



UNIVERSIDAD TECNICA
FEDERICO SANTA MARIA

Departamento de Obras Civiles

Comparación de Técnicas de Machine Learning con el Modelo Hidrológico USM, para la predicción de caudales en una cuenca de alta montaña.

Memoria de Título presentada por

Benjamín Alejandro González Calvo

como requisito parcial para optar al título de la carrera de

Ingeniería Civil

Profesor Guía:

Luis Andrés De La Fuente Contreras.

Profesora Co referente:

Lina Mabel Castro Heredia.

Fecha: 21/07/2026



CONSTANCIA DE VALIDACIÓN Y CONFIDENCIALIDAD DE MONOGRAFÍA A REPOSITORIO ACADÉMICO

1.- IDENTIFICACIÓN DEL TRABAJO ACADÉMICO

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

Título del trabajo: Comparación de Técnicas de Machine Learning con el Modelo Hidrológico USM, para la predicción de caudales en una cuenca de alta montaña.

Nombre del candidato(a): Benjamín Alejandro González Calvo

Carrera / Grado: Ingeniería Civil / Pregrado

Campus: San Joaquín **Departamento:** Obras Civiles

2.- VALIDACIÓN DEL PROFESOR GUÍA/DIRECTOR DE TESIS

Yo, Luis De la Fuente Contreras, en mi calidad de profesor(a) guía/director(a) del trabajo académico mencionado anteriormente **DEJO CONSTANCIA** que:

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

3.- EVALUACIÓN DE CONFIDENCIALIDAD POR PROPIEDAD INDUSTRIAL (marcar una opción)

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

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

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

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

4.- FIRMAS

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

Fecha: 21-07-2026

Firma:

Estudiante o Candidato(a):

Fecha: 20-07-2026

Firma:

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



UNIVERSIDAD TECNICA
FEDERICO SANTA MARIA

TITULO DE LA TESIS:

Comparación de Técnicas de Machine Learning con el Modelo Hidrológico USM, para la predicción de caudales en una cuenca de alta montaña.

AUTOR:

Benjamín Alejandro González Calvo

TRABAJO DE MEMORIA, presentado como requisito parcial para optar al título de la carrera de INGENIERIA CIVIL de la Universidad Técnica Federico Santa María.

	Nombre	Firma
Prof. Guía	Luis De La Fuente Contreras	
Miembro 1 Comisión		
Miembro 2 Comisión		

Valparaíso/Santiago, Chile, Julio de 2026

Comparación de Técnicas de Machine Learning con el Modelo Hidrológico USM, para la predicción de caudales en una cuenca de alta montaña.

Benjamín Alejandro González Calvo, Luis De La Fuente Contreras

Universidad Técnica Federico Santa María

1. Resumen

Las cuencas de alta montaña son fundamentales para la gestión hídrica en Chile, pero su dinámica climática, caracterizada por una alta variabilidad y procesos difíciles de modelar, representa un desafío constante para los modelos hidrológicos tradicionales. Frente a ello, los modelos de Machine Learning (ML) permiten capturar relaciones no lineales y aprender patrones complejos directamente desde los datos, sin imponer explícitamente formulaciones físicas para representar los procesos hidrológicos. Este estudio evalúa el desempeño de tres modelos de ML, Random Forest (RF), Long Short-Term Memory (LSTM) y Convolutional Neural Networks (CNN) en la predicción de caudales para una cuenca de alta montaña en Chile central. Los resultados obtenidos se comparan con el modelo hidrológico USM y con una implementación posterior del mismo desarrollada en Python.

Para ello, se utilizaron variables meteorológicas y radiación extraterrestre, organizadas mediante variables rezagadas y acumuladas con el propósito de incorporar información antecedente del sistema hidrológico. Entre los modelos evaluados, LSTM presentó el mejor desempeño global, alcanzando un KGE de 0.80, seguido por CNN y RF, ambos con valores de 0.72. En todos los casos, los modelos de ML superaron el desempeño del modelo hidrológico USM. A nivel interanual también se observaron diferencias en la capacidad de representar eventos extremos y distintos regímenes hidrológicos, siendo LSTM el modelo que mostró mayor consistencia general.

Junto con el análisis predictivo, el estudio incorporó herramientas de interpretabilidad, como curvas de dependencia parcial (PDP), trayectorias individuales (ICE) y análisis de sensibilidad mediante perturbaciones controladas de $\pm 10\%$. Esto permitió examinar tanto la magnitud como la forma de las relaciones aprendidas por cada arquitectura. Los resultados evidencian diferencias en la sensibilidad y representación de variables asociadas al almacenamiento antecedente, la temperatura y los procesos de fusión nival, destacando el potencial de estas técnicas para complementar la modelación hidrológica tradicional en sistemas complejos.

Tabla de contenido

1. Resumen.....	3
2. Introducción	6
3. Metodología	8
3.1. Zona de estudio.....	8
3.2. Modelo hidrológico USM.....	8
3.2.1. Modelo de Balance de Energía de Utah (UEB).....	8
3.2.2. Modelo de Grado-Índice Mejorado (ETI)	9
3.2.3. Modelo de Escorrentía de Míguels	9
3.3. Modelo hidrológico distribuido de base física en Python	9
3.4. Base de datos	10
3.5. Diseño experimental.....	11
3.6. Modelos de ML	12
3.6.1. Random Forest (RF)	12
3.6.2. Long Short-Term Memory (LSTM)	13
3.6.3. Convolutional Neural Networks (CNN).....	16
3.7. Evaluación de desempeño	19
3.8. Análisis de interpretabilidad y sensibilidad modelos de ML.....	20
3.8.1. Importancia global de variables mediante permutación	20
3.8.2. Análisis PDP + ICE para variables dominantes	21
3.8.3. Análisis de sensibilidad mediante perturbaciones proporcionales ($\pm 10\%$)	22
4. Resultados	22
4.1. Configuración final modelos de ML	22
4.1.1. Configuración RF	22
4.1.2. Configuración LSTM	23
4.1.3. Configuración CNN.....	24
4.2. Desempeño comparativo	25
4.2.1. Métricas globales en periodo de prueba final.....	26
4.2.2. Hidrograma comparativo enfoques ML vs Enfoques físicos	26
4.2.3. Curva de duración de caudales	28
4.3. Análisis desempeño en años secos, normales y húmedos	29
4.3.1. Análisis años secos	30
4.3.2. Análisis años normales	32
4.3.3. Análisis años húmedos	35
4.4. Análisis de la respuesta hidrológica simulada bajo dinámicas contrastantes	37
4.4.1. Respuesta ante eventos extremos aislados del año 2013	37
4.4.2. Respuesta ante transición estacional seca-húmeda (año 2008).....	39
4.4.3. Sesgo acumulativo en volumen anual para años 2015 y 2019	40

4.5. Interpretabilidad de las arquitecturas de ML y sensibilidad a perturbaciones.....	42
4.5.1. Resultado del análisis curvas PDP+ICE.....	42
4.5.2. Respuestas de modelos de ML frente a sensibilidad de perturbaciones	46
4.5.3. Síntesis del análisis PDP+ICE y perturbaciones	48
5. Discusión	50
5.1. Complejidad arquitectónica y representación hidrológica	50
5.2. Desempeño bajo distintos regímenes hidrológicos	50
5.3. Representación de dinámicas hidrológicas contrastantes	52
5.4. Implicancias del análisis de sensibilidad e interpretabilidad.....	53
6. Conclusiones	54
7. Agradecimientos	57
8. Referencias.....	58
9. Anexos	62

2. Introducción

Las cuencas de alta montaña cumplen un rol clave en el abastecimiento de agua potable y de riego en poblaciones aguas abajo, y actúan como reservorios naturales que regulan el régimen hídrico anual (Blöschl et al., 2019). Sin embargo, la predicción de caudales en estos sistemas representa un desafío significativo para la modelación hidrológica debido a la complejidad de los procesos involucrados. En estas cuencas, la precipitación, el deshielo nival y glaciar son determinantes clave en la escorrentía, lo que las convierte en sistemas altamente variables y sensibles al cambio climático (Sanllehi, 2023).

Desde hace décadas, los modelos hidrológicos físicos han sido la piedra angular en la predicción del caudal de los ríos y la gestión de los recursos hídricos, ya que, al estar basados en ecuaciones matemáticas y principios físicos, han permitido representar los procesos de escorrentía, evaporación y percolación en las distintas cuencas hidrográficas existentes del mundo (Blöschl et al., 2019). Sin embargo, pese a sus buenos resultados, presentan sus limitaciones. Sus ecuaciones, aunque útiles, son simplificaciones de los procesos hidrológicos reales, lo que puede llevar a la pérdida de información crucial (Singh et al., 2023a). Por ejemplo, estos modelos suelen asumir relaciones lineales entre variables que en realidad tienen una naturaleza no lineal altamente compleja (Yang et al., 2019).

Muchos de los modelos hidrológicos tradicionales han sido calibrados bajo condiciones climáticas históricas que ya no reflejan con fidelidad el comportamiento actual ni futuro de las cuencas hidrográficas, especialmente considerando el impacto creciente del cambio climático (Yang et al., 2019; Singh et al., 2015a; Ali et al., 2018; Lutz et al., 2019; Singh et al., 2022). De acuerdo con este criterio, se ha advertido que los modelos calibrados podrían perder su capacidad predictiva si no se adaptan estructuralmente a nuevas condiciones hidroclimáticas futuras (Blöschl et al., 2019).

Frente a los escenarios actuales, ha emergido con fuerza el uso de técnicas de Machine Learning (ML), o conocidos también como modelos de aprendizaje automático, las cuales permiten modelar fenómenos hidrológicos altamente no lineales y aprender patrones directamente desde grandes volúmenes de datos, sin la necesidad de suposiciones previas sobre la estructura de los procesos físicos subyacentes (Nearing et al., 2021; Shen, 2018). A diferencia de los modelos hidrológicos tradicionales, que dependen de ecuaciones predeterminadas, los modelos de Machine Learning ajustan sus parámetros durante el proceso de entrenamiento a partir del uso de grandes volúmenes de datos, quedando posteriormente fijos para la etapa de validación final (Fu et al., 2020). De este modo, la relación entre sus variables no es impuesta por formulación de ecuaciones físicas explícitas, sino que son aprendidas automáticamente. Esto representa una ventaja importante en escenarios donde los modelos físicos suelen enfrentar dificultades, ya sea por la complejidad de los procesos hidrológicos o por la limitada disponibilidad de información detallada sobre las características de la cuenca.

Las técnicas de ML han mostrado desempeños comparables, e incluso superiores a los modelos hidrológicos tradicionales, particularmente en la representación de respuestas hidrológicas complejas y en la capacidad de reproducir la variabilidad del caudal bajo distintas condiciones del sistema. Con su capacidad para descifrar patrones complejos y manejar datos incompletos o ruidosos, las técnicas de ML son cada vez más populares y tienen más reconocimiento en la hidrología (Rozos et al., 2022; Shen, 2018). Modelos como Long Short Term Memory (LSTM), un tipo especial de redes neuronales recurrentes, son altamente adecuados para la tarea de modelado de lluvia-escorrentía. En un estudio preliminar, Kratzert et al. (2018) demostraron que un modelo LSTM podía alcanzar desempeños comparables, e incluso superiores, al modelo conceptual SAC-SMA + Snow-17 en la simulación de caudales para múltiples cuencas, utilizando principalmente series meteorológicas como variables de en-

trada. Más adelante, Kratzert et al. (2019) ampliaron este enfoque hacia un modelado regional más robusto, incorporando atributos estáticos de cuenca mediante una arquitectura Entity-Aware LSTM (EA-LSTM). Esta arquitectura permitió representar tanto comportamientos hidrológicos similares entre cuencas como respuestas específicas asociadas a sus características particulares, algo especialmente relevante en sistemas donde dos cuencas aparentemente similares pueden reaccionar de manera distinta frente a un mismo evento meteorológico (Kratzert et al., 2019).

Otros ejemplos de comparación de modelos lo encontramos con Singh et al. (2023), quien evaluó distintos modelos de Machine Learning, incluyendo Gaussian Linear Model (GLM), Gaussian Additive Model (GAM), Multivariate Adaptive Regression Splines (MARS), Artificial Neural Networks (ANN), Random Forest (RF) y 1D Convolutional Neural Networks (1D-CNN) con enfoques hidrológicos tradicionales en la cuenca del río Sutlej, encontrando que los modelos de inteligencia artificial obtenían mejores valores en métricas como el coeficiente de determinación (R^2), la eficiencia Kling-Gupta (KGE) y la eficiencia Nash-Sutcliffe (NSE) (Singh et al., 2023). También otro estudio realizó una evaluación similar en la cuenca del río Qingliu en China, donde Artificial Neural Networks (ANN), Support Vector Machines (SVR) y Random Forest (RF) superaron significativamente a los modelos hidrológicos basados en procesos físicos en términos de error cuadrático medio (RMSE) y error absoluto medio (MAE) (Yang et al., 2019).

Otras investigaciones han reforzado estos hallazgos en diferentes escenarios hidrológicos. Shortridge (2016) evidenciaron que los modelos de aprendizaje automático lograron menores errores de predicción en comparación con modelos físicos desarrollados para la región del lago Tana en Etiopía, destacando la flexibilidad de aquellos enfoques para capturar relaciones no lineales complejas (Shortridge et al., 2016). Por su parte, Adnan (2020) y Rahimzad (2021) analizaron el rendimiento de los modelos LSTM en la predicción diaria de caudales y concluyeron que estos modelos no solo reducían el error de predicción, sino que también mostraban una mayor capacidad de adaptación a condiciones climáticas cambiantes (Adnan et al., 2020; Rahimzad et al., 2021). Además, los modelos de Convolutional Neural Networks (CNN) y LSTM han demostrado ser efectivos para manejar conjuntos de datos hidrológicos complejos y mejorar la precisión de las predicciones en comparación con modelos físicos calibrados (Duan et al., 2020; Li et al., 2022).

El presente estudio busca abordar esta problemática comparando el desempeño de modelos de aprendizaje automático, con la última calibración del modelo hidrológico USM realizada en el estudio de Alekos Sanllehi (2023), así como con su implementación distribuida en Python desarrollada recientemente por Riveros (2025), para la predicción de caudales en una cuenca de alta montaña. Se entrenarán y validarán tres modelos de aprendizaje automático, RF, LSTM y CNN. Estos modelos utilizarán datos hidrometeorológicos de la cuenca, y se evaluará su capacidad predictiva mediante métricas de desempeño como la eficiencia Kling-Gupta (KGE), el error cuadrático medio (RMSE) y el error absoluto medio (MAE). Posteriormente, los resultados obtenidos se compararán con los resultados de los dos modelos base. También se analizarán fortalezas y debilidades de cada modelo. Este análisis permitirá determinar qué tan bueno es el desempeño de las técnicas de ML para la predicción de caudales y proporcionará recomendaciones sobre la aplicabilidad de estos métodos en la modelación hidrológica.

Además del análisis de desempeño, el estudio incorpora un enfoque de interpretabilidad mediante curvas de dependencia parcial (PDP), trayectorias individuales (ICE) y análisis de sensibilidad ante perturbaciones controladas. Estas herramientas permiten examinar no solo la precisión de los modelos, sino también la forma y magnitud de las relaciones que aprenden entre variables de entrada y caudal.

3. Metodología

3.1. Zona de estudio

La cuenca del río Juncal se encuentra en la Región de Valparaíso de Chile, y es un afluente importante del río Aconcagua. Con una longitud de aproximadamente 35 kilómetros, el río Juncal nace en el glaciar Juncal Norte, ubicado en el cerro Juncal en la cordillera de los Andes, entre las coordenadas $32^{\circ}43'$ y $33^{\circ}3'$ latitud sur y meridianos $69^{\circ}59'$ y $70^{\circ}10'$. Presenta una superficie de drenaje correspondiente a 343.1 [km²], con un 7,5% de la superficie correspondiente a zona con presencia de glaciares.

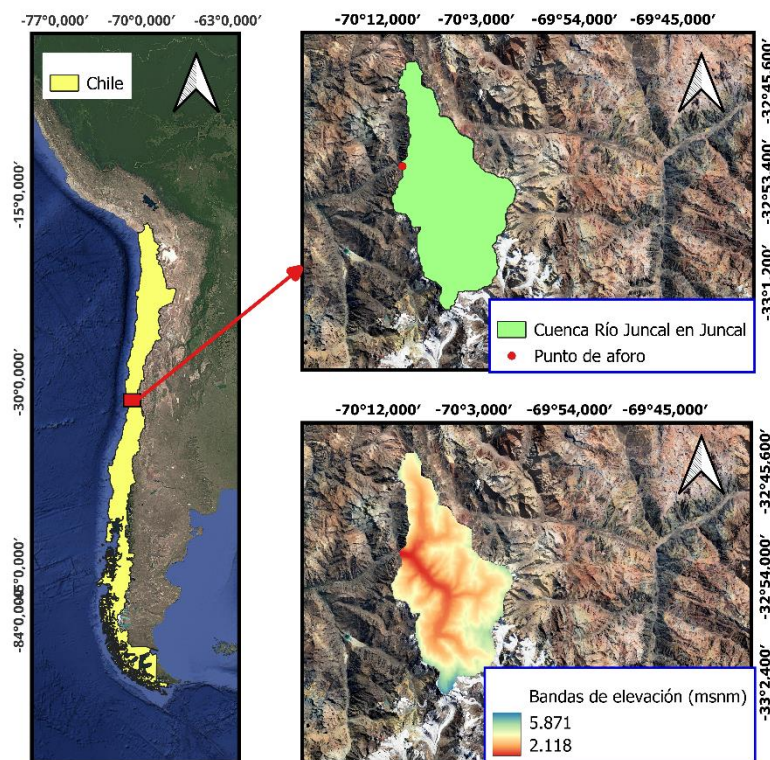


Figura 1. Ubicación geográfica de la cuenca Río Juncal en Juncal (*Fuente: elaboración propia*).

El clima predominante tipo tundra con verano seco para la zona cordillerana, junto con un clima de tipo mediterráneo en la zona de valle de acuerdo con la clasificación climática de Köppen–Geiger (Sarricolea et al., 2017), caracterizado por bajas temperaturas y precipitaciones en forma de nieve, lo que resulta en la presencia permanente de hielo y nieve en la región. Este régimen nival provoca que los caudales más altos del río ocurran durante el deshielo en los meses de primavera y verano, mientras que los estiajes se presentan en invierno.

3.2. Modelo hidrológico USM

El modelo Hidrológico USM fue originalmente propuesto por Stowhas (1975), con el objetivo de mejorar la representación del derretimiento nival y glaciar en cuencas de alta montaña. En estudios posteriores, como la investigación de Sanllehi Placencia (2023) propuso una reestructuración del modelo hidrológico USM, permitiendo mejorar la estimación de los aportes hídricos provenientes de la nieve, los glaciares y la precipitación líquida. Para ello, se integraron tres submodelos, cada uno con un rol específico en la simulación del ciclo hidrológico de las cuencas estudiadas.

3.2.1. Modelo de Balance de Energía de Utah (UEB)

Uno de los mayores desafíos en la modelación hidrológica de cuencas de alta montaña radica en capturar la dinámica de acumulación y fusión de la nieve. Tradicionalmente, el modelo USM utilizaba un submodelo semi-distribuido de nieve, aunque funcionaba, presentaba limitaciones en su capacidad de representar los procesos energéticos asociados al manto nival. Para superar estas limitaciones, se incorporó el Modelo de Balance de Energía de Utah, un enfoque basado en las ecuaciones de balance de energía y de masa (Tarboton et al., 1994). Este modelo permite estimar con mayor precisión la evolución del manto de nieve al considerar factores determinantes como lo es la radiación solar, la temperatura del aire y el contenido energético de la nieve. Gracias a esta modificación, fue posible capturar con mayor realismo el comportamiento del derretimiento nival y su influencia en la escorrentía (Sanllehi, 2023).

3.2.2. Modelo de Grado-Índice Mejorado (ETI)

Si bien el derretimiento nival constituye una fuente predominante de escorrentía, el retroceso glaciar y su contribución al caudal de los ríos se han vuelto cada vez más relevantes en el contexto de cambio climático. Para abordar los procesos de fusión glaciar, Sanllehi integró el Modelo de Grado e Índice Mejorado (ETI) (Pellicciotti et al., 2005) que introduce una combinación del enfoque clásico de grado-día con la radiación de onda corta como un factor adicional para estimar el derretimiento del hielo glaciar.

A diferencia de los modelos tradicionales que solo consideran la temperatura del aire como el factor determinante de los derretimientos glaciares, el modelo ETI incorpora la radiación solar como un impulsor clave para el proceso de fusión glaciar. Este cambio permitió mejorar la estimación de los aportes de los glaciares al caudal y evaluar su importancia relativa dentro del balance hídrico de las cuencas analizadas.

3.2.3. Modelo de Escorrentía de Migueles

Finalmente, la estimación de los aportes hídricos no estaría completa sin un esquema robusto para simular la escorrentía generada por la combinación de derretimiento nival, glaciar y la precipitación líquida. Para ello, se implementó el Modelo de Escorrentía de Migueles (Migueles et al., 1994), un modelo conceptual, de características semi distribuido diseñado para representar la transformación del agua disponible en caudal efectivo.

Este modelo permite procesar de manera integrada los diferentes aportes hídricos y simular su transporte a través del sistema hidrológico. Las variables clave consideradas en su formulación incluyen la cantidad de agua derretida, la infiltración del suelo y el escurrimiento superficial y subterráneo. De esta manera, el modelo no solo permite evaluar la respuesta de la cuenca ante diferentes forzantes climáticos, sino también identificar la proporción de cada fuente de agua en la escorrentía total.

3.3. Modelo hidrológico distribuido de base física en Python

En un estudio reciente, Riveros (2025) desarrolló una implementación en Python del modelo glacio hidrológico USM modificado en el trabajo de Sanllehi (2023), originalmente programado en MATLAB. El objetivo principal fue reproducir el funcionamiento del modelo en un entorno de código abierto, manteniendo tanto su formulación física como la estructura conceptual propuesta en la versión original (Riveros et al., 2025).

La implementación conserva la lógica modular del modelo base, integrando de manera independiente los submodelos UEB para la acumulación y derretimiento nival, ETI para el derretimiento glaciar y el modelo de escorrentía de Migueles. Cada componente fue desarrollado como un módulo autónomo, lo que permitió realizar su ejecución, calibración y validación por separado antes de integrarlos en el

esquema completo de simulación. Esta organización no solo facilita el análisis del modelo, sino que también entrega una mayor flexibilidad para futuras modificaciones o mejoras.

El trabajo incorpora un módulo de ruteo distribuido, basado en la resolución bidimensional de las ecuaciones de Saint Venant, con el propósito de representar de forma distribuida el transporte superficial del agua dentro de la cuenca. Gracias a esta incorporación, fue posible acoplar los submodelos de derretimiento nival y glaciar a un esquema de ruteo completamente distribuido, ampliando así la capacidad del modelo para representar la dinámica espacial del sistema hidrológico.

La implementación desarrollada por Riveros (2025) conserva la formulación física del modelo glacio hidrológico USM original, pero reorganiza su arquitectura dentro de un entorno modular en Python e incorpora un esquema de ruteo distribuido basado en ecuaciones de Saint Venant. Al igual que en el trabajo de Sanllehi (2023), esta adaptación será utilizada como caso base para la comparación con los modelos de Machine Learning.

3.4. Base de datos

Para la implementación y evaluación de los modelos de ML, es necesario contar con un conjunto de datos hidrometeorológicos representativos de la cuenca de estudio, para describir tanto las condiciones climáticas como la respuesta hidrológica del sistema. En este trabajo, las variables consideradas como entrada incluyen precipitación, temperatura del aire, velocidad del viento, caudales observados y radiación extraterrestre, cada una de ellas vinculada a procesos que influyen directamente en la dinámica hidrológica de la cuenca.

Para las variables meteorológicas obtenidas a partir de productos grillados, se extrajeron todas las celdas ubicadas dentro de los límites de la cuenca del río Juncal. Luego, en cada paso temporal, se calculó el promedio espacial de estas celdas para generar una única serie temporal representativa de las condiciones medias de la cuenca. Esta serie se utilizó como datos de entrada para el entrenamiento y la validación de los modelos de aprendizaje automático. A continuación, se describen con mayor detalle las características y el tratamiento de estas variables.

- Precipitación diaria (mm): Se utilizarán datos del producto satelital CR2MET (Alvarez-Garreton et al., 2018), con una resolución espacial de 0.05° y temporal diario. La precipitación es una de las variables más críticas, ya que determina el volumen de agua disponible en la cuenca.
- Temperatura máxima y mínima ($^\circ\text{C}$): Extraída de CR2MET (Alvarez-Garreton et al., 2018), esta variable es clave en la modelación del derretimiento nival y glaciar, además de influir en la evapotranspiración.
- Velocidad del viento (m/s): Datos con frecuencia horaria, ajustados mediante el Explorador de Energía Eólica (Muñoz et al., 2018), esenciales para la evaporación y el transporte de humedad.
- Caudales diarios observados (m^3/s): Se obtienen de la Dirección General de Aguas (DGA), representando los valores reales medidos en la cuenca (Dirección General de Aguas, 2023). Estos datos corresponden a la variable objetivo-utilizada en el entrenamiento y validación.
- Radiación solar extraterrestre (W/m^2): Esta variable fue incorporada como representación de energía solar incidente potencial, permitiendo caracterizar la disponibilidad energética para procesos de evapotranspiración y fusión nival. Se calculó a partir de las coordenadas geográficas (latitud y

longitud) mediante un procedimiento propio basado en fórmulas astronómicas (Iqbal, M. et al., 2012), que considera el día juliano, la declinación solar y la distancia Tierra-Sol. El cálculo fue implementado en Python y permitió generar una serie diaria continua, que se integró al set de variables predictoras empleadas en los modelos de aprendizaje automático.

3.5. Diseño experimental

Para garantizar una adecuada evaluación de desempeño entre los modelos de ML y el modelo hidrológico USM, se empleará el mismo criterio de división de datos utilizado en el estudio de Sanllehi (2023). En dicho estudio, la selección de los períodos de calibración y validación se realizó asegurando que ambas fases incluyeran la misma cantidad de años secos y húmedos, con el propósito de representar adecuadamente la variabilidad climática y minimizar la incertidumbre en los resultados obtenidos. Esta clasificación se fundamentó en la caracterización hidrometeorológica de la cuenca, asegurando que los datos de validación permitieran evaluar de manera realista la capacidad predictiva del modelo sin sesgos derivados de una distribución desbalanceada de condiciones climáticas. Esta estrategia permite reducir sesgos asociados a distribuciones climáticas desbalanceadas y mejorar la capacidad de generalización de los modelos.

