Desarrollo e integración de un modelo predictivo para la calidad del agua en Esval

Aplicación del modelo Random Forest con variables hidrológicas y climáticas en la PTAP San Juan, Región de Valparaíso

Carla Silva Saavedra

Profesora guía: Karine Bertin

INSTITUTO DE INGENIERÍA MATEMÁTICA, FACULTAD DE INGENIERÍA
UNIVERSIDAD DE VALPARAÍSO

Valparaíso, Enero 2026

Agradecimientos

Quiero expresar mi más sincero agradecimiento a todas las personas e instituciones que, de distintas maneras, hicieron posible la realización de esta memoria de título y me acompañaron a lo largo de todo mi proceso universitario.

En primer lugar, agradezco profundamente a mis padres, Mónica y Roberto, por su apoyo incondicional durante toda mi formación universitaria. Su constante preocupación, dedicación y confianza fueron un pilar fundamental para que pudiera avanzar y perseverar en este camino. Gracias por estar siempre presentes, por el cariño, los consejos y el acompañamiento constante que me brindaron en cada etapa, permitiéndome alcanzar esta importante meta.

A mis hermanos, Orlando, Karen y Pascale, les agradezco por el apoyo, el cariño y la contención entregada a lo largo de estos años. Cada uno, a su manera, contribuyó a hacer este camino más llevadero, recordándome la importancia de reír, desconectarse y seguir confiando en mí misma.

Quiero dedicar un agradecimiento especial a Akira, quien estuvo a mi lado de manera incondicional durante innumerables jornadas de estudio. Su compañía constante, silenciosa y fiel fue un apoyo invaluable en los momentos buenos y difíciles de esta etapa, transformándose en un pilar emocional fundamental durante mi vida universitaria.

Agradezco también a Matías, por su compañía, comprensión y palabras de aliento durante gran parte de este proceso. Su apoyo fue fundamental en los momentos de mayor exigencia, entregándome tranquilidad, motivación y la fuerza necesaria para seguir adelante.

A mis amigas y compañeras de la universidad, agradezco el apoyo, la compañía y los momentos compartidos durante esta etapa. Las jornadas de estudio, las risas y las conversaciones hicieron de este proceso una experiencia más llevadera.

Agradezco a mi profesora guía, Karine Bertin, por su acompañamiento, dedicación y compromiso durante el desarrollo de esta memoria. Su constante disposición, sus valiosas observaciones y el tiempo que dedicó a orientarme fueron clave para mi crecimiento académico y profesional, especialmente en la etapa final de mi formación.

Asimismo, agradezco a los integrantes de mi comisión evaluadora, Cristian Meza y Daniel Madrid, por su disposición a formar parte de este proceso. En particular, agradezco a Daniel por la confianza depositada en mí a lo largo de mis distintas etapas en Esval, así como por su apoyo, orientación y por brindarme la oportunidad de desarrollar este proyecto en un entorno profesional enriquecedor.

Extiendo mis agradecimientos a la empresa Esval, por abrirme las puertas en distintas instancias de mi formación y por permitirme desarrollar este trabajo en colaboración con su equipo. Agradezco especialmente la disposición, el apoyo recibido y el acceso a la información necesaria para la realización de esta memoria, esperando que este trabajo también represente un aporte para la organización.

Agradezco sinceramente a mi jefe de carrera, Héctor Olivero, por su permanente apoyo durante toda mi trayectoria universitaria. Su disposición, cercanía y orientación fueron fundamentales para superar diversas etapas del proceso académico, convirtiéndose en un respaldo constante a lo largo de la carrera.

A la profesora Andrea Jiménez, le agradezco por su confianza, apoyo y por considerarme en distintas instancias de difusión y actividades asociadas a la carrera, contribuyendo significativamente a mi desarrollo académico y profesional.

Asimismo, agradezco a los profesores y a la secretaría de la carrera de Ingeniería Civil Matemática, por la orientación y el acompañamiento brindados durante todos estos años, los cuales fueron fundamentales en mi proceso formativo.

Finalmente, agradezco al Proyecto FONDECYT N°1221373, por el financiamiento otorgado, el cual permitió apoyar actividades académicas vinculadas a este trabajo, contribuyendo a su difusión y fortalecimiento.

