



UNIVERSIDAD NACIONAL AUTÓNOMA DE MÉXICO

FACULTAD DE INGENIERÍA

**Regionalización del modelo GR2M para la
estimación de los volúmenes de agua de entrada
a las presas del Sistema Cutzamala mediante
técnicas de Machine Learning**

TESIS

Que para obtener el título de

Ingeniero Geofísico

P R E S E N T A

Edgar Orozco Aguilar

DIRECTOR DE TESIS

Dr. Saúl Arciniega Esparza



Ciudad Universitaria, Cd. Mx., 2025



**PROTESTA UNIVERSITARIA DE INTEGRIDAD Y
HONESTIDAD ACADÉMICA Y PROFESIONAL
(Titulación con trabajo escrito)**



De conformidad con lo dispuesto en los artículos 87, fracción V, del Estatuto General, 68, primer párrafo, del Reglamento General de Estudios Universitarios y 26, fracción I, y 35 del Reglamento General de Exámenes, me comprometo en todo tiempo a honrar a la institución y a cumplir con los principios establecidos en el Código de Ética de la Universidad Nacional Autónoma de México, especialmente con los de integridad y honestidad académica.

De acuerdo con lo anterior, manifiesto que el trabajo escrito titulado REGIONALIZACION DEL MODELO GR2M PARA LA ESTIMACION DE LOS VOLUMENES DE AGUA DE ENTRADA A LAS PRESAS DEL SISTEMA CUTZAMALA MEDIANTE TECNICAS DE MACHINE

LEARNING que presenté para obtener el título de INGENIERO GEOFÍSICO es original, de mi autoría y lo realicé con el rigor metodológico exigido por mi Entidad Académica, citando las fuentes de ideas, textos, imágenes, gráficos u otro tipo de obras empleadas para su desarrollo.

En consecuencia, acepto que la falta de cumplimiento de las disposiciones reglamentarias y normativas de la Universidad, en particular las ya referidas en el Código de Ética, llevará a la nulidad de los actos de carácter académico administrativo del proceso de titulación.

EDGAR OROZCO AGUILAR
Número de cuenta: 317123087

“Juro por mi vida y por mi amor a ella que jamás viviré para otro hombre ni permitiré que ningún hombre viva para mí.”

— Ayn Rand, *La Rebelión de Atlas*

Agradecimientos

Agradezco profundamente al Dr. Saúl Arciniega Esparza, director de esta tesis, por haberme guiado desde el primer día. Su acompañamiento constante, su orientación académica y su apoyo en cada etapa marcaron de forma decisiva este trabajo. Extiendo también mi gratitud por su guía más allá del ámbito académico, por ayudarme a encontrar claridad en mi camino profesional y personal. Mi agradecimiento hacia él es, y será siempre, eterno.

Al comité de sinodales —Dr. Jorge López Alvis, Dr. José Antonio Hernández Espriú, M.C. Manuel Ortiz Osio y el Ing. Gabriel Salinas Calleros— por su tiempo, su dedicación y por los comentarios y sugerencias que enriquecieron profundamente esta investigación.

Al Grupo de Hidrología de la Facultad de Ingeniería, por brindarme un espacio de trabajo y un entorno de aprendizaje durante todo este proceso, así como por su constante disposición para compartir conocimiento y experiencia.

A mi mamá, Isabel, por su amor inagotable, su fortaleza y el ejemplo que ha guiado cada uno de mis pasos; a Alejandro, por su compañía y por el apoyo constante que hizo posible este camino; y a mi hermano, Alberto, por su presencia constante y su manera única de recordarme quién soy. A ellos, y al resto de mi familia, gracias por la fortaleza y la sabiduría que me han brindado.

A Jacqueline Garduño, cuya presencia ha sido una brújula en los momentos decisivos. Su compañía en este proceso, su fortaleza inquebrantable y la claridad con la que, incluso entre dudas, señaló la dirección hacia la que valía la pena avanzar fueron una guía invaluable en este camino.

A Erick, Jesús, Mario, Jessica y Samara, quienes compartieron conmigo los años de universidad y la hicieron más ligera, más divertida y más significativa; y a Emiliano, Mary y Pame, cuya presencia, amistad y apoyo en distintas etapas contribuyeron también a este camino. A todas las personas que, de una u otra forma, estuvieron presentes en este camino. Su compañía, su amistad y los momentos compartidos forman parte del trayecto que me trajo hasta aquí.

ÍNDICE

Resumen.....	1
Abstract	2
I. Introducción.....	3
A. Objetivos	5
II. Antecedentes.....	6
A. El Sistema Cutzamala	6
III. Marco teórico.....	9
A. Fundamentos de la modelación hidrológica.....	9
B. Calibración y análisis de incertidumbre en modelos hidrológicos.....	11
C. Regionalización de parámetros hidrológicos	12
D. Datos satelitales y percepción remota aplicados a hidrología.....	13
E. Principio del Balance Hídrico	15
F. Modelo Hidrológico.....	15
G. Optimización.....	19
H. Coeficiente de Eficiencia de Nash-Sutcliffe (NSE).....	22
I. Normalización y Estandarización de los datos.....	22
J. Análisis de Correlación.....	24
K. Aprendizaje de Máquina (<i>Machine Learning</i>).....	25
1. Sobreajuste (<i>overfitting</i>).....	25
2. Algoritmos de Regresión en el Aprendizaje Supervisado.....	26
L. Validación cruzada	31
IV. Descripción de la zona de estudio	32
A. Ubicación	32
B. Geología.....	33
C. Clima	35
D. Tipo de Suelo	35
E. Hidrogeología e Hidrología	36
V. Bases de datos y metodología	37

A.	Bases de datos.....	37
B.	Metodología	41
1.	Descripción general de la metodología.....	41
2.	Calibración del modelo en cuencas con información hidrométrica	42
3.	Validación en las cuencas del Sistema Cutzamala	42
4.	Preprocesamiento de datos y Análisis de Correlación	44
5.	Calibración Regional del modelo GR2M	45
6.	Estimación espacial de parámetros	51
VI.	Resultados y discusión.....	54
A.	Calibración del modelo GR2M para cada cuenca.....	54
B.	Análisis de correlación entre variables y parámetros del modelo.....	55
C.	Regionalización de los parámetros x_1 y x_2 del modelo GR2M	57
D.	Comparación del caudal simulado en cuencas con registros hidrométricos	63
E.	Validación cruzada	68
F.	Simulación del caudal de entrada a las presas del Sistema Cutzamala.....	75
VII.	Conclusiones y trabajo futuro	81
A.	Conclusiones.....	81
B.	Trabajos futuros y líneas de investigación	83
VIII.	Anexo	86
IX.	Referencias	89

Resumen

El Sistema Cutzamala constituye una de las infraestructuras hidráulicas más relevantes de México, ya que aporta hasta un 30% del agua potable consumida en la Zona Metropolitana del Valle de México. Sin embargo, la limitada información hidrométrica disponible complica la gestión del almacenamiento de sus principales presas, por lo cual resulta fundamental contar con metodologías que permitan estimar los volúmenes de entrada a partir de información existente de cuencas vecinas.

En este trabajo se aplicó el modelo hidrológico conceptual GR2M debido a su sencillez y a su capacidad para simular el balance hídrico mensual con un número reducido de parámetros, lo que lo hace adecuado en contextos con información hidrométrica limitada. El modelo fue alimentado con productos satelitales (CHIRPS y FLDAS), y sus parámetros se regionalizaron en cuencas aforadas y no aforadas a partir del entrenamiento realizado con información de cuencas vecinas. Dado que las técnicas tradicionales de transferencia de parámetros basadas en regresiones lineales o métodos espaciales suelen presentar limitaciones al capturar relaciones no lineales entre variables fisiográficas e hidrológicas, se propusieron nuevos algoritmos de aprendizaje automático (SVR y XGBoost). Estos permiten modelar interacciones más complejas y aumentar la capacidad predictiva del modelo, especialmente cuando existen pocos registros de datos observados en la región. Los resultados muestran que el modelo calibrado individualmente alcanza valores superiores a 0.8 con el criterio de eficiencia de Nash-Sutcliffe en la mayoría de las cuencas, lo que confirma su capacidad para reproducir el comportamiento hidrológico mensual. Durante la regionalización, los métodos de aprendizaje automático superaron a las regresiones tradicionales, aunque sin alcanzar la precisión de la calibración directa. La estimación espacial, por su parte, permitió generar mapas continuos de parámetros útiles para la simulación en cualquier cuenca no aforada dentro del dominio.

Con esta aproximación, se lograron estimar los caudales de entrada a las presas Valle de Bravo, Villa Victoria y El Bosque, siendo las primeras dos las que muestran un mejor desempeño al comparar las estimaciones contra un balance de agua derivado de datos de la presa. Por otro lado, el bajo desempeño en la presa El Bosque podría estar asociado a problemas en sus registros y a pérdidas no contabilizadas debido a infiltraciones. Derivado del análisis de resultados de estas tres presas, se observa que los años más críticos se han presentado en 1997 y 2005, asociados a años con poca precipitación.

El enfoque propuesto permite evaluar la capacidad de predicción de las distintas metodologías, así como explorar implicaciones hidrológicas de los patrones observados, con el objetivo de ofrecer indicios para mejorar la seguridad hídrica de la zona de estudio.

Abstract

The Cutzamala System is one of the most important hydraulic infrastructures in Mexico, providing up to 30% of the drinking water consumed in the Mexico Valley Metropolitan Zone. However, the limited hydrometric information available complicates the management of storage at the system's main dams, which makes it essential to develop methodologies that can estimate inflow volumes based on information from neighboring basins.

In this study, the GR2M conceptual hydrological model was employed due to its simplicity and capacity to simulate monthly water balances with a limited number of parameters, rendering it suitable for contexts with scarce hydrometric data. The model was fed with satellite products (CHIRPS and FLDAS) and trained with information from neighboring basins, with the aim of regionalizing its parameters in ungauged basins.

In light of the frequently encountered limitations of conventional parameter transfer methodologies, such as linear regressions or spatial analysis techniques, in adequately capturing the nonlinear interrelationships between physiographic and hydrological variables, the integration of machine algorithms, specifically Support Vector Machines (SVR) and Extreme Gradient Boosting (XGBoost), was employed. These models enable the representation of more intricate interactions and enhance the predictive capability of the model, particularly in regions with limited observed data.

The results show that the individually calibrated model achieves values greater than 0.8 according to the Nash-Sutcliffe efficiency criterion in most catchments, validating its ability to represent monthly hydrological behavior. In regionalization, machine learning methods outperformed traditional regressions, although they did not surpass the accuracy obtained through direct calibration. Spatial estimation, on the other hand, allowed the generation of continuous parameter maps that are useful for simulation in any basin within the domain.

This approach enabled the inflows to the Valle de Bravo, Villa Victoria, and El Bosque dams to be estimated, with the first two showing better performance when the estimates were compared with a water balance derived from dam data. Conversely, the inferior performance of the El Bosque dam could be associated with problems in the dam's records and unaccounted losses due to leakage. Based on the analysis of these three dams, the most critical years were 1997 and 2005, which coincided with years of low precipitation.

The proposed approach makes it possible to evaluate the predictive capacity of different methodologies, as well as to explore the hydrological implications of the observed patterns, with the aim of providing useful insights for water security in the study region.

I. Introducción

La disponibilidad y calidad de la información hidrométrica constituyen un reto central en la hidrología contemporánea. En gran parte del mundo, principalmente en países en desarrollo, los registros de caudal son limitados, discontinuos o inexistentes. Esta carencia dificulta la comprensión de los procesos hidrológicos a escala de cuenca y limita la capacidad para apoyar la toma de decisiones en escenarios de creciente variabilidad climática y aumento en la demanda de agua.

Las presas constituyen infraestructuras clave para el almacenamiento y distribución del agua, así como para la generación de energía y la regulación de avenidas (National Geographic Society, 2024). Su operación depende en gran medida de una estimación adecuada de las entradas de agua, es decir, de los caudales que llegan a los embalses desde sus cuencas tributarias. Sin registros directos de caudal, resulta complejo anticipar los volúmenes disponibles y, por lo tanto, planificar de forma eficiente su aprovechamiento (Boulange et al., 2021). Esta limitante hace que el desarrollo de métodos que permiten estimar aportaciones hídricas en cuencas no aforadas se vuelve esencial para la gestión de los sistemas de presas (Hrachowitz et al., 2013). Dicha necesidad es particularmente crítica, ya que la operación segura y eficiente de estas infraestructuras depende de predicciones de caudal confiables, especialmente ante la incertidumbre que plantean el cambio climático y el envejecimiento de las infraestructuras (Ho et al., 2017).

Entre las herramientas más empleadas para este propósito destacan los modelos hidrológicos, que representan de manera simplificada los procesos del ciclo hidrológico. Su aplicación permite simular la relación entre la precipitación y la escorrentía, considerando características físicas y climáticas de las cuencas (Devia et al., 2015). Sin embargo, el uso de estos modelos requiere de una gran cantidad de datos para su ejecución, así como datos observados para su calibración, es decir, el ajuste de sus parámetros para lograr que las simulaciones reproduzcan con la mayor precisión posible las observaciones disponibles (Rajat et al., 2021). Aun así, la incertidumbre en los datos de entrada y en la propia estructura de los modelos plantea desafíos adicionales para su aplicación práctica (Beven et al., 2024).

La percepción remota ha emergido en las últimas décadas como una fuente complementaria y cada vez más confiable de información hidrológica (Huang et al., 2020). El acceso a datos satelitales de precipitación, temperatura, humedad del suelo y cobertura vegetal ha permitido reducir la dependencia de los registros de campo. Estos productos satelitales no solo extienden la cobertura espacial y temporal de la información, sino que también facilitan la

aplicación de modelos en extensas regiones sin estaciones de medición, contribuyendo así a disminuir las limitaciones impuestas por la falta de datos *in situ*.

En este marco, la problemática de los sistemas de almacenamiento de agua adquiere particular relevancia en México. El Sistema Cutzamala, que abastece entre el 25 y el 30% del agua potable de la Zona Metropolitana del Valle de México, es uno de los casos más importantes del país. Este sistema carece de registros hidrométricos en sus principales presas (Valle de Bravo, Villa Victoria y El Bosque), lo que dificulta la planeación y operación del sistema. Además, episodios recientes de sequías, como la de 2021, derivaron en una reducción en los niveles de almacenamiento, lo que evidenció su vulnerabilidad ante la variabilidad climática y la necesidad de contar con herramientas para mejorar la estimación de aportaciones hídricas.

En este contexto, la presente tesis propone una metodología para combinar un modelo hidrológico conceptual con datos de percepción remota y algoritmos de machine learning para estimar los caudales de entrada al Sistema Cutzamala. Además de aportar un marco metodológico para la regionalización de parámetros en cuencas no aforadas, se plantea el desarrollo de una herramienta que permita el monitoreo continuo de caudales en el sistema sin la necesidad de mediciones directas en sitio. Este enfoque busca contribuir a la generación de información hidrológica en regiones con escasa instrumentación y fortalecer la gestión del recurso hídrico en una de las regiones más críticas del país.

A. Objetivos

Objetivo General

- Determinar los volúmenes de entrada de las principales presas del Sistema Cutzamala (Villa Victoria, Valle de Bravo y El Bosque) mediante modelación hidrológica y datos derivados de percepción remota.

Objetivos Particulares

- Implementar y evaluar los datos del producto global FLDAS para la simulación hidrológica en las presas del Sistema Cutzamala y cuencas vecinas.
- Caracterizar geomorfológicamente las cuencas del Sistema Cutzamala y sus cuencas vecinas, así como obtener parámetros a partir de productos globales y de percepción remota.
- Calibrar y validar el modelo hidrológico GR2M para simular los caudales mensuales en las cuencas vecinas al Sistema Cutzamala.
- Regionalizar los parámetros $x1$ y $x2$ del modelo GR2M para la simulación en cuencas sin datos observados de caudal comparando distintas metodologías: Optimización, Regresión Multilineal, modelos de Machine Learning (SVR y XGBoost) y Regresión Espacial.
- Comparar las estimaciones del caudal en las presas del Sistema Cutzamala a partir del modelo GR2M regionalizado contra las estimaciones obtenidas con un balance de agua en el embalse de las presas.

II. Antecedentes

A. El Sistema Cutzamala

El Sistema Cutzamala es uno de los sistemas hidráulicos más importantes de México y de América Latina, encargado de abastecer una parte significativa del agua potable que consume la Zona Metropolitana del Valle de México (ZMVM), una de las áreas urbanas más pobladas del mundo, con más de 21 millones de habitantes para el año 2020 (INEGI, 2020). Diseñado para trasladar agua desde zonas del oeste con mayor disponibilidad hacia regiones con alta demanda y escasez relativa (Escolero, 2016), este sistema se extiende a lo largo de los estados de Michoacán, Estado de México y la Ciudad de México, atravesando regiones montañosas con una compleja orografía, bombeando el agua por más de 1 km de diferencia de elevación (Huerta-Vergara et al., 2025).

La historia del Sistema Cutzamala se remonta a finales de la década de 1970, cuando se iniciaron los trabajos de planeación y construcción como respuesta a la creciente presión sobre el acuífero del Valle de México por el aumento de la población de la Ciudad de México, debido, principalmente, a la inmigración, así como a la necesidad de diversificar las fuentes de abastecimiento. Como obra del gobierno federal, su desarrollo ha sido progresivo y se ha dividido en cuatro etapas generales, cada una con el objetivo de aumentar la capacidad de suministro de agua ante el acelerado crecimiento poblacional y urbano de la zona metropolitana (World Bank Group, 2015).

El sistema comprende una red altamente sofisticada de infraestructura hidráulica que incluye presas, estaciones de bombeo, plantas de tratamiento y acueductos. En conjunto, estos componentes permiten el transporte de agua desde las partes altas de la cuenca del río Cutzamala, ubicadas a más de 2000 metros sobre el nivel del mar, hasta el Valle de México, en un recorrido que supera los 300 kilómetros y que requiere vencer desniveles de más 1100 metros (World Bank Group, 2015).

El sistema aprovecha el agua de la cuenca alta del río Cutzamala, el cual está conformado por las presas Tuxpan y El Bosque, ubicadas en el municipio de Zitácuaro, Michoacán; y por las presas Colorines, Ixtapan del Oro, Valle de Bravo, Villa Victoria y Chilesdo, situadas en diferentes municipios del Estado de México. Cabe destacar que la presa Chilesdo se construyó para aprovechar el caudal del río Malacatepec, mientras que las demás formaban parte originalmente del Sistema Hidroeléctrico Miguel Alemán (World Bank Group, 2015).

Actualmente, el Sistema Cutzamala aporta entre el 20% y el 36% del agua que consume la Ciudad de México, abasteciendo de forma directa a 13 de las 16 alcaldías de la capital y a varios municipios conurbados del Estado de México. Su operación y mantenimiento están a cargo de la Comisión Nacional del Agua (CONAGUA) en coordinación con el Sistema de

Aguas de la Ciudad de México (SACMEX) y la Comisión del Agua del Estado de México (CAEM).

La infraestructura del Sistema Cutzamala trabaja de manera integrada para transportar el agua desde las cuencas altas de los ríos Cutzamala y Tuxpan hasta la ZMVM. Como se observa en la Figura 1, el sistema inicia en la presa Tuxpan, continuando a las presas de El Bosque, Valle de Bravo y Villa Victoria, a través de un recorrido que incluye presas más pequeñas como Ixtapan del Oro, Sabaneta y Donato Guerra. Desde esta última, el agua se desvía hacia el acueducto de Chilesdo y el Canal de Villa Victoria, siguiendo su camino hacia la Ciudad de México, atravesando desniveles en la Sierra de las Cruces.

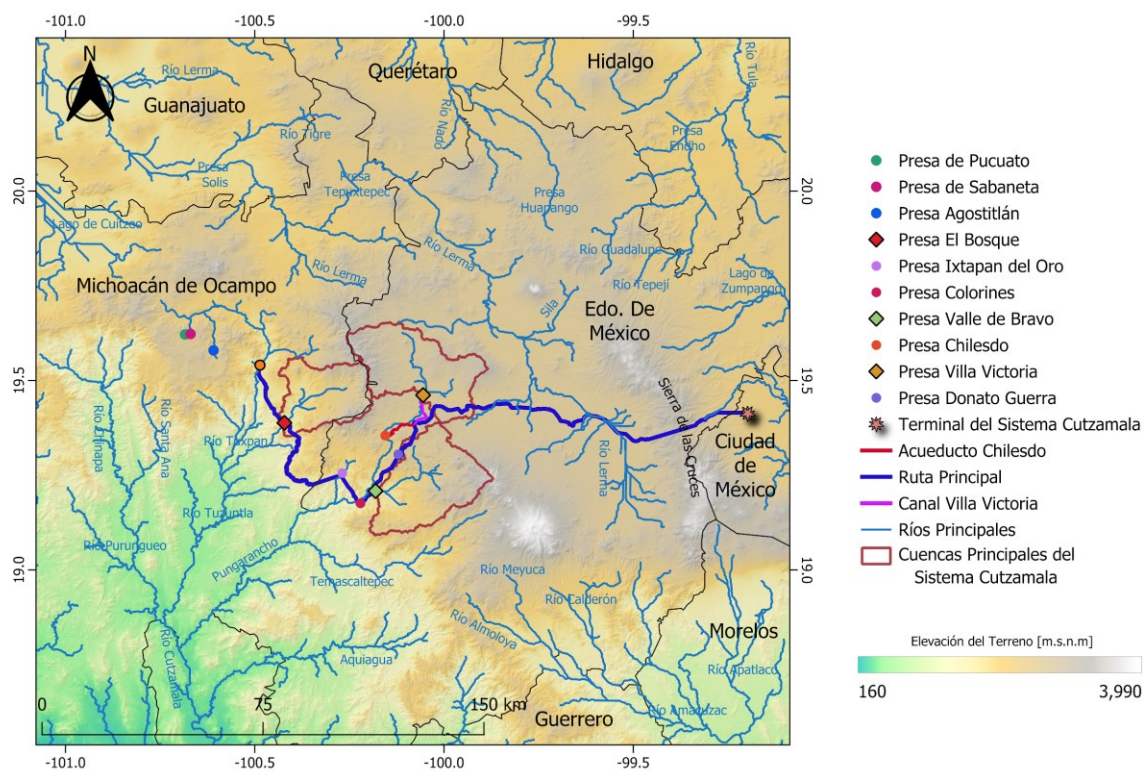


Figura 1. Infraestructura del Sistema Cutzamala. (Fuente: Elaboración propia a partir de datos de la Auburn University, 2022)

El sistema de transporte facilita el movimiento del agua entre las diferentes presas y cuencas. Por ejemplo, los canales Tuxpan - El Bosque y El Bosque - Colorines, con una longitud total de 73.7 km, permiten el trasvase por gravedad, superando las barreras topográficas mediante estructuras como sifones y túneles. La red de conducción también enfrenta desafíos como las filtraciones en la presa El Bosque y la sedimentación en los vasos de las presas, problemas que requieren un monitoreo constante (World Bank Group, 2015).

Una vez almacenada, el agua es impulsada a través de seis plantas de bombeo que en conjunto permiten vencer el desnivel mencionado anteriormente, utilizando potentes motores de hasta 22,000 caballos de fuerza. Estas plantas consumen aproximadamente 2200 GWh (Gigavatios hora) anuales, lo que representa el 80% de los costos operativos del sistema. Posteriormente, el agua es conducida a través de 322 km de tuberías, canales y túneles, incluyendo el túnel Analco - San José, una obra crítica de 15 km que atraviesa la Sierra de las Cruces. Este túnel, construido en 1978, requiere un mantenimiento constante debido a grietas y filtraciones, lo que subraya la importancia de evaluar su estabilidad geológica y estructural.

Finalmente, el agua llega a la planta potabilizadora Los Berros, ubicada a 2560 msnm, donde se somete a un tratamiento convencional de floculación, sedimentación, filtración y cloración. Con una capacidad de diseño de $20 \text{ m}^3/\text{s}$, la planta enfrenta desafíos relacionados con la presencia de algas y contaminantes emergentes, lo que reduce su eficiencia y demanda mejoras tecnológicas. Una vez potabilizada, el agua es entregada en bloques a 13 alcaldías de la Ciudad de México y 14 municipios del Estado de México a través de puntos clave como el tanque Pericos y la lumbrera Cruz de la Misión.

III. Marco teórico

A. Fundamentos de la modelación hidrológica

Un modelo hidrológico puede definirse como una representación simplificada de los procesos que ocurren en un sistema hidrológico real, como la precipitación, la infiltración, la escorrentía o el almacenamiento de agua. El modelo más adecuado es aquel que reproduce con mayor precisión el comportamiento del sistema empleando el menor número posible de parámetros y evitando una complejidad innecesaria. En el caso del caudal, un modelo puede entenderse como un conjunto de ecuaciones destinadas a estimarlo a partir de diversos parámetros que describen las características de una cuenca (van Esse et al., 2013).

Aunque los datos de precipitación y el área de drenaje constituyen la base de cualquier modelo, también suelen considerarse otras variables relevantes como las propiedades del suelo, la cobertura vegetal, la topografía, la humedad inicial del terreno y las características del acuífero. Actualmente, los modelos hidrológicos son herramientas fundamentales para la gestión de los recursos y ambientales (Devia, 2015).

Los modelos lluvia-escorrentía pueden clasificarse de distintas maneras. Una de ellas considera cómo se representan en el modelo los parámetros y las entradas en el espacio y el tiempo, lo que lleva a distinguir entre modelos agrupados y modelos distribuidos. También es común diferenciar entre modelos deterministas, que para un conjunto de datos de entrada siempre producen el mismo resultado, y modelos estocásticos, en los que puede obtenerse más de una salida posible para una misma entrada (Yevjevich, 1987).

En un modelo agrupado, toda la cuenca se trata como una única unidad, ignorando la variabilidad espacial y calculando la salida sin considerar los procesos distribuidos en el espacio. En contraste, un modelo distribuido divide la cuenca en unidades más pequeñas, por ejemplo, celdas regulares o redes triangulares, en las que parámetros, entradas y salidas pueden variar de manera independiente (Horton et al., 2022).

Si bien estos modelos capturan la heterogeneidad espacial de la cuenca, su configuración tradicionalmente ha sido compleja y dependiente de software GIS propietario, lo que dificulta su accesibilidad y reproducibilidad. Recientemente, frameworks de código abierto y basados en Python, como pyTRIBS (Raming et al., 2025), están surgiendo para automatizar y estandarizar el flujo de trabajo de los modelos distribuidos, desde el preprocesamiento de datos hasta el análisis de resultados, superando estas barreras mediante scripts reproducibles y contenedores software.

Otra forma de clasificación se basa en el tratamiento del tiempo: los modelos estáticos prescinden de la dimensión temporal, mientras que los modelos dinámicos la incorporan. Sorooshian et al. (2008) clasificaron los modelos en modelos basados en eventos y modelos

continuos. Los primeros producen resultados solo para periodos específicos, mientras que los segundos producen resultados continuos.

Una de las clasificaciones ampliamente utilizada en hidrología distingue tres enfoques principales: modelos empíricos, modelos conceptuales y modelos basados en procesos físicos.

- Modelos Empíricos (Modelos basados en datos)

Estos modelos se fundamentan exclusivamente en la información contenida en los datos observados, sin representar explícitamente los procesos ni las características físicas del sistema. Por este motivo, también se conocen como modelos basados en datos. Suelen contribuir a partir de relaciones matemáticas derivadas de series temporales de entrada y salida, prescindiendo de ecuaciones que describen los procesos hidrológicos.

La validez de estos modelos está limitada al rango de condiciones cubierto por los datos con los que fueron calibrados o entrenados, lo que significa que no son confiables fuera de esos límites. Ejemplos típicos son el hidrograma unitario, los modelos de regresión y correlación, y diversas técnicas de aprendizaje automático como las redes neuronales artificiales (ANN) o la lógica difusa aplicadas a la hidrología computacional (Kambalimath, 2020).

- Modelos conceptuales (Modelos Paramétricos)

En este tipo de modelos se representan de manera esquemática los distintos procesos hidrológicos que intervienen en una cuenca. Habitualmente se emplean un sistema de reservorios interconectados que simboliza los diferentes componentes físicos, los cuales se recargan por procesos como la lluvia, infiltración o percolación, y se vacían mediante la evaporación, escorrentía o drenaje (Jajarmizadeh, 2014).

Los modelos conceptuales emplean ecuaciones semi empíricas y combinan la estimación de parámetros a partir de datos de campo con procedimientos de calibración. Este proceso de calibración, que suele implicar el ajuste de curvas, requiere series temporales extensas de información meteorológica e hidrológica. Una limitación de este enfoque es que los efectos de cambios en el uso de suelo no siempre pueden predecirse con certeza (Dal Molin et al., 2021).

Un ejemplo clásico es el Modelo de Cuenca Hidrográfica de Stanford IV (SWM), desarrollado por Crawford y Linsley en 1966, que emplea entre 16 y 20 parámetros. El modelo GR2M, utilizado en el presente proyecto, pertenece a esta categoría.

- Modelos basados en física

También denominados modelos mecanicistas o de base física, constituyen representaciones matemáticas idealizadas que describen los procesos hidrológicos a partir de principios físicos fundamentales, como las ecuaciones de conservación de masa y momento. Emplean variables de estado medibles que dependen tanto del tiempo como del espacio, y suelen

resolverse mediante métodos numéricos como diferencias o volúmenes finitos (Loague et al., 2006).

A diferencia de los modelos conceptuales o agregados, que dependen críticamente de la calibración de parámetros efectivos, la principal fortaleza de los modelos físicos reside en que sus parámetros, en teoría son medibles en campo (e. g., conductividad hidráulica, porosidad topografía). Sin embargo, en la práctica, la extrema heterogeneidad espacial de las cuencas hace que la caracterización precisa de estos parámetros sea un desafío mayor, pudiendo ser necesario un ajuste para representar correctamente la respuesta integrada de la cuenca (Loague et al., 2006). Esta base física les permite simular la respuesta hidrológica de manera distribuida y ofrece el potencial de realizar predicciones bajo condiciones diferentes a las observadas históricamente, ampliando su aplicabilidad para estudiar cambios de uso de suelo o clima. Un ejemplo de esta clase de modelos es el Sistema Hidrológico Europeo (SHE) posteriormente comercializado como MIKE SHE (Abbott et al., 1986).

B. Calibración y análisis de incertidumbre en modelos hidrológicos.

La calibración de modelos hidrológicos constituye un proceso esencial para garantizar que las simulaciones reproduzcan adecuadamente el comportamiento observado en una cuenca. En términos generales, calibrar un modelo implica ajustar los parámetros de su estructura conceptual o matemática hasta minimizar la discrepancia entre las variables simuladas (por ejemplo, caudal) y las observadas. Para ello, se emplean algoritmos de optimización junto con criterios de desempeño como el coeficiente de Nash-Sutcliffe, el error absoluto medio (MAE) o el sesgo porcentual (PBIAS) (Muste et al., 2012).

Uno de los principales retos en la calibración es la equifinalidad (Beven, 2006), es decir, la existencia de múltiples combinaciones de parámetros que producen resultados similares, dificultando la identificación de un único conjunto óptimo (Knoben et al., 2020). Este fenómeno resalta que la calibración no solo busca un ajuste matemático, sino también evaluar la plausibilidad física de los parámetros y la estructura del modelo (Muste et al., 2012).

El análisis de incertidumbre complementa la calibración de modelos hidrológicos al cuantificar cómo los errores en los datos de entrada, la estructura del modelo o los parámetros influyen en las salidas (Knoben et al., 2020). Este proceso busca estimar el rango de posibles valores dentro del cual se espera que se encuentren los resultados con un nivel de confianza determinado (Muste et al., 2012). En hidrología, una de las metodologías más empleadas para este fin es la simulación de Monte Carlo, que permite construir intervalos de predicción mediante la ejecución del modelo bajo múltiples combinaciones de parámetros (Beven, 2014). Aunque existen enfoques más recientes basados en técnicas de machine learning que reducen los costos computacionales de estos análisis (Lim, 2022; Saha, 2024).

C. Regionalización de parámetros hidrológicos

La regionalización de parámetros hidrológicos es una técnica utilizada para estimar el comportamiento hidrológico en cuencas no aforadas, es decir, aquellas que carecen de registros directos de caudal. Este enfoque se fundamenta en la transferencia de información desde cuencas con mediciones hacia cuencas sin mediciones, con el objetivo de calibrar modelos y generar predicciones de caudal. La regionalización permite trasladar parámetros calibrados de modelos hidrológicos desde cuencas donantes hacia cuencas receptoras en función de su similitud física, climática o hidrológica (Beck et al., 2020; Guo et al., 2021).

