UNIVERSIDAD DE COSTA RICA
SISTEMA DE ESTUDIOS DE POSGRADO

# PREDICCIÓN DEL CAUDAL PROMEDIO HORARIO DE LA ESTACIÓN HIDROLÓGICA PALMAR, UTILIZANDO MODELOS DE MACHINE LEARNING BASADOS EN ÁRBOLES DE DECISIÓN

Trabajo final de investigación aplicada sometido a la consideración de la Comisión del Programa de Estudios de Posgrado en Estadística para optar al grado y título de Maestría Profesional en Estadística

ANÍBAL BRENES JIMÉNEZ

Ciudad Universitaria Rodrigo Facio, Costa Rica
2020

"Este trabajo final de investigación aplicada fue aceptado por la Comisión del Programa de Estudios de Posgrado en Estadística de la Universidad de Costa Rica, como requisito parcial para optar al grado y título de Maestría Profesional en Estadística."

PhD Marcela Álfaro Córdoba

**Profesora Guía**

PhD Guaner Rojas Rojas

**Lector**

MSc Juan José Leitón Montero

**Lector**

Aníbal Brenes Jiménez

**Sustentante**

# Tabla de contenidos

# Lista de cuadros

# Lista de figuras

## RESUMEN

Se realizó una comparación de la capacidad predictiva de los modelos de Árboles de decisión, Bosques aleatorios y XGBoosting al modelar el caudal promedio horario en la estación hidrológica Palmar, ubicada sobre el río Grande de Térraba en la región Pacífico-Sur de Costa Rica.

Para el ajuste del modelo, se trabajó con datos a nivel horario de cuatro estaciones hidrológicas y treinta y seis estaciones meteorológicas, todas propiedad del Instituto Costarricense de Electricidad.

Se comparó la capacidad predictiva de los modelos en dos escenarios: en el primero, se imputaron los datos faltantes de precipitación mediante la ponderación por el cuadrado del inverso de la distancia y los de caudal por el método de área-lluvia. En el segundo, se utilizaron los datos sin ningún tipo de imputación. Como medidas para evaluar el desempeño de los modelos en ambos escenarios, se utilizaron el Coeficiente de Eficiencia de Nash-Sutcliffe, la raíz cuadrada de error cuadrático medio, el error medio absoluto, el error cuadrático medio y el porcentaje de error absoluto medio.

El resultado de la investigación es que para este caso particular, el modelo de Bosques aleatorios es el que tiene un mejor desempeño en ambos escenarios. Además, la imputación de datos faltantes en las variables predictoras generó mejoras en el desempeño de todos los modelos.

UNIVERSIDAD DE COSTA RICA

SEP Sistema de Estudios de Posgrado

**Autorización para digitalización y comunicación pública de Trabajos Finales de Graduación del Sistema de Estudios de Posgrado en el Repositorio Institucional de la Universidad de Costa Rica.**

Yo, Aníbal Brenes Jiménez, con cédula de identidad 114790959, en mi condición de autor del TFG titulado Predicción del caudal promedio horario de la estación hidrológica Palmar, utilizando modelos de Machine Learning basados en Árboles de decisión

Autorizo a la Universidad de Costa Rica para digitalizar y hacer divulgación pública de forma gratuita de dicho TFG a través del Repositorio Institucional u otro medio electrónico, para ser puesto a disposición del público según lo que establezca el Sistema de Estudios de Posgrado. **SI** [X] **NO** *[ ]

***En caso de la negativa favor indicar el tiempo de restricción: _____ año (s).**

Este Trabajo Final de Graduación será publicado en formato PDF, o en el formato que en el momento se establezca, de tal forma que el acceso al mismo sea libre, con el fin de permitir la consulta e impresión, pero no su modificación.

Manifiesto que mi Trabajo Final de Graduación fue debidamente subido al sistema digital Kerwá y su contenido corresponde al documento original que sirvió para la obtención de mi título, y que su información no infringe ni violenta ningún derecho a terceros. El TFG además cuenta con el visto bueno de mi Director (a) de Tesis o Tutor (a) y cumplió con lo establecido en la revisión del Formato por parte del Sistema de Estudios de Posgrado.

**INFORMACIÓN DEL ESTUDIANTE:**

Nombre Completo: Aníbal Brenes Jiménez.

Número de Carné: A91044 Número de cédula: 11470959.

Correo Electrónico: anibalbrenesj@gmail.com.

Fecha: 18/11/2020. Número de teléfono: 88363642.

Nombre del Director (a) de Tesis o Tutor (a): Marcela Alfaro Córdoba.

**FIRMA ESTUDIANTE**

Nota: El presente documento constituye una declaración jurada, cuyos alcances aseguran a la Universidad, que su contenido sea tomado como cierto. Su importancia radica en que permite abreviar procedimientos administrativos, y al mismo tiempo genera una responsabilidad legal para que quien declare contrario a la verdad de lo que manifiesta, puede como consecuencia, enfrentar un proceso penal por delito de perjurio, tipificado en el artículo 318 de nuestro Código Penal. Lo anterior implica que el estudiante se vea forzado a realizar su mayor esfuerzo para que no sólo incluya información veraz en la Licencia de Publicación, sino que también realice diligentemente la gestión de subir el documento correcto en la plataforma digital Kerwá.

# 1 Introducción

Esta investigación evalúa la capacidad predictiva de modelos de aprendizaje de máquina o *Machine Learning* (de ahora en adelante se escribirá como *ML*) al modelar datos de caudal promedio horario en la estación hidrológica Palmar en la región Pacífico-Sur de nuestro país.

Los registros de caudal presentan con frecuencia grandes cantidades de datos faltantes, o bien, no son lo suficientemente extensos, por lo que se requiere contar con modelos que permitan completar estos vacíos de información. Actualmente, se han desarrollado investigaciones en el uso modelos de *ML* en cuencas en países como Estados Unidos, Canadá, Turquía, entre otros. Sin embargo, las características climáticas de estas regiones son diferentes a las de Costa Rica, por lo que es necesario verificar que estos modelos tengan desempeños aceptables para nuestras condiciones.

El uso de modelos de *ML* en datos de caudal promedio horario ha sido poco investigado, a diferencia de caudales a escalas diaria o mensual. Al ser la variabilidad del caudal horario mucho mayor que en las otras escalas temporales, es importante estudiar si los modelos de *ML* pueden reproducir esta variabilidad de manera satisfactoria.

Esta investigación arroja que que los modelos de *ML* basados en Árboles de decisión pueden modelar de manera satisfactoria el caudal horario en la estación hidrológica Palmar, en la cuenca del río Grande de Térraba. Esto abre las puertas al uso de dichos modelos para completar y extender registros hidrológicos en esta estación y para su aplicación en otras regiones de nuestro país.

# 2 Justificación

Beauchamp et al. (1989) mencionan que la síntesis de datos de caudal para periodos largos de datos faltantes es un problema común en hidrología. Los usos de datos de caudal varían desde la planificación, el diseño y la operación de sistemas complejos de recursos hídricos (Gyau-Boakye and Schultz (1994)) hasta su uso como información inicial en modelos ambientales (Kim and Pachepsky (2010)). Por lo general, para estos diferentes usos se requiere que las series no presenten vacíos de información, o bien, extender el registro disponible hasta una longitud adecuada para el respectivo análisis.

Dentro de las razones que explican la ausencia de datos, Mwale et al. (2012) mencionan la falta temporal del personal observador, mal funcionamiento del equipo y falta de recursos económicos. Adicional a estas, en nuestro país es común que los equipos fallen ante una creciente extrema, que sean removidos para ser ubicados en otros sitios o incluso por vandalismo.

Al ser este un problema recurrente en hidrología, a lo largo del tiempo se han propuesto una gran cantidad de métodos para darle solución. Los métodos varían desde utilizar promedios ponderados como los referenciados por Gyau-Boakye and Schultz (1994) hasta modelos más modernos y complejos como lo son los de *ML*, utilizados recientemente por Shortridge et al. (2016) o Tongal and Booij (2018).

El uso de modelos de *ML* en hidrología inició en la década de los noventas con la investigación de Karunanithi et al. (1994), en la cual se utilizan las redes neuronales para modelar el caudal diario del río Huron en Michigan, Estados Unidos. A partir de este estudio se han desarrollado una gran cantidad de investigaciones en las cuales se utilizan modelos de *support vector regression* para la estimación de caudal en ríos.

De acuerdo con Ardabili et al. (2020) el uso de modelos físicos y estadísticos, como por ejemplo las series de tiempo, tiene una larga tradición de aplicación en fenómenos hidrológicos. Los autores mencionan que existen una serie de inconvenientes en el uso de estos modelos como, por ejemplo: precisión, debilidad en análisis de incertidumbre, alto costo computacional, entre otros. Ante estas debilidades, el uso de modelos de *ML* se ha incrementado en los últimos años.

El Instituto Costarricense de Electricidad (ICE) es el principal desarrollador de proyectos hidroeléctricos en Costa Rica y posee la red hidrometeorológica más grande del país, por lo que se torna de vital importancia que pueda contar con modelos estadísticos para completar o extender los registros disponibles que minimicen los errores de estimación, asegurando la calidad de la información utilizada para los diferentes fines de la institución.

Actualmente, el Área de Hidrología del ICE únicamente completa los registros faltantes a nivel diario. Para esto, utiliza principalmente modelos de regresión lineal simple, en los cuales se utiliza como variable independiente el caudal en alguna otra estación aguas arriba o aguas abajo de la estación con el registro faltante. Con la presente investigación se pretende, en primer lugar, verificar si es posible completar los datos a una escala menor (nivel horario). En segundo lugar, utilizar una amplia gama de modelos para identificar cuál de esos genera una mejor estimación de

los datos faltantes y, finalmente, ajustar modelos que incluyan una mayor cantidad de información hidrometeorológica con el fin de disminuir los errores de estimación.

La estación hidrológica Palmar se encuentra ubicada en una de las cuencas con mayor instrumentación en el país, lo cual es beneficioso ya que existe una gran cantidad de información hidrometeorológica para ser utilizada como covariables de los modelos. Además, la información hidrológica disponible en esta cuenca tiene la ventaja de que no ha sido afectada por el desarrollo de proyectos hidroeléctricos. Esto es de suma importancia ya que la existencia de un proyecto hidroeléctrico altera el comportamiento natural de la escorrentía, lo cual podría afectar el ajuste de los modelos. Otro aspecto de importancia para trabajar con datos de esta cuenca es que en ella se tiene planeada la construcción del Proyecto Hidroeléctrico Diquís, el cual es de gran importancia para la institución.

La presente investigación tiene como fin estimar el caudal del río Grande de Térraba a nivel horario (a la altura de la estación hidrológica Palmar), mediante el uso de tres modelos de *ML*, a saber, Árboles de decisión, Bosques aleatorios y Potenciación, con el propósito de comparar su capacidad predictiva. Además, se desea comprobar si estos modelos generan resultados aceptables en una cuenca con las características climáticas de la región Pacífico-Sur de Costa Rica.

# 3 Objetivos

## 3.1 Objetivo general

Comparar la capacidad predictiva de diferentes modelos estadísticos para los datos de caudal promedio horario de la estación hidrológica Palmar, en el río Grande de Térraba.

## 3.2 Objetivos específicos

- Realizar un análisis exploratorio de los datos de caudal y precipitación registrados en la cuenca del río Grande de Térraba.
- Completar los registros hidrológicos y meteorológicos de la cuenca previo al modelado.
- Modelar el caudal en la estación Palmar utilizando como covariables la precipitación horaria de la cuenca y el caudal horario registrado en otras estaciones aguas arriba.
- Comparar la capacidad predictiva de los diferentes modelos.
- Comparar la capacidad predictiva al utilizar registros completos y registros con datos faltantes.

# 4 Marco Teórico

A continuación se presentará una descripción de los modelos de *ML* que fueron utilizados para la estimación del caudal en la estación hidrológica Palmar. En primer lugar, se describirá el modelo de Árboles de decisión, el cual es la base para los dos modelos restantes. En segundo lugar, se explicará el modelo de Bosques aleatorios, y finalmente, el modelo de Potenciación.

## 4.1 Árboles de decisión

Los árboles de decisión pueden ser utilizados tanto para problemas de clasificación (variable respuesta categórica) como para problemas de regresión (variable respuesta continua). Para esta investigación únicamente los segundos son de interés. A grandes rasgos, este modelo consiste en estratificar o segmentar el espacio de las variables independientes (predictoras) en un número de regiones simples. Para esto, se utiliza una serie de reglas binarias que permiten generar la partición del espacio y las cuales pueden ser representadas en forma de árbol.

