{i,p}\right)}{\xi_{i,p}}, \quad i = 1, 2, \dots, n_p;\ p = 1, 2, \dots, P \tag{16}$$

where $PMP_{i,p}$ corresponds to the PMP at the point i from the cluster p.

All steps of the proposed methodology were implemented in R (R Core Team, 2017).

4. Case Study

4.1. Domain

Chile extends from approximately 17.8°S to 55°S, spanning over 4,300 km along the western edge of South America (Figure 4-1a). This pronounced latitudinal extent, coupled with a steep topographic gradient—comprising the Coastal Range to the west, the Intermediate Depression in the center, and the Andes Mountains to the east (with elevations exceeding 6,000 m)— results in a remarkable diversity of hydroclimatic and geographic conditions (Aceituno et al., 2021; Alvarez-Garreton et al., 2018; Vásquez et al., 2025). Annual precipitation varies dramatically across the country, from less than 10 mm/yr in the Atacama Desert (one of the driest regions globally) to more than 3,000 mm/yr in the south area (Gateño et al., 2024; Vásquez et al., 2024, 2025), reflecting a suite of hydroclimatic regimes that pose challenges for hydrological and climatological analyses.

Following Alvarez-Garreton et al. (2018), continental Chile can be subdivided into six macro-zones —Far North, Near North, Central Zone, Southern Zone, Austral Zone, and Southern Patagonia— that reflect pronounced climatic and topographic gradients. These spatial variations are critical for hydrological applications, as extreme precipitation is driven by distinct mechanisms (e.g., convective storms in the north and frontal systems in the south). These macro-zones broadly align with the Köppen-Geiger climate classification (Sarricolea et al., 2017) and provide a coherent framework for interpreting the country's diverse precipitation regimes (Figure 4-1b). The Far North (17.8°S–25.7°S) ranges from cold desert (BWk) at low elevations to tundra-like (ET) climates in the high Andes. The Near North (25.7°S–32.2°S) remains predominantly cold desert in areas like the Atacama region, transitioning into cold semi-arid (BSk) regimes (e.g., Coquimbo region), where precipitation events are typically short and intense. The Central Zone (32.2°S–36°S) features a sub-humid Mediterranean

climate (Csb) characterized by more frequent winter rains. In the Southern Zone (36°S–42.5°S), areas such as Biobío and Araucanía exhibit humid Mediterranean conditions. At the same time, the adjacent Los Ríos and Los Lagos regions shift toward temperate oceanic (Cfb) climates with substantially higher annual precipitation totals. Further south, the Austral (42.5°S–49°S) and Southern Patagonia (49°S–55°S) regions are predominantly characterized by temperate oceanic (Cfb), subpolar oceanic (Cfc), tundra (ET), and ice cap (EF) climates, marked by persistently low temperatures and precipitation occurring throughout the year in both liquid and solid forms.

Such heterogeneity makes Chile particularly well-suited for testing Probable Maximum Precipitation (PMP) estimation techniques, given the need to account for the pronounced latitudinal gradient, coastal influences in the west, and the orographic effects of the Andes. Hence, we aim to demonstrate the transferability and robustness of our method in regions with varied topography and hydroclimatic regimes. These insights have implications not only for Chile but also for any area facing limited observational data and substantial hydroclimatic diversity.

4.2. Data

We compiled daily precipitation records from the Center for Climate and Resilience Research (CR2). This database comprises 879 meteorological stations across continental Chile (Figure 4-1c), primarily operated by two government agencies: the Chilean General Water Directorate (DGA, 845 stations) and the Chilean Meteorological Office (DMC, 34 stations). The measurement records span from January 1900 to June 2020, although many stations exhibit varying degrees of missing data and differing record lengths.

Initially, each station's Annual Maximum Daily Precipitation (AMDP) series was constructed by selecting the highest recorded daily precipitation value per year. We only retained stations with at least 25 years of record length to ensure sufficient temporal coverage and data reliability. Because some stations contained incomplete or intermittent measurements, we then applied two additional filters (adapted from Papalexiou & Koutsoyiannis, 2013) to mitigate the potential bias that can arise from either discarding entire years or uncritically retaining years with excessive data gaps:

i. **Missing-Data Threshold:** A year was excluded if more than 33% of its daily records were absent.
ii. **Percentile Criterion:** If the AMDP exceeded the 40th percentile of the station's entire AMDP series, that year was still retained, even if it had partial gaps.

After applying these filters, we retained 428 stations (out of the original 879) with record lengths ≥25 years (black circles in Figure 4-1c). To ensure consistency with hydrological practice.

To implement the GEV shape parameter bias correction approach proposed by Papalexiou & Koutsoyiannis (2013), we selected only stations with record lengths of at least 40 years. After this filter, we retained 243 stations, with record lengths varying between 40 and 99 years. Each record was then partitioned into subseries with lengths ranging from 10 to 55 years, increasing by a step of 3 years. The 55-year upper limit was chosen because longer lengths significantly reduce the number of generated subseries (fewer than 100 records). For all these records at each record length, we estimated the GEV shape parameter using the L-moments method.

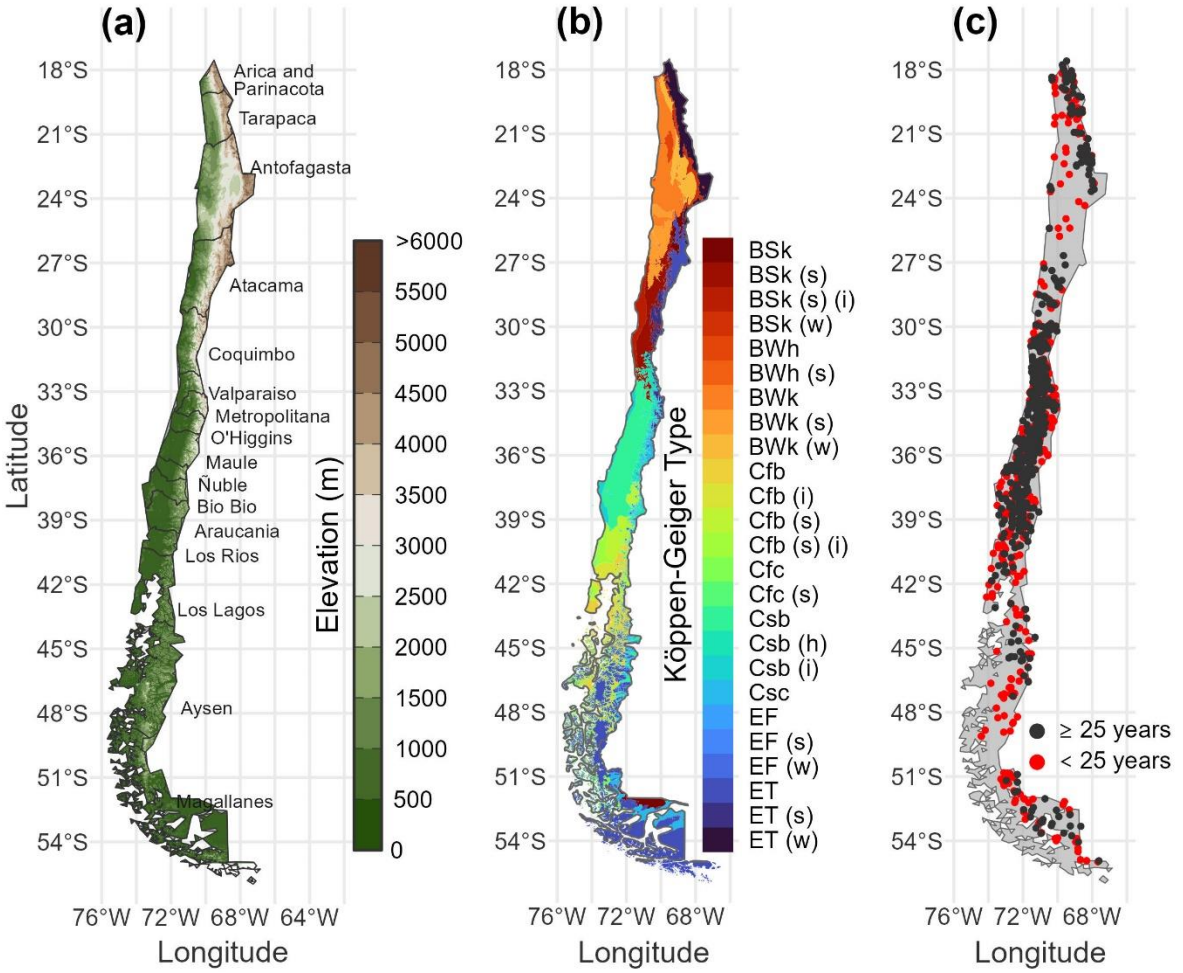


Figure 4-1. Map of mainland Chile showing three variables. (a) Elevation map (meters) displaying the topography of Chile, regional boundaries, and names. (b) Köppen-Geiger climate classification across the country, where BSk: cold semi-arid, BWh: hot desert, BWk: cold desert, Cfb: temperate oceanic, Cfc: subpolar oceanic, Csb: warm-summer Mediterranean, Csc: cold-summer Mediterranean, EF: ice cap, and ET: tundra. (c) Spatial distribution of data points categorized by AMDP record length (≥ 25 years in black and < 25 years in red).