El proceso inicia con la calibración de un modelo hidrológico en una o varias cuencas con datos observados, lo que permite estimar un conjunto de parámetros que representan las condiciones y procesos hidrológicos de dichas cuencas. Posteriormente, esos parámetros se asignan a cuencas no aforadas siguiendo diferentes criterios de transferencia. Los métodos más habituales incluyen la proximidad geográfica, bajo el supuesto de que cuencas cercanas comparten condiciones climáticas y fisiográficas; la similitud de atributos físicos, como topografía, uso de suelo, geología y propiedades del suelo; y la coincidencia en firmas hidrológicas, que son indicadores estadísticos del régimen de caudales, como el índice de escorrentía, estacionalidad o duración de crecidas (Guo et al., 2021).

Existen diferentes enfoques para la regionalización. Guo et al. (2020) clasifican los métodos en diferentes categorías:

- Métodos Basados en Similitud Física

Estos métodos seleccionan cuencas donantes con características físicas y climáticas similares a la cuenca no aforada objetivo, asumiendo que cuencas con atributos comparables tendrán respuestas hidrológicas análogas. Entre las técnicas utilizadas destacan los índices de similitud, que calculan distancias euclidianas o ponderadas entre descriptores como pendiente, tipo de suelo o cobertura vegetal, y el agrupamiento (clustering) mediante algoritmos como *k-means* o *fuzzy c-means* para identificar regiones hidrológicamente homogéneas. Su principal ventaja radica en capturar procesos hidrológicos subyacentes, lo que mejora la representatividad de los parámetros transferidos. Sin embargo, su efectividad depende de la selección adecuada de variables y escalas espaciales, y puede verse limitada en zonas con alta homogeneidad no capturada por los descriptores disponibles.

- Métodos Basados en Proximidad Espacial

Estos enfoques parten del principio de que cuencas cercanas geográficamente tienden a exhibir comportamientos hidrológicos similares. Las técnicas más comunes incluyen el vecino más cercano, donde los parámetros se transfieren directamente de la cuenca aforada más próxima, y métodos de interpolación espacial como kriging o ponderación por distancia inversa (IDW), que asignan pesos a múltiples cuencas donantes según su proximidad. Su principal fortaleza es la simplicidad y bajo costo computacional, especialmente útil en regiones

con alta densidad de estaciones. No obstante, su rendimiento se degrada en áreas con fuertes gradientes climáticos o topográficos, donde la proximidad geográfica no garantiza similitud hidrológica.

- Métodos Basados en Regresión

Estos métodos establecen relaciones funcionales entre los parámetros del modelo hidrológico y atributos físicos o climáticos de la cuenca (ej.: área, índice de aridez, pendiente media). Los enfoques emplean regresión lineal múltiple (RLM), mientras que técnicas avanzadas como *Gaussian Process Regression* (GPR) o *Random Forest* permiten modelar relaciones no lineales y reducir incertidumbres. Son particularmente valiosos en escalas regionales con abundantes datos, ya que pueden generalizar patrones complejos. Sin embargo, requieren una cuidadosa selección de predictores para evitar sobreajuste, y su desempeño puede verse comprometido en cuencas con condiciones extremas no representadas en los datos de calibración.

- Métodos Basados en Firmas Hidrológicas

Estas técnicas utilizan indicadores cuantitativos del comportamiento hidrológico, como curvas de duración de caudales (FDC), índices de estiaje o frecuencia de crecidas, para transferir información a cuencas no aforadas. Un enfoque común es la calibración inversa mediante el método de Monte Carlo, donde se generan múltiples conjuntos de parámetros y se seleccionan aquellos que reproducen las firmas observadas en cuencas donantes. Su ventaja es la capacidad de incorporar múltiples aspectos del régimen hidrológico (ej.: flujos bajos y altos), pero su implementación puede ser computacionalmente costosa y sensible a errores en los datos de entrada.

Una vez transferidos o estimados los parámetros para la cuenca no aforada, el modelo se ejecuta utilizando las series disponibles de variables de entrada, como precipitación y evapotranspiración, para generar las predicciones de caudal. El éxito de este procedimiento depende de la calidad y representatividad de las cuencas donantes, así como de la idoneidad del método de regionalización elegido (Guo et al., 2021; Alebachew et al., 2014).

La regionalización de parámetros es especialmente valiosa en regiones con infraestructura de monitoreo limitada, ya que proporcionan información esencial para gestión de recursos hídricos, diseño de infraestructura hidráulica, elaboración de mapas de riesgo, implementación de sistemas de alerta temprana y planificación de medidas para el manejo de sequías.

D. Datos satelitales y percepción remota aplicados a hidrología

La percepción remota se define como la adquisición de información sobre un objeto o proceso a distancia sin necesidad de contacto directo, en contraste con las mediciones *in situ*, donde los instrumentos de registros se encuentran en contacto con el sistema observado (Schultz,

1988). En el ámbito de la hidrología, esta técnica se basa principalmente en el uso de la radiación electromagnética, que constituye el medio fundamental a través del cual se realizan las observaciones (Tang et al., 2009). Dicha radiación puede ser emitida o reflejada por distintos elementos de la superficie terrestre, como cuerpos de agua, suelos, vegetación o nubes, lo que permite inferir información sobre ellos a partir de diferentes longitudes de onda. Existen dos grandes categorías de percepción remota: pasiva y activa. La primera mide únicamente la energía natural emitida o reflejada por los efectos, como ocurre en la fotografía aérea convencional o en sensores ópticos instalados en satélites. La segunda implica la emisión de una señal artificial (por ejemplo, microondas en radares), de la cual se mide la fracción reflejada terrestre; un ejemplo de esta categoría son los radares meteorológicos (Schultz, 1988). En teoría, todo el espectro electromagnético podría emplearse en percepción remota, aunque en la práctica sólo ciertas regiones, denominadas ventanas atmosféricas, permiten la transmisión eficaz de la señal a través de la atmósfera.

Cada tipo de sensor opera en diferentes bandas espectrales, visible, infrarrojo, infrarrojo cercano, microondas, y aporta información específica sobre procesos hidrológicos. Por ejemplo, las microondas pasivas son especialmente útiles para la estimación de la humedad del suelo debido a su capacidad para penetrar parcialmente la superficie, mientras que los sensores ópticos permiten caracterizar la cobertura vegetal y apoyar la estimación de la evapotranspiración mediante índices espectrales (Becker, 2006).

La percepción remota satelital presenta ventajas significativas frente a las observaciones tradicionales: ofrece mediciones de área en lugar de puntos individuales, permite obtener datos en regiones remotas donde no existen estaciones de monitoreo, proporciona resoluciones espaciales y temporales adecuadas para el análisis multiescala y genera productos en formato digital que facilitan su integración en sistemas de información geográfica y modelos matemáticos (Schultz, 1988; Becker, 2006). En los últimos años, esta capacidad ha evolucionado hacia la integración directa con modelos hidrológicos permitiendo desarrollar representaciones más realistas de los procesos del ciclo hidrológico e incluso la construcción de gemelos digitales de cuenca para evaluar escenarios futuros y apoyar la gestión adaptativa del agua (Fatichi et al., 2024). De esta manera, la percepción remota ha pasado de ser un recurso complementario para convertirse en un componente esencial de la hidrología moderna, ya que contribuye a superar una de sus principales limitaciones históricas: la escasez y dispersión de datos *in situ*.

Entre los productos satelitales desarrollados para aplicaciones hidrológicas destaca el *Famine Early Warning Systems Network Land Data Assimilation System* (FLDAS), creado por la NASA y sus colaboradores como herramienta para monitorear y predecir condiciones de sequía y disponibilidad hídrica en regiones vulnerables (McNally et al., 2017). FLDAS integra datos de teledetección y datos terrestres con modelos globales usando técnicas de

asimilación y proporciona información clave sobre variables como la humedad del suelo, la evapotranspiración, el balance hídrico y el caudal superficial. Estos datos se generan a escalas espaciales y temporales que permiten evaluar tanto tendencias regionales como dinámicas locales, lo que convierte a FLDAS en un insumo fundamental para la investigación hidrológica en cuencas con escaso o nulo monitoreo directo.

Los productos satelitales como FLDAS se integran a un repositorio más amplio de insumos provenientes de sensores espaciales, los cuales abarcan variables fundamentales para la hidrología: precipitación, evapotranspiración, humedad del suelo, extensión de nieve, temperatura superficial terrestre y cobertura vegetal, entre otros (Tang et al., 2009).

La aplicación de estos datos en hidrología es diversa y de gran impacto. Los productos satelitales permiten la estimación de parámetros de modelos hidrológicos y mejoran la predicción de caudales, especialmente en cuencas no aforadas. Asimismo, facilitan la evaluación del almacenamiento en nieve y glaciares, el monitoreo de sequías, la cuantificación de la evapotranspiración, la detección de cambios en la calidad del agua y el seguimiento de inundaciones (Schultz, 1988, Tang et al., 2009). No obstante, a pesar de estas ventajas, subsisten desafíos importantes como la resolución insuficiente de algunos satélites para aplicaciones locales, la necesidad de validación continua con mediciones *in situ* y la propagación de incertidumbre en productos derivados, especialmente en regiones montañosas o semiáridas (Rodell et al., 2019; Terpanelli et al., 2023).

E. Principio del Balance Hídrico

El balance hídrico se fundamenta en el principio de la conservación de la masa, el cual establece que toda el agua que ingresa a un sistema debe salir o almacenarse. En el contexto hidrológico, este principio se expresa como la diferencia entre las entradas (precipitación, escorrentía superficial, infiltración, entre otras) y las salidas (evapotranspiración, descarga y extracción), reflejándose en el cambio de almacenamiento del sistema.

Este principio constituye la base conceptual sobre la cual se desarrollan la mayoría de los modelos lluvia-escorrentía, como el modelo GR2M, que representa el comportamiento hidrológico de una cuenca a partir de la relación entre los flujos de entrada, salida y almacenamiento.

F. Modelo Hidrológico

El modelo utilizado para la simulación hidrológica en las cuencas de interés es el GR2M (Génie Rural à 2 paramètres Mensuel), de origen francés que representa un enfoque de modelación hidrológica mensual, destacándose por su equilibrio entre simplicidad conceptual y capacidad predictiva. Desarrollado por Mouelhi et al. (2006), este modelo de dos parámetros (x_1 y x_2) ha demostrado su eficacia en la simulación del comportamiento hidrológico de

cuencas bajo diversas condiciones climáticas, desde regiones áridas hasta tropicales húmedas (Ditthakit et al., 2023, (Méndez et al., 2022).

El GR2M estructura el proceso de transformación lluvia-esorrentía en dos componentes: el primero corresponde al módulo de producción, donde el parámetro $x1$ está asociado a la capacidad máxima de almacenamiento del suelo y controla a la dinámica de almacenamiento y liberación de agua en la zona no saturada del suelo; la segunda componente del modelo involucra el proceso de enrutamiento, donde el parámetro $X2$ incorporara los intercambios hídricos subterráneos con sistemas externos a la cuenca. Este enfoque reconoce que, a escala mensual, los flujos laterales y verticales no contabilizados por las mediciones convencionales pueden tener un impacto sustancial en el balance hídrico.

En contraste con modelos como *abcd* (Alley, 1985) de cuatro parámetros o el modelo T-alpha de Thornthwaite and Mather (Arnel, 1992) de tres parámetros, el modelo GR2M logra una precisión comparable o mejor con una estructura más simple. Esta característica resulta particularmente valiosa en cuencas con limitaciones de información, donde modelos más complejos tienden a presentar problemas de sobreajuste.

La Figura 4 presenta un esquema conceptual del funcionamiento del modelo GR2M. El parámetro $x1$ muestra una correlación con las características fisiográficas de la cuenca, y sus valores típicos oscilan entre 140 mm para suelos con baja capacidad de retención (como en zonas áridas con suelos arenosos) hasta más de 2600 mm para cuencas con suelos profundos y alta capacidad de almacenamiento (como en regiones tropicales húmedas). El parámetro $X2$ actúa como un factor que compensa flujos no contabilizados de la interacción agua superficial-subterránea. Valores inferiores a 1 indican pérdidas netas del sistema (por ejemplo, por filtración profunda), mientras que valores superiores a 1 sugieren aportes externos (como recarga subterránea desde acuíferos regionales). Su rango típico varía entre 0.21 y 1.31.

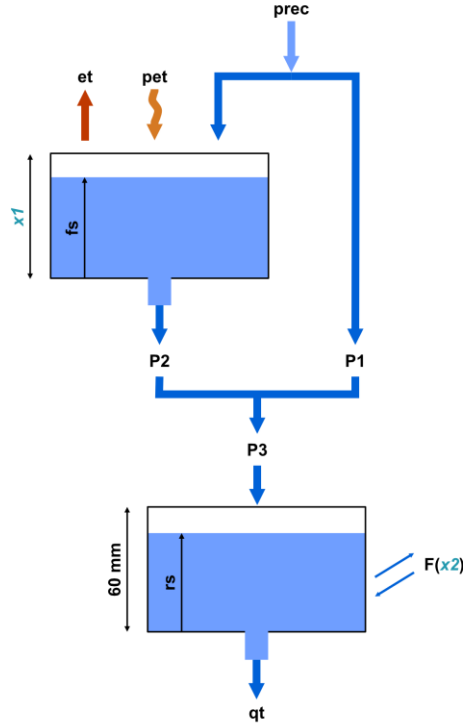


Figura 2. Esquema del modelo GR2M. Fuente: LuMod (<https://zaul-ae.gitlab.io/lumod-docs/models/gr2m/>)

Matemáticamente, las dinámicas internas en un paso de tiempo del modelo GR2M se representan mediante las siguientes ecuaciones:

$$V = \tanh \frac{prec(t)}{x1} \quad (1)$$

Esta ecuación calcula un factor de escorrentía basado en la precipitación mensual ($prec(t)$) y el parámetro $x1$ (capacidad máxima del almacenamiento de humedad del suelo).

- $\tanh$ Es la función tangente hiperbólica, que limita el valor entre 0 y 1.
- Si la precipitación es alta en comparación con $x1$, V se acerca a 1, indicando que el suelo está saturado y genera escorrentía.
- Si la precipitación es baja, V se acerca a 0, indicando que la mayoría del agua se infiltra.

$$W = \tanh \frac{pet(t)}{x1} \quad (2)$$

Similar a la anterior, pero usa la evapotranspiración potencial ($pet(t)$).

- Si la evapotranspiración potencial es alta en comparación con $x1$, W se acerca a 1, indicando una mayor pérdida de agua del suelo por evaporación.

- Si es baja, W se acerca a 0.

$$H1 = \frac{H + x1 * V}{1 + H * V/x1} \quad (3)$$

Actualiza el almacenamiento de humedad del suelo (H) después de agregar la precipitación.

- El numerador ($H + x1 * V$) representa el agua entrante.
- El denominador ajusta el almacenamiento para evitar superar la capacidad máxima $x1$.
- Esta es una forma no lineal de modelar la infiltración y retención del suelo.

$$H2 = \frac{H1 * (1 - W)}{1 + W * (1 - H1/x1)} \quad (4)$$

Actualiza el almacenamiento de humedad ($H1$) después de considerar las pérdidas por evapotranspiración.

- El término $(1 - W)$ reduce el almacenamiento según la evapotranspiración.
- El denominador ajusta la reducción para evitar valores negativos.

$$H = \frac{H2}{[1 + (\frac{H2}{x1})^{x2}]^{1/x2}} \quad (5)$$

Calcula el nuevo almacenamiento de humedad (H) después de la percolación (flujo de agua hacia capas más profundas).

- $x2$ Es un parámetro que representa el intercambio neto de agua subterránea con el exterior de la cuenca.
- Valores altos de $x2$ indican mayor aporte neto desde acuíferos externos.

$$fs(t) = \frac{H}{x1} \quad (6)$$

Esta ecuación calcula la fracción de almacenamiento de humedad del suelo en relación con su capacidad máxima ($x1$).

- Si $H = x1$, entonces $fs(t) = 1$: El suelo está completamente saturado.
- Si $H = 0$, entonces $fs(t) = 0$: El suelo está completamente seco.
- Valores intermedios ($0 < fs(t) < 1$) indican el grado de saturación en el modelo.

$$P1 = prec(t) + H - H1 \quad (7)$$

Calcula el exceso de precipitación ($P1$) que no puede ser absorbido por el suelo y se convierte en escorrentía directa.

- Si $H1$ es mayor que $H + prec(t)$, $P1$ será negativo (no ocurre en condiciones normales).
- En la práctica, $P1$ representa la escorrentía superficial.

$$P3 = P1 + P2 \quad (7)$$

Suma la escorrentía superficial ($P1$) y el flujo subterráneo ($P2$) para obtener la entrada total al almacenamiento de routing ($P3$).

$$S1 = S + P3 \quad (8)$$

Actualiza el almacenamiento de routing (S) agregando la entrada $P3$.

$$S2 = x2 * S1 \quad (9)$$

Aplica un intercambio de agua subterránea con el exterior del sistema.

- $x2$ Es un parámetro que representa ganancias ($x2 > 1$) o pérdidas ($x2 < 1$) de agua.
- Si $x2 = 1$, no hay intercambio.

$$S = S1 - qt(t) \quad (10)$$

Actualiza el almacenamiento de routing restando el caudal generado ($qt(t)$).

$$rs(t) = \frac{S}{60} \quad (11)$$

Calcula la fracción de llenado del almacenamiento de routing (S) en relación con su capacidad fija (60 mm).

- Si $S = 60$, entonces $rs(t) = 1$: el reservorio del río está lleno.
- Si $S = 0$, entonces $rs(t) = 0$: el reservorio está vacío.
- Valores intermedios indican el grado de llenado ($0 < rs(t) < 1$).

$$qt(t) = \frac{s^2}{S2 + x4} \quad (12)$$

Calcula el caudal generado por el almacenamiento de enrutamiento ($S2$).

- La ecuación es similar a un reservorio no lineal, donde el caudal aumenta con $S2$ pero está limitado por el denominador ($S2 + 60$).

G. Optimización

La optimización numérica constituye una disciplina fundamental en la ingeniería y ciencias computacionales, cuyo objetivo es sistemáticamente encontrar la mejor solución a un

problema de diseño, definido típicamente mediante la minimización o maximización de una función objetivo, sujeta a un conjunto de restricciones. Los problemas de optimización en ingeniería suelen caracterizarse por simulaciones intensivas desde el punto de vista computacional, la dificultad para estimar sensibilidades, cuando existen, la presencia de múltiples restricciones y, de manera crítica, la existencia de una multiplicidad de soluciones locales. Esta última característica, la multimodalidad, impulsa la necesidad de desarrollar estrategias de búsqueda global que eviten quedar atrapadas en óptimos locales subóptimos. Sin embargo, los algoritmos puramente globales conllevan a menudo un costo computacional prohibitivo para problemas reales, lo que ha motivado el desarrollo de métodos híbridos que combinen la eficiencia de los algoritmos locales con la exhaustividad de las búsquedas globales.

- Optimización Nelder-Mead

Dentro del espectro de los métodos de optimización sin derivadas, el algoritmo Nelder-Mead, también conocido como “método simplex del *Downhill*” (Nelder- Mead, 1965), ocupa un lugar prominente por su simplicidad conceptual y robustez práctica. Al no requerir el cálculo de derivadas de la función objetivo, se hace especialmente adecuado para optimizar funciones con superficies irregulares, ruidosas o no diferenciables, donde los métodos basados en gradientes pueden fallar o ser inaplicables.

Este algoritmo opera mediante la evaluación iterativa de un conjunto de $n+1$ puntos, conocidos como vértices de un simplex, en un espacio de parámetros de dimensión n . Un simplex es una figura geométrica que generaliza el concepto de triángulo o tetraedro a dimensiones superiores (Luersen, 2004). En cada iteración, el método explora sistemáticamente la superficie de la función objetivo a través de una serie de operaciones heurísticas diseñadas para reemplazar el peor vértice del simplex por uno mejor. Estas operaciones incluyen:

- Reflexión: Se proyecta el peor vértice a través del centroide del simplex opuesto, explorando una región prometedora en la dirección al punto de peor desempeño.
- Expansión: Si el punto reflejado es significativamente mejor, el simplex se expande en esa dirección para acelerar el progreso.
- Contracción: Si la reflexión no produce una mejora suficiente, el simplex se contrae a lo largo de esa dimensión, permitiendo un ajuste más fino de la búsqueda.
- Reducción (o Encogimiento): En casos donde la contracción no es efectiva, todo el simplex se reduce hacia su mejor vértice, reenfocando la búsqueda en la vecindad más prometedora.

El efecto acumulativo de estas operaciones es estirar el simplex a lo largo de las direcciones de descenso y ajustar su tamaño cerca de los óptimos locales. Si bien el método original fue concebido para dominios no acotados, en la práctica es incorporar el manejo de restricciones,

particularmente límites en las variables. Una estrategia común y sencilla es la proyección, mediante la cual cualquier variable que exceda sus límites se fija al valor del límite correspondiente.

Cabe destacar que, a pesar de su utilidad, el algoritmo Nelder-Mead clásico puede ocasionalmente fallar en converger a un óptimo local, por ejemplo, cuando el simplex colapsa en un subespacio de búsqueda. Para mitigar estas limitaciones, se han propuesto diversas mejoras, como la detección y reinicialización de simplex degenerados, así como criterios de convergencia más robustos que los basados únicamente en la dispersión de los valores de la función en los vértices (Luersen, 2004).

- Optimización L-BFGS-B

El algoritmo L-BFGS-B (*Limited-memory Broyden-Fletcher-Goldfarb-Shanno with Bounds*) pertenece a la familia de métodos cuasi-Newtonianos, diseñados para resolver problemas de optimización suaves y diferenciables de manera eficiente, especialmente en presencia de restricciones de tipo caja (*box constraints*) sobre los parámetros (Zhu et al., 1997).

Los métodos cuasi-Newton se basan en aproximar la matriz Hessiana (segunda derivada de la función objetivo) a partir de información proveniente de los gradientes calculados en iteraciones anteriores, sin necesidad de evaluarla explícitamente. Esto permite capturar la curvatura local de la superficie de optimización con un costo computacional significativamente menor que los métodos Newtonianos exactos.

La versión “*limited-memory*” del método BFGS está optimizada para problemas de gran dimensión, ya que, en lugar de almacenar y actualizar la matriz Hessiana completa, conserva únicamente una pequeña cantidad de pares de vectores de gradiente y desplazamiento de iteraciones recientes. Este enfoque reduce drásticamente el uso de memoria y mejora la estabilidad numérica, manteniendo una aproximación efectiva de la información de curvatura. Una de las principales ventajas de L-BFGS-B radica en su capacidad para manejar restricciones de tipo caja, que establecen límites inferiores y superiores para cada parámetro a optimizar. Este mecanismo resulta especialmente útil en contextos físicos o hidrológicos, donde los parámetros deben permanecer dentro de rangos físicamente plausibles.

El procedimiento general del algoritmo puede resumirse en las siguientes etapas:

- Cálculo del gradiente: se evalúa el gradiente de la función objetivo en el punto actual.
- Actualización cuasi-Newtoniana: se actualiza la aproximación de la Hessiana utilizando información de los gradientes y desplazamientos previos.
- Búsqueda lineal (*line search*): se determina el paso óptimo en la dirección de descenso que satisface las condiciones de Wolfe o Armijo (Wolfe, 1969; Wasi et al., 2021).
- Aplicación de límites: se ajustan los parámetros a sus límites inferiores o superiores en caso de excederlos.

El método continúa iterativamente hasta cumplir un criterio de convergencia típicamente basado en la magnitud del gradiente o en la variación relativa de la función objetivo.

Por su balance entre eficiencia, robustez y capacidad de manejo de restricciones, L-BFGS-B es uno de los métodos más utilizados en optimización numérica moderna, tanto en problemas de calibración de modelos hidrológicos como en contextos de *machine learning* y ajuste de modelos no lineales.

H. Coeficiente de Eficiencia de Nash-Sutcliffe (NSE)

El coeficiente de eficiencia de Nash-Sutcliffe (NSE) (Nash-Sutcliffe, 1970) es una métrica estadística ampliamente utilizada para evaluar el desempeño de modelos hidrológicos. Su propósito es cuantificar la correspondencia entre los caudales simulados y los observados, expresando qué tan bien las predicciones del modelo reproducen la variabilidad de los datos reales.

El NSE se define mediante la siguiente expresión:

$$NSE = 1 - \frac{\sum_{t=1}^T (Q_0^t - Q_m^t)^2}{\sum_{t=1}^T (Q_0^t - \underline{Q}_0)^2} \quad (13)$$

En donde:

- $\underline{Q}_0$ Es la media de las descargas observadas a lo largo del periodo analizado.
- Q_m^t Es el caudal simulado por el modelo en un tiempo t .
- Q_0^t Es el caudal observado en el tiempo t .

El valor del NSE puede variar desde $-\infty$ hasta 1. Un valor de 1 indica una concordancia perfecta entre los valores simulados y observados. Un valor de 0 señala que el modelo tiene la misma capacidad predictiva que utilizar la media de los valores observados como estimación constante. Valores negativos implican que el modelo es menos efectivo que una estimación basada en la media, lo que indica una representación deficiente del sistema hidrológico.

Esta métrica, por su simplicidad e interpretación directa, es ampliamente empleada como función objetivo en la calibración de modelos lluvia-escorrentía, ya que permite evaluar el ajuste entre las simulaciones del modelo y las observaciones hidrométricas reales.

I. Normalización y Estandarización de los datos

En el análisis de datos científicos y ambientales, es común explicar técnicas de normalización y estandarización para mejorar la estabilidad numérica durante el procesamiento y garantizar la comparabilidad entre variables con distintas unidades o escalas. Estas transformaciones resultan especialmente útiles al integrar los datos en modelos estadísticos o de *machine*

learning, ya que ayudan a evitar problemas derivados de problemas extremos, negativos o cercanos a cero.

Una de las técnicas más empleadas es la normalización Min-Max (Mazziotta-Pareto, 2022), que consiste en reescalar los valores de una variable a un intervalo definido, comúnmente entre $[0,1]$. Este método preserva la forma de la distribución original de los datos, lo que lo hace adecuado para modelos sensibles a la escala relativa de las variables. Sin embargo, su principal desventaja es su sensibilidad a valores atípicos (*outliers*), los cuales pueden comprimir artificialmente el rango de la mayoría de los datos, reduciendo su resolución relativa.

La fórmula general de la normalización Min-Max es la siguiente:

$$X_{norm} = \frac{X_t - X_{min}}{X_{max} - X_{min}} \quad (14)$$

Donde X_t es el valor de la variable en el tiempo o posición t , y X_{min} y X_{max} corresponden a los valores mínimo y máximo del conjunto de datos respectivamente.

Por otro lado, la estandarización *Z-score* (Patro, 2015) permite identificar valores que se desvían significativamente de la media de la serie, calculando para cada dato su puntuación Z , definida como la diferencia entre el valor de la variable en el tiempo (X_t) y la media (μ), dividida por la desviación estándar (σ):

$$Z_t = \frac{X_t - \mu}{\sigma} \quad (15)$$

La estandarización *Z-score* supone que los datos provienen de una distribución aproximadamente normal, de modo que valores $|Z| > 2.5$ o $|Z| > 3$ pueden considerarse potencialmente anómalos. Sin embargo, cuando los datos presentan distribuciones sesgadas o multimodales, esta suposición puede generar interpretaciones erróneas, enmascarar verdaderos valores atípicos o producir falsas detecciones (*masking effect*).

En disciplinas como la hidrología, la ecología o las geociencias, estas transformaciones son esenciales para garantizar la coherencia de las variables y evitar que diferencias de escala influyan en la calibración de los modelos o en las métricas de desempeño (Thirel et al., 2024). Adicionalmente, cuando los datos presentan una alta variabilidad o distribución sesgada, se recurre con frecuencia a la transformación logarítmica natural, la cual contribuye a estabilizar la varianza, mejorar la normalidad y reducir la influencia de valores extremos. Esta transformación se expresa como:

$$X_{t,log} = \ln(X_t + \varepsilon) \quad (16)$$

Donde:

- X_t Representa el valor del parámetro (x_1 y x_2) en el tiempo
- $\ln(\cdot)$ Es el logaritmo natural
- ε Es un pequeño valor positivo (por ejemplo, 10^{-6}) que se suma para evitar problemas cuando $X_t = 0$.

Además, esta transformación tiene una implicación metodológica clave: al trabajar en el espacio logarítmico, es posible convertir un problema de ajuste de tipo no lineal en uno lineal respecto a los parámetros. Por ejemplo, una relación potencialmente no lineal entre un parámetro y una variable predictora del tipo $Y = a \cdot X^b$ puede expresarse de forma lineal como:

$$\ln(Y) = \ln(a) + b \cdot \ln(X) \quad (17)$$

De esta manera, el uso de transformaciones logarítmicas y normalizaciones no solo tiene un papel estadístico, sino también metodológico, pues facilita la aplicación de algoritmos lineales y mejora la robustez del ajuste (Pushpalatha, 2012).

En conjunto, las técnicas de normalización, estandarización y transformación logarítmica son pasos fundamentales en el procesamiento de datos científicos, al garantizar la estabilidad numérica, la comparabilidad entre variables y la interpretabilidad física de los resultados.

J. Análisis de Correlación

El análisis de correlación constituye una técnica estadística fundamental para evaluar el grado de asociación entre dos variables cuantitativas. Su objetivo principal es medir la fuerza y dirección de la relación lineal, proporcionando una primera aproximación sobre cómo varía una variable respecto a otra (Gogtay, 2017).

En el ámbito hidrológico, este análisis se utiliza ampliamente para explorar relaciones entre variables hidroclimáticas, geomorfológicas y parámetros de modelos hidrológicos, permitiendo detectar patrones que pueden orientar la selección de predictores en modelos empíricos o de aprendizaje automático (Benesty et al., 2009).

El coeficiente de correlación de Pearson (r) es una de las métricas más utilizadas para este propósito. Este coeficiente mide la relación lineal entre dos variables continuas, tomando valores en el rango $[-1, 1]$:

- $r \approx 1$: Correlación positiva fuerte.
- $r \approx -1$: Correlación negativa fuerte
- $r \approx 0$: Ausencia de correlación lineal.

Matemáticamente, se expresa como:

$$r = \frac{\sum_{i=1}^n (x_i - \underline{x})(y_i - \underline{y})}{\sqrt{\sum_{i=1}^n (x_i - \underline{x})^2} \sqrt{\sum_{i=1}^n (y_i - \underline{y})^2}} \quad (18)$$

Donde x_i y y_i representan los valores observados de las dos variables, y $\underline{x}$ y $\underline{y}$ sus medias respectivas.

Este coeficiente es adimensional y no depende de la escala de variables, por lo que no requiere normalización previa.

K. Aprendizaje de Máquina (*Machine Learning*)

El aprendizaje de máquina es una subdisciplina de la Inteligencia Artificial que comprende un conjunto de técnicas computacionales las cuales, a través de algoritmos, permiten que las máquinas aprendan patrones a partir de datos y realicen predicciones o decisiones sin estar explícitamente programadas para cada tarea (Shinde et al., 2018).

Se clasifica principalmente en tres tipos:

1. Aprendizaje supervisado: Se dispone de un conjunto de datos con entradas (variables predictoras) y salidas conocidas (variables dependientes). El objetivo es aprender una función que prediga la salida a partir de las entradas. Ejemplos: RLM, SVR, XGBoost.
2. Aprendizaje no supervisado: No se cuenta con una variable dependiente. El objetivo es identificar estructuras, patrones o agrupaciones en los datos. Ejemplos: clusterización con KMeans o MeanShift.
3. Otras modalidades: Combinan aspectos de los dos anteriores o utilizan retroalimentación del entorno para aprender políticas de acción.

1. Sobreajuste (*overfitting*)

El sobreajuste (*overfitting*) constituye uno de los desafíos más críticos en el desarrollo de modelos de machine learning, particularmente en aplicaciones hidrológicas donde los datos suelen ser limitados y presentar ruido (Dietterich, 1995; Ying, 2019). Este fenómeno ocurre cuando un modelo complejo se ajusta excesivamente a los datos de entrenamiento, capturando no solo los patrones subyacentes sino también el ruido y las variaciones aleatorias, lo que perjudica su capacidad de generalización.

La detección del sobreajuste se basa típicamente en la discrepancia entre un alto rendimiento en el conjunto de entrenamiento y un bajo rendimiento en un conjunto de prueba independiente. Entre las estrategias más comunes para mitigarlo se encuentran la validación cruzada, la técnica de *early stopping* y los métodos de regularización, los cuales están incorporados en algoritmos como SVR y XGBoost a través de sus hiperparámetros.

2. Algoritmos de Regresión en el Aprendizaje Supervisado

a) Regresión Lineal Múltiple

La regresión lineal múltiple (RLM) es un modelo estadístico fundamental en el análisis predictivo, cuyo propósito es describir la relación entre una variable dependiente continua y un conjunto de variables independientes o predictoras (James et al., 2023). Su objetivo principal es entender cómo cambia la variable de interés a medida que cambian los valores de los predictores, y construir un modelo que permita realizar estimaciones o predicciones sobre nuevos datos.