Estos modelos tienen la ventaja de que son simples e interpretables, lo cual puede llegar a ser muy útil dependiendo de los objetivos que se quieran cumplir con el modelo. Por otra parte, tienen la desventaja de que por lo general no tienen tan buen desempeño al compararles con otros modelos de *ML*. Sin embargo, estos modelos son la base de otros más complejos, como por ejemplo el de bosques aleatorios.

De acuerdo con James et al. (2013), para construir un modelo de árboles de decisión se requiere solamente de dos pasos:

1. Dividir el espacio de las variables predictoras $(X_1, X_2, ..., X_P)$ en J regiones sin traslape $(R_1, R_2, ..., R_J)$.

2. A cada observación que esté en una región $R_j$, se le asigna la misma predicción, la cual es el promedio de los valores de las observaciones del conjunto de datos de entrenamiento que se ubicaron en esta región a la hora de realizar el modelo.

Para ejecutar el primer paso, se elige dividir el espacio predictor en una serie de rectángulos multidimensionales, esto por simplicidad y facilidad de interpretación del modelo resultante. El objetivo es encontrar los rectángulos $R_1, R_2, ..., R_J$ que minimicen la suma del error cuadrático *RSS* dado por:

$$\sum_{j=1}^{J} \sum_{i \in R_j} (y_i - \hat{y}_{R_j})^2$$

Donde $\hat{y}_{R_j}$ es el valor promedio de la repuesta de las observaciones de entrenamiento dentro del rectángulo $j$. En términos de capacidad computacional, no es factible considerar todas las posibles particiones del espacio predictor en $J$. Ante esta situación se utiliza un método de partición recursivo binario.

El método comienza con todas las observaciones contenidas dentro de un primer rectángulo. Este espacio se divide en dos partes procurando obtener la mejor partición posible

independientemente de las posibles particiones futuras. En caso de ser necesario, las nuevas regiones conformadas se vuelven a dividir de forma binaria y se continua así sucesivamente hasta obtener un resultado satisfactorio.

Para realizar la partición de cada región, primeramente se selecciona un predictor $X_j$ y un punto de corte $s$ tal que el espacio predictor sea divido en las regiones $\{X|X_j < s\}$ y $\{X|X_j \geq s\}$ de forma que se minimice el *RSS*. Es decir, se consideran todos los posibles predictores $X_1, X_2, ..., X_p$ y todos los posibles puntos de corte $s$ para cada predictor y se escoge el predictor cuyo árbol resultante tenga el mejor *RSS*.

Seguidamente, el proceso se repite buscando un nuevo mejor predictor y un nuevo punto de corte, que dividan una de las dos regiones creadas anteriormente. Este proceso se repite hasta que se alcance algún criterio de tolerancia.

Una vez creadas las regiones $R_1, R_2, ..., R_J$, la respuesta de los individuos de un set de prueba o de un nuevo set de individuos se calcula como el promedio de las observaciones de entrenamiento contenidas en la región a la cual pertenece el individuo en análisis.

El proceso descrito anteriormente, el cual fue propuesto por James et al. (2013), puede producir buenas predicciones en la base de datos de entrenamiento. Sin embargo, el método es propenso a sobreajustar estos datos, lo que lleva a un desempeño deficiente en los datos de prueba. Esto se debe a que el árbol resultante podría ser muy complejo. Un árbol más pequeño con menos particiones (menor cantidad de regiones $R_1, ..., R_J$) podría disminuir la varianza y mejorar la interpretación del modelo a cambio de un pequeño sesgo.

Una estrategia para evitar el sobreajuste del modelo, según James et al. (2013), consiste en construir un árbol $T_0$ lo suficientemente largo y posteriormente "podarlo" para obtener un subárbol de menor tamaño. El objetivo es seleccionar un subárbol que genere la menor tasa de error. Para un determinado subárbol, es posible estimar esta tasa mediante validación cruzada. Sin embargo, realizar esta labor para todos los posibles subárboles no es factible a nivel computacional. En vez de esto, se necesita una forma de seleccionar una cantidad pequeña de subárboles para analizar.

La "poda por el nodo más débil" es una forma de solucionar el problema mencionado anteriormente. En vez de considerar todo subárbol posible, se considera una secuencia de árboles indexados por un parámetro no negativo $\alpha$. A cada valor de $\alpha$ le corresponde un subárbol $T \subset T_0$ tal que:

$$\sum_{m=1}^{|T|} \sum_{i:x_i \in R_m} (y_i - \hat{y}_{R_m})^2 + \alpha|T|$$

es tan pequeño como sea posible. $|T|$ indica el número de nodos terminales del árbol $T$, $R_m$ es el rectángulo correspondiente al nodo terminal $m$ y $\hat{y}_{R_m}$ es la media de las observaciones de entrenamiento en $R_m$. El parámetro de ajuste $\alpha$ controla la complejidad del árbol y su ajuste a los datos de entrenamiento. Cuando $\alpha = 0$, el subárbol $T$ será simplemente igual a $T_0$. Conforme $\alpha$ incremente su valor, tener un árbol con muchos nodos será penalizado ya que la cantidad de nodos ($|T|$) se multiplica por este parámetro. Con lo cual la cantidad de nodos tenderá a ser minimizada.

## 4.2 Bosques aleatorios

Para adentrarse en el modelo de bosques aleatorios, es necesario iniciar detallando el concepto de *Bagging*.

Según James et al. (2013), los árboles de decisión mencionados anteriormente sufren de una alta variabilidad. Esto se debe a que al dividir los datos de entrenamiento en dos partes de forma aleatoria y ajustar un árbol a cada una, el resultado que se obtendría podría ser distinto. Por otra parte, un procedimiento con baja variabilidad podría generar resultados similares si se aplica repetidamente a distintos datos. Por ejemplo, la regresión lineal tiende a tener una baja varianza si la razón entre el número de observaciones y variables es moderadamente grande. La técnica de *bagging* tiene como propósito reducir la varianza de métodos de aprendizaje. Es particularmente utilizada en el contexto de bosques aleatorios.

Recordemos que para un grupo de $n$ observaciones independientes $Z_1, ..., Z_n$, cada una con varianza $\sigma^2$, la varianza de la media $\overline{Z}$ de las observaciones está dada por $\sigma^2/n$. Es decir, promediar un grupo de observaciones reduce la varianza. James et al. (2013) parten de esta idea y proponen que una forma de reducir la varianza y por ende incrementar la precisión de un modelo es tomar muchos sets de entrenamiento de la población, construir un modelo para cada uno de estos y promediar sus predicciones. Esto es calcular $\hat{f}^1(x), \hat{f}^2(x), ..., \hat{f}^B(x)$, donde cada $f$ es un árbol de decisión ajustado para $B$ set de entrenamiento, y promediarlos para obtener un único modelo de varianza baja dado por:

$$\hat{f}_{avg}(x) = \frac{1}{B}\sum_{b=1}^{B}\hat{f}^b(x)$$

Este procedimiento tiene la limitación de que usualmente no se tiene acceso a múltiples sets de entrenamiento. En su lugar, es posible re-muestrear diferentes sets a partir de un único set de entrenamiento (*bootstrap*). En este enfoque, se generan $B$ diferentes sets de entrenamiento remuestreados a partir del conjunto de datos original. Posteriormente, se entrena el modelo en cada $b$-ésimo set de entrenamiento para calcular $\hat{f}^{*b}(x)$ y finalmente promediar todas las predicciones para obtener:

$$\hat{f}_{bag}(x) = \frac{1}{B}\sum_{b=1}^{B}\hat{f}^{*b}(x)$$

Esto es conocido como *bagging*.

De acuerdo con James et al. (2013) los bosques aleatorios proporcionan una mejora al método de *bagging* al utilizar árboles no correlacionados. Esta no correlación se logra al utilizar una muestra de tamaño $m$ del total de variables $p$ en cada división del árbol. Un valor típico es $m \approx \sqrt{p}$. En otras palabras, al construir un bosque aleatorio, en cada división del árbol, el algoritmo no puede considerar a la mayoría de los predictores disponibles. Esto tiene la ventaja de que en caso de que exista un predictor mucho más fuerte que los demás, éste probablemente sería seleccionado como primera división en la mayoría de los árboles, ocasionando que éstos se parezcan mucho entre sí.

Lo anterior genera que la reducción en la varianza no sea grande, ya que promediar cantidades altamente correlacionadas no genera una diferencia significativa a un árbol único.

Reducir la cantidad de posibles predictores a la hora de construir un bosque aleatorio, da una mayor oportunidad a variables que no tengan un poder predictivo tan fuerte y, por ende, se tendrá una mayor variabilidad entre los árboles generados. Este procedimiento es particularmente útil cuando se tiene una cantidad alta de predictores correlacionados.

## 4.3 Potenciación

Al igual que *bagging*, potenciación es un método general que puede ser aplicado a distintos modelos de regresión o clasificación. El contexto en el que se detallará este método es el de los árboles de decisión.

James et al. (2013) proponen que la técnica de potenciación funciona de forma similar a *bagging*, con la excepción de que los árboles se construyen de forma secuencial. Es decir, cada árbol es ajustado con información de árboles previamente construidos. A diferencia del *bagging*, cada árbol es ajustado en una versión modificada del set de datos original y no en un set obtenido mediante remuestreo (*bootstrap*).

Al igual que *bagging*, potenciación involucra combinar un número de árboles de decisión $\hat{f}^1(x), ..., \hat{f}^B(x)$. Sin embargo, en vez de ajustar un árbol de decisión grande para el set de datos, se construye uno utilizando como respuesta los residuos del modelo en vez de la variable $Y$. Cada uno de estos árboles puede ser relativamente pequeño con pocos nodos terminales. Al ajustar árboles pequeños a los residuos, se mejora $\hat{f}$ en áreas donde este tenía un desempeño deficiente. Nótese que a diferencia de *bagging*, en potenciación la construcción de cada árbol depende fuertemente de los árboles que fueron previamente ajustados.

El algoritmo para realizar el método de potenciación, propuesto por James et al. (2013), está dado por:

1. Fije $\hat{f}(x) = 0$ y $r_i = y_i$ para todo $i$ en el set de entrenamiento
2. Para $b = 1, 2, ..., B$, repita:
   (a) Ajuste el árbol $\hat{f}^b$ con $d$ particiones al set de entrenamiento $(X, r)$.
   (b) Actualice $\hat{f}$ agregando una nueva versión reducida del árbol:
   $$\hat{f}(x) = \hat{f}(x) + \lambda \hat{f}^b(x)$$
   (c) Actualice los residuos:
   $$r_i = r_i - \lambda \hat{f}^b(x_i)$$
3. Obtenga el modelo potenciado:
   $$\hat{f}(x) = \sum_{b=1}^{B} \lambda \hat{f}^b(x)$$

Los modelos descritos serán utilizados en la sección 6.3 para realizar el modelado de la serie de caudal de la estación hidrológica Palmar. A continuación, se presentará una descripción geográfica de la zona de estudio.

## 4.4 Revisión bibliográfica

Actualmente existen investigaciones enfocadas en la comparación de modelos de *ML*. Sin embargo, la mayoría de estas fueron realizadas a partir de datos diarios. Un ejemplo de este tipo de investigación es la realizada por Karakurt et al. (2013), en la cual se compara el desempeño de ensambles de árboles de decisión y de redes neuronales para la predicción del caudal diario en el río Seyhan en Turquía. Este estudio tuvo como resultado que tanto los ensambles de árboles de decisión como los de redes neuronales tienen una mejor capacidad predictiva que un modelo de redes neuronales simple. Además, se encontró que el ensamble de redes neuronales genera mejores resultados que el de árboles de decisión.

Otro ejemplo es el estudio realizado por Rasouli et al. (2012). En él se compara el desempeño de las redes neuronales bayesianas, *support vector regression* y procesos gaussianos al estimar el caudal diario en el río Stave en British Columbia, Canadá. Los modelos de *ML* utilizados se compararon con un modelo de regresión lineal múltiple; se obtuvo que los tres generan mejores resultados que la regresión en términos de la raíz cuadrada del error cuadrático medio *RMSE*, el error medio absoluto *MAE* y el Coeficiente de Nash-Sutcliffe *NSE*. Además, las redes neuronales bayesianas produjeron mejores resultados que los dos modelos restantes.