Tabla 1. Tabla resumen de periodo de años de entrenamiento y periodos de validación.

Años de entrenamiento	2002	2003	2004	2006	2009	2011	2012	2014	2017	2018
Años de validación	2000	2001	2005	2007	2008	2010	2013	2015	2016	2019

Para optimizar el desempeño de los tres modelos, se implementa un procedimiento sistemático de búsqueda de la mejor configuración de hiperparámetros, con el objetivo de poder armar una arquitectura que maximice los resultados de cada técnica de ML. Para no comprometer la evaluación final del modelo, se extrajeron los años 2006 y 2018 del conjunto de datos de entrenamiento, reservándolos exclusivamente para este proceso. Estos dos años fueron seleccionados por su representatividad climática (al representar condiciones contrastantes dentro de la serie analizada, correspondientes a un año seco y un año húmedo) y funcionaron como un subconjunto de prueba intermedio. Así, los modelos fueron entrenados inicialmente con los datos de entrenamiento restantes (sin 2006 y 2018) y luego evaluados sobre este conjunto de validación interna para seleccionar la arquitectura u configuración más robusta.

Con el propósito de complementar el análisis global mediante una evaluación diferenciada según los distintos regímenes hidrológicos, se realiza una clasificación interanual del período de validación final para años secos, normales y húmedos. Esta segmentación permite analizar cómo varía el desempeño estadístico cuando cambian las condiciones de disponibilidad hídrica, evitando que el análisis global oculte comportamientos específicos asociados a eventos extremos o períodos de escasez.

Para clasificar los años en secos, normales y húmedos, se utiliza como indicador el caudal medio anual de la serie observada. A partir de esta variable, se construye una distribución de valores anuales y se definieron umbrales basados en percentiles, de manera de representar distintos niveles de disponibilidad hídrica. En particular, los años con caudales bajo el percentil 33 se clasifican como secos, aquellos entre los percentiles 33 y 66 como normales, y los superiores al percentil 66 como húmedos. Este criterio permite segmentar la serie de forma objetiva, capturando condiciones representativas de déficit, normalidad y abundancia hídrica dentro de la cuenca.

3.6. Modelos de ML

3.6.1. Random Forest (RF)

El modelo Random Forest (RF), propuesto por Breiman (2001), es un método de aprendizaje supervisado, de carácter no paramétrico basado en la construcción de un ensamble de árboles de decisiones. A diferencia de los enfoques paramétricos no asume una distribución específica de los datos, lo que lo hace más flexible para problemas de clasificaciones y regresiones. RF busca reducir la alta variabilidad y sensibilidad al ruido de los árboles de decisión individuales al crear múltiples árboles con muestras aleatorias obtenidas mediante la técnica de “Bootstrapping”. Al hacer divisiones con subconjuntos aleatorios de datos, se reducen las correlaciones entre los árboles y se mejora la estabilidad del conjunto. Este enfoque reduce la varianza de las predicciones y mejora la precisión general del modelo (Breiman, 2001).

El modelo se construye a partir de un conjunto de “**árboles de decisión**”, donde cada árbol genera una predicción individual $\hat{y}_i$ definida por la función de decisión $f(x)$, lo cual corresponde a la predicción del caudal diario a partir de particiones sucesivas del espacio de predictores. La predicción final del modelo se obtiene como el promedio de las predicciones generadas por todos los árboles que componen el bosque, lo que permite reducir la varianza y mejorar la estabilidad del resultado final. De forma general, la predicción de cada árbol puede expresarse como:

$$\hat{y}_i = f(x) \quad (1)$$

Mientras que la estimación final del modelo corresponde a:

$$y = \frac{1}{N} \sum_{i=1}^N \hat{y}_i \quad (2)$$

donde $\hat{y}_i$ corresponde a la predicción generada por el i -ésimo árbol de decisión y N corresponde al número total de árboles que conforman el bosque.

Al promediar las predicciones individuales de cada árbol, el modelo logra atenuar la varianza asociada a estimaciones individuales y mejora la estabilidad y capacidad de generalización del resultado final (Breiman, 2001). De esta manera, la predicción obtenida no depende de una única estructura de decisión, sino del consenso entre un conjunto diverso de modelos entrenados a partir de distintas particiones del conjunto de datos.

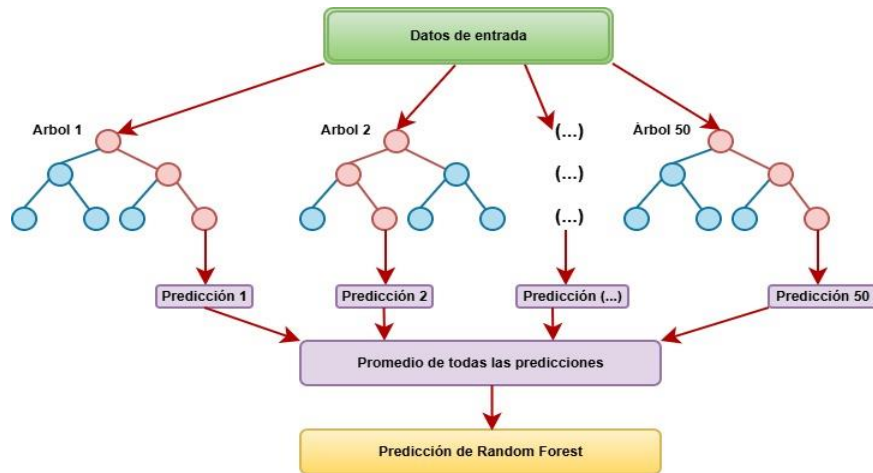


Figura 2. Esquema conceptual del modelo Random Forest. A partir de un conjunto de datos de entrada, múltiples árboles de decisión generan predicciones individuales, las cuales son posteriormente agregadas mediante un promedio para obtener la predicción final del modelo (Fuente: elaboración propia).

El modelo RF fue implementado utilizando la librería de Python Scikit-learn (Pedregosa et al., 2011), junto con herramientas complementarias como pandas, Matplotlib y NumPy (Harris et al., 2020) para el procesamiento de datos, visualización y cálculo de métricas. La implementación consideró una configuración explícita de los principales hiperparámetros del modelo, tales como el número de árboles del bosque, la profundidad máxima de cada árbol y el número mínimo de muestras requeridas para realizar particiones internas.

Debido a su naturaleza no secuencial y su falta de memoria interna del algoritmo, por esta razón fue necesario una reconstrucción explícita de la información pasada para capturar la dinámica del caudal. Para ello, se deben generar variables acumuladas y rezagadas a partir de los datos meteorológicos diarios, como sumas móviles de precipitación, medias térmicas de días anteriores. De este modo, se le entrega al modelo una representación indirecta de la memoria hidrológica que compensa su incapacidad de almacenamiento de memoria pasada. Con el fin de solucionar dicha limitación del modelo, fue necesario reconstruir la información pasada a partir de las series meteorológicas diarias. Para ello, se incorporaron retardos diarios de hasta siete días para cada una de las variables de entrada, lo cual permite al modelo aprender relaciones de tipo “memoria corta”, lo que permitió representar la influencia inmediata de forzantes recientes sobre el caudal diario y capturar procesos de respuesta rápida del sistema hidrológico.

Adicionalmente, se incorporaron variables agregadas con el objetivo de ayudar al modelo a representar procesos hidrológicos de evolución lenta y acumulativa, vinculados principalmente a la persistencia de condiciones meteorológicas y al derretimiento progresivo nival y glaciar. En particular, se calcularon acumulados de precipitación y promedios de temperatura mínima, temperatura máxima, velocidad del viento y radiación solar extraterrestre sobre distintas escalas temporales, las cuales abarcaron intervalos de 8 a 15, 16 a 31, 32 a 62, 63 a 123 y 124 a 224 días, permitiendo al modelo capturar tanto la respuesta estacional del sistema como efectos retardados asociados a la acumulación de nieve y a la liberación gradual de caudal durante los períodos de fusión. De este modo, la inclusión de estas nuevas variables permitió representar una forma de memoria hidrológica de largo plazo, que complementa la memoria de corto plazo generada mediante los retardos diarios.

3.6.2. Long Short-Term Memory (LSTM)

Uno de los principales desafíos en la predicción hidrológica es la capacidad de modelar series temporales a largo plazo, donde el estado actual del sistema depende fuertemente de lo ocurrido en días, semanas o incluso meses. En cuencas de montaña, el caudal de hoy no es simplemente una respuesta inmediata a la lluvia reciente, sino el resultado acumulado de procesos de almacenamiento, fusión nival y liberación gradual de agua. Para abordar esta necesidad, se implementa el modelo Long Short-Term Memory (Hochreiter & Schmidhuber, 1997), una red neuronal recurrente diseñada para aprender dependencias a largo plazo en series de datos. A diferencia de las redes neuronales tradicionales que pueden sufrir de problemas de desvanecimiento de memoria o explosión del gradiente, las LSTM cuentan con unidades de memoria interna que almacenan o descartan información a lo largo del tiempo, y lo hace de mejor manera que otras redes neuronales.

Cada unidad LSTM contiene la compuerta de entrada, la compuerta de olvido y la compuerta de salida, que son sus tres mecanismos principales de control. Estas compuertas regulan qué información se incorpora a la memoria, cuál se descarta para la predicción final y qué parte del contenido almacenado en la celda de memoria, es el que se propaga hacia la salida del modelo (Hochreiter & Schmidhuber, 1997).

La **puerta de entrada** controla la cantidad de información nueva que se incorpora a la celda de memoria y se define como:

$$i_t = \sigma(W_i * [h_{t-1}, x_t] + B_i) \quad (3)$$

donde i_t es la activación de la puerta de entrada, W_i son los pesos, h_{t-1} es la salida de la celda anterior, x_t es la entrada actual y B_i es el sesgo.

La **puerta del olvido** decide qué información se debe eliminar de la celda de memoria, donde su ecuación corresponde a la siguiente expresión:

$$f_t = \sigma(W_f * [h_{t-1}, x_t] + B_f) \quad (4)$$

donde f_t corresponde a la activación de la puerta del olvido para el tiempo t , los términos W_f y B_f representan los pesos y sesgos asociados a esta compuerta.

La actualización del **estado interno de la celda** se realiza combinando la información retenida y la nueva información candidata, según la expresión:

$$C_t = f_t * C_{t-1} + i_t * \tilde{C}_t \quad (5)$$

donde C_t es el estado de la celda en el tiempo t , C_{t-1} corresponde al estado previo y $\tilde{C}_t$ es la información candidata a ser incorporada a la celda.

Finalmente, la **puerta de salida** controla que parte de la información de C_t se entrega para la salida del modelo, se calcula como:

$$o_t = \sigma(W_o * [h_{t-1}, x_t] + B_o) \quad (6)$$

$$h_t = o_t * \tanh(C_t) \quad (7)$$

donde o_t es la puerta de salida, h_t corresponde el estado oculto del modelo en el tiempo t y representa la salida interna de la unidad LSTM. La Figura 3 resume el funcionamiento descrito anteriormente, mostrando el flujo de información a través de la unidad LSTM y la manera en que interactúan las compuertas de entrada, olvido y salida para gestionar el procesamiento de la información.

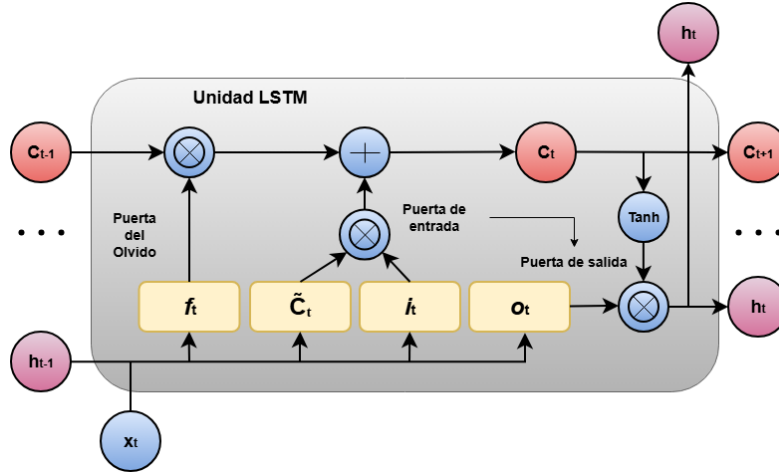


Figura 3. Esquema conceptual de unidad LSTM y flujo de información mediante sus compuertas de entrada, olvido y salida. (adaptado de Hochreiter y Schmidhuber, 1997).

El modelo LSTM fue implementado en la librería PyTorch (Paszke et al., 2019), utilizando además bibliotecas auxiliares como NumPy (Harris et al., 2020), Pandas, Matplotlib y Scikit-learn (Pedregosa et al., 2011) para el procesamiento de datos, normalización, visualización y cálculo de métricas. Previo a la fase de entrenamiento, cada variable se normaliza mediante un escalamiento del tipo Min-Max, con el objetivo de estabilizar el proceso de optimización. Asimismo, para garantizar la reproducibilidad de los resultados, se estableció una semilla aleatoria común ($seed = 42$) en los generadores pseudoaleatorios de NumPy y PyTorch, de modo que las ejecuciones del experimento produjeran resultados consistentes bajo las mismas condiciones.

Para ajustar el modelo LSTM, se exploraron las siguientes combinaciones de hiperparámetros, los cuales son claves para controlar la capacidad representativa del modelo, su estabilidad numérica y riesgo a sobreajustes. En particular, se evaluaron distintas configuraciones del número de unidades de neuronas de la capa LSTM ($hidden_size$), este parámetro determina la capacidad del modelo para aprender representaciones internas complejas, el número de capas recurrentes apiladas (num_layers) y la longitud de la secuencia de entrada (seq_length), el cual define cuántos días consecutivos de información se entregan al modelo para predecir el siguiente caudal. Este último es un parámetro crítico, ya que un valor muy corto puede hacer perder contexto relevante, mientras que uno muy largo puede introducir ruido o redundancia. Asimismo, se incorporó el uso de “dropout” como mecanismo de regularización, aplicado únicamente en arquitecturas con más de una capa, con el fin de reducir la dependencia excesiva entre neuronas y mejorar la capacidad de generalización del modelo.

El proceso de entrenamiento se llevó a cabo minimizando el error cuadrático medio mediante el optimizador Adam (Kingma & Ba, 2017), el cual logra una actualización adaptativa de los pesos y ha demostrado un buen desempeño en problemas de optimización no lineal asociados a redes neuronales profundas. Con el fin de garantizar reproducibilidad y estabilidad en los resultados, se fijaron semillas

aleatorias en las distintas librerías utilizadas, asegurando que cada configuración evaluada partiera desde unas condiciones iniciales controladas.

Para poder mejorar la capacidad del modelo LSTM y no sobredimensionar el número de unidades de memoria de su arquitectura, se diseñó un conjunto de entradas que incorporan información tanto de corto como de largo plazo. Como ya hemos mencionado que LSTM es capaz de aprender dependencias temporales desde su secuencia interna, sin embargo, se observó que aumentar en exceso la longitud de su secuencia o ventana temporal (por ejemplo, usar un `Seq_length` muy largos de más de 300 días) reducía el desempeño del modelo. Esto podría deberse a una pérdida de información relevante cuando se tiene una ventana de memoria muy larga, dificultad de entrenamiento o sobreajuste para las características de la cuenca. Por esta razón, se optó por una estrategia mixta, que consiste en mantener una secuencia de entrada de días (`seq_length`), y complementarla con variables agregadas, que actúan como una memoria adicional que no depende del estado interno del modelo.

Se incorporaron cinco variables adicionales, estas variables incluyen el acumulado de precipitación y de radiación extraterrestre, así como los promedios de temperatura máxima, temperatura mínima y velocidad del viento en dicha ventana de tiempo encontrada. Estas variables fueron diseñadas para resumir tendencias lentas en el comportamiento climático e hidrológico, como lo pueden ser la acumulación de humedad en el suelo, procesos de derretimiento nival o efectos de almacenamiento subterráneo, que no necesariamente se podrían capturar bien con la secuencia directa de LSTM. Así, se ayuda y evita forzar al modelo a retener información muy antigua dentro de su estado interno, entregándosela directamente como entrada adicional. Este enfoque se alinea con estrategias utilizadas en la literatura reciente, donde se han propuesto modelos LSTM que integran información multiescalar, en lugar de solo tener una escala diaria en los datos. Por ejemplo, Gauch et al. (2021) desarrollaron arquitecturas con entradas a distintas escalas temporales, como el uso de escalas horarias y diarias, y demostraron que la combinación mejora la predicción del caudal, especialmente en cuencas complejas con respuesta retardada (Gauch et al., 2021).

Aunque el enfoque propuesto en este trabajo no constituye una arquitectura multiescalar explícita, como la propuesta por Gauch et al. (2021), se alinea conceptualmente con su lógica de trabajo. En dicho estudio, los autores desarrollaron el modelo MTS-LSTM, que procesa entradas a dos escalas temporales mediante ramas LSTM paralelas, una para forzantes diarios y otra para forzantes horarios, transfiriendo información entre ambas a través de estados intermedios (Gauch et al., 2021). A pesar de la diferencia estructural, ambos enfoques comparten el objetivo de incorporar la información en múltiples escalas temporales para mejorar la capacidad de predicción del modelo. En nuestro caso, la memoria no solo se aprende secuencialmente, sino que también se sintetiza en variables agregadas externas, lo cual permite representar procesos lentos sin exigir mayor capacidad a la red neuronal o generar una arquitectura muy compleja que se ajuste demasiado a los datos que vio en el entrenamiento.

3.6.3. Convolutional Neural Networks (CNN)

En los últimos años, las CNN han demostrado ser herramientas eficaces para la detección de patrones en datos estructurados (Lecun, 1998). Aunque tradicionalmente han sido altamente utilizadas en el procesamiento de imágenes, en estudios recientes, esta arquitectura ha sido adaptada para analizar series temporales como las que se utilizan en hidrología, extrayendo información de variables como la precipitación y la temperatura sin necesidad de modelar relaciones explícitas entre los datos.

Investigaciones recientes han utilizado este tipo de arquitecturas junto con variables hidrometeorológicas, como precipitación y temperatura, para representar la dinámica de escorrentía en cuencas hidro-

gráficas. Singh et al. (2023) , por ejemplo, implementaron una arquitectura 1D-CNN en la cuenca del río Sutlej, obteniendo resultados competitivos frente a otros enfoques de ML aplicados a la modelación hidrológica.

En este estudio, hemos implementado una CNN con un filtro unidimensional (1D-CNN) para trabajar específicamente con la dimensión temporal. En lugar de analizar imágenes, la red procesa secuencias de datos meteorológicos para aprender cómo las variaciones en los distintos parámetros afectan la cantidad de agua que fluye en la cuenca. De esta forma, podemos detectar patrones recurrentes en los datos y mejorar las predicciones de caudal sin necesidad de recurrir a modelos más costosos computacionalmente.

La convolución constituye el núcleo de las CNN, consiste en una operación matemática que combina un filtro o “Kernel”, con una entrada para producir un mapa de características. La idea central es extraer información relevante localmente, reduciendo las conexiones entre todas las neuronas. El proceso de aprendizaje de la CNN comienza con la aplicación de filtros de convolución, que recorren los datos meteorológicos en ventanas de tiempo, buscando características relevantes en cada intervalo. Estos filtros funcionan de manera similar a cómo el ojo humano reconoce patrones en una imagen, pero en este caso, identifican estructuras en las variaciones de temperatura, viento y precipitación.

$$y_{(i,j)} = \sum_{m=1}^k \sum_{n=1}^k x_{(i+m-1,j+n-1)} * w_{(m,n)} + b \quad (8)$$

- $x_{(i+m-1,j+n-1)}$: Representa los valores de la entrada (como una imagen o serie de datos).
- $w_{(m,n)}$: Es el kernel o filtro que se desplaza por la entrada.
- k : Es el tamaño del filtro.
- b : Es el sesgo que permite ajustar la salida.
- $y_{(i,j)}$: Es el valor del mapa de características en la posición (i,j).

Para introducir no linealidad en el modelo, las CNN emplean funciones como “Rectified Linear Unit” (ReLU), conocida también como función de activación (Nair & Hinton, 2010). Su principal ventaja radica en que evita en gran medida el desvanecimiento del gradiente, ya que mantiene un gradiente constante cuando las activaciones son positivas (Glorot & Bengio, 2010; Nair & Hinton, 2010). No obstante, esta misma propiedad puede generar limitaciones, ya que para valores negativos la función anula completamente el gradiente, lo que puede llevar a que algunas neuronas queden inactivas durante el entrenamiento (He et al., 2015).

$$f(x) = \text{Max}(0, x) \quad (9)$$

Después de esta fase, se emplea una técnica llamada “max-pooling”, que selecciona los valores más representativos de cada ventana de tiempo, reduciendo la cantidad de información sin perder los detalles más importantes. Finalmente, la red neuronal combina todos estos datos procesados para generar una predicción del caudal esperado en función de los patrones que ha aprendido a lo largo del entrenamiento.

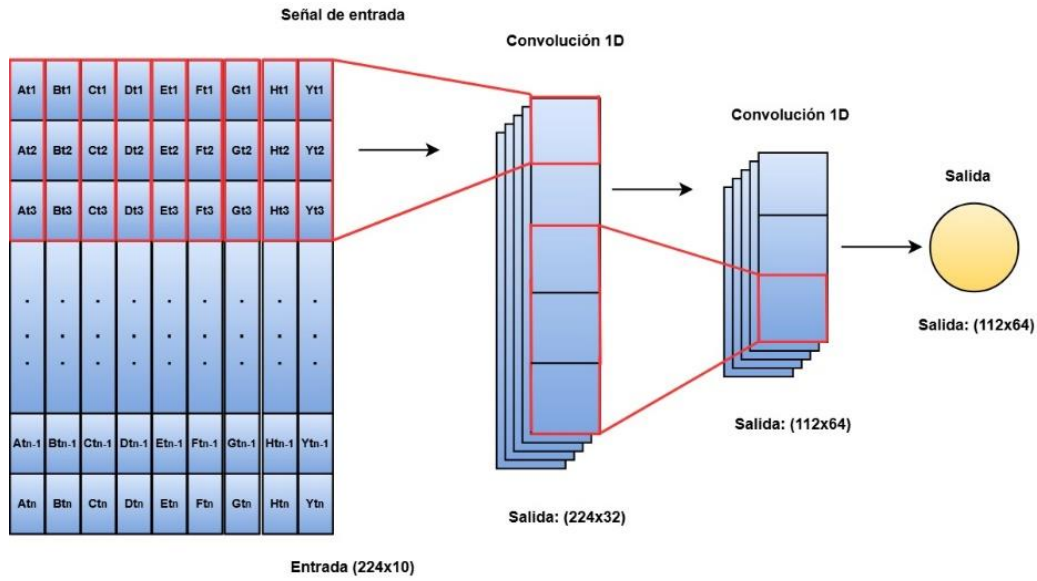


Figura 4. Esquema conceptual de la arquitectura CNN utilizada para la predicción de caudales. La red recibe como entrada una ventana temporal de variables meteorológicas y aplica convoluciones unidimensionales para extraer patrones relevantes en la serie. Las características obtenidas se integran para generar la estimación final del caudal (*Fuente: elaboración propia*).

El modelo CNN, al igual que LSTM, fue desarrollado en la librería PyTorch, aprovechando su capacidad para definir arquitecturas de redes profundas de forma eficiente y flexible. Para el tratamiento de datos, normalización, visualización y evaluación de métricas, también se utilizaron bibliotecas auxiliares como NumPy, Pandas, Scikit-learn y Matplotlib, manteniendo consistencia con el flujo de trabajo utilizado para LSTM.

En concordancia con el procedimiento realizado con la red LSTM, el entrenamiento del modelo CNN se realizó utilizando el optimizador Adam. Previo al entrenamiento, tanto las predictoras como el caudal objetivo fueron normalizados mediante un escalamiento Min–Max. Con el fin de garantizar la reproducibilidad de los resultados, se fijó una semilla aleatoria común ($\text{seed}=42$) para los generadores de números pseudoaleatorios de NumPy y PyTorch, asegurando consistencia en la inicialización de pesos y en el muestreo de los lotes durante el entrenamiento.

Para el ajuste del modelo convolucional se exploraron distintas configuraciones de hiperparámetros que controlan tanto la escala temporal de análisis como la capacidad de extracción de características del modelo. En particular, se evaluaron diferentes valores de la longitud de la secuencia de entrada (*sequence length*), la cual define el tamaño de la ventana temporal total entregada al modelo como bloque de información, a diferencia del enfoque secuencial que presentaba la ventana de LSTM. En el caso de la CNN, esta ventana no representa una memoria interna, sino un intervalo fijo sobre el cual se aplican operaciones convolucionales para identificar patrones temporales.

Dentro de esta ventana se analizaron distintos tamaños del kernel convolucional (*kernel size*), definido como una sub-ventana deslizante que recorre la secuencia temporal y controla el número de pasos temporales que se consideran al aplicar cada convolución, por lo que un valor pequeño permite capturar detalles locales, mientras que uno mayor abarca patrones todavía más prolongados. Asimismo, se evaluaron distintas configuraciones del número de filtros en la primera y segunda capa convolucional (*filtros_1* y *filtros_2*), los cuales determinan la cantidad de mapas de activación generados en cada nivel de la red y afectan directamente la capacidad de extracción de características del modelo. Un mayor número de filtros permite capturar una mayor diversidad de patrones, aunque también incre-

menta la carga computacional. Adicionalmente, se consideró la tasa de dropout, aplicada entre las capas convolucionales y la capa totalmente conectada con el objetivo de reducir el sobreajuste, así como distintos valores de la tasa de aprendizaje (learning rate) asociada al optimizador, la cual controla la magnitud de las actualizaciones de los pesos del modelo durante el proceso de entrenamiento.

Al igual que en el enfoque adoptado para el modelo LSTM, la arquitectura CNN se complementó con un conjunto de variables agregadas multiescales, diseñadas para representar procesos hidrológicos de respuesta lenta que no se capturan adecuadamente a partir de ventanas cortas. Estas variables también fueron construidas mediante acumulados móviles y promedios calculados sobre un intervalo temporal previo al día de predicción, cuya extensión fue tratada como un parámetro del modelo y ajustada durante el proceso de entrenamiento. En particular, se consideraron agregados asociados a la precipitación y la radiación extraterrestre, así como promedios de temperatura máxima, temperatura mínima y velocidad del viento.