La forma general del modelo lineal múltiple es:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_p X_p + \epsilon \quad (19)$$

Donde:

- Y : Variable dependiente
- $X_1, \dots, X_p$: Variables independientes o predictoras
- β_0 : Intercepto o término constante.
- β_i : Coeficientes de regresión que cuantifican el efecto de cada predictor sobre la variable dependiente.
- ϵ : Término de error aleatorio, que representa la variabilidad no explicada por el modelo.

El modelo asume una relación lineal entre la variable objetivo y los predictores. En este contexto, el término lineal no implica necesariamente que las variables predictoras sean lineales entre sí, sino que los parámetros (β_i) lo son dentro del modelo.

Uno de los métodos más comunes para estimar los coeficientes de la regresión es el método de Mínimos Cuadrados Ordinarios (*Ordinary Least Squares*, OLS), el cual busca minimizar la suma de los errores al cuadrado entre los valores observados y los predichos por el modelo:

$$\min \sum_{i=1}^n (y_i - \hat{y}_i)^2 \quad (20)$$

Donde y_i son los valores observados y $\hat{y}_i$ los estimados.

El método OLS es ampliamente utilizado debido a su simplicidad, interpretabilidad y eficiencia computacional, y bajo ciertos supuestos (linealidad, independencia, homocedasticidad, normalidad y ausencia de multicolinealidad), se considera un estimador óptimo en términos de varianza mínima y ausencia de sesgo.

Es importante distinguir que la regresión lineal múltiple constituye el modelo estadístico, mientras que OLS es uno de los algoritmos de estimación utilizados para calcular sus parámetros (Gareth et al., 2021).

En contextos hidrológicos, la RLM permite identificar relaciones entre los parámetros de modelo conceptuales (como el modelo GR2M) y las variables físico-climáticas de la cuenca. No obstante, su efectividad depende de que las relaciones entre predictores y parámetros sean aproximadamente lineales y que no exista multicolinealidad significativa entre las variables predictoras; además, se asume cierto comportamiento adecuado de los residuos (independencia y homocedasticidad), condiciones necesarias para una estimación estable e interpretación coherente de los coeficientes.

b) Máquinas de Soporte Vectorial (SVR)

Las Máquinas de Soporte Vectorial para Regresión pertenecen al aprendizaje supervisado y son una extensión de las Máquinas de Soporte Vectorial (SVM) utilizadas en problemas de clasificación (Awad & Khanna, 2015).

El principio teórico de SVR se centra en encontrar una función que se desvíe de los valores observados en una cantidad no mayor que ϵ para cada punto de datos, manteniéndose al mismo tiempo lo más plana posible. Esto se logra mediante la minimización de una función de pérdida que penaliza las predicciones que se alejan más de ϵ de los valores reales.

La función objetivo de SVR se expresa como:

$$\min \frac{1}{2} ||w||^2 + C \sum_{i=1}^n (\xi_i + \xi_i^*) \quad (21)$$

Sujeta a:

$$\begin{cases} y_i - (w \cdot x_i + b) \leq \epsilon + \xi_i \\ (w \cdot x_i + b) - y_i \leq \epsilon + \xi_i^* \\ \xi_i \xi_i^* \geq 0 \end{cases} \quad (22)$$

Donde:

- w : Vector de pesos que define la orientación del hiperplano.
- b : Término de sesgo.
- C : Parámetro de regularización que establece un balance entre la complejidad del modelo (representado por el término $\frac{1}{2} ||w||^2$) y la tolerancia a los errores de predicción (capturados por el término de penalización $\xi_i + \xi_i^*$).

- ξ_i y ξ_i^* : Variables de holgura no negativas que permiten manejar errores fuera del margen ϵ .

Estas restricciones definen explícitamente el papel de ϵ , el cual determina la “banda” alrededor del hiperplano dentro de la cual las predicciones se consideran aceptables. Los puntos de datos que quedan fuera de esta banda contribuyen al término de penalización a través de las variables $\xi_i + \xi_i^*$.

Funciones Kernel

Para manejar relaciones no lineales, SVR utiliza funciones *kernel* que transforman los datos a un espacio de mayor dimensionalidad donde pueden ser separados linealmente. Un *kernel* comúnmente utilizado es el *kernel* radial o de base gaussiana (RBF):

$$K(x_i x_j) = \exp(-\gamma \|x_i - x_j\|^2) \quad (23)$$

Donde γ controla la influencia de cada punto de datos: valores pequeños de γ generan una superficie de decisión más suave (influencia lejana), mientras que valores grandes de γ pueden llevar al sobreajuste.

Por otro lado, el *kernel* lineal opera en el espacio original de características, asumiendo una relación directa entre variables, lo que lo hace más adecuado para problemas con patrones simples o cuando se prioriza la interpretabilidad del modelo:

$$K(x_i x_j) = x_i^\top x_j \quad (24)$$

Donde: $x_i^\top x_j$ representa el producto punto entre los vectores x_i y x_j , lo que mide su similitud lineal en el espacio de características.

Hiperparámetros en SVR

Los hiperparámetros son parámetros cuyo valor se establece antes del proceso de aprendizaje y que no pueden ser aprendidos directamente de los datos. Su selección es crucial para el rendimiento del modelo.

Hiperparámetros Importantes en SVR

1. C (Regularización): Controla el equilibrio entre lograr un modelo simple (valores bajos de C) y el ajuste a los datos de entrenamiento (valores altos de C).
2. ϵ (Margen de tolerancia): Define el margen dentro del cual no se penalizan los errores. Un ϵ mayor permite un modelo más general y menos sensible al ruido.

3. γ (Alcance del *kernel* RBF): Define qué tan lejana es la influencia de un solo ejemplo de entrenamiento. Valores bajos implican una influencia más amplia, mientras que valores altos restringen la influencia a una región pequeña, lo que puede provocar sobreajuste.

c) eXtreme Gradient Boosting

XGBoost es un algoritmo de aprendizaje supervisado basado en el principio de *gradient boosting*, que combina secuencialmente múltiples modelos débiles (típicamente árboles de decisión) para formar un modelo robusto. Pertenece a la familia de métodos de ensamble por *boosting* y se distingue por su eficiencia computacional y su capacidad para manejar grandes conjuntos de datos (Chen & Guestrin, 2016; Arif et al., 2023).

Fundamento Teórico

La base teórica de XGBoost se centra en la optimización de una función objetivo que combina una función de pérdida (que mide el error de predicción) con un término de regularización (que controla la complejidad del modelo). La función objetivo en el paso t se define como:

$$L^T = \sum_{i=1}^n l(y_i, \hat{y}_i^{(t-1)} + f_t(x_i)) + \sum_{k=1}^t \Omega(f_k) \quad (25)$$

Donde:

- $l(y_i, \hat{y}_i)$ Es la función de pérdida que mide la diferencia entre los valores observados y_i y los predichos $\hat{y}_i$.
- $y_i, \hat{y}_i^{(t-1)} + f_t(x_i)$ Representa la predicción actual, que es la predicción anterior más la corrección del nuevo árbol.
- $\Omega(f_k) = \gamma T + \frac{1}{2} \lambda ||\omega||^2$ Es el término de regularización que penaliza la complejidad del árbol K -ésimo.
 - T : número de hojas en el árbol
 - ω : scores de las hojas
 - γ : complejidad por hoja
 - λ : parámetro de regularización L2

Además, cada nueva predicción se actualiza de manera incremental según:

$$\hat{y}_i^{(t)} = \hat{y}_i^{(t-1)} + \eta f_t(x_i) \quad (26)$$

Donde η (tasa de aprendizaje) es un hiperparámetro que controla la magnitud con la que cada árbol contribuye a la predicción final. Valores pequeños de η hacen que el

modelo aprenda de forma más gradual, reduciendo el riesgo de sobreajuste, pero requiriendo un mayor número de árboles (Chen & Guestrin, 2016).

Hiperparámetros importantes en XGBoost

- Número de árboles (*n_estimators*): Cantidad de árboles en el ensamble.
- Profundidad máxima (*max_depth*): Controla la complejidad individual de cada árbol.
- Tasa de aprendizaje (*learning_rate*): Factor de escalado para las contribuciones de cada árbol.
- Submuestreo de características (*colsample_bytree*): Proporción de características consideradas para cada árbol.
- Regularización L2 (*reg_lambda*): Controla la penalización sobre la magnitud de los scores de las hojas.

Ventajas

XGBoost es computacionalmente eficiente, robusto frente a datos ruidosos y valores faltantes, capaz de capturar interacciones complejas entre variables, e incluye regularización nativa que reduce el sobreajuste en comparación con otros algoritmos de boosting tradicionales.

d) Métricas de Evaluación de Modelos

El coeficiente de determinación, denotado como R^2 , es una métrica fundamental para evaluar el rendimiento de modelos de regresión en el aprendizaje supervisado. Cuantifica la proporción de la varianza en la variable dependiente que es explicada por las variables independientes (Di Bucchianico, 2008).

Desde una perspectiva, el R^2 mide que tan bien las predicciones del modelo se aproximan a los datos reales, en comparación con un modelo básico que siempre predice la media de los valores observados.

La formulación matemática del R^2 es la siguiente:

$$R^2 = 1 - \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{\sum_{i=1}^n (y_i - \underline{y})^2} \quad (27)$$

Donde:

- y_i Representa los valores observados reales.
- $\hat{y}_i$ Representa los valores predichos por el modelo.
- $\underline{y}$ Es la media aritmética de los valores observados.
- El numerador $\sum_{i=1}^n (y_i - \hat{y}_i)^2$ corresponde a la suma de cuadrados de los residuos (SCR), que mide la discrepancia entre los datos reales y las predicciones.

- El denominador $\sum_{i=1}^n (y_i - \bar{y})^2$ es la suma total de cuadrados (STC), que representa la variabilidad total de los datos.

Interpretación:

- Un valor de $R^2 = 1$ indica que el modelo explica el 100% de la variabilidad de los datos, lo que representa un ajuste perfecto.
- Un valor de $R^2 = 0$ significa que el modelo no explica ninguna de las variaciones y no es mejor que simplemente predecir la media.
- Si bien un R^2 más alto generalmente indica un mejor ajuste, es crucial considerar que un valor cercano a 1 en los datos de entrenamiento puede ser indicio de sobreajuste, especialmente si no se replica en un conjunto de prueba independiente.

L. Validación cruzada

La validación cruzada es una técnica utilizada en estadística y *machine learning* para evaluar modelos predictivos, particularmente cuando se trabaja con conjuntos de datos reducidos. Su objetivo principal es estimar la capacidad de generalización de un modelo mientras se minimiza el riesgo de sobreajuste, garantizando que las métricas de evaluación no dependan de una partición específica de los datos (Hastie et al., 2009; James et al., 2013).

Entre los diversos métodos de validación cruzada el enfoque de *Leave-One-Out (LOO)* destaca por su especial utilidad en estudios con muestras limitadas. Este método representa una variante extrema de la validación cruzada *k-fold*, en la cual $k = n$, siendo n el número total de observaciones (Stone, 1974). En este esquema, el modelo se entrena repetidamente usando todas las muestras excepto una, que se reserva para validación, repitiendo este proceso hasta que cada observación haya servido exactamente una vez como conjunto de prueba.

Este procedimiento ofrece una evaluación exhaustiva e independiente de divisiones aleatorias, lo que reduce sesgos en la estimación de error (Kohavi, 1995). Además, resulta particularmente útil cuando el tamaño muestral es limitado, pues permite aprovechar al máximo la información disponible, siendo recomendada en contextos con escasez de datos (Hastie et al., 2009; James et al., 2003).

No obstante, el método LOO presenta también ciertas limitaciones: (1) alto costo computacional, ya que requiere ajustar el modelo tantas veces como observaciones existan (Stone, 1974); y (2) sensibilidad a valores atípicos, pues cada iteración de validación depende de una sola observación (James et al., 2003). A pesar de estos inconvenientes, sus ventajas lo convierten en una opción robusta para la evaluación de modelos bajo condiciones de escasez de datos.

IV. Descripción de la zona de estudio

A. Ubicación

El Sistema Cutzamala se localiza dentro de la región del Eje Neovolcánico Transversal, zona caracterizada por su topografía montañosa, con una altitud promedio que oscila entre los 1700 y 2700 msnm, y una compleja estructura geológica. Las coordenadas geográficas aproximadas del sistema se encuentran entre los 19.0° y 19.5° de latitud norte, y entre los 99.5° y 100.5° de longitud oeste. Esta área incluye varios municipios clave, como Zitácuaro en Michoacán; Donato Guerra, Ixtapan del Oro, Valle de Bravo, Villa Victoria y Amanalco en el Estado de México; así como a las 13 de las 16 alcaldías de la Ciudad de México que reciben agua a partir de este sistema.

La zona de estudio corresponde a las cuencas de aportación de caudal a las principales presas del Sistema Cutzamala (polígonos de color verde en la Figura 2): El Bosque, Villa Victoria y Valle de Bravo. Adicionalmente, se analizan cuencas aledañas que contienen información hidrométrica (polígonos en color azul en la Figura 2). Estas cuencas sirvieron para entrenar el modelo hidrológico para la predicción en cuencas sin información hidrométrica.

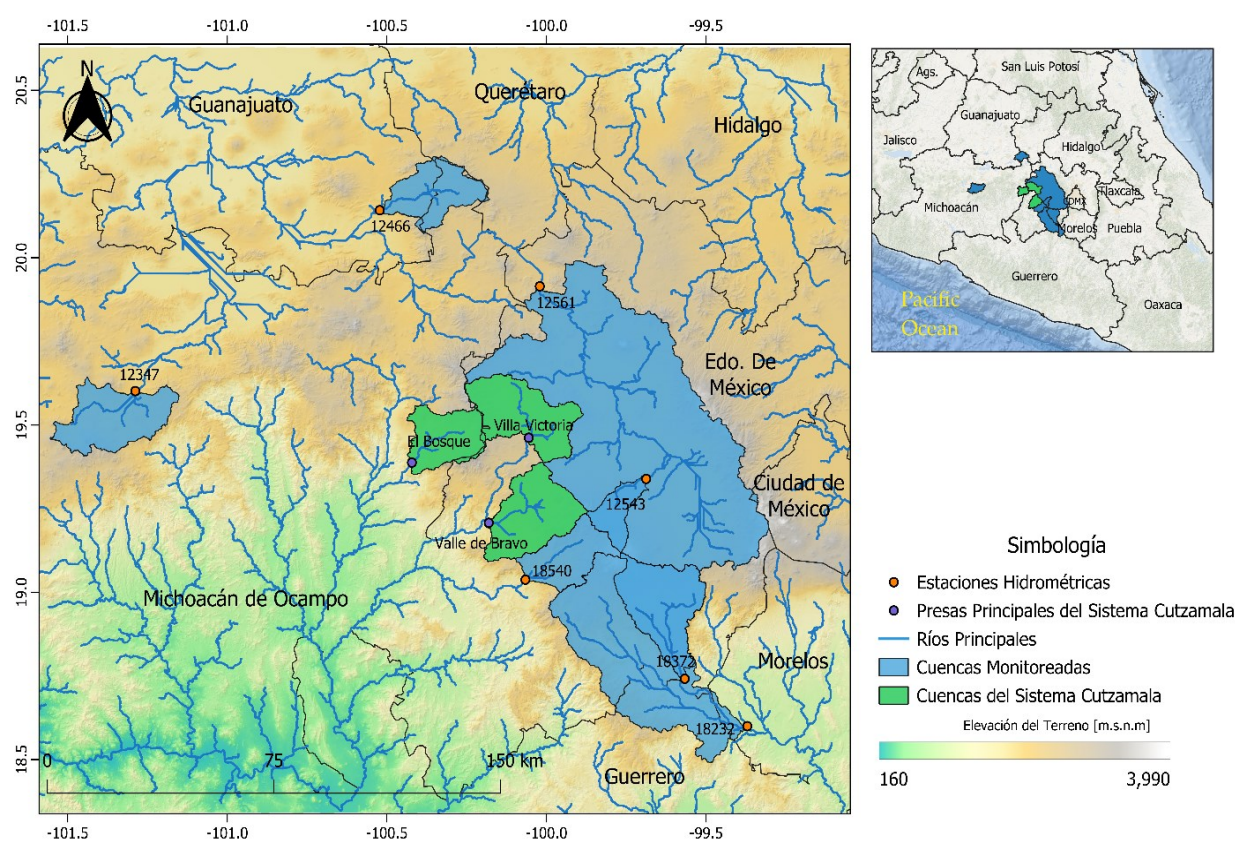


Figura 2. Ubicación de las cuencas de análisis.

La zona de estudio y el Sistema Cutzamala se encuentran principalmente dentro de dos Regiones Hidrológicas: la Región Hidrológica Lerma - Santiago y la Región Hidrológica del Río Balsas. La primera abarca la porción sureste, donde se localizan las cuencas correspondientes a las estaciones hidrométricas: 18232, 18372, 18540, 12543, 12561 y 12466. La segunda se ubica al noroeste e incluye las presas del Sistema Cutzamala, la estación hidrométrica 12347, presas importantes como Tilostoc y Colorines, así como Laguna Seca y Laguna Verde.

B. Geología

El Sistema Cutzamala se encuentra en una región geológicamente compleja que abarca partes de los estados de Michoacán y Estado de México. Esta área forma parte del Eje Neovolcánico Transversal, cuya antigüedad se remonta a 11-17 millones de años y presenta un basamento de la era Mesozoica relacionado con magmatismo marino y al terreno Guerrero, de la Sierra Madre del Sur, la cuál es un terreno originalmente alóctono a la masa continental que conformaba México, compuesto por bloques corticales delimitados por fallas y unidades litológicas distintas, los cuales llegan a formar terrenos tectono-estratigráficos con características específicas. El límite entre la Sierra Madre del Sur y el Eje Neovolcánico lo conforma la subprovincia “Depresión del Tepalcatepec”, un área formada por una secuencia vulcano-sedimentaria del Pleistoceno - Holoceno y por depósitos fluviales que proceden de la Meseta Purépecha. Su relieve más abundante son las planicies con algunas rampas de piedemonte erosivo-denudativas que se encuentran en el oeste y el este de la subprovincia, así como una pequeña cantidad de montañas volcánicas de Acahuato y conos de ceniza con derrames de lava del este.

Una unidad geológica significativa en esta región es la Formación Cutzamala, la cual consiste en una gruesa sucesión de conglomerados, areniscas y limolitas. En la base, predominan los conglomerados con guijarros de rocas volcánicas, areniscas y calizas en una matriz arcillosa de color roja. Las areniscas y las limolitas, de estratificación cruzada, dominan en la cima (Keeman y Estrada-Rodarte, 1999), indicando depósitos de ambientes fluviales y de llanuras de inundación (ver Figura 3).

Esta zona se encuentra ubicada en la parte occidental del Estado de México, particularmente en la cercana a Toluca, producto del vulcanismo asociado al Eje Neovolcánico Transversal. El Nevado de Toluca, uno de los principales volcanes de la zona, ha generado flujos de lava dacítica, depósitos piroclásticos y lahares que forman parte del relleno volcánico que cubre amplias extensiones del Valle de Toluca. Las secuencias volcánicas se encuentran sobre unidades más antiguas del Mioceno y Plioceno, constituidas por tobas soldadas, brechas volcánicas y sedimentos fluviales. Esta zona presenta una sismicidad moderada relacionada con estructuras de fallamiento normal como la falla Tepetates y otras fallas menores orientadas en dirección NO-SE, responsables de la formación de grabens y horsts que

condicionan el escurrimiento superficial y el flujo subterráneo. La combinación de depósitos piroclásticos, paleosuelos y cenizas interestratificadas influye directamente en la capacidad de recarga de los acuíferos locales, siendo la relevancia para el abastecimiento de agua en la región.

En la zona norte del estado de Michoacán, que abarca zonas cercanas a Morelia, Zirahuén y Taretan, predomina un paisaje estructuralmente complejo controlado por el volcanismo monogenético del Eje Neovolcánico Transversal. Esta área forma parte del Campo Volcánico Michoacano-Guanajuatense, caracterizado por la presencia de más de 1200 centros volcánicos monogenéticos, como conos de ceniza, domos de lava y maares. Ejemplos representativos incluyen el Parícutín, que emergió en 1943, y los domos de Zirahuén y Tancítaro. Litológicamente, se observan basaltos, andesitas y tobas, muchas veces intercaladas con paleosuelos ricos en materia orgánica, que han sido datados mediante técnicas de radiocarbono y estudios estratigráficos. Además, esta región está atravesada por estructuras de fallamiento normal como la falla de Taretan, asociada a la formación de valles alargados y lagos tectónicos como el Zirahuén y el Pátzcuaro.

El entendimiento de la configuración geológica es clave para analizar el comportamiento del agua superficial y subterránea en la región. Los materiales volcánicos fracturados, los paleosuelos interestratificados y las estructuras tectónicas han condicionado la formación de acuíferos libres y zonas de escurrimiento superficial, esenciales para el funcionamiento del Sistema Cutzamala.

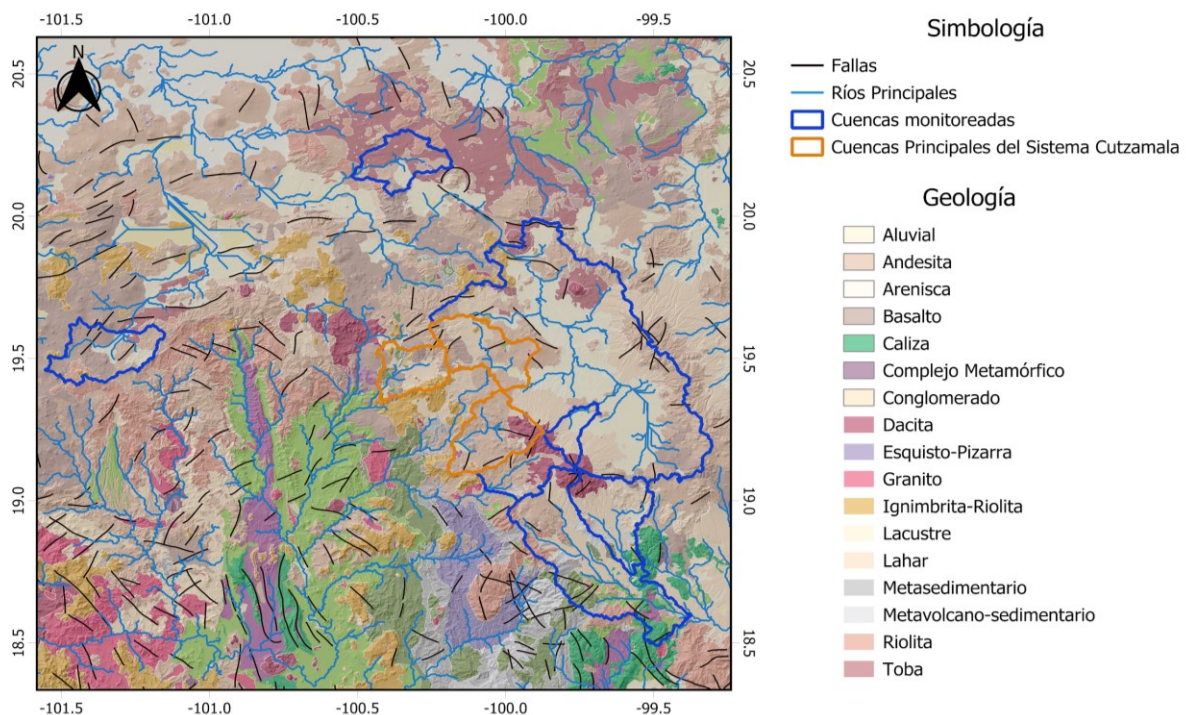


Figura 3. Mapa Geológico de la Región del Sistema Cutzamala.

C. Clima

El clima en la zona de estudio es diverso, presentando variaciones estacionales y microclimáticas asociadas a la altitud y a la configuración topográfica. Sin embargo, el área de influencia del Sistema Cutzamala se caracteriza principalmente por un clima templado subhúmedo con lluvias en verano, de acuerdo con la clasificación de Köppen modificada por Enriqueta García (García, 1964) e INEGI (1980). Esta condición climática, junto con la distribución espacial de la precipitación y los patrones de evapotranspiración, tiene una influencia directa en la dinámica de escurrimiento superficial y recarga de los acuíferos en la región.

La precipitación media anual varía dependiendo de la subregión. En la porción correspondiente a la Región Hidrológica Lerma-Santiago, los valores oscilan entre 800 y 1000 mm anuales, mientras que, en la Región Hidrológica del Balsas, que presenta un relieve más accidentado y mayor cobertura boscosa, las precipitaciones alcanzan valores entre 1150 y 1250 mm anuales, lo cual representa hasta un 50% más que el promedio nacional (Secretaría del Agua del Estado de México, 2023).

En términos de temperatura, la región muestra un gradiente térmico asociado a la altitud: las zonas montañosas presentan temperaturas medias anuales entre 12 y 16 °C, mientras que, en las zonas más bajas o protegidas, como los valles intermontanos, las temperaturas pueden oscilar entre 16 y 20 °C. Esta variabilidad influye en la evapotranspiración potencial, un parámetro clave en la modelación hidrológica y en la estimación de balances hídricos (Secretaría del Agua del Estado de México, 2023).

Asimismo, las lluvias se concentran en la temporada de verano, principalmente entre junio y septiembre, lo cual genera una marcada estacionalidad en la disponibilidad de agua. Esta estacionalidad es crítica para el manejo del Sistema Cutzamala, ya que implica una fuerte dependencia del almacenamiento en presas durante los meses de lluvia para garantizar el abasto en la temporada seca.

D. Tipo de Suelo

En la Región Hidrológica Lerma - Santiago predomina un paisaje de montaña-valle. Las sierras están compuestas por rocas volcánicas: andesitas y basalto, mientras que las planicies son de origen lacustre, compuestas por depósitos de aluvión, materiales piroclásticos y suelos con alto contenido de materia orgánica (Figura 3). Debido a la presencia histórica de cuerpos lacustres, en esta región se encuentran suelos como el Phaeozem, Luvisol, Vertisol, y en particular, el Andosol, (Secretaría del Agua del Estado de México, 2023), este último presente en la zona del Sistema Cutzamala, que constituye el área principal de estudio. Estos suelos, además de su fertilidad, presentan una elevada capacidad de retención de humedad, lo cual influye de manera directa en la dinámica de recarga y escurrimiento (González-Ortigoza et al., 2024).

Por otro lado, en la Región Hidrológica del Balsas coexisten formaciones sedimentarias de origen marino con sierras formadas por materiales volcánicos. En esta región predominan los Andosoles, de origen volcánico, ricos en materia orgánica, con buena capacidad de retención de agua. También se encuentran Regosoles, aunque en menor proporción, suelos jóvenes, claros y pobres en materia orgánica, con baja retención de humedad (Secretaría del Agua del Estado de México, 2023).

Respecto al uso de suelo, dominan los bosques, aunque se enfrentan a una intensa presión por el cambio de uso, principalmente hacia la agricultura de temporal (20%), y en menor medida, hacia pastizales inducidos (15%) para actividades de pastoreo (Secretaría del Agua del Estado de México, 2023).

E. Hidrogeología e Hidrología

Esta región cuenta con varios acuíferos de importancia hidrogeológica e hidrológica. Entre los más relevantes dentro del Estado de México destacan Ixtlahuaca - Atlacomulco, Valle de Toluca, Tenancingo y Villa Victoria-Valle de Bravo. Estos acuíferos contribuyen al balance hídrico regional, aunque no todos están directamente vinculados con las presas principales del sistema.

Las presas más importantes del Sistema Cutzamala —El Bosque, Villa Victoria y Valle de Bravo— se localizan principalmente en los acuíferos Villa Victoria - Valle de Bravo (Estado de México) y Huetamo (Michoacán) (CONAGUA, 2024).

Asimismo, en esta región se encuentran cuerpos de agua relevantes como la presa Temascaltepec, que forma parte del sistema de almacenamiento complementario.

Por otra parte, aunque los acuíferos Ixtlahuaca-Atlacomulco y Valle de Toluca poseen importancia regional y están parcialmente dentro del área de estudio, no contienen las presas principales del sistema. Sin embargo, el Sistema Cutzamala atraviesa otros acuíferos a lo largo de su trazado, aun cuando no se realicen extracciones directas de ellos. Un ejemplo relevante es la presa Tuxpan, donde inicia el sistema, que se ubica en el acuífero Hidalgo-Tuxpan (CONAGUA, 2024).

Aunque estos acuíferos no están directamente asociados a la extracción de agua para el Sistema Cutzamala, su caracterización hidrogeológica e hidrológica es relevante en el proceso de simulación y entrenamiento del modelo hidrológico GR2M, ya que influyen en la disponibilidad, recarga y almacenamiento de agua en la región (Secretaría del Agua del Estado de México, 2023).

V. Bases de datos y metodología

A. Bases de datos

Para la caracterización física e hidrológica de las cuencas de estudio se utilizaron diversas fuentes de información espacial e hidroclimática, lo cual permite alimentar al modelo hidrológico y realizar análisis comparativos entre las condiciones naturales y la infraestructura existente.

La información se organizó en cuatro tablas principales:

1. Estaciones hidrométricas y su localización (Tabla 1)
2. Características geomorfológicas de las cuencas (Tabla 2)
3. Infraestructura hidráulica del sistema (Tabla 3)
4. Variables hidroclimáticas obtenidas de productos satelitales y modelaciones globales (Tabla 4).

A continuación, se describe el contenido de cada tabla y su relevancia dentro del análisis desarrollado en esta investigación.

Para el análisis de los caudales registrados en las cuencas cercanas al Sistema Cutzamala, se cuenta con información proveniente de las estaciones hidrométricas, como se presenta en la Tabla 1. Estos datos fueron obtenidos a partir del Banco Nacional de Datos de Agua Superficiales (BANDAS) de la CONAGUA. Se seleccionaron estaciones hidrométricas con un registro mayor a 5 años dentro del periodo de 1981 al 2024 (periodo común de los datos satelitales), y se removieron las estaciones con registros atípicos o con caudales fuertemente regulados por infraestructura.

Tabla 1. Características de las estaciones hidrométricas utilizadas para la simulación.

ID Estación	Nombre de Estación	Coordenadas		Periodo de Registro
18540	Temascaltepec	-99.933	19.107	01/03/1973 - 07/12/2024
12347	Santiago Undameo	-101.355	19.520	01/08/1939 - 30/09/2002
12466	Jerécuaro	-100.505	20.221	01/02/1972 - 31/12/1986
12561	Temascalcingo	-99.725	19.486	01/08/1962 - 31/12/2024
12543	Calixtlahuaca	-99.778	19.240	01/08/1961 - 31/12/2024
18232	Amacuzac	-99.428	18.674	01/02/1955 - 31/12/2024
18372	Coatepequito	-99.635	18.960	01/12/1964 - 07/12/2024

A partir de la ubicación de las estaciones hidrométricas se delimitaron las cuencas hidrográficas y se estimaron sus parámetros geomorfológicos, como el área, perímetro, elevación media, pendiente media y centroide, los cuales fueron calculados utilizando los programas SAGA GIS y QGIS. Las cuencas fueron delimitadas a partir de herramientas de procesamiento del terreno con el programa SAGA GIS utilizando el modelo de elevación digital *Shuttle Radar Topography Mission* (SRTM) (Bamler, 1999). Estos parámetros se presentan en la Tabla 2 que proporciona información detallada sobre las características geomorfológicas de las cuencas estudiadas.

Tabla 2. Propiedades de las cuencas asociadas a las estaciones hidrométricas.

ID Cuenca	Área (km^2)	Perímetro (km)	Elevación (msnm)	Pendiente (km)	Centroide
18540	415.57	143.953	2682.8	21.046	19.107
12347	615.20	182.460	2318.9	12.7436	19.520
12561	4812.36	573.496	2749.6	10.084	19.486
12543	208.77	102.575	3068.9	11.0828	19.240
12466	460.62	87.753	2188.1	9.8266	20.194
18232	2242.57	59.633	1110	8.054	18.840
18372	673.79	182.326	2355.1	201.017	18.960

En cuanto a las presas del Sistema Cutzamala, se recopiló información fundamental para la simulación hidrológica y para la estimación del balance hídrico y almacenamiento. Entre los datos considerados se encuentran las coordenadas geográficas, el año de construcción, la capacidad de almacenamiento y la elevación de las cortinas de las presas, los cuales se presentan en la Tabla 3. Esta información fue obtenida del Sistema de Información Hidrológica (SIH) de CONAGUA.

Cabe señalar que, para la implementación del modelo hidrológico, únicamente se emplearon las coordenadas de las presas y el área de las presas, mientras que los datos de capacidad de almacenamiento fueron empleados en el cálculo del balance hídrico.

Tabla 3. Propiedades de las principales presas del Sistema Cutzamala

Nombre	Coordenadas	Año de Construcción	Elevación de Cortina [m]	Capacidad de Almacenamiento [hm ³]	Área [km ²]
Villa Victoria	19.4686, -100.0033	1944	19	79.639	617.915
Valle de Bravo	19.1970, -100.1612	1947	56	261.323	619.685
El Bosque	19.3903, -100.4104	1954	70	113.282	419.322

Finalmente, la Tabla 4 presenta las variables hidroclimáticas empleadas en esta investigación obtenidas a partir de productos satelitales y modelaciones globales. Estas variables complementan los datos observacionales y permiten una caracterización más completa de las condiciones hidrológicas e hidrogeológicas de la zona de estudio.