Resumen

La gestión eficiente del agua potable requiere anticipar tanto la disponibilidad del recurso como la evolución de los parámetros que determinan su calidad, especialmente en contextos de creciente variabilidad climática. En este escenario, la capacidad de proyectar el comportamiento del caudal y de contaminantes relevantes constituye una herramienta clave para apoyar la toma de decisiones operativas en los sistemas de tratamiento de agua potable.

El presente trabajo desarrolla una metodología predictiva basada en aprendizaje automático para la estimación y proyección del caudal y de los parámetros de calidad del agua sólidos disueltos totales (SDT), nitratos (NO_3^-) y turbiedad, utilizando información histórica y variables externas de carácter climático e hidrológico. En particular, el análisis considera registros mensuales de contaminantes disponibles para el período enero de 2021 a mayo de 2025, una serie histórica de caudal correspondiente al período enero de 2000 a mayo de 2025, y variables climáticas externas (temperatura, precipitaciones y nieve acumulada) que incluyen información histórica desde enero de 2000 y proyecciones futuras hasta junio de 2030. El estudio se realiza en la Planta de Tratamiento de Agua Potable (PTAP) San Juan, ubicada en el sector de Llolleo, comuna de San Antonio, región de Valparaíso, y se desarrolla en colaboración con la empresa sanitaria Esval, en el marco de un trabajo aplicado a un sistema real de operación.

La metodología propuesta se basa en la implementación del modelo Random Forest, el cual permite capturar relaciones no lineales y dependencias temporales complejas entre las variables analizadas. Para cada variable de interés se evalúan distintas configuraciones del modelo, considerando la selección del número óptimo de variables predictoras y del número de árboles del bosque, mediante el uso de métricas de error absoluto y relativo. Asimismo, se analiza la importancia relativa de las variables seleccionadas, se generan proyecciones a horizontes de 1 y 5 años y se valida el desempeño predictivo del modelo mediante la comparación entre valores proyectados y observados.

Los resultados obtenidos muestran que el modelo Random Forest es capaz de representar adecuadamente la dinámica temporal del caudal y de los contaminantes analizados, entregando proyecciones coherentes con el comportamiento histórico del sistema y niveles de error aceptables en los períodos de validación. En conjunto, este trabajo aporta una base metodológica replicable que puede ser extendida a otras plantas de tratamiento y utilizada como apoyo en la gestión preventiva y operativa de la calidad del agua potable.

Índice general

Capítulo 1

Introducción General

La disponibilidad y calidad del agua potable constituyen un desafío prioritario para la gestión de los recursos hídricos, particularmente en regiones que enfrentan una creciente presión climática y ambiental, como la Región de Valparaíso en Chile. En los últimos años, la disminución sostenida de los caudales, la ocurrencia de sequías prolongadas y el aumento en la frecuencia de eventos climáticos extremos han incrementado la incertidumbre respecto del comportamiento futuro tanto de la cantidad como de la calidad del agua disponible para consumo humano.

En este contexto, las empresas sanitarias requieren herramientas que les permitan anticipar escenarios de riesgo asociados a variaciones en los parámetros de calidad del agua, de modo de apoyar la toma de decisiones operativas y fortalecer una gestión preventiva del recurso. Entre los parámetros de mayor interés se encuentran los sólidos disueltos totales (SDT), los nitratos (NO_3^-) y la turbiedad, cuya evolución depende de la interacción entre factores hidrológicos, climáticos y temporales, y que pueden presentar variaciones significativas en períodos de corto y mediano plazo.

El presente estudio se desarrolla en colaboración con la empresa sanitaria Esval y se centra específicamente en la Planta de Tratamiento de Agua Potable (PTAP) San Juan, ubicada en el sector de Llolleo, comuna de San Antonio, región de Valparaíso. Esta planta cumple un rol relevante en el abastecimiento de agua potable para la zona y opera a partir de agua cruda proveniente de fuentes superficiales, las cuales pueden experimentar cambios importantes tanto en caudal como en calidad. En este escenario, resulta fundamental contar con estimaciones anticipadas de la concentración de contaminantes en el agua que ingresa a la planta, con el fin de apoyar la planificación operativa, evaluar posibles escenarios de riesgo y contribuir a una gestión más robusta

del proceso de tratamiento.