Tongal and Booij (2018) realizaron una investigación en la cual simulan y pronostican caudales diarios en cuatro diferentes cuencas hidrográficas de Estados Unidos mediante regresores de soporte vectorial, redes neuronales y bosques aleatorios. Dicha investigación es de gran relevancia para este trabajo ya que presenta una amplia descripción del proceso metodológico utilizado para realizar la comparación de las capacidades predictivas de los modelos. Los autores encontraron que el modelo de árboles de decisión fue el que generó peores resultados para el periodo de calibración. Además, las redes neuronales fueron el mejor modelo en tres de las cuatro cuencas que fueron estudiadas. Otro hallazgo fue que, para el periodo de validación, el peor modelo fueron las redes neuronales.

Con respecto a estudios realizados con caudales horarios, se encontró que, actualmente existen pocos intentos por modelar este tipo de información mediante modelos de *ML*. Uno de los artículos existentes es el elaborado por Asefa et al. (2006), en el cual utilizan *Support Vector Machines* para la predicción de caudales mensuales y horarios. Los autores concluyen que estos modelos generaron resultados satisfactorios en ambos escenarios. Otra investigación de relevancia es la presentada por Wu and Lin (2015) en la que pronostican los caudales horarios mediante distintos tipos de redes neuronales, *Support Vector Machines* y *self-organinzing maps*. Al realizar la comparación, los investigadores hallaron que los modelos de redes neuronales tienen un mejor desempeño que los otros dos.

Al realizar esta revisión bibliográfica, se encontró que los modelos de *ML* basados en árboles de decisión han sido previamente utilizados en el área de la hidrología. Otro hallazgo de interés es

que no existe un claro modelo de *ML* que sea mejor que los demás en términos de estadísticos como el NSE, RMSE o MAE, de ahí que sea pertinente ajustar distintos modelos para cada problema específico con el fin de encontrar el mejor modelo posible. Es necesario resaltar que la mayoría de estos resultados obtenidos a partir de estudios de *ML* fueron obtenidos en datos de caudal diario, mientras que para datos de caudal horario existen pocas investigaciones. Además, la gran mayoría de análisis existentes fueron realizados en cuencas de países con características climatológicas muy distintas a las que presenta Costa Rica (Estados Unidos, Canadá, Turquía, entre otros). Lo anterior podría implicar que los resultados obtenidos en estas investigaciones no son extrapolables a las condiciones de nuestro país.

## 4.5 Descripción de la zona de estudio

Previo a iniciar con la exploración y análisis de datos, es necesario realizar una breve descripción del área de estudio para lograr una mejor comprensión de la relación existente entre las variables meteorológicas e hidrológicas con la variable respuesta.

La Figura 1 muestra la ubicación general de la cuenca del río Grande de Térraba a la altura de la estación hidrológica Palmar. Como se observa, está ubicada en el Pacífico Sur de nuestro país y se encuentra mayoritariamente en la provincia de Puntarenas, aunque la parte este de la cuenca se localiza en la provincia de San José.

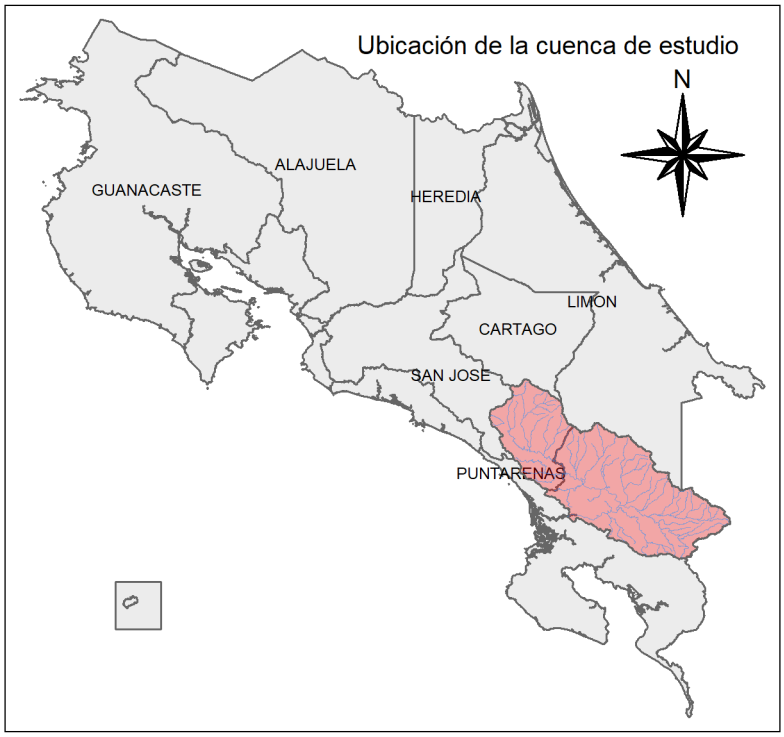


Figura 1: Ubicación general del área de estudio

Fuente: Elaboración propia

En la Figura 2 se muestra con mayor detalle el área de estudio. Se pueden observar las cuencas

de los ríos General, Coto Brus y Cabagra a las alturas de las estaciones El Brujo, Caracucho y Cabagra respectivamente. El Cuadro 1 muestra el área de drenaje de cada una de ellas en kilómetros cuadrados. El área de la cuenca Palmar está compuesta por la suma de las áreas de las otras tres cuencas más el área intermedia (área blanca en la Figura 2). De manera similar, el caudal observado en la estación Palmar está compuesto por el registrado en las otras tres estaciones más el que se produce en el área intermedia y en otros afluentes al río Grande de Térraba que no están contenidos en las cuencas de las otras estaciones.

Cuadro 1: Área de drenaje de cuencas hidrográficas

| | **Nombre** | **Área (km2)** |
|---|---|---|
| 1 | Cabagra | 371.7 |
| 2 | Caracucho | 1133.9 |
| 3 | El Brujo | 2400.2 |
| 4 | Palmar | 4765.2 |

Fuente: Elaboración propia con datos del ICE

La Figura 2 también muestra la ubicación de todas las estaciones meteorológicas. Se observa que la mayoría están ubicadas en la cuenca del río General. Sin embargo, se cuenta con mediciones en todas las cuencas, lo cual es de gran importancia para hacer estimaciones como la precipitación media sobre el área de drenaje.

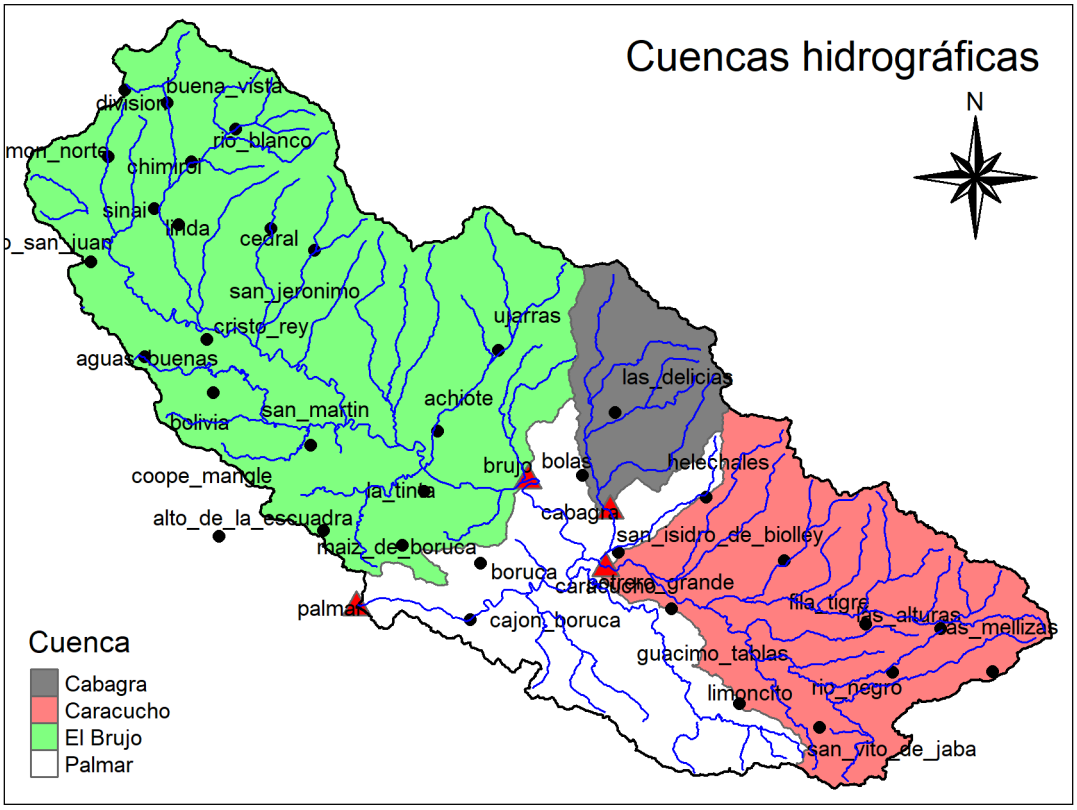


Figura 2: Cuencas en el área de estudio

Fuente: Elaboración propia con datos del ICE

# 5 Metodología

## 5.1 Revisión bibliográfica

Como paso inicial para el desarrollo de esta investigación se realizó una revisión bibliográfica con el fin de identificar estudios previos que abordaran el uso de modelos de *ML* basados en Árboles de decisión en datos de caudal, ya fuera a nivel horario o diario. Además, se investigó acerca de estudios de este tipo realizados en Costa Rica. Posterior a la revisión, se procedió a plantear los objetivos de esta investigación.

## 5.2 Recopilación de información

La información hidrometeorológica utilizada en esta investigación fue proporcionada por el Área de Hidrología del Centro de Servicios Estudios Básicos de Ingeniería del Instituto Costarricense de Electricidad. Además, se utilizaron algunas capas del Atlas Digital de Costa Rica del Tecnológico de Costa Rica en su versión 2014 para la elaboración de mapas.

## 5.3 Análisis exploratorio

Se realizó un análisis exploratorio tanto de la información hidrológica como meteorológica con dos fines. Primero, describir la calidad del registro en términos de cantidad de datos faltantes; segundo, estudiar la correlación entre las variables predictoras y la variable respuesta. Tanto el análisis exploratorio como los demás análisis estadísticos posteriores, fueron elaborados en el programa *R* en su versión 4.0.0.

## 5.4 Imputación de datos faltantes

En el caso de los datos meteorológicos, la imputación se realizó por medio de una ponderación por el cuadrado del inverso de la distancia con las tres estaciones más cercanas (distancia euclídea).

Los datos hidrológicos fueron completados por medio del método área - lluvia. Para esto se calculó la precipitación promedio sobre las cuencas utilizando la ponderación por polígonos de Voronoi.

## 5.5 Modelado

En la sección de modelado se trabajó en la selección de las variables según su importancia, el ajuste de los parámetros de los modelos y la comparación del desempeño del modelo.

Para los primeros dos pasos se trabajó con un conjunto de datos de entrenamiento del 77% del total. El 23% restante se utilizó para conjunto de prueba para evaluar la capacidad predictiva de los modelos ajustados.

## 5.6 Importancia de variables

Se utilizó un modelo de Bosques aleatorios para cuantificar la importancia de cada variable independiente. Para esto se utilizaron los criterios de ganancia de información basada en entropía, el estadístico de independencia Chi Cuadrado y, por último, la ganancia de información basada en permutaciones de bosques aleatorios. Una vez establecida se importancia de cada variable se seleccionaron las cuarenta variables más importantes de acuerdo con cada indicador.

## 5.7 Ajuste de parámetros

Se realizó un ajuste de los parámetros de cada modelo utilizado con el fin de lograr el mejor ajuste posible sobre el conjunto de entrenamiento y con esto una adecuada comparación del desempeño. Como indicador de la calidad del ajuste se utilizó el coeficiente de eficiencia de Nash-Sutcliffe, el cual es ampliamente utilizado a la hora de medir el desempeño de modelos hidrológicos.

## 5.8 Comparación de desempeño

Una vez obtenidos los parámetros óptimos para cada modelo, se procedió a ajustar los diferentes modelos con estos valores y predecir el caudal promedio del conjunto de prueba. El resultado de la predicción de cada modelo se evaluó mediante el coeficiente de eficiencia de Nash-Sutcliffe y medidas de error como el RMSE, MAE, MSE y MAPE.

## 5.9 Modelos con datos faltantes

En esta sección se repitieron los pasos de la sección de modelado, con la diferencia de que se utilizó el set de datos previo a la imputación de datos faltantes. Lo anterior con el fin de identificar el impacto que tiene el realizar la imputación en el desempeño final de los modelos utilizados.