Se utilizaron datos del *Famine Early Warning Systems Network Land Data Assimilation System* (FLDAS), desarrollado por la NASA y la Universidad de Princeton, el cual combina observaciones satelitales y modelos terrestres para determinar variables como precipitación, temperatura, evapotranspiración, escurrimientos y humedad del suelo a escala regional. De estas variables, la precipitación y la temperatura son esenciales para alimentar el modelo hidrológico y evaluar el balance hídrico en las cuencas del Sistema Cutzamala.

Como variables adicionales para la estimación de los parámetros del modelo hidrológico en cuencas sin información, se implementaron datos de vegetación, propiedades hidráulicas del suelo, y propiedades del flujo derivado de modelos globales. Se empleó el producto MODIS13, generado por el sensor *Moderate Resolution Imaging Spectroradiometer* (MODIS) a bordo de los satélites Terra y Aqua de la NASA, que ofrece índices de vegetación como el NDVI (*Normalized Difference Vegetation Index*) a alta resolución espacial y temporal. La caracterización de propiedades del suelo se apoyó en información del proyecto *Covariates-based GeoTransfer Functions* (CoGTF), desarrollado por la ETH Zürich (*Eidgenössische Technische Hochschule Zürich*), el cual predice la conductividad hidráulica saturada (KSat) a partir de variables ambientales mediante algoritmos de aprendizaje supervisado, específicamente *Random Forest*. Finalmente, se emplearon datos del *Global Streamflow Characteristics Dataset* (GSCD) disponible a través del portal GloH2O, que proporciona mapas globales derivados de observaciones de escurrimiento, incluyendo la constante de recesión del caudal base (K), coeficientes de escurrimiento (RC) y caudal medio por unidad de área (QMean).

Tabla 4. Descripción de las bases de datos satelitales y globales.

Dataset	Variable	Periodo	Resolución Temporal	Resolución Espacial [km]	Fuente
FLDAS	Temperatura (Tair)	1982 - Presente	Mensual	Global 11.1	McNally et al. (2017)
FLDAS	Precipitación (Rain)	1982 - Presente	Mensual	Global 11.1	McNally et al. (2017)
FLDAS	Evapotranspiración (Evap)	1982 - Presente	Mensual	Global 11.1	McNally et al. (2017)
FLDAS	Velocidad del Viento (Wind)	1982 - Presente	Mensual	Global 11.1	McNally et al. (2017)
FLDAS	Humedad del Suelo (SM)	1982 - Presente	Mensual	Global 11.1	McNally et al. (2017)
FLDAS	Escorrentía subterránea y subsuperficial (Qt)	1982 - Presente	Mensual	Global 11.1	McNally et al. (2017)
FLDAS	Humedad Específica (Qair)	1982 - Presente	Mensual	Global 11.1	McNally et al. (2017)
MODIS13	Índice de Vegetación de Diferencia Normalizada (NDVI)	2000 - Presente	16 días	Global 1	Didan, K. (2021)
CoGTF	Conductividad Hidráulica Saturada (KSat)	-	-	Global 1	Gupta et al. (2020)
GSCD	Contenido de limo (Silt)	-	-	Global 1	Beck et al. (2015)
GSCD	Contenido de arcilla (Clay)	-	-	Global 1	Beck et al. (2015)
GSCD	Índice de Flujo Base (BFI)	-	-	Global 1	Beck et al. (2015)
GSCD	Contenido de arena (Sand)	-	-	Global 1	Beck et al. (2015)
GSCD	Coefficiente de Escorrentía (RC)	-	-	Global 1	Beck et al. (2015)
GSCD	Caudal Medio (QMean)	-	-	Global 1	Beck et al. (2015)
GSCD	Recesión de caudal base (K)	-	-	Global 1	Beck et al. (2015)

B. Metodología

1. Descripción general de la metodología

La Figura 4 resume la metodología empleada. En primer lugar, se recopilaron los datos de caudal provenientes de las estaciones hidrométricas (Tabla 2), así como las variables climáticas de precipitación y temperatura del producto global FLDAS, y las variables fisiográficas derivadas de los productos MODIS, CoGTF y GSCD.

El modelo GR2M fue calibrado mediante el algoritmo de optimización Nelder-Mead, estimando los parámetros x_1 y x_2 con base en la métrica de desempeño NSE como función objetivo. Una vez calibrado, los parámetros fueron regionalizados a partir del análisis de correlación con variables fisiográficas y climáticas, empleando modelos de utilizando LRM, SVR y XGBoost. Adicionalmente, se aplicó una regresión espacial utilizando las variables fisiográficas en formato ráster para generar mapas continuos de los parámetros del modelo. Posteriormente, se realizaron simulaciones hidrológicas con el modelo GR2M utilizando los distintos enfoques de regionalización y evaluando su desempeño mediante validación cruzada Leave-One-Out (LOO).

Finalmente, los parámetros regionalizados obtenidos se aplicaron a las cuencas de las presas del Sistema Cutzamala, estimando los valores de x_1 y x_2 a partir de los modelos de *machine learning*. En este caso, la regresión espacial sirvió como base al proporcionar mapas de parámetros que permitieron obtener un valor promedio de x_1 y x_2 para cada cuenca. A partir de dichos valores fisiográficos específicos, los algoritmos de LRM, SVR y XGBoost estimaron los parámetros del modelo sin incluir estas cuencas en el proceso de entrenamiento utilizando para ello la información de las cuencas vecinas y las variables fisiográficas correspondientes a las presas principales del Sistema Cutzamala.

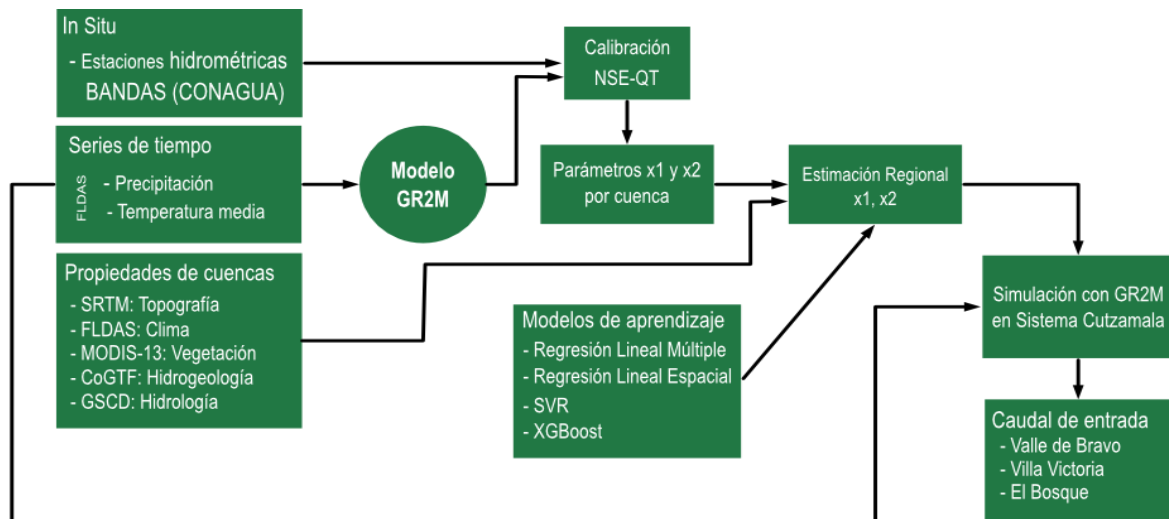


Figura 4. Esquema general de la metodología

2. Calibración del modelo en cuencas con información hidrométrica

La calibración del modelo GR2M consiste en optimizar los parámetros x_1 y x_2 con el fin de reproducir con la mayor fidelidad el comportamiento hidrológico observado en cada cuenca. Este proceso requiere datos de caudal mensual, así como series temporales de precipitación y temperatura media, que actúan como insumos climáticos del modelo. En la implementación utilizada a partir de la librería *LuMod*, la evapotranspiración potencial (PET, ver Figura 2), no se proporciona como entrada, sino que se calcula internamente a partir de la temperatura media mensual mediante el método empírico de Thornthwaite (1948), el cual estima la PET en función de la temperatura y la duración media del día según la latitud del centroide de la cuenca. Adicionalmente, se incorporan atributos morfométricos de la cuenca en cuestión, área de captación (km^2) y la ubicación del centroide (latitud en grados), con fines de caracterización y análisis espacial.

Se definieron rangos iniciales para los parámetros: $x_1 \in [50, \infty)$ y $x_2 \in [0, 1]$, y se procedió a la búsqueda de los valores óptimos mediante el algoritmo de Nelder-Mead, implementado a través de la librería *SciPy* de Python. La función objetivo utilizada para este caso fue el NSE, el cual permitió cuantificar la similitud entre los caudales simulados y los observados. El proceso iterativo permitió identificar el conjunto de parámetros x_1 y x_2 que maximizan el NSE (es decir, minimizan el error cuadrático entre simulaciones y observaciones), garantizando así una representación consistente del comportamiento hidrológico de cada cuenca durante el periodo de calibración.

3. Validación en las cuencas del Sistema Cutzamala

En el caso de las presas del Sistema Cutzamala no se cuenta con registros de caudal observado. No obstante, es posible realizar una estimación del caudal utilizando datos abiertos proporcionados por el Sistema de Información Hidrológica (SIH) de la CONAGUA. Estos datos incluyen series temporales de almacenamiento del embalse de la presa, precipitación, evaporación y volúmenes extraídos por diferentes medios.

A partir de esta información, se aplicó un enfoque basado en el balance hídrico para estimar el caudal neto de entrada de la presa, el cual puede utilizarse como referencia para validar el desempeño del modelo hidrológico y las metodologías propuestas.

El balance hídrico considera que el cambio en el volumen almacenado del embalse refleja el efecto neto entre las entradas (como la precipitación directa) y las salidas (como la evaporación, extracciones y descargas por vertedero).

Para cuantificar el caudal mensual estimado Q_t se utilizó la siguiente ecuación:

$$Q_t [hm^3] = dV - \frac{P \cdot A}{1e5} + \frac{Et \cdot A}{1e5} + EOT * 2.592 + EV * 2.592 \quad (28)$$

Donde:

$Qt [hm^3] \rightarrow$ Es el caudal en hm^3 .

$dV \rightarrow$ Diferencia de volumen $V2 - V1$, representa el cambio en el almacenamiento de la presa.

Se utilizó el valor promedio.

$P \rightarrow$ Precipitación medida directamente en la presa. Se utilizó la suma de los valores registrados.

$A \rightarrow$ Corresponde al área de la presa.

$Et \rightarrow$ Evaporación registrada en la presa. Se utilizó la suma de los valores registrados.

$EOT \rightarrow$ Volumen extraído a través de la toma de obra.

$EV \rightarrow$ Volumen extraído a través del vertedero.

Los factores de conversión (10^{-5} y 2.592) homogeniza las unidades a hm^3 , considerando que:

- $1 mm \cdot m^2 = 10^{-3}m^3 \rightarrow 1mm \cdot km^2 = 10^{-3}hm^3$ (para Precipitación y Evaporación)
- $1m^3/s \cdot día = 86,400m^3 = 0.0864 hm^3$ (para volumen extraído por toma de obra y por vertedero)

Esta formulación permite estimar el volumen neto de entrada del sistema a partir de variables observadas, utilizando unidades compatibles y factores de conversión adecuados.

Limitaciones y posibles fuentes de error

Si bien esta ecuación (14) permite obtener una aproximación razonable del caudal mensual, es importante considerar sus limitaciones:

- Calidad de los datos: La estimación depende directamente de datos de acceso público, los cuales pueden contener errores de medición, vacíos de información o inconsistencias en la escala temporal y espacial.
- Procesos no incluidos: No se consideran entradas subterráneas, escurrimientos superficiales laterales, infiltración, ni pérdidas estructurales, las cuales pueden ser significativas en algunos embalses.
- Supuestos de homogeneidad: Se asume que la precipitación y evaporación son uniformes en toda la superficie del embalse, y que el área de la presa permanece constante, lo cual puede no reflejar condiciones reales.

Por estas razones, los valores de Qt obtenidos deben interpretarse como aproximaciones de primer orden, útiles para evaluar tendencias generales del sistema y validar el comportamiento global del modelo, más que como representaciones exactas del caudal real.

4. Preprocesamiento de datos y Análisis de Correlación

a) Preprocesamiento de datos

Previo al análisis de correlación y a la aplicación de los modelos de predicción, se realizaron distintas transformaciones a las variables involucradas con el objetivo de garantizar la comparabilidad, mejorar la distribución estadística y reducir la influencia de valores atípicos. Estas transformaciones se agrupan en tres etapas: normalización, estandarización y transformación logarítmica.

a) Normalización de variables espaciales

Se aplicó la normalización Min-Max a las variables espaciales derivadas de información de ráster (NDVI, KSat, QMean, Evapotranspiración, entre otras), empleando el programa SAGA-GIS.

Este procedimiento reescaló los valores al intervalo $[0, 1]$ (véase ecuación 14), eliminando las diferencias de magnitud entre las variables y permitiendo su integración en análisis comparativos y posteriores procesos de modelado.

La normalización fue indispensable para poder realizar el análisis de correlación entre las capas ráster y los valores de los parámetros x_1 y x_2 del modelo GR2M, asegurando que las variables mantuvieran proporciones coherentes sin alterar su estructura espacial original.

b) Estandarización y depuración de series temporales

Durante el procesamiento de las series temporales de caudal observado (Qobs), se aplicó un criterio de remoción de valores atípicos utilizando la estandarización *Z-Score* (véase ecuación 15).

Este método se utilizó asumiendo una distribución normal, aunque dicha condición no fue verificada formalmente, ya que el objetivo principal fue eliminar valores anómalos evidentes. La estandarización se implementó en Python, utilizando la librería *SciPy*, y permitió identificar observaciones cuyo valor absoluto del puntaje *Z* excediera el umbral de 2.5, las cuales fueron consideradas *outliers*, lo que permitió mejorar la calidad de los datos utilizados en el proceso de la calibración del modelo y reduciendo la influencia de estas observaciones anómalas en la estimación de los parámetros

c) Transformación logarítmica de parámetros del modelo

Finalmente, se aplicó una transformación logarítmica natural a los parámetros calibrados del modelo hidrológico (x_1 y x_2) (véase ecuación 16).

Esta transformación tuvo como objetivo reducir la asimetría, estabilizar la varianza y disminuir la influencia de valores extremos, mejorando la linealidad de las relaciones entre los parámetros y las variables predictoras.

El uso de la transformación logarítmica no constituye una estandarización, sino una técnica de transformación estadística que convierte relaciones lineales respecto a los parámetros

(véase ecuación 17). Esto facilitó la aplicación de métodos lineales, como la regresión de Mínimos Cuadrados Ordinarios (OLS), y permitió mantener la interpretabilidad física de los parámetros del modelo GR2M tras aplicar la transformación inversa y devolverlos a su escala original.

b) Correlación

Una vez realizados los procesos de preprocesamiento y depuración de datos, se llevó a cabo el análisis de correlación de Pearson para identificar las variables con mayor relación lineal respecto a los parámetros x_1 y x_2 .

El coeficiente de correlación de Pearson se seleccionó por su capacidad de cuantificar la fuerza y dirección de las relaciones lineales entre las variables predictores normalizadas y los parámetros del modelo transformados logarítmicamente.

Se eligió este coeficiente porque:

- Evalúa dependencias lineales directas, coherentes con los métodos de regresión empleados posteriormente.
- Supone una distribución aproximadamente gaussiana de las variables, condición aproximada tras las transformaciones aplicadas.
- Es una métrica adimensional y simétrica, adecuada para comparar variables de distinta naturaleza física.

El análisis se implementó en Python, utilizando las funciones *corr()* de la librería Pandas y *heatmap()* de Seaborn para la representación gráfica de las correlaciones. Los resultados derivados de este análisis se presentan en la sección de Resultados. A partir de ellos, se identificaron y seleccionaron como variables predictoras aquellas que mostraron una correlación lineal significativa con los parámetros del modelo GR2M.

5. Calibración Regional del modelo GR2M

a) Regresión

Para la estimación de los parámetros del modelo GR2M (x_1 y x_2), se implementó una regresión lineal múltiple (RLM) utilizando la biblioteca *Scikit-Learn* en Python. La estimación de los coeficientes se realizó mediante el método de Mínimos Cuadrados Ordinarios (OLS), siguiendo la formulación descrita en la sección correspondiente del marco teórico (véase ecuación 20).

En este modelo, las variables dependientes correspondieron a los parámetros x_1 y x_2 del modelo GR2M, previamente obtenidos mediante el proceso de optimización por el método de Nelder-Mead, mientras que las variables predictoras fueron seleccionadas a partir del análisis de correlación descrito anteriormente.

La implementación del modelo permitió obtener los coeficientes de regresión y evaluar la influencia de cada variable predictora sobre los parámetros x_1 y x_2 , constituyendo la base para la estimación espacial de dichos parámetros en las presas del Sistema Cutzamala.

Selección de Predictores y Contexto Hidrológico

Las variables predictoras utilizadas en este estudio fueron aquellas que mostraron correlaciones estadísticamente significativas con los parámetros x_1 y x_2 del modelo GR2M y que, además, resultaron hidrológicamente coherentes para describir los procesos dominantes del balance hídrico de la cuenca.

Cada variable fue incorporada considerando su relación en los procesos de almacenamiento, escorrentía e infiltración:

- **NDVI (*Normalized Difference Vegetation Index*):** Este índice espectral cuantifica la densidad y vigor de la vegetación. Una mayor cobertura vegetal implica una mayor transpiración y retención de humedad en el suelo, lo que se traduce en una relación positiva con la capacidad de almacenamiento hídrico (parámetro x_1). Además, el NDVI actúa como un indicador indirecto de la evapotranspiración, variable clave en la dinámica del almacenamiento superficial.
- **Evapotranspiración (Evap):** Representa la pérdida total de agua hacia la atmósfera a través de la evaporación del suelo y la transpiración de la vegetación. Esta variable integra los efectos combinados del clima, la cobertura vegetal y la disponibilidad de agua en el suelo, por lo que influye directamente en el balance hídrico superficial. Una mayor evapotranspiración se asocia con una disminución en la capacidad de almacenamiento (x_1), y en ciertos casos, con una modificación en el parámetro (x_2), el cual regula el intercambio entre los almacenes superficial y subterráneo del modelo. Este último está estrechamente vinculado con el coeficiente de recesión (K), que se deriva de x_2 y controla la liberación del flujo base en las simulaciones hidrológicas.
- **KSat (Conductividad Hidráulica Saturada):** Representa la capacidad del suelo para transmitir agua en condiciones saturadas, controlando la infiltración y la recarga del flujo subterráneo. Su asociación con x_1 se explica por su influencia directa sobre la capacidad de almacenamiento y percolación del agua hacia capas más profundas.
- **QMean (Caudal Medio):** Expresa la respuesta integrada del sistema hidrológico a las condiciones climáticas, el uso del suelo y la geología. Actúa como una variable de síntesis del comportamiento hidrológico de la cuenca, relacionada tanto con la capacidad de almacenamiento (x_1) como con el coeficiente de recesión (x_2), al reflejar la variabilidad estacional y la persistencia del flujo.
- **RC (Coeficiente de escorrentía):** Define la fracción de la precipitación que se convierte en escorrentía superficial. Este coeficiente permite captar el grado de retención o

infiltración del suelo, complementando la información del caudal medio y contribuyendo a la caracterización del comportamiento superficial del flujo.

- **K (Constante de Recesión del Flujo Base):** Describe la tasa de disminución del caudal subterráneo en ausencia de recarga, reflejando la respuesta retardada del flujo base y su relación con el parámetro x_2 . Esta variable proporciona una conexión directa entre los procesos de descarga subterránea y la capacidad de almacenamiento del acuífero.

Estas variables se mantuvieron como predictoras por su pertinencia hidrológica y su capacidad para describir de forma adecuada los componentes principales del ciclo hidrológico en la zona de estudio, permitiendo representar las interacciones entre clima, vegetación, suelo y escorrentía de manera consistente con los fundamentos del modelo GR2M. Con el fin de identificar de manera más precisa las variables más relevantes para la estimación de los parámetros del modelo, se realizó un análisis de correlación lineal entre cada variable explicativa y los parámetros calibrados x_1 y x_2 .

Para el caso de $\ln(x_1)$, se seleccionaron tres variables: NDVI, Evapotranspiración (Evap) y Conductividad Hidráulica Saturada (KSat), las cuales presentaron los coeficientes de correlación más altos y consistentes con la interpretación hidrológica del parámetro, reflejando principalmente la capacidad de almacenamiento hídrico superficial y la influencia de la vegetación.

Para el caso $\ln(x_2)$, se seleccionaron tres variables: Caudal Medio Anual (QMean), Coeficiente de escorrentía (RC) y Constante de Recesión del Caudal Base (K), que mostraron las relaciones más significativas con este parámetro, representando la dinámica del flujo base y la recesión del caudal en la cuenca.

Este procedimiento permitió reducir la dimensionalidad del modelo y garantizar que las variables utilizadas tuvieran una relación física y estadística significativa con los parámetros a estimar, asegurando consistencia hidrológica y robustez en las estimaciones.

Evaluación del Modelo

La RLM permitió estimar los coeficientes de cada predictor y analizar su influencia sobre los parámetros x_1 y x_2 . Los resultados de las métricas de desempeño (como el NSE) mostraron que, aunque el modelo capturó las tendencias generales, no se observó una mejora significativa frente a la calibración individual de cada cuenca, por lo que se evaluaron métodos alternativos basados en aprendizaje automático (SVR y XGBoost), capaces de modelar relaciones lineales más complejas entre las variables predictoras y los parámetros hidrológicos.

b) Configuración, Entrenamiento y Selección de Modelos

Para la predicción de los parámetros x_1 y x_2 , se exploraron dos algoritmos de aprendizaje supervisado: SVR y XGBoost. La selección de estos modelos se basó explícitamente en su capacidad reconocida para capturar relaciones no lineales complejas entre las variables predictoras y los parámetros objetivo. Ambos modelos fueron aplicados de forma comparativa a cada parámetro para finalmente escoger el que mejor comportamiento tuviera basándose en el valor más alto del coeficiente de determinación (R^2) en el conjunto de prueba.

El conjunto de variables predictoras utilizado fue idéntico al empleado en el enfoque de regresión lineal, manteniendo las siguientes por su correlación significativa y su fundamento hidrológico para explicar los procesos de retención de agua y regulación de flujos en la cuenca: evapotranspiración, el NDVI, KSat, QMean, RC y K.

c) Proceso de Optimización de Hiperparámetros

- SVR

Para la estimación de los parámetros x_1 y x_2 mediante Máquinas de Soporte Vectorial para Regresión (SVR), se utilizó la librería *Scikit-Learn* en Python. El modelo se configuró con un kernel radial (RBF) debido a su capacidad para representar relaciones no lineales entre las variables predictoras y los parámetros hidrológicos.

La definición de los hiperparámetros del modelo se realizó mediante un proceso de calibración empírica, complementado con una búsqueda en malla (*Grid Search*).

Los hiperparámetros ajustados fueron:

- C : *Parámetro* de regularización que controla la complejidad del modelo.
- γ : *Parámetro* gamma del kernel RBF que define el alcance de la influencia de los puntos de entrenamiento.
- ϵ : *Margen* de tolerancia que regula la sensibilidad del modelo ante los errores de predicción.

El ajuste de los hiperparámetros se realizó considerando un equilibrio entre precisión y capacidad de generalización, evitando configuraciones que pudieran inducir sobreajuste, además de que busca acercar el R^2 a 1, reflejando un modelo capaz de explicar gran parte de la varianza en los datos.

Finalmente, se seleccionaron los valores $C = 10$, $\gamma = scale$, $\epsilon = 0.1$. Es importante destacar que el uso de $\gamma = scale$ permite que este parámetro se ajuste automáticamente a la escala de los datos, lo que hace al modelo más robusto frente a variables con diferentes rangos. Esta configuración ofreció un equilibrio óptimo entre desempeño predictivo y estabilidad frente a variaciones en los datos.

- **XGBoost**

Para complementar la estimación de los parámetros del modelo GR2M, tanto para las cuencas vecinas como las presas del Sistema Cutzamala, con un modelo no lineal y de mayor complejidad estructural, se implementó el algoritmo XGBoost, el cual construye un ensamble de árboles de decisión optimizados de manera secuencial.

El modelo fue implementado en Python mediante la librería *XGBoost*, aplicando los mismos conjuntos de variables predictoras y valores normalizados empleados en la regresión lineal múltiple.

Entre los hiperparámetros clave configurados se encuentran aquellos que controlan la arquitectura del modelo y su capacidad de generalización.

El modelo final se configuró con los siguientes valores:

n_estimators=100, max_depth=5, learning_rate=0.1, subsample=0.8, colsample_bytree=0.8 y reg_lambda=1.

Esta configuración permite adaptarse a particularidades hidrológicas como las que presentan las cuencas analizadas y las cuencas de las presas del Sistema Cutzamala, además de que la utilización de ellos permite asegurar un buen rendimiento predictivo y la eficiencia computacional.

d) Control del Sobreajuste

La calibración de parámetros hidrológicos del modelo GR2M corre el riesgo de sobreajuste, especialmente debido a la cantidad limitada de cuencas disponibles con información hidrométrica.

En el caso del SVR, la regularización mediante el parámetro C y la restricción del margen ϵ limitaron la flexibilidad del modelo, evitando que este memorizara particularidades de los datos de entrenamiento.

En el XGBoost, la profundidad de los árboles se restringió a un máximo de cinco niveles y se aplicaron tasas de aprendizaje moderadas ($\eta = 0.1$) ya que una profundidad excesiva o demasiados árboles puede ajustarse a variaciones irrelevantes (Ying, 2019).

Estas estrategias garantizaron que los modelos mantuvieran una adecuada capacidad de generalización y evitaran un ajuste excesivo a las variaciones de los datos hidrológicos.

e) Implementación y evaluación

La implementación de ambos modelos se realizó en Python, utilizando las librerías *Scikit-Learn* (para SVR) y *XGBoost* (para el algoritmo del mismo nombre). En ambos casos se emplearon los conjuntos de datos normalizados y transformados previamente en la etapa previa a la aplicación de la regresión lineal múltiple.

Cabe señalar que, aunque XGBoost tiene la capacidad de manejar variables en diferentes escalas de manera intrínseca, se mantuvo la normalización por consistencia metodológica. El desempeño de cada modelo se evaluó mediante el coeficiente de determinación (R^2), seleccionando para cada parámetro x_1 y x_2 aquel modelo que presentó el mayor valor de dicha métrica y una interpolación hidrológica coherente.

f) Selección final de modelos

Tras la evaluación comparativa de desempeño entre los algoritmos implementados, se observó que el modelo SVR presentó un mejor ajuste para la estimación del parámetro x_1 , asociado con la capacidad de almacenamiento de agua en el suelo. Esto se debió a su adecuada respuesta ante relaciones moderadamente no lineales y a su estabilidad en conjuntos de datos de tamaño reducido.

Por otro lado, el modelo XGBoost mostró un desempeño superior en la estimación del parámetro x_2 , asociado al intercambio neto de agua entre la cuenca y sistemas externos, el cual permite compensar pérdidas o ganancias no representadas explícitamente en el balance hídrico mensual. Su capacidad para capturar interacciones más complejas entre las variables predictoras permitió representar con mayor precisión los procesos hidrológicos asociados.

En consecuencia, SVR fue seleccionado para la predicción del parámetro x_1 y XGBoost para la predicción del parámetro x_2 , configurando así la estructura final del modelo de predicción adoptado en este estudio.

g) Implementación de la validación cruzada *Leave-One-Out*

Para la estimación de los parámetros x_1 y x_2 del modelo GR2M, se implementó la validación cruzada *Leave-One-Out* (LOO) tanto para la regresión espacial como para los modelos de *machine learning*. Esta elección se basó en el tamaño reducido del conjunto de datos, compuesto por siete cuencas con información hidrométrica que rodean a las presas del Sistema Cutzamala.

El procedimiento consistió en realizar siete iteraciones, donde en cada una de ellas una cuenca fue excluida del conjunto de calibración y posteriormente utilizada como conjunto de validación. De esta forma, los modelos, tanto la regresión espacial como los algoritmos de *machine learning*, se ajustaron con seis cuencas y se evaluaron con la restante, garantizando una validación completa y sin pérdida de información.

Este esquema se implementó para la estimación de los parámetros x_1 y x_2 del modelo GR2M, los cuales fueron calculados en cada iteración a partir de los modelos ajustados con las demás cuencas. Posteriormente, se simuló el caudal con base en los parámetros estimados y se evaluó el desempeño de cada modelo mediante el NSE permitiendo determinar el grado de ajuste y generalización en cada caso.

6. Estimación espacial de parámetros

La estimación espacial de los parámetros (x_1 y x_2) se realizó mediante un enfoque geoestadístico que utilizó las mismas variables predictoras en las etapas previas (RLM y *machine learning* aplicadas a valores tabulados), pero transformadas a formato ráster. Estas capas ráster, representan espacialmente los predictores con correlación identificada previamente, se emplearon en un esquema de regresión análogo al utilizado en fases anteriores, considerando tres variables independientes.

A partir de la combinación lineal de las capas ráster, se generaron mapas espaciales de los parámetros x_1 y x_2 . Posteriormente, mediante un proceso de corte espacial y cálculo de medias ponderadas por área, se obtuvieron los valores representativos de cada parámetro para cada cuenca de interés (Figura 5).

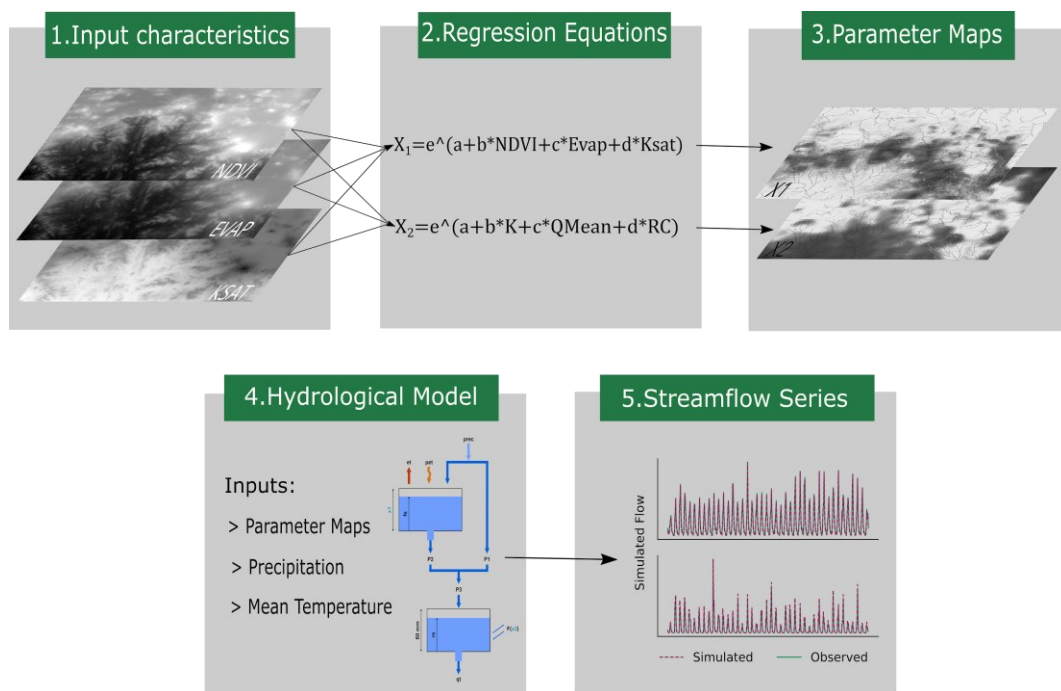


Figura 5. Metodología de Regresión Espacial

La calibración del modelo consistió en optimizar los coeficientes de la regresión lineal múltiple de los ráster, de forma que se maximizara la eficiencia del modelo, medida con la métrica NSE entre los caudales observados y simulados.

Esta tarea se formuló como un problema de minimización de una función de pérdida (función objetivo), que fue resuelto utilizando el algoritmo *L-BFGS-B*, implementado en la librería *SciPy*.

El uso de *L-BFGS-B* se justificó principalmente por dos razones.

En primer lugar, debido a la naturaleza espacial de los datos, ya que los ráster contienen miles de valores por píxel, lo que implica un volumen de información elevado incluso para una sola variable predictora. En este contexto, *L-BFGS-B* resulta más eficiente que otros métodos

de optimización, ya que aprovecha la información del gradiente y requiere menor capacidad de almacenamiento, lo que reduce considerablemente el tiempo de cómputo.