5. Results

5.1. GEV fitting

Table 5-1 displays the summary statistics of the dataset (428 stations) used for PMP estimation. All statistical metrics show significant variation –for example, the mean ranges from 1.55 mm to 182.79 mm, and the standard deviation ranges from 1.3 mm to 72.39 mm. This variability reflects the diverse hydroclimatic conditions across Chile (Figure 4-1b). Notably, the significant variation in shape characteristics (skewness, L-skewness, and L-kurtosis) suggests that distributions with a fixed shape parameter will be inadequate for describing the AMDP data. This observation is confirmed in Figure 5-1, which displays Observed L-kurtosis versus L-Skewness points of the 428 AMDP records along with the theoretical point and line of the Gumbel and GEV distribution, respectively. The GEV distribution's flexibility in capturing the data's wide range of shape features is evident. In contrast, the Gumbel distribution, represented by a single point, cannot adequately describe the diverse shape features present in the dataset.

Table 5-1. Basic summary statistics of the 428 AMDP records considered for the PMP estimation.

	Record Length	mean	SD	Skewness	L-variation	L-skewness	L-curtosis
Min	25	1.55	1.30	-0.32	0.07	-0.07	-0.07
Q_5	26	12.34	7.52	0.19	0.14	0.05	0.06
Q_{25}	31	28.33	15.57	0.64	0.18	0.13	0.12
Q_{50}	41	54.45	21.25	1.01	0.23	0.20	0.16
Q_{75}	53	72.83	27.26	1.42	0.29	0.26	0.20
Q_{95}	72	112.77	38.70	2.35	0.46	0.38	0.31
Max	99	182.79	72.39	4.36	0.71	0.60	0.53
Mean	43.9	55.01	21.96	1.10	0.25	0.20	0.17
SD	14.9	30.98	9.72	0.70	0.10	0.11	0.07

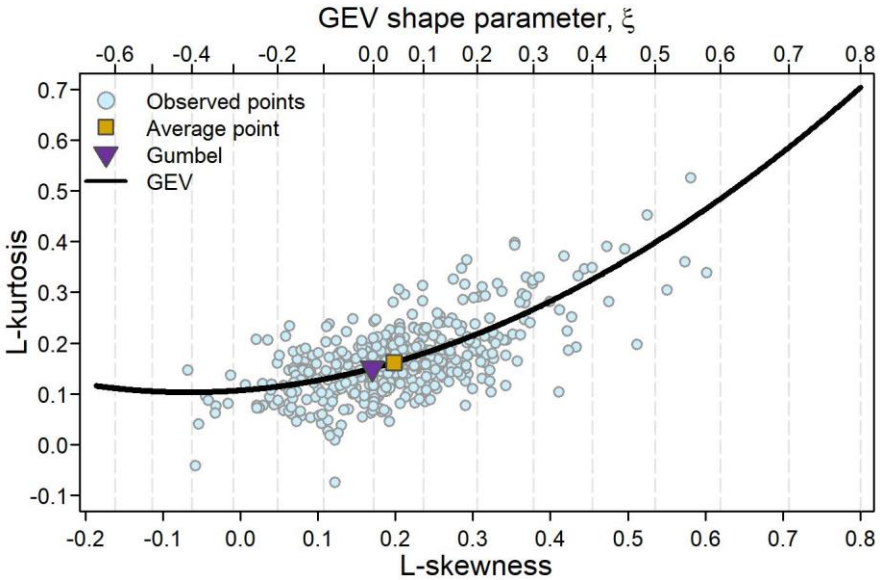


Figure 5-1. Observed L-kurtosis versus L-Skewness points of the 428 AMDP records and the theoretical point and line of the Gumbel and GEV distribution, respectively.

Table 5-2 depicts the summary statistics of the GEV parameters estimated from fitting the GEV distribution to the 428 AMDP records using the L-moments method. The location parameter (μ) ranges from 0.83 to 162.52, with a mean of 45.02 and a standard deviation of 27.65, reflecting substantial variability in the magnitude of extreme daily precipitation across the stations. The scale parameter (σ) exhibits values between 0.86 and 56.93, with a mean of 16.45, indicating a wide range of dispersion in the AMDP data. The shape parameter (ξ) also shows a high variation, with its values ranging from -0.41 to 0.58, with an average of 0.04; the 90% empirical confidence interval (ECI) is clearly smaller. Figure 5-2 displays the empirical distributions of the estimated GEV parameters. In the case of the GEV shape parameter, the plot also shows a fitted normal distribution with a mean of 0.04 and a standard deviation of 0.16. The shape parameter distribution is approximately symmetric and closely follows a fitted normal distribution (solid yellow line), which supports the assumption of a normal distribution for the shape bias correction method (Papalexiou & Koutsoyiannis, 2013).

Table 5-2. Summary statistics of the estimated parameters of the fitted GEV to the 428 AMDP records using the L-moments.

	μ	σ	ξ
Min	0.83	0.86	-0.41
Q_5	7.94	5.28	-0.20
Q_{25}	21.03	10.56	-0.07
Q_{50}	44.80	16.12	0.04
Q_{75}	60.17	20.64	0.14
Q_{95}	96.91	31.08	0.30
Max	162.52	56.93	0.58
Mean	45.02	16.45	0.04
SD	27.65	8.21	0.16

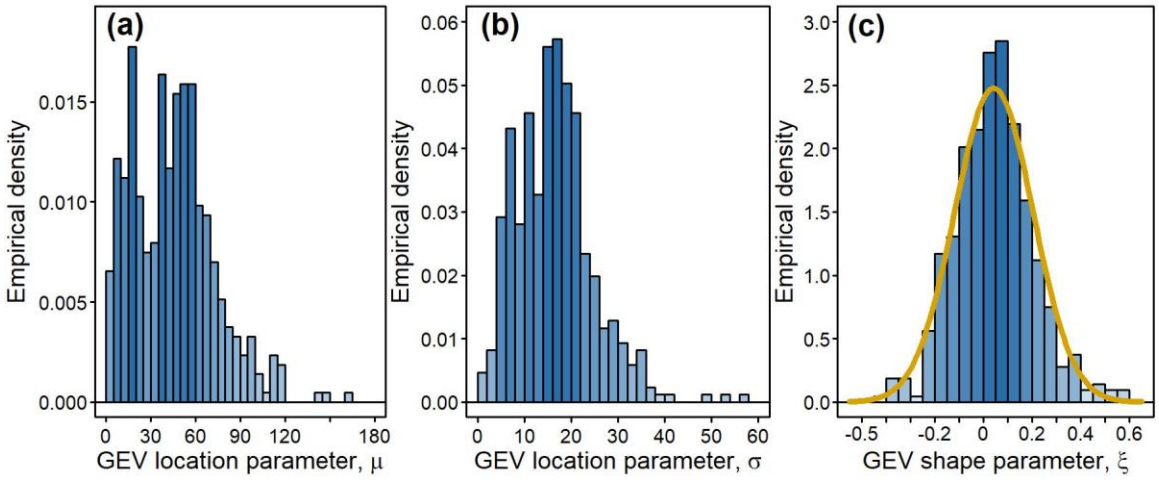


Figure 5-2. Empirical distributions of the GEV estimated parameters resulted from fitting the GEV distribution to the 428 AMDP records using the L-moments. (a) The location parameter (μ). (b) The scale parameter (σ). (c) The shape parameter (ξ). The solid yellow line denotes a fitted normal distribution.