Las variables multiescales fueron integradas directamente como canales adicionales de entrada al modelo, permitiendo incorporar información de memoria larga de forma explícita, sin necesidad de aumentar la longitud de la secuencia temporal principal utilizada por la CNN.

3.7. Evaluación de desempeño

La evaluación del desempeño de los modelos se realizó mediante tres métricas complementarias, que permiten evaluar distintos aspectos del error de simulación. La métrica principal seleccionada fue el Kling-Gupta Efficiency (KGE) (Gupta et al., 2009), ampliamente utilizada en hidrología por su capacidad de descomponer el error en componentes interpretables, lo que permite una mejor comprensión del ajuste logrado por los modelos. Como apoyo, también se utilizaron el Error Absoluto Medio (MAE) (Willmott, C. J., 2005) y la Raíz del Error Cuadrático Medio (RMSE) (Chai & Draxler, 2014), con el objetivo de tener una visión más completa del desempeño de cada modelo.

Kling-Gupta Efficiency (KGE)

Propuesto por Gupta et al. (2009), busca superar las limitaciones del clásico coeficiente de eficiencia de Nash-Sutcliffe (NSE), particularmente su sensibilidad al sesgo y su tendencia a subestimar la variabilidad del caudal. A diferencia del NSE, el KGE permite descomponer el error en tres componentes, la correlación lineal entre los valores observados y predichos (r), el sesgo medio (β), y la dispersión relativa (α), facilitando así una evaluación más robusta y transparente.

La expresión general de la KGE se define como:

$$KGE = 1 - \sqrt{(r - 1)^2 + (\alpha - 1)^2 + (\beta - 1)^2} \quad (10)$$

Valores de KGE cercanos a 1 indican una predicción precisa y equilibrada, mientras que valores negativos sugieren una predicción menos fiable.

Mean Absolute Error (MAE)

El MAE cuantifica el error promedio en términos absolutos entre los valores observados y los simulados. Es una métrica intuitiva y robusta frente a valores atípicos, útil para interpretar el error promedio en unidades físicas [m^3/s]. Se calcula como:

$$MAE = \frac{1}{n} \sum_{i=1}^n |Q_{observado\ i} - Q_{simulado\ i}| \quad (11)$$

Root Mean Squared Error (RMSE)

El RMSE penaliza de forma más severa los errores grandes al elevar las diferencias al cuadrado antes de promediarlas. Esto lo convierte en una métrica sensible a outliers, útil para detectar eventos extremos mal representados por el modelo.

Se calcula de la siguiente manera:

$$RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^n (Q_{observado\ i} - Q_{simulado\ i})^2} \quad (12)$$

3.8. Análisis de interpretabilidad y sensibilidad modelos de ML

Más allá del desempeño predictivo alcanzado por los modelos de ML, resulta fundamental analizar cómo estos representan las relaciones entre las variables meteorológicas y la respuesta hidrológica de la cuenca. No solo interesa evaluar la capacidad de los modelos para reproducir correctamente el caudal, sino también examinar si las relaciones aprendidas presentan coherencia con el comportamiento físico esperado del sistema hidrológico. Con este objetivo, se incorporó un análisis de interpretabilidad y sensibilidad orientado a estudiar de qué manera cada arquitectura construye sus predicciones.

El análisis se desarrolló en dos niveles complementarios. En primer lugar, se estimó la importancia global de las variables mediante un enfoque de permutación (Fisher et al., 2019). Posteriormente, para las variables más influyentes, se exploró su efecto detallado sobre la predicción utilizando Partial Dependence Plots (PDP) (Friedman, 2001) e Individual Conditional Expectation (ICE) (Goldstein et al., 2014). Finalmente, se incorporó un análisis de sensibilidad mediante perturbaciones controladas, con el fin de evaluar la estabilidad del modelo ante cambios sistemáticos en las variables con mayor peso en los modelos.

3.8.1. Importancia global de variables mediante permutación

La importancia por permutación permite cuantificar cuál es la importancia o peso de cada variable de entrada en la obtención de la variable objetivo, evaluando la degradación del error al alterar artificialmente la información contenida en dicha variable. Dicho en otras palabras, si una variable es relevante, alterar su información debería deteriorar su capacidad predictiva (Fisher et al., 2019). En esta etapa, la métrica utilizada fue el **RMSE** calculado sobre el conjunto de prueba final.

Sea " $RMSE_{base}$ " el error cuadrático del modelo sin perturbación, y " $RMSE_j^{perm}$ " el error obtenido tras permutar la variable j . La importancia absoluta que tendrá cada variable se define mediante la expresión:

$$\Delta RMSE_j = RMSE_j^{perm} - RMSE_{base} \quad (13)$$

En el caso de los modelos neuronales (LSTM y CNN), la permutación se realizó para su ventana temporal completa, considerando cada ventana temporal como una unidad, y dichas ventanas fueron reordenadas aleatoriamente entre muestras, sin modificar el orden interno de los valores dentro de cada secuencia. Esto implica que no se desplaza la serie en el tiempo ni se altera la dependencia temporal interna, pero sí se rompe la correspondencia entre cada secuencia y su instante temporal original, generando combinaciones inconsistentes que permiten evaluar la importancia de la variable analizada.

Para facilitar la comparación entre variables, las importancias se normalizaron y expresaron en porcentaje según:

$$I_j = \frac{\Delta RMSE_j}{\sum_{k=1}^p \Delta RMSE_k} * 100 \quad (14)$$

La expresión anterior define la “importancia relativa” de cada variable j como la proporción de la degradación total del desempeño del modelo atribuible a dicha variable. Asimismo, $\Delta RMSE_j$ cuantifica el aumento del error al romper la relación entre la variable j y el caudal, mientras que el denominador representa la suma de las degradaciones del error asociadas a la permutación de todas las variables de entrada. De este modo, el índice I_j permite comparar directamente la contribución relativa de cada variable al desempeño predictivo del modelo, independientemente de la escala absoluta del error.

3.8.2. Análisis PDP + ICE para variables dominantes

Con el fin de profundizar en la interpretación del comportamiento de los modelos, se desarrolló un análisis adicional mediante “Partial Dependence Plots (PDP) (Friedman, 2001) e Individual Conditional Expectation (ICE) (Goldstein et al., 2014)” para las tres variables con mayor importancia global identificadas en la etapa anterior. La curva PDP representa el efecto promedio de una variable x_j sobre la predicción del caudal $\hat{y}$, manteniendo constantes el resto de las variables de entrada:

$$PDP_j(z) = E_{x_{-j}} * [f(z, x_{-j})] \quad (15)$$

donde $f()$ corresponde al modelo entrenado, la variable x_j corresponde a la variable de interés cuyo efecto se desea analizar, mientras que z denota un valor específico dentro de su rango de variación. El término x_{-j} agrupa al resto de las variables de entrada distintas de x_j , las cuales se consideran según su distribución observada en el conjunto de datos. El operador de esperanza $E_{x_{-j}}[]$ indica que la respuesta del modelo se promedia respecto a todas las configuraciones posibles de dichas variables, permitiendo así estimar el efecto promedio de la variable analizada sobre la predicción del caudal.

Dado que los modelos operan sobre ventanas temporales, el valor representativo de cada variable dentro de la ventana se resumió mediante funciones físicas coherentes con su naturaleza, tales como el **promedio temporal** o el **acumulado**, según correspondiera. A partir de estos valores se construyó una grilla de evaluación basada en percentiles (5–95 %), con el fin de evitar la influencia de valores extre-

mos.

Las curvas ICE se calcularon para un subconjunto aleatorio de trayectorias, permitiendo evaluar la variabilidad de la respuesta del modelo bajo distintas condiciones hidrológicas. El PDP final se estimó como la mediana de las trayectorias ICE, lo que reduce la sensibilidad a valores atípicos y mejora la estabilidad interpretativa.

El objetivo de este enfoque es traducir el comportamiento interno del modelo, que opera como una función no lineal, en relaciones interpretables entre la variable de entrada y la predicción del caudal. Mientras que la permutación indica qué variables son relevantes en las arquitecturas de los modelos, las curvas PDP+ICE permiten comprender cómo actúan dichas variables sobre las respuestas del modelo, permitiendo identificar la forma funcional de la relación (lineal, no lineal, convexa, saturante), la existencia de umbrales críticos, regiones de mayor o menor sensibilidad, y posibles cambios de régimen en la respuesta.

3.8.3. Análisis de sensibilidad mediante perturbaciones proporcionales ($\pm 10\%$)

Como última etapa de este análisis, utilizando nuevamente las tres variables con mayor importancia de cada modelo, se realizó un estudio de sensibilidad local mediante perturbaciones de variables. Para cada modelo de ML, se modificó individualmente cada variable dominante en $\pm 10\%$ respecto a su valor original dentro de las ventanas temporales, y el resto de las variables de entrada se mantienen constantes.

Posteriormente, se evaluó el cambio promedio en el caudal predicho respecto al escenario base, definido como la predicción sin perturbaciones. La variación en el caudal se cuantificó como:

$$\Delta Q = Q_{\pm 10\%} - Q_{base} \quad (16)$$

En donde $Q_{\pm 10\%}$ corresponde al caudal medio sujeto bajo la perturbación, el Q_{base} al caudal medio predicho en condiciones originales.

Es importante destacar que los modelos no fueron reentrenados tras aplicar las perturbaciones. Las modificaciones se realizaron exclusivamente sobre los datos de entrada del conjunto de evaluación, manteniendo fijos los parámetros previamente calibrados del modelo. De este modo, el análisis refleja la sensibilidad de la función aprendida por el modelo ante variaciones controladas de las variables meteorológicas, sin alterar su estructura ni sus pesos internos.

4. Resultados

4.1. Configuración final modelos de ML

A continuación, se presentan las configuraciones finales seleccionadas para cada modelo de ML. Estas arquitecturas corresponden a las combinaciones que mostraron el mejor desempeño durante el proceso de ajuste y validación interna, luego de múltiples pruebas y modificaciones en sus configuraciones.

4.1.1. Configuración RF

Para la arquitectura de RF, el proceso de ajuste de hiperparámetros mediante la función de búsqueda “GridSearchCV” de la librería scikit-learn, se seleccionó la configuración que entregó el mejor desempeño evaluado en el conjunto de validación interna. El modelo final resultó compuesto por 100 árboles de decisión, sin restricción en su profundidad, y con reglas que controlan el número mínimo de muestras necesarias para dividir nodos o formar hojas. Además, se utilizó una fracción aleatoria de predictores en cada división (max_features igual a log2) y muestreo con reemplazo para aumentar la diversidad del conjunto.

Tabla 2. Resultados de la mejor iteración de parámetros de la arquitectura de RF.

Hiperparámetros RF	valor
N_estimators	100
Max_depth	None.
Min_samples_split	2
Min_samples_leaf	4
Max_features	Log2
Bootstrap	True

La profundidad seleccionada evitó que el modelo generara particiones excesivamente específicas sobre los datos de entrada, manteniendo una representación más general de las relaciones entre las variables meteorológicas y la respuesta hidrológica de la cuenca. Por otra parte, el uso de 100 árboles permitió estabilizar las predicciones del ensamble, reduciendo la variabilidad asociada a árboles individuales y generando una respuesta temporal más suave en la simulación del caudal.

Durante las primeras pruebas, cuando se utilizaron únicamente las variables meteorológicas bases, el comportamiento de RF mostró dificultades para representar de manera consistente la dinámica del caudal. Las configuraciones obtenidas resultaban inestables y, además, muy sensibles a pequeños cambios en los hiperparámetros. En cierto modo, el modelo reaccionaba bien a condiciones puntuales, pero le costaba capturar la continuidad propia del comportamiento hidrológico de la cuenca. Esto afirmaba que la información instantánea, por sí sola, no bastaba para describir adecuadamente la persistencia del sistema. La incorporación de variables acumuladas y rezagadas permitió entonces añadir una representación antecedente de las condiciones meteorológicas e hidrológicas, entregándole al modelo una visión temporal más completa. Como resultado, mejoró la estabilidad de las arquitecturas obtenidas, y también una mejor capacidad para representar la variabilidad temporal del caudal.

4.1.2. Configuración LSTM

La configuración final de LSTM incorpora una secuencia de entrada de 180 días consecutivos y un conjunto adicional de variables agregadas descritas en la metodología, fueron calculadas sobre una ventana de 185 días, integradas como entradas complementarias al modelo. De esta manera, la arquitectura final combina una memoria secuencial explícita a través de la LSTM (180 días), e información agregada de escala similar (185 días), incorporada en la base de datos como entrada al modelo.

Tabla 3. Resultados de la mejor iteración de parámetros de la arquitectura de LSTM.

Hiperparámetros LSTM	valor
Hidden_size	4
Num_layers	2
Seq_length	180
dropout	0.1
Learning_rate	0.001

Durante el periodo de entrenamiento se observó que ventanas más extensas (por ejemplo, superiores a 250 días) tendían a reducir la estabilidad del modelo y no mejoraban la capacidad predictiva. Esto evidencia que, para la cuenca analizada, el modelo se ve perjudicado al tener una memoria muy larga de días y no necesariamente una arquitectura más densa va a entregar el mejor rendimiento y va a aportar más información relevante. Por esta razón, la ventana de tiempo mediante la incorporación de variables rezagadas se convierte en un nuevo parámetro del modelo a encontrar.

El valor de esta ventana externa se obtuvo en la misma etapa de iteración en que se determinó la longitud óptima de la secuencia LSTM, funcionando como una extensión estadística de la ventana de memoria. Mientras los 180 días permiten al modelo capturar dependencias secuenciales explícitas, los 185 días agregados resumen tendencias acumulativas que no necesariamente requieren ser aprendidas exclusivamente por la secuencia temporal interna del modelo. Sin embargo, este enfoque no trata de reemplazar la capacidad de memoria del LSTM, sino de organizar la información temporal de modo que el modelo no dependa exclusivamente de su estado interno para representar procesos acumulativos. Esta decisión reduce la carga sobre la red y define con mayor claridad la escala temporal de los procesos que se busca modelar.

El número de neuronas (Hidden size) define la dimensionalidad del estado interno del modelo y, en consecuencia, la cantidad de información temporal que puede retener durante el aprendizaje. En este caso, la configuración seleccionada corresponde a una arquitectura relativamente compacta, indicando que la dinámica hidrológica de la cuenca podía representarse adecuadamente sin recurrir a estados internos demasiado extensos. Durante el proceso iterativo se observó que configuraciones con mayor número de neuronas aumentaban la variabilidad entre ejecuciones sin mejorar de manera consistente las métricas de validación. Esto deja ver que, para esta cuenca, incorporar más capacidad al modelo no necesariamente aportaba información útil, sino que tendía a introducir arquitecturas más inestables y sensibles al ajuste.

La utilización de dos capas recurrentes (Num_layers) permitió organizar la información temporal en distintos niveles. Mientras la primera capa capturaba la evolución inmediata de las forzantes meteorológicas, la segunda trabajaba sobre patrones ya transformados, permitiendo representar relaciones temporales más complejas sin depender únicamente de un aumento en la dimensionalidad del estado interno.

4.1.3. Configuración CNN

La configuración final del modelo CNN se presenta en la Tabla 4. La arquitectura seleccionada quedó definida por una ventana de entrada de 224 días, dos capas convolucionales con 32 y 64 filtros respectivamente, un tamaño de kernel de 7 días, y una capa densa final para la predicción del caudal.

Tabla 4. Resultados de la arquitectura más óptima de CNN.

Hiperparámetros CNN	Valor
Kernel_sizes	7
Filtro_1	32
Filtro_2	64
dropout	0.1
Learning_rate	0.001
Seq_length	224

La ventana de 224 días define el horizonte temporal total sobre el cual el modelo recibe información secuencial. A diferencia de la red LSTM, que procesa esta secuencia de forma recurrente mediante un estado interno, la CNN analiza la secuencia completa aplicando filtros que recorren la señal temporal.

El tamaño de kernel igual a 7 implica que cada filtro opera sobre sub-ventana semanal dentro de los 224 días. En términos prácticos, la red va desplazando esta ventana de 7 días a lo largo de toda la secuencia y extrayendo patrones locales asociados a combinaciones de precipitación, temperatura, viento o de cualquier relación entre las forzantes de entrada que puedan ser relevantes para la generación de caudal. Esto permite identificar estructuras temporales de corta escala, por ejemplo, acumulaciones breves de precipitación o cambios térmicos sostenidos por algunos días antes de combinarlas en representaciones más complejas en las capas siguientes.

La arquitectura final seleccionada mostró una capacidad consistente para detectar patrones temporales de distinta escala dentro de las series meteorológicas. La primera capa convolucional, compuesta por 32 filtros, permitió capturar patrones elementales de corto plazo asociados a cambios recientes en las forzantes meteorológicas. Cada filtro aprendió configuraciones temporales diferentes dentro de una ventana de 7 días, mientras que la segunda capa convolucional reorganizaba esas respuestas iniciales en representaciones temporales más complejas. En la práctica, esto permitió que el modelo reconociera relaciones temporales de distinta escala sin depender de una memoria acumulativa explícita.

Además, se incorporaron variables agregadas calculadas sobre una ventana de 110 días, cuyo valor surgió directamente del proceso iterativo de ajuste. Estas variables ayudaron al modelo a incorporar información de evolución más lenta y acumulativa, asociada principalmente a la persistencia de condiciones meteorológicas y a procesos de derretimiento nival y glaciar. A diferencia de LSTM, donde la memoria temporal se construye progresivamente a lo largo de la secuencia, en la CNN la representación temporal surge principalmente de la combinación jerárquica de patrones detectados por los filtros convolucionales. Por esta razón, las variables agregadas de 110 días no reemplazan el rol de la convolución, sino que complementan la representación temporal entregando información acumulativa que difícilmente podría capturarse únicamente mediante relaciones locales de corto plazo.

4.2. Desempeño comparativo

Más allá de las diferencias particulares observadas en cada arquitectura, el análisis comparativo global permitió evaluar cómo los distintos modelos representaron la dinámica hidrológica de la cuenca a lo largo de todo el período de validación final. La comparación no se limitó únicamente a métricas globales de desempeño, también consideró la capacidad de cada enfoque para reproducir la variabilidad temporal del caudal, la forma general del hidrograma y el comportamiento de la distribución de cau-

dales en distintos niveles de excedencia. Bajo esta perspectiva, las arquitecturas de Machine Learning (RF, LSTM y CNN) fueron evaluadas junto al modelo hidrológico USM y su implementación en Python.

4.2.1. Métricas globales en periodo de prueba final

Tal como se observa en la Tabla 5, al evaluar el desempeño sobre el conjunto completo del período de testing final, el modelo LSTM alcanza el mejor desempeño con un KGE de 0.80, mientras que RF y CNN presentan también un muy buen desempeño, con valores de 0.72. Los modelos Hidrológico USM, tanto la variante implementada por Sanllehi como la adaptación a Python presentan desempeños más regulares, con valores de KGE de 0.54 y 0.65 respectivamente.

Tabla 5. Resultado de métricas globales de modelos de ML y modelos base en el conjunto de testing final.

Modelo	KGE Global	r	α	β	RMSE	MAE
LSTM	0.8	0.8142	0.9738	1.0554	2.55	1.59
RF	0.72	0.8278	0.7844	1.0171	2.37	1.39
CNN	0.72	0.7701	0.9779	1.1758	2.95	1.89
Modelo de Python	0.65	0.8023	1.2523	1.1351	3.21	2.03
Modelo Hidrológico USM	0.54	0.6136	1.2323	1.0731	4.23	2.73

El mejor desempeño de LSTM se explica por un equilibrio más consistente entre correlación, variabilidad y sesgo ($r = 0.814$, $\alpha = 0.974$ y $\beta = 1.055$), permitiendo representar de forma más estable tanto la dinámica temporal como la magnitud y dispersión de los caudales simulados. El comportamiento de RF fue distinto, mostrando la mayor correlación ($r = 0.828$), pero acompañado de una subestimación importante de la variabilidad ($\alpha = 0.784$), lo que se refleja en una tendencia a suavizar los extremos del hidrograma. Por otra parte, CNN logra representar adecuadamente la variabilidad general del caudal ($\alpha = 0.978$), aunque con una menor correlación y una mayor sobreestimación promedio ($\beta = 1.176$), afectando su desempeño global.

Al analizar las métricas de error absoluto se observan diferencias adicionales entre las arquitecturas. RF presenta el menor RMSE (2.37), indicando menores desviaciones en eventos de gran magnitud, mientras que LSTM, pese a obtener un RMSE levemente superior (2.55), mantiene el mejor equilibrio general reflejado en el KGE. En el caso de CNN, presenta los mayores valores de RMSE y MAE (2.95 y 1.89 respectivamente), lo cual evidencian una respuesta más sensible frente a eventos de alta variabilidad y una mayor dispersión de los errores a lo largo de la serie temporal evaluada.

Tanto el modelo hidrológico USM original como el modelo de Python presentan valores de RMSE y MAE superiores a los modelos de ML. Aun así, el modelo adaptado a Python logra mantener un KGE global competitivo y cercano al de RF y CNN, mostrando que la estructura hidrológica general del régimen sigue siendo representada de manera razonable. Las principales diferencias aparecen en la representación de la magnitud y variabilidad de los caudales, donde ambos modelos físicos tienden a perder consistencia frente a cambios más marcados en la dinámica del sistema.

4.2.2. Hidrograma comparativo enfoques ML vs Enfoques físicos

Luego del análisis de las métricas KGE y de sus distintos componentes, resulta necesario observar cómo esas diferencias se traducen en la representación temporal del caudal. Después de todo, un buen desempeño estadístico no garantiza necesariamente que el modelo reproduzca de forma adecuada la

dinámica hidrológica de la cuenca. Por esta razón, el análisis del hidrograma comparativo adquiere un papel central en esta etapa de evaluación.

Esta representación permite examinar de manera directa cómo cada modelo reproduce la evolución temporal del caudal durante el período de validación final. Más allá de los valores agregados de desempeño, el hidrograma revela la forma en que cada arquitectura responde frente a crecidas, recesiones y variaciones en la magnitud del caudal en comparación con los valores observados. Así, es posible identificar diferencias en el comportamiento temporal de las simulaciones que las métricas globales, por sí solas, no siempre logran capturar completamente.

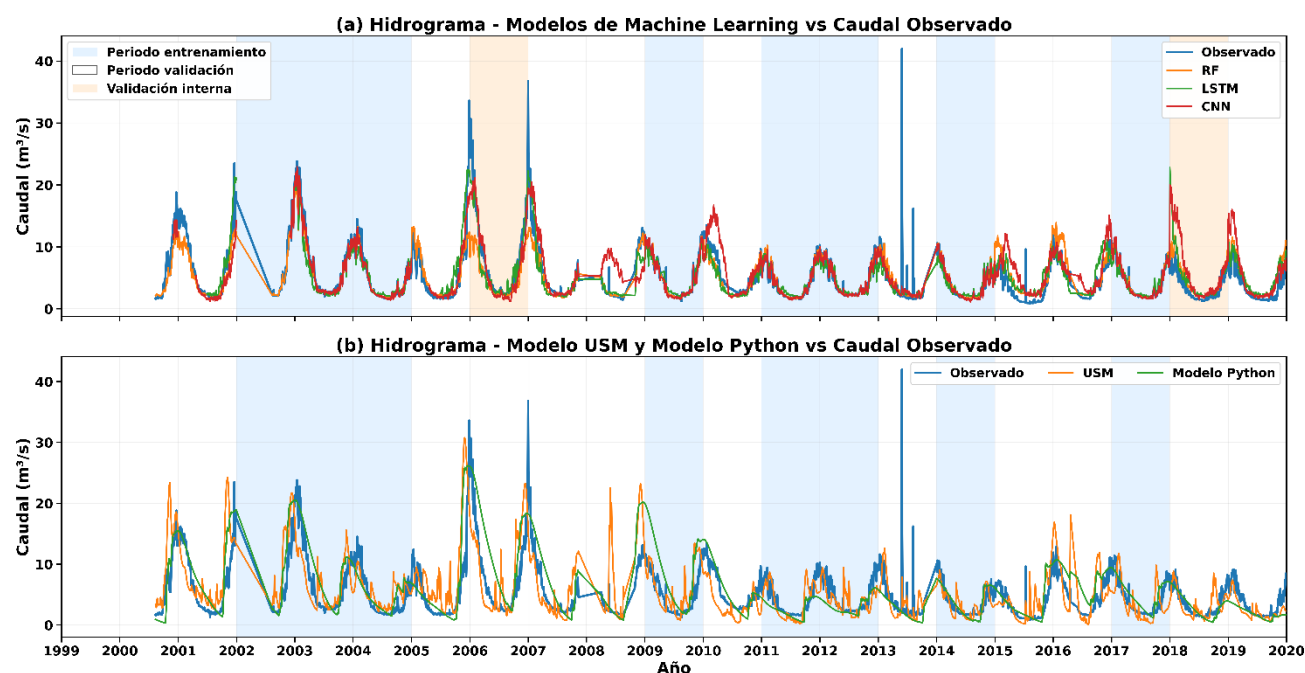


Figura 5. Comparación entre caudales observados y simulados por (a) modelos de ML (LSTM, RF y CNN) y (b) modelo hidrológico USM calibrada por Sanllehi y la adaptación a Python, durante el período de validación final.

Desde una perspectiva global, todos los modelos logran reproducir la estacionalidad de la cuenca, caracterizada por un aumento de caudal durante el período de deshielo y recesiones posteriores. Sin embargo, se observan diferencias relevantes en la forma en que cada modelo representa tanto la magnitud como la sincronización temporal de los eventos. En términos de capacidad para seguir la evolución de los caudales en el tiempo, el modelo LSTM muestra el comportamiento más consistente, lo que es coherente con su alto valor de correlación observado previamente. A lo largo de la mayor parte del período, los caudales simulados siguen adecuadamente tanto la forma como la magnitud de la serie observada, sin evidenciar desviaciones sistemáticas muy marcadas. Por otro lado, el modelo RF también reproduce correctamente la forma general del hidrograma, pero presenta una menor amplitud en algunos eventos de crecida, lo que se traduce en una atenuación de los máximos respecto a la serie de tiempo observada, coherente con lo evidenciado en las métricas de desempeño.

Por otra parte, CNN presenta un comportamiento particular, aunque logra reproducir la forma general del hidrograma, muestra una tendencia a sobreestimar los caudales, la cual resulta especialmente evidente en los años 2008 y 2010, donde los caudales simulados se mantienen de forma persistente por encima de los valores observados durante períodos prolongados, incluso cuando no se registran crecidas de magnitud comparable en la serie observada. En comparación con los demás modelos de ML, la CNN presenta una tendencia más pronunciada a sobreestimar los caudales.