En segundo lugar, se seleccionó este método tras realizar pruebas comparativas con el algoritmo Nelder-Mead, utilizado inicialmente para mantener coherencia metodológica con etapas previas del estudio. Sin embargo, el tiempo de ejecución de Nelder-Mead resultó sustancialmente mayor, por lo que se optó por *L-BFGS-B*, logrando reducir el tiempo de cálculo aproximadamente a una tercera parte sin afectar la calidad del ajuste de los parámetros x_1 y x_2 del modelo GR2M.

Asimismo, la posibilidad de establecer límites inferiores y superiores de los parámetros permitió restringir los valores estimados a intervalos físicamente plausibles, evitando soluciones no realistas y manteniendo la estabilidad numérica del modelo durante el proceso de calibración.

7. Simulación de caudal de entrada a las presas del Sistema Cutzamala

Para estimar los parámetros x_1 y x_2 en las presas del Sistema Cutzamala se siguió un enfoque coherente con las metodologías desarrolladas para las cuencas calibradas.

Es importante destacar que los valores de estos parámetros no fueron incluidos en el conjunto de entrenamiento de los modelos, sino que se utilizaron únicamente como conjunto de prueba (*testing set*). Esto permitió evaluar la capacidad de generalización de los modelos ante sitios no considerados durante la calibración.

Las variables predictoras empleadas fueron las mismas utilizadas en las cuencas vecinas:

- Para x_1 : NDVI, Evapotranspiración (Evap) y KSat.
- Para x_2 : K, QMean y RC.

1. Regresión Empírica (Regresión Lineal Múltiple):

Se aplicaron los modelos ajustados previamente de RLM previamente ajustados para x_1 y x_2 sobre las variables predictoras correspondientes a las presas. Las predicciones se realizaron en el espacio logarítmico y posteriormente se transformaron a la escala original mediante la función exponencial para su uso en las simulaciones hidrológicas.

2. Regresión Espacial

Se emplearon los mapas de variación espacial de x_1 y x_2 generados mediante la metodología descrita y procesados en SAGA GIS. Para cada presa se extrajo el valor medio de los parámetros en la ubicación geográfica correspondiente, obteniendo así una estimación espacial directa de x_1 y x_2 .

3. Modelos de *Machine Learning*

Los modelos SVR (para x_1) y XGBoost (para x_2), previamente entrenados con las cuencas calibradas y utilizando las mismas variables predictoras, se aplicaron a las presas para obtener predicciones independientes.

Dado que únicamente se dispone de tres observaciones (una por presa), no fue posible aplicar validación cruzada ni calcular métricas de error como R^2 ; por ello, los hiperparámetros se definieron de forma conservadora, priorizando la estabilidad y la capacidad de generalización sobre el ajuste fino.

VI. Resultados y discusión

En este capítulo se presentan los resultados obtenidos a partir de la implementación del modelo GR2M en las cuencas que rodean y que alimentan al Sistema Cutzamala, así como la evaluación de distintas metodologías para la estimación de sus parámetros y la simulación de caudales. Se abordan tanto los aspectos estadísticos como hidrológicos del desempeño del modelo, considerando múltiples escalas temporales y espaciales.

Los análisis comprenden la comparación del desempeño de los diferentes enfoques de estimación, optimización individual, regresión lineal múltiple, algoritmos de *machine learning* y regresión espacial, junto con la validación cruzada del modelo, la evaluación del comportamiento hidrológico simulado en las cuencas y el análisis específico del balance de agua en las presas del Sistema Cutzamala.

A. Calibración del modelo GR2M para cada cuenca

La Tabla 5 muestra los parámetros x_1 y x_2 y el valor del criterio NSE obtenidos a partir de las cuatro metodologías presentadas: Optimización Individual (cuenca por cuenca), RLM, Regresión Espacial y Machine Learning. Los resultados obtenidos por Regresión Lineal Múltiple, Regresión Espacial y Machine Learning se describen en los siguientes subcapítulos. Se consideró como modelo base la metodología de Optimización Individual, ya que esta ajusta directamente los parámetros del modelo GR2M para cada cuenca con el algoritmo de Nelder-Mead, buscando maximizar el NSE entre los caudales simulados y observados.

Tabla 5. Parámetros x_1 y x_2 estimados con distintas metodologías para cada cuenca con registros hidrométricos.

Método	Variable	18372	18232	12347	12466	18540	12561	12543
Optimización Individual	x_1	1332.82	1241.16	2001.98	176.07	3692.53	1168.43	1788.99
	x_2	0.6919	0.6298	0.4419	0.4146	0.8056	0.5685	0.5094
	NSE	0.8459	0.8112	0.7623	0.6480	0.6459	0.6258	0.4521
Regresión Lineal Múltiple	x_1	1161.48	1503.05	1429.70	327.91	3338.37	597.09	1518.41
	x_2	0.6670	0.6941	0.4784	0.5358	0.8270	0.5707	0.5996
	NSE	0.8284	0.7976	0.7554	0.5803	0.6371	0.5973	0.4137

Método	Variable	18372	18232	12347	12466	18540	12561	12543
Regresión Espacial	x1	1364.70	1896.59	1688.24	350.80	4104.11	1088.84	2139.38
	x2	0.6615	0.6553	0.3680	0.4435	0.8134	0.5013	0.5442
	NSE	0.8425	0.7902	0.7549	0.5981	0.6375	0.5929	0.4064
Machine Learning	x1	1205.32	1371.71	1811.05	194.57	3342.57	1057.69	1911.68
	x2	0.6918	0.6296	0.4419	0.4154	0.8037	0.5680	0.5101
	NSE	0.8446	0.8100	0.7606	0.6462	0.6423	0.6246	0.4518

De las siete cuencas analizadas, tres de ellas (18372, 18232 y 12347) obtuvieron un buen ajuste ($NSE > 0.7$), mientras que la cuenca 12543 obtuvo el menor desempeño en la simulación de caudales mensuales ($NSE \sim 0.45$). La explicación y comparación de los errores obtenidos para cada cuenca se abordará en los siguientes subcapítulos.

B. Análisis de correlación entre variables y parámetros del modelo

Para identificar relaciones entre las variables explicativas y los parámetros x_1 y x_2 , se elaboró una matriz de correlación de Pearson (Figura 6) (Ver Tabla A.1 y A.2).

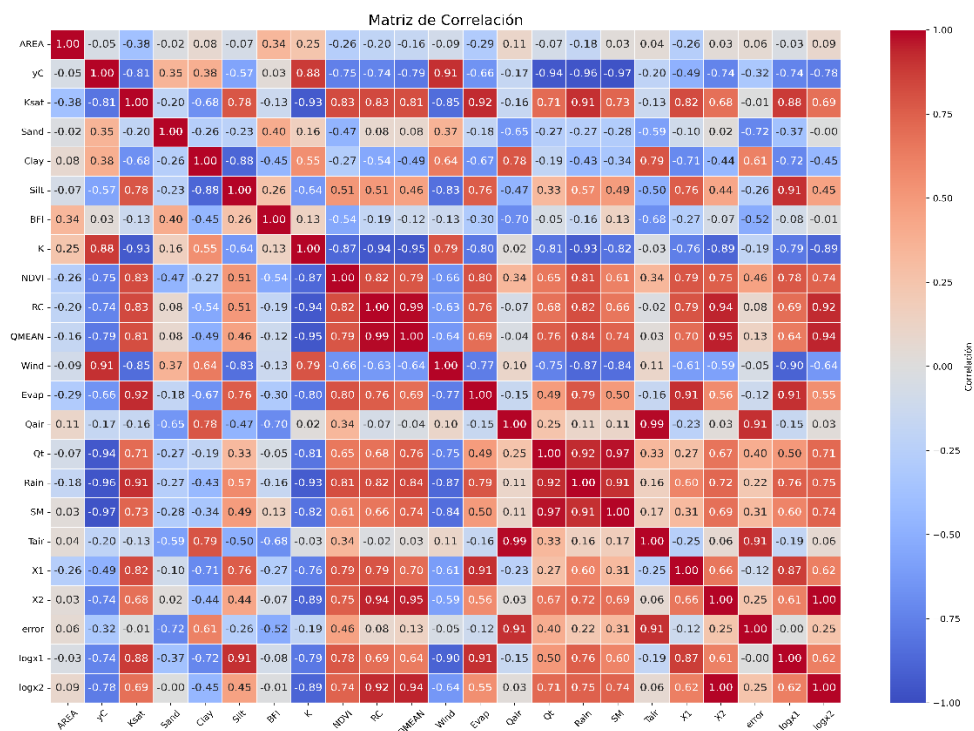


Figura 6. Matriz de correlación de variables estandarizadas explicativas con los parámetros $\ln(x_1)$ y $\ln(x_2)$ del modelo GR2M.

La matriz fue generada a partir de las librerías de *Pandas* y *Seaborn*, mediante un mapa de calor en el que los tonos más cálidos indican mayor correlación positiva, y los tonos más fríos indican mayor correlación negativa (Ver Tabla A.1 y A.2). A partir de esta matriz, se seleccionaron las variables más relevantes para cada parámetro considerando tanto su magnitud de correlación como su justificación física.

En el caso de $\ln(x1)$, se presentan los valores de correlación obtenidos para las variables seleccionadas y su relación con el parámetro calibrado, destacando la correspondencia física entre la cobertura vegetal, la evapotranspiración y la capacidad de almacenamiento superficial.

1. Evapotranspiración (Evap)

La evapotranspiración media anual estandarizada presentó una correlación positiva alta ($r = 0.91$), lo que sugiere que los suelos con mayor capacidad de almacenamiento (alto $x1$) presentan una mayor disponibilidad hídrica para los procesos evapotranspirativos. El valor promedio de evapotranspiración en la zona de estudio ($816.08 \text{ [kg/m}^2\text{s]}$) indica una demanda evaporativa sostenida que es abastecida principalmente por el almacenamiento edáfico (Ali et al., 2020). Este comportamiento es particularmente relevante en períodos interlluvias, cuando el suelo actúa como reservorio clave, zonas con vegetación nativa adaptada a condiciones semiáridas y sistemas agrícolas con manejo eficiente del agua.

2. Conductividad Hidráulica Saturada (KSat)

Muestra una correlación alta ($r = 0.88$), similar a lo encontrado por (van Genuchten, 1980). Este comportamiento podría explicarse por la presencia de estructuras geológicas particulares (e.g., fracturas o fallas preferenciales), efecto de escala en mediciones de *KSat* (valores puntuales contra representativos de cuenca) que combinan alta permeabilidad con capacidad de retención (Tuller, 2020), o interacción con procesos hidrogeológicos profundos no considerados en el modelo (Ver Tabla A.1)

3. Índice de Vegetación de Diferencia Normalizada (NDVI)

Muestra una correlación positiva significativa ($r = 0.78$). Esto indica que la capacidad de almacenamiento del suelo aumenta con la cobertura vegetal. Estudios previos han demostrado que la vegetación mejora la retención hídrica al promover la formación de poros en el suelo a través de raíces y materia orgánica (Wei y Wan, 2022). El NDVI promedio de la zona fue moderado ($0.55 [-]$), este valor indica una cobertura vegetal presente pero no completamente densa. Esto sugiere que incluso incrementos limitados (ej: restauración de pastizales) podrían mejorar significativamente la capacidad de almacenamiento del suelo. No obstante, el valor de la correlación puede estar modulada por factores locales como:

- Tipo de Vegetación: Cultivos o especies de alto consumo hídrico pueden aumentar el NDVI, pero reducir el almacenamiento efectivo.

- Intervención humana: Actividades como el riego artificial pueden enmascarar la relación natural entre vegetación y retención de agua.

En el caso de $\ln(x_2)$, se presentan los valores de correlación obtenidos para las variables seleccionadas y su relación con el parámetro calibrado, destacando la correspondencia física entre el caudal medio anual, el coeficiente de escorrentía y la permeabilidad del suelo en la dinámica del flujo base y la recesión del caudal en la cuenca.

1. Caudal Medio Anual (QMean)

Presentó una correlación positiva excepcional ($r = 0.91$), indicando que cuencas con mayor escorrentía (QMean promedio en la zona de estudio = 369.89 [mm/year]) tienden a exhibir valores de $x_2 > 1$. Este comportamiento sugiere la presencia de aportes extremos no contabilizados en el balance hídrico superficial como recarga subterránea desde acuíferos adyacentes, flujos laterales intercuenas a través de unidades geológicas permeables, transferencias antropogénicas (p.e. retornos de riesgo).

2. Coeficiente de escorrentía (RC)

La correlación positiva ($r = 0.88$) refuerza que $\ln(x_2)$ está modulado por procesos subsuperficiales que potencian la generación de escorrentía (RC promedio = 0.37 [–]), posiblemente relacionados con conectividad hidrogeológica entre acuíferos y cauces, descarga subterránea en zonas riparias o litologías fracturadas.

3. Permeabilidad del suelo (K)

Se observa una correlación negativa significativa ($r = -0.77$), coherente con el marco teórico que asocia alta permeabilidad (K promedio = 0.48 [–]) con descargas laterales y percolación del sistema ($x_2 < 1$). No obstante, se identificaron excepciones en áreas con suelos volcánicos fracturados (alta K, pero $x_2 > 1$), zonas de recarga artificial intensiva o acuíferos someros desconectados del drenaje superficial.

Estas variables fueron posteriormente utilizadas como predictores en el modelo de regresión lineal y algoritmos de machine learning para estimar los parámetros x_1 y x_2 previamente optimizados.

Es importante mencionar que ni el área ni la latitud tienen una correlación significativa (-0.03 y -0.02 respectivamente) con los parámetros $\ln(x_1)$ y $\ln(x_2)$ a pesar de que son necesarios para que el modelo GR2M funcione correctamente.

C. Regionalización de los parámetros x_1 y x_2 del modelo GR2M

Como parte del análisis predictivo de los parámetros del modelo hidrológico, una vez realizados los ajustes individuales mediante el algoritmo de optimización Nelder-Mead, se desarrollaron tres enfoques distintos de modelado: regresión empírica (RLM), regresión espacial y modelos de machine learning (SVR y XGBoost). En los tres casos se utilizaron las

mismas variables predictoras: NDVI, Evapotranspiración (Evap), Conductividad Hidráulica Saturada (KSat) para x_1 ; y Conductividad Hidráulica (K), Caudal Medio Anual (QMean) y Coeficiente de Escorrentía (RC) para x_2 . La selección de estas variables se basó en el análisis de correlación descrito previamente.

Los modelos de regresión, tanto empírica como espacial, dieron lugar a expresiones explícitas que permiten estimar los parámetros x_1 y x_2 a partir de sus respectivas variables predictoras. Estas ecuaciones, presentadas en la Tabla 7, constituyen una herramienta clave para la transferencia de parámetros a cuencas no aforadas, es decir, aquellas donde no se cuenta con mediciones directas de caudal.

Por su parte, los modelos de machine learning no generan ecuaciones analíticas fáciles de representar debido a la gran cantidad de parámetros utilizados. Aunque SVR y XGBoost emplean un gran número de parámetros internos aumentando su complejidad, su inclusión permite comparar el desempeño de enfoques tradicionales de regresión con técnicas computacionalmente más sofisticadas. A continuación, se presentan los modelos utilizados, sus variables de entrada y los hiperparámetros seleccionados, mostrando de manera resumida la configuración aplicada durante el entrenamiento:

Tabla 6. Configuración de hiperparámetros de los modelos SVR y XGBoost utilizados en la estimación de los parámetros del modelo GR2M.

Modelo	Variables de Entrada	Hiperparámetros principales
SVR	NDVI, Evap, KSat	<i>Kernel RBF</i> , $C = 10$, $\gamma = scale$, $\varepsilon = 0.1$
XGBoost	QMean, RC, K	$n_estimators = 100$, $max_depth = 5$, $learning_rate = 0.1$, $subsample = 0.8$, $colsample_bytree = 0.8$, $reg_lambda = 1$, $random_state = 42$

La selección de los hiperparámetros para SVR y XGBoost (Tabla 6) se realizó mediante un enfoque exploratorio, considerando dos aspectos fundamentales: la complejidad del modelo y su capacidad predictiva. En el caso de SVR, se evaluaron distintas combinaciones de C , γ y ε , analizando su efecto sobre la estabilidad de las predicciones frente a variaciones en los datos de entrada y su influencia en el R^2 , con el objetivo de maximizarlo.

Para XGBoost, se ajustaron el número de árboles, la profundidad máxima y la tasa de aprendizaje, priorizando configuraciones que ofrecieran un desempeño predictivo sin inducir sobreajuste, también buscando optimizar R^2 . Este procedimiento, aunque cualitativo, permitió

identificar valores de hiperparámetros que equilibran precisión y generalización de los modelos.

Tabla 7. Ecuaciones de regresión correspondientes a los modelos optimizados. NDVI, Evap, KSat, K, QMean, y RC corresponden a las variables estandarizadas del Índice Normalizado de Diferencia de Vegetación, Evapotranspiración, Conductividad Hidráulica Saturada, Caudal

	Parámetro	Ecuación
Regresión Empírica	x_1	$\ln(x_1) = 6.0603 + 0.1137 NDVI + 1.0403 Evap + 0.2989 Ksat$
	x_2	$\ln(x_2) = -0.2977 + 0.6108K + 0.4316QMean - 0.1030 RC$
Regresión Espacial	x_1	$\ln(x_1) = 6.06712 + 0.1184 NDVI + 1.0269 Evap + 0.2926 Ksat$
	x_2	$\ln(x_2) = -0.3478 + 0.8064K + 0.5139QMean - 0.0781 RC$

Medio Anual y Coeficiente de Escorrentía, respectivamente.

El rendimiento de todos los modelos fue evaluado mediante la métrica Nash-Sutcliffe Efficiency (NSE), lo que permitió cuantificar la capacidad de las metodologías para reproducir el caudal a partir de los valores optimizados de x_1 y x_2 . Esta comparación se muestra en la Tabla 5 y es clave para comparar el desempeño relativo entre enfoques y seleccionar los modelos más adecuados para la predicción y simulación del caudal en las cuencas del Sistema Cutzamala. Todas las cuencas, con excepción de la cuenca 12466, presentan valores de x_1 superiores a 1000 mm con todas las metodologías empleadas (Tabla 5).

Para la estimación espacial de los parámetros x_1 y x_2 con la RLM espacial, se tomó la intersección y los coeficientes mostrados en la Tabla 7, y se usaron las variables predictoras en formato ráster. La variación espacial de x_1 y x_2 se muestra en la Figura 7, en donde se muestran valores de x_1 entre 3 y 6040 mm y x_2 entre 0.37 y 1.35. Estos resultados se encuentran dentro de los rangos reportados por la literatura, donde x_1 varía típicamente entre 140 a 2640 mm y x_2 entre 0.21 a 1.31 (Mouelhi et al. (2006)).

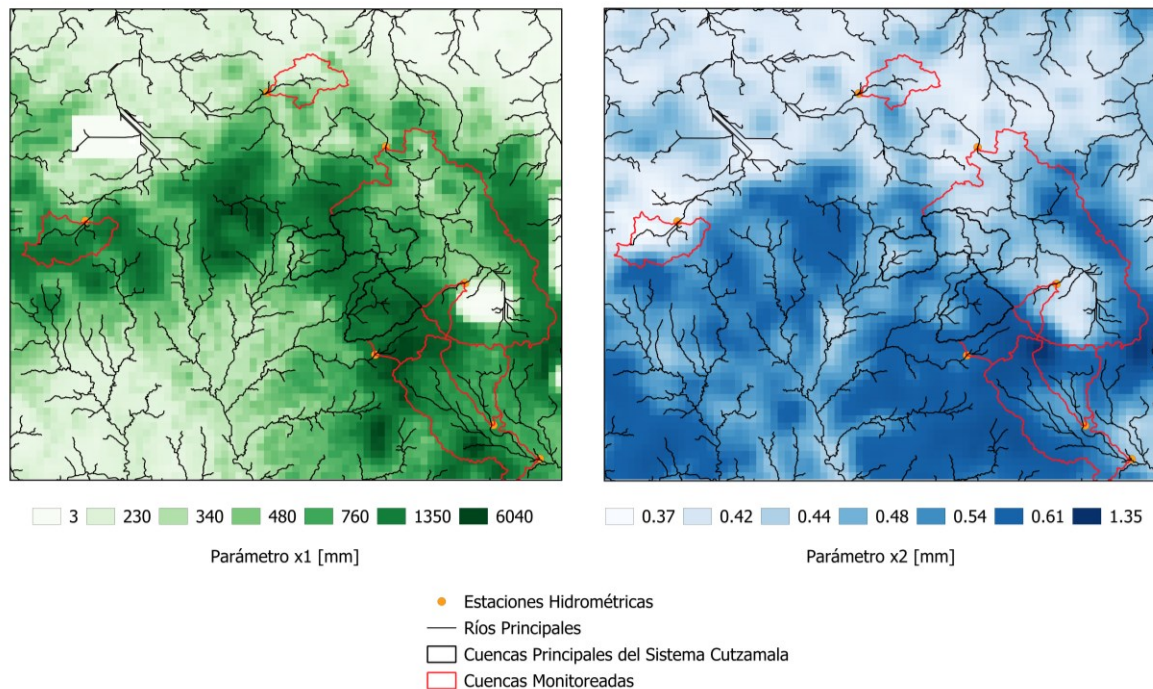


Figura 7. Mapa de variaciones de los parámetros x_1 y x_2 obtenidos a partir de las RLM espaciales.

El patrón espacial de x_1 (Figura 7) sugiere una alta capacidad de almacenamiento de agua en el suelo en las zonas altas, un comportamiento típicamente observado en regiones tropicales húmedas o en áreas con acuíferos extensos. Por otro lado, la geología y edafología de la zona podrían estar asociadas a estos altos valores de x_1 debido a que esta región es atravesada por el Eje Neovolcánico, dominado por suelos derivados de cenizas volcánicas, como andosoles y luvisoles, los cuales presentan una alta capacidad de retención de humedad debido a su estructura porosa y su alto contenido de materia orgánica. La cuenca 12466, ubicada entre los estados de Guanajuato y Querétaro, presenta el valor de x_1 más bajo ($x_1 < 351$ mm). Este valor puede estar asociado con materiales volcánicos como tobas riolíticas y andesitas, junto con suelos lacustres (Figura 7). Estos materiales comparten mecanismos que limitan o modifican la retención de agua efectiva: (1) porosidad no conectada en tobas, (2) rápida infiltración por fracturas en andesitas, y (3) la retención estática superficial en arcillas lacustres. Estos comportamientos coinciden con lo reportado por Mouelhi et al. (2006), en donde cuencas con suelos poco profundos o rocosos presentan valores de bajos de x_1 .

En cuanto al parámetro x_2 , el cual representa el intercambio subterráneo neto de agua con el exterior de la cuenca, se observan valores menores a 1 en todas las cuencas analizadas. De acuerdo con la literatura, valores de x_2 inferiores a 1 sugieren pérdidas de agua subterránea hacia acuíferos profundos o cuencas vecinas, un comportamiento común en

regiones semiáridas o montañosas. Esta interpretación es coherente con el comportamiento hidrológico regional, donde numerosos acuíferos están en estado de sobreexplotación y el relieve montañoso favorece el drenaje rápido, lo que limita la retención prolongada de agua subterránea dentro de las cuencas. Aunque los suelos volcánicos permiten una alta infiltración, esta agua puede estar migrando rápidamente fuera del sistema si existen fracturas o conexiones con unidades hidrogeológicas regionales. Desde una perspectiva topográfica, se trata además de cuencas con gradientes pronunciados (Tabla 2), en donde los procesos de escorrentía e intercambio lateral tienden a ser dirigidos hacia el exterior. Los valores bajos de x_2 también pueden estar asociados con una mayor vulnerabilidad a sequías, ya que indican que parte del agua disponible se pierde de forma estructural hacia zonas profundas o vecinas.

Para las estimaciones de x_1 y x_2 , los modelos predictivos como la regresión lineal o los algoritmos de machine learning, intentan generalizar relaciones a partir de un conjunto limitado de datos y variables explicativas, lo que introduce una capa adicional de incertidumbre, ya que la calidad de la estimación dependerá no solo del comportamiento hidrológico, sino también de la representatividad de las variables predictoras, la calidad de los datos de entrenamiento y la capacidad del modelo para capturar relaciones no lineales o complejas. En otras palabras, mientras que la optimización individual se enfoca en encontrar la mejor combinación de parámetros por cuenca, los modelos de regionalización predictivos buscan aproximar estos valores optimizados mediante funciones de aprendizaje en todo el conjunto de datos.

En las Figuras 8 y 9 se presentan los valores estimados frente a los valores observados para $\ln(x_1)$ y $\ln(x_2)$, utilizando las tres metodologías presentadas. Cada punto representa una cuenca, y se incluye una línea de referencia 1:1, así como el valor de R^2 correspondiente a cada modelo.

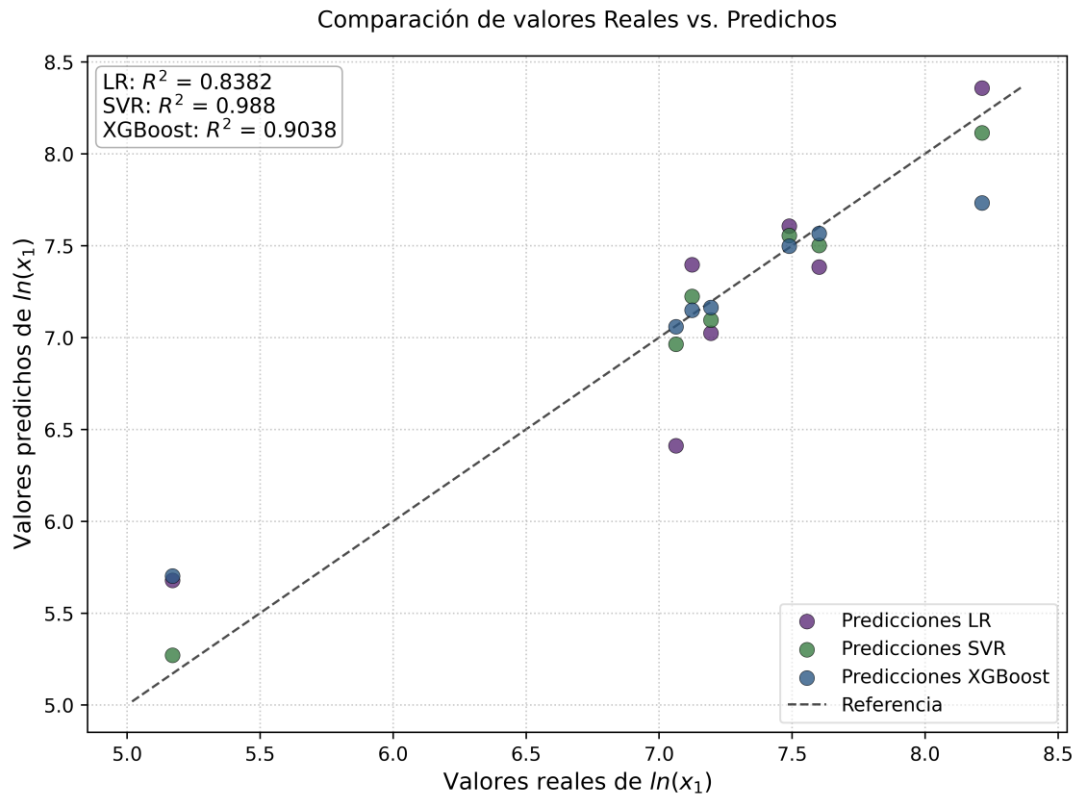


Figura 8. Comparación entre valores observados y estimados de $\ln(x_1)$ utilizando LR, SVR y XGBoost.

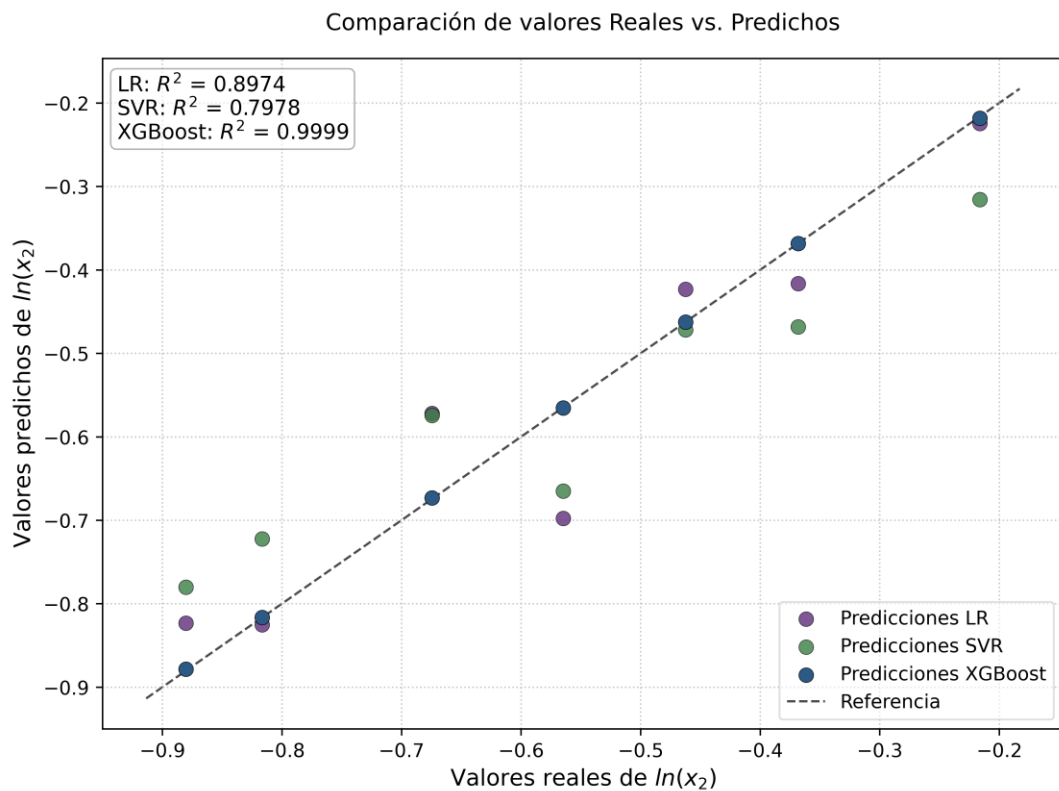


Figura 9. Comparación entre valores observados y estimados de $\ln(x_2)$ utilizando LR, SVR y XGBoost.

Los resultados mostraron que SVR y XGBoost superaron, en general, a la regresión lineal múltiple en la predicción de los parámetros x_1 y x_2 . En el caso de x_1 , SVR alcanzó un R^2 de 0.988, mientras que XGBoost obtuvo un R^2 de 0.9038, en contraste con la regresión lineal que tuvo un valor de 0.8382. Para x_2 , el comportamiento fue diferente: XGBoost presentó un R^2 de 0.999, muy por encima de SVR (0.7978) y regresión lineal múltiple (0.8974).

Estos valores sugieren que SVR es el modelo más adecuado para x_1 , mientras que XGBoost mostró un desempeño excepcional en la estimación de x_2 . Cabe destacar que los hiperparámetros de los modelos fueron ajustados de forma independiente para cada parámetro, lo cual pudo haber influido en las diferencias de desempeño observadas. El ajuste casi perfecto de XGBoost para x_2 sugiere un riesgo de sobreajuste (*overfitting*) al conjunto de calibración. Este problema podría mitigarse mediante técnicas como la regularización de los árboles, la limitación de la profundidad máxima o el uso de un conjunto de validación con *early stopping* para controlar el entrenamiento y mejorar la capacidad de generalización del modelo.

En contraste, la regresión lineal múltiple mostró un desempeño menor, especialmente en el caso de x_1 . Por otro lado, a pesar de los mejores ajustes en la estimación de x_1 y x_2 , no se aprecian mejoras significativas en la simulación de caudal según el NSE (Tabla 5), resultado en valores promedio de NSE para todas las cuencas de:

- 0.6844 para la Optimización Individual
- 0.6663 para la Regresión Multilineal
- 0.6828 para los modelos de *Machine Learning*
- 0.6603 para la Regresión Espacial

Cabe destacar que el error observado no solo está relacionado con la estimación de los parámetros, sino también con la calidad de los datos de entrada, incluyendo la presencia de valores atípicos y posibles errores de medición en los caudales observados. Por ejemplo, la cuenca 18372 obtuvo el NSE más alto (0.8459), lo cual puede atribuirse a la gran cantidad de datos disponibles y a la ausencia significativa de outliers. En contraste, la cuenca 12543 presentó el menor NSE (0.4143), posiblemente debido a la presencia de outliers o a errores sistemáticos en los registros de caudal, lo cual afecta negativamente el ajuste del modelo y a la métrica NSE, independientemente de la metodología empleada.

D. Comparación del caudal simulado en cuencas con registros hidrométricos

Con el objetivo de evaluar el desempeño de los modelos propuestos en la simulación del caudal en las cuencas que rodean a las presas del Sistema Cutzamala, se construyeron visualizaciones detalladas para cada una de las cuencas analizadas.