5.1.1. Correction of the GEV Parameters

Figure 5-3 shows the observed mean, 95% confidence interval (CI), and standard deviation of the GEV shape parameter using the L-moments method across different record lengths. The figure also includes the corresponding fitted theoretical functions of the form $g(n) = a + bn^{-c}$, alongside the fitted curves obtained from Papalexiou & Koutsoyiannis (2013) for comparison. All observed values exhibit downward trends that converge to limiting values as record length increases. Table 5-3 displays the estimated parameters for the fitted curves. The fitted trends are similar to those of Papalexiou & Koutsoyiannis (2013), but differences in magnitude are evident. Specifically, the limiting values for the mean and 95% CI curves are lower than those reported by Papalexiou & Koutsoyiannis (2013) —0.114, 0.021, and 0.195, respectively. In contrast, the standard deviation exhibits higher limiting values —0.045 in Papalexiou & Koutsoyiannis (2013) — However, the mean limiting value of 0.049 closely aligns with the value of 0.044 reported by Carney (2016) for California —a region with hydroclimatic conditions that are comparable to large portions of Chile (Mooney et al., 1970).

Based on these findings, the GEV shape parameters are corrected by substituting the following functions into Eq. (7): $\mu_\xi = 0.049$, $\sigma_\xi = 0.07$, $\mu_\xi(n) = 0.049 - 1.243 \cdot n^{-1.295}$, and $\sigma_\xi(n) = 0.07 + 1.596 \cdot n^{-0.824}$. Subsequently, the location and scale parameters are recalculated using Eqs. (3)-(6). As expected, a significant reduction in the variability of the corrected shape parameter (ξ) was achieved (Figure 5-4 and Table 5-4), which now ranges from -0.16 to 0.31, with a mean of 0.05 —a value consistent with the previously estimated $\mu_\xi = 0.05$ —and a standard deviation of 0.08, which is approximately half of the initial standard deviation reported in Table 5-2. In contrast, the location (μ) and scale (σ) parameters exhibit relatively minor changes compared to their initial estimates (Figure 5-2 and Table 5-2). Finally, the high variability of corrected GEV parameters is likely attributable to the hydroclimatic heterogeneity of the study region.

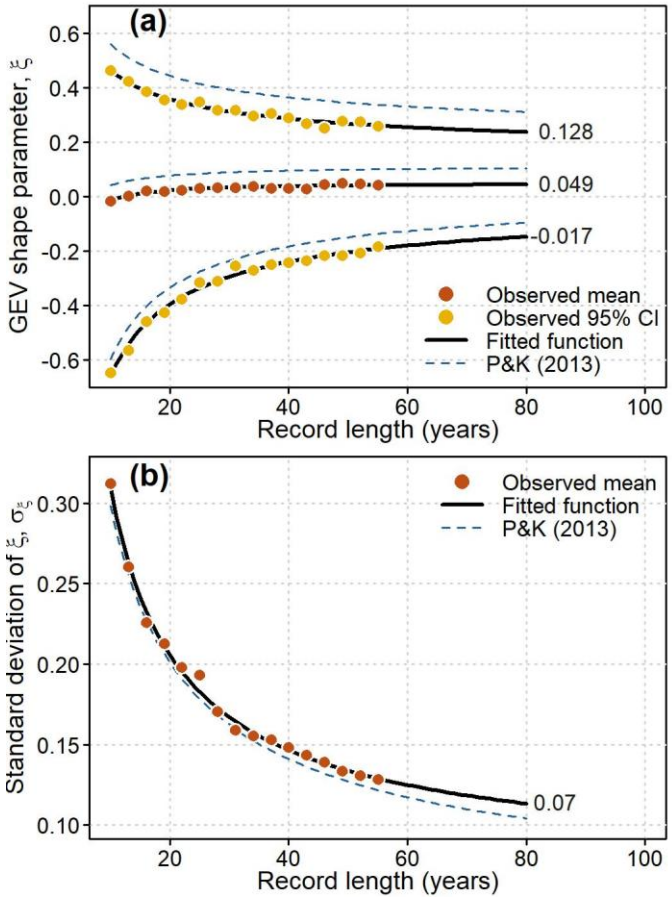


Figure 5-3. Record length analysis of the Generalized Extreme Value (GEV) shape parameter (ξ). **(**a**)** Mean of ξ and **(**b**)** standard deviation of ξ. Observed values (dots) are compared with the fitted function (solid black line) and Papalexiou & Koutsoyiannis (2013) fitted curves (dashed blue line).

Table 5-3. Parameters estimated for the curves fitted in Figure 5-3.

Statistic	*a*	*b*	*c*
Mean	0.049	-1.243	1.295
Lower 95% CI	-0.017	-3.754	0.767
Upper 95% CI	0.128	1.170	0.543
Standard Deviation	0.070	1.596	0.824

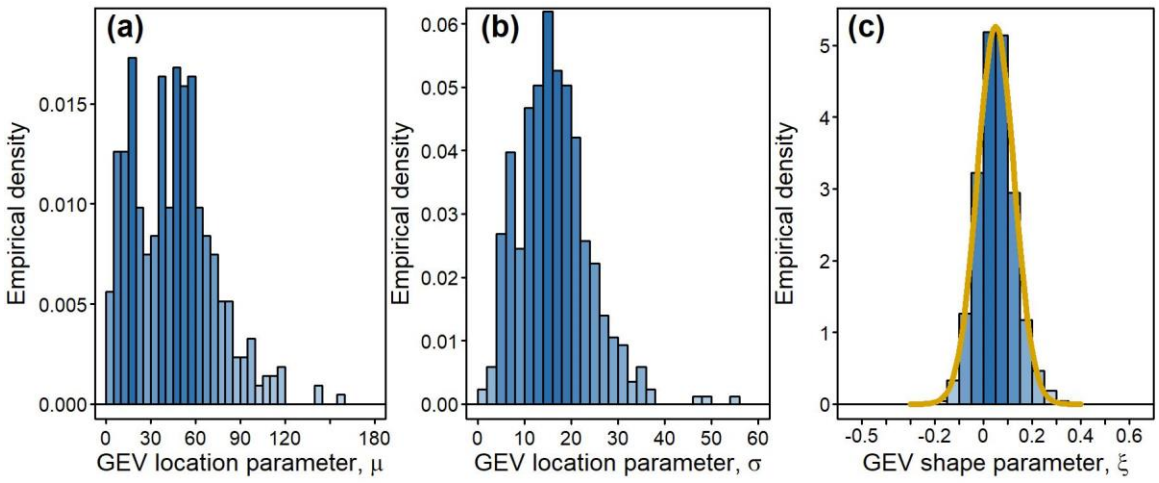


Figure 5-4. Empirical distributions of the GEV corrected parameters from the 428 AMDP records. (a) The location parameter (μ). (b) The scale parameter (σ). (c) The shape parameter (ξ). The solid yellow line denotes a fitted normal distribution.

Table 5-4. Summary statistics of the GEV corrected parameters of the 428 AMDP records.

	μ	σ	ξ
Min	0.89	0.93	-0.16
Q_5	7.92	5.47	-0.07
Q_{25}	21.05	11.19	0.00
Q_{50}	44.99	15.86	0.05
Q_{75}	60.43	20.43	0.10
Q_{95}	95.24	29.84	0.17
Max	159.99	55.53	0.31
Mean	44.84	16.34	0.05
SD	27.38	7.67	0.08

5.2. Spatial Clustering

Before defining the clustering approach, we conducted an exploratory analysis to evaluate which variables best distinguished distinct precipitation regimes. From this assessment, we selected latitude, elevation, and L-variation as our core attributes for applying KMA. This combination effectively captures the spatial gradients and statistical behavior of extreme precipitation events, a strategy consistent with common practices in hydrological regionalization (Farquharson et al., 1987; Hosking, 1990; Viglione et al., 2007).

Figure 5-5a shows the within-cluster sum of squares (WCSS) and its percentage change (%ΔWCSS) as a function of the number of clusters. The so-called elbow in the WCSS curve appears at around six clusters, where further increases in the number of clusters provide only marginal reductions in WCSS. This balance between model parsimony and goodness of fit guided us toward adopting six clusters for the subsequent analyses.

To further validate the number of clusters, we also considered L-CV and L-Skewness—proposed by Hosking (1990) and widely used for hydrological classification (Farquharson et al., 1987; Viglione et al., 2007). Figure 5-5b contrasts the L-variation coefficient (L-CV) versus L-skewness for all stations, color-coded by the assigned cluster. As noted in Lu & Stedinger (1992) —where high values of L-CV and L-CS were similarly observed for arid-zone streamflows— the spatial grouping in this plane highlights six distinct "clouds" of points, each reflecting different precipitation regimes. These patterns —which range from very low L-CV and L-skewness in some central-southern areas (clusters 4 and 5) to relatively high variability and skewness in the northern arid zones (clusters 2 and 3)— reinforce the partitioning suggested by the WCSS elbow method and further support the robustness of the six-cluster configuration. Notably, Clusters 1 and 6 partially overlap, which is consistent with the tundra-like climates characteristic of the high Andes in the Far North and Southern Patagonia— regions associated with these clusters.