# 6 Resultados

## 6.1 Información inicial

Para llevar a cabo esta investigación, se solicitó información hidrológica y meteorológica al ICE. Específicamente, se solicitó el caudal horario de las estaciones Caracucho, El Brujo, Cabagra y Palmar. Con respecto a la información meteorológica, se solicitó información horaria de 36 pluviómetros localizados dentro de la cuenca del río Grande de Térraba (ver Cuadro 2). El periodo de información solicitado corresponde del 01/01/2013 al 31/12/2018 lo que corresponde a 52584 observaciones.

Cuadro 2: Estaciones meteorológicas en la cuenca del río Grande de Térraba

| **Estaciones meteorológicas** | | | |
|---|---|---|---|
| Volcán B. A. | San Jerónimo | La Tinta | Alto de la Esc. |
| San Vito de Jaba | Limoncito | Potrero Grande | División |
| Cedral | Cajón Boruca | Bolas | Las Alturas |
| Bolivia | Las Mellizas | Ujarrás | Helechales |
| Río Negro | Fila Tigre | Chimirol | Linda |
| San Martín | Aguas Buenas | Cristo Rey | San Isidro de B. |
| Achiote | San Ramón N. | Alto San Juan | El Ángel |
| Río Blanco | Las Delicias | Guácimo Tablas | Coope Mangle |
| Buena Vista | Sinaí | Maíz de Boruca | Boruca |

Fuente: Elaboración propia con datos del ICE

### 6.1.1 Información hidrológica

Como paso inicial para la exploración de los datos hidrológicos, se cuantificó la cantidad de datos faltantes presente en cada estación. La Figura 3 muestra que la estación Caracucho posee más de un 70% de datos faltantes en el periodo de estudio. Mientras que la estación Palmar, la cual se utilizó como variable respuesta en los modelos, cuenta con porcentaje de información faltante cercano al 40%.

La Figura 4 permite observar la cantidad de datos faltantes a nivel mensual en cada estación, además de identificar cuáles son los periodos que tienen una mayor cantidad de información disponible. En el gráfico se aprecia que la estación Caracucho tiene un periodo común muy corto con las estaciones de Cabagra y Palmar. Exceptuando la estación Caracucho, el periodo con información más completa está comprendido entre el 2014 y 2017.

La correlación entre estaciones es otro aspecto de importancia que debe considerarse en etapas iniciales. Es de principal interés la correlación entre las variables predictoras con la variable respuesta, es decir, la correlación entre Palmar y las demás estaciones. La Figura 5 muestra

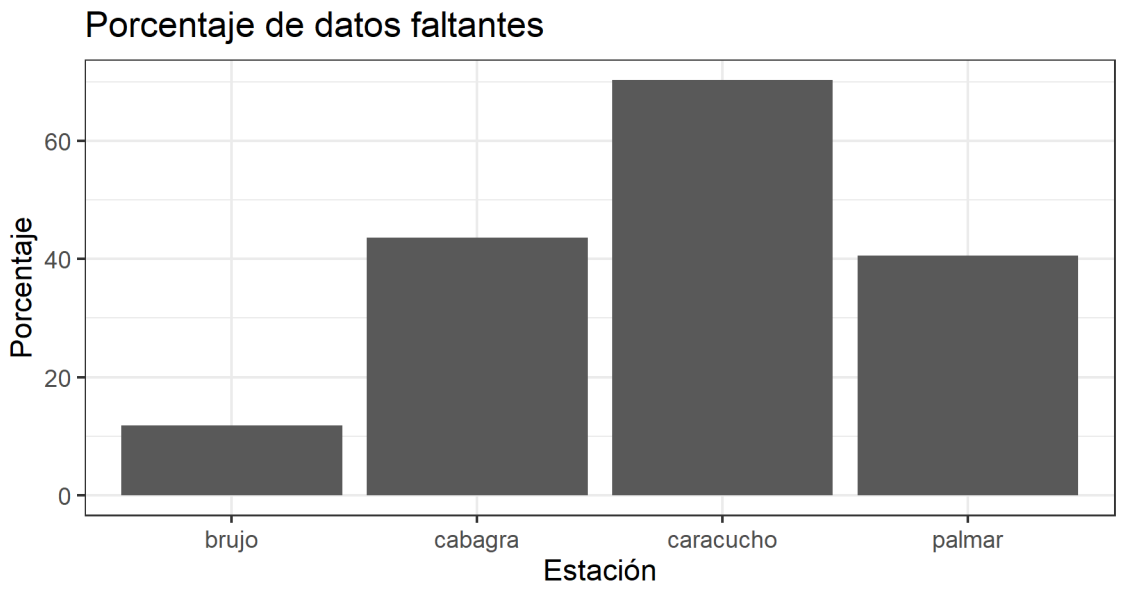


Figura 3: Porcentaje de datos faltantes por estación hidrológica

Fuente: Elaboración propia con datos del ICE

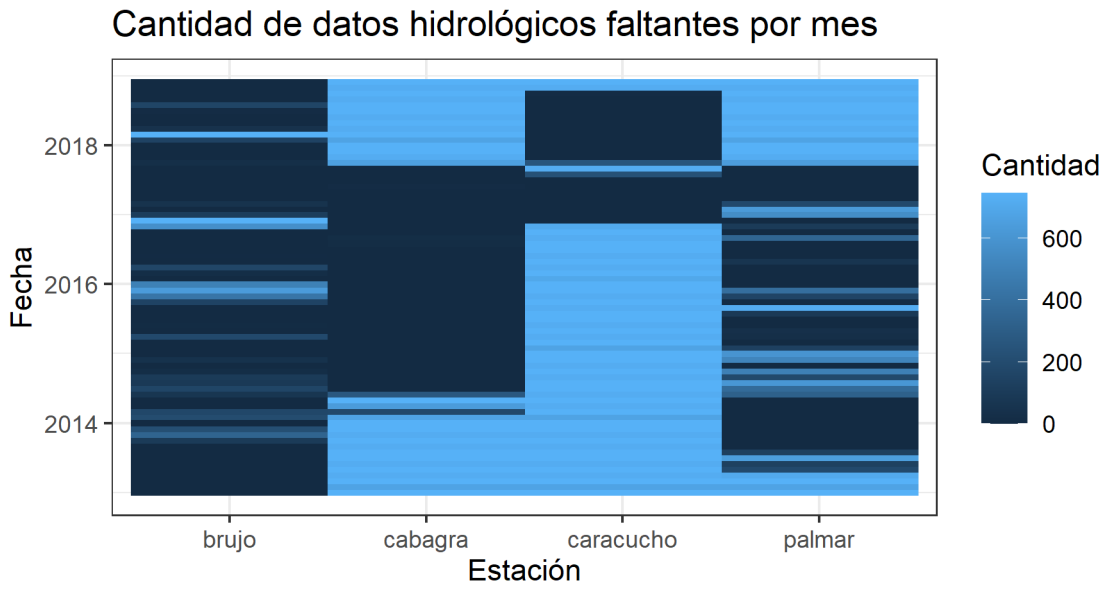


Figura 4: Cantidad de datos faltantes por mes en cada estación hidrológica

Fuente: Elaboración propia con datos del ICE

que en general existen correlaciones positivas altas entre todas las estaciones (superiores a 0.59). Específicamente, para Palmar se tienen correlaciones de 0.9, 0.7 y 0.67 con las estaciones El Brujo, Caracucho y Cabagra respectivamente.

Con el fin de observar las distribuciones de los datos de cada estación, se construyó la Figura 6, la cual contiene el diagrama de caja de cada estación. Se observa que a nivel diario, las estaciones de El Brujo y Caracucho tienen un caudal promedio similar, cercano a los 100 m3/s. Sin embargo,

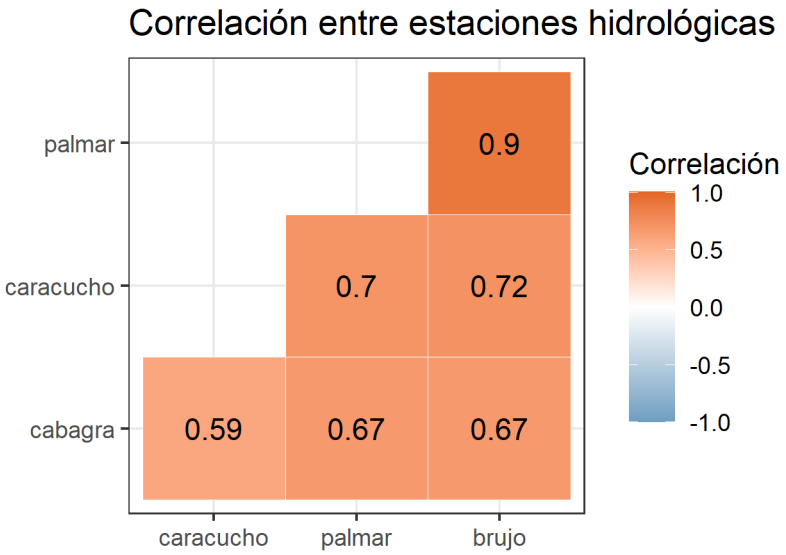


Figura 5: Correlación entre estaciones hidrológicas

Fuente: Elaboración propia con datos del ICE

se observa que en El Brujo se han registrado valores diarios más altos que en Caracucho. Como es de esperar, los valores de Cabagra y Palmar son los más bajos y altos, al ser estas las cuencas de menor y mayor área respectivamente.

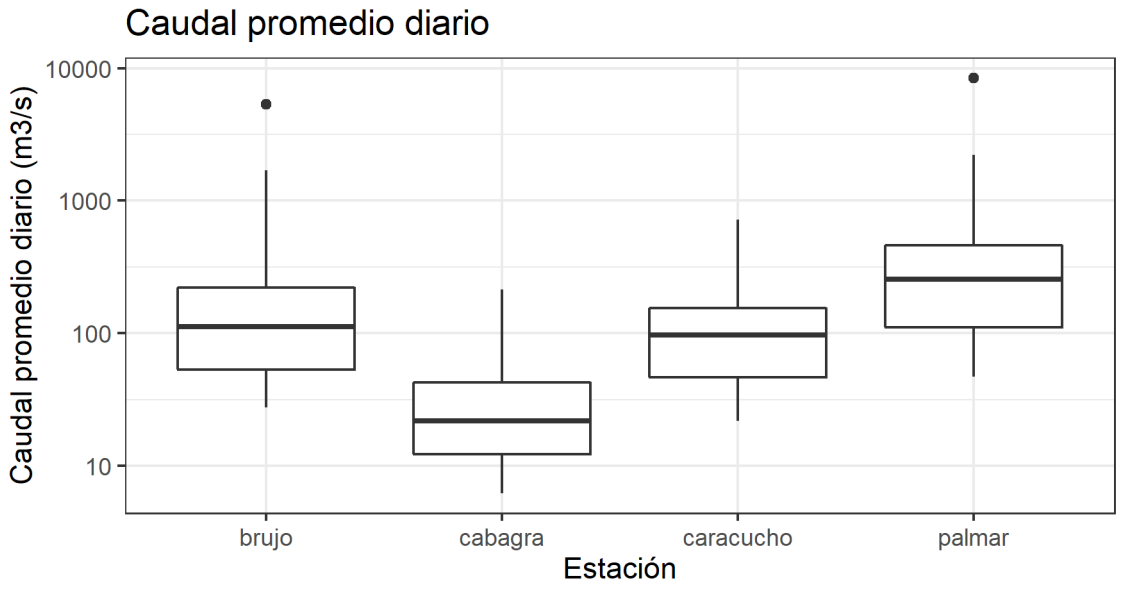


Figura 6: Gráficos de cajas para las estaciones hidrológicas

Fuente: Elaboración propia con datos del ICE

### 6.1.2 Información meteorológica

Según se detalló en el Cuadro 7, se cuenta con el registro de 36 estaciones meteorológicas dentro de la cuenca del río Grande de Térraba. La Figura 7 muestra el porcentaje de datos faltantes en cada estación para el periodo en estudio. Destacan cuatro estaciones con más de un 20% del registro incompleto, estas son (orden descendente): Volcán Buenos Aires, El Ángel, Las Alturas y Maíz de Boruca.