Las Figuras 10 a 16 presentan los resultados correspondientes a cada cuenca individual. Cada figura incluye cuatro paneles: (1) series de precipitación acumulada y temperatura

media en escala mensual, (2) precipitación y temperatura media mensual, (3) comparación entre el caudal observado y los caudales simulados por los cuatro enfoques (Optimización, Regresión, Regresión Espacial y Machine Learning) a escala mensual, en el lado izquierdo de la figura se encuentra la escala en mm y en el lado derecho en Hm^3 , la cual fue determinada a partir del área de la cuenca y (4) caudal medio mensual. Esta representación gráfica permite contrastar visualmente el comportamiento hidrológico observado (líneas de color negro) y simulado, identificar los periodos de mayor caudal, y analizar el ajuste de cada modelo en distintos contextos climáticos.

En las gráficas de los valores medios mensuales se observa un desfase temporal entre los picos de precipitación y los máximos en el caudal, fenómeno característico en muchas cuencas hidrológicas. Aunque la precipitación comienza a incrementarse a partir de junio, los valores máximos de caudal no se alcanzan hasta septiembre, coincidiendo con el segundo pico de precipitación. Este desfase puede atribuirse a los procesos de almacenamiento en el suelo, infiltración y humedad del suelo que retrasan la respuesta de la cuenca. Asimismo, en todas las cuencas analizadas se identifica un patrón común: la temperatura media alcanza su máximo entre abril y mayo, mientras que el máximo de la precipitación ocurre durante el verano (de junio a septiembre). Este desfase temporal entre los máximos térmicos y las precipitaciones estivales explica el incremento subsecuente del caudal, que posteriormente declina con la llegada del invierno.

A continuación, se describe de manera individual el comportamiento observado en cada cuenca:

- **Cuenca 18372**

Esta cuenca presenta un valor alto de eficiencia ($NSE = 0.84$), a pesar de tener interrupciones en el periodo de registro (1980-2015), con un intervalo continuo entre 1993 y 2006. Las cuatro metodologías se comportan de forma muy similar y reproducen adecuadamente la estacionalidad observada del caudal. En septiembre, el promedio de caudal alcanza los 80 mm/mes, equivalente a aproximadamente $\sim 50 Hm^3$, lo que sugiere una capacidad robusta del modelo GR2M y de los diferentes métodos de estimación de parámetros para generalizar el comportamiento hidrológico de esta cuenca (Figura 10).

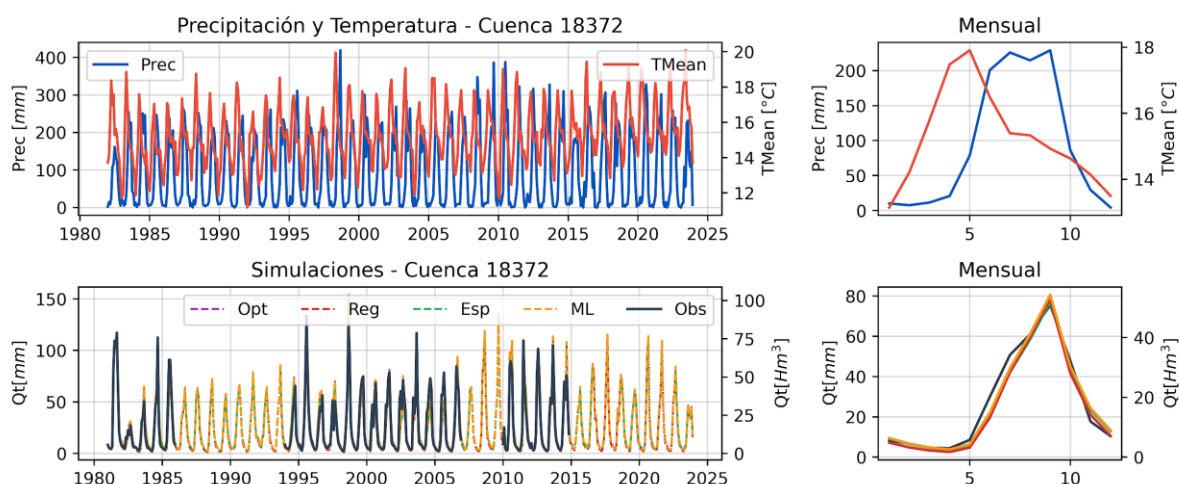


Figura 10. Simulación de caudal en la cuenca 18372.

- Cuenca 18232

El periodo de observación en esta cuenca es más continuo que en la anterior y presenta un alto valor de eficiencia ($NSE = 0.81$). Se trata además de una de las cuencas con mayores volúmenes de caudal, superando los 150 mm en diferentes años, como en el 2010, respaldado por la precipitación observada. En el panel mensual, todos los modelos replican adecuadamente la forma estacional, y alcanzan un caudal promedio de 80 mm, equivalente a 175 Hm^3 . La optimización muestra el mejor ajuste, mientras que los enfoques de regresión, espacial y machine learning presentan comportamientos prácticamente idénticos (Figura 11).

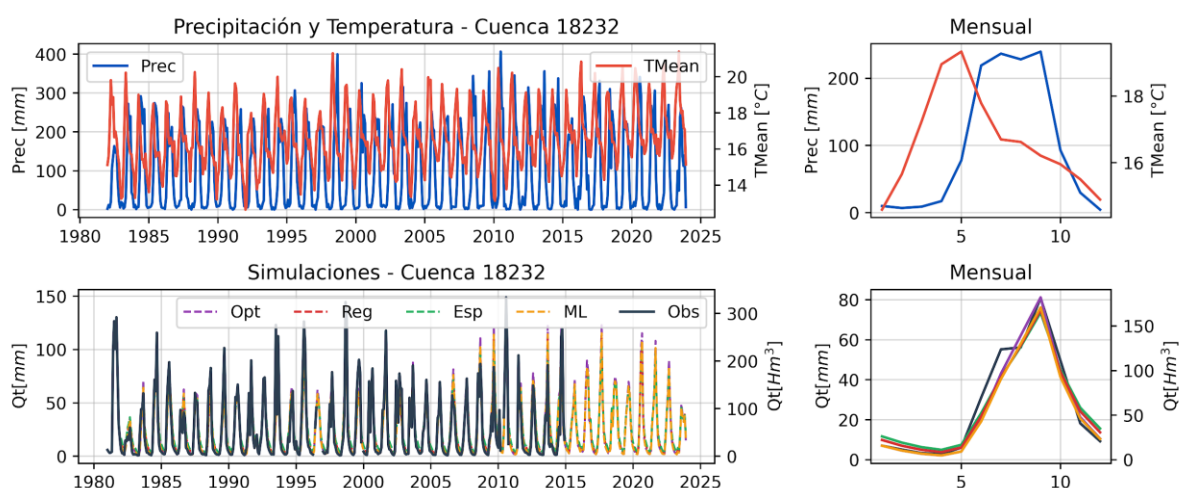


Figura 11. Simulación de caudal en la cuenca 18232.

- Cuenca 12347

Aunque el periodo de datos observados (1981-2002) es más corto que en otras cuencas, la eficiencia obtenida es satisfactoria ($NSE = 0.76$). No se presentan eventos extremos atípicos y los picos de caudal (con máximos de hasta ~60 mm) son consistentes con la precipitación registrada. En escala media mensual, los modelos de regresión, espacial y machine learning coinciden mejor con los datos observados (~25 mm o ~ 17 Hm^3), mientras que la optimización

individual tiende a sobreestimar ligeramente (~ 29 mm), sin embargo, esto no indica un ajuste menos preciso, ya que esta metodología fue la que alcanzó el NSE más alto a comparación de las otras (Figura 12).

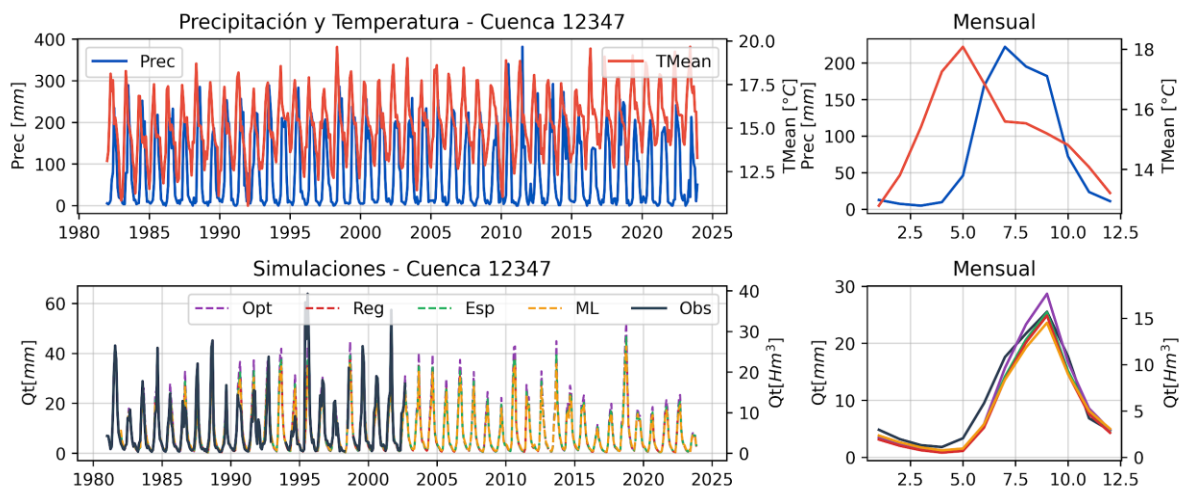


Figura 12. Simulación de caudal en la cuenca 12347.

- Cuenca 12466

Dos eventos destacan en esta cuenca: uno en julio de 1991 con ~ 170 mm en el mes, respaldado por la precipitación, y otro caudal de ~ 200 mm sin justificación meteorológica clara, lo que sugiere un posible *outlier* o problemas de los datos satelitales para detectar eventos extremos en esa fecha. Con un $NSE = 0.64$, los modelos logran compartir parcialmente el comportamiento del caudal observado. En el panel de los datos medios mensuales, los modelos de regresión multilínea, espacial y machine learning coinciden con un máximo común de ~ 20 mm, mientras que la optimización se aproxima mejor a los valores observados (~ 30 mm o ~ 12 Hm^3)(Figura 13).

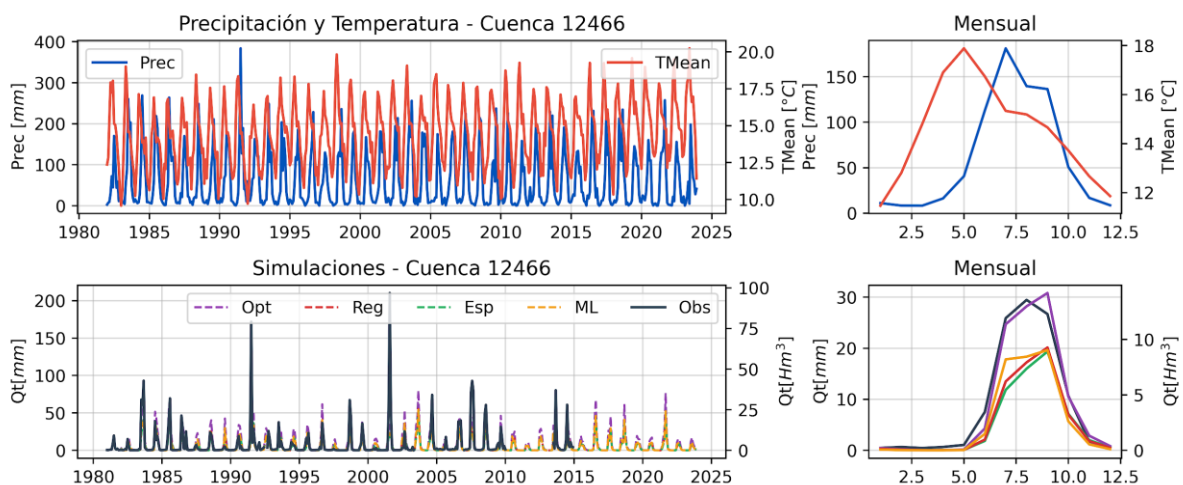


Figura 13. Simulación de caudal en la cuenca 12466.

- Cuenca 18540

Esta cuenca muestra un desempeño moderado con un $NSE = 0.64$, posiblemente afectado por las múltiples interrupciones en su serie de datos (1980 - 2015). En 2013 se observa un caudal elevado ($\sim 150 \text{ mm}$) que no se asocia a la precipitación registrada, lo cual sugiere un posible dato atípico. El modelo no simula un caudal tan elevado para ese año, reforzando esta idea. En el análisis de los valores medios mensuales, los modelos muestran una coherencia general, en donde la optimización individual predice los mayores valores ($\sim 85 \text{ mm}$, $\sim 35 \text{ Hm}^3$), seguida por el enfoque de machine learning ($\sim 80 \text{ mm}$), que se ajusta mejor a los datos observados (Figura 14).

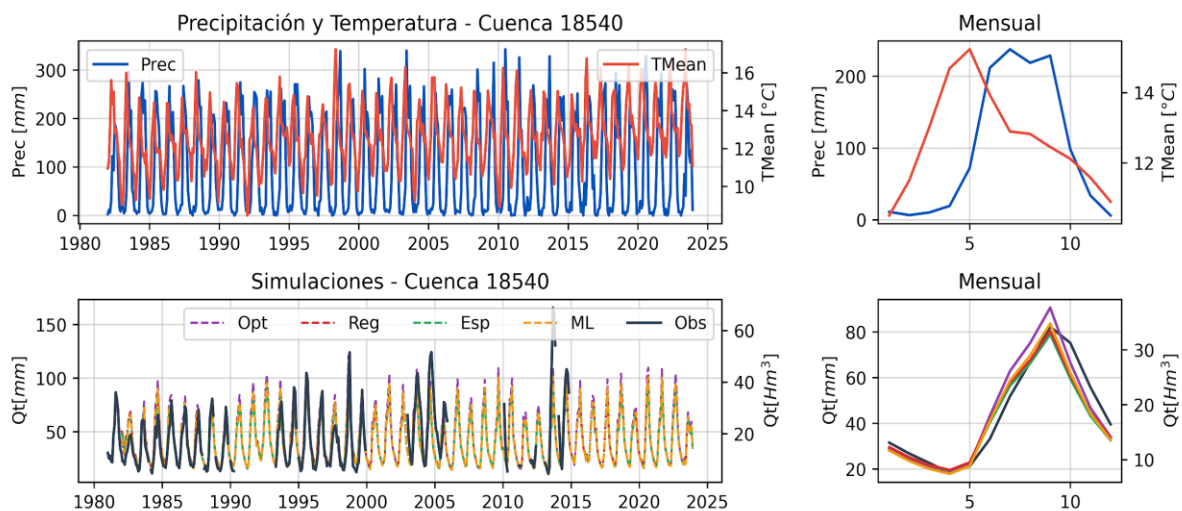


Figura 14. Simulación de caudal en la cuenca 18540.

- Cuenca 12561

El ajuste en esta cuenca es moderado ($NSE = 0.62$), lo cual podría estar asociado a la intermitencia de la serie y a que del 2010 al 2015 el caudal observado en los meses secos incrementa sin asociarse al comportamiento de años anteriores. Se identifica un posible valor atípico en 2003, donde el caudal elevado no está respaldado por la precipitación. Los modelos coinciden en la forma estacional del caudal, pero difieren en su magnitud. En el análisis mensual, la optimización individual y el modelo de machine learning reproducen patrones similares, alcanzando un caudal promedio de $\sim 35 \text{ mm}$ ($\sim 160 \text{ Hm}^3$), aunque sin alinearse completamente con los valores observados. Los modelos de regresión y la regresión espacial muestran una mayor coincidencia en magnitud con los datos reales (Figura 15).

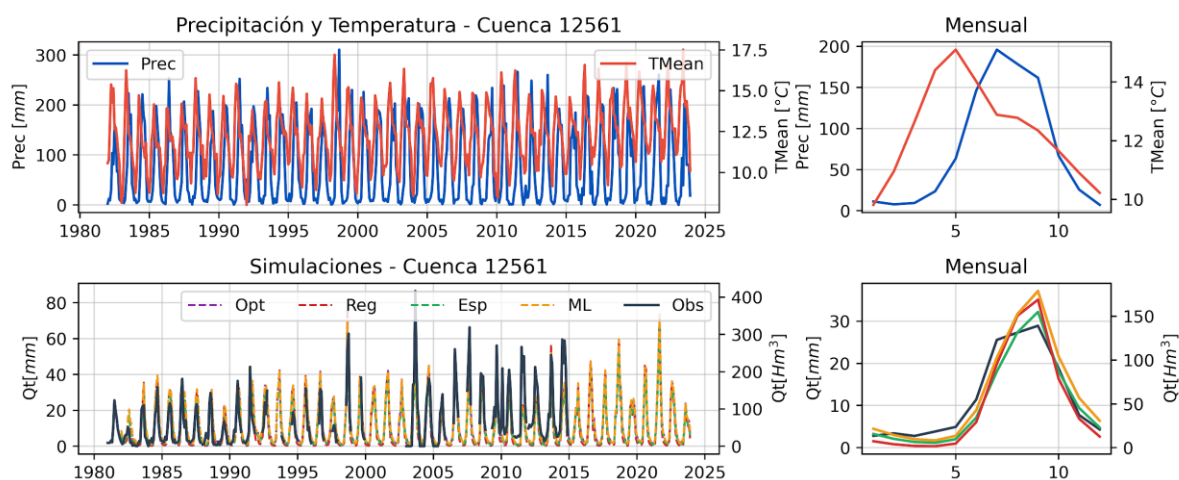


Figura 15. Simulación de caudal en la cuenca 12561.

- Cuenca 12543

Esta es la cuenca con el menor rendimiento obtenido ($NSE = 0.45$), probablemente afectada por la presencia de valores atípicos no asociados con una precipitación de magnitud proporcional. Exceptuando un evento en 1998, los picos extremos (>200 mm) carecen de justificación climática y reducen el NSE. Esta cuenca también presenta los menores valores promedio de caudal. En el panel mensual, la optimización individual tiende a sobrestimar el pico (~ 60 mm o ~ 12 Hm^3), mientras que la regresión y la regresión espacial muestran mejor coincidencia con los datos reales (~ 45 mm o ~ 8 Hm^3). El modelo de machine learning queda ligeramente por debajo del valor observado. La alta variabilidad no explicada sugiere que sería necesario revisar la calidad de los datos observados (Figura 16).

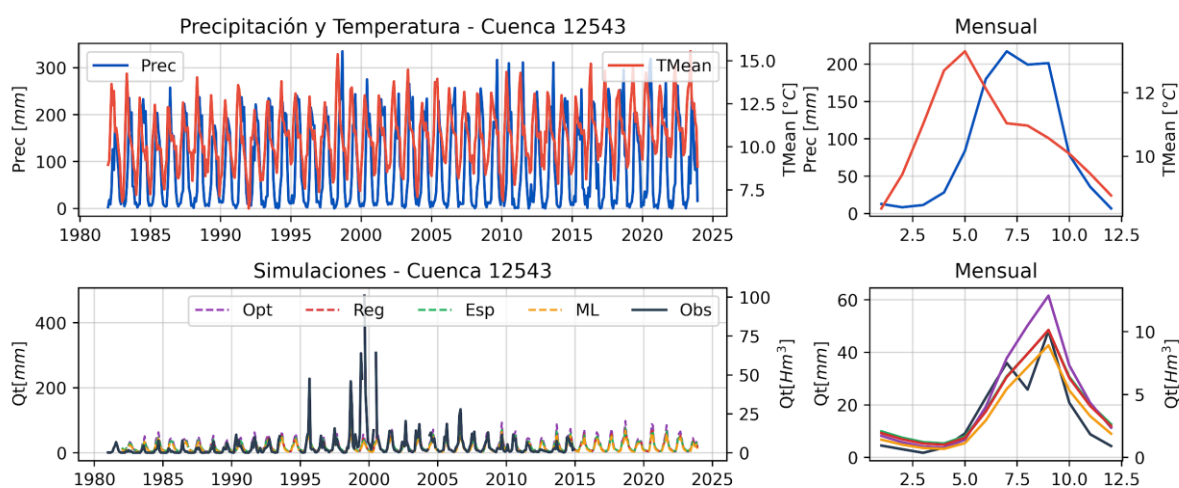


Figura 16. Simulación de caudal en la cuenca 12543.

E. Validación cruzada

Con el propósito de evaluar la capacidad de generalización de los modelos utilizados para la estimación de los parámetros $x1$ y $x2$ del modelo GR2M, se aplicó la estrategia de validación

cruzada Leave-One-Out (LOO) en dos de las metodologías exploradas: la regresión lineal múltiple y los algoritmos de *Machine Learning*.

Durante este proceso, los valores de x_1 y x_2 obtenidos previamente en la calibración individual de cada cuenca se mantuvieron como referencia. Para la validación, se construyeron nuevamente las relaciones entre los parámetros transformados logarítmicamente ($\ln X_1, \ln X_2$) y las mismas variables predictoras utilizadas para la estimación inicial, es decir: NDVI, Evapotranspiración y KSat para x_1 , y K, RC y QMean para x_2 . En el caso de los modelos de *Machine Learning*, la validación cruzada se aplicó tanto al algoritmo Support Vector Regression (SVR), utilizado para la estimación del parámetro x_1 , como a XGBoost, empleado para x_2 . En ambos casos se mantuvieron los mismos hiperparámetros definidos durante el entrenamiento inicial, con el fin de preservar la consistencia del modelo y garantizar que las variaciones en el desempeño se debieran únicamente al proceso de validación. Si bien el método LOO puede emplearse también para la selección de los hiperparámetros, en este estudio se optó por mantener los valores previamente calibrados para evitar sobreajuste y asegurar que la evaluación reflejara la capacidad de generalización de los modelos, sin efectos asociados a una nueva optimización sobre un conjunto de datos reducido.

Los resultados obtenidos de este ejercicio se presentan en la Tabla 8, la cual muestra los valores estimados x_1 y x_2 para cada cuenca al ser excluida del conjunto de calibración, junto con el valor del NSE correspondiente. Esta información permite observar el comportamiento de cada modelo al excluir datos y su capacidad de generalización.

Tabla 8. Resultados de la validación cruzada para cada cuenca para los métodos de regresión espacial y con los algoritmos de *Machine Learning* (SVR y XGBoost).

ID Cuenca	Resultado Validación Cruzada (Regresión Espacial)			Resultado Validación Cruzada (Machine Learning)		
	x_1	x_2	NSE	x_1	x_2	NSE
18232	1787.96	0.6685	0.7914	1783.00	0.6917	0.7895
18372	81.37	0.6089	0.4540	985.20	0.8036	0.6633
12347	1425.28	0.4126	0.6416	1381.33	0.5100	0.6546
12561	445.91	0.4648	0.5745	696.14	0.4420	0.4942
12466	676.40	0.5085	0.4009	1518.77	0.4864	0.2661

	Resultado Validación Cruzada (Regresión Espacial)			Resultado Validación Cruzada (Machine Learning)		
18540	5084.85	0.5931	0.0661	713.64	0.6908	0.1304
12543	26474.02	0.5793	-0.4771	1914.18	0.6297	0.1970

Las Figuras 17 a 23 muestran los resultados de la validación cruzada para cada una de las cuencas analizadas, utilizando únicamente el algoritmo SVR para la estimación de ambos parámetros. En cada figura se presenta la comparación entre los caudales observados (color azul) y los estimados (color verde) por la validación cruzada. Se discute brevemente el comportamiento individual de cada cuenca, haciendo énfasis en los cambios en el NSE y su posible interpretación.

- **Cuenca 18232**

Esta cuenca muestra un desempeño muy similar entre el modelo original y la validación cruzada (NSE de 0.81 vs. 0.78), lo que indica que los modelos de *machine learning* logran generalizar adecuadamente sus parámetros a partir de los datos de las cuencas vecinas. Esto puede atribuirse tanto a la similitud de sus variables explicativas con las cuencas cercanas como al hecho de pertenecer a la misma región geográfica y climática, lo que implica condiciones fisiográficas comparables. Además, el buen desempeño puede estar relacionado con un periodo de observación más continuo y completo en comparación con otras cuencas (Figura 17).

- **Cuenca 18372**

Se observa una disminución del NSE durante la validación cruzada (0.84 vs. 0.66). Esta diferencia podría explicarse por la presencia de valores de caudal observados relativamente altos que el modelo GR2M no es capaz de estimar en ausencia de la cuenca de entrenamiento (Figura 18). Aunque el modelo mantiene una buena capacidad predictiva, la caída de desempeño podría atribuirse principalmente a la predicción del parámetro x1, el cuál mostró mayor variación respecto a su valor global (Tabla 5). En contraste, la estimación de x2 presenta una menor discrepancia.

- **Cuenca 12347**

Se observa una ligera disminución en el NSE al ser excluida del conjunto de entrenamiento (0.76 contra 0.65). Este resultado se atribuye a un mejor ajuste en el modelo calibrado globalmente, además de sus características particulares (como un NDVI intermedio y QMean bajo) (Ver Tabla A.3) las cuales fueron mejor interpretados en un contexto de generalización. A pesar del ruido presente en los datos observados (Figura 19), el modelo mantiene un desempeño aceptable, lo que sugiere una capacidad de generalización razonable.

- **Cuenca 12561**

Se observa un cambio ligero en el NSE al aplicar la validación cruzada (0.62 contra 0.49). Este resultado podría estar influido por cambios en el caudal observado en el periodo de 2009 al 2015, lo cual afecta negativamente al cálculo del NSE (Figura 20). A pesar de ello, el modelo logró generalizar los parámetros x_1 y x_2 .

- **Cuenca 12466**

Al igual que la 12347, se encuentra geográficamente separada del resto. Presenta una disminución en el NSE (0.64 contra 0.26), posiblemente relacionada con una condición más árida respecto a las otras cuencas (meses de enero a mayo) y que los registros presentan dos eventos que carecen de justificación meteorológica clara, lo que afecta negativamente esta métrica (Figura 21). Además, es la única cuenca donde todas las variables estandarizadas explicativas presentan valores negativos (Ver Tabla A.3). Los modelos de *machine learning* logran cierta capacidad de extrapolación basada en similitudes estructurales más que en cercanía espacial.

- **Cuenca 18540**

Esta cuenca presenta una caída significativa en su desempeño (NSE de 0.6423 a 0.1304). Además de las inconsistencias en los datos observados, el modelo GR2M no logró realizar una estimación correcta adecuada del caudal, lo cual afectó notablemente la eficiencia (Figura 22). Esta cuenca presenta los valores más altos entre todas las variables estandarizadas explicativas (por ejemplo, NDVI, KSat, $RC > 1$) (Ver Tabla A.3) lo que sugiere que los modelos de machine learning tuvieron dificultades para extrapolar en regiones del espacio de variables escasamente representadas en el conjunto de calibración, afectando su capacidad predictiva en este caso particular.

- **Cuenca 12543**

Esta cuenca muestra una disminución notable en el NSE (0.45 contra 0.19). La presencia de valores atípicos, probablemente causados por errores de medición, y la inconsistencia en los datos observados afectaron significativamente la estimación (Figura 23) Esta cuenca también muestra una combinación atípica de variables explicativas: un NDVI bajo junto con valores altos de KSat y RC medianamente altos (Ver Tabla A.3). Esta configuración puede haber dificultado la predicción de parámetros coherentes por parte del modelo, especialmente si no hay cuencas en el conjunto de entrenamiento.

Este análisis evidencia que ambos enfoques experimentan una disminución general en su desempeño al aplicar LOO, lo cual es esperable dada la reducción en el conjunto de entrenamiento. No obstante, la caída es mucho más pronunciada en el caso de la regresión multilineal, evidenciando su baja capacidad de extrapolación. En cambio, el modelo basado en machine learning mantiene valores positivos de NSE en todos los casos, lo cual indica una

mayor capacidad de generalización incluso bajo condiciones de exclusión total de la cuenca objetivo.

Además, la estrategia de validación cruzada LOO no solo permite evaluar la capacidad predictiva de los modelos, sino que también permite simular y asumir un escenario común en hidrología: la estimación de parámetros en una cuenca sin observaciones directas, utilizando únicamente información de otras cuencas similares. Esta aproximación es especialmente útil en casos donde la disponibilidad de datos hidrológicos es limitada. En este caso, las cuencas analizadas, junto con las del Sistema Cutzamala, pertenecen a las mismas regiones hidrológicas y a su vez comparten similitudes climáticas, fisiográficas y geológicas, lo cual favorece la transferencia de información entre ellas. A pesar del reducido número de cuencas disponibles para entrenamiento, los modelos de machine learning demostraron una notable capacidad de generalización, lo que resalta su aplicación en contextos con datos escasos.

Precipitación y Simulación de Caudal para cuenca 18232

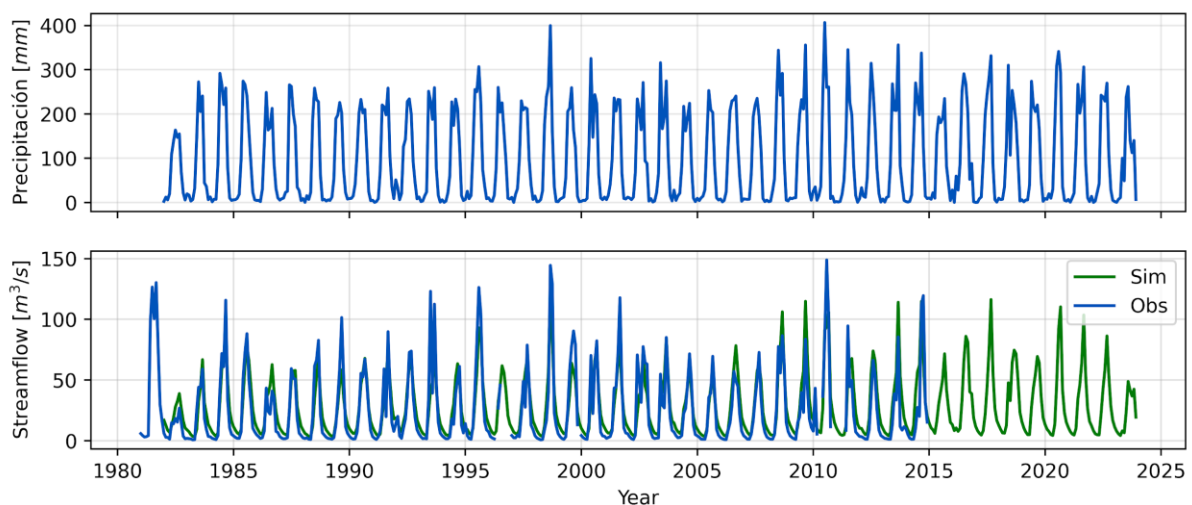


Figura 17. Resultado de la validación cruzada para la cuenca 18232.

Precipitación y Simulación de Caudal para cuenca 18372

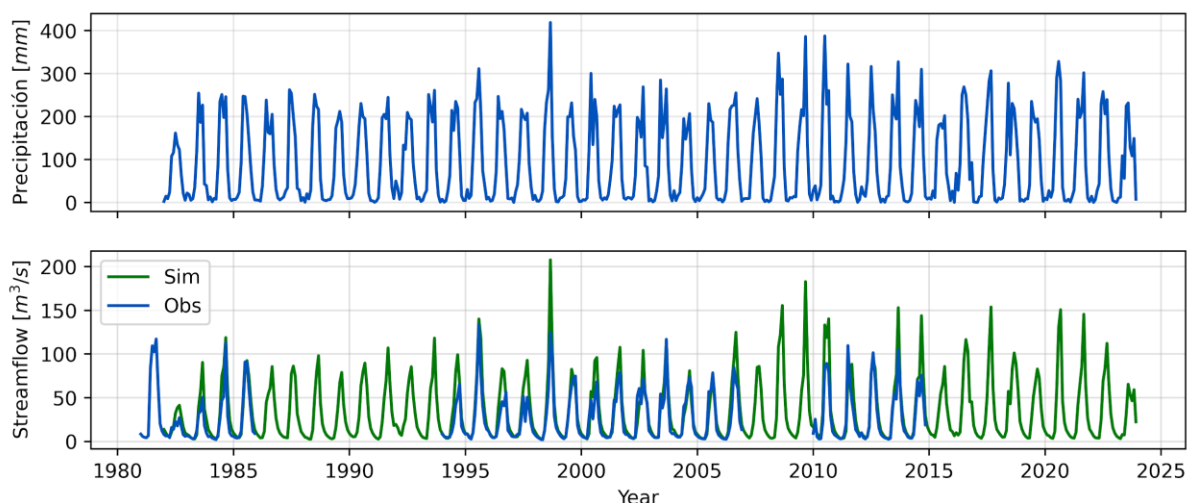


Figura 18. Resultado de la validación cruzada para la cuenca 18372.