Figure 5-5c provides a geographical perspective of these six clusters across mainland Chile. Distinct latitudinal bands emerge —from the hyper-arid Far North to the temperate-humid Southern Zone — underscoring how hydroclimatic gradients and orographic features (e.g., the Coastal Range, Intermediate Depression, and Andes Mountains) drive variations in extreme precipitation. The distribution of clusters aligns closely with known macro-regional and hydroclimatic delineations (Alvarez-Garreton et al., 2018; Sarricolea et al., 2017). Notably, the map also highlights transitional areas —such as around 30°S— where cluster boundaries coincide with steep shifts in topography and precipitation patterns.

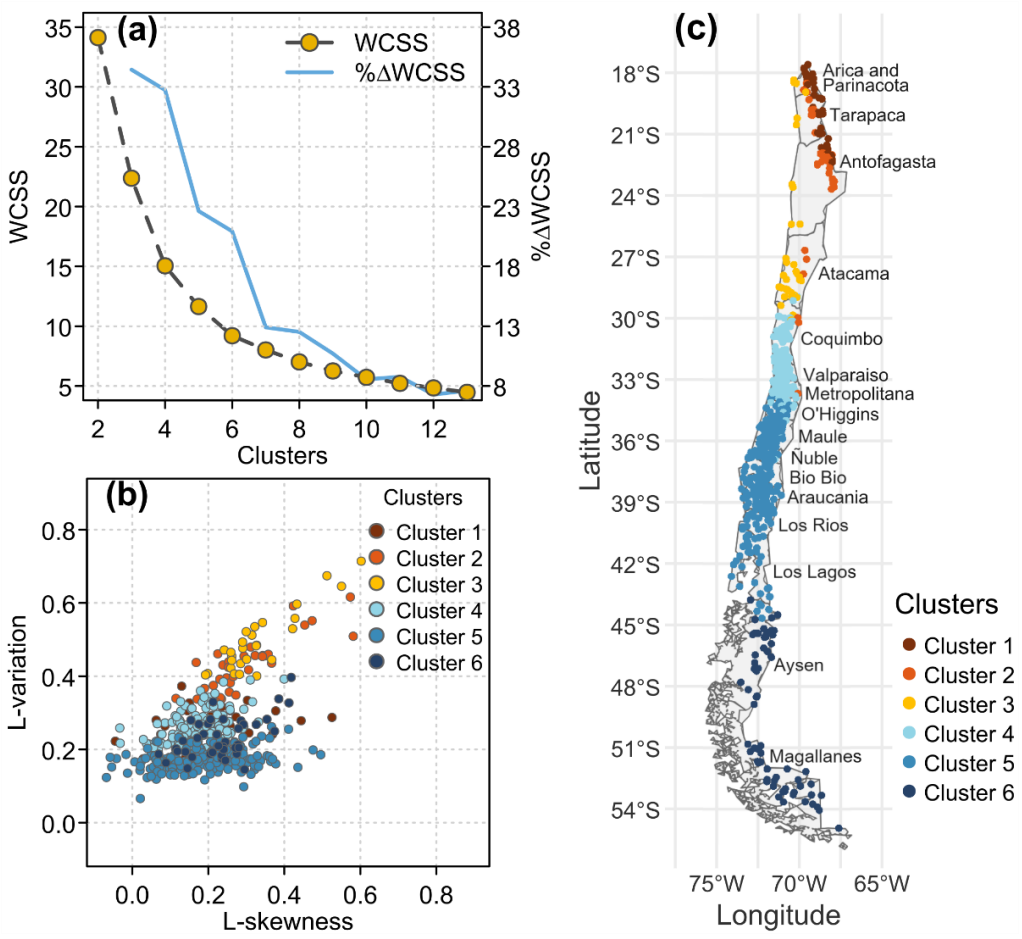


Figure 5-5. Clustering evaluation and results. **(**a) The within-cluster sum of squares (WCSS) and its percentage change (%ΔWCSS) as a function of the number of clusters used to determine the optimal number of clusters based on the elbow method; (b) scatter plot of L-variation coefficient (L-CV) versus L-skewness coefficient, with data points colored according to their assigned clusters; and (c) spatial distribution of clusters across the study area.

Table 5-6 further underscores the distinct precipitation regimes captured by the six-cluster scheme. For instance, Cluster 2 has the lowest mean AMDP (16.74 mm) yet exhibits relatively high variability (CV ≈ 0.85) and the largest skewness (1.62), reflecting the sporadic but intense precipitation events typical in arid or semi-arid regions. By contrast, Cluster 5 presents the highest mean AMDP (80.86 mm) and the lowest CV (0.32), indicating more frequent and relatively uniform precipitation patterns characteristic of temperate-humid climates. These contrasting statistics across clusters help explain why L-CV and L-skewness —combined with geographical variables— prove effective for differentiating extreme precipitation patterns. Moreover, the alignment of these metrics with known hydroclimatic gradients (e.g., from hyper-arid north to humid south) confirms that the six-cluster configuration balances complexity and representativeness, bolstering confidence in the chosen regionalization approach.

Table 5-5. Basic summary statistics of the AMDP records for each cluster

Cluster	1	2	3	4	5	6
Number of records	34	33	31	127	162	41
Mean	20.41	16.74	20.69	55.59	80.86	36.55
SD	9.93	13.28	17.8	25.72	25.76	15.4
CV	0.49	0.85	0.94	0.47	0.32	0.45
Skewness	1.09	1.62	1.56	0.88	1	1.43

5.3. Regional Maximum Frequency Factors Estimation

Table 5-6 summarizes the maximum modified frequency factor ($K_{EV,m}$) for the six clusters defined previously. The mean $K_{EV,m}$ values range from 4.12 to 4.49, with an overall mean of 4.21. Minimum and maximum values span 2.27 to 6.43. The standard deviation varies between 0.54 and 0.90 (overall 0.79), while skewness values indicate predominantly right-skewed distributions, ranging from -0.12 to 0.83 (0.65 without clustering).

To compare the proposed approach for estimating the PMP, we also calculated the frequency factor (K) proposed by (Hershfield, 1961a), where the maximum frequency factor for station j ($K_{j,m}$) is given by:

$$K_{j,m} = \frac{\left(x_{j,m} - \bar{x}_j\right)}{SD_j} \tag{17}$$

where $x_{j,m}$, $\bar{x}_j$, and SD_j correspond to the maximum value, mean, and standard deviation of the AMDP at the station j, respectively. $\bar{x}_j$, and SD_j were estimated without the maximum value of the AMDP series.

Table 5-7 presents the summary statistics of K_m from the 428 records, showing that this metric exhibits even higher variability within and between clusters. This is reflected in the high skewness values ranging from 00.13 to 2.46 (2.30 without clustering).

To estimate the regional maximum frequency factors (K_{EV}^{MAX} and K^{MAX}) —representing the fixed upper limits used to calculate PMP values for each cluster— we tested the three upper-bounded distributions described in section 2.4 (EV4, LN4, and TDF). Among these, the EV4 distribution provided the best fit —the highest ML value— for both frequency factors across all clusters and the entire study region. Table 5-8 presents the values of regional maximum frequency factors for each cluster. The proposed maximum frequency factor ($K_{EV,m}$) shows coherent values across clusters, ranging from 5.48 (Cluster 3) to 9.57 (Cluster 2). This coherence highlights the robustness of the proposed approach in providing reliable regional frequency factor estimates. In contrast, the traditional maximum frequency factor (K^{MAX}) exhibits substantial variability among clusters, with values ranging from 6.77 (Cluster 3) to a notably high 60.17 (Cluster 2). The extreme value observed in Cluster 2 reflects the sensitivity of the traditional method to outliers.

This comparison underscores the advantage of the proposed method, which mitigates the influence of extreme values and ensures more coherent and consistent regional frequency factors for PMP estimation.

Table 5-6. Summary statistics of the maximum modified frequency factor ($K_{EV,m}$) for the number of clusters selected.