Al comparar el modelo USM y el modelo de Python, se observa que ambos logran representar adecuadamente las características generales del hidrograma de la cuenca, particularmente el patrón de crecidas y recesión. Sin embargo, aparecen diferencias claras en la forma en que se representa la magnitud de los eventos. En el caso del modelo USM, se aprecia una mayor variabilidad, con situaciones en las que los caudales simulados sobreestiman los valores observados y otras en que los subestiman, mostrando un comportamiento poco uniforme entre eventos. Por su parte, el modelo en Python presenta una respuesta algo más suavizada, lo que también se aprecia en el hidrograma, pero igualmente mantiene una dispersión importante en la magnitud de los caudales, alternando entre sobreestimaciones y subestimaciones. En términos de sincronización, no se observa un desfase sistemático a lo largo de la serie, aunque sí existen desalineaciones puntuales en algunos eventos, esto deja en evidencia que las principales limitaciones de ambos modelos están más asociadas a la representación de la magnitud y la variabilidad de los caudales que a la estructura temporal general del régimen.

4.2.3. Curva de duración de caudales

La curva de duración de caudales ofrece una forma distinta y, en cierto modo, más reveladora de evaluar el desempeño de los modelos hidrológicos. Mientras el hidrograma muestra cómo varía el caudal a lo largo del tiempo, esta representación se enfoca en la capacidad del modelo para reproducir la distribución completa de los caudales y su comportamiento frente a distintos niveles de excedencia.

No basta con evaluar si un modelo reproduce adecuadamente una crecida puntual. También es necesario analizar su capacidad para representar, de manera consistente, los caudales altos, medios y bajos a lo largo de toda la serie. La curva de duración permite identificar cómo responde cada modelo en los distintos rangos de caudal, revelando fortalezas y limitaciones que el análisis temporal, por sí solo, puede pasar por alto. Es común que un modelo represente con precisión las crecidas, pero presente mayores diferencias durante los períodos de estiaje, o que ocurra la situación inversa. Por ello, la curva de duración complementa la información proporcionada por el hidrograma y ofrece una evaluación más completa del desempeño del modelo en todo el rango de caudales.

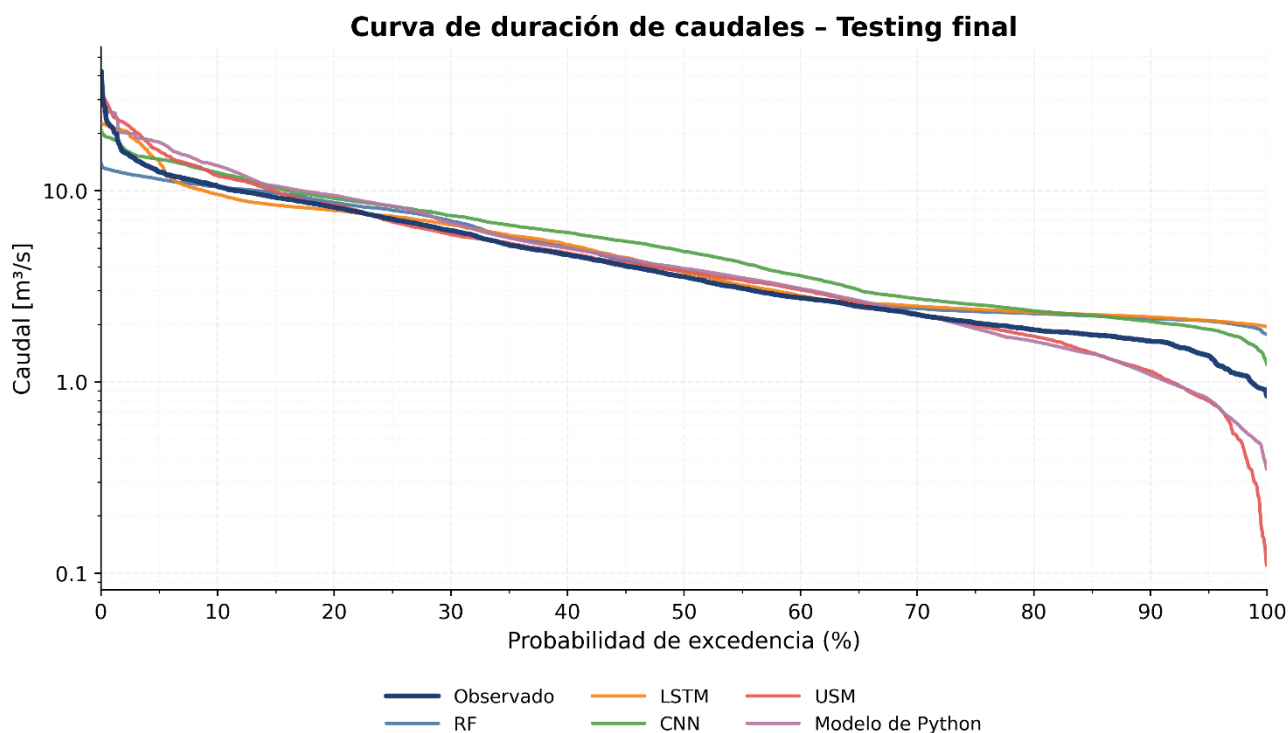


Figura 6. Curvas de duración de caudales observados y simulados por los modelos evaluados durante el período de testing final.

En el rango de bajas probabilidades de excedencia, correspondiente a los eventos de mayor magnitud, los modelos LSTM, CNN y RF muestran un comportamiento bastante similar y logran reproducir de forma adecuada la pendiente general de la curva observada. En este tramo no se aprecian diferencias marcadas entre los modelos, aunque sí aparecen pequeñas variaciones en la amplitud de los máximos. En algunos casos, los picos extremos aparecen ligeramente más suavizados o sobrestimados que en los registros observados, lo que indica que ciertos modelos tienden a reducir parte de la variabilidad presente en los eventos más intensos.

En el rango intermedio de excedencia, asociado al régimen dominante de la cuenca, todos los modelos siguen bastante bien la forma de la curva observada. La principal excepción es el modelo CNN, que en ciertos sectores tiende a amplificar los caudales simulados. Aun así, tanto los enfoques de aprendizaje automático como los modelos de base física representan adecuadamente la señal hidrológica principal del sistema, fuertemente controlada por el deshielo estacional. Las diferencias en este rango son relativamente menores y no modifican de manera significativa la distribución general de los caudales medios.

Las discrepancias más importantes aparecen en el extremo de altas probabilidades de excedencia, asociado a los caudales bajos. En esta zona, el modelo USM presenta una desviación más evidente respecto de la curva observada, mostrando una pendiente más pronunciada y una subestimación de los valores mínimos. Algo similar ocurre con el Modelo de Python, cuya curva desciende con mayor intensidad en el tramo final y se aleja de la forma observada. Este comportamiento sugiere dificultades para representar adecuadamente los procesos de almacenamiento y liberación gradual de agua durante los períodos de estiaje. En cuanto a los modelos de ML, se mantienen sistemáticamente por sobre la curva observada en el tramo entre 90 y 100 % de excedencia, lo que evidencia una tendencia a sobreestimar los caudales de estiaje. Esto indica que, aunque estos modelos conservan de buena manera la forma general del régimen hidrológico en gran parte del rango analizado, todavía presentan limitaciones al momento de reproducir con precisión la cola asociada a los caudales mínimos.

4.3. Análisis desempeño en años secos, normales y húmedos

El análisis por régimen hidrológico complementa la evaluación global del desempeño de los modelos, ya que su capacidad de representación puede variar en función de las condiciones hidrológicas de la zona. En cuencas de alta montaña, donde el régimen de caudales está condicionado por la interacción entre la acumulación de nieve, el derretimiento y la precipitación líquida, es esperable que el desempeño de los modelos varíe según el contexto climático predominante. Por ello, el análisis desagregado no solo permite identificar fortalezas y debilidades relativas de cada enfoque, sino también examinar si los modelos mantienen coherencia en la representación de procesos tanto en condiciones de déficit como de abundancia hídrica.

De acuerdo con la clasificación definida en la Sección 3.5, se analizaron los siguientes años representativos:

- Años secos: 2008, 2013 y 2015.
- Años normales: 2007, 2016 y 2019.
- Años húmedos: 2001, 2005 y 2010.

Para cada año y modelo, se analizan los resultados obtenidos a partir de los valores de KGE, sus componentes (r , α y β), además de RMSE y MAE. A continuación, se analiza el comportamiento dentro de cada régimen hidrológico.

4.3.1. Análisis años secos

Los años secos se caracterizan por presentar caudales medios más bajos y una menor variabilidad intraanual, predominando períodos prolongados de recesión y una presencia mucho más limitada de crecidas significativas. Bajo estas condiciones, la atención ya no se centra tanto en la capacidad del modelo para capturar eventos extremos, sino en qué tan bien representa el régimen base y la evolución gradual del caudal a lo largo del tiempo. Resulta especialmente importante evaluar si los modelos logran reproducir las recesiones de manera estable, sin introducir oscilaciones artificiales ni pérdidas excesivas de variabilidad. Pequeñas desviaciones en este tipo de escenarios pueden alterar de forma importante la representación del comportamiento hidrológico dominante durante períodos secos.

Tabla 6. Métricas de desempeño anual para los modelos evaluados durante los años secos del período de testing final.

Año	Modelo	KGE	r	α	β	RMSE	MAE
2008	RF	0.895	0.987	0.903	0.964	0.725	0.475
	LSTM	0.613	0.957	0.642	0.859	1.783	1.158
	CNN	-0.648	-0.492	0.418	1.39	5.2	5.002
	Modelo USM	-0.325	0.744	1.989	1.844	6.82	4.662
	Modelo de Python	-0.102	0.988	1.977	1.509	4.581	2.891
2013	RF	0.388	0.52	0.621	0.99	3.678	1.154
	LSTM	0.214	0.536	0.393	0.814	3.746	1.278
	CNN	0.329	0.55	0.511	0.91	3.595	1.021
	Modelo USM	0.414	0.496	0.715	0.907	3.849	2.129
	Modelo de Python	0.115	0.509	0.388	0.592	4.107	1.812
2015	RF	0.399	0.919	1.235	1.548	2.183	1.902
	LSTM	0.538	0.878	0.896	1.433	1.849	1.656
	CNN	0.215	0.655	1.116	1.696	3.165	2.584
	Modelo USM	0.636	0.684	1.147	1.105	2.184	1.678
	Modelo de Python	0.721	0.794	1.153	1.11	1.799	1.159

En 2008 se observa uno de los contrastes más claros entre los modelos evaluados. El modelo RF presenta el mejor desempeño del año, con un KGE de 0.895, acompañado de una alta correlación ($r = 0.987$) y valores de α (0.903) y β (0.964) cercanos a 1. Este comportamiento refleja una representación equilibrada tanto de la variabilidad como del volumen medio, lo que también se traduce en los menores errores del conjunto (RMSE = 0.725; MAE = 0.475). Por su parte, el modelo LSTM mantiene una correlación elevada ($r = 0.957$), pero su menor valor de α (0.642) evidencia una subestimación importante de la variabilidad, lo que explica su desempeño inferior y el aumento en sus errores (RMSE = 1.783).

El caso de la CNN es distinto. Su KGE negativo (-0.648) responde a una combinación de correlación

negativa ($r = -0.492$), subestimación marcada de la variabilidad ($\alpha = 0.418$) y una sobreestimación del volumen medio ($\beta = 1.39$). En términos prácticos, esto implica que el modelo no solo pierde coherencia en la forma del hidrograma, sino que además representa una escala de caudales que se aleja a la observada.

En cuanto a los modelos hidrológicos, tanto el USM como su implementación en Python muestran valores elevados de α (1.989 y 1.977) y β (1.844 y 1.509), asociado a una sobreamplificación del régimen. A pesar de que el modelo en Python mantiene una correlación alta ($r = 0.988$), el deterioro del KGE en ambos casos se explica principalmente por distorsiones en la magnitud del caudal más que por problemas en la estructura temporal.

Los errores absolutos terminan de mostrar el mismo patrón. En la CNN y en el modelo hidrológico, los valores elevados de RMSE (5.2 y 6.82, respectivamente) nos dicen que las mayores diferencias aparecen concentradas en ciertos eventos donde el caudal es claramente sobredimensionado. Más que pequeñas discrepancias, se observan episodios donde la simulación se separa con bastante fuerza del comportamiento observado. El MAE, que también se mantiene alto en ambos casos, apunta además a un problema más persistente. La desviación no ocurre solo durante algunos picos extremos, sino que acompaña buena parte de la evolución anual del caudal, indicando un sesgo más sostenido en la representación del régimen hidrológico.

En 2013, el comportamiento de los modelos cambia de forma evidente. Ninguno alcanza un desempeño alto y las diferencias entre ellos se explican principalmente por la representación de la magnitud de los caudales más que por la correlación temporal. En términos generales, los modelos logran reproducir parcialmente la dinámica de la serie, pero presentan dificultades para capturar la amplitud real de las variaciones observadas. El modelo RF muestra una subestimación de la variabilidad ($\alpha = 0.621$), mientras que LSTM ($\alpha = 0.393$) y CNN ($\alpha = 0.511$) acentúan esta tendencia, evidenciando una respuesta más suavizada del régimen hidrológico, lo que genera que los caudales aparezcan amortiguados respecto a la serie observada, reduciendo la dispersión de los caudales simulados.

En este escenario, el modelo hidrológico USM logra el mejor desempeño relativo del año ($KGE = 0.414$), con una representación más estable del volumen ($\alpha = 0.715$; $\beta = 0.907$). Sin embargo, su baja correlación ($r = 0.496$) indica que, aunque la magnitud es razonable, la dinámica temporal no es bien capturada. Por el contrario, el modelo en Python presenta el menor desempeño ($KGE = 0.115$), asociado a una subestimación tanto de la variabilidad ($\alpha = 0.388$) como del caudal medio ($\beta = 0.592$).

Los valores de RMSE observados en este año están fuertemente influenciados por un evento de gran magnitud, superior a 40 m³/s, que domina el error cuadrático. Dado que el RMSE penaliza de forma desproporcionada los errores asociados a eventos extremos, este episodio puntual tiene como consecuencia la amplificación de la magnitud, incluso cuando el comportamiento general del hidrograma es razonablemente representado durante el resto del año. En contraste, el MAE menos sensible a valores extremos permite observar que el error promedio diario, aunque elevado, no difiere tan abruptamente entre modelos. Esto explica que el USM, pese a alcanzar el mayor KGE del año por equilibrio relativo entre componentes, no necesariamente presente una menor magnitud de error en términos absolutos.

El año 2015 introduce un comportamiento distinto dentro del régimen seco. En este caso, el modelo en Python alcanza el mejor desempeño ($KGE = 0.721$), con un equilibrio adecuado entre variabilidad ($\alpha = 1.153$) y volumen ($\beta = 1.11$). El modelo hidrológico USM también muestra un buen ajuste ($KGE = 0.636$), aunque con una leve tendencia a amplificar la variabilidad ($\alpha = 1.147$). Entre los modelos de ML, la red LSTM ($KGE = 0.538$) mantiene un comportamiento relativamente equilibrado ($\alpha = 0.896$; $\beta = 1.433$), mientras que RF ($KGE = 0.399$), a pesar de tener una muy buena correlación ($r = 0.919$),

tiende a amplificar los picos de caudal ($\alpha = 1.235$) y a sobreestimar el volumen medio ($\beta = 1.548$). El modelo CNN en cambio, vuelve a presentar el desempeño más bajo dentro de este grupo ($KGE = 0.215$), debido a que presenta una sobreestimación importante del régimen ($\alpha = 1.116$; $\beta = 1.696$).

Los valores de RMSE y MAE en 2015 disminuyen respecto a 2013, reflejando un mejor ajuste de la estructura anual del caudal. Aun así, continúan apareciendo sesgos relacionados principalmente con la magnitud de los caudales simulados. El modelo en Python presenta el menor error promedio ($MAE = 1.159$), mostrando una representación diaria más estable y cercana a la observada, mientras que la CNN sigue registrando desviaciones comparativamente más altas.

En términos generales, los años secos concentran los desempeños más bajos y también los más variables entre los modelos evaluados, mostrando que este régimen representa uno de los escenarios más complejos de reproducir dentro de la serie analizada. Sin embargo, la disminución del KGE no responde a una única causa dominante. Más bien, cada año presenta limitaciones distintas según el comportamiento hidrológico y la forma en que cada modelo representa el caudal. En 2008, por ejemplo, los principales problemas se asocian a distorsiones de escala en los modelos hidrológicos y a una pérdida de coherencia temporal en CNN. En cambio, durante 2013, las limitaciones se relacionan principalmente con una marcada subestimación de la variabilidad del sistema. Para 2015, el deterioro del desempeño aparece vinculado sobre todo a una sobreestimación del volumen medio ($\beta > 1$), incluso en modelos que conservan correlaciones relativamente altas.

Además, no se observa una jerarquía estable entre arquitecturas dentro del régimen seco. Aunque RF alcanza el mejor desempeño promedio entre los modelos de ML y destaca claramente en el año 2008, esa ventaja no se mantiene de manera consistente en los demás años analizados. De forma similar, el modelo hidrológico USM presenta el mejor desempeño relativo en 2013, mientras que el modelo de Python lidera en 2015.

Los resultados obtenidos para este régimen demuestran que, bajo condiciones de baja disponibilidad hídrica, pequeñas diferencias en la representación de la magnitud y la variabilidad del caudal pueden tener un efecto considerable sobre el desempeño final. Como consecuencia, la estabilidad de cada enfoque cambia de forma importante entre distintos años secos, dificultando la existencia de un modelo consistentemente superior en este tipo de escenarios.

4.3.2. Análisis años normales

Los años clasificados como normales representan un escenario intermedio, donde el régimen hidrológico mantiene una señal estacional relativamente definida, pero sin alcanzar las condiciones extremas observadas en años secos o húmedos. En estos periodos conviven eventos de magnitud moderada con una dinámica más estable del caudal, permitiendo observar con mayor claridad las diferencias en el comportamiento de cada modelo frente a variaciones más habituales del sistema.

Tabla 7. Métricas de desempeño anual para los modelos evaluados durante los años normales del período de testing final.

Año	Modelo	KGE	r	α	β	RMSE	MAE
2007	RF	0.576	0.938	0.638	0.788	2.733	1.317
	LSTM	0.904	0.946	1.045	0.935	1.828	1.15
	CNN	0.802	0.914	1.115	1.137	2.52	1.476
	Modelo USM	0.588	0.715	0.73	0.875	3.754	2.677
	Modelo de Python	0.71	0.872	1.024	1.259	3.118	2.531
2016	RF	0.705	0.923	1.22	1.181	1.747	1.243
	LSTM	0.772	0.807	1.005	1.121	1.977	1.483
	CNN	0.721	0.862	1.092	1.224	2.053	1.491
	Modelo USM	0.355	0.614	1.507	1.1	3.612	2.786
	Modelo de Python	0.495	0.875	0.774	1.434	2.75	2.467
2019	RF	0.468	0.95	1.292	1.442	1.904	1.586
	LSTM	0.592	0.935	1.128	1.382	1.652	1.374
	CNN	0.056	0.919	1.796	1.5	2.838	1.839
	Modelo USM	0.516	0.808	0.697	0.675	1.799	1.342
	Modelo de Python	0.225	0.767	0.416	0.548	2.33	1.634

En 2007, el mejor desempeño entre los modelos de ML fue alcanzado por LSTM (KGE = 0.904). El modelo logró un equilibrio bastante sólido entre una alta correlación ($r = 0.946$), una representación muy cercana de la variabilidad observada ($\alpha = 1.045$) y un volumen simulado consistente con los registros reales ($\beta = 0.935$). A esto se suman los menores valores de RMSE y MAE, reflejando una reproducción más estable y precisa del caudal diario. La arquitectura de CNN también mostró un ajuste favorable (KGE = 0.802), aunque con una tendencia a sobreamplificar la variabilidad ($\alpha = 1.115$) y a sobreestimar el volumen total ($\beta = 1.137$). RF, en cambio, mantuvo una correlación alta ($r = 0.938$), pero presentó una subestimación marcada de la variabilidad ($\alpha = 0.638$), reduciendo así su desempeño global (KGE = 0.576).

En los modelos base, el modelo hidrológico USM presenta una menor correlación ($r = 0.715$) junto con una subestimación de la variabilidad ($\alpha = 0.73$). Por su parte, el modelo implementado en Python alcanza un desempeño competitivo (KGE = 0.71), con buena correlación ($r = 0.872$), aunque con una tendencia a sobreestimar el volumen ($\beta = 1.259$).

En 2016 el comportamiento de los modelos de ML es más homogéneo en términos de KGE, con valores cercanos entre LSTM (0.772), CNN (0.721) y RF (0.705). Sin embargo, al analizar sus componentes se observa que las diferencias no están en la correlación, que se mantiene alta en todos los casos, sino en la forma en que cada modelo representa la escala del sistema. Para el modelo RF, a pesar de su alta correlación ($r = 0.923$), tiende a amplificar tanto la variabilidad ($\alpha = 1.22$) como el volumen medio ($\beta = 1.181$). La red CNN presenta un comportamiento similar, con un aumento moderado de la variabilidad ($\alpha = 1.092$) y un mayor sesgo en el volumen ($\beta = 1.224$). En cambio, LSTM mantiene valores más cercanos a la unidad en términos de variabilidad ($\alpha = 1.005$), lo que implica una representación más estable de la dinámica anual, aunque presenta una leve sobreestimación del volumen ($\beta = 1.121$).

En los modelos físicos, las diferencias son más marcadas. El modelo USM exhibe el menor desempeño ($KGE = 0.355$), asociado a una amplificación excesiva de la variabilidad ($\alpha = 1.507$). El modelo de Python mejora parcialmente el ajuste ($KGE = 0.495$), pero aún presenta distorsiones relevantes en la escala ($\alpha = 0.774$; $\beta = 1.434$).

Los errores ayudan a entender mejor estas diferencias. Aunque los valores de KGE entre los modelos de ML son similares, el RMSE muestra que las principales discrepancias se concentran en eventos de mayor magnitud. En particular, el modelo USM combina el mayor RMSE con un MAE elevado (2.786), mostrando que el problema principal no está en la sincronización temporal, sino en la representación de la magnitud del caudal.

Para el año 2019 introduce una situación más contrastante, donde las diferencias entre modelos no se explican por la correlación, que se mantiene alta en varios casos, sino más bien por la forma en la que cada uno representa la magnitud de la señal hidrológica. RF mantiene buena correlación ($r = 0.95$), pero tiende a exagerar tanto la variabilidad ($\alpha = 1.292$) como el volumen medio ($\beta = 1.442$), lo que como consecuencia reduce su valor de KGE (0.468). LSTM logra un equilibrio mejor ($KGE = 0.592$), aunque también con α y β superiores a 1. En cuanto a CNN, aun pese a una correlación elevada ($r = 0.919$), presenta la mayor distorsión de escala entre los algoritmos de ML ($\alpha = 1.796$; $\beta = 1.5$), y esa diferencia termina explicando su bajo valor de KGE (0.056).

En los modelos base, el comportamiento resulta bastante más uniforme. El modelo hidrológico USM alcanza un desempeño intermedio ($KGE = 0.516$) y muestra una tendencia a subestimar tanto la variabilidad ($\alpha = 0.697$) como el volumen medio ($\beta = 0.675$). En la práctica, esto se traduce en una representación más conservadora del régimen hidrológico, con una señal menos intensa que la observada. El modelo de Python acentúa todavía más esa tendencia. Su KGE desciende a 0.225 y presenta una reducción más marcada en la amplitud de la señal ($\alpha = 0.416$) y en el caudal medio ($\beta = 0.548$), reflejando una simulación considerablemente más amortiguada del comportamiento hidrológico anual.

De manera general, los años normales presentan un comportamiento más estable que los años secos, lo que permite distinguir con mayor claridad las diferencias entre las arquitecturas evaluadas. En este régimen, es común observar correlaciones altas en varios modelos, sin embargo, los resultados muestran que una buena correlación no siempre se traduce en un buen desempeño global. El ajuste final depende en gran medida del equilibrio entre la variabilidad (α) y el volumen medio (β), de modo que pequeñas amplificaciones o atenuaciones sostenidas de la señal terminan afectando de forma importante el valor del KGE.

El año 2019 refleja bien esta situación, ya que, aunque varios modelos logran reproducir adecuadamente la evolución temporal del hidrograma, presentan dificultades al representar la magnitud real del régimen anual. En otras palabras, los modelos siguen correctamente la dinámica de la señal, pero exageran o reducen su intensidad, y esa diferencia termina penalizando el desempeño final. Así, el caso de 2019 muestra con bastante claridad cómo una amplificación excesiva de la señal puede deteriorar el KGE incluso cuando la correlación permanece elevada.

Al promediar el desempeño de los años normales, el modelo LSTM presenta el mayor KGE medio entre las arquitecturas evaluadas, manteniendo además un comportamiento más equilibrado entre correlación, variabilidad y volumen medio. A nivel anual, el mayor KGE de cada uno de los tres años normales evaluados es alcanzado por una arquitectura de ML, aunque esto no implica que todos los enfoques de ML superen sistemáticamente a los modelos base en cada caso particular. Más bien, los resultados muestran que los modelos de ML tienden a adaptarse con mayor facilidad a condiciones

4.3.3. Análisis años húmedos

Los años húmedos presentan, en general, un mejor desempeño de los modelos respecto a otros regímenes hidrológicos. Bajo condiciones de mayor disponibilidad hídrica, la señal del caudal tiende a ser más clara y consistente, lo que facilita que las distintas arquitecturas reproduzcan con mayor precisión la dinámica anual. Aun así, un mejor ajuste global no significa que todos los modelos representen el sistema de la misma forma. Siguen apareciendo diferencias relevantes en la estimación de caudales altos, en la variabilidad simulada y en el volumen acumulado.

El análisis de estos años permite ir más allá de los valores globales de KGE y observar qué errores persisten cuando la respuesta hidrológica de la cuenca se vuelve más intensa. De esta manera, la comparación deja de enfocarse solo en la correlación temporal y permite examinar con mayor detalle cómo cada modelo ajusta la magnitud y la escala del caudal simulado.

Tabla 8. Métricas de desempeño anual para los modelos evaluados durante los años húmedos del período de testing final.