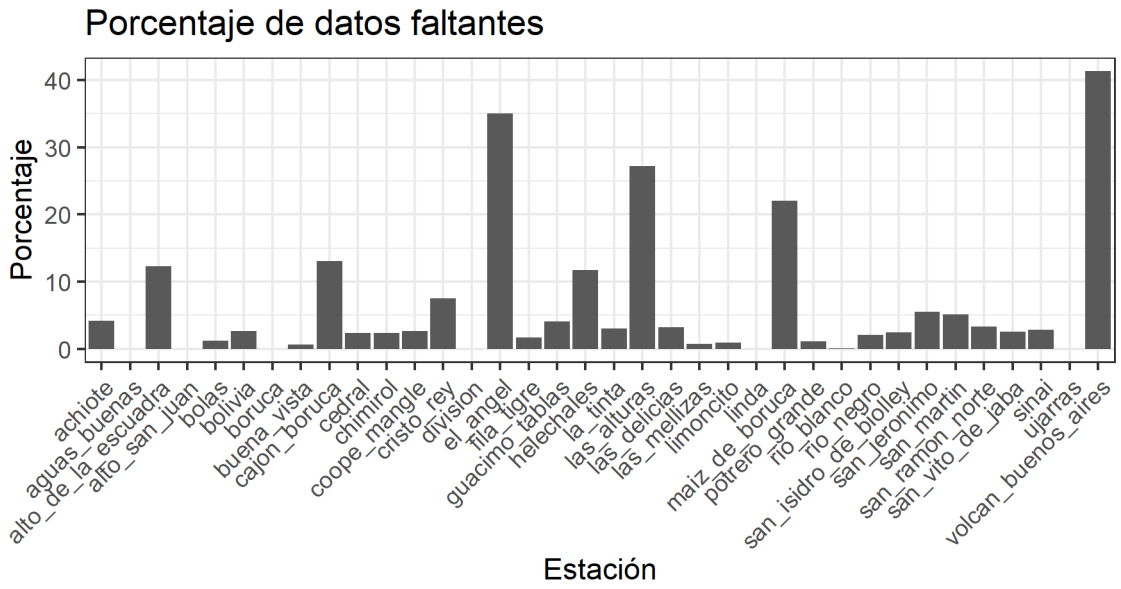


Figura 7: Porcentaje de datos faltantes por estación meteorológica

Fuente: Elaboración propia con datos del ICE

Al igual que para el caudal, se creó la Figura 8, la cual muestra la cantidad de datos faltantes a nivel mensual, en este caso para las estaciones meteorológicas. Se observa que a grandes rasgos el registro se encuentra bastante completo. El principal periodo con datos faltantes ocurre posterior al 01/01/2018, ya que al menos ocho estaciones tienen datos faltantes. Se observa también que la estaciones El Ángel y Volcán Buenos Aires estuvieron fuera de operación por periodos prolongados.

La Figura 9 muestra las correlaciones entre las series de precipitación horaria. Se observa que en general estas son menores que en los datos de caudal. Se obtuvo un rango coeficiente de correlación de 0.71 a 0.06 con un valor promedio igual a 0.20.

Es de gran interés conocer la correlación entre las variables predictoras y la variable respuesta. La Figura 10 presenta la correlación entre dichas variables. Se observa que en términos generales las correlaciones son bajas (inferiores a 0.15). El promedio de las correlaciones es de 0.09. Las estaciones con mayores correlaciones con la variable respuesta son: Aguas Buenas, Alto San Juan y Alto de la Escuadra.

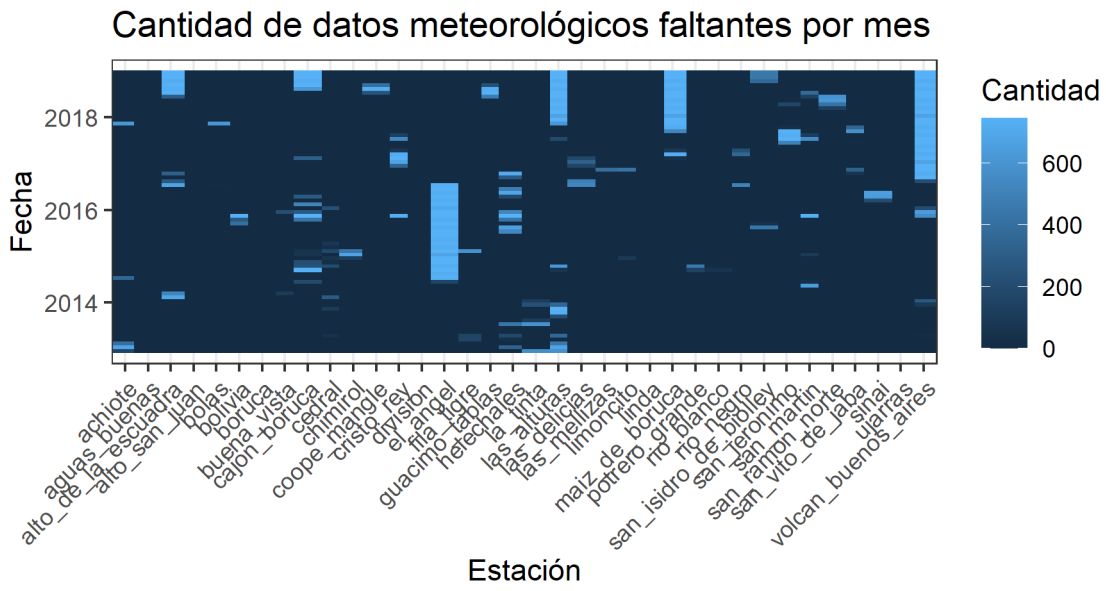


Figura 8: Cantidad de datos faltantes por mes en cada estación meteorológica

Fuente: Elaboración propia con datos del ICE

## 6.2 Imputación de datos faltantes

Como paso previo al modelado de los datos, fue necesario completar los datos faltantes tanto en los registros de caudal como de precipitación. A continuación, se presentará el resultado de la imputación de datos.

### 6.2.1 Imputación de datos meteorológicos

Se decidió iniciar con la imputación de los datos faltantes de las estaciones meteorológicas ya que estos serán un insumo para completar los registros hidrológicos. Previo a la imputación, se excluyeron del las estaciones de El Ángel y Volcán Buenos Aires, ya que como se observa en las Figuras 3 y 4 estas tienen un alto porcentaje de valores faltantes, además de que estuvieron fuera de operación por periodos prolongados.

Como método de imputación se utilizó la ponderación inversa de la distancia o *IDW*. Este método consiste en utilizar las $n$ estaciones más cercanas a estación con el dato faltante (en una hora $i$) para la estimación. La fórmula para el cálculo es:

$$p_i = \sum_{j=1}^{n} \frac{\frac{p_{j,i}}{d_j^2}}{\frac{1}{d_j^2}}$$

Donde:

- $p_{j,i}$: Precipitación en la estación $j$, en una hora $i$.
- $d_j$: Distancia entre la estación $j$ y la estación con el dato faltante.

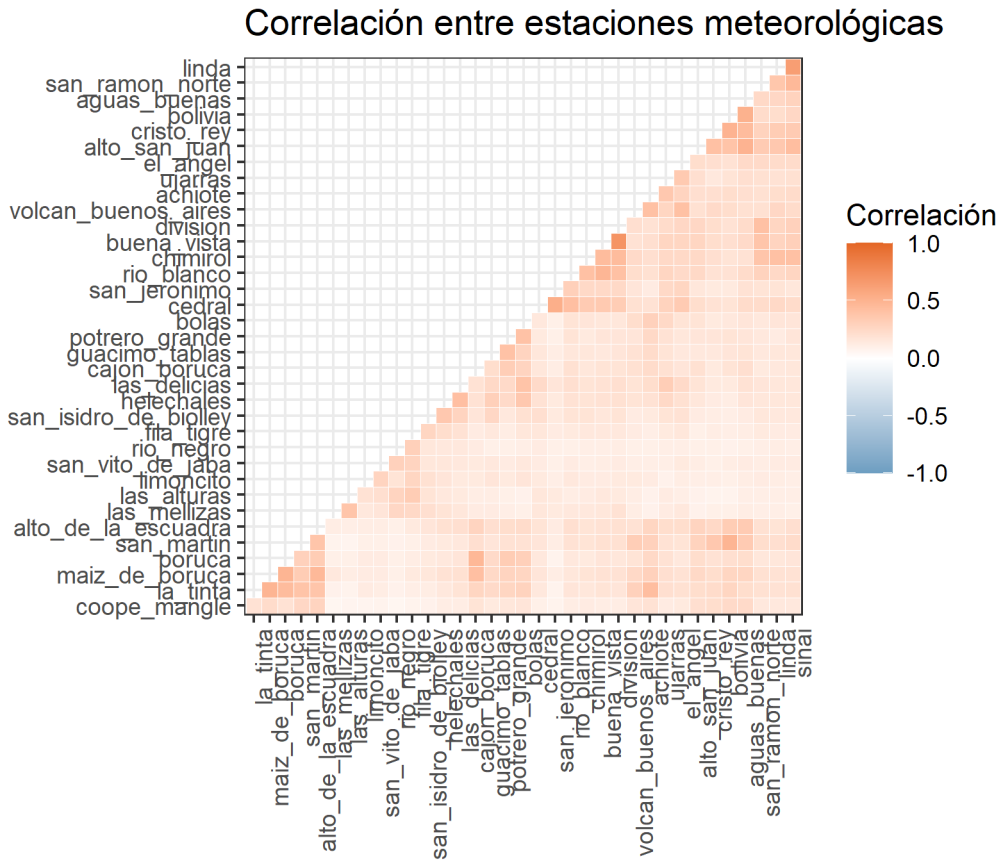


Figura 9: Correlación entre estaciones meteorológicas

Fuente: Elaboración propia con datos del ICE

Esta fórmula fue aplicada para completar los datos de todas las estaciones. En el Cuadro 3 se muestra la comparación entre los promedios y desviaciones estándar de los datos de cada estación previo y posterior a la imputación, con el fin de verificar que estos no hayan variado considerablemente debido a ésta. Se puede observar que la variación en cada estación es poca, por lo que se consideró que el resultado fue satisfactorio.

### 6.2.2 Imputación de datos hidrológicos

Previo a iniciar con el completado de datos faltantes, se decidió lo siguiente:

1. Dado el alto porcentaje de datos faltantes en la estación Caracucho (ver Figura 3), esta fue removida del análisis.

2. Al ser la estación Palmar la variable respuesta del modelo, se optó por no completar este registro pues se desea medir el error de los modelos únicamente sobre datos observados y no estimados. Los casos (horas) en los que esta estación cuenta con un dato faltante fueron removidos del análisis en su totalidad.

Al aplicar el segundo criterio, el conjunto de datos se redujo de 52584 a 31223 observaciones.

Cuadro 3: Comparación de promedio y desviación estándar de datos meteorológicos antes y después de la imputación de datos faltantes

| **Estación** | **Media inicial** | **Desv. Est. inicial** | **Media final** | **Desv. Est. final** |
|---|---|---|---|---|
| san_vito_de_j. | 0.37 | 2.09 | 0.40 | 2.09 |
| cedral | 0.46 | 2.64 | 0.51 | 2.79 |
| bolivia | 0.18 | 1.29 | 0.21 | 1.39 |
| potrero_grande | 0.22 | 1.71 | 0.25 | 1.88 |
| rio_negro | 0.32 | 2.11 | 0.36 | 2.26 |
| bolas | 0.39 | 2.57 | 0.43 | 2.66 |
| san_martin | 0.19 | 1.40 | 0.21 | 1.48 |
| ujarras | 0.37 | 2.61 | 0.43 | 2.85 |
| achiote | 0.32 | 2.30 | 0.34 | 2.31 |
| chimirol | 0.49 | 2.74 | 0.57 | 3.00 |
| rio_blanco | 0.30 | 1.76 | 0.35 | 1.94 |
| cristo_rey | 0.29 | 1.86 | 0.31 | 1.97 |
| buena_vista | 0.30 | 1.75 | 0.34 | 1.89 |
| alto_san_juan | 0.26 | 1.65 | 0.28 | 1.71 |
| san_jeronimo | 0.32 | 2.20 | 0.43 | 2.48 |
| guacimo_tablas | 0.22 | 1.63 | 0.28 | 1.82 |
| limoncito | 0.48 | 2.65 | 0.50 | 2.67 |
| maiz_de_boruca | 0.20 | 1.46 | 0.22 | 1.50 |
| cajon_boruca | 0.26 | 2.07 | 0.29 | 2.16 |
| alto_de_la_esc. | 0.27 | 1.98 | 0.29 | 2.02 |
| boruca | 0.20 | 1.49 | 0.21 | 1.56 |
| division | 0.40 | 2.24 | 0.48 | 2.55 |
| las_mellizas | 0.34 | 2.22 | 0.37 | 2.30 |
| las_alturas | 0.30 | 2.06 | 0.31 | 2.02 |
| fila_tigre | 0.41 | 2.35 | 0.45 | 2.52 |
| helechales | 0.35 | 2.19 | 0.40 | 2.30 |
| aguas_buenas | 0.29 | 1.98 | 0.31 | 2.10 |
| linda | 0.49 | 2.72 | 0.54 | 2.90 |
| san_ramon_n. | 0.40 | 2.26 | 0.46 | 2.44 |
| san_isidro_de_b. | 0.42 | 2.45 | 0.48 | 2.65 |
| las_delicias | 0.46 | 2.86 | 0.52 | 3.06 |
| sinai | 0.37 | 2.13 | 0.43 | 2.31 |
| coope_mangle | 0.40 | 2.64 | 0.44 | 2.86 |
| la_tinta | 0.22 | 1.61 | 0.23 | 1.60 |