	Cluster	1	2	3	4	5	6	Total
Summary Statistics of $K_{EV,m}$	Min	2.76	3.41	3.22	2.77	2.28	2.89	2.28
	Q_5	3.08	3.60	3.41	3.04	2.91	3.41	3.02
	Q_{25}	3.43	4.07	3.79	3.54	3.62	3.78	3.65
	Q_{50}	3.96	4.43	4.34	4.06	4.14	4.22	4.14
	Q_{75}	4.87	4.73	4.63	4.52	4.57	4.83	4.65
	Q_{95}	5.65	5.72	5.12	5.57	5.61	6.32	5.69
	Max	6.43	6.15	5.26	6.13	7.19	6.80	7.19
	Mean	4.22	4.49	4.27	4.10	4.16	4.42	4.21
	SD	0.89	0.64	0.54	0.74	0.83	0.94	0.79
	Skew	0.52	0.75	-0.12	0.63	0.65	0.80	0.65

Table 5-7. Summary statistics of the maximum frequency factor (K) proposed by (Hershfield, 1961a) for the number of clusters selected.

	Cluster	1	2	3	4	5	6	Total
Summary Statistics of K	Min	1.83	2.61	2.41	1.93	1.5	2.12	1.5
	Q_5	2.19	2.83	2.62	2.2	2.1	2.6	2.15
	Q_{25}	2.66	3.48	3.24	2.74	2.77	3.07	2.83
	Q_{50}	3.33	4.1	4.18	3.09	3.3	3.58	3.37
	Q_{75}	4.19	4.91	4.89	3.69	4.06	4.5	4.19
	Q_{95}	6.13	8.99	5.64	5	6.17	8.12	6.34
	Max	7.77	12.92	6.27	8.15	9.41	10.89	12.92
	Mean	3.65	4.67	4.15	3.35	3.59	4.27	3.71
	SD	1.37	2.21	1.05	0.96	1.31	1.91	1.42
	Skew	1.19	2.46	0.13	1.69	1.75	1.6	2.3

Table 5-8. Regional frequency factor for Hershfield (K^{MAX}) and proposed (K_{EV}^{MAX}) methods.

Groups	1	2	3	4	5	6
K^{MAX}	12.91	60.17	6.77	17.52	13.02	19.12
K_{EV}^{MAX}	7.78	8.49	5.48	6.50	9.57	7.97

5.4. PMP Estimation

We estimated the Probable Maximum Precipitation (PMP) for each cluster using both the traditional Hershfield and the proposed GEV-based approach, applying the regional maximum frequency factors derived in the previous section (Table 5-8). Figure 5-6 compares the PMP estimates from both methods across all six clusters and boxplots of the maximum observed AMDP values (yellow) for each cluster. A striking discrepancy emerges in Cluster 2, where Hershfield's approach yields unrealistically large PMP values. This issue arises from a highly skewed dataset (skewness ≈ 2.5). In contrast, the proposed GEV-based methodology demonstrates better numerical stability in Cluster 2. It provides lower, more plausible PMP estimates in other arid or data-scarce areas.

A similar inconsistency is observed between Clusters 4 and 5. Although both clusters exhibit comparable skewness and CV values (Table 5-5), Cluster 4 has a lower mean precipitation. Nevertheless, the Hershfield method yields PMP estimates for Cluster 4 that are comparable to —or even higher than— those in Cluster 5. This contradicts expectations based on the regional climate (Figure 4-1b), as Cluster 5 represents the humid southern region, where extreme precipitation events should be more frequent and intense. These results suggest that Hershfield's reliance on only the mean and standard deviation is insufficient to differentiate regions with similar variability but distinct precipitation regimes. By contrast, the proposed GEV-based method yields estimates that are more consistent with regional hydroclimatic characteristics, highlighting its improved capacity to reflect physical precipitation processes across diverse hydroclimatic zones.

Figure 5-7 maps the spatial distribution of PMP across Chile, interpolated using the Inverse Distance Weighting (IDW) method based on estimates from the proposed approach. Compared with Hershfield's estimates (not shown in the figure but indicated in Table 5-8 and Figure 5-6), it becomes evident that the GEV-based framework avoids significant overestimations, particularly in clusters exhibiting elevated skewness. The proposed PMP values also display a smoother latitudinal progression, aligning more closely with known precipitation gradients from Chile's hyper-arid north to its humid south.

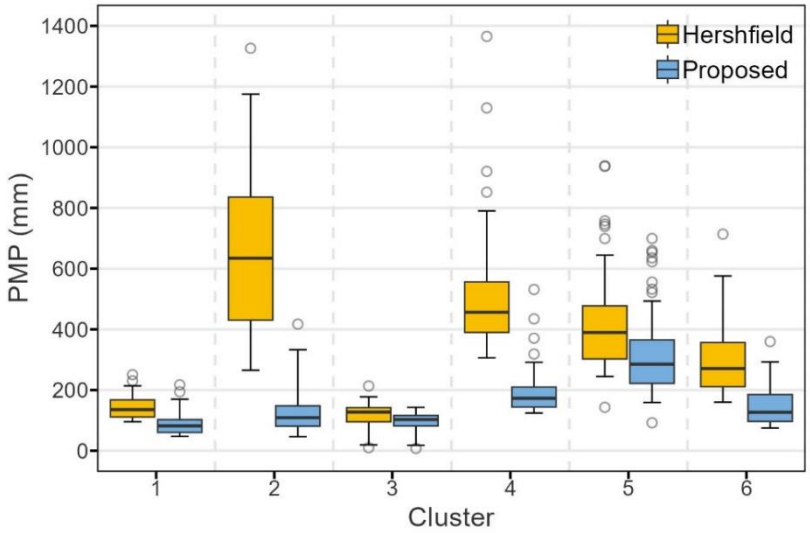


Figure 5-6. Comparison of Probable Maximum Precipitation (PMP) estimates using the Hershfield method (yellow) and the proposed methodology (blue) across different clusters.

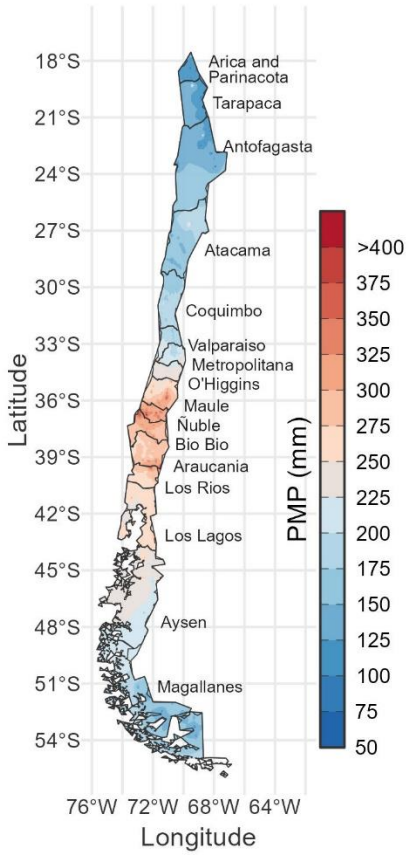


Figure 5-7. Spatial distribution of Probable Maximum Precipitation (PMP) estimates across Chile using the proposed method.

6. Discussion

6.1. GEV Shape Parameter Correction

A key challenge in estimating PMP stems from short data records, which often bias the shape parameter in GEV-based models. Our bias-correction procedure (Section 3.2) helps stabilize parameter estimates for stations with fewer years of data, mitigating the risk of overestimating extreme precipitation events. In particular, the correction becomes important in regions where historical measurements are sparse or discontinuous, ensuring that the GEV shape parameter better reflects the underlying precipitation distribution rather than artifacts of record length.

To illustrate this effect, Figure 6-1 compares PMP estimates across clusters using maximum regional frequency factors derived with and without the shape parameter correction. The largest differences occur in Clusters 4 and 5, with average PMP differences of 143 mm and 113 mm, respectively. These clusters also have the highest proportion of stations with record length fewer than 35 years (17% in Cluster 4 and 24% in Cluster 5) and a high incidence of negative shape parameters before the correction (50% and 45%, respectively), which are unrealistic for block maxima series. As shown in Figure 5, the bias correction has a higher effect when the uncorrected shape parameter is negative and the sample size is limited.

Our results are consistent with those reported by Carney (2016) for California —a region with a large hydroclimatic diversity comparable to significant portions of Chile (Mooney et al., 1970). This finding suggests that the correction method may be transferable to regions with similar precipitation regimes. However, for other regions with different hydroclimatic characteristics, the specific correction can be derived using the same methodological framework proposed in this study.