Año	Modelo	KGE	r	α	β	RMSE	MAE
2001	RF	0.637	0.978	0.7	0.798	2.28	1.521
	LSTM	0.772	0.828	0.927	0.87	3.07	2.063
	CNN	0.769	0.941	0.853	0.832	2.146	1.314
	Modelo USM	0.531	0.571	1.06	1.181	5.069	3.497
	Modelo de Python	0.675	0.887	1.091	1.29	3.304	2.404
2005	RF	0.641	0.837	0.68	0.996	3.093	1.425
	LSTM	0.529	0.875	1.209	1.403	3.913	2.559
	CNN	0.807	0.913	0.854	1.091	2.299	1.312
	Modelo USM	0.258	0.659	1.379	1.538	6.408	4.176
	Modelo de Python	0.524	0.804	1.397	1.174	4.662	2.689
2010	RF	0.779	0.932	0.847	0.856	1.404	1.164
	LSTM	0.719	0.942	0.799	0.813	1.539	1.221
	CNN	0.453	0.836	1.466	1.237	2.89	2.084
	Modelo USM	0.485	0.772	0.766	0.601	2.975	2.435
	Modelo de Python	0.728	0.937	1.254	1.073	1.51	1.215

Para el año 2001, los modelos de ML de carácter neuronal presentan desempeños relativamente cercanos. LSTM (KGE = 0.772) logra un buen equilibrio entre correlación y magnitud, con valores de variabilidad ($\alpha = 0.927$) y volumen ($\beta = 0.87$) próximos a la unidad. La CNN muestra un comportamiento similar (KGE = 0.769), mientras que RF, a pesar de presentar la mayor correlación ($r = 0.978$), tiende a suavizar la variabilidad ($\alpha = 0.7$), lo que limita su eficiencia global.

En los modelos físicos, el USM presenta una ligera intensificación de la señal ($\alpha = 1.06$) junto con una sobreestimación del volumen ($\beta = 1.181$). Por su parte, el modelo de Python mantiene una buena coherencia temporal, aunque con un sesgo positivo en el volumen acumulado ($\beta = 1.29$).

Cabe destacar que el año 2001 incluye eventos de mayor magnitud, lo que exige una representación adecuada de la escala del caudal. En este contexto, las métricas de error permiten diferenciar con mayor claridad los desempeños. La CNN presenta los menores valores de error (RMSE = 2.146; MAE = 1.314), lo que confirma que su equilibrio en α y β se traduce en una menor desviación absoluta. El modelo USM, en cambio, presenta los mayores errores (RMSE = 5.069; MAE = 3.497), evidenciando que la sobreestimación del volumen genera desviaciones acumuladas, especialmente en eventos extremos.

En 2005, la CNN alcanza el mejor desempeño global (KGE = 0.807), reflejando una representación relativamente equilibrada del régimen anual. En cambio, LSTM, aunque mantiene coherencia temporal, tiende a intensificar la variabilidad ($\alpha = 1.209$) y el volumen medio ($\beta = 1.403$), lo que incrementa los errores en magnitud. RF muestra el comportamiento opuesto, con una reducción de la amplitud interanual ($\alpha = 0.68$), lo que afecta su capacidad para reproducir adecuadamente los picos del año.

El modelo USM presenta nuevamente el menor desempeño (KGE = 0.258), asociado a una intensificación considerable de la dinámica anual ($\alpha = 1.379$) y una sobreestimación del volumen ($\beta = 1.538$), lo que se traduce en los mayores errores (RMSE = 6.408; MAE = 4.176). El modelo implementado en Python mejora parcialmente este ajuste (KGE = 0.524), pero mantiene una amplitud del régimen superior a la observada ($\alpha = 1.397$), lo que también impacta en los errores absolutos.

En 2010 se observa una mejora general en la representación del régimen húmedo. Los valores de RMSE y MAE son considerablemente menores en la mayoría de los modelos en comparación con 2005, lo que indica una mayor consistencia en la simulación del caudal. Dentro de este escenario, RF alcanza los menores errores (RMSE = 1.404; MAE = 1.164), lo que confirma que su buen desempeño no solo se refleja en el KGE, sino también en una menor magnitud de error diario.

La CNN, en cambio, presenta errores más elevados (RMSE = 2.89; MAE = 2.084), lo que refleja discrepancias en la magnitud de la señal simulada. Al analizar las componentes del KGE, se observa que RF mantiene valores cercanos a la unidad en variabilidad y volumen, lo que respalda su consistencia estructural. LSTM también muestra un comportamiento estable, aunque con una leve atenuación de la variabilidad ($\alpha = 0.799$).

Por su parte, la CNN tiende a intensificar la señal respecto a la observada ($\alpha = 1.466$) y a incrementar el volumen anual ($\beta = 1.237$), lo que explica el aumento en los errores absolutos. En los modelos base, el USM presenta un desempeño intermedio bajo (KGE = 0.485), con una subestimación del volumen medio ($\beta = 0.601$). Distinto es el caso del modelo de Python, ya que logra un ajuste competitivo (KGE = 0.728), aunque con una mayor amplitud en la señal anual ($\alpha = 1.254$).

De acuerdo con este análisis, los años húmedos muestran que la principal dificultad no está tanto en reproducir la forma temporal del hidrograma, sino en representar correctamente la magnitud de los eventos extremos y el volumen acumulado. En varios casos, los modelos logran seguir adecuadamente la dinámica general del caudal, pero presentan desviaciones importantes al momento de capturar la intensidad real de los picos o la amplitud de la señal. Como resultado, cuando la variabilidad es exagerada o excesivamente atenuada, el KGE se deteriora incluso bajo correlaciones elevadas. Esto evidencia que reproducir correctamente la evolución temporal del sistema no garantiza, por sí solo, una representación adecuada del régimen hidrológico.

Resulta consistente que el mayor KGE de los tres años húmedos haya sido alcanzado por arquitecturas de ML, siendo LSTM el modelo con mejor desempeño en 2001, CNN en 2005 y RF en 2010. Aunque

la arquitectura dominante cambia según el año analizado, los resultados muestran una tendencia relativamente clara. Los enfoques de ML tienden a responder de mejor manera bajo condiciones de alta disponibilidad hídrica, donde la señal del caudal presenta patrones más intensos, marcados y estructurados.

Sin embargo, el análisis conjunto de α , β y de los errores absolutos también indica que esta ventaja debe interpretarse con cautela. Los modelos de ML no están exentos de sesgos y, en varios casos, continúan presentando problemas asociados a sobreestimaciones o subestimaciones de la variabilidad y del volumen medio. Más bien, su mejor desempeño parece relacionarse con una mayor estabilidad para representar simultáneamente la dinámica temporal, la magnitud del caudal y los eventos extremos, superando en estos escenarios a los modelos conceptuales evaluado.

4.4. Análisis de la respuesta hidrológica simulada bajo dinámicas contrastantes

Si bien en los apartados anteriores ya se discutieron las diferencias entre modelos a partir de métricas globales, hidrogramas y análisis de los modelos bajo los distintos regímenes hidrológicos, ciertos eventos particulares siguen resultando especialmente relevantes para comprender cómo responde cada arquitectura frente a dinámicas hidrológicas más complejas o menos comunes. En algunos casos, pequeñas diferencias en la representación temporal del caudal terminan revelando comportamientos que, al analizar la serie de manera global, pueden pasar más desapercibidos.

Bajo esta premisa, se seleccionaron años representativos cuyo comportamiento hidrológico permite analizar con mayor detalle la respuesta estructural de los modelos frente a dinámicas contrastantes. El foco no está únicamente en identificar qué modelo alcanza una mejor eficiencia anual, sino en observar cómo cada enfoque, ya sea basado en ML o en formulaciones físicas, responde ante distintos patrones de variabilidad intraanual, transiciones estacionales y eventos de alta magnitud.

De este modo, el análisis complementa la evaluación estadística previa y permite interpretar los mecanismos que están detrás de las diferencias observadas en las métricas. Más que comparar resultados, busca ofrecer una lectura más profunda de las fortalezas y limitaciones estructurales de cada modelo en la representación del sistema hidrológico.

4.4.1. Respuesta ante eventos extremos aislados del año 2013

El año 2013 presenta un evento puntual que alcanza valores superiores a 40 m³/s, muy por sobre el comportamiento predominante del régimen seco. Si bien en los días previos se registran precipitaciones importantes, el máximo observado no coincide con precipitación en el mismo día y su magnitud resulta desproporcionada respecto al patrón general del año. Por ello, este episodio debe interpretarse con cautela, ya que podría estar asociado tanto a una respuesta hidrológica retardada como a una anomalía en la serie observada.

Las condiciones meteorológicas previas refuerzan esa incertidumbre. Durante los dos días anteriores al máximo caudal se registraron precipitaciones cercanas a 69.5 mm el 27 de mayo y 29.8 mm el 28 de mayo. Sin embargo, las temperaturas mínimas permanecieron bajo 0 °C (−3,20 y −3,59 °C, respectivamente) y las máximas alcanzaron solo 0,39 y −0,93 °C, dificultando asociar el aumento abrupto del caudal a una transformación inmediata de la precipitación en escorrentía líquida. A esto se suma otro elemento relevante, el máximo caudal ocurre el 29 de mayo, precisamente en una jornada sin

precipitación registrada. Todo esto apunta a una posible respuesta diferida del sistema hidrológico o, quizá, a cierta inconsistencia puntual entre las forzantes meteorológicas y el caudal observado.

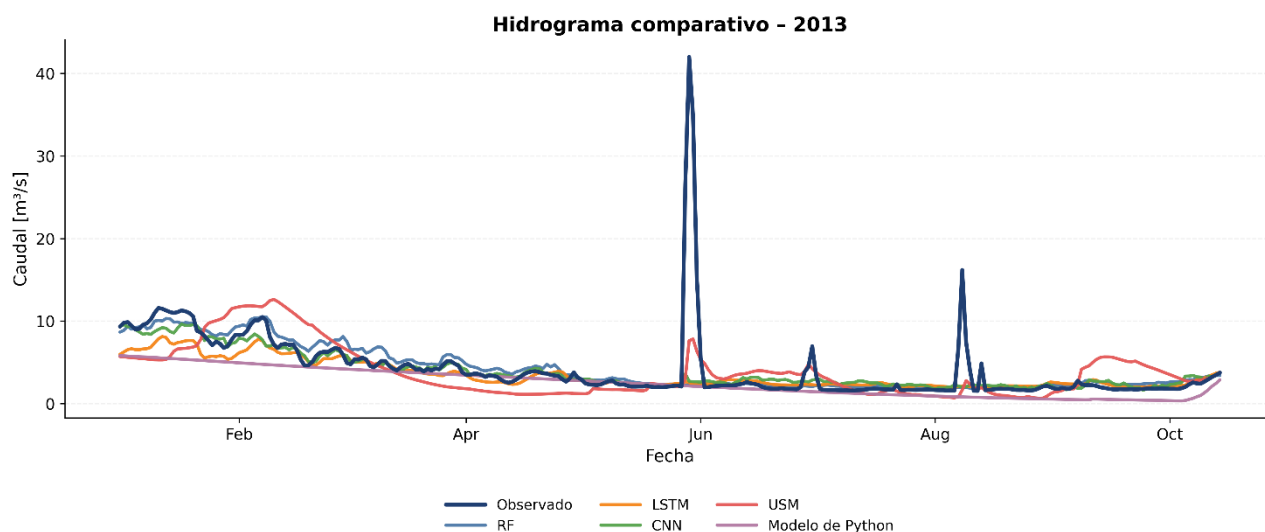


Figura 7. Hidrograma comparativo del año 2013 entre el caudal observado y las simulaciones de los modelos de ML (RF, LSTM y CNN), el modelo hidrológico USM y el Modelo de Python.

Al analizar el hidrograma del año 2013, se observa que, si bien ninguno de los modelos reproduce la magnitud de ciertos eventos máximos, sí existen diferencias claras en la forma en que responden a los eventos de menor escala presentes en la serie. En particular, durante los episodios registrados a finales de mayo y agosto, el modelo hidrológico USM responde de manera más marcada a algunos incrementos del caudal, mientras que RF y Python muestran una señal más estable. Los modelos LSTM y CNN también reflejan variaciones durante esos episodios, aunque de menor amplitud, las cuales quedan parcialmente ocultas por la escala impuesta por el evento extremo principal.

Este patrón coincide con las condiciones meteorológicas registradas durante ese período. Las precipitaciones ocurrieron con temperaturas mínimas inferiores a 0 °C, mientras que el máximo caudal se alcanzó en una jornada sin lluvia. Esto indica que la respuesta hidrológica no estuvo asociada únicamente a la escorrentía superficial inmediata, sino también a procesos como la acumulación y posterior fusión de nieve o a un desfase en la respuesta de la cuenca. Bajo estas condiciones, la relación entre precipitación y caudal deja de ser inmediata, lo que aumenta la dificultad para representar adecuadamente la evolución observada.

Parte de estas diferencias puede explicarse por la forma en que cada enfoque representa los procesos hidrológicos. El modelo USM incorpora de manera explícita procesos físicos de almacenamiento y liberación del agua, permitiendo capturar respuestas asociadas a precipitaciones invernales aun cuando el aumento del caudal ocurra con cierto retraso. Si bien el Modelo de Python se basa en el mismo enfoque conceptual, en este evento presenta una respuesta más amortiguada. Los modelos de ML, en cambio, basan sus predicciones en las relaciones aprendidas durante el entrenamiento. Cuando enfrentan eventos poco frecuentes o situaciones en las que el vínculo entre precipitación y caudal presenta un desfase temporal, tienden a generar respuestas más suavizadas y a subestimar la intensidad de estos episodios. Desde esta perspectiva, la incapacidad de los modelos para alcanzar dicho valor no necesariamente representa una falla estructural completa, sino que refleja una tendencia a suavizar eventos atípicos que se alejan del comportamiento dominante aprendido durante el entrenamiento.

4.4.2. Respuesta ante transición estacional seca–húmeda (año 2008)

El año 2008 presenta un comportamiento típico de régimen seco durante gran parte del período, con caudales bajos y una evolución relativamente estable en los primeros meses. Esta condición se mantiene sin cambios bruscos, mostrando una recesión progresiva posterior al periodo de deshielo.

En la figura 8 se observa que, hacia en el último tramo del año se observa un cambio en la dinámica del caudal, sin embargo, este comportamiento debe interpretarse considerando la ausencia de caudales observados durante los meses de septiembre y octubre. Dado que esas fechas fueron excluidas del conjunto de datos utilizado en la evaluación, tanto las observaciones como las predicciones de los modelos presentan una discontinuidad en ese período, por lo que no es posible analizar de forma continua la transición previa al incremento registrado en noviembre.

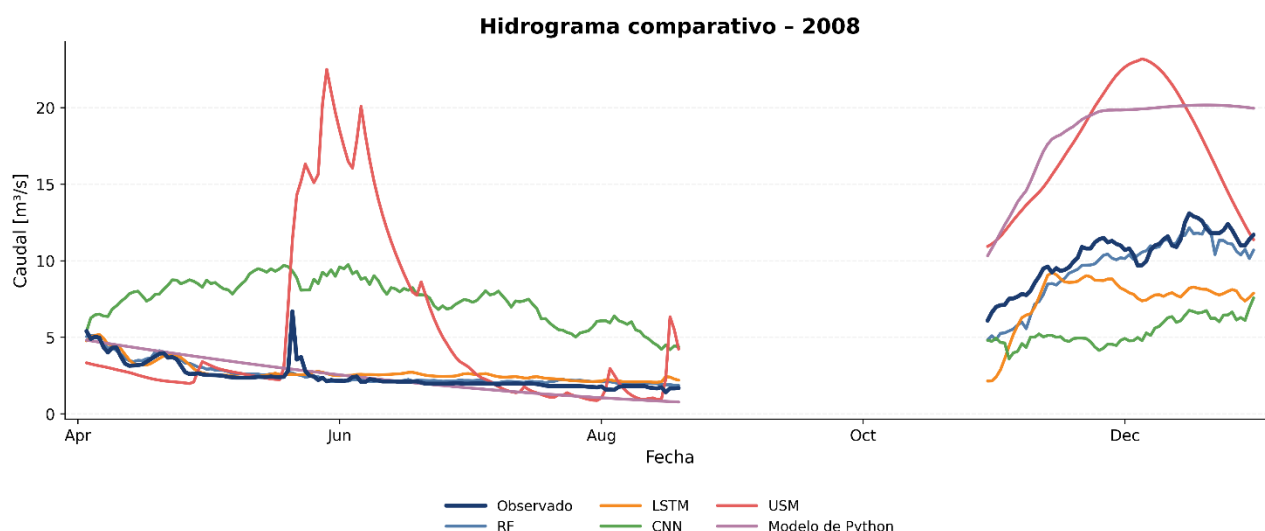


Figura 8. Hidrograma comparativo del año 2008 entre el caudal observado y las simulaciones de los modelos de ML (RF, LSTM y CNN), el modelo hidrológico USM y el Modelo de Python.

Al observar la evolución del año 2008, las diferencias entre modelos se hacen evidentes a medida que el régimen cambia desde una fase de caudales bajos hacia el incremento del último trimestre. Durante la primera mitad del año, donde predomina la recesión, RF y LSTM mantienen un ajuste cercano a la señal observada, siguiendo de forma razonable tanto la disminución del caudal como su estabilidad posterior. En cambio, CNN tiende a mantenerse por sobre los valores observados, sin representar con claridad la fase de descenso.

Esta diferencia se vuelve más notoria al acercarse al último trimestre. Mientras RF logra acompañar el ascenso del caudal, LSTM sigue la tendencia general, aunque con una leve atenuación en la variabilidad. La red CNN, en cambio, no logra representar adecuadamente esta transición, manteniendo valores bajos durante gran parte del incremento y reaccionando de forma tardía, con una respuesta que no alcanza a capturar ni la magnitud ni el ritmo del cambio observado.

El comportamiento de los modelos físicos introduce otro matiz. Tanto el modelo hidrológico USM como el modelo de Python reaccionan con un aumento más pronunciado que el observado, generando una sobreestimación clara en el ascenso final. No es que no vean el cambio, lo ven y con claridad, pero lo amplifican como si el sistema respondiera con más ímpetu del que realmente muestra el hidrograma.

Un evento particular se observa a comienzos de junio, cuando el caudal presenta un incremento abrupto en medio de la fase de recesión. Frente a este evento, RF y LSTM muestran una respuesta poco reactiva, manteniéndose casi cercano a los caudales previos y sin reproducir claramente el aumento observado. El modelo CNN tampoco logra representar este pulso, manteniendo una señal relativamente sobreestimada respecto del caudal observado durante ese período. El Modelo de Python, por su parte, presenta una respuesta aún más suave, prácticamente sin reaccionar al incremento del caudal, mientras que el modelo hidrológico USM si detecta ese aumento en el caudal, pero lo amplifica considerablemente, generando un pico muy superior al observado. En consecuencia, estos resultados muestran que los incrementos repentinos del caudal continúan siendo uno de los escenarios más difíciles de representar para los modelos evaluados. Aunque cada uno responde de manera diferente, ninguno logra reproducir adecuadamente la magnitud y la dinámica de este tipo de eventos.

4.4.3. Sesgo acumulativo en volumen anual para años 2015 y 2019

Los años 2015 y 2019 muestran un comportamiento distinto a los casos anteriores, donde las diferencias entre modelos no están dominadas por eventos extremos puntuales, sino por desviaciones que se van acumulando de forma progresiva a lo largo del año. En ambos casos, el hidrograma presenta un régimen relativamente ordenado, con una fase de recesión bien definida y un incremento hacia el último trimestre, sin la presencia de pulsos abruptos que condicionen la señal completa.

Sin embargo, es precisamente en esta aparente normalidad donde emergen las diferencias más sutiles. Las simulaciones no fallan en momentos específicos, sino que se van separando gradualmente de la señal observada, generando una sobreestimación o subestimación del volumen acumulado. Es un desajuste menos evidente a simple vista, pero más persistente, casi como una deriva que se instala sin hacer ruido y termina marcando la diferencia al final del año.

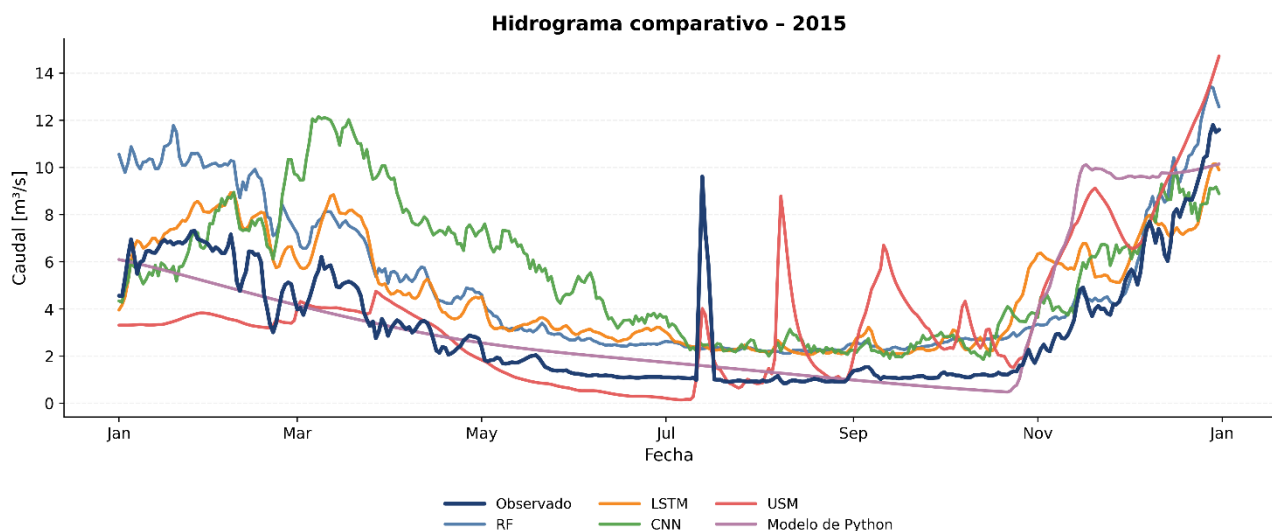


Figura 9. Hidrograma comparativo del año 2015 entre el caudal observado y las simulaciones de los modelos de ML (RF, LSTM y CNN), el modelo hidrológico USM y el Modelo de Python.

En 2015, el hidrograma muestra una disminución progresiva desde los primeros meses hacia el invierno, seguida de un aumento hacia el final del año. A diferencia de otros casos, aquí las diferencias entre modelos aparecen desde el inicio. Durante el primer trimestre, CNN y RF tienden a sobreestimar los caudales, elevando la línea base por sobre lo observado, mientras que LSTM se mantiene más cercano, aunque sin reproducir completamente la amplitud de las variaciones.

Durante el periodo invernal, el modelo USM introduce pulsos aislados con magnitudes superiores a las observadas, lo que rompe la continuidad de la señal en momentos específicos. Hacia el tramo final del año, varios modelos logran seguir el aumento del caudal, pero difieren en la forma en que lo representan. RF y USM muestran un ascenso más pronunciado, mientras que el modelo de Python mantiene una respuesta más amortiguada y reacciona de manera tardía, sin capturar completamente la variabilidad del incremento.

En este sentido, el año 2015 muestra que las diferencias entre modelos no siempre se originan en eventos específicos, sino que pueden desarrollarse de forma progresiva a lo largo del tiempo, afectando de manera conjunta la representación del régimen anual.

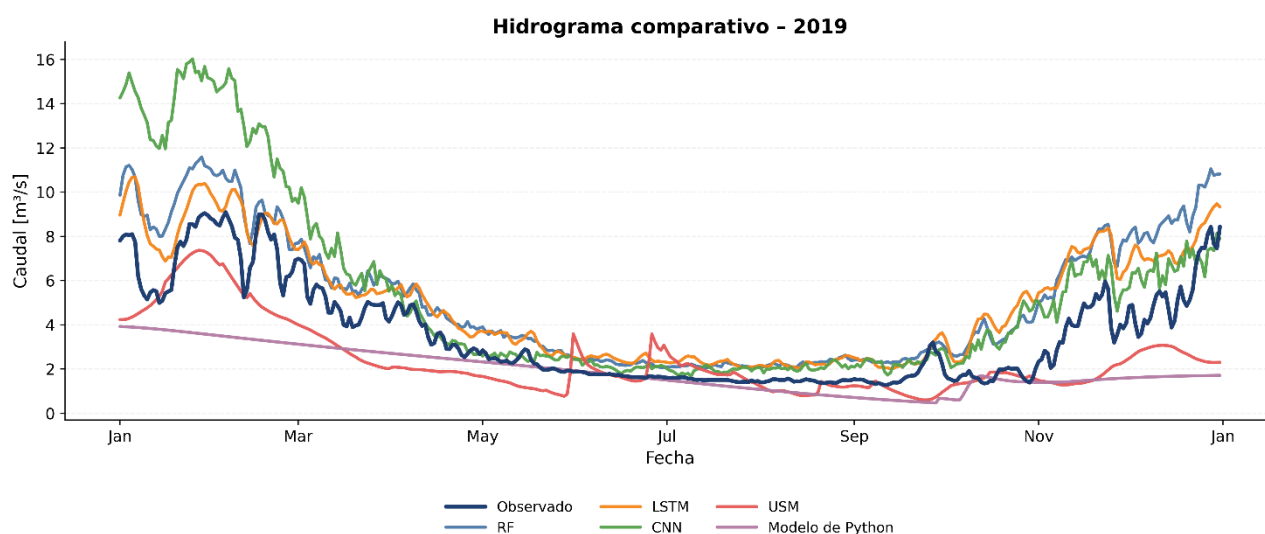


Figura 10. Hidrograma comparativo del año 2019 entre el caudal observado y las simulaciones de los modelos de ML (RF, LSTM y CNN), el modelo hidrológico USM y el Modelo de Python.

En 2019 el patrón es distinto. Durante el primer trimestre la CNN alcanza valores considerablemente superiores al observado, superando los 15 [m³/s] cuando el caudal real se mantiene por debajo de 9 [m³/s]. RF y LSTM también parten por sobre la señal observada, aunque con menor diferencia, pero a partir del último trimestre del año, RF y LSTM vuelven a sobreestimar el ascenso. En el periodo medio del año la mayoría de los modelos converge hacia valores similares, pero se puede observar que el modelo de Python permanece consistentemente por debajo del observado durante casi todo el año, y sus valores son demasiado suaves.

El comportamiento del modelo USM sigue una lógica distinta. Tras iniciar cercano a la señal observada, entra rápidamente en una fase de subestimación que se prolonga durante el otoño e invierno, con valores incluso inferiores a los registrados. Hacia el último trimestre logra captar el cambio de tendencia, pero lo hace con un ascenso mucho más suavizado, sin alcanzar la magnitud observada ni la de otros modelos.