Ante esta necesidad, los modelos de aprendizaje automático surgen como una alternativa adecuada para el análisis de sistemas ambientales complejos, ya que permiten capturar relaciones no lineales, incorporar múltiples variables explicativas y adaptarse a dinámicas altamente variables. En particular, el modelo Random Forest destaca por su capacidad predictiva, su robustez frente al ruido en los datos y su potencial para analizar la importancia relativa de las variables que influyen en el comportamiento del sistema.

Este trabajo tiene como propósito desarrollar y aplicar una metodología predictiva para la estimación y proyección del caudal y de los principales contaminantes de la calidad del agua en la PTAP San Juan, utilizando información histórica de carácter hidrológico y climático. En particular, las series de caudal y variables climáticas externas (temperatura, precipitaciones y nieve acumulada) se encuentran disponibles para el período comprendido entre enero de 2000 y mayo de 2025, mientras que los registros de los contaminantes sólidos disueltos totales (SDT), nitratos (NO_3^-) y turbiedad abarcan el período enero de 2021 a mayo de 2025. A partir de este caso de estudio, se busca generar una base metodológica que pueda ser extendida posteriormente a otras plantas de tratamiento y a otros parámetros de interés para la empresa sanitaria.

En el contexto de la generación de proyecciones futuras, este estudio integra información proveniente de distintas fuentes. En particular, se dispone de proyecciones climáticas de temperatura del aire, precipitaciones y nieve acumulada obtenidas desde la plataforma ARCLIM, las cuales permiten contar con valores futuros de estas variables externas. En el caso del caudal, no se dispone de proyecciones directas para el período de interés, por lo que se desarrolla un modelo predictivo específico basado en Random Forest con el fin de estimar su evolución futura. De este modo, las proyecciones climáticas obtenidas desde ARCLIM, junto con las proyecciones de caudal generadas en este estudio, constituyen la base de entrada para la proyección de los contaminantes sólidos disueltos totales, nitratos y turbiedad mediante modelos Random Forest.

En las siguientes secciones de esta introducción se presenta, en primer lugar, la motivación del estudio y su relevancia en el contexto de la gestión del recurso hídrico. Posteriormente, se formulan el objetivo general y los objetivos específicos que guían el desarrollo del trabajo. Finalmente, se describe la estructura del documento, detallando la organización de los capítulos y el contenido abordado en cada uno de ellos.

1.1. Motivación del estudio

La gestión eficiente del agua potable depende tanto de la disponibilidad del recurso como de la capacidad de anticipar cambios en su calidad, especialmente en sistemas de abastecimiento que dependen de fuentes superficiales sujetas a una alta variabilidad hidrológica y climática. En este contexto, la modelación predictiva de parámetros de calidad del agua adquiere un rol fundamental como herramienta de apoyo a la toma de decisiones operativas y estratégicas.

En la PTAP San Juan, ubicada en la comuna de San Antonio, la calidad del agua cruda que ingresa a la planta puede experimentar variaciones relevantes asociadas a cambios en el caudal, eventos de precipitación, arrastre de material particulado y procesos de dilución o concentración de sustancias disueltas. Estos factores pueden impactar directamente parámetros como los sólidos disueltos totales, los nitratos y la turbiedad, los cuales resultan críticos para la operación del proceso de tratamiento y el cumplimiento de los estándares de calidad del agua potable.

Desde el punto de vista operativo, contar con estimaciones anticipadas del comportamiento de estos contaminantes permite mejorar la planificación del tratamiento, optimizar el uso de insumos, anticipar escenarios de riesgo y fortalecer una gestión preventiva frente a episodios desfavorables. En este sentido, la capacidad de proyectar la evolución futura de los parámetros de calidad del agua representa una ventaja significativa para las empresas sanitarias, particularmente en un escenario de creciente incertidumbre climática.

Asimismo, la motivación de este estudio se sustenta en la necesidad de incorporar metodologías modernas de análisis de datos que permitan abordar la complejidad a los sistemas ambientales. Los modelos de aprendizaje automático, como Random Forest, ofrecen una alternativa estable frente a enfoques tradicionales, al ser capaces de capturar relaciones no lineales, integrar múltiples variables explicativas y adaptarse a series temporales con comportamiento irregular.

Finalmente, este trabajo se enmarca en una colaboración directa con la empresa sanitaria Esval, lo que refuerza su carácter aplicado y su orientación hacia un problema real de interés público. El desarrollo de una metodología predictiva basada en datos reales de operación aporta un valor directo a la gestión de la PTAP San Juan y establece las bases para la extensión del enfoque a otras plantas de tratamiento y a la evaluación de nuevos parámetros de calidad del agua en el futuro.