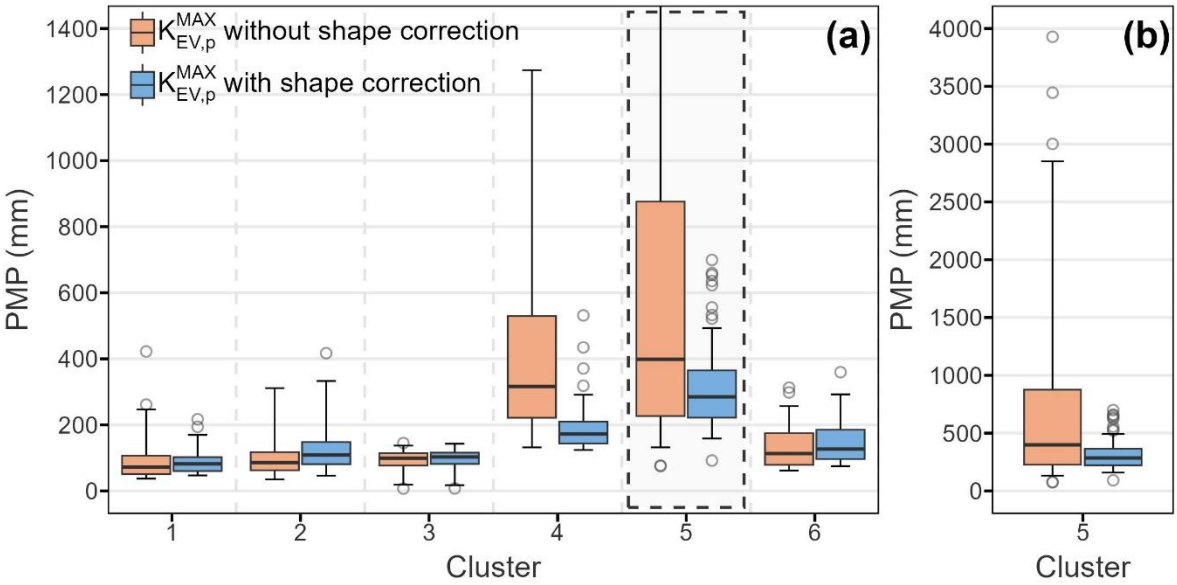


Figure 6-1. Comparison of Probable Maximum Precipitation (PMP) estimates across clusters obtained with (blue) and without (orange) applying a bias correction to the shape parameter of the GEV distribution. (a) The distribution of PMP values for each cluster. (b) Cluster 5.

6.2. Clustering

One of the significant challenges in estimating extreme precipitation is the management of heavy-tailed distributions, which are characterized by a pronounced skewness. This phenomenon is particularly pronounced in Chile, where the diversity of hydroclimates stretches from the northern regions' hyper-arid conditions to humid environments in the south. These varying precipitation patterns create complexities in accurately forecasting extreme precipitation events.

We addressed this heterogeneity through KMA clustering of stations (Section 3.3), using attributes —such as latitude, elevation, and L-variation— to demarcate zones with broadly comparable precipitation regimes. Indeed, as evidenced by the close alignment of clusters with established hydroclimatic zones (Sarricolea et al., 2017), the method can capture latitudinal gradients, orographic effects, and the transitions from arid to humid precipitation regimes. By delineating hydroclimatic homogeneous areas, the approach curbs systematic biases arising from applying a single frequency factor across diverse hydroclimates and bolsters confidence in the resulting PMP estimates.

The approach introduced in this work —combining GEV-based frequency factors with estimating their regional maxima using upper-bounded distributions (e.g., EV4)— yields more coherent PMP values across these clusters and aligns more naturally with local hydroclimatic patterns. In contrast, the Hershfield method tends to inflate estimates in highly skewed datasets (particularly in Clusters 2 and 4) because its reliance on mean and standard deviation alone fails to adequately capture heavy-tailed precipitation extremes. It is worth noting that (Stohwas, 1983) —whose values have been

widely used in the design of hydraulic infrastructure in Chile— explicitly excluded stations from the Far North in applying the Hershfield method precisely due to the challenges posed by high skewness in those regions. This skewness results from the strong presence of very low or zero precipitation values in annual maxima series, which compromises the applicability and stability of the Hershfield formulation.

By explicitly clustering stations with similar hydroclimatic signatures, the proposed method keeps local biases in check, thereby improving the internal consistency of PMP estimates. To further illustrate the adequacy of this approach, Figure 6-2 presents the differences in PMP estimates obtained using a fixed maximum frequency factor for the entire country ($K_{EV}^{MAX} = 9.52$) versus using cluster-specific $K_{EV,p}^{MAX}$ values derived from the proposed methodology. With the exception of Cluster 5, whose regional frequency factor closely matches the country value, using a fixed coefficient leads to notable overestimations, reaching up to 350 mm in Cluster 4. It is also important to note that although the differences in PMP estimates in Clusters 1 and 2 appear relatively small (27 mm and 32 mm, respectively), they are substantial when assessed in relation to the mean annual maximum precipitation in those regions, which does not exceed 21 mm (Table 5-5). These results underscore the importance of using regionally tailored frequency factors to avoid distortions in PMP estimation, particularly in arid or topographically complex areas.

A comparative analysis of PMP estimates from both the proposed method and the traditional Hershfield approach, relative to the maximum historically observed precipitation at each station, further highlights the implications of methodological selection. As shown in Figure 6-3, the Hershfield method consistently produces relative increases that in some clusters exceed 300%, and in extreme cases, even surpass 1200% (e.g., Cluster 2). These values vastly exceed the bounds of expected natural variability and are inconsistent with climate model projections. Such substantial increases are far beyond the magnitudes typically reported in studies analyzing historical trends and climate change projections for extreme precipitation events in Chile (González-Reyes et al., 2021; Lagos-Zúñiga et al., 2024). This contrast underscores a critical limitation of the Hershfield method, which may lead to substantial overdesign in hydraulic infrastructure, particularly in data-sparse or highly skewed hydroclimatic zones. By comparison, the proposed GEV-based approach yields more moderated and spatially coherent PMP estimates that align more closely with both historical extremes and plausible future scenarios. These findings reinforce the importance of adopting regionally adaptive and distribution-sensitive methodologies in PMP estimation, ensuring the technical reliability and economic viability of infrastructure design under evolving climatic conditions.

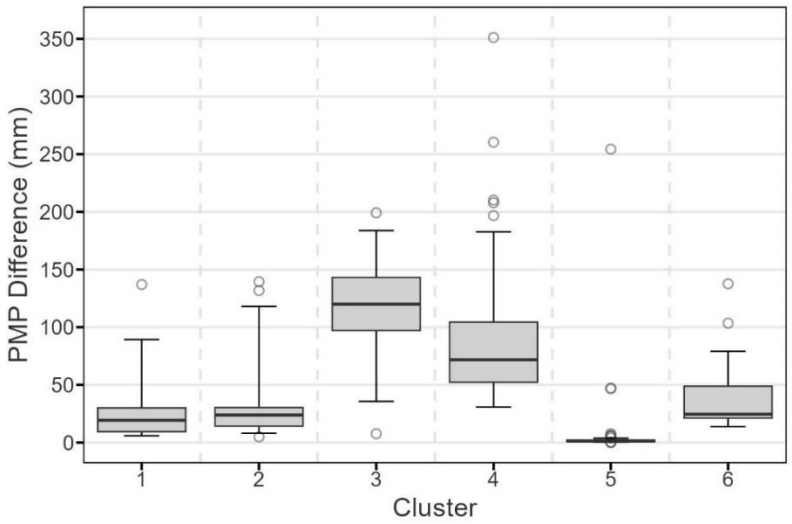


Figure 6-2. Difference in PMP estimates between using a fixed K_{EV}^{MAX} for the entire study area ($K_{EV}^{MAX} = 9.52$) and cluster-specific $K_{EV,p}^{MAX}$ values across regions.