En el tramo final del año, cuando el caudal aumenta, varios modelos vuelven a separarse de la señal observada. RF y LSTM tienden a sobreestimar el ascenso, mientras que otros enfoques mantienen respuestas más contenidas. Bajo este comportamiento, el problema no radica en la sincronización temporal, ya que la mayoría logra identificar correctamente los momentos de cambio, sino en la representación de la magnitud. Así, pequeñas diferencias sostenidas en el tiempo terminan

acumulándose y generan discrepancias significativas en el volumen anual.

Estos resultados revelan una diferencia clara en la forma en que los modelos representan el comportamiento del caudal. Los modelos RF, LSTM y CNN priorizan la reproducción de la variabilidad temporal del hidrograma y logran seguir gran parte de las oscilaciones observadas a lo largo del año. Sin embargo, presentan diferencias en la magnitud de los caudales, especialmente al inicio del año, y durante el período de ascenso del último trimestre. Por otro lado, el modelo hidrológico USM y el modelo de Python tendieron a generar una respuesta más suavizada, capturando la tendencia media del hidrograma, pero sin representar buena parte de la variabilidad registrada.

4.5. Interpretabilidad de las arquitecturas de ML y sensibilidad a perturbaciones

Una de las principales críticas que han tenido los modelos de ML, pese a sus buenos resultados de desempeño, es su carácter de “caja negra”, es decir, la dificultad para interpretar de manera explícita las relaciones internas que realizan estos algoritmos para llegar a una predicción determinada. A diferencia de los modelos hidrológicos conceptuales, donde cada componente representa un proceso físico identificable, las arquitecturas de aprendizaje automático operan a partir de relaciones internas que no siempre son evidentes. En lugar de describir procesos, construyen patrones a partir de los datos, lo que dificulta establecer una conexión directa entre la predicción y los mecanismos físicos que gobiernan el sistema. Este debate ha sido ampliamente discutido en la literatura reciente sobre hidrología basada en datos (Anderson & Radić, 2022; Nearing et al., 2021; Shen, 2018), donde el foco ya no está solo en qué tan bien predicen los modelos, sino en qué tan coherentes son las relaciones que aprenden.

A partir de esta discusión, y como complemento a los resultados obtenidos, en este trabajo se analiza si las relaciones capturadas por los modelos de ML son consistentes con los procesos físicos dominantes en la cuenca del río Juncal. Más allá del ajuste numérico, el interés está en entender si los modelos están representando dinámicas que tengan sentido desde el punto de vista hidrológico.

4.5.1. Resultado del análisis curvas PDP+ICE

Las curvas PDP e ICE permiten observar con mayor detalle cómo cada arquitectura relaciona las variables meteorológicas con el caudal simulado. Más allá de identificar qué variables son importantes dentro de cada modelo, el interés de este análisis está en entender cómo esas relaciones se organizan internamente y qué tan coherente resulta el comportamiento aprendido frente a la dinámica esperada en una cuenca de régimen nivo-glacial.

El análisis se centra en las variables más influyentes identificadas previamente mediante importancia por permutación, considerando las tres variables con mayor peso dentro de cada arquitectura. A partir de ellas, se busca interpretar qué patrones terminan guiando el comportamiento de los modelos y cómo cada enfoque representa procesos asociados a acumulación, fusión y generación de escorrentía dentro de la cuenca.

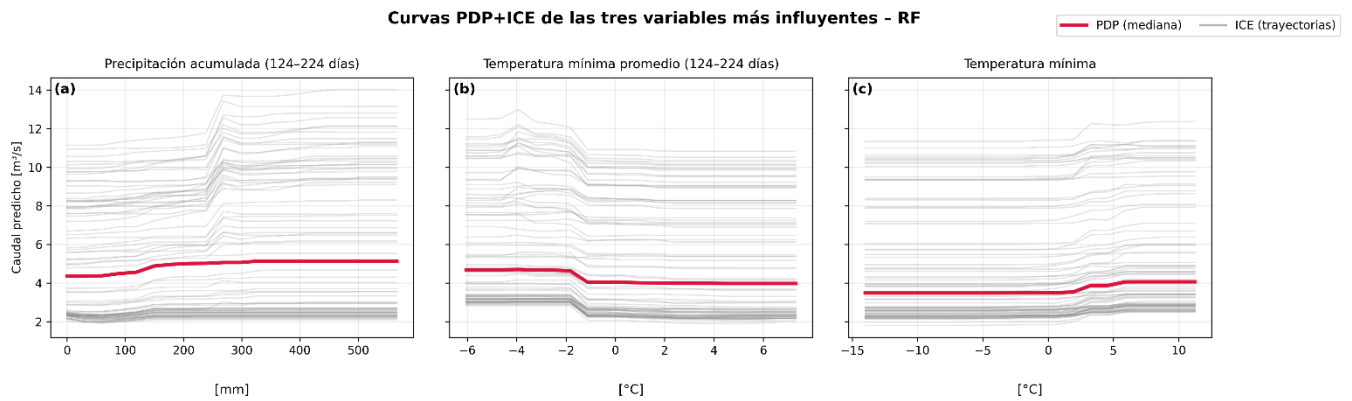


Figura 11. Curvas de dependencia parcial (PDP) y trayectorias individuales (ICE) para las tres variables más influyentes del modelo RF. Las variables más influyentes corresponden a precipitación acumulada (124–224 días), temperatura mínima promedio (124–224 días) y temperatura mínima diaria.

En la arquitectura RF, según el panel (a) de la figura 11, la variable con mayor influencia corresponde a la precipitación acumulada en una ventana larga de 124 a 224 días. La curva PDP muestra una relación relativamente clara, ya que, a medida que aumenta la acumulación previa, también aumenta el caudal predicho. Sin embargo, ese crecimiento no es indefinido, debido a que la respuesta se intensifica principalmente en rangos intermedios y luego comienza a estabilizarse, como si el sistema alcanzara cierto nivel de saturación a partir del cual nuevas acumulaciones ya no generan cambios proporcionales en la descarga simulada. Las trayectorias ICE muestran además que la respuesta no es uniforme entre todos los casos analizados. Algunas curvas reaccionan con mayor intensidad que otras frente a niveles similares de acumulación, evidenciando que el efecto de esta variable depende también de otras condiciones presentes en el sistema hidrológico, por lo que la precipitación acumulada, por sí sola, no determina completamente la respuesta del modelo, sino que interactúa con el estado previo de la cuenca y con otras variables meteorológicas relevantes.

La temperatura mínima promedio en la misma ventana temporal introduce un matiz más interesante, casi contraintuitivo a primera vista. La curva muestra una leve estabilidad en valores negativos y luego una disminución del caudal predicho al acercarse a temperaturas cercanas o superiores a 0 °C. Leída superficialmente, podría parecer que “más temperatura implica menos caudal”, lo que sería absurdo si se tomara de forma literal. Pero aquí la clave está en entender qué representa realmente esta variable, no un gatillante inmediato, sino una memoria térmica del sistema. Temperaturas promedio más altas en ese periodo sugieren menor acumulación nival previa o incluso pérdida anticipada de esa nieve, lo que reduce la disponibilidad de agua para generar caudal posteriormente. En otras palabras, no es la temperatura la que quita agua, sino lo que esa temperatura implica en términos de historia del sistema. Lo interesante es que las curvas ICE muestran una transición relativamente consistente en este punto, como si el modelo identificara un umbral difuso asociado al cambio de estado entre acumulación y pérdida de nieve.

La temperatura mínima diaria, en cambio, se comporta de manera más directa, aunque no por ello más simple. La curva PDP se mantiene prácticamente plana en valores bajos y comienza a elevarse suavemente cuando se superan temperaturas cercanas a 0 °C. Aquí sí aparece la lógica más inmediata, ya que condiciones térmicas que favorecen la fusión se traducen en aumentos de caudal, siempre y cuando exista agua disponible. Pero incluso en este caso, la respuesta es moderada y heterogénea. Las curvas ICE no siguen un patrón único, lo que deja en evidencia algo que suele olvidarse con facilidad. La temperatura por sí sola no genera caudal, del mismo modo en que el calor no derrite lo que no existe. Su efecto depende, inevitablemente, del estado previo del sistema, en particular de la disponibilidad de nieve o almacenamiento hídrico.

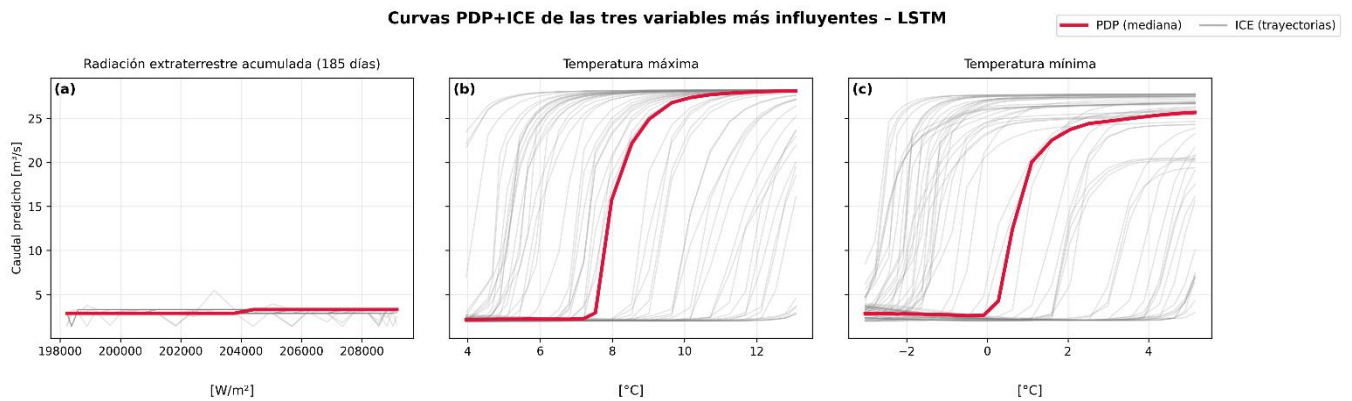


Figura 12. Curvas de dependencia parcial (PDP) y trayectorias individuales (ICE) para las tres variables más influyentes del modelo LSTM. Las variables más influyentes corresponden a radiación extraterrestre acumulada (185 días), temperatura máxima y temperatura mínima.

Para el modelo LSTM, de acuerdo con el panel (a) de la Figura 12, la radiación extraterrestre acumulada en una ventana de 185 días muestra una influencia bastante acotada sobre el caudal simulado. La curva PDP se mantiene prácticamente plana a lo largo de todo el rango, con variaciones menores que no cambian de forma sustantiva la respuesta del modelo, es decir, aunque la variable está presente, no cumple un rol determinante en la predicción. Tiene sentido si se piensa que no aporta agua al sistema, sino energía potencial, y que su efecto real queda mediado por variables más directas como la temperatura. En la práctica, una mayor disponibilidad de energía puede favorecer procesos como la fusión, pero solo en la medida en que exista nieve disponible en la cuenca. Las trayectorias ICE acompañan esta lectura, ya que no muestran cambios estructurales relevantes, sino una dispersión acotada que refuerza la idea de una contribución limitada.

El comportamiento cambia de forma evidente al observar la temperatura máxima. La curva PDP muestra un aumento muy marcado del caudal en un rango relativamente estrecho de temperaturas, lo que a primera vista podría interpretarse como un umbral claro. Sin embargo, al mirar las curvas individuales, esa transición no ocurre en un punto fijo, sino que se desplaza según cada trayectoria. Más que un umbral único, lo que aparece es una especie de “zona de activación”, donde la respuesta del modelo se intensifica dependiendo del contexto previo, por lo que el modelo pudo identificar esa activación del caudal. En otras palabras, la temperatura no actúa como un interruptor, sino como una condición que habilita la generación de caudal cuando el sistema ya tiene agua disponible.

La temperatura mínima sigue una lógica similar, aunque con una transición más gradual. La curva PDP muestra un aumento del caudal a medida que se pasa desde valores negativos a positivos, lo que es coherente con condiciones más favorables para la fusión. Pero, al igual que antes, las trayectorias ICE revelan que esta respuesta no es uniforme. Algunas reaccionan antes, otras después, lo que refuerza la idea que el modelo no está respondiendo únicamente al valor instantáneo de la temperatura, más bien está integrando información previa del sistema, donde temperaturas bajas en periodos anteriores pueden haber favorecido la acumulación de nieve y, por lo tanto, condicionan la disponibilidad de agua en el presente.

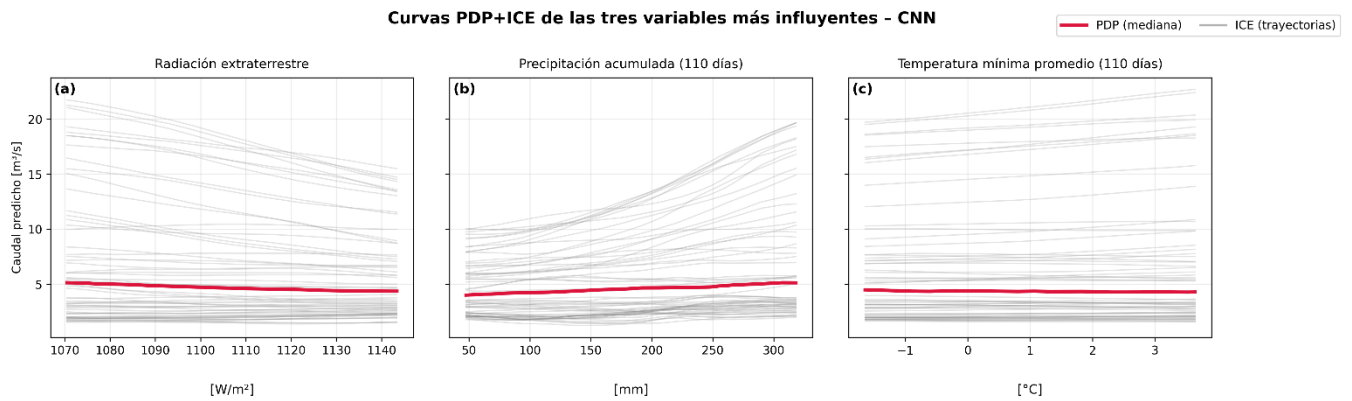


Figura 13. Curvas de dependencia parcial (PDP) y trayectorias individuales (ICE) para las tres variables más influyentes del modelo CNN. Las variables más influyentes corresponden a radiación extraterrestre, la precipitación acumulada (110 días) y la temperatura mínima promedio (110 días).

En la figura 13 se observa que las variables más influyentes del modelo CNN muestran un comportamiento algo distinto a lo observado en los otros enfoques, con respuestas que, en general, tienden a ser más suaves y dependientes del contexto. Al analizar la radiación extraterrestre, se observa una relación levemente negativa con el caudal predicho. La curva PDP disminuye de forma gradual a medida que aumenta la radiación, aunque el cambio es acotado. Más que interpretar esto como una relación directa, conviene entenderlo como una asociación interna del modelo. Al igual que en otros casos, la radiación no aporta agua al sistema, sino energía potencial, por lo que su efecto depende de la disponibilidad previa de nieve, lo cual se puede apreciar en las trayectorias ICE, donde algunas respuestas presentan descensos más marcados mientras otras permanecen relativamente estables, evidenciando que el efecto de la radiación no es uniforme y depende del contexto hidrológico representado en cada trayectoria. En este sentido, no resulta extraño que su rol aparezca como más secundario frente a variables térmicas que representan de mejor manera el estado energético real de la cuenca.

La precipitación acumulada en 110 días, en cambio, presenta una relación más clara con el caudal predicho, ya que a medida que aumenta la acumulación también lo hace el caudal estimado por la CNN, coherente con una mayor disponibilidad hídrica dentro de la ventana temporal analizada. Sin embargo, la respuesta no evidencia quiebres abruptos ni umbrales claramente definidos, sino un aumento progresivo que parece representar la acumulación previa como una condición antecedente del sistema más que como un detonante puntual de crecidas. Esta tendencia aparece de forma relativamente consistente en gran parte de las trayectorias ICE, aunque con distintas magnitudes de respuesta entre casos, reflejando que el efecto final de la variable continúa dependiendo de las condiciones hidrológicas presentes y de la interacción con las demás variables de entrada.

Por su parte, la temperatura mínima promedio en 110 días presenta una respuesta mucho más tenue. La curva PDP se mantiene prácticamente plana, con una leve tendencia negativa dentro del rango analizado, lo que indica que, en promedio, el modelo no modifica de forma importante el caudal frente a cambios en esta variable. Sin embargo, al observar las trayectorias ICE, la respuesta deja de ser uniforme, ya que algunas muestran aumentos suaves del caudal, otras permanecen estables y algunas incluso presentan tendencias opuestas. Al tratarse de una variable promediada en una ventana temporal extensa, es razonable asociarla a condiciones antecedentes vinculadas a acumulación o fusión nival, cuyo efecto promedio aparece atenuado en la curva PDP, pero emerge con mayor claridad en la dispersión de las curvas ICE.

4.5.2. Respuestas de modelos de ML frente a sensibilidad de perturbaciones

Mientras las curvas PDP e ICE permiten interpretar las relaciones aprendidas entre las variables de entrada y el caudal predicho, el análisis de perturbaciones complementa esa información evaluando cómo cambia la respuesta del modelo frente a modificaciones controladas de las variables más influyentes. A partir de estos resultados es posible identificar cuáles producen las mayores variaciones en el caudal estimado y comparar la sensibilidad de las diferentes arquitecturas de aprendizaje automático.

En el modelo RF, la precipitación acumulada entre 124 y 224 días aparece como la variable que genera la respuesta más marcada frente a las perturbaciones analizadas. Una reducción del 10 % produce una disminución cercana a $-0.06 \text{ [m}^3/\text{s]}$ en el caudal, mientras que un aumento equivalente genera un incremento aproximado de $+0.03 \text{ [m}^3/\text{s]}$. Esta asimetría en la respuesta es relevante, ya que muestra que el modelo es considerablemente más sensible a déficits de precipitación que a incrementos de igual magnitud. Sin embargo, la precipitación acumulada no controla el caudal de manera aislada. En una cuenca de alta montaña, la disponibilidad efectiva de agua depende también de las condiciones térmicas que permiten transformar esa acumulación en deshielo y escorrentía. En ese contexto, una disminución en la acumulación previa puede restringir con mayor fuerza el caudal potencial del sistema, mientras que aumentos adicionales de precipitación no necesariamente producen una respuesta proporcional. Si las condiciones energéticas no favorecen la fusión de nieve, o si la cuenca ya dispone de un almacenamiento importante, parte de ese exceso puede no traducirse inmediatamente en un aumento del caudal simulado.

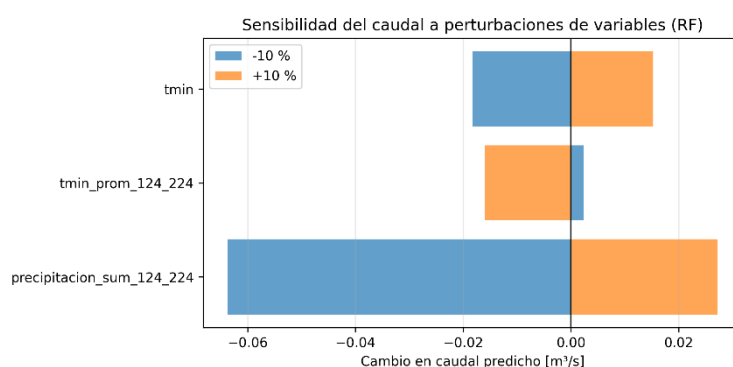


Figura 14. Sensibilidad del caudal predicho por el modelo RF ante perturbaciones de $\pm 10\%$ en las tres variables más influyentes.

En las variables térmicas se observa un comportamiento diferenciado, en línea con lo identificado previamente en las curvas PDP+ICE. La temperatura mínima diaria presenta una respuesta positiva frente a las perturbaciones, algo coherente con su influencia más inmediata sobre los procesos de fusión. Aun así, el efecto sigue siendo moderado, ya que un aumento de la temperatura, por sí solo, no necesariamente se traduce en incrementos importantes de caudal si el sistema no dispone previamente de agua almacenada en forma de nieve o precipitación acumulada.

Por otro lado, la temperatura mínima promedio en la ventana de 124 a 224 días muestra una respuesta negativa ante incrementos, reforzando la idea de que esta variable funciona más como una señal de memoria térmica del sistema que como un desencadenante directo del deshielo. Temperaturas invernales más altas durante ese período podrían estar asociadas a una menor acumulación nival o a pérdidas anticipadas de nieve, reduciendo así el volumen de agua disponible durante la temporada de fusión.

En base a estos resultados, el modelo RF presenta respuestas de menor magnitud frente a perturbaciones uniformes en comparación con los otros modelos. En la práctica, esto se traduce como un compor-

tamiento más conservador, con una sensibilidad más acotada frente a cambios abruptos en las condiciones meteorológicas.

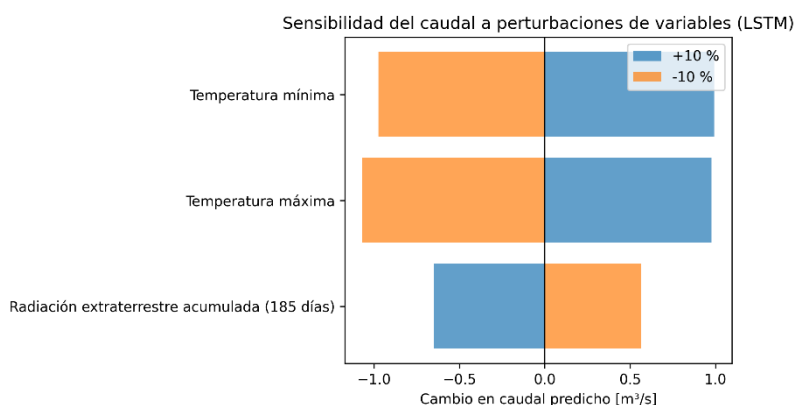


Figura 15. Sensibilidad del caudal predicho por el modelo LSTM ante perturbaciones de $\pm 10\%$ en las tres variables más influyentes.

En el caso del modelo LSTM, las mayores sensibilidades se concentran en las variables térmicas, donde variaciones de un 10 % generan cambios cercanos a $\pm 1 \text{ m}^3/\text{s}$ en el caudal estimado. Este resultado es consistente con lo observado anteriormente en las curvas PDP+ICE, en las que las temperaturas máxima y mínima mostraban las transiciones más marcadas en la respuesta del modelo. En otras palabras, el LSTM parece haber incorporado de manera muy clara la influencia del balance energético del sistema, haciendo que pequeñas variaciones de temperatura se traduzcan en cambios apreciables en la generación de caudal, siempre que la cuenca presente disponibilidad previa de agua almacenada en forma de nieve o precipitación. A diferencia de lo observado en el modelo RF, aquí la respuesta es bastante simétrica, reflejando un comportamiento más estable y continuo alrededor del punto analizado.

La radiación extraterrestre acumulada en 185 días también muestra una respuesta coherente en términos de tendencia, aunque con una influencia menor que la de las variables térmicas. Esto resulta consistente con lo observado previamente en las curvas PDP+ICE, donde la variable presentaba una respuesta relativamente plana y trayectorias ICE con baja dispersión, indicando una contribución más contextual dentro de la predicción, donde la radiación puede asociarse a la disponibilidad energética del sistema y favorecer procesos de fusión siempre que exista nieve o agua disponible en la cuenca. Sin embargo, debido a su marcada estacionalidad, también es posible que el LSTM la utilice parcialmente como una señal temporal, ayudando al modelo a identificar en qué etapa del régimen nivoglaciar se encuentra el sistema.

Por otro lado, llama la atención que la precipitación acumulada no aparezca entre las variables más sensibles. Esto evidencia que, al menos para el entorno local analizado, el modelo otorga mayor importancia a las condiciones energéticas y a la información temporal del sistema que al almacenamiento explícito asociado a las acumulaciones de precipitación.

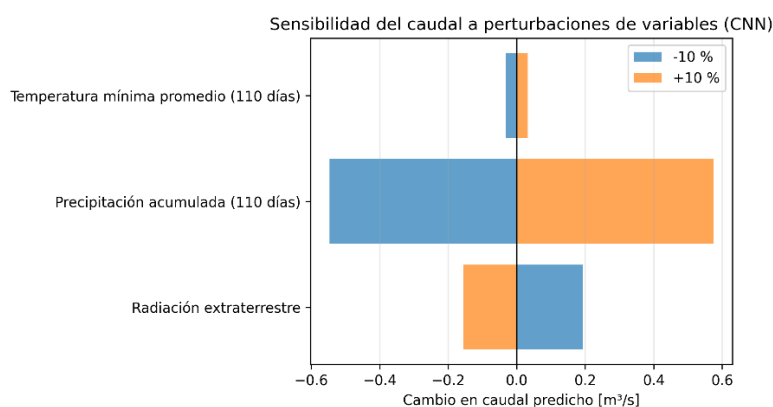


Figura 16. Sensibilidad del caudal predicho por el modelo CNN ante perturbaciones de $\pm 10\%$ en las tres variables más influyentes.

En el modelo CNN, la mayor sensibilidad se concentra en la precipitación acumulada en 110 días, donde perturbaciones de $\pm 10\%$ generan cambios cercanos a $\pm 0.6 \text{ m}^3/\text{s}$ en el caudal estimado. La respuesta, además, es bastante simétrica frente a incrementos y disminuciones de la variable, lo que tiene un comportamiento estable y sin cambios abruptos ante variaciones uniformes en las condiciones de entrada. En términos prácticos, esto indica que el almacenamiento reciente sigue siendo uno de los principales factores que modulan la respuesta hidrológica dentro de esta arquitectura, aunque de una manera más gradual que en el modelo RF.

La radiación extraterrestre, en cambio, muestra una respuesta negativa frente a incrementos de la variable. Este resultado parece estar más relacionado con el contexto temporal que con un efecto energético directo sobre el caudal. Dado que la radiación sigue un patrón fuertemente estacional, valores elevados pueden coincidir con períodos en los que parte importante del almacenamiento nival ya ha disminuido. Bajo esa lógica, el modelo ajusta el caudal hacia valores menores cuando esta variable aumenta.

Por otro lado, la temperatura mínima promedio en 110 días presenta una sensibilidad relativamente baja en comparación con las demás variables analizadas. Esto refleja que, dentro del CNN, el efecto térmico actúa más como un ajuste complementario que como un factor dominante en la generación de caudal. El contraste con el modelo LSTM es bastante claro, ya que en ese caso las perturbaciones térmicas producían variaciones mucho más marcadas en la respuesta predicha.