1.2. Objetivos del estudio

El presente trabajo tiene como propósito central desarrollar una metodología de carácter predictivo que permita anticipar el comportamiento del caudal y de distintos parámetros de calidad del agua en la PTAP San Juan, utilizando información histórica y variables externas de carácter hidrológico y climático. A través de este enfoque, se busca aportar una herramienta cuantitativa que apoye la gestión preventiva y operativa del sistema, permitiendo una mejor comprensión de la dinámica temporal de los contaminantes y su relación con el comportamiento del caudal. Asimismo, este estudio se plantea como un caso piloto desarrollado en colaboración con Esval, con el objetivo de establecer una base metodológica replicable y extensible a otros sistemas de tratamiento de agua potable.

1.2.1. Objetivo general

Desarrollar una metodología predictiva para la estimación y proyección del caudal y de los parámetros de calidad del agua sólidos disueltos totales (SDT), nitratos (NO_3^-) y turbiedad en la PTAP San Juan de Esval, considerando variables externas de carácter climático e hidrológico, con el fin de aportar una herramienta de apoyo a la gestión preventiva y operativa del sistema.

1.2.2. Objetivos específicos

- Procesar, limpiar y estructurar los datos históricos de caudal, contaminantes y variables externas disponibles, asegurando su consistencia temporal y calidad para el análisis.
- Analizar la relación entre el caudal y variables climáticas externas, tales como temperatura, precipitaciones y nieve acumulada, mediante herramientas de análisis exploratorio y correlacional.
- Desarrollar modelos predictivos basados en Random Forest para el caudal como etapa previa y fundamental para la proyección de los parámetros de calidad del agua.
- Implementar modelos predictivos basados en Random Forest para los contaminantes sólidos disueltos totales (SDT), nitratos (NO_3^-) y turbiedad, evaluando su

desempeño mediante métricas de error absolutas y relativas.

- Analizar la importancia relativa de las variables predictoras en cada modelo, con el fin de interpretar los principales factores que influyen en la dinámica de cada parámetro.
- Generar proyecciones de corto y mediano plazo para el caudal y los contaminantes analizados, evaluando su coherencia respecto del comportamiento histórico observado.
- Validar las proyecciones generadas mediante la comparación entre valores proyectados y observados, utilizando el error porcentual absoluto medio (MAPE) como medida de desempeño predictivo.
- Identificar posibles líneas de extensión del estudio hacia otras plantas de tratamiento de Esval, la incorporación de nuevas variables explicativas y el uso de series temporales con mayor resolución temporal.

1.3. Estructura del documento

El presente documento se estructura en cuatro capítulos principales, los cuales permiten desarrollar de manera ordenada el marco conceptual, la metodología aplicada, los resultados obtenidos y las conclusiones derivadas de este estudio.

En el capítulo **Marco teórico: modelo Random Forest** se presentan los fundamentos conceptuales que sustentan el desarrollo del trabajo. En particular, se describe el funcionamiento del algoritmo Random Forest, sus principales características y ventajas, así como su aplicación en problemas de regresión asociados a sistemas ambientales e hidrológicos. Además, se introducen las métricas de error empleadas para evaluar el desempeño de los modelos predictivos, proporcionando el marco conceptual necesario para la interpretación de los resultados obtenidos.

El capítulo **Metodología y descripción de los datos** presenta el enfoque metodológico del estudio y las fuentes de información utilizadas. En particular, se describen los datos históricos de caudal, contaminantes y variables climáticas externas asociadas a la PTAP San Juan, junto con los procesos de preprocesamiento y análisis exploratorio realizados. Asimismo, se expone el procedimiento de implementación del modelo Random Forest, la generación de proyecciones a 1 y 5 años y el esquema de validación empleado para evaluar el desempeño de las proyecciones.

En el capítulo **Resultados para la PTAP San Juan** se presentan y analizan los resultados obtenidos a partir de la aplicación del modelo Random Forest. Este capítulo aborda la modelación del caudal y de los parámetros de calidad del agua considerados, sólidos disueltos totales (SDT), nitratos (NO_3^-) y turbiedad, incluyendo la selección de configuraciones óptimas del modelo, el análisis de la importancia relativa de las variables predictoras, las proyecciones a uno y cinco años y la validación de las proyecciones mediante métricas de error absoluto y relativo. Además, se incorpora una comparación general de los resultados obtenidos para las distintas variables analizadas.