Fuente: Elaboración propia con datos del ICE

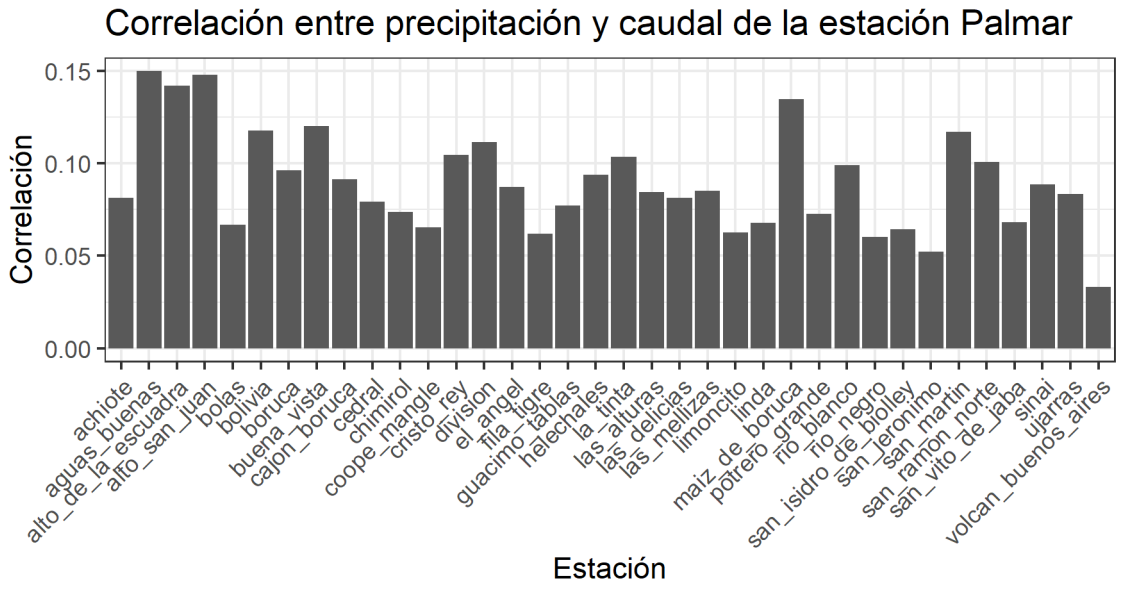


Figura 10: Correlación entre estaciones meteorológicas y la estación hidrológica Palmar

Fuente: Elaboración propia con datos del ICE

Para esta nueva cantidad de datos, el porcentaje de datos faltantes en El brujo y Cabagra es de 13.7% y 25.9% respectivamente.

Para completar los datos faltantes, se utilizó una ponderación de por Área - Lluvia. Sea $q_{i,f}$ el dato faltante de caudal en una hora $i$ de una estación $f$ y $Q_{i\pm t,c}$ el dato observado en la estación $c$ en una hora $i \pm t$ utilizado para la imputación, la ponderación está dada por:

$$q_{i,f} = \frac{A_f}{A_c}\frac{P_f}{P_c}Q_{i\pm t,c}$$

Donde:

- $A_f$: Área de la cuenca de la estación $f$.
- $A_c$: Área de la cuenca de la estación $c$.
- $P_f$: Precipitación media sobre la cuenca de la estación $f$.
- $P_c$: Precipitación media sobre la cuenca de la estación $c$.
- $t$: Tiempo de traslado entre estaciones

Es importante aclarar que en caso de que la precipitación media sobre alguna de las cuencas sea igual a cero, la fórmula anterior se ve reducida a:

$$q_{i,f} = \frac{A_f}{A_c}Q_{i\pm t,c}$$

Es decir, se pondera únicamente de acuerdo al área superficial de las cuencas.

Tanto el registro de El Brujo como el de Cabagra fueron imputados a partir del registro de Palmar, por lo que $A_c, P_c$ y $Q_{j,c}$ corresponden a datos de esta estación.

Para la estimación de la precipitación media de cuenca, se utilizó una ponderación de área de acuerdo con los polígonos de Voronoi. La fórmula utilizada para la ponderación es:

$$P_i = \frac{\sum_{i=1}^{n} A_i p_i}{\sum_{i=1}^{n} A_i}$$

Donde:

- $P$: Precipitación media sobre la cuenca.
- $A_i$: Área del polígono de Voronoi de la estación $i$.
- $p_i$: Precipitación en la estación $i$.
- $n$: Número de polígonos de Voronoi en la cuenca.

Para realizar la imputación, es necesario considerar el tiempo de tránsito que existe entre dos estaciones. Por ejemplo, el caudal observado en una hora $i$ en la estación El Brujo, se observará en Palmar en un tiempo $i + t$ donde $t$ es el tiempo que tarda el agua recorriendo la distancia entre estaciones o de tránsito. La Figura 11 muestra un ejemplo del tiempo de tránsito entre las estaciones Cabagra, El Brujo y Palmar. Se puede observar como para este evento particular, el cual ocurrió el 5 de noviembre de 2014, el tiempo de traslado (diferencia en tiempo entre caudales pico) entre Cabagra y Palmar es de aproximadamente 4 horas, mientras que con El Brujo es de 3 horas.

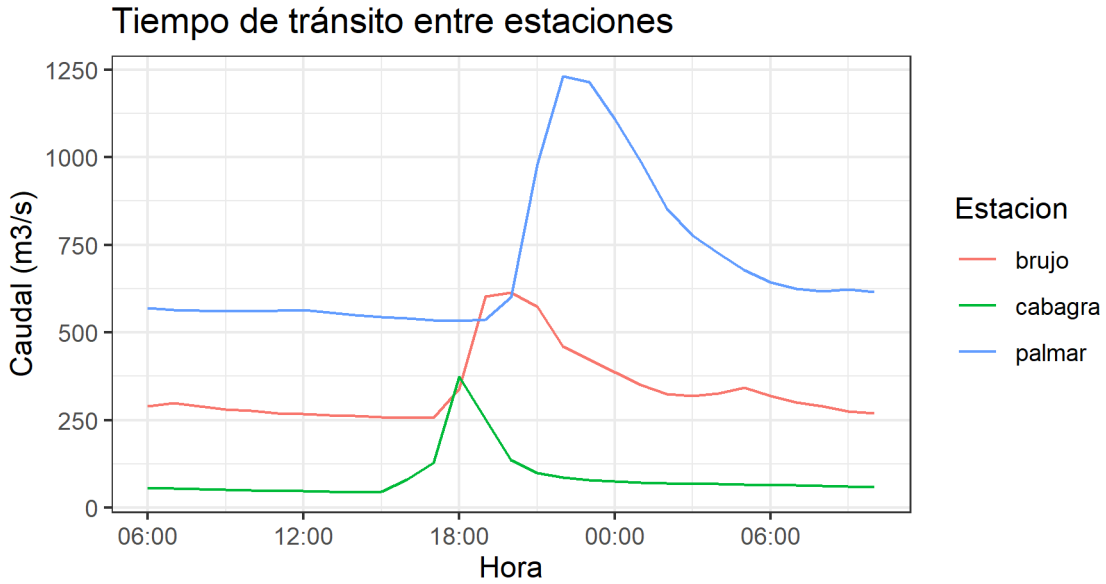


Figura 11: Ejemplo de tiempo de tránsito entre estaciones

Fuente: Elaboración propia con datos del ICE

El tiempo de tránsito no es constante. Existe una relación inversa entre éste y el caudal. Al incrementarse el flujo, se incrementa también su velocidad y, por ende, el tiempo de traslado es menor. Con el objetivo de encontrar una función que permita calcular $t$ como función del caudal,

se extrajo de forma manual diferentes eventos como el presentado en la Figura 11, se calculó el tiempo entre caudales máximos y se ajustó una curva de la forma $t = aQ^b$ mediante mínimos cuadrados para los datos obtenidos. La Figura 12 muestra las curvas ajustadas para obtener las funciones de tiempo de tránsito para ambas estaciones.

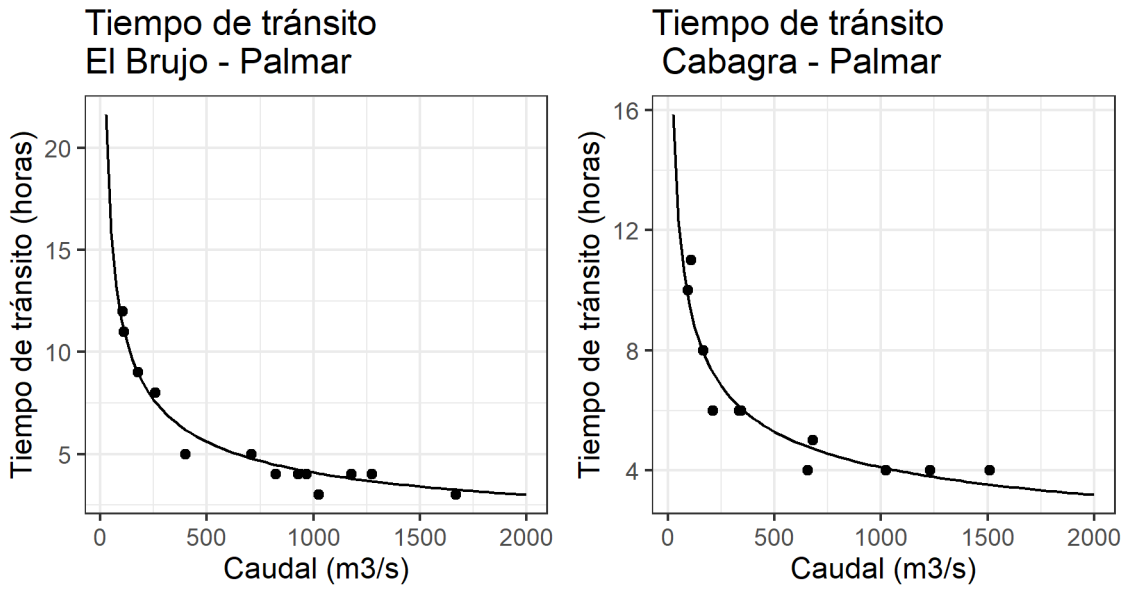


Figura 12: Funciones de tiempo de tránsito

Fuente: Elaboración propia con datos del ICE

A continuación, se presentan las ecuaciones de las curvas de ajuste de Figura 12. Estas fueron utilizadas para la estimación de los tiempos de traslado.

$$t_{ElBrujo} = 92.47 Q_{Palmar}^{-0.451}$$

y

$$t_{Cabagra} = 51.51 Q_{Palmar}^{-0.366}$$

Una vez obtenidas las áreas de las cuencas (Cuadro 1), calculado la precipitación media sobre ellas y estimado las curvas de tiempo de tránsito, se procedió a utilizar el caudal en la estación Palmar para imputar los datos en El Brujo y Cabagra. El Cuadro 4 muestra la variación en el promedio y desviación estándar de los datos de las estaciones previo y posterior a la imputación. Se observa que la variación en estos estadísticos es poca, por lo que se consideró que el resultado fue satisfactorio.

Cuadro 4: Comparación de promedio y desviación estándar de datos hidrológicos antes y después de la imputación de datos faltantes

| **Estación** | **Media inicial** | **Desv. Est. inicial** | **Media final** | **Desv. Est. final** |
|---|---|---|---|---|
| brujo | 165.40 | 159.96 | 161.38 | 161.31 |
| cabagra | 32.39 | 29.14 | 33.92 | 34.22 |

Fuente: Elaboración propia con datos del ICE

## 6.3 Modelado

Posterior a la finalización del imputado de datos faltantes, se procedió a utilizar las tablas de datos completas para modelar el caudal horario en la estación Palmar. El modelado se dividió en tres grandes partes: selección de variables de acuerdo con su importancia, optimización de parámetros de los modelos y comparación de los modelos optimizados. A continuación, se detallará cada una de las partes.

### 6.3.1 Importancia de variables

Como paso previo para realizar la estimación de la importancia de las variables independientes en la predicción, se calcularon rezagos de todos los predictores, ya que como se evidenció en la Figura 12 existe un tiempo de tránsito entre las estaciones. Ante esta situación se decidió incluir 16 rezagos de todas las variables.