4.5.3. Síntesis del análisis PDP+ICE y perturbaciones

En función de los resultados obtenidos, el análisis conjunto mediante curvas PDP, ICE y perturbaciones permitió identificar que cada arquitectura aprende relaciones diferentes entre las variables meteorológicas y la respuesta hidrológica de la cuenca, utilizando distintas escalas temporales y diferentes niveles de dependencia del contexto hidrológico. En el caso de RF, las relaciones observadas tendieron a separar con mayor claridad las variables asociadas al almacenamiento antecedente de aquellas vinculadas a señales térmicas más inmediatas relacionadas con procesos de fusión. Además, el análisis de perturbaciones mostró una respuesta marcadamente asimétrica frente a cambios uniformes, donde reducciones en la acumulación previa generaban efectos más notorios que incrementos equivalentes. Esto deja ver que el modelo reacciona con mayor sensibilidad frente a pérdidas de almacenamiento que frente a aumentos graduales de las condiciones antecedentes.

El modelo LSTM presentó un comportamiento distinto, mucho más ligado a la memoria temporal del

sistema. En este caso, las variables térmicas no actúan como umbrales fijos, sino como señales cuyo efecto cambia según el estado previo de la cuenca. Esa dependencia temporal también apareció en el análisis de sensibilidad, donde pequeñas perturbaciones térmicas generaron variaciones importantes en el caudal predicho. Más que responder únicamente al valor instantáneo de las variables, el modelo incorpora la secuencia y persistencia de las condiciones previas dentro de su representación hidrológica. Por su parte, la CNN tendió a mostrar respuestas promedio más graduales en varias de sus variables influyentes, particularmente en acumulaciones y promedios temporales. A esto se suman sensibilidades más estables y aproximadamente simétricas frente a perturbaciones positivas y negativas, dando la impresión de una representación más suavizada de ciertas condiciones antecedentes dentro de la ventana temporal analizada. Los resultados indican que CNN parece haber aprendido a representar principalmente la condición media del sistema hidrológico, privilegiando el efecto de variables de baja frecuencia y de carácter acumulativo por sobre aquellas asociadas a cambios rápidos del caudal.

A su vez, los resultados muestran que una variable altamente importante según el análisis de permutación no necesariamente genera respuestas PDP intensas ni variaciones abruptas sobre el caudal. En distintos casos, variables identificadas como muy influyentes exhibieron relaciones promedio relativamente suaves e incluso cercanas a respuestas planas. Sin embargo, esto no significa que la variable tenga un rol menor dentro del modelo. En varias situaciones, las trayectorias ICE revelaron respuestas individuales mucho más marcadas que terminaban suavizándose al observar la curva promedio representada en la PDP. El análisis de perturbaciones complementó esta interpretación, mostrando que algunas de estas variables seguían produciendo modificaciones relevantes en el caudal cuando eran alteradas de manera controlada. En otras palabras, la importancia de una variable no depende únicamente de cuánto modifica el valor promedio del caudal, sino también de cómo contribuye a representar distintos estados hidrológicos y patrones de interacción dentro del sistema.

Además, las curvas ICE permitieron observar que muchas de estas relaciones no siguen un comportamiento homogéneo entre trayectorias, ya que la respuesta del modelo cambia considerablemente según el contexto hidrológico en que se evalúa cada caso. Esto se hizo especialmente evidente en variables térmicas y energéticas, donde algunas trayectorias reaccionaban de forma marcada mientras otras permanecían estables. En cuencas nivo glaciales, este comportamiento resulta coherente con la propia dinámica del sistema, considerando que variables como temperatura o radiación rara vez actúan de manera aislada. Su efecto depende en gran medida de la disponibilidad previa de nieve, hielo o agua almacenada dentro de la cuenca.

Finalmente, resulta importante distinguir entre la suavidad observada en algunas curvas PDP y el comportamiento predictivo global de los modelos. Aunque la CNN presentó respuestas promedio medianamente suaves para varias de sus variables más influyentes, eso no implica necesariamente que su dinámica hidrológica global también sea suave. En análisis anteriores aparecieron amplificaciones de magnitud y desfases temporales durante ciertos eventos, especialmente en crecidas importantes. Por lo tanto, la suavidad observada en las curvas PDP debe interpretarse únicamente como una característica de la respuesta promedio frente a determinadas variables y no como una representación completa del comportamiento predictivo del modelo.

5. Discusión

5.1. Complejidad arquitectónica y representación hidrológica

Las configuraciones finales obtenidas para las arquitecturas evaluadas muestran que el desempeño de los modelos no estuvo necesariamente ligado a un aumento extremo de complejidad estructural. En los tres casos analizados, las configuraciones óptimas correspondieron a arquitecturas compactas, capaces de capturar gran parte de la dinámica hidrológica dominante de la cuenca sin recurrir a redes excesivamente profundas ni a modelos con una cantidad muy elevada de parámetros. Este comportamiento resulta particularmente interesante en este tipo de cuencas, donde muchos de los procesos hidrológicos mantienen cierta organización temporal asociada a la acumulación de nieve, el almacenamiento antecedente y los ciclos estacionales de fusión.

En el caso de LSTM, las pruebas realizadas durante el ajuste mostraron que aumentar demasiado la longitud de memoria o el tamaño interno de la red no se tradujo en mejoras consistentes del desempeño. De hecho, las configuraciones más extensas tendieron a aumentar la variabilidad entre ejecuciones y a reducir la estabilidad del modelo. Para la cuenca estudiada, gran parte de la información temporal relevante pareció concentrarse en escalas más bien definidas, por lo que incorporar secuencias demasiado largas no aportó relaciones significativamente nuevas para la predicción. El uso de variables acumuladas y promedios móviles en distintas ventanas temporales probablemente permitió incorporar parte importante de la memoria hidrológica directamente en los datos de entrada, disminuyendo la necesidad de utilizar secuencias extensas para representar procesos de almacenamiento y persistencia temporal.

Una situación parecida se observó en CNN y RF. A pesar de utilizar estrategias de aprendizaje muy distintas, ambas arquitecturas lograron desempeños competitivos mediante configuraciones moderadas. Al igual que en LSTM, el uso de variables acumuladas y promedios móviles en distintas escalas temporales facilitó el aprendizaje de relaciones hidrológicas relevantes, especialmente aquellas asociadas a respuestas retardadas del sistema. En cierta medida, parte de la información temporal que normalmente tendría que ser aprendida internamente por arquitecturas más complejas ya estaba representada previamente en las series utilizadas para el entrenamiento. Bajo este escenario, no solo LSTM se vio beneficiado por la presencia de memoria hidrológica en los datos, sino también arquitecturas como RF y CNN, que lograron capturar relaciones temporales importantes sin requerir estructuras excesivamente profundas.

Los resultados muestran que no fue casualidad que los tres modelos mejoraran su desempeño al incorporar variables acumuladas y promedios móviles en distintas escalas temporales. Este tipo de representación parece especialmente útil en sistemas hidrológicos donde los procesos de almacenamiento, acumulación y fusión presentan respuestas retardadas en el tiempo. Más allá de la arquitectura utilizada, la incorporación de variables construidas bajo una lógica hidrológica permitió entregar información temporal ya organizada, facilitando la identificación de patrones relevantes para la predicción de caudales.

5.2. Desempeño bajo distintos regímenes hidrológicos

El análisis de años secos, normales y húmedos muestra que el desempeño de los modelos no depende únicamente de la arquitectura utilizada, sino también del régimen hidrológico bajo el cual son evaluados. A medida que cambian la magnitud de los caudales, la persistencia de las recesiones o la intensidad de los eventos anuales también cambia el tipo de dificultad que cada modelo debe enfrentar. Desde

esa perspectiva, los resultados evidencian que las exigencias de modelación no son constantes entre regímenes, ya que cada escenario parece requerir distintas capacidades para representar almacenamiento, variabilidad y magnitud del caudal.

Uno de los patrones más notorios es que, bajo condiciones húmedas, las arquitecturas de ML tienden a presentar desempeños relativamente consistentes y cercanos entre sí. Esto resulta interesante porque indica que las arquitecturas evaluadas responden mejor cuando la señal hidrológica es más intensa y presenta patrones más definidos. Bajo condiciones de mayor disponibilidad hídrica, las relaciones entre las variables meteorológicas y el caudal parecen volverse más reconocibles para los modelos, facilitando el aprendizaje de la dinámica anual. Bajo esas condiciones, parte importante del desempeño de los enfoques de aprendizaje automático parece depender de qué tan clara y estructurada es la señal hidrológica disponible, y no solo de la complejidad de la arquitectura utilizada. Sin embargo, los años húmedos también dejaron en evidencia que representar correctamente la magnitud de los eventos sigue siendo uno de los principales desafíos. En varios casos, los modelos mantuvieron correlaciones elevadas aun cuando persistían diferencias importantes en la intensidad de los caudales y en el volumen total representado por los modelos. Esto muestra que seguir adecuadamente la evolución temporal del hidrograma puede ser más sencillo que reproducir correctamente la escala de las crecidas y la intensidad real de los aportes hídricos durante eventos de gran magnitud. Desde un punto de vista hidrológico, el problema no parece estar únicamente en identificar cuándo ocurre un evento, sino en cuantificar de forma adecuada cuánto agua moviliza realmente el sistema.

Por otro lado, los años secos presentaron un comportamiento bastante más heterogéneo y, en general, más difícil de representar tanto para los modelos de ML como para los enfoques físicos evaluados. Este resultado es relevante porque muestra que las limitaciones observadas no dependen únicamente del tipo de arquitectura utilizada, sino también de la complejidad propia del régimen seco. Bajo condiciones de bajo caudal, el sistema depende mucho más de procesos lentos asociados al agua almacenada en la cuenca y a la forma en que esta se libera progresivamente durante los períodos de estiaje, mientras que la señal hidrológica disponible se vuelve más débil y sensible a pequeñas perturbaciones. Bajo estas condiciones, incluso pequeñas desviaciones en la magnitud simulada pueden deteriorar de manera importante métricas globales como el KGE. La dificultad observada durante los años secos también podría indicar que las variables meteorológicas utilizadas no contienen toda la información necesaria para representar adecuadamente este tipo de régimen. Esto abre una discusión relevante sobre la posible incorporación de variables adicionales asociadas al almacenamiento, humedad antecedente o procesos subterráneos, especialmente considerando que escenarios más secos podrían volverse cada vez más frecuentes en cuencas de alta montaña bajo condiciones futuras de cambio climático.

Este comportamiento también apareció en la curva de duración de caudales, particularmente en el tramo asociado a altas probabilidades de excedencia. En esa zona, los modelos de ML tendieron a mantenerse por sobre la curva observada, amplificando parcialmente la señal de estiaje, mientras que los modelos físicos mostraron una tendencia más marcada a subestimar los caudales mínimos. Más allá de las diferencias métricas, este comportamiento refleja que ambos enfoques representan de manera distinta los procesos de agotamiento y almacenamiento hídrico durante períodos secos. Mientras los modelos físicos tienden a hacer caer el caudal más rápidamente durante los períodos de estiaje, las arquitecturas de ML logran mantener parte de la variabilidad observada incluso bajo condiciones de bajo caudal.

En los años normales, las diferencias entre modelos dejaron de estar marcadas principalmente por la sincronización temporal y pasaron a depender, sobre todo, de la forma en que cada arquitectura representó la magnitud y variabilidad del caudal. En varios casos, arquitecturas con sincronizaciones similares presentaron desempeños globales distintos debido a pequeñas amplificaciones o atenuaciones

sistemáticas de la señal anual. Lo interesante es que una correlación alta no garantiza necesariamente una representación hidrológica adecuada, sobre todo cuando persisten distorsiones en la magnitud del caudal simulado.

Dentro de este comportamiento variable, el mejor desempeño promedio alcanzado por LSTM en los años normales resulta particularmente interesante porque aparece precisamente bajo condiciones donde la señal hidrológica no está dominada ni por extremos intensos ni por caudales excesivamente bajos. En este régimen intermedio, las diferencias entre modelos parecen depender menos de la capacidad de reaccionar frente a eventos aislados y más de mantener equilibrio entre persistencia temporal, variabilidad y volumen acumulado a lo largo del año. Desde esta perspectiva, los años normales permiten evaluar con mayor claridad qué arquitecturas logran representar de manera más equilibrada el funcionamiento hidrológico de la cuenca.

Otro aspecto relevante es que no existió una arquitectura claramente dominante para todos los regímenes hidrológicos evaluados. Algunos modelos mostraron respuestas más estables durante años húmedos, mientras otros lograron desempeños relativamente mejores bajo condiciones secas o intermedias. Esto abre una discusión metodológica importante, ya que evaluar modelos únicamente mediante métricas promedio globales puede ocultar fortalezas y debilidades que solo aparecen bajo ciertos contextos hidrológicos específicos.

Dentro de este comportamiento variable, el modelo LSTM destacó por mantener una respuesta más equilibrada frente a distintos regímenes. Aunque en algunos años también presentó deterioros importantes, en términos generales logró conservar desempeños consistentes tanto en años secos como normales y húmedos. Más que sobresalir únicamente en un escenario particular, su principal fortaleza parece estar asociada a una mayor capacidad de adaptación frente a cambios en la estructura temporal y en la intensidad de la señal hidrológica anual. En varios casos consiguió mantener una representación razonable de los eventos principales incluso cuando la dinámica de la cuenca se volvía más compleja o menos persistente.

5.3. Representación de dinámicas hidrológicas contrastantes

Los eventos hidrológicos contrastantes analizados muestran que modelos con desempeños globales similares pueden responder de manera muy distinta frente a anomalías, transiciones estacionales o desviaciones acumulativas del régimen anual. Esto es relevante porque evidencia que un análisis de KGE no siempre alcanzan a reflejar cómo cada arquitectura representa realmente la estabilidad y variabilidad del sistema hidrológico. Bajo condiciones promedio, muchas de estas diferencias pasan desapercibidas, pero aparecen con mayor claridad cuando el sistema enfrenta eventos extremos, cambios progresivos de régimen o períodos con comportamientos menos persistentes.

El evento extremo observado durante el año 2013 resulta particularmente interesante porque, aunque ninguna arquitectura logra reproducir completamente la sobreamplificación registrada en el caudal observado. Este resultado indica que los modelos de ML construyen sus predicciones a partir de los patrones predominantes presentes en los datos de entrenamiento, por lo que cuando contienen eventos excepcionales o poco representados en ese conjunto, su capacidad para estimar correctamente la magnitud del caudal disminuye. Por ello, las diferencias observadas no pueden atribuirse únicamente a las características de cada arquitectura, sino también a la ausencia de este tipo de eventos durante el entrenamiento.

La transición seca-húmeda observada durante el año 2008 permitió analizar un escenario distinto,

donde las diferencias entre modelos no aparecen asociadas principalmente a eventos extremos aislados, sino a la capacidad de representar cambios progresivos de régimen. Mientras algunos modelos logran seguir razonablemente el ascenso gradual del caudal hacia el último trimestre del año, otros reaccionan más tarde o mantienen respuestas demasiado amortiguadas. Esto muestra que representar transiciones hidrológicas sostenidas requiere capacidades distintas a las necesarias para reproducir eventos extremos puntuales. Mientras las crecidas puntuales dependen, en gran medida, de la capacidad del modelo para responder a perturbaciones específicas, mientras que las transiciones de régimen requieren sostener una representación coherente de la evolución temporal del sistema, incluyendo cambios graduales en la acumulación, almacenamiento y liberación de agua dentro de la cuenca.

Los años 2015 y 2019 exponen una dificultad distinta a la observada en escenarios dominados por extremos o cambios abruptos de régimen. En estos casos, la mayoría de las arquitecturas consigue reproducir de manera razonable la forma general del hidrograma y seguir correctamente la evolución temporal de la señal. Sin embargo, pequeñas diferencias sostenidas en la magnitud del caudal terminan acumulándose a lo largo del año y generan discrepancias importantes en el volumen anual simulado. Lo interesante de este comportamiento es que las diferencias entre modelos no aparecen asociadas a fallas evidentes o eventos particularmente conflictivos, sino a desajustes mucho más sutiles y persistentes. En varios casos, los hidrogramas mantienen una apariencia coherente e incluso una buena sincronización temporal, pero leves sobreestimaciones o atenuaciones sostenidas modifican gradualmente la respuesta acumulada del sistema. Esto muestra que una simulación visualmente adecuada no siempre garantiza una representación correcta del balance hídrico anual, especialmente cuando los errores en la magnitud del caudal se mantienen durante períodos prolongados.

5.4. Implicancias del análisis de sensibilidad e interpretabilidad

Uno de los hallazgos más relevantes del análisis mediante curvas PDP, ICE y perturbaciones corresponde a que las arquitecturas evaluadas lograron representar relaciones coherentes con el funcionamiento hidrológico esperado de la cuenca aun sin incorporar ecuaciones físicas explícitas. Variables asociadas al aporte de masa, como la precipitación acumulada, tendieron a relacionarse con condiciones de almacenamiento previo y mayor disponibilidad de agua dentro del sistema. En cambio, las variables térmicas y energéticas mostraron respuestas más ligadas a procesos de activación de fusión y a cambios en las condiciones energéticas de la cuenca. Resulta muy interesante que estas relaciones emergieran a partir de patrones estadísticos aprendidos desde los datos y no desde formulaciones hidrológicas predefinidas, evidenciando que los enfoques de ML pueden reconstruir parte importante de la organización física dominante del sistema hidrológico mediante una base de datos bien estructurada.

Esta capacidad adquiere especial interés en escenarios donde el comportamiento hidrológico de la cuenca puede cambiar progresivamente con el tiempo. En sistemas de montaña, pequeñas modificaciones en la acumulación de nieve, la duración de los períodos de fusión o la persistencia de los estiajes pueden alterar las relaciones tradicionales entre las variables meteorológicas y el caudal. Bajo ese tipo de condiciones, la capacidad de las arquitecturas de ML para aprender patrones directamente desde los datos podría facilitar una adaptación más flexible frente a señales hidrológicas que ya no responden exactamente al comportamiento observado durante los períodos de calibración. En lugar de depender únicamente de relaciones fijas previamente definidas, estos enfoques pueden reajustar internamente las relaciones aprendidas a medida que evoluciona la estructura de la señal hidrológica disponible.

Otro aspecto importante corresponde a que las distintas arquitecturas no solo identificaron variables influyentes diferentes, sino que además representaron el sistema hidrológico utilizando enfoques implícitos distintos entre sí. Mientras algunos modelos parecieron responder con mayor sensibilidad

frente a almacenamiento antecedente, otros privilegiaron relaciones más asociadas a memoria temporal o patrones progresivos de evolución de la señal hidrológica. Aun así, varias de estas arquitecturas lograron desempeños competitivos en la predicción de caudales, lo que resulta particularmente interesante porque muestra que no parece existir una única forma de representar adecuadamente la dinámica hidrológica de la cuenca. Desde esta perspectiva, las diferencias observadas entre arquitecturas no necesariamente deben interpretarse únicamente como ventajas o desventajas predictivas, sino también como distintas formas de aproximarse al mismo sistema hidrológico. Esto abre una discusión relevante respecto al potencial de utilizar enfoques complementarios para explorar distintas dimensiones de la dinámica de una cuenca, particularmente en sistemas complejos donde la respuesta hidrológica no depende de un único proceso dominante. En cuencas de montaña, por ejemplo, observamos que la generación de escorrentía resulta de la interacción entre acumulación nival, fusión, almacenamiento temporal de agua y persistencia de condiciones meteorológicas previas, procesos que además operan en distintas escalas temporales. Bajo este tipo de dinámicas, distintas arquitecturas pueden terminar capturando componentes diferentes del funcionamiento hidrológico.

Aunque las arquitecturas representaron el funcionamiento hidrológico mediante enfoques diferentes, todas coincidieron en un aspecto importante, ya que la precipitación diaria no figuró entre las variables de mayor influencia. En todos los modelos, las variables más relevantes correspondieron principalmente a acumulaciones y promedios calculados sobre ventanas temporales extensas, por lo que el caudal estimado estuvo más condicionado por el estado antecedente de la cuenca que por forzantes meteorológicas de respuesta inmediata.

6. Conclusiones

La simulación de caudales en cuencas de alta montaña representa un desafío debido a la fuerte influencia de procesos de acumulación y liberación de nieve, así como a la alta variabilidad climática que caracteriza estos sistemas. La comparación realizada entre modelos hidrológicos basados en procesos y enfoques de aprendizaje automático permitió evaluar distintas formas de representar la dinámica hidrológica de una cuenca con régimen nival-glaciar en Chile central. A partir de los resultados obtenidos, se evidencia que los modelos de ML logran capturar con mayor precisión las relaciones no lineales entre variables hidrometeorológicas y caudal.

En términos globales, los resultados muestran que los modelos de ML alcanzaron desempeños predictivos superiores a las dos versiones modelo físico USM en termino de métricas como KGE, RMSE y MAE. El modelo LSTM presentó el mejor desempeño general y mayor estabilidad interanual, seguido por CNN y RF. Sin embargo, los resultados mostraron que las diferencias entre arquitecturas no pueden interpretarse únicamente desde un ranking global de desempeño. Mientras algunas configuraciones mostraron respuestas más consistentes bajo señales hidrológicas intensas, otras presentaron comportamientos relativamente más estables frente a condiciones secas o intermedias. Esto evidenció que la dificultad del problema cambia según el régimen hidrológico dominante y que las fortalezas de cada modelo dependen fuertemente del tipo de dinámica que debe representar.

El contraste observado entre los modelos físicos y las arquitecturas de ML mostró diferencias importantes en la forma de representar la dinámica hidrológica de la cuenca. Mientras el modelo USM mantiene una representación explícita de procesos como acumulación nival, fusión y escorrentía, los resultados evidenciaron que ciertos eventos y cambios de régimen tendieron a degradar con mayor facilidad su desempeño relativo frente a las arquitecturas de aprendizaje automático. Esto se hizo particularmente visible en escenarios donde pequeñas distorsiones en la magnitud del caudal o en la variabilidad anual producían deterioros importantes en métricas globales. Desde esta perspectiva, los resultados

sugieren que las diferencias entre enfoques no dependen únicamente de la precisión predictiva alcanzada, sino también del grado de flexibilidad con que cada modelo logra adaptarse a dinámicas hidrológicas complejas o cambiantes dentro de la cuenca. Mientras los modelos físicos representan explícitamente procesos internos mediante formulaciones predefinidas, las arquitecturas de ML mostraron una mayor capacidad para ajustarse a relaciones emergentes presentes en los datos bajo las condiciones evaluadas.

El modelo LSTM destacó por mantener un comportamiento más consistente entre distintos años y por representar con mayor estabilidad la evolución temporal del caudal, especialmente durante los períodos de mayor variabilidad estacional. Parte de este comportamiento podría estar relacionada con su capacidad para incorporar dependencias temporales de más largo plazo dentro de la dinámica hidrológica de la cuenca, permitiéndole conservar con mayor continuidad la información asociada a procesos de acumulación y respuesta progresiva del sistema. El modelo CNN también alcanzó un buen desempeño global, aunque con una mayor variabilidad entre años en comparación con LSTM y RF, ya que en algunos periodos el modelo mostró mayor ruido o pequeñas desviaciones en la secuencia del hidrograma, lo que se reflejó en un desempeño más irregular en ciertos años del periodo analizado. Por su parte, RF mostró una representación más estable de los patrones medios del caudal y de la interacción entre variables meteorológicas e hidrológicas, particularmente mediante el uso de variables acumuladas y rezagadas que permitieron incorporar información antecedente del sistema dentro de las entradas del modelo. Sin embargo, se observaron algunas debilidades del modelo frente a ciertos eventos extremos aislados, los cuales tendieron a ser suavizados o subestimados.

El análisis interanual mostró que el desempeño de los modelos cambia de manera importante según el régimen hidrológico dominante. En los años húmedos, las arquitecturas de ML mantuvieron desempeños relativamente altos y consistentes entre sí, mostrando una buena capacidad para representar señales hidrológicas intensas y estructuradas. Dentro de este régimen, RF presentó el mayor desempeño promedio, aunque con diferencias relativamente acotadas respecto a LSTM y CNN. En los años normales, el modelo LSTM destacó con mayor claridad frente al resto de las arquitecturas, manteniendo el comportamiento más equilibrado entre correlación, variabilidad y volumen medio. En los años secos, en cambio, el desempeño general disminuyó tanto para modelos físicos como para enfoques basados en datos, aunque RF mostró una representación relativamente más estable frente a señales de baja magnitud.

Los resultados también mostraron que ciertos eventos extremos aislados fueron difíciles de representar tanto para modelos físicos como para arquitecturas de ML. En particular, algunos máximos observados tendieron a aparecer sobreamplificados respecto al comportamiento general de la serie y no lograron ser reproducidos completamente por los distintos enfoques evaluados, aun cuando varios modelos sí reaccionaban frente a la perturbación hidrológica. Esto resulta especialmente interesante porque indica que la dificultad no parece depender únicamente de la arquitectura utilizada, sino también de qué tan representadas se encuentran esas dinámicas dentro del período de entrenamiento y de la coherencia entre las variables meteorológicas disponibles y la magnitud real del evento observado. Desde esta perspectiva, el análisis interanual permitió evidenciar que métricas globales relativamente buenas no siempre garantizan una representación adecuada de eventos excepcionales dentro del régimen hidrológico.

El análisis mediante curvas PDP, trayectorias ICE y perturbaciones permitió ver que los modelos no aprendieron relaciones arbitrarias entre las variables y el caudal. En los resultados apareció de forma bastante clara que las arquitecturas lograban distinguir cuándo una variable estaba asociada al aporte directo de agua al sistema y cuándo su efecto dependía más bien de condiciones energéticas o de almacenamiento previo dentro de la cuenca. Variables como la precipitación acumulada tendieron a

relacionarse con disponibilidad hídrica y almacenamiento antecedente, mientras que temperatura y radiación mostraron respuestas más ligadas a procesos de activación de fusión. También se observó que varias respuestas cambiaban según el estado hidrológico previo de la cuenca, evidenciando que los modelos no reaccionaban únicamente al valor instantáneo de una variable, sino al contexto completo en que esa condición ocurría.