Finalmente, el capítulo **Conclusiones generales** sintetiza los principales hallazgos del estudio, discute sus implicancias para la gestión del sistema de tratamiento de agua potable de la PTAP San Juan y plantea los principales desafíos y líneas de trabajo futuro.

Capítulo 2

Marco teórico: modelo Random Forest

En este capítulo se presenta el marco teórico asociado al modelo Random Forest, el cual constituye la base metodológica utilizada para la construcción de los modelos predictivos desarrollados en este estudio. Se describen los conceptos fundamentales que permiten comprender el funcionamiento del algoritmo, su estructura interna y los principios que explican su funcionamiento en problemas de regresión.

Random Forest es un método de aprendizaje automático basado en la combinación de múltiples modelos simples, cuyo objetivo es mejorar la capacidad predictiva y la estabilidad de las predicciones frente a la variabilidad de los datos. Este enfoque construye un conjunto de predictores que, al combinarse, permiten capturar relaciones complejas y no lineales entre las variables, reduciendo el riesgo de sobreajuste y mejorando la capacidad de generalización del modelo. Esta característica resulta adecuada para el análisis de sistemas ambientales e hidrológicos, donde las relaciones entre variables suelen ser no lineales, ruidosas y altamente dependientes del tiempo.

En primer lugar, el enfoque adoptado en este trabajo se enmarca en la resolución de problemas de regresión, cuyo objetivo es modelar la relación entre una variable respuesta continua y un conjunto de variables predictoras. En este contexto, el modelo Random Forest se construye a partir de árboles de decisión, que actúan como unidades básicas encargadas de aprender reglas de partición a partir de los datos. Estos árboles utilizan criterios de impureza para realizar divisiones sucesivas del espacio de predictores, buscando generar regiones cada vez más homogéneas en términos de la variable de interés, es decir, subconjuntos de observaciones en los que los valores de la variable que

se quiere predecir presentan un menor variabilidad entre sí.

Posteriormente, se aborda el enfoque de bagging como estrategia para reducir la variabilidad de las predicciones y mejorar la estabilidad del modelo, mediante la combinación de múltiples árboles entrenados sobre diferentes subconjuntos de los datos. Sobre esta base, se presenta formalmente el modelo Random Forest, destacando sus principales características en problemas de regresión.

Finalmente, se presentan elementos para la evaluación e interpretación del modelo, entre ellos el análisis de la importancia de variables, que permite identificar los predictores más relevantes en la explicación de la variable objetivo, y el uso de métricas de error para cuantificar su desempeño predictivo. En este trabajo se emplean el error absoluto medio (MAE) y el error porcentual absoluto medio (MAPE), los cuales permiten evaluar el comportamiento del modelo tanto en la etapa de ajuste como en la comparación entre valores estimados y observados.

2.1. Problemas de regresión

Un problema de regresión consiste en modelar la relación existente entre una variable respuesta continua y un conjunto de variables explicativas, con el objetivo de aproximar dicha relación a partir de los datos disponibles y utilizarla posteriormente con fines predictivos.
De manera general, la relación entre la variable de interés y el conjunto de predictores puede representarse como:

$$y_t = f(x_t) + e_t \quad , \quad t = 1, \ldots, T,$$

donde:

- y_t corresponde a la variable que se desea predecir en el instante t, ya sea el caudal o alguna de los contaminantes analizados.

- x_t representa el vector de variables predictoras disponibles en el instante t, el cual incluye variables climáticas, hidrológicas y sus respectivos rezagos temporales.

- $f(\cdot)$ es una función desconocida que describe la relación entre las variables predictoras y la variable respuesta.

- e_t corresponde a un término de error aleatorio que recoge la variabilidad no explicada por el modelo.

El objetivo central de un modelo de regresión es obtener una aproximación adecuada de la función $f(\cdot)$ a partir de un conjunto finito de observaciones, de modo que sea posible explicar el comportamiento histórico de la variable de interés y generar predicciones confiables. En la práctica, la forma funcional de $f(\cdot)$ no es conocida y puede ser altamente compleja, especialmente en sistemas ambientales e hidrológicos, donde las relaciones entre variables suelen ser no lineales, ruidosas y dependientes del tiempo.