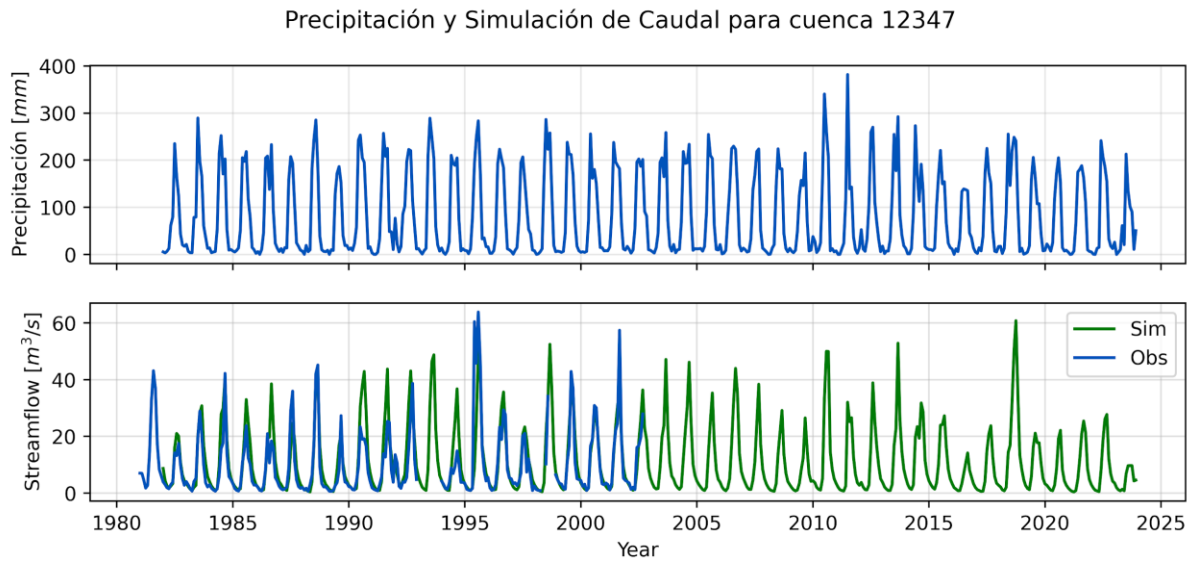


Figura 19. Resultado de la validación cruzada para la cuenca 12347.

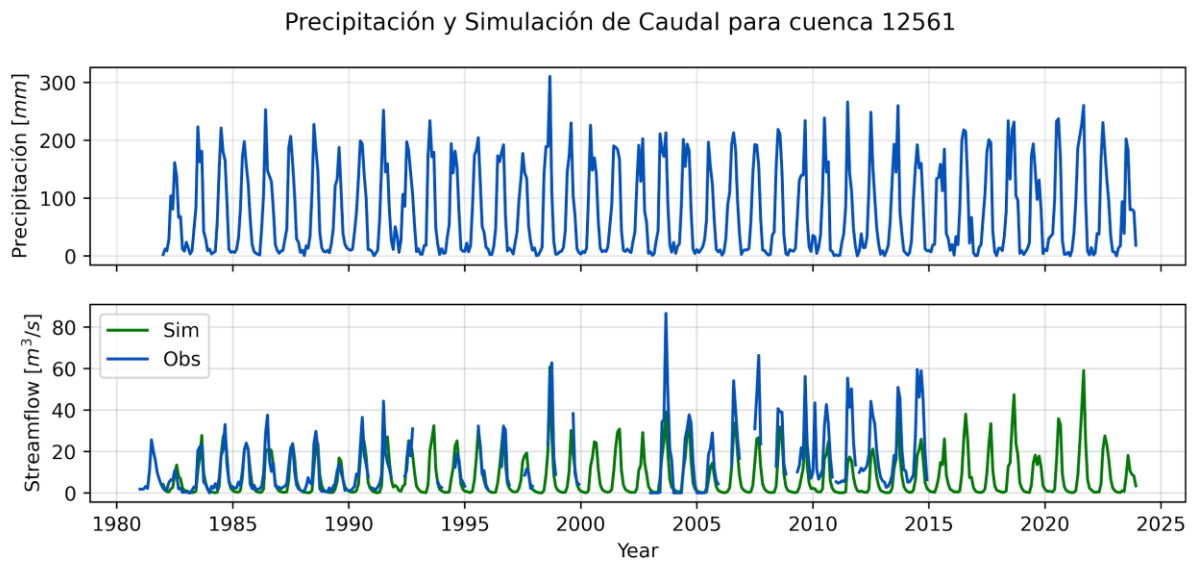


Figura 20. Resultado de la validación cruzada para la cuenca 12561.

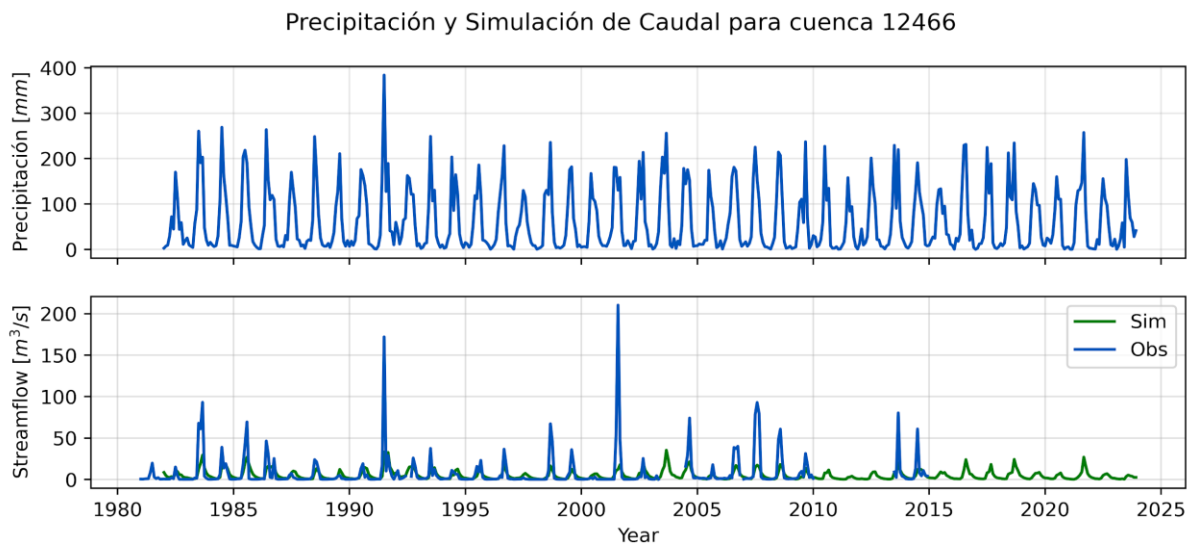


Figura 21. Resultado de la validación cruzada para la cuenca 12466.

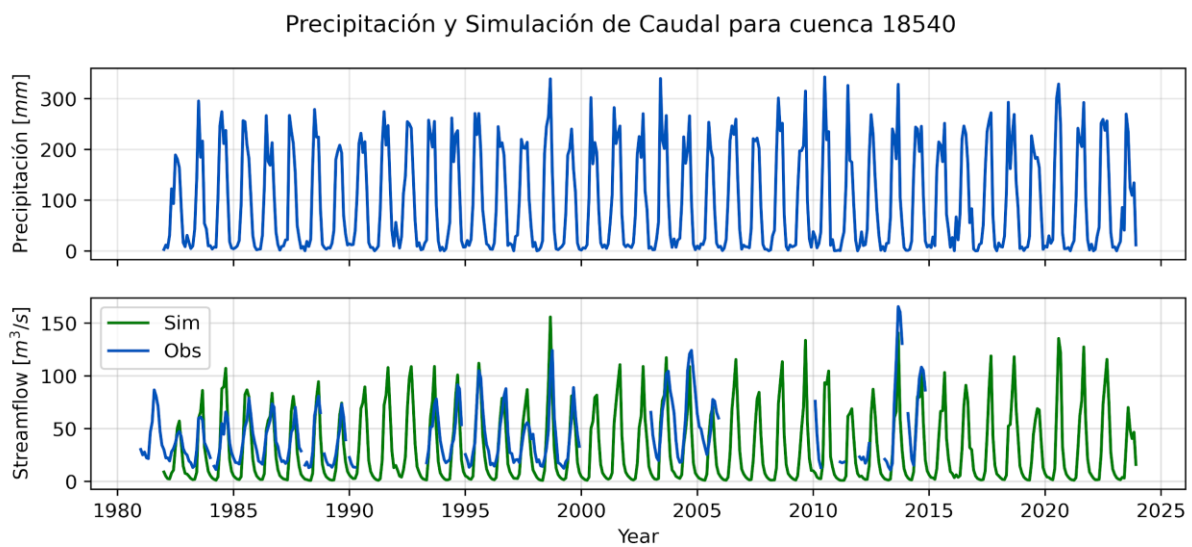


Figura 22. Resultado de la validación cruzada para la cuenca 18540.

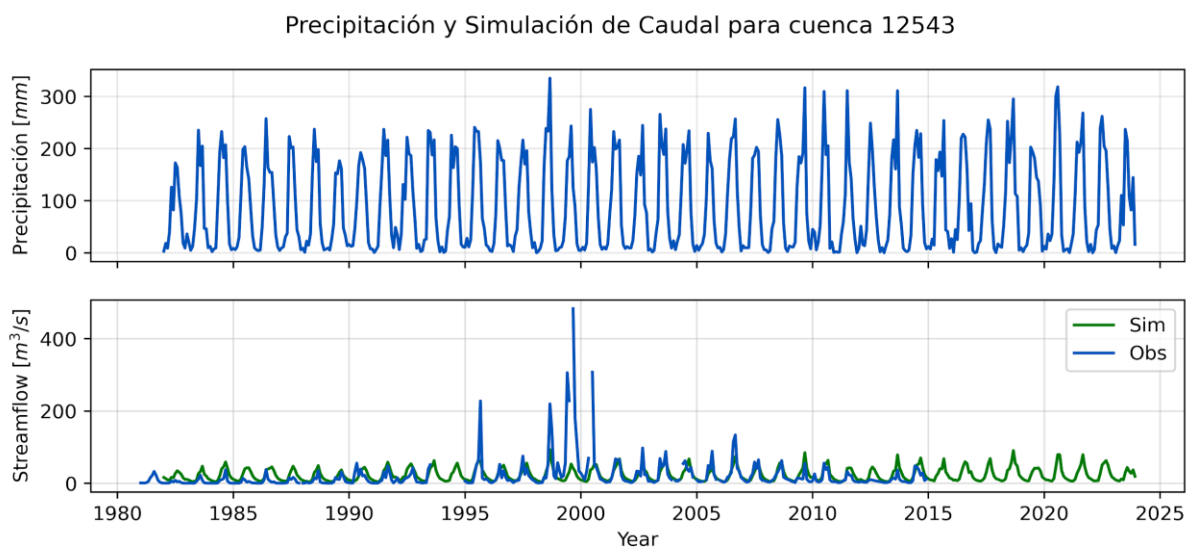


Figura 23. Resultado de la validación cruzada para la cuenca 12543.

F. Simulación del caudal de entrada a las presas del Sistema Cutzamala

Los valores estimados de los parámetros x_1 y x_2 para cada cuenca asociada a las presas del Sistema Cutzamala, obtenidos mediante las tres metodologías evaluadas (Regresión Empírica (RLM), Regresión Espacial y modelos de *Machine Learning*), se presentan en la Tabla 8.

Para la estimación de los parámetros x_1 y x_2 en las presas se siguió un procedimiento coherente con los modelos desarrollados a partir de las cuencas calibradas. El caudal de ingreso a las presas calculado mediante el balance hídrico inverso se empleó como referencia para validar las simulaciones del modelo GR2M. El objetivo fue evaluar la capacidad predictiva de cada metodología cuando se aplica a unidades hidrológicas no incluidas en la calibración inicial, y explorar su desempeño en escenarios donde no existen registros de caudal observados.

En el caso de la Regresión Empírica, se aplicaron directamente los modelos ajustados previamente (ver Tabla 7) a las variables predictoras de cada presa, y los resultados se transformaron de la escala logarítmica a la original mediante la función exponencial.

Para la Regresión Espacial, los valores de x_1 y x_2 se obtuvieron a partir de los mapas generados con esta metodología (Figura 7), extrayendo los valores medios correspondientes a la ubicación de cada presa dentro del entorno SIG (SAGA GIS).

Respecto a los modelos de *Machine Learning*, se emplearon los mismos modelos entrenados con las cuencas calibradas, SVR para x_1 y XGBoost para x_2 , aplicándolos sobre las variables predictoras de las presas. Dado que solo se dispone de tres observaciones (una por presa), no es posible aplicar validación cruzada ni métricas de error. Por ello se utilizaron hiperparámetros conservadores que priorizan la estabilidad y la capacidad de generalización del modelo. En el caso del SVR, las variables fueron estandarizadas antes del entrenamiento y la predicción.

Los hiperparámetros utilizados fueron los mismos que se emplearon para cada modelo de manera individual (ver Tabla 6), con el propósito de mantener coherencia en el proceso de validación cruzada y asegurar una comparación equitativa del desempeño.

La Tabla 9 muestra los valores estimados de x_1 y x_2 para cada presa según las 3 metodologías presentadas.

Tabla 9. Parámetros x_1 y x_2 para las cuencas de las presas del Sistema Cutzamala.

Presa	Regresión Empírica		Regresión Espacial		Machine Learning	
	x_1	x_2	x_1	x_2	x_1	x_2
El Bosque	1187.7409	0.5025	1442.41	0.4964	1627.2604	0.5471

	Regresión Empírica		Regresión Espacial		Machine Learning	
Valle de Bravo	2664.9880	0.6374	2808.53	0.6229	2443.9031	0.6950
Villa Victoria	1295.3090	0.4752	1457.84	0.4715	1400.6171	0.5227

La cuenca de la presa El Bosque muestra las mayores diferencias en la estimación de x_1 , mientras que las diferencias de x_2 entre métodos son más sutiles.

El comportamiento hidrológico observado en las cuencas también se mantiene en las presas analizadas (Figura 24 a 26), lo cual es esperable dado que pertenecen a la misma región climática e hidrológica. Al igual que en las cuencas vecinas, el caudal simulado alcanza su máximo en septiembre mientras que la temperatura media anual presenta su valor más alto entre abril y mayo.

Las figuras correspondientes a cada presa (Figura 24 a 26) mantienen el mismo esquema organizativo empleado en las figuras por cuenca. Cada una presenta cuatro paneles: (1) precipitación acumulada y temperatura media a escala mensual, (2) precipitación y temperatura media mensual, (3) comparación de caudales estimados por tres enfoques (regresión, regresión espacial y machine learning), junto con el caudal derivado del balance hídrico con datos de la presa indicado como caudal observado, en escalas en mm (izquierda) y Hm^3 (derecha), y (4) comparación de las metodologías a escala mensual.

- Presa Villa Victoria

La cuenca presenta un valor promedio de x_1 de 1384.58 mm, atribuido a la presencia de suelos arcillosos (luvisoles) y depósitos lacustres de baja permeabilidad, así como a una intensa presión agrícola que favorece la compactación del suelo y reduce la infiltración. El parámetro x_2 , con un valor de 0.4898 sugiere pérdidas netas de agua, probablemente asociadas con extracciones agrícolas, abastecimiento urbano y drenaje subterráneo fuera del sistema.

El balance hídrico muestra consistencia con el modelo GR2M y con los parámetros estimados indirectamente, reflejando un comportamiento hidrológico estable. No obstante, el balance a partir de los datos de la presa presenta valores elevados de caudal (hasta 60 mm), explicables por eventos máximos de precipitación mensual (~200 mm). En contraste, las simulaciones con Machine Learning alcanzan un máximo de 25 mm (~15 Hm^3), mientras que las regresiones lineal y espacial estiman valores cercanos a ~20 mm (~12.5 Hm^3), con diferencias mínimas entre ambas metodologías (Figura 24).

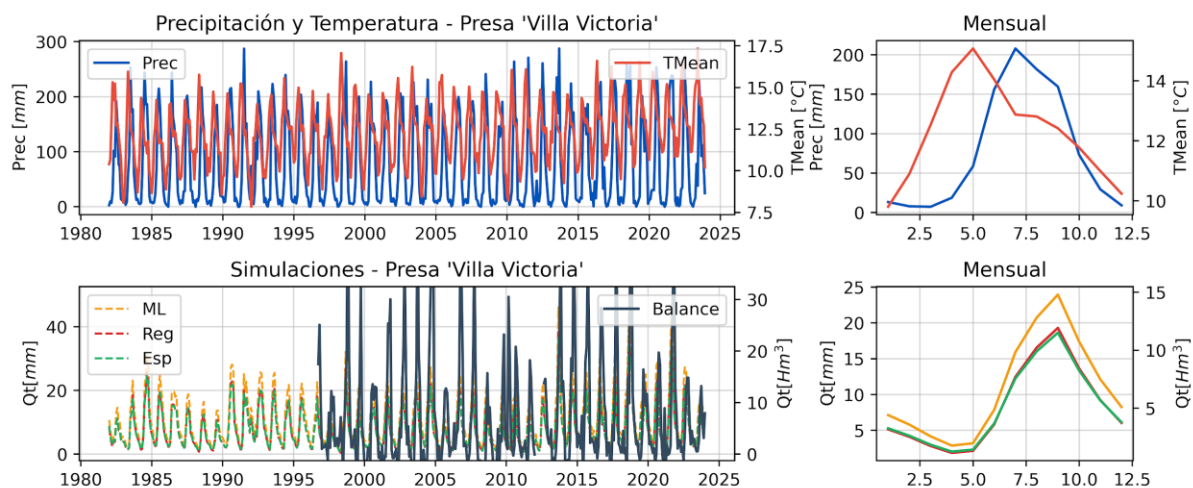


Figura 24. Simulación y Balance en la Presa “Villa Victoria”.

- Presa Valle de Bravo

Considerada el corazón del Sistema Cutzamala, esta presa posee la mayor capacidad de almacenamiento (393.07 millones de m^3), aportando entre el 45 y 50% del caudal del sistema. Su valor del parámetro $x1$ de ~ 2500 mm, indica una elevada capacidad de almacenamiento en la zona no saturada, asociada a la presencia de andosoles volcánicos (Figura 3), suelos altamente porosos que permiten una recarga gradual, comportamiento posiblemente interpretado por el modelo GR2M como flujo subterráneo lento. La densa cobertura y un ecosistema boscoso bien conservado contribuyen a reducir la evapotranspiración y mantener la humedad.

El parámetro $x2$, con un valor promedio de 0.6517, lo cual sugiere menos pérdidas hídricas en comparación con las otras presas, favorecido por la geología (basaltos fracturados y andosoles) y una menor presión por extracción subterránea. La simulación del modelo GR2M y el balance hídrico presentan buena concordancia entre metodologías. Aunque se obtuvieron caudales observados elevados (~ 60 mm), estos podrían estar afectados por errores en las mediciones para estimar el balance. La precipitación mensual máxima (~ 225 mm) es la más alta entre las presas. Los métodos de Machine Learning simulan un caudal máximo de ~ 45 mm ($\sim 27.5 Hm^3$), mientras que las regresiones estiman un máximo de ~ 25 mm ($\sim 23.5 Hm^3$) (Figura 25).

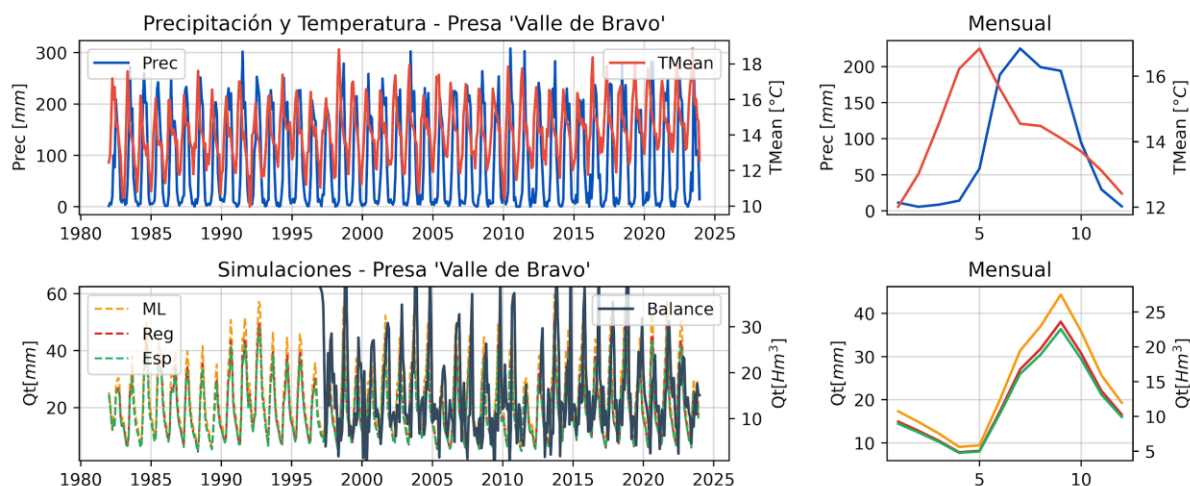


Figura 25. Simulación y Balance en la Presa “Valle de Bravo”.

- Presa El Bosque

Esta cuenca presenta un valor promedio de $x1$ de 1419.14 mm, relacionado con la presencia de leptosoles sobre rocas volcánicas fracturadas (Figura 3), que permiten cierto almacenamiento en fracturas que favorecen un drenaje rápido. Las pendientes pronunciadas incrementan la escorrentía superficial, reduciendo el tiempo de retención hídrica. El parámetro $x2$ (~ 0.5153) indica una mayor contribución subterránea a la presa, apoyada por la baja densidad poblacional, que reduce la presión sobre los acuíferos y favorece la recarga.

A diferencia de las otras presas, el balance hídrico y las simulaciones del modelo no presentan una concordancia clara. Esto puede explicarse por filtraciones documentadas en la infraestructura de la presa (World Bank Group, 2015), lo que altera los resultados del balance. Por otro lado, los datos derivados del balance de agua con datos de la presa no concuerdan con la capacidad de la presa y con la variación del almacenamiento comparado con la presa Valle de Bravo, lo cual refuerza la idea de que existen errores en los datos reportados para la presa. El modelo GR2M, con parámetros derivados de la estimación con Machine Learning, estima un caudal de ~ 20 mm ($\sim 8 \text{ Hm}^3$), mientras que las regresiones arrojan valores ligeramente menores (~ 17.5 mm o $\sim 7.5 \text{ Hm}^3$). El Bosque presenta los valores de temperatura más altos entre las presas ($\sim 18^\circ\text{C}$ promedio anual) y una precipitación media mensual ligeramente menor (~ 200 mm) (Figura 26).

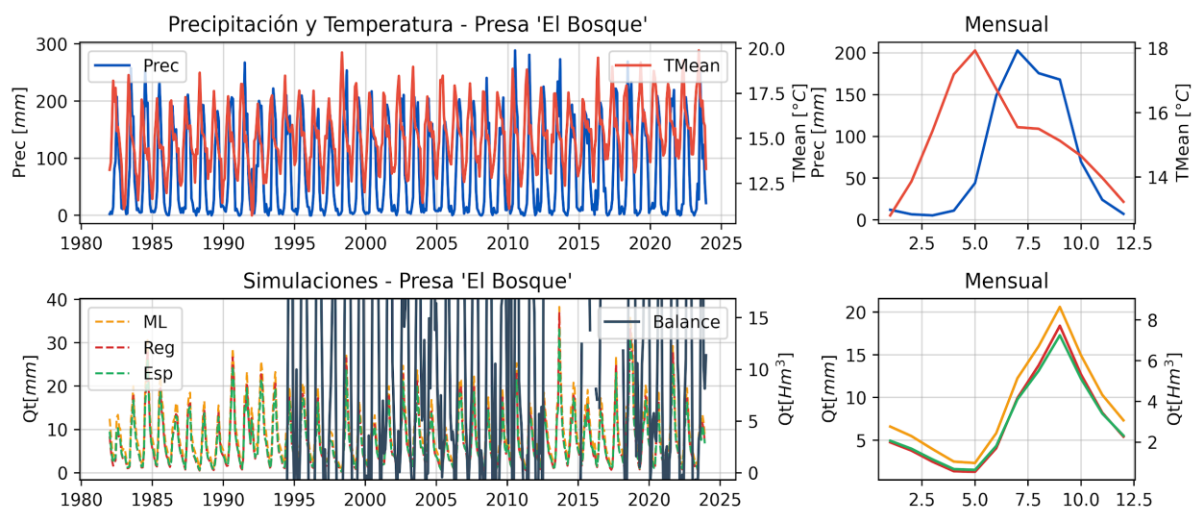


Figura 26. Simulación y Balance en la Presa “El Bosque”.

La Figura 27 presenta la comparación anual de la precipitación anual y los caudales simulados para las tres presas principales del Sistema Cutzamala.

Valle de Bravo registra consistentemente los valores más altos de precipitación, superando los 100 mm en años como 2010 y 2022. En cambio, Villa Victoria y El Bosque muestran valores similares, generalmente inferiores a 1100 mm anuales, con disminuciones marcadas en 2005 y 2011; este último asociado a una de las sequías más severas en siete décadas (Domínguez, 2016). Estas diferencias reflejan no solo la localización y capacidad de almacenamiento (Tabla 3), sino también condiciones orográficas y microclimáticas particulares.

En cuanto al caudal simulado:

- Valle de Bravo mantiene los valores más altos alcanzando hasta $\sim 250 \text{ Hm}^3$ con el enfoque de Machine Learning, $\sim 200 \text{ Hm}^3$ con la regresión lineal y ligeramente menor con la regresión espacial. Incluso en años secos, mantiene una respuesta hidrológica más estable, atribuida a su alta capacidad de almacenamiento.
- Villa Victoria muestra caudales intermedios, con mayor oscilación anual. Las tres metodologías se comportan de forma similar, siendo la regresión espacial la que estima los valores más bajos.
- El Bosque registra los caudales más bajos ($\sim 50 \text{ Hm}^3$), pese a que la precipitación es comparable con Villa Victoria. Las diferencias metodológicas son sutiles, aunque Machine Learning tiende a producir estimaciones ligeramente mayores.

El año 2017 destaca por un incremento simultáneo en el caudal de las tres presas, en concordancia con un aumento notable en la precipitación: Valle de Bravo alcanza entre 175 y 200 Hm^3 ; Villa Victoria $\sim 150 \text{ Hm}^3$, y El Bosque $\sim 90 \text{ Hm}^3$, respectivamente. Por el contrario,

2005 refleja un año críticamente seco, con caudales mínimos estimados, coherentes con las condiciones climáticas de ese periodo.

A partir de este análisis se identifica que, pese a que del 2020 al 2023 se presentaron condiciones adversas en las presas, las entradas de agua a las presas se han recuperado para el 2024. Las condiciones más adversas, se observan en los años de 1988, 1997, 2005.

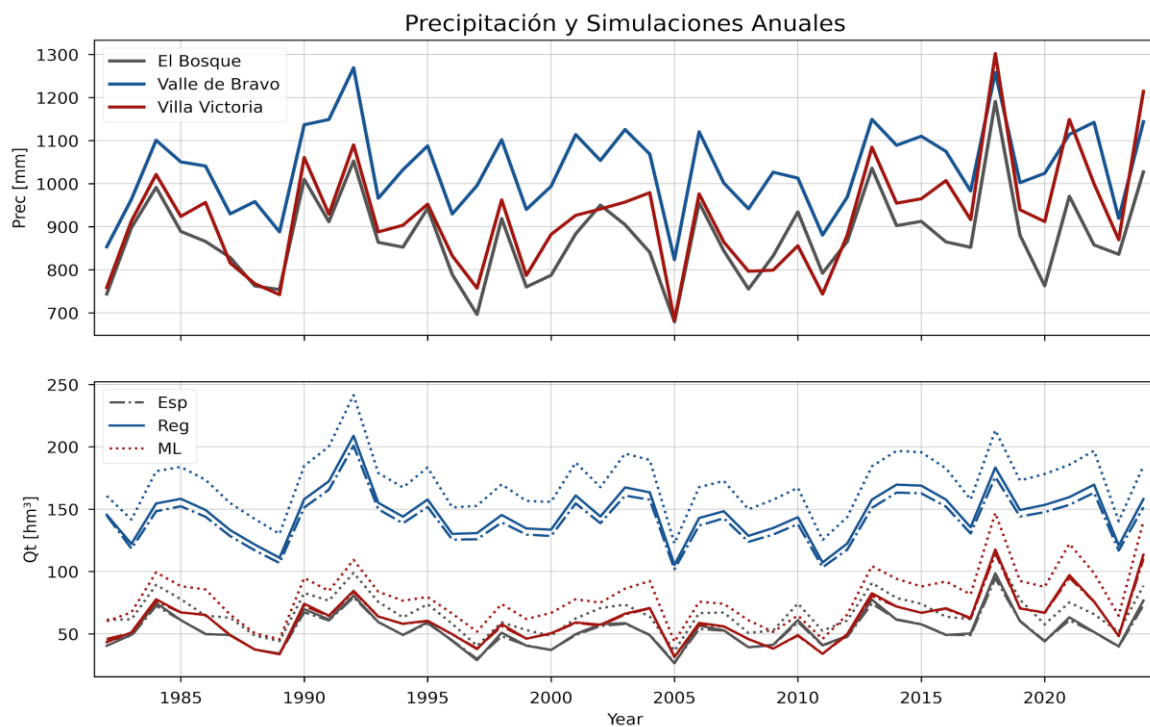


Figura 27. Precipitación y simulaciones de caudal en escala anual para cada una de las presas del Sistema Cutzamala.

VII. Conclusiones y trabajo futuro

A. Conclusiones

Esta investigación logró cumplir con el objetivo general de estimar los volúmenes de entrada a las presas principales del Sistema Cutzamala, Villa Victoria, Valle de Bravo y El Bosque, mediante la aplicación del modelo hidrológico conceptual GR2M, la integración de datos derivados de percepción remota y la implementación de técnicas de regionalización basadas en algoritmos de Machine Learning.

Entre los principales hallazgos y aportes se destacan:

1. Uso de datos globales y percepción remota:

Se comprobó la utilidad de los datos del producto FLDAS para representar adecuadamente la dinámica climática de la región de estudio, tanto en precipitación como en temperatura. A pesar de la incertidumbre inherente, derivada de factores como la topografía compleja, la variabilidad climática regional y la escasa densidad de estaciones meteorológicas (una sola estación meteorológica en Valle de Bravo), estos datos permitieron alimentar de manera adecuada al modelo GR2M en ausencia de información local detallada.

Además, el uso de productos de percepción remota como el NDVI del producto MODIS-13 y la conductividad hidráulica saturada del suelo del producto CoGTF facilitó la caracterización indirecta de variables biofísicas clave, como el tipo de cobertura vegetal y la capacidad de infiltración del suelo, fundamentales para la estimación de los parámetros del modelo.

2. Caracterización geomorfológica e hidrogeológica:

La caracterización geomorfológica de las cuencas mediante software especializado permitió evidenciar la relación entre las condiciones físicas del terreno y los parámetros hidrológicos del modelo. Factores como tipo de suelo, pendiente, cobertura vegetal y presión antrópica influyen directamente en la capacidad de almacenamiento ($x1$) y las pérdidas o ganancias subterráneas ($x2$). Esta comprensión espacial fue clave para sustentar la regionalización de parámetros y entender los contrastes entre presas.

3. Modelación hidrológica confiable con GR2M:

La calibración y validación del modelo GR2M en cuencas vecinas de las presas mostró un buen desempeño, con valores de eficiencia altos, como en las cuencas: 18372 y 18232 cuyos coeficientes NSE alcanzaron 0.84 y 0.81 respectivamente. No obstante, también se identificaron cuencas con menor ajuste como la 12561 y la 12543, atribuibles en gran medida a inconsistencias en los datos observados, incluyendo valores atípicos no justificables a partir de los datos de precipitación disponibles en FLDAS. Esto sugiere que el bajo desempeño en estos casos no necesariamente refleja un fallo en la simulación del modelo, sino limitaciones en la calidad de los datos de referencia utilizados para la validación.

En conjunto, la simplicidad del modelo GR2M, su base conceptual y su capacidad para representar procesos de almacenamiento y flujo subterráneo, lo convierten en una herramienta eficaz para contextos donde la disponibilidad de datos es limitada.

4. Regionalización de parámetros mediante técnicas comparativas:

Se exploraron y compararon diversas metodologías para estimar los parámetros x_1 y x_2 en cuencas sin datos observados, incluyendo regresión múltiple, regresión espacial y algoritmos de Machine Learning como SVR y XGBoost. Aunque estas últimas técnicas presentaron un mejor desempeño predictivo que los métodos estadísticos tradicionales, ninguno superó el ajuste ni la robustez alcanzados mediante la optimización individual con el algoritmo Nelder-Mead.

Esto se explica por la diferencia fundamental en el objetivo de cada método. Los algoritmos de optimización ajustan directamente los parámetros del modelo GR2M para minimizar la diferencia entre el caudal simulado y observado, actuando sobre la función objetivo definida en la calibración. Por el contrario, los algoritmos de *machine learning* (SVR y XGBoost) no se utilizan para predecir caudales, sino para resolver el problema de regionalización. Su propósito es establecer las relaciones estadísticas entre los parámetros previamente calibrados del modelo (x_1 y x_2) y un conjunto de atributos fisiográficos y climáticos de las cuencas. De esta manera, el modelo mantiene su rol central en la simulación hidrológica, mientras que los algoritmos amplían su aplicabilidad a cuencas sin datos, permitiendo la transferencia de parámetros en cuencas con características similares.