En varias variables donde el efecto sobre la generación de caudal era más evidente, también aparecieron zonas de activación relativamente marcadas dentro de las curvas analizadas, especialmente en variables térmicas asociadas a procesos de fusión. Esto permitió observar que los modelos identificaban ciertos rangos donde pequeñas variaciones comenzaban a generar cambios más notorios en la respuesta hidrológica, algo coherente con el comportamiento esperado en una cuenca de régimen nivo-glaciar. En ese sentido, uno de los resultados más interesantes del trabajo es que estas relaciones emergieron únicamente a partir de los datos, sin que las arquitecturas incorporaran explícitamente procesos físicos de acumulación, fusión o escorrentía. Más que aprender una relación fija entre variables y caudal, los modelos parecieron reconocer parte de la lógica hidrológica dominante del sistema, incluyendo procesos asociados a almacenamiento, persistencia temporal y activación de umbrales de respuesta.

En este estudio, más allá de identificar qué modelo obtuvo el mejor desempeño bajo las condiciones evaluadas, uno de los aspectos más interesantes observados en este trabajo es que los enfoques físicos y las arquitecturas de ML no necesariamente aparecen como herramientas excluyentes entre sí. Bajo distintos escenarios hidrológicos, ambos enfoques mostraron fortalezas diferentes, ya que, mientras los modelos físicos permitieron mantener una representación explícita de procesos internos de la cuenca, varias arquitecturas de ML mostraron una mejor capacidad para adaptarse a cambios en la magnitud del caudal, variaciones interanuales y ciertos comportamientos complejos de la señal hidrológica.

Más que la búsqueda de un único modelo óptimo, el análisis deja ver la importancia de comprender las fortalezas y limitaciones de cada arquitectura para aprovechar de mejor manera su potencial predictivo. El desempeño observado indica que distintos modelos capturan aspectos complementarios de la dinámica hidrológica, por lo que la exploración de estrategias de combinación o enfoques híbridos aparece como una línea prometedora para futuras investigaciones. La integración de modelos con distintas capacidades, Por ejemplo, arquitecturas con buena representación de la dinámica temporal podrían complementarse con enfoques más robustos para capturar relaciones no lineales o representar eventos extremos, permitiendo ensamblar esquemas predictivos más equilibrados. Este tipo de estrategias abre la posibilidad de desarrollar modelos que no solo mantengan una adecuada representación de la evolución temporal del hidrograma, sino que también mejoren la capacidad de respuesta frente a eventos de mayor magnitud.

Asimismo, los resultados obtenidos muestran que este tipo de enfoques podría resultar especialmente útil en estudios futuros orientados a evaluar múltiples cuencas o sistemas hidrológicos con dinámicas similares. Sin embargo, avanzar hacia aplicaciones regionales requiere previamente disponer de bases de datos suficientemente representativas y consistentes que permitan capturar adecuadamente la diversidad hidrológica entre cuencas. En el contexto de la hidrología chilena, donde la aplicación sistemática de técnicas de aprendizaje automático aún es escasa, por lo que este tipo de estudios contribuye a ampliar el repertorio metodológico disponible y a evaluar de manera crítica las ventajas y limitaciones de cada enfoque.

Para finalizar, los resultados obtenidos muestran que los modelos de ML constituyen herramientas predictivas altamente eficaces para la simulación de caudales en cuencas complejas de régimen nival y glaciar, pudiendo superar a enfoques físico en desempeño global bajo las condiciones evaluadas. La

aplicación de estos enfoques ya sea de forma independiente o integrados con modelos de base física, ofrece nuevas oportunidades para mejorar la capacidad predictiva de los sistemas hidrológicos y aprovechar de manera más eficiente la información contenida en los registros hidrometeorológicos. En un contexto de creciente variabilidad hidroclimática a nivel mundial, que afectan la disponibilidad de agua en regiones de montaña y en las distintas partes del mundo, por lo que el desarrollo y la integración de estas herramientas adquiere especial relevancia para fortalecer la capacidad de análisis, planificación y gestión de los recursos hídricos en escenarios futuros.

7. Agradecimientos

Esta memoria se realizó en el marco de dos proyectos de investigación: Proyecto FONDECYT Iniciación en investigación No.11220482 y el proyecto USM Interno Multidisciplinario PI_M_24_03.

Asimismo, quisiera expresar mi especial agradecimiento al profesor Luis De La Fuente Contreras y a la profesora Lina Castro Heredia por su constante apoyo, orientación y disposición a lo largo del desarrollo de esta memoria. Su acompañamiento durante este proceso no solo fue fundamental para el desarrollo de la investigación, sino también para mi formación académica y profesional. Del mismo modo, agradezco la oportunidad de haber formado parte de este proyecto de investigación, experiencia que me permitió profundizar significativamente en el área de la modelación hidrológica y las técnicas de aprendizaje automático, ampliando mi mirada sobre los desafíos y posibilidades que existen en este campo.

8. Referencias

- Alvarez-Garretón, C., Mendoza, P. A., Pablo Boisier, J., Addor, N., Galleguillos, M., Zambrano-Bigiarini, M., Lara, A., Puelma, C., Cortes, G., Garreaud, R., McPhee, J., & Ayala, A. (2018). The CAMELS-CL dataset: Catchment attributes and meteorology for large sample studies-Chile dataset. *Hydrology and Earth System Sciences*, 22(11), 5817–5846. <https://doi.org/10.5194/hess-22-5817-2018>
- Anderson, S., & Radić, V. (2022). Interpreting Deep Machine Learning for Streamflow Modeling Across Glacial, Nival, and Pluvial Regimes in Southwestern Canada. *Frontiers in Water*, 4. <https://doi.org/10.3389/frwa.2022.934709>
- Anilkumar, R., Bharti, R., Chutia, D., & Aggarwal, S. P. (2023). Modelling point mass balance for the glaciers of the Central European Alps using machine learning techniques. *Cryosphere*, 17(7), 2811–2828. <https://doi.org/10.5194/tc-17-2811-2023>
- Blöschl, G., Bierkens, M. F. P., Chambel, A., Cudennec, C., Destouni, G., Fiori, A., Kirchner, J. W., McDonnell, J. J., Savenije, H. H. G., Sivapalan, M., Stumpp, C., Toth, E., Volpi, E., Carr, G., Lupton, C., Salinas, J., Széles, B., Viglione, A., Aksoy, H., ... Zhang, Y. (2019). Twenty-three unsolved problems in hydrology (UPH)—a community perspective. *Hydrological Sciences Journal*, 64(10), 1141–1158. <https://doi.org/10.1080/02626667.2019.1620507>
- Breiman. (2001). *Random Forests*. *Machine Learning*, 45(1), 5-32.
- Chai, T., & Draxler, R. R. (2014). Root mean square error (RMSE) or mean absolute error (MAE)? -Arguments against avoiding RMSE in the literature. *Geoscientific Model Development*, 7(3), 1247–1250. <https://doi.org/10.5194/gmd-7-1247-2014>
- CLIMATE RESEARCH Clim Res. (n.d.). Retrieved www.int-res.com
- Dirección General de Aguas (DGA). (2023). Sistema Nacional de Información del Agua (SNIA): Base de datos hidrológicos de Chile. Ministerio de Obras Públicas, Chile. <https://snia.mop.gob.cl>
- Duan, S., Ullrich, P., & Shu, L. (2020). Using Convolutional Neural Networks for Streamflow Projection in California. *Frontiers in Water*, 2. <https://doi.org/10.3389/frwa.2020.00028>
- Fisher, A., Rudin, C., & Dominici, F. (2019). All Models are Wrong, but Many are Useful: Learning a Variable's Importance by Studying an Entire Class of Prediction Models Simultaneously. In *Journal of Machine Learning Research* (Vol. 20). <http://jmlr.org/papers/v20/18-760.html>.
- Francisco, M., Hidalgo, R., & Guía, P. (2025). *MODELO HIDROLÓGICO DISTRIBUIDO DE BASE FÍSICA EN PYTHON CON CAPACIDAD DE RUTEO PARA CUENCAS NIVOGIACIARES Memoria de Título presentada por Ingeniería Civil*.
- Friedman, J. H. (2001). 999 REITZ LECTURE GREEDY FUNCTION APPROXIMATION: A GRADIENT BOOSTING MACHINE 1. In *The Annals of Statistics* (Vol. 29, Number 5).
- Fu, M., Fan, T., Ding, Z., Salih, S. Q., Al-Ansari, N., & Yaseen, Z. M. (2020). Deep Learning Data-Intelligence Model Based on Adjusted Forecasting Window Scale: Application in Daily Streamflow Simulation. *IEEE Access*, 8, 32632–32651. <https://doi.org/10.1109/ACCESS.2020.2974406>

- Gauch, M., Kratzert, F., Klotz, D., Nearing, G., Lin, J., & Hochreiter, S. (2021). Rainfall-runoff prediction at multiple timescales with a single Long Short-Term Memory network. *Hydrology and Earth System Sciences*, 25(4), 2045–2062. <https://doi.org/10.5194/hess-25-2045-2021>
- Glorot, X., & Bengio, Y. (n.d.). *Understanding the difficulty of training deep feedforward neural networks*. Retrieved <http://www.iro.umontreal>.
- Goldstein, A., Kapelner, A., Bleich, J., & Pitkin, E. (2014). *Peeking Inside the Black Box: Visualizing Statistical Learning with Plots of Individual Conditional Expectation*. <http://arxiv.org/abs/1309.6392>
- Gupta, H. V., Kling, H., Yilmaz, K. K., & Martinez, G. F. (2009). Decomposition of the mean squared error and NSE performance criteria: Implications for improving hydrological modelling. *Journal of Hydrology*, 377(1–2), 80–91. <https://doi.org/10.1016/j.jhydrol.2009.08.003>
- Harris, C. R., Millman, K. J., van der Walt, S. J., Gommers, R., Virtanen, P., Cournapeau, D., Wieser, E., Taylor, J., Berg, S., Smith, N. J., Kern, R., Picus, M., Hoyer, S., van Kerkwijk, M. H., Brett, M., Haldane, A., del Río, J. F., Wiebe, M., Peterson, P., ... Oliphant, T. E. (2020). Array programming with NumPy. In *Nature* (Vol. 585, Number 7825, pp. 357–362). Nature Research. <https://doi.org/10.1038/s41586-020-2649-2>
- He, K., Zhang, X., Ren, S., & Sun, J. (n.d.). *Delving Deep into Rectifiers: Surpassing Human-Level Performance on ImageNet Classification*.
- Hochreiter, S., & Schmidhuber, J. (1997). Long Short-Term Memory. *Neural Computation*, 9(8), 1735–1780. <https://doi.org/10.1162/neco.1997.9.8.1735>
- Iqbal, M. (2012). *Introducción a la radiación solar* ([edición no disponible]). Elsevier Science. Recuperado de <https://www.perlego.com/book/1834395/an-introduction-to-solar-radiation-pdf> (Obra original publicada en 2012).
- Kingma, D. P., & Ba, J. (2017). *Adam: A Method for Stochastic Optimization*. <http://arxiv.org/abs/1412.6980>
- Kratzert, F., Klotz, D., Brenner, C., Schulz, K., & Herrnegger, M. (2018). Rainfall-runoff modelling using Long Short-Term Memory (LSTM) networks. *Hydrology and Earth System Sciences*, 22(11), 6005–6022. <https://doi.org/10.5194/hess-22-6005-2018>
- Kratzert, F., Klotz, D., Shalev, G., Klambauer, G., Hochreiter, S., & Nearing, G. (2019). Towards learning universal, regional, and local hydrological behaviors via machine learning applied to large-sample datasets. *Hydrology and Earth System Sciences*, 23(12), 5089–5110. <https://doi.org/10.5194/hess-23-5089-2019>
- lecun-89e. (n.d.).
- Li, Z., Liu, F., Yang, W., Peng, S., & Zhou, J. (2022). A Survey of Convolutional Neural Networks: Analysis, Applications, and Prospects. *IEEE Transactions on Neural Networks and Learning Systems*, 33(12), 6999–7019. <https://doi.org/10.1109/TNNLS.2021.3084827>
- Muñoz, R. C., Falvey, M. J., Arancibia, M., Astudillo, V. I., Elgueta, J., Ibarra, M., Santana, C., & Vásquez, C. (2018). Wind energy exploration over the atacama desert : A numerical model-guided observational program. *Bulletin of the American Meteorological Society*, 99(10), 2079–2092. <https://doi.org/10.1175/BAMS-D-17-0019.1>
- Nair, V., & Hinton, G. E. (n.d.). *Rectified Linear Units Improve Restricted Boltzmann Machines*.

- Nearing, G. S., Kratzert, F., Sampson, A. K., Pelissier, C. S., Klotz, D., Frame, J. M., Prieto, C., & Gupta, H. V. (2021). What Role Does Hydrological Science Play in the Age of Machine Learning? In *Water Resources Research* (Vol. 57, Number 3). Blackwell Publishing Ltd.
<https://doi.org/10.1029/2020WR028091>
- Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury Google, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., Desmaison, A., Xamla, A. K., Yang, E., Devito, Z., Raison Nabla, M., Tejani, A., Chilamkurthy, S., Ai, Q., Steiner, B., ... Chintala, S. (2019). *PyTorch: An Imperative Style, High-Performance Deep Learning Library*.
- Pedregosa FABIANPEDREGOSA, F., Michel, V., Grisel OLIVIERGRISEL, O., Blondel, M., Prettenhofer, P., Weiss, R., Vanderplas, J., Cournapeau, D., Pedregosa, F., Varoquaux, G., Gramfort, A., Thirion, B., Grisel, O., Dubourg, V., Passos, A., Brucher, M., Perrot and Édouardand, M., Duchesnay, and Édouard, & Duchesnay EDOUARDDUCHESNAY, Fré. (2011). Scikit-learn: Machine Learning in Python Gaël Varoquaux Bertrand Thirion Vincent Dubourg Alexandre Passos PEDREGOSA, VAROQUAUX, GRAMFORT ET AL. Matthieu Perrot. In *Journal of Machine Learning Research* (Vol. 12). <http://scikit-learn.sourceforge.net>.
- Pellicciotti, F., Brock, B., Strasser, U., Burlando, P., Funk, M., y Corripio, J. (2005). An enhanced temperature-index glacier melt model including the shortwave radiation balance: Development and testing for haut glacier d'arolla, switzerland. *Journal of Glaciology*, 51 (175), 573–587. doi: 10.3189/172756505781829124
- Pellicciotti, F., Helbing, J., Rivera, A., Favier, V., Corripio, J., Araos, J., . . . Carenzo, M. (2008, sep). A study of the energy balance and melt regime on juncal norte glacier, semi-arid andes of central chile, using melt models of different complexity. *Hydrological Processes*, 22 (19), 3980–3997. doi: 10.1002/hyp.7085
- Rozos, E., Dimitriadis, P., & Bellos, V. (2022). Machine learning in assessing the performance of hydrological models. *Hydrology*, 9(1). <https://doi.org/10.3390/hydrology9010005>
- Sanllehi. (2023). *Integración y evaluación de modelo de derretimiento y acumulación de nieve y hielo glaciar a modelo hidrológico usm*.
- Sarricolea, P., Herrera-Ossandon, M., & Meseguer-Ruiz, O. (2017). *Climatic regionalisation of continental Chile*. *Journal of Maps*, 13(2), 66–73.
<https://doi.org/10.1080/17445647.2016.1259592>
- Shen, C. (2018). A Transdisciplinary Review of Deep Learning Research and Its Relevance for Water Resources Scientists. In *Water Resources Research* (Vol. 54, Number 11, pp. 8558–8593). Blackwell Publishing Ltd. <https://doi.org/10.1029/2018WR022643>
- Shortridge, J. E., Guikema, S. D., & Zaitchik, B. F. (2016). Machine learning methods for empirical streamflow simulation: A comparison of model accuracy, interpretability, and uncertainty in seasonal watersheds. *Hydrology and Earth System Sciences*, 20(7), 2611–2628. <https://doi.org/10.5194/hess-20-2611-2016>
- Singh, D., Vardhan, M., Sahu, R., Chatterjee, D., Chauhan, P., & Liu, S. (2023). Machine-learning- and deep-learning-based streamflow prediction in a hilly catchment for future scenarios using CMIP6 GCM data. *Hydrology and Earth System Sciences*, 27(5), 1047–1075. <https://doi.org/10.5194/hess-27-1047-2023>
- Stowhas, L. (1975). *Simulation of Energy Budget and Melting Processes in Snowpack Systems under Data Constrains*. Davis: University of California

- Tarboton, D., y Luce, C. (1996). Utah energy balance snow accumulation and melt model (ueb). computer model technical description and users guide (Inf. Téc.).
- Yang, S., Yang, D., Chen, J., Santisirisomboon, J., Lu, W., & Zhao, B. (2020). A physical process and machine learning combined hydrological model for daily streamflow simulations of large watersheds with limited observation data. *Journal of Hydrology*, 590, 125206.
<https://doi.org/10.1016/J.JHYDROL.2020.125206>

9. Anexos

Tabla A.1. Métricas de desempeño durante el período de validación interna para los modelos RF, LSTM y CNN.

Métricas de entrenamiento	RF	LSTM	CNN
KGE	0.48	0.82	0.78
RSME	3.74	2.23	3.49
MAE	2.2	1.54	2.23

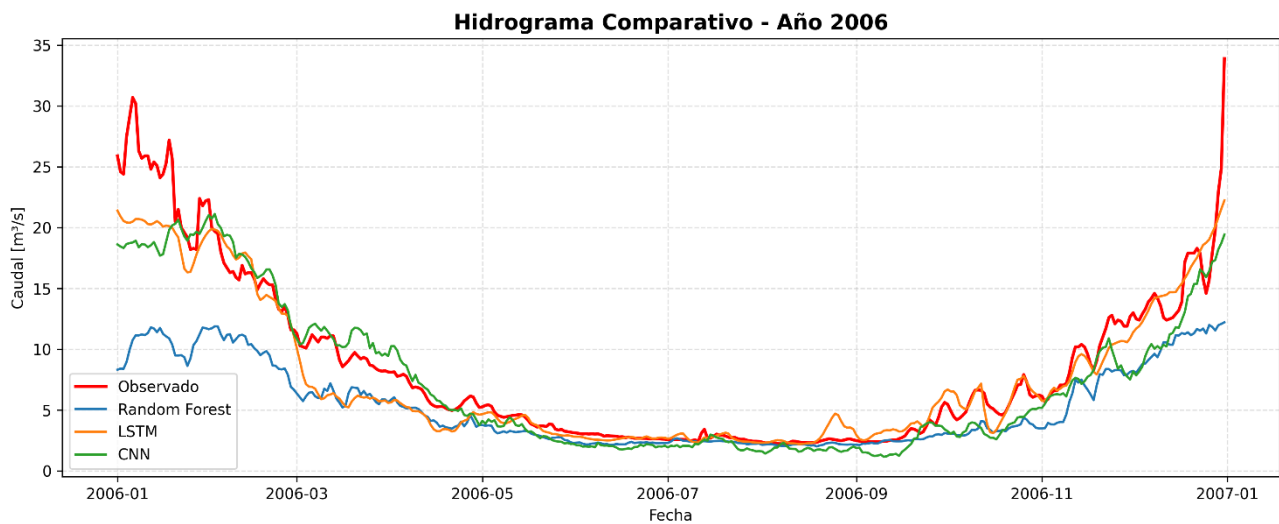


Figura A.1. Hidrograma comparativo entre caudal observado y simulado por los modelos de ML durante el período de validación interna correspondiente al año 2006.

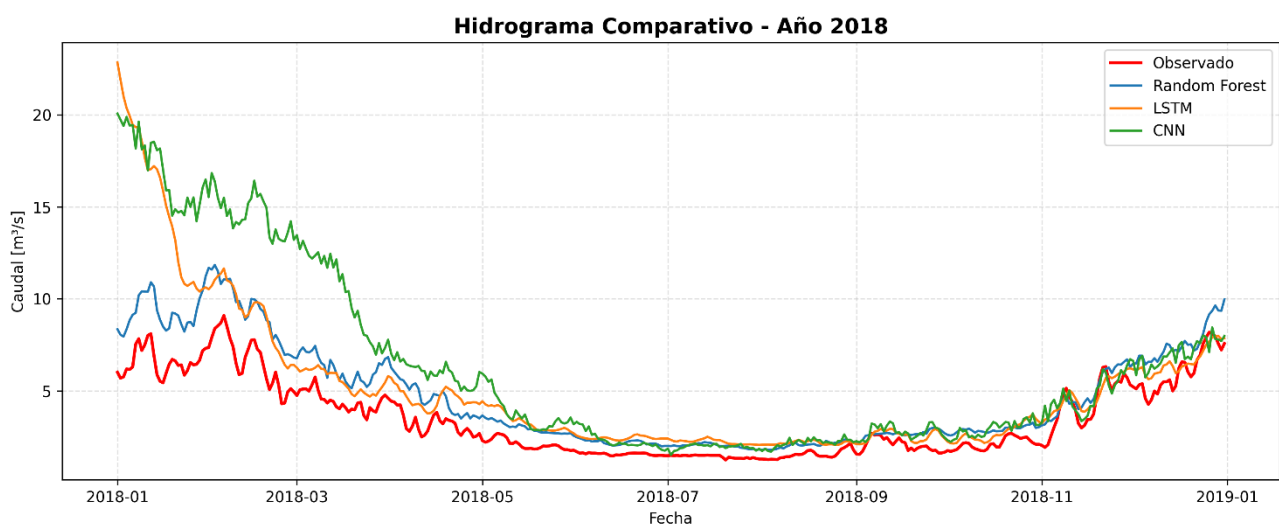


Figura A.2. Hidrograma comparativo entre caudal observado y simulado por los modelos de ML durante el período de validación interna correspondiente al año 2018.

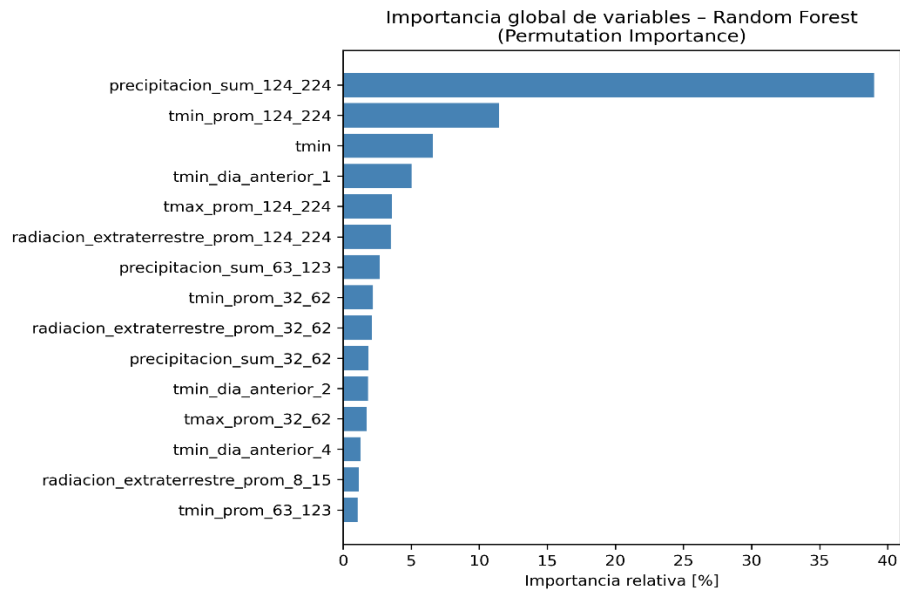


Figura A.3. Importancia global de variables estimada mediante el análisis de permutación para el modelo Random Forest.

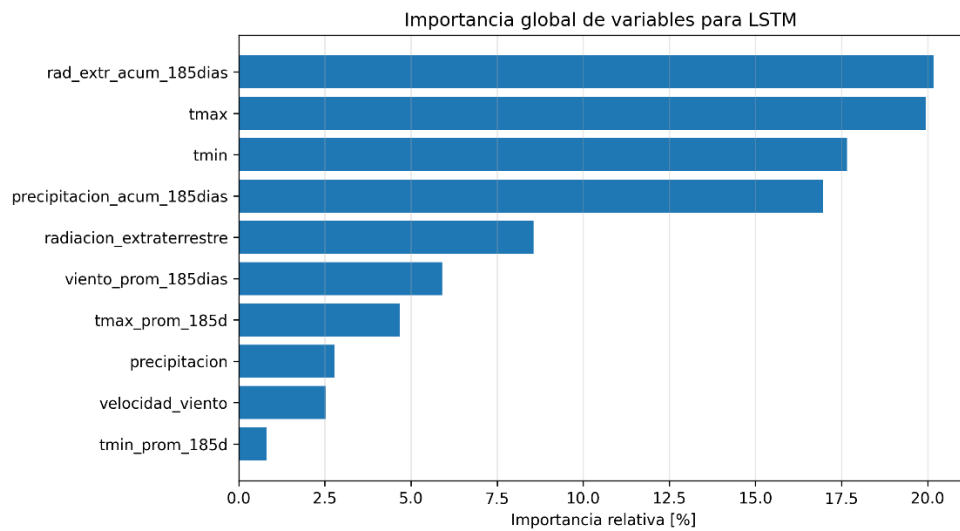


Figura A.4. Importancia global de variables estimada mediante el análisis de permutación para el modelo LSTM.

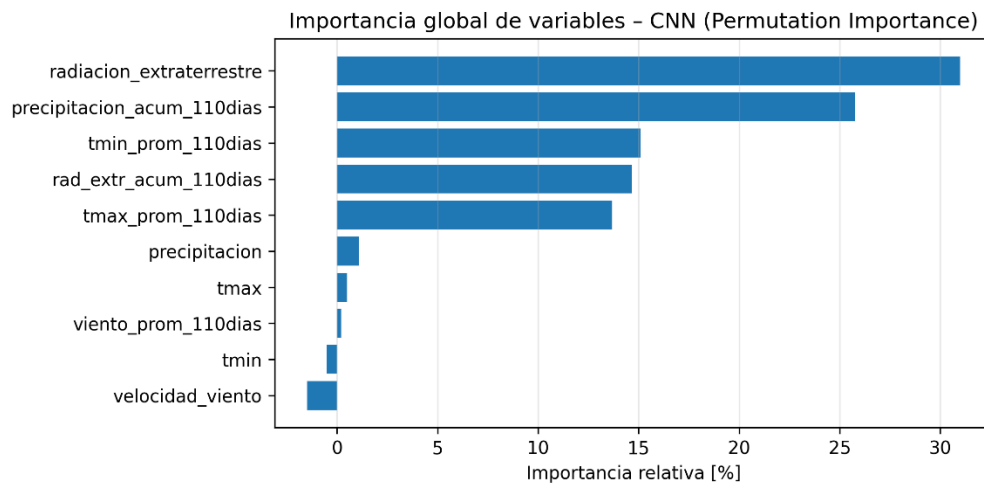


Figura A.5. Importancia global de variables estimada mediante el análisis de permutación para el modelo CNN.