Al realizar el análisis de importancia se encontró que los datos de El Brujo, Cabagra y sus respectivos rezagos, superan de manera considerable la importancia de todas las variables meteorológicas. Además, al incluir 16 rezagos de cada variable la cantidad de variables predictoras pasó de 34 a 578. Por estos dos motivos se decidió sustituir la información meteorológica puntual en cada estación por la precipitación promedio sobre las cuencas de El Brujo, Cabagra y Palmar utilizando la ponderación descrita en la Sección 6.2.2.

El reemplazo de los datos de las estaciones meteorológicas por la precipitación promedio redujo la cantidad de variables predictoras de 34 a 5. Al incluir los 16 rezagos de estas variables se alcanzó un total de 85, las cuales fueron sometidas al análisis de importancia.

Para el análisis, los datos fueron divididos en entrenamiento y prueba. El set de entrenamiento está compuesto por los datos desde las 08:00 del 07/05/2013 hasta las 00:00 del 05/10/2016. Esto representa un total de 23 645 observaciones, es decir, un 76.6% del total. Por otra parte, el set de prueba está conformado por las observaciones registradas entre la 01:00 del 05/10/2016 y las 00:00 del 05/10/2017. Es decir, se decidió utilizar toda la información contenida en el último año del registro. En este periodo se cuenta con un total de 7 231 observaciones.

Para cuantificar la importancia de las variables se utilizaron tres diferentes criterios: primero, la ganancia de información basada en entropía; segundo, el estadístico de independencia entre la

variable objetivo y las independientes Chi Cuadrado; tercero, la ganancia de información basada en permutaciones de bosques aleatorios. La Figura 13 muestra el resultado de las 40 variables con mayor puntaje en cada indicador. Se observa que en los tres gráficos las primeras 34 variables corresponden a los datos de El Brujo, Cabagra y sus respectivos rezagos. Las restantes seis son ocupadas por rezagos de la precipitación media en la cuenca de la estación Palmar y de El Brujo.

Para realizar la optimización de parámetros, se seleccionaron todas las variables que aparecieran listadas en alguno de los tres criterios de importancia. Lo anterior generó un total de 50 variables predictoras.

### 6.3.2 Ajuste de parámetros

Para lograr una adecuada comparación del desempeño de los modelos predictivos, es necesario realizar previamente un ajuste a los diferentes parámetros de cada uno de estos con el fin de aumentar, lo más posible, la calidad de su ajuste en el set de entrenamiento.

Como indicador de la calidad del ajuste de los modelos se utilizó el coeficiente de eficiencia de Nash-Sutcliffe, el cual es ampliamente utilizado para evaluar modelos hidrológicos. Su rango es de $]-\infty, 1]$ siendo 1 el valor óptimo. El coeficiente se calcula de la siguiente forma:

$$NSE = 1 - \frac{\sum_{t=1}^{T}(Q_m^t - Q_o^t)^2}{\sum_{t=1}^{T}(Q_o^t - \overline{Q_o})^2}$$

Donde:

- $Q_m^t$: Caudal modelado en el tiempo $t$
- $Q_o^t$: Caudal observado en el tiempo $t$
- $\overline{Q_o}$: Promedio del caudal observado

Como estrategia de remuestreo para la optimización del $NSE$, se utilizó una validación cruzada de ventana fija, con una ventana inicial del 20% de las observaciones del set de entrenamiento (4 729 horas) y un horizonte de pronóstico de dos semanas (336 horas).

A continuación, se presentarán los parámetros sometidos a ajuste para cada modelo, el rango evaluado y el resultado obtenido:

Modelo de árboles de decisión (DT)

- Complejidad [0.001, 0.006]: 0.001000125
- Número mínimo de observaciones en un nodo terminal [3, 12]: 9
- Máxima profundidad de cualquier nodo en el árbol final [5, 30]: 21

Modelo de bosques aleatorios (RF)

- Número posible de variables a dividir en cada nodo [2, 24]: 24
- Tamaño mínimo de nodo [2, 15]: 2

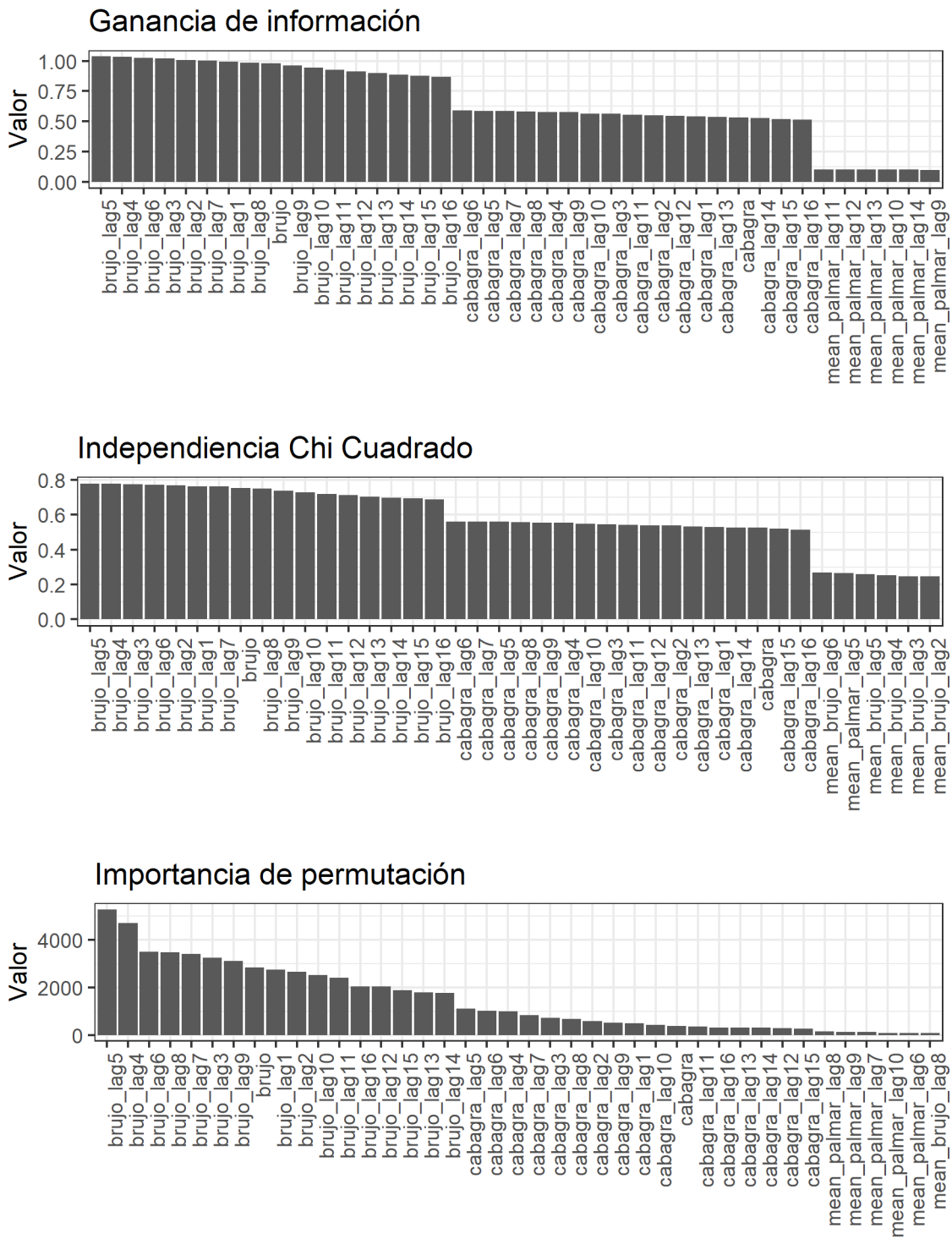


Figura 13: Importancia de variables

Fuente: Elaboración propia con datos del ICE

- Número de árboles [50, 800]: 624
- Fracción de observaciones a muestrear [0.3, 1]: 0.300026

Modelo de XGBoosting (XGB)

- Número máximo de iteraciones [100, 700]: 200
- Razón del número de columnas para el ajuste de cada árbol [0.3, 0.9]: 0.4698696
- Razón de submuestreo para la instancia de entrenamiento [0.5, 0.9]: 0.820163
- Tasa de aprendizaje [0.01, 0.35]: 0.01023758
- Máxima profundidad de un árbol [4, 12]: 11

### 6.3.3 Comparación de modelos

Una vez construidos los tres modelos, se procedió a comparar el desempeño de estos a la hora de predecir el caudal registrado en la estación Palmar en el set de prueba. Como se mencionó anteriormente, este set está compuesto por todos los registros del último año del conjunto de datos (de la 01:00 del 05/10/2016 a las 00:00 del 05/10/2017). Como medidas de comparación de desempeño se utilizaron: el Coeficiente Eficiencia de Nash-Sutcliffe (NSE), el cual se desea aproximar a su valor máximo de 1; la raíz cuadrada de error cuadrático medio (RMSE), el error medio absoluto (MAE), el error cuadrático medio (MSE) y el porcentaje de error absoluto medio (MAPE), los cuales se busca sean mínimos.

En el Cuadro 5 se muestra la comparación de todas las medidas para los diferentes modelos. Se observa que el modelo de Bosques Aleatorios cuenta con un desempeño superior a los otros modelos en cuatro de los cinco indicadores. Únicamente en el MAPE el modelo de *Boosting* supera al de Bosques Aleatorios. Este modelo (*Boosting*) también superó al de árboles de decisión en cuatro medidas, pues únicamente en el NSE obtuvo el mismo desempeño que el de Árboles de Decisión. Por esta razón se selecciona como el segundo mejor modelo para predecir este problema particular.

Cuadro 5: Comparación de medidas de eficiencia de predicción

| **Modelo** | **NSE** | **RMSE** | **MAE** | **MSE** | **MAPE** |
|---|---|---|---|---|---|
| Árboles de decisión | 0.81 | 148.01 | 84.30 | 21907.42 | 0.19 |
| Bosques Aleatorios | 0.86 | 126.14 | 69.92 | 15910.85 | 0.17 |
| XGBoosting | 0.81 | 145.18 | 71.85 | 21077.38 | 0.16 |

Fuente: Elaboración propia con datos del ICE

La Figura 14 muestra la comparación de la serie real y las modeladas para el conjunto de prueba. Se puede observar a grandes rasgos que las series presentan un comportamiento similar entre sí. Se observa que al inicio de la época seca (enero 2017) todos los modelos fallan en la predicción del comienzo de la recesión del caudal (disminución por falta de lluvia); sin embargo, rápidamente se logran ajustar y mejoran la predicción en los meses secos restantes (febrero, marzo

y abril). En términos generales el ajuste de los tres modelos es satisfactorio de acuerdo con lo observado en la figura, además de los altos valores de NSE presentados en el Cuadro 5.

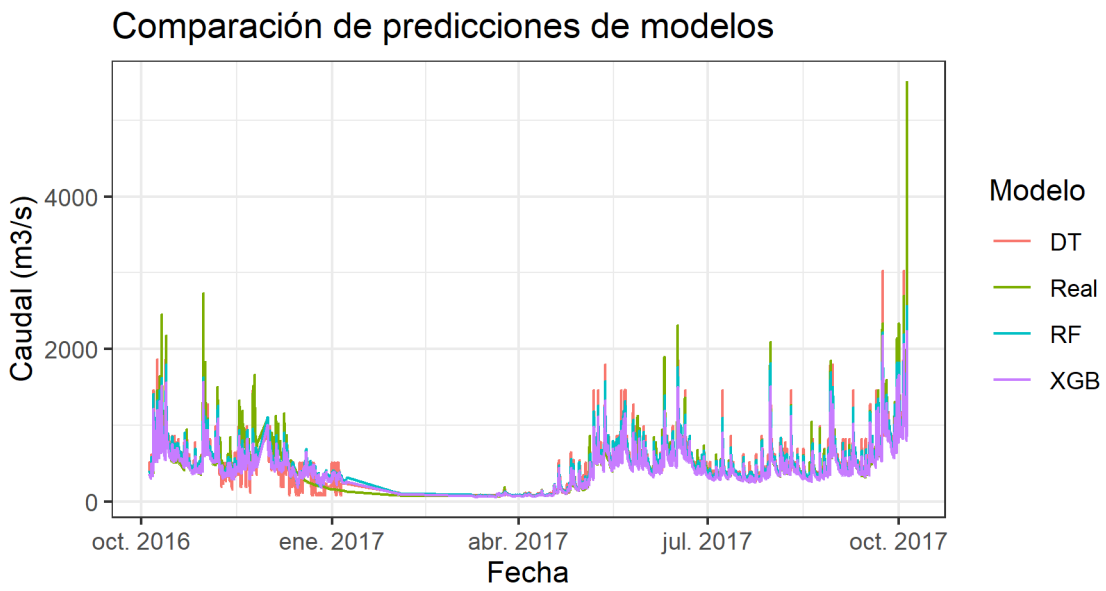


Figura 14: Comparación de resultados de predicción

Fuente: Elaboración propia con datos del ICE

### 6.3.4 Modelos con datos faltantes

La sección anterior muestra los resultados de la comparación de los modelos al utilizar los conjuntos de datos completos, es decir, los resultantes del proceso de imputación de datos faltantes. En esta sección se desea comparar los resultados previos (Cuadro 5) con los que se obtienen al utilizar la información previo a la imputación de datos.