En este trabajo, la función $f(\cdot)$ es aproximada mediante el modelo Random Forest, el cual construye una estimación de la relación existente entre las variables predictoras y la variable de interés a partir de la combinación de múltiples árboles de regresión. Este enfoque permite capturar relaciones no lineales y efectos de interacción entre las variables predictoras, lo que resulta especialmente adecuado para el análisis de series temporales ambientales con múltiples fuentes de variabilidad.

2.2. Árboles de decisión

Los árboles de decisión constituyen la estructura básica sobre la cual se construye el modelo Random Forest. A través de ellos, el algoritmo aprende reglas de partición que permiten relacionar una variable de interés con un conjunto de variables explicativas, organizando dichas decisiones de manera jerárquica y secuencial.

Un árbol se estructura a partir de un nodo inicial que contiene todas las observaciones disponibles. A partir de este nodo, los datos se van dividiendo sucesivamente mediante nodos intermedios, cada uno asociado a una condición sobre alguna variable predictora. Este proceso continúa hasta alcanzar nodos finales o terminales, en los cuales se asigna una predicción. De esta forma, cada observación sigue un recorrido definido por las divisiones del árbol, lo que permite representar la relación entre las variables mediante reglas. Con el fin de ilustrar gráficamente la estructura descrita, en la figura 2.1 se presenta un esquema general de un árbol de decisión:

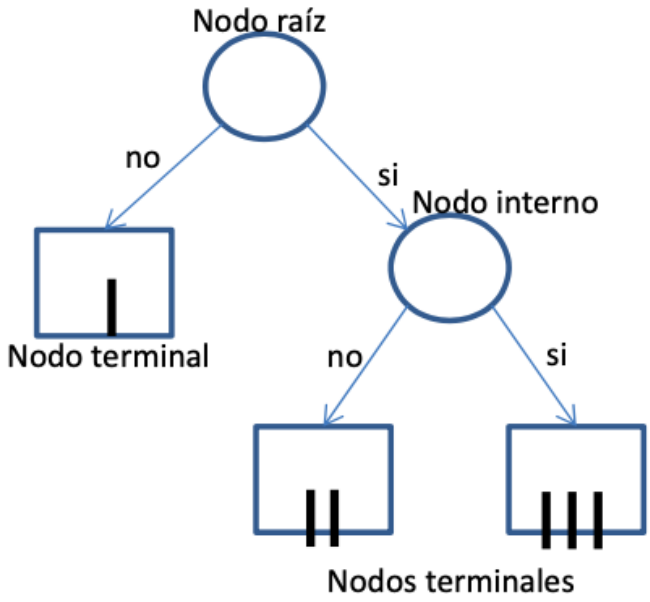

Figura 2.1: Esquema general de un árbol de decisión.

En la figura 2.1 se observa cómo el conjunto total de observaciones se organiza de manera jerárquica a partir de un nodo raíz, dando lugar a sucesivas particiones representadas por nodos internos. Cada partición corresponde a una restricción sobre una variable predictora, definida mediante un punto de corte numérico, lo que permite dividir el espacio de predictores en regiones más específicas. A través de estas particiones, cada observación sigue un único recorrido desde el nodo raíz hasta un nodo terminal, quedando asociada a una única región del espacio de variables.

El tipo de árbol se define en función de la variable respuesta, determinando si el problema corresponde a un árbol de clasificación o de regresión. En los árboles de clasificación, la variable de interés es categórica y la predicción corresponde a la clase predominante en cada nodo terminal. En cambio, en los árboles de regresión, la variable respuesta es continua y la predicción asignada en cada nodo terminal corresponde al valor promedio de dicha variable calculado sobre las observaciones contenidas en el nodo.

Más precisamente, cuando se consideran variables predictoras numéricas, cada partición de un nodo se construye a partir de reglas del tipo:

$$x^{(k)} \leq a \quad \text{y} \quad x^{(k)} > a,$$

donde $x^{(k)}$ denota una de las variables predictoras y a corresponde al punto de corte numérico que define la división. Estas reglas permiten separar el conjunto de observaciones que alcanza el nodo en dos subconjuntos disjuntos, de modo que cada observación pertenece a una única rama del árbol.