El desempeño de los modelos de *machine learning* en esta tarea depende críticamente de la calidad, distribución y cobertura de los datos de entrenamiento, así como de un ajuste fino de sus hiperparámetros. Si bien el modelo XGBoost, aplicado a la estimación del parámetro x_2 , produjo resultados muy similares e incluso idénticos a los de la optimización directa en algunas cuencas, este éxito no radica en que el algoritmo haya aprendido la física hidrológica subyacente. En cambio, indica su capacidad para identificar y replicar patrones estadísticos robustos presentes en los datos de entrenamiento, que vinculan las características de las cuencas con los valores óptimos de los parámetros.

Es importante destacar dicha distinción, ya que la fortaleza de dichos algoritmos reside en su poder de patrones complejos para la regionalización, no en la interpretación de procesos físicos. Esta limitación puede hacerse más evidente en contextos con conjuntos de datos reducidos, donde el riesgo de sobreajuste es mayor y puede generar una falsa sensación de robustez en la generalización a condiciones no vistas.

5. Comparación con el balance hídrico:

Las simulaciones realizadas mediante GR2M, utilizando parámetros regionalizados, presentaron una buena correspondencia con las estimaciones de volumen derivadas del balance hídrico en las presas. Si bien se observan algunas discrepancias, atribuidas a errores

en los datos para estimar el balance o a filtraciones no contabilizadas, como en el caso de la presa El Bosque, las tendencias generales fueron coherentes, lo que valida la aplicabilidad del modelo y la estrategia de regionalización empleada.

6. Aplicabilidad y escalabilidad del enfoque:

Este trabajo demuestra que es posible generar estimaciones razonables de caudal en regiones sin monitoreo hidrológico directo, mediante una combinación de modelos hidrológicos simples, productos de percepción remota y técnicas de análisis de datos. La metodología desarrollada puede ser fácilmente replicable y escalable a otras regiones del país o del mundo con problemáticas similares de monitoreo limitado.

Además, se resalta que modelos conceptuales como GR2M pueden representar una alternativa útil frente a enfoques de Machine Learning o Deep Learning, especialmente en contextos como el de este estudio, donde se cuenta con un número limitado de observaciones para entrenar los modelos. A diferencia de los algoritmos de aprendizaje computacional, que requieren una base de entrenamiento amplia para capturar patrones complejos, GR2M permite obtener resultados interpretables, robustos y vinculados a procesos físicos conocidos, incluso con datos limitados.

B. Trabajos futuros y líneas de investigación

A partir de los resultados obtenidos en este trabajo, se identifican varias líneas de investigación que podrían ampliarse o profundizarse en estudios futuros:

1. Modelado hidrológico con redes neuronales y enfoques híbridos

Explorar el uso de algoritmos de *deep learning*, en particular redes LSTM (Long Short-Term Memory) permitiría modelar el comportamiento temporal del caudal a partir de series históricas de precipitación y temperatura. Las LSTM son especialmente adecuadas para series de tiempo hidrológicas por su capacidad de retener información a largo plazo y capturar relaciones no lineales. Este enfoque podría aplicarse a cuencas con escasa información para predecir el comportamiento hidrológico a partir de patrones aprendidos en otras regiones.

Asimismo, se sugiere investigar modelos híbridos, que integren componentes físicos con machine learning, como las Redes Neuronales Informadas por Física (PINNs). Estos modelos tienen el potencial de respetar principios hidrológicos fundamentales al mismo tiempo que aprenden de datos observados, lo cual podría mejorar la capacidad de generalización y la interpretabilidad de las predicciones.

2. Análisis de incertidumbre mediante enfoques bayesianos

Dado que los datos empleados provienen de productos globales (como FLDAS o CoGTF), es relevante incorporar métodos bayesianos para estimar la incertidumbre asociada a dichos insumos y su propagación hacia los parámetros calibrados del modelo (x_1 y x_2). Métodos como MCMC (*Markov Chain Monte Carlo*) o DREAM (*Differential Evolution Adaptive Metropolis*) podrían utilizarse para generar distribuciones a posteriori de los parámetros

permitiendo no solo obtener valores óptimos, sino también intervalos de confianza, lo cual sería particularmente útil para evaluar la confiabilidad del modelo en escenarios de escasez o exceso hídrico.

Este análisis permitiría, además, cuantificar cómo errores sistemáticos o sesgos en datos satelitales (como FLDAS o KSat) pueden amplificarse en la estimación del caudal, lo cual resulta esencial en contextos de planificación hídrica.

3. Implementación de sistemas de monitoreo y pronóstico en tiempo real

A partir del modelo calibrado y validado en esta tesis, es posible proponer el desarrollo de una red de monitoreo en tiempo real para el Sistema Cutzamala, que integre datos provenientes de productos satelitales como CHIRPS, GPM IMERG y estaciones meteorológicas automáticas. Estos datos podrían alimentar mensualmente el modelo GR2M, ya calibrado, para generar pronósticos de caudal y apoyar la toma de decisiones en tiempo real.

Este sistema permitiría:

- Monitorear la disponibilidad hídrica de manera continua.
- Prevenir situaciones críticas como escasez de agua en temporadas secas.
- Detectar anomalías como eventos de precipitación extrema o posibles fallas en la infraestructura.
- Fortalecer la toma de decisiones en materia de política pública y gestión de riesgos.

4. Ampliación y fusión de fuentes de datos climáticos

Para reducir la incertidumbre en las entradas del modelo GR2M, una estrategia consiste en ampliar la base de datos climática incorporando múltiples productos satelitales o reanálisis. La combinación de fuentes como FLDAS, ERA5, CHIRPS, IMERG, MERRA-2 o estaciones meteorológicas podría permitir una estimación más robusta de las variables requeridas (precipitación y temperatura media mensual), mediante técnicas de fusión de datos o ponderación por calidad.

Esto no solo ayudaría a minimizar los errores en la entrada del modelo, sino que también podría mejorar la estimación de los parámetros hidrológicos al disponer de series más completas y precisas.

5. Integración de bases de datos hidrológicas extensas para mejorar la regionalización de parámetros

Otra línea de investigación relevante consiste en emplear bases de datos hidrológicas de gran escala, como CAMELS (*Catchment Attributes and Meteorology for Large-sample Studies*), para mejorar la estimación y regionalización de los parámetros x_1 y x_2 del modelo GR2M.

Estas bases de datos contienen información estandarizada de cientos de cuencas con atributos climáticos, topográficos, geológicos y de uso de suelo, lo que permitiría entrenar los modelos de *machine learning* en conjuntos de datos más amplios y representativos.

Además, sería posible restringir el entrenamiento a cuencas con condiciones fisiográficas y climáticas similares, fortaleciendo la capacidad del modelo para generalizar de manera más coherente en regiones con características afines.

La integración de CAMELS en la metodología propuesta abriría la puerta a estudios comparativos entre distintas regiones y al desarrollo de modelos regionales de transferencia hidrológica con mayor robustez y aplicabilidad.

6. Evaluación de escalabilidad del modelo en nuevas regiones

Finalmente, una línea de trabajo relevante sería aplicar la metodología desarrollada en esta tesis a otras regiones hidrográficas de México o América Latina, particularmente aquellas con condiciones similares de escasez de datos o presión hídrica. Esto permitiría evaluar la portabilidad del modelo GR2M calibrado con datos remotos y las metodologías de regionalización propuestas, así como comparar su desempeño en diferentes contextos fisiográficos, climáticos y de uso de suelo.

VIII. Anexo

Tabla A.1. Variables físicas para cada una de las cuencas vecinas al Sistema Cutzamala

Variable/ Source	18232	18540	12347	12561	12543	18372	12466
KSat [m/s]	2.04	2.14	2.02	1.92	2.09	2.06	1.84
Sand [%]	36.53	37.83	34.93	37.42	38.14	36.12	38.48
Clay [%]	40.83	35.54	39.45	37.62	34.93	38.73	41.87
Silt [%]	22.66	26.64	25.62	24.98	26.93	25.16	19.62
Tair [°C]	17.45	12.97	15.16	12.47	10.36	15.14	14.35
Wind [m/s]	2.95	2.98	3.19	3.19	2.87	2.98	4.27
Evap [kg/m ² s]	838.31	925.94	845.51	760.42	863.77	785.75	692.88
Qair [kg/ kg]	0.0084	0.0072	0.0080	0.0072	0.0064	0.0077	0.0076
Qt [kg/m ² s]	359.03	228.72	116.73	116.76	228.45	332.24	3.31
Rain [kg/ m ² s]	1197.06	1153.42	959.84	877.50	1093.48	1118.04	726.59
SM [m ³ /m ³]	746.48	701.59	65.26	671.84	712.12	758.75	59.76
NDVI [–]	0.6051	0.6636	0.5888	0.4890	0.5154	0.5933	0.4493
BFI [–]	0.6224	0.6279	0.6196	0.6763	0.6745	0.6607	0.6401
K [1/day]	0.0504	0.4063	0.5987	0.6274	0.5315	0.4675	0.7121
QMEAN [mm/year]	461.18	625.76	208.80	275.67	368.81	465.74	183.30
RC [–]	0.4246	0.5838	0.2803	0.3057	0.3724	0.4207	0.2387

Tabla A.2. Variables físicas para cada una de las presas del Sistema Cutzamala

Variable/ Source	El Bosque	Valle de Bravo	Villa Victoria
KSat [m/s]	1.182	1.258	0.4387
Sand [%]	-3.794	-0.6845	-6.102
Clay [%]	-1.17	-1.40	-0.8929
Silt [%]	1.423	1.370	1.364
Tair [°C]	-0.5867	-0.8149	-1.25
Wind [m/s]	-0.7590	-0.7737	-0.3991
Evap [kg/m ² s]	0.5170	1.29	0.9565
Qair [kg/kg]	-0.5477	-0.7368	-1.07
Qt [kg/m ² s]	-0.2836	0.2699	-0.3248
Rain [kg/m ² s]	0.1471	0.1067	0.4423
SM [m ³ /m ³]	0.1001	0.1588	0.0559
NDVI [–]	1.124	0.9281	-0.1774
BFI [–]	0.3739	0.3705	1.06
K [1/day]	-0.7682	-1.004	-0.5788
QMEAN [mm/year]	0.4736	1.454	0.0891
RC [–]	1.030	1.684	0.4996

Tabla A.3. Variables físicas estandarizadas utilizadas para la predicción de los modelos para cada una de las cuencas vecinas al Sistema Cutzamala

Variable/ Source	18232	18540	12347	12561	12543	18372	12466
Evap [<i>kg/m²s</i>]	0.93	1.57	0.98	0.35	1.12	0.54	-0.13
KSat [<i>m/s</i>]	0.94	1.70	0.77	0.07	1.32	1.09	-0.49
NDVI [–]	0.72	1.28	0.56	-0.39	-0.13	0.60	-0.77
QMean [<i>mm/year</i>]	1.27	2.34	-0.36	0.07	0.67	1.30	-0.52
RC [–]	0.97	2.20	-0.14	0.55	0.57	0.94	-0.46
K [<i>1/day</i>]	-0.90	-1.18	-0.64	-0.56	-0.83	-1.00	-0.32

Tabla A.4. Variables físicas estandarizadas utilizadas para la predicción de los modelos para cada una de las cuencas del Sistema Cutzamala

Variable/ Source	El Bosque	Valle de Bravo	Villa Victoria
Evap [<i>kg/m²s</i>]	0.51	1.29	0.95
KSat [<i>m/s</i>]	1.18	1.25	0.43
NDVI [–]	1.12	0.92	-0.17
QMean [<i>mm/year</i>]	0.47	1.45	0.08
RC [–]	1.03	1.68	0.49
K [<i>1/day</i>]	-0.76	-1.00	-0.57

IX. Referencias

- Abbott, M. B., Bathurst, J. C., Cunge, J. A., O'Connell, P. E., & Rasmussen, J. (1986). An introduction to the European Hydrological System — Systeme Hydrologique Europeen, "SHE", 1: History and philosophy of a physically-based, distributed modelling system. *Journal of Hydrology*, 87(1), 45–59. [https://doi.org/https://doi.org/10.1016/0022-1694\(86\)90114-9](https://doi.org/https://doi.org/10.1016/0022-1694(86)90114-9)
- Alebachew, M. A., Ye, S., Li, H. yi, Huang, M., Leung, L. R., Fiori, A., & Sivapalan, M. (2014). Regionalization of subsurface stormflow parameters of hydrologic models: Up-scaling from physically based numerical simulations at hillslope scale. *Journal of Hydrology*, 519(PA), 683–698. <https://doi.org/10.1016/j.jhydrol.2014.07.018>
- Ali, A., Sanaei, A., Li, M., Nalivan, O. A., Ahmadaali, K., Pour, M. J., Valipour, A., Karami, J., Aminpour, M., Kaboli, H., & Askari, Y. (2020). Impacts of climatic and edaphic factors on the diversity, structure and biomass of species-poor and structurally-complex forests. *Science of The Total Environment*, 706, 135719. <https://doi.org/https://doi.org/10.1016/j.scitotenv.2019.135719>
- Alley, W. M. (1985). Water Balance Models in One-Month-Ahead Streamflow Forecasting. *Water Resources Research*, 21(4), 597–606. <https://doi.org/10.1029/WR021i004p00597>
- Arif Ali, Z., H. Abduljabbar, Z., A. Tahir, H., Bibo Sallow, A., & Almufti, S. M. (2023). eXtreme Gradient Boosting Algorithm with Machine Learning: a Review. *Academic Journal of Nawroz University*, 12(2), 320–334. <https://doi.org/10.25007/ajnu.v12n2a1612>
- Arnell, N. W. (1992). Factors controlling the effects of climate change on river flow regimes in a humid temperate environment. *Journal of Hydrology*, 132(1–4), 321–342. [https://doi.org/10.1016/0022-1694\(92\)90184-W](https://doi.org/10.1016/0022-1694(92)90184-W)
- Awad, M., & Khanna, R. (2015). Support Vector Regression. In M. Awad & R. Khanna (Eds.), *Efficient Learning Machines: Theories, Concepts, and Applications for Engineers and System Designers* (pp. 67–80). Apress. https://doi.org/10.1007/978-1-4302-5990-9_4
- Bamler, R. (1999). The SRTM mission: A world-wide 30m resolution DEM from SAR interferometry in 11 days. Photogrammetric Week.

Beck, H. E., Roo, D. E., & van Dijk, A. I. J. M. (2015). *Global Maps of Streamflow Characteristics Based on Observations from Several Thousand Catchments**. <https://doi.org/10.1175/JHM-D-14-0155.s1>

Beck, H. E., Pan, M., Lin, P., Seibert, J., van Dijk, A. I. J. M., & Wood, E. F. (2020). Global Fully Distributed Parameter Regionalization Based on Observed Streamflow From 4,229 Headwater Catchments. *Journal of Geophysical Research: Atmospheres*, 125(17). <https://doi.org/10.1029/2019JD031485>

Becker, M. W. (2006). Potential for satellite remote sensing of ground water. In *Ground Water* (Vol. 44, Issue 2, pp. 306–318). <https://doi.org/10.1111/j.1745-6584.2005.00123.x>

Benesty, J., Chen, J., Huang, Y., & Cohen, I. (2009). Pearson correlation coefficient. In *Springer Topics in Signal Processing* (Vol. 2, pp. 1–4). Springer Science and Business Media B.V. https://doi.org/10.1007/978-3-642-00296-0_5

Beven, K. (2006). A manifesto for the equifinality thesis. *Journal of Hydrology*, 320(1–2), 18–36. <https://doi.org/10.1016/J.JHYDROL.2005.07.007>

Beven, K., & Binley, A. (2014). GLUE: 20 years on. *Hydrological Processes*, 28(24), 5897–5918. <https://doi.org/10.1002/hyp.10082>

Beven, K., Page, T., Smith, P., Kretschmar, A., Hankin, B., & Chappell, N. (2024). UPH Problem 20 – reducing uncertainty in model prediction: a model invalidation approach based on a Turing-like test. *Proceedings of IAHS*, 385, 129–134. <https://doi.org/10.5194/piahs-385-129-2024>

Boulange, J., Hanasaki, N., Yamazaki, D., & Pokhrel, Y. (2021). Role of dams in reducing global flood exposure under climate change. *Nature Communications*, 12(1), 417. <https://doi.org/10.1038/s41467-020-20704-0>

Chen, T., & Guestrin, C. (2016). XGBoost: A Scalable Tree Boosting System. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 785–794. <https://doi.org/10.1145/2939672.2939785>

CONAGUA. (2024). *Actualización de la Disponibilidad Media Anual de Agua en el Acuífero Ciudad Hidalgo (1610), Estado de Michoacán*. CONAGUA. https://sigagis.conagua.gob.mx/gas1/sections/Disponibilidad_Acuiferos.html

CONAGUA. (2024). *Actualización de la Disponibilidad Media Anual de Agua en el Acuífero Huetamo (1612), Estado de Michoacán*. CONAGUA. https://sigagis.conagua.gob.mx/gas1/sections/Disponibilidad_Acuiferos.html

CONAGUA. (2024). *Actualización de la Disponibilidad Media Anual de Agua en el Acuífero Zacatepec (1703), Estado de Morelos*. CONAGUA. https://sigagis.conagua.gob.mx/gas1/sections/Disponibilidad_Acuiferos.html

Dal Molin, M., Kavetski, D., & Fenicia, F. (2021). SuperflexPy 1.3.0: an open-source Python framework for building, testing, and improving conceptual hydrological models. *Geoscientific Model Development*, 14(11), 7047–7072. <https://doi.org/10.5194/gmd-14-7047-2021>

Devia, G. K., Ganasri, B. P., & Dwarakish, G. S. (2015). A Review on Hydrological Models. *Aquatic Procedia*, 4, 1001–1007. <https://doi.org/https://doi.org/10.1016/j.agpro.2015.02.126>

Di Bucchianico, A. (2008). Coefficient of Determination (2). In *Encyclopedia of Statistics in Quality and Reliability*. John Wiley & Sons, Ltd. <https://doi.org/https://doi.org/10.1002/9780470061572.eqr173>

Didan, K.. *MOD13Q1 MODIS/Terra Vegetation Indices 16-Day L3 Global 250m SIN Grid V006*. 2015, distributed by NASA EOSDIS Land Processes Distributed Active Archive Center, <https://doi.org/10.5067/MODIS/MOD13Q1.006>. Accessed 2025-04-28.

Dietterich, T. (1995). *Overfitting and Undercomputing in Machine Learning*. ACM Computing Surveys, 27(3), 326-337

Domínguez, Judith. (2016). Revisión histórica de las sequías en México: de la explicación divina a la incorporación de la ciencia. *Tecnología y ciencias del agua*, 7(5), 77-93. http://www.scielo.org.mx/scielo.php?script=sci_arttext&pid=S2007-24222016000500077&lng=es&tlng=es

Gayathri K. Devia, B.P. Ganasri, G.S. Dwarakish (2015). A Review on Hydrological Models, Aquatic Procedia, Volume 4, Pages 1001-1007, ISSN 2214-241X, <https://doi.org/10.1016/j.aqpro.2015.02.126>

Gogtay, N. J., & Thatte, U. M. (2017). Principles of Correlation Analysis. In *Journal of The Association of Physicians of India* ■ (Vol. 65).

González-Ortigoza, S., Hernández-Espriú, A., & Arciniega-Esparza, S. (2023). Regional modeling of groundwater recharge in the Basin of Mexico: new insights from satellite observations and global data sources. *Hydrogeology Journal*, 31(7), 1971–1990. <https://doi.org/10.1007/s10040-023-02667-w>

Group, W. B. (2015). *Cutzamala Diagnóstico integral*. Banco Mundial. <https://www.worldbank.org/ext/en/home>

Guo, Y., Zhang, Y., Zhang, L., & Wang, Z. (2021). Regionalization of hydrological modeling for predicting streamflow in ungauged catchments: A comprehensive review. In *Wiley Interdisciplinary Reviews: Water* (Vol. 8, Issue 1). John Wiley and Sons Inc. <https://doi.org/10.1002/wat2.1487>

Gupta, S., Hengl, T., Lehmann, P., Bonetti, S., Papritz, A. and Or, D.: [Global prediction of soil saturated hydraulic conductivity using random forest in a Covariate-based Geo Transfer Functions \(CoGTF\) framework](#). manuscript submitted to Journal of Advances in Modeling Earth Systems (JAMES).

Hidrológico-Administrativa, R. (n.d.). *Programa Hídrico Regional 2021-2024*.

Horton, P., Schaefli, B., & Kauzlaric, M. (2022). Why do we have so many different hydrological models? A review based on the case of Switzerland. In *Wiley Interdisciplinary Reviews: Water* (Vol. 9, Issue 1). John Wiley and Sons Inc. <https://doi.org/10.1002/wat2.1574>

Hrachowitz, M., Savenije, H. H. G., Blöschl, G., McDonnell, J. J., Sivapalan, M., Pomeroy, J. W., ... Cudennec, C. (2013). A decade of Predictions in Ungauged Basins (PUB)—a review. *Hydrological Sciences Journal*, 58(6), 1198–1255. <https://doi.org/10.1080/02626667.2013.803183>

Huang, Q., Qin, G., Zhang, Y., Tang, Q., Liu, C., Xia, J., Chiew, F. H. S., & Post, D. (2020). Using Remote Sensing Data-Based Hydrological Model Calibrations for Predicting Runoff in Ungauged or Poorly Gauged Catchments. *Water Resources Research*, 56(8). <https://doi.org/10.1029/2020WR028205>

Huerta-Vergara, A. R., Arciniega-Esparza, S., Salinas-Calleros, G., & Hernández-Espriú, A. (2025). Addressing circular water approaches for sustainable groundwater management in Mexico City. *Groundwater for Sustainable Development*, 29, 101435. <https://doi.org/https://doi.org/10.1016/j.gsd.2025.101435>

INEGI (2020) Censos y Conteos de Población y Vivienda 2020. Instituto Nacional de Estadística, Geografía e Informática, México. [Population and Housing Censuses and Counts 2020. National Institute of Statistics, Geography and Informatics, Mexico]. <https://www.inegi.org.mx/programas/ccpv/2020/>

Jajarmizadeh, M. (2014). *A Review on Theoretical Consideration and Types of Models in Hydrology*. <https://www.researchgate.net/publication/258768700>

James, G., Witten, D., Hastie, T., Tibshirani, R., & Taylor, J. (2023). Linear regression. In *An introduction to statistical learning: With applications in python* (pp. 69–134). Springer.

Kambalimath, S., & Deka, P. C. (2020). A basic review of fuzzy logic applications in hydrology and water resources. In *Applied Water Science* (Vol. 10, Issue 8). Springer Science and Business Media Deutschland GmbH. <https://doi.org/10.1007/s13201-020-01276-2>

Knoben, W. J. M., Freer, J. E., Peel, M. C., Fowler, K. J. A., & Woods, R. A. (2020). A Brief Analysis of Conceptual Model Structure Uncertainty Using 36 Models and 559 Catchments. *Water Resources Research*, 56(9). <https://doi.org/10.1029/2019WR025975>

Kohavi, R. K., & Kohavi, R. (1995). *A Study of Cross-Validation and Bootstrap for Accuracy Estimation and Model Selection*. <http://robotics.stanford.edu/~ronnyk>

Lim, T., & Wang, K. (2022). Comparison of machine learning algorithms for emulation of a gridded hydrological model given spatially explicit input. *Computers & Geosciences*, 159, 105025. <https://doi.org/https://doi.org/10.1016/j.cageo.2021.105025>

Loague, K., Heppner, C. S., Mirus, B. B., Ebel, B. A., Ran, Q., Carr, A. E., BeVille, S. H., & VanderKwaak, J. E. (2006). Physics-based hydrologic-response simulation: Foundation for hydroecology and hydrogeomorphology. In *Hydrological Processes* (Vol. 20, Issue 5, pp. 1231–1237). <https://doi.org/10.1002/hyp.6179>

Luersen, M. A., & le Riche, R. (2004). Globalized nelder-mead method for engineering optimization. *Computers and Structures*, 82(23–26), 2251–2260. <https://doi.org/10.1016/j.compstruc.2004.03.072>

Mazziotta, M., & Pareto, A. (2022). Normalization methods for spatio-temporal analysis of environmental performance: Revisiting the Min–Max method. *Environmetrics*, 33(5), e2730. <https://doi.org/https://doi.org/10.1002/env.2730>

McNally, A., Arsenault, K., Kumar, S., Shukla, S., Peterson, P., Wang, S., Funk, C., Peters-Lidard, C. D., & Verdin, J. P. (2017). A land data assimilation system for sub-Saharan Africa food and water security applications. *Scientific Data*, 4. <https://doi.org/10.1038/sdata.2017.12>

Mendez, M., Calvo-Valverde, L. A., Imbach, P., Maathuis, B., Hein-Grigg, D., Hidalgo-Madriz, J. A., & Alvarado-Gamboa, L. F. (2022). Hydrological Response of Tropical Catchments to Climate Change as Modeled by the GR2M Model: A Case Study in Costa Rica. *Sustainability (Switzerland)*, 14(24). <https://doi.org/10.3390/su142416938>

Moradkhani, H., & Sorooshian, S. (2008). General Review of Rainfall-Runoff Modeling: Model Calibration, Data Assimilation, and Uncertainty Analysis. *UC Irvine*. Retrieved from <https://escholarship.org/uc/item/42d5d6c9>

Mouelhi, S., Michel, C., Perrin, C., & Andréassian, V. (2006). Stepwise development of a two-parameter monthly water balance model. *Journal of Hydrology*, 318(1), 200–214. <https://doi.org/https://doi.org/10.1016/j.jhydrol.2005.06.014>

Muste, M., Lee, K., & Bertrand-Krajewski, J. L. (2012). Analyse normalisée des incertitudes en hydrométrie: Examen des approches pertinentes et exemples d'application. In *Hydrological Sciences Journal* (Vol. 57, Issue 4, pp. 643–667). <https://doi.org/10.1080/02626667.2012.675064>

Nash, J. E., & Sutcliffe, J. v. (1970). River flow forecasting through conceptual models part I — A discussion of principles. *Journal of Hydrology*, 10(3), 282–290. [https://doi.org/https://doi.org/10.1016/0022-1694\(70\)90255-6](https://doi.org/https://doi.org/10.1016/0022-1694(70)90255-6)

National Geographic Society. (2024). Dams. National Geographic Education. <https://education.nationalgeographic.org/resource/dams/>

Nelder, J. A., & Mead, R. (1965). A Simplex Method for Function Minimization. *The Computer Journal*, 7(4), 308–313. <https://doi.org/10.1093/comjnl/7.4.308>

Park, Y., & Ho, J. C. (2021). Tackling Overfitting in Boosting for Noisy Healthcare Data. *IEEE Transactions on Knowledge and Data Engineering*, 33(7), 2995–3006. <https://doi.org/10.1109/TKDE.2019.2959988>

Patro, S. G. K., & Sahu, K. K. (2015). Normalization: A Preprocessing Stage. *CoRR*, abs/1503.06462. <http://arxiv.org/abs/1503.06462>

Pushpalatha, R., Perrin, C., Moine, N. le, & Andréassian, V. (2012). A review of efficiency criteria suitable for evaluating low-flow simulations. *Journal of Hydrology*, 420–421, 171–182. <https://doi.org/https://doi.org/10.1016/j.jhydrol.2011.11.055>

Rajat, & Athira, P. (2021). Calibration of hydrological models considering process interdependence: A case study of SWAT model. *Environmental Modelling & Software*, 144, 105131. <https://doi.org/https://doi.org/10.1016/j.envsoft.2021.105131>

Raming, L. W., Vivoni, E. R., Cederstrom, C. J., Hossain, M. A., & Becerra, J. A. (2025). pyTRIBS: An open, modular, and reproducible python-based framework for distributed hydrologic modeling. *Environmental Modelling and Software*, 188. <https://doi.org/10.1016/j.envsoft.2025.106432>

Ricard, S., Bourdillon, R., Roussel, D., & Turcotte, R. (2013). Global Calibration of Distributed Hydrological Models for Large-Scale Applications. *Journal of Hydrologic Engineering*, 18(6), 719–721. [https://doi.org/10.1061/\(asce\)he.1943-5584.0000665](https://doi.org/10.1061/(asce)he.1943-5584.0000665)

SACMEX. (2018). *Diagnóstico Logros y Desafíos*. SACMEX. <https://aplicaciones.sacmex.cdmx.gob.mx/fut/>

- Saha, A., & Chandra Pal, S. (2024). Application of machine learning and emerging remote sensing techniques in hydrology: A state-of-the-art review and current research trends. In *Journal of Hydrology* (Vol. 632). Elsevier B.V. <https://doi.org/10.1016/j.jhydrol.2024.130907>
- Schultz, G. A. (1988). Remote sensing in hydrology. *Journal of Hydrology*, 100(1), 239–265. [https://doi.org/10.1016/0022-1694\(88\)90187-4](https://doi.org/10.1016/0022-1694(88)90187-4)
- Secretaría del Agua del Estado de México. (2023). *Atlas de las Cuencas Hidrológicas del Estado de México*. Secretaría del Agua del Estado de México. <https://agua.edomex.gob.mx/>
- Shinde, P. P., & Shah, S. (2018). A Review of Machine Learning and Deep Learning Applications. *2018 Fourth International Conference on Computing Communication Control and Automation (ICCUBE)*. <https://doi.org/10.1109/ICCUBE.2018.8697857>
- Stone, M. (1974). Cross-Validatory Choice and Assessment of Statistical Predictions. *Journal of the Royal Statistical Society: Series B (Methodological)*, 36(2), 111–133. <https://doi.org/https://doi.org/10.1111/j.2517-6161.1974.tb00994.x>
- Tang, Q., Gao, H., Lu, H., & Lettenmaier, D. P. (2009). Remote sensing: Hydrology. *Progress in Physical Geography*, 33(4), 490–509. <https://doi.org/10.1177/0309133309346650>
- Thirel, G., Santos, L., Delaigue, O., & Perrin, C. (2024). On the use of streamflow transformations for hydrological model calibration. *Hydrology and Earth System Sciences*, 28(21), 4837–4860. <https://doi.org/10.5194/hess-28-4837-2024>
- Thornthwaite, C. W., & Mather, J. R. (1951). The role of evapotranspiration in climate. *Archiv Für Meteorologie, Geophysik Und Bioklimatologie, Serie B*, 3(1), 16–39. <https://doi.org/10.1007/BF02242588>
- Tuller, M., Or, D., & Tuller, M. (2002). Unsaturated Hydraulic Conductivity of Structured Porous Media: A Review of Liquid Configuration-Based Models. Published in *Vadose Zone Journal* (Vol. 1). www.vadosezonejournal.org
- Van Genuchten, M. Th. (1980). A Closed-form Equation for Predicting the Hydraulic Conductivity of Unsaturated Soils. *Soil Science Society of America Journal*, 44(5), 892–898. <https://doi.org/https://doi.org/10.2136/sssaj1980.03615995004400050002x>

van Esse, W. R., Perrin, C., Booij, M. J., Augustijn, D. C. M., Fenicia, F., Kavetski, D., & Lobligeois, F. (2013). The influence of conceptual model structure on model performance: a comparative study for 237 French catchments. *Hydrology and Earth System Sciences*, 17(10), 4227–4239. <https://doi.org/10.5194/hess-17-4227-2013>

Wasi, H. A., & Shiker, M. A. K. (2021). Nonlinear Conjugate Gradient Method with Modified Armijo Condition to Solve Unconstrained Optimization. *Journal of Physics: Conference Series*, 1818(1), 012021. <https://doi.org/10.1088/1742-6596/1818/1/012021>

Wei, Z., & Wan, X. (2022). Spatial and Temporal Characteristics of NDVI in the Weihe River Basin and Its Correlation with Terrestrial Water Storage. *Remote Sensing*, 14(21). <https://doi.org/10.3390/rs14215532>

Wolfe, P. (1969). CONVERGENCE CONDITIONS FOR ASCENT METHODS*. In *SIAM REVIEW* (Vol. 11, Issue 2). <https://epubs.siam.org/terms-privacy>

Yevjevich, V. (1987). Stochastic Hydrology and Hydraulics Stochastic models in hydrology. In *Stochastic Hydrol. Hydraul* (Vol. 1).

Ying, X. (2019). An Overview of Overfitting and its Solutions. *Journal of Physics: Conference Series*, 1168(2). <https://doi.org/10.1088/1742-6596/1168/2/022022>

Zhu, C., Byrd, R. H., Lu, P., & Nocedal, J. (1997). Algorithm 778: L-BFGS-B: Fortran subroutines for large-scale bound-constrained optimization. *ACM Trans. Math. Softw.*, 23(4), 550–560. <https://doi.org/10.1145/279232.279236>