Para esto, se siguió el mismo procedimiento que el presentado en las secciones 6.3.1, 6.3.2 y 6.3.3 con la diferencia de que el set de datos presenta faltantes. A pesar de esto, los datos cumplen con las dos consideraciones de la Sección 6.2.2, es decir, se eliminaron los datos de la estación Caracucho y únicamente se trabajó con las horas en donde el caudal en la estación Palmar fue registrado. Por esta razón, las dimensiones de este conjunto con datos faltantes son las mismas que para el caso anterior y se mantienen los porcentajes de datos faltantes para las estaciones de El Brujo y Cabagra descritos en la sección 6.2.2. En el caso de la precipitación media horaria sobre las distintas cuencas, esta fue calculada con solamente las estaciones que tuvieran información para la hora en cuestión, es decir, tampoco se realizó ningún completado de información meteorológica.

Para cálculo de la importancia de variables, fue necesario utilizar otro indicador que permitiera el cálculo a pesar de la presencia de valores faltantes. El indicador escogido fue la importancia de acuerdo con el modelo de Bosques Aleatorios. La Figura 15 muestra resultados similares a lo obtenido anteriormente (Figura 13) ya que las primeras 34 variables más importantes son los caudales en El Brujo y Cabagra más todos sus rezagos. Se decidió seleccionar

un total de 50 variables para mantener la misma cantidad que en las secciones anteriores. Las 16 variables restantes incluidas son precipitaciones medias correspondientes a los rezagos del 5 al 12 de diferentes cuencas.

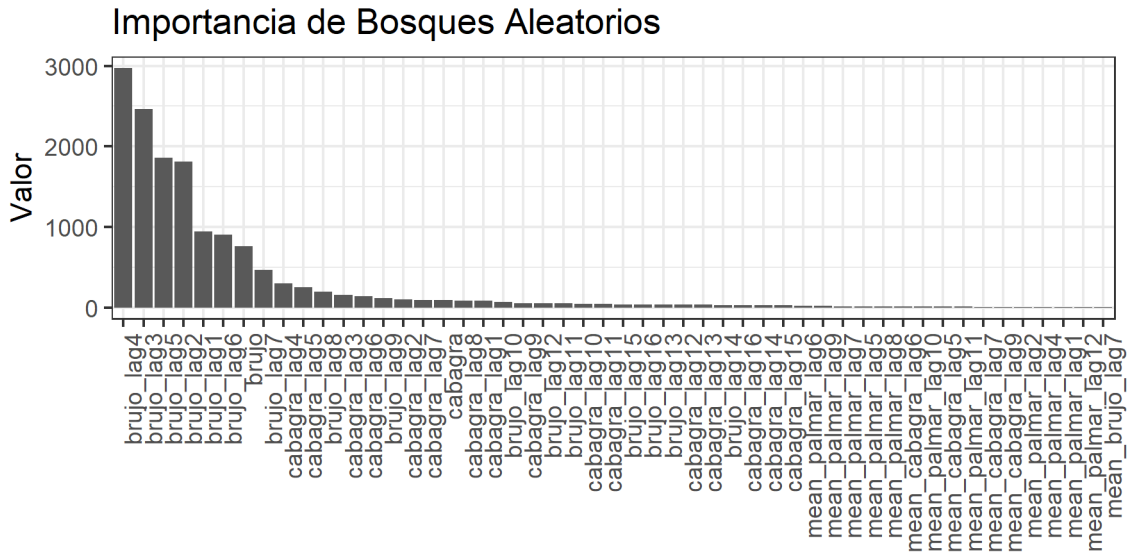


Figura 15: Importancia de variables en set con datos faltantes

Fuente: Elaboración propia con datos del ICE

Una vez seleccionadas las variables se realizó el ajuste de los mismos parámetros de los modelos que se utilizaron previamente (6.3.2). Los resultados obtenidos se detallan a continuación.

Modelo de árboles de decisión (DT)

- Complejidad [0.001, 0.006]: 0.001000046
- Número mínimo de observaciones en un nodo terminal [3, 12]: 5
- Máxima profundidad de cualquier nodo en el árbol final [5, 30]: 16

Modelo de bosques aleatorios (RF)

- Número posible de variables a dividir en cada nodo [2, 24]: 2
- Tamaño mínimo de nodo [2, 15]: 2
- Número de árboles [50, 800]: 329
- Fracción de observaciones a muestrear [0.3, 1]: 0492868

Modelo de XGBoosting (XGB)

- Número máximo de iteraciones [100, 700]: 138
- Razón del número de columnas para el ajuste de cada árbol [0.3, 0.9]: 0.4497482

- Razón de submuestreo para la instancia de entrenamiento [0.5, 0.9]: 0.5916034
- Tasa de aprendizaje [0.01, 0.35]: 0.01006512
- Máxima profundidad de un árbol [4, 12]: 10

Una vez definidos los parámetros óptimos para cada modelo, se procedió a calcular los indicadores de desempeño para cada uno. El Cuadro 6 muestra los resultados obtenidos. Al igual que en el caso anterior, el modelo de Bosques Aleatorios es el que presenta un mejor desempeño. En todos los indicadores presenta una diferencia considerable con respecto a los dos modelos restantes. El segundo mejor modelo es el de árboles de decisión, el cual supera en cuatro de los cinco indicadores al modelo de *Boosting*. Al comparar el desempeño de los modelos al utilizar datos faltantes en los conjuntos de datos contra los sets completos, se observa que estos tienen un menor desempeño. Únicamente el modelo de Bosques Aleatorios presenta medidas levemente inferiores a las obtenidas en el Cuadro 5, mientras que los dos modelos restantes presentaron resultados considerablemente inferiores. Estos resultados evidencian la importancia de la imputación de los datos faltantes.

Cuadro 6: Comparación de medidas de eficiencia de predicción en modelos con datos faltantes

| **Modelo** | **NSE** | **RMSE** | **MAE** | **MSE** | **MAPE** |
|---|---|---|---|---|---|
| Árboles de decisión | 0.66 | 197.82 | 113.12 | 39133.09 | 0.36 |
| Bosques Aleatorios | 0.80 | 149.33 | 80.65 | 22300.00 | 0.22 |
| XGBoosting | 0.65 | 198.70 | 119.40 | 39480.90 | 0.28 |

Fuente: Elaboración propia con datos del ICE

# 7 Conclusiones y recomendaciones

Al realizar el análisis exploratorio de los datos, se encontró que la correlación entre la precipitación y el caudal a nivel horario en la estación Palmar es baja si esta se compara con la correlación que existe entre las series hidrológicas.

El análisis exploratorio también evidenció que los registros de precipitación presentan una cantidad mucho menor de datos faltantes que los hidrológicos, lo cual obedece al riesgo de daño que presentan las estaciones hidrológicas ante eventos extremos y un mantenimiento más complejo.

El resultado del análisis de importancia de variables para la predicción del caudal horario arrojó como resultado que los registros de caudal y sus respectivos rezagos, tienen un peso mucho mayor que el de los registros de precipitación media, en todos los escenarios analizados. Lo anterior refleja la importancia de contar con diferentes registros de caudal dentro de un área de estudio.

El análisis anterior también confirmó la importancia de considerar el tiempo de tránsito que existe entre las diferentes estaciones, ya que se observó que, en términos generales, muchos de los rezagos de ambas estaciones hidrológicas tienen una mayor importancia que la variable sin rezagar.

Al analizar el desempeño de los modelos predictivos, se consideró, que en términos generales, los modelos de *ML* basados en árboles de decisión tienen desempeños aceptables para este caso particular. Lo anterior basado en que todos los valores de NSE obtenidos fueron superiores a 0.8.

El modelo que permitió modelar de mejor forma el caudal en la estación Palmar, de acuerdo con el NSE, fue el de Bosque Aleatorios, utilizando como predictores el conjunto de variables posterior a la imputación de datos faltantes. En segundo lugar, se ubicó el modelo de Potenciación y tercer lugar el de Árboles de Decisión. Este resultado es congruente con lo expuesto en el marco teórico, ya que tanto el modelo de Bosques Aleatorios como el de Potenciación surgen como mejoras al modelo de Árboles de Decisión, por lo que se esperaba que su desempeño fuera superior.

Con el planteamiento del escenario del ajuste de modelos sin realizar la imputación de datos faltantes, se encontró que el completado de datos mejora la predicción de los modelos, principalmente en los modelos de Árboles de Decisión y Potenciación donde el NSE subió en aproximadamente 0.2.

A pesar de haber excluido el registro de la estación Caracucho, el cual contiene la información del sector oeste de la cuenca del río Grande de Térraba, el modelo genera resultados satisfactorios. Sin embargo, se recomienda analizar toda la información disponible para determinar si es posible encontrar un periodo en el cual se cuente con información en todas las estaciones y así observar el aporte que podrían realizar los datos de esta estación en el modelo.

Ante los resultados obtenidos, se recomienda continuar con el mantenimiento de la red hidrometeorológica, dado que la información generada permite obtener conjuntos de datos completos que son el insumo para otros modelos hidrológicos. A pesar de que la información de precipitación no es tan relevante para los modelos utilizados, esta podría ser de mayor importancia para otro nivel de información (diaria, mensual o anual), por lo que es importante continuar con su registro.

# References


Ardabili, S., Mosavi, A., Dehghani, M., and Várkonyi-Kóczy, A. R. (2020). Deep learning and machine learning in hydrological processes climate change and earth systems a systematic review. In Várkonyi-Kóczy, A. R., editor, *Engineering for Sustainable Future*, volume 101, pages 52–62. Springer International Publishing.

Asefa, T., Kemblowski, M., McKee, M., and Khalil, A. (2006). Multi-time scale stream flow predictions: The support vector machines approach. *Journal of Hydrology*, 318(1-4):7–16.

Beauchamp, J., Downing, D., and Railsback, S. (1989). Comparison of regression and time-series methods for synthesizing missing streamflow records. *Journal of the American Water Resources Association*, 25(5):961–975.

Gyau-Boakye, P. and Schultz, G. A. (1994). Filling gaps in runoff time series in West Africa. *Hydrological Sciences Journal*, 39(6):621–636.

James, G., Witten, D., Hastie, T., and Tibshirani, R. (2013). *An Introduction to Statistical Learning*, volume 103 of *Springer Texts in Statistics*. Springer New York.

Karakurt, O., Erdal, H. I., Namli, E., and Yumurtaci-Aydogmus, H. (2013). Comparing ensembles of decision trees and neural networks for one-day-ahead streamflow prediction. *Scientific Research Journal*, page 13.

Karunanithi, N., Grenney, W. J., Whitley, D., and Bovee, K. (1994). Neural Networks for River Flow Prediction. *Journal of Computing in Civil Engineering*, page 20.

Kim, J.-W. and Pachepsky, Y. A. (2010). Reconstructing missing daily precipitation data using regression trees and artificial neural networks for SWAT streamflow simulation. *Journal of Hydrology*, 394(3-4):305–314.

Mwale, F., Adeloye, A., and Rustum, R. (2012). Infilling of missing rainfall and streamflow data in the Shire River basin, Malawi – A self organizing map approach. *Physics and Chemistry of the Earth, Parts A/B/C*, 50-52:34–43.

Rasouli, K., Hsieh, W. W., and Cannon, A. J. (2012). Daily streamflow forecasting by machine learning methods with weather and climate inputs. *Journal of Hydrology*, 414-415:284–293.

Shortridge, J. E., Guikema, S. D., and 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.

Tongal, H. and Booij, M. J. (2018). Simulation and forecasting of streamflows using machine learning models coupled with base flow separation. *Journal of Hydrology*, 564:266–282.

Wu, M.-C. and Lin, G.-F. (2015). An Hourly Streamflow Forecasting Model Coupled with an Enforced Learning Strategy. *Water*, 7(11):5876–5895.