En los árboles de decisión, los nodos terminales representan las regiones finales del espacio de predictores. En el caso de árboles de regresión, dichas regiones se denotan por $R_1, R_2, \ldots, R_M$, donde M corresponde al número total de nodos terminales. Cada región R_m se asocia a una predicción constante c_m, correspondiente al valor asignado a todas las observaciones que pertenecen a dicha región. En consecuencia, el árbol de regresión define una función por tramos de la forma:

$$\hat{f}(x) = \sum_{m=1}^{M} c_m \, \mathbf{1}\{x \in R_m\},$$

donde R_m representa las regiones en que el árbol divide el espacio de predictores y c_m es el valor asignado en cada una de ellas. En particular, la predicción asociada a la región R_m corresponde al valor promedio de la variable respuesta sobre las observaciones contenidas en dicha región, es decir,

$$c_m = \overline{y}_{R_m},$$

donde, si R es una región que contiene N_R observaciones, se define:

$$\overline{y}_R = \frac{1}{N_R} \sum_{i \in R} y_i.$$

En este estudio se utilizan árboles de regresión, dado que las variables analizadas corresponden a magnitudes continuas, como el caudal y las concentraciones de contaminantes. Esta elección permite capturar relaciones no lineales entre variables climáticas, hidrológicas y la calidad del agua, manteniendo una estructura que favorece la interpretación de los resultados.

A modo ilustrativo, se presenta en la figura 2.2 un ejemplo de un árbol de regresión construido para la variable turbiedad, entrenado con una profundidad reducida y cuyo objetivo es visualizar de manera concreta el criterio de división descrito anteriormente. En particular, el esquema mostrado corresponde a un árbol con cuatro niveles de partición, presentado únicamente con fines explicativos, de modo de facilitar la comprensión del mecanismo de selección de variables y puntos de corte.

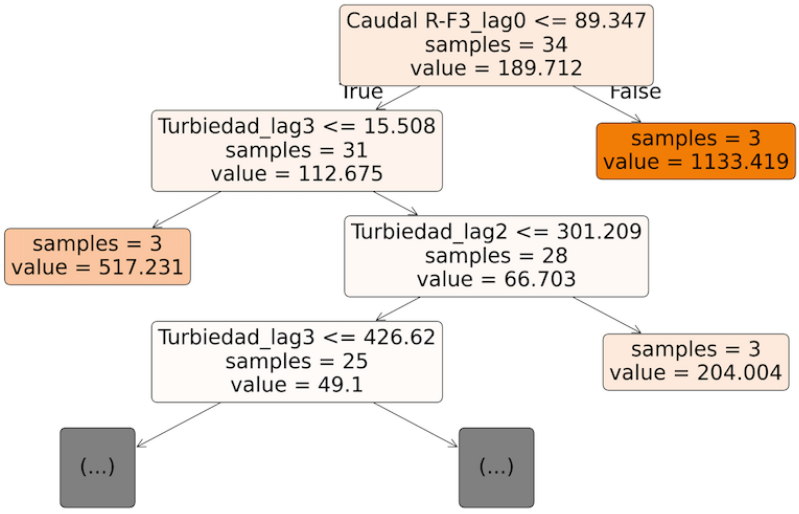

Figura 2.2: Ejemplo de árbol de regresión para turbiedad y su criterio de división.

En la figura 2.2, cada nodo del árbol presenta la siguiente información:

- *Samples*: número de observaciones que alcanzan el nodo correspondiente.
- *Value*: valor promedio de la variable respuesta, en este caso la turbiedad, calculado sobre las observaciones contenidas en el nodo, el cual constituye la predicción asociada a dicho nodo.
- La expresión ubicada en la parte superior del árbol de regresión (en el ejemplo, Caudal R-F3 lag0 $\leq$ 89,347) indica la variable predictora utilizada para realizar la partición y el punto de corte numérico que define la división del conjunto de datos en dicho nodo.

En el caso de las variables predictoras utilizadas en el árbol, la notación lag hace referencia a rezagos temporales. En particular, lag0 corresponde al valor de la variable en el mismo instante de tiempo en que se realiza la predicción, mientras que lag1, lag2, lag3, etc., representan los valores de dicha variable en uno, dos o tres instantes de tiempo anteriores, respectivamente. La inclusión de estos rezagos permite incorporar información temporal al modelo, capturando posibles efectos retardados de las variables climáticas e hidrológicas sobre la turbiedad.