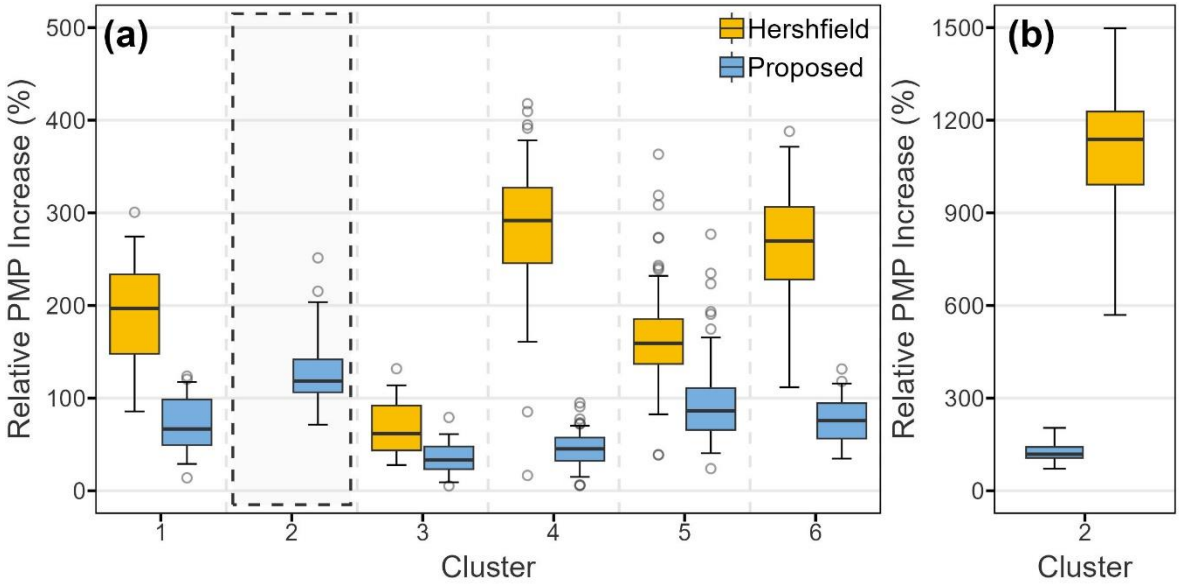


Figure 6-3. (a) The relative increase in Probable Maximum Precipitation (PMP) estimates with respect to the maximum observed AMDP value, expressed as a percentage, for the traditional Hershfield method (yellow) and the proposed methodology (blue) across six clusters. (b) Cluster 2.

7. Summary and Conclusion

This study presents a novel statistical approach for estimating Probable Maximum Precipitation (PMP) by integrating the Generalized Extreme Value (GEV) distribution with a KMA clustering-based regionalization approach. A key innovation is the bias correction for the GEV shape parameter, which substantially improves the reliability of extreme precipitation estimates in regions with short or incomplete data records. Additionally, the framework ensures that PMP estimates remain within physically meaningful limits by employing upper-bounded distributions, mitigating the risk of unrealistically large extrapolations. The methodology was applied to 428 meteorological stations in Chile —a country with diverse hydroclimatic conditions— demonstrating that the proposed approach produces more consistent and regionally coherent PMP estimates than the Hershfield´s method (Hershfield, 1961b).

The key contributions of this study are:

- Improved PMP Estimation: Through GEV-based modeling and shape-parameter bias correction, the proposed framework offers a statistically robust means of quantifying extreme precipitation. It addresses the limitations of short record lengths and heavy-tailed distributions by reducing the influence of outliers and negative shape parameter estimates, thereby enhancing the stability of PMP estimates.

- By employing KMA clustering on both geographic and statistical attributes, the methodology delineates hydroclimatic homogeneous regions. This not only aligns PMP estimates with recognized climatic zones (Sarricolea et al., 2017) but also avoids the distortions that arise when using a single maximum frequency factor for all stations. As a result, highly skewed or arid clusters, which often produce extreme outliers with traditional methods, are more accurately represented.

The combination of bias-corrected GEV modeling, upper-bounded distributions, and data-driven clustering addresses key limitations in PMP estimation, providing an adaptable framework for regions with significant hydroclimatic variability. Notably, the proposed approach can also be used to establish a physically meaningful upper bound for hydraulic infrastructure design, even in cases where standard frequency analysis identifies a different probability distribution as the best fit to the data, since it only requires estimating L-moment statistics from the observed data.

Future research should focus on validating the proposed methodology across hydroclimatic contexts and exploring its integration with high-resolution climate model projections to assess potential variations in PMP estimates under changing climate conditions. In the case of continental Chile, it is advisable to periodically update the bias correction expressions applied to the GEV shape parameter. This practice would enable detecting and adjusting potential shifts in the parameter's behavior over time, particularly in light of climate change. Currently, the mean value of the unbiased shape parameter is significantly influenced by the spatially uneven distribution of meteorological stations, which are predominantly concentrated in the central-southern region of the country. Regular updates to the correction formulas would enhance the robustness and spatial representativeness of extreme precipitation estimates, ensuring their continued relevance under changing climatic conditions.

Appendix A: K-means algorithm (KMA)

Let $\boldsymbol{X} = \{\boldsymbol{x}_1, \boldsymbol{x}_2, \dots, \boldsymbol{x}_M\}$ be a dataset of M data points, where each data point $\boldsymbol{x}_j$ is an n-dimensional vector of attributes, i.e., $\boldsymbol{x}_j = \{x_{1j}, x_{2j}, \dots, x_{nj}\}$ for $j = 1, 2, \dots, M$. KMA partitions the dataset $\boldsymbol{X}$ into P groups or clusters, each defined by a centroid $\boldsymbol{u}_k = \{u_{1p}, u_{2p}, \dots, u_{np}\}$, which has the same dimensionality as the original data, being $p = 1, \dots, P$. The clustering process begins with a random initialization of the centroids $\{\boldsymbol{u}_1^0, \boldsymbol{u}_2^0, \dots, \boldsymbol{u}_P^0\}$. At each iteration i, each data point $\boldsymbol{x}_r$ is assigned to the cluster whose centroid is closest in terms of Euclidean distance:

$$p = \arg\min_p \left\{ \left\| \boldsymbol{x}_i - \boldsymbol{u}_p^{(i-1)} \right\| \right\}, \quad p = 1, \dots, P \tag{9}$$

where $\|\cdot\|$ denotes the Euclidean norm, and $\boldsymbol{u}_p^{i-1}$ is the centroid p from the previous iteration $(i-1)$.

Once all data points have been assigned to clusters, the centroids are updated as the mean of the assigned data points:

$$\boldsymbol{u}_p^i = \frac{1}{n_p} \sum_{\boldsymbol{x}_i \in C_p} \boldsymbol{x}_i \tag{10}$$

where n_p represents the number of data points in the cluster p, and C_p is the subset of data points assigned to cluster p. The KMA iteratively updates the centroids to minimize the overall within-cluster distance. This iterative process continues until convergence, which occurs when the cluster assignments stabilize (For more details, refer to Hastie et al., 2009). Note that before applying KMA, normalization must be applied to the attributes since they are in different scales (Camus et al., 2011; Wu et al., 2022).

References

Aceituno, P., Boisier, J. P., Garreaud, R., Rondanelli, R., & Rutllant, J. A. (2021). Climate and Weather in Chile. In B. Fernández & J. Gironás (Eds.), *Water Resources of Chile* (pp. 7–29). Springer International Publishing. https://doi.org/10.1007/978-3-030-56901-3_2

Alvarez-Garreton, C., Mendoza, P. A., Pablo Boisier, J., Addor, N., Galleguillos, M., Zambrano-Bigiarini, M., Lara, A., Puelma, C., Cortes, G., Garreaud, R., McPhee, J., & Ayala, A. (2018). The CAMELS-CL dataset: Catchment attributes and meteorology for large sample studies-Chile dataset. *Hydrology and Earth System Sciences*, *22*(11), 5817–5846. https://doi.org/10.5194/HESS-22-5817-2018

Botero, B. A., & Francés, F. (2010). Estimation of high return period flood quantiles using additional non-systematic information with upper bounded statistical models. *Hydrology and Earth System Sciences*, *14*(12), 2617–2628. https://doi.org/10.5194/hess-14-2617-2010

Carney, M. C. (2016). Bias Correction to GEV Shape Parameters Used to Predict Precipitation Extremes. *Journal of Hydrologic Engineering*, *21*(10). https://doi.org/10.1061/(ASCE)HE.1943-5584.0001416

Elíasson, J. (1994). Statistical estimates of PMP values. *Nordic Hydrology*, *25*(4), 301–312.

Elíasson, J. (1997). A statistical model for extreme precipitation. *Water Resources Research*, *33*(3), 449–455. https://doi.org/10.1029/96WR03531

Farquharson, F. A. K., Green, C. S., Meigh, J. R., & Sutcliffe, J. V. (1987). Comparison of flood frequency curves for many different regions of the world. *International Symposium on Flood Frequency and Risk Analyses*, 223–256.

Fırat, M., Dikbaş, F., Koç, A. C., & Güngör, M. (2012). Classification of Annual Precipitations and Identification of Homogeneous Regions using K-Means Method. *Teknik Dergi*, *23*(115), 1609–1622.

Gateño, F., Mendoza, P. A., Vásquez, N., Lagos-Zúñiga, M., Jiménez, H., Jerez, C., Vargas, X., Rubio-Álvarez, E., & Montserrat, S. (2024). Screening CMIP6 models for Chile based on past performance and code genealogy. In *Climatic Change* (Vol. 177, Issue 6). https://doi.org/10.1007/s10584-024-03742-1

González-Reyes, Á., Jacques-Coper, M., & Muñoz, A. A. (2021). Seasonal precipitation in south-central Chile: Trends in extreme events since 1900. *Atmosfera*, *34*(4), 371–384. https://doi.org/10.20937/ATM.52871

Gu, X., Ye, L., Xin, Q., Zhang, C., Zeng, F., Nerantzaki, S. D., & Papalexiou, S. M. (2022). Extreme Precipitation in China: A Review on Statistical Methods and Applications. *Advances in Water Resources*, *163*, 104144. https://doi.org/10.1016/j.advwatres.2022.104144

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The Elements of Statistical Learning*. https://doi.org/10.1007/978-0-387-84858-7

Hershfield, D. M. (1961a). Estimating the Probable Maximum Precipitation. *Journal of the Hydraulics Division*, *87*(5), 99–116. https://doi.org/10.1061/JYCEAJ.0000651

Hershfield, D. M. (1961b). Estimating the Probable Maximum Precipitation. *Journal of the Hydraulics Division*, *87*(5), 99–116. https://doi.org/10.1061/JYCEAJ.0000651

Hosking, J. R. M. (1990). L-Moments: Analysis and Estimation of Distributions Using Linear Combinations of Order Statistics. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, *52*(1), 105–124. https://doi.org/10.1111/j.2517-6161.1990.tb01775.x

Hosking, J. R. M., & Wallis, J. R. (1997a). *Regional Frequency Analysis: An approach based on L-moments*. Cambridge University Press. https://doi.org/10.1017/CBO9780511529443

Hosking, J. R. M., & Wallis, J. R. (1997b). *Regional Frequency Analysis*. Cambridge University Press. https://doi.org/10.1017/CBO9780511529443

Hosking, J. R. M., Wallis, J. R., & Wood, E. F. (1985). Estimation of the Generalized Extreme-Value Distribution by the Method of Probability-Weighted Moments. *Technometrics*, *27*(3), 251–261. https://doi.org/10.1080/00401706.1985.10488049

Hossain, I., Khastagir, A., Aktar, M. N., Imteaz, M. A., Huda, D., & Rasel, H. M. (2022). Comparison of estimation techniques for generalised extreme value (GEV) distribution parameters: a case study with Tasmanian rainfall. *International Journal of Environmental Science and Technology*, *19*(8), 7737–7750. https://doi.org/10.1007/s13762-021-03693-5

Humaira, H., & Rasyidah, R. (2020). Determining The Appropiate Cluster Number Using Elbow Method for K-Means Algorithm. *Proceedings of the Proceedings of the 2nd Workshop on Multidisciplinary and Applications (WMA) 2018, 24-25 January 2018, Padang, Indonesia*. https://doi.org/10.4108/eai.24-1-2018.2292388

Kanda, J. (1981). A new extreme distribution with lower and upper limits for earthquake motions and wind speeds. *Proc. of the 31st Japan National Congress for Applied Mechanics*, 351–360.

Kömüşcü, A. Ü., Turgu, E., & DeLiberty, T. (2022). Dynamics of precipitation regions of Turkey: A clustering approach by K-means methodology in respect of climate variability. *Journal of Water and Climate Change*, *13*(10), 3578–3606. https://doi.org/10.2166/wcc.2022.186

Lagos-Zúñiga, M., Mendoza, P. A., Campos, D., & Rondanelli, R. (2024). Trends in seasonal precipitation extremes and associated temperatures along continental Chile. *Climate Dynamics*, *62*(5), 4205–4222. https://doi.org/10.1007/s00382-024-07127-z

Lu, L. H., & Stedinger, J. R. (1992). Sampling variance of normalized GEV/PWM quantile estimators and a regional homogeneity test. *Journal of Hydrology*, *138*(1–2), 223–245. https://doi.org/10.1016/0022-1694(92)90166-S

Martins, E. S., & Stedinger, J. R. (2000). Generalized maximum-likelihood generalized extreme-value quantile estimators for hydrologic data. *Water Resources Research*, *36*(3), 737–744. https://doi.org/10.1029/1999WR900330

Mooney, H. A., Dunn, E. L., Shropshire, F., & Song, L. (1970). Vegetation Comparisons between the Mediterranean Climatic Areas of California and Chile. *Flora*, *159*(5), 480–496. https://doi.org/10.1016/S0367-2530(17)31053-8

Papalexiou, S. M., & Koutsoyiannis, D. (2013). Battle of extreme value distributions: A global survey on extreme daily rainfall. *Water Resources Research*, *49*(1), 187–201. https://doi.org/10.1029/2012WR012557

R Core Team. (2017). *R: A Language and Environment for Statistical Computing* (3.6.2). R Foundation for Statistical Computing.

Roushangar, K., & Alizadeh, F. (2018). A multiscale spatio-temporal framework to regionalize annual precipitation using k-means and self-organizing map technique. *Journal of Mountain Science*, *15*(7), 1481–1497. https://doi.org/10.1007/s11629-017-4684-5

Salas, J. D., Anderson, M. L., Papalexiou, S. M., & Frances, F. (2020). PMP and Climate Variability and Change: A Review. *Journal of Hydrologic Engineering*, *25*(12), 03120002. https://doi.org/10.1061/(ASCE)HE.1943-5584.0002003

Sarricolea, P., Herrera-Ossandon, M., & Meseguer-Ruiz, Ó. (2017). Climatic regionalisation of continental Chile. *Journal of Maps*, *13*(2), 66–73. https://doi.org/10.1080/17445647.2016.1259592

Šimková, T., & Picek, J. (2017). A comparison of L-, LQ-, TL-moment and maximum likelihood high quantile estimates of the GPD and GEV distribution. *Communications in Statistics - Simulation and Computation*, *46*(8), 5991–6010. https://doi.org/10.1080/03610918.2016.1188206

Slade, J. (1936). An asymmetric probability function. *Transactions of the American Society of Civil Engineers*, *101*(1), 35–104.

Stohwas, L. (1983). Precipitaciones Máximas Diarias En Chile. *VI Congreso Chileno de Ingeniería Hidráulica*, 455–473.

Syakur, M. A., Khotimah, B. K., Rochman, E. M. S., & Satoto, B. D. (2018). Integration K-Means Clustering Method and Elbow Method For Identification of The Best Customer Profile Cluster. *IOP Conference Series: Materials Science and Engineering*, *336*, 012017. https://doi.org/10.1088/1757-899X/336/1/012017

Takara, K., & Loebis, J. (1996). Frequency Analysis Introducing Probable Maximum Hydrologic Events: Preliminary Studies in Japan and in Indonesia. *Proceedings of International Symposium on Comparative Research on Hydrology and Water Resources in Southeast Asia and the Pacific*, 67–76.

Takara, K., & Tosa, K. (1999). Storm and flood frequency analysis using PMP/PMF estimates. *Proceedings of International Symposium on Floods and Droughts, UNESCO-IHP, Nanjing, China*, 18–20.

Tosunoğlu, F. (2018). Accurate estimation of T year extreme wind speeds by considering different model selection criterions and different parameter estimation methods. *Energy*, *162*, 813–824. https://doi.org/10.1016/j.energy.2018.08.074

Tuel, A., & Martius, O. (2022). Subseasonal Temporal Clustering of Extreme Precipitation in the Northern Hemisphere: Regionalization and Physical Drivers. *Journal of Climate*, *35*(11), 3537–3555. https://doi.org/10.1175/JCLI-D-21-0562.1

Vásquez, N. A., Mendoza, P. A., Knoben, W. J. M., Arnal, L., Lagos-Zúñiga, M., Clark, M., & Vargas, X. (2024). The Key Role of Temporal Stratification for GCM Bias Correction in Climate Impact Assessments. *Earth's Future*, *12*(8), e2023EF004242. https://doi.org/10.1029/2023EF004242;WGROUP:STRING:PUBLICATION

Vásquez, N. A., Mendoza, P. A., Lagos-Zuñiga, M., Scaff, L., Muñoz-Castro, E., & Vargas, X. (2025). Robust spatial changes in climate classes: insights from bias-corrected CMIP6 models across Chile. *Environmental Research Letters*, *20*(1), 014061. https://doi.org/10.1088/1748-9326/ad9d5b

Viglione, A., Laio, F., & Claps, P. (2007). A comparison of homogeneity tests for regional frequency analysis. *Water Resources Research*, *43*(3). https://doi.org/10.1029/2006WR005095