En particular, en la figura 2.2 se pueden identificar algunas de las regiones finales o nodos terminales del árbol de decisión. A modo ilustrativo, se presentan a continuación tres de dichas regiones, definidas a partir de la intersección de las reglas de partición

que se aplican a lo largo del recorrido desde el nodo raíz hasta cada nodo terminal:

$$R_1 = \{\text{Caudal R-F3 lag0} > 89{,}347\},$$

$$R_2 = \{\text{Caudal R-F3 lag0} \leq 89{,}347,\ \text{Turbiedad lag3} \leq 15{,}508\},$$

$$R_3 = \{\text{Caudal R-F3 lag0} \leq 89{,}347,\ \text{Turbiedad lag3} > 15{,}508,\ \text{Turbiedad lag2} > 301{,}209\}.$$

Cada una de estas regiones corresponde a un nodo terminal del árbol y agrupa observaciones que cumplen simultáneamente las condiciones impuestas por las particiones sucesivas. En cada región R_m, la predicción asignada por el árbol corresponde a un valor constante c_m, definido como el promedio de la variable respuesta sobre las observaciones contenidas en dicha región.

En el ejemplo considerado, cada una de estas regiones contiene tres observaciones, por lo que las predicciones asociadas vienen dadas por:

$$c_1 = \overline{y}_{R_1} = 1133{,}419, \qquad c_2 = \overline{y}_{R_2} = 517{,}231, \qquad c_3 = \overline{y}_{R_3} = 204{,}004.$$

Cabe destacar que estas regiones corresponden únicamente a una parte del árbol mostrado en la figura, mientras que en niveles inferiores el árbol puede continuar creciendo y definir nuevos nodos terminales y regiones adicionales del espacio de predictores.

En el ejemplo de la figura 2.2, el nodo raíz contiene 34 observaciones y entrega como predicción el valor promedio de la turbiedad correspondiente a dicho conjunto. La primera partición del árbol se realiza utilizando la variable Caudal R-F3 lag0, con un punto de corte cercano a 89.347. Esta división separa el conjunto inicial de observaciones en dos subconjuntos, los cuales presentan comportamientos distintos de la turbiedad en función del caudal.

A partir de esta primera división, el árbol continúa aplicando el mismo procedimiento de manera recursiva. Las observaciones que satisfacen la condición impuesta por la partición avanzan por la rama correspondiente y constituyen el nuevo conjunto de datos sobre el cual se evalúan nuevamente distintas variables predictoras y posibles puntos de corte. En cada nivel del árbol, el número de *samples* se actualiza de acuerdo con la cantidad de observaciones que alcanzan el nodo, mientras que el valor *value* se recalcula como el promedio de la turbiedad sobre dicho subconjunto.

En los siguientes niveles del árbol aparecen particiones basadas en la turbiedad y en sus rezagos temporales, lo que indica que, una vez condicionado el rango de caudal,

la información histórica de la turbiedad resulta relevante para caracterizar los distintos comportamientos de la variable respuesta. De este modo, el árbol construye progresivamente regiones del espacio de predictores cada vez más específicas.

Cabe destacar que el árbol ilustrado continúa creciendo más allá de los niveles mostrados en la figura. Sin embargo, para efectos de este análisis, se presenta únicamente un número acotado de niveles con el objetivo de ilustrar la estructura del árbol y el mecanismo de partición del espacio de predictores. Cada nodo terminal representa una región del espacio de predictores dentro de la cual el modelo asigna una predicción constante, correspondiente al promedio de la turbiedad calculado sobre las observaciones que pertenecen a dicha región.

Si bien un árbol individual permite modelar relaciones complejas entre las variables, su principal relevancia en este trabajo radica en su integración dentro del modelo Random Forest, donde múltiples árboles se combinan para construir un predictor más robusto y estable. Los mecanismos que permiten mejorar el desempeño mediante esta combinación se abordan en los apartados siguientes.

2.3. Impureza y criterios de división

En un principio, el proceso de construcción de un árbol de decisión se basa en la realización de divisiones sucesivas del conjunto de datos, con el objetivo de organizar las observaciones en nodos donde la variable de interés presente una menor variabilidad. Para evaluar la calidad de estas divisiones y elegir las divisiones óptimas en cada nodo se introducen las medidas de impureza, las cuales permiten cuantificar qué tan dispersos son los valores de la variable respuesta dentro de un nodo.

En el contexto de los árboles de regresión, un nodo se considera más homogéneo cuando los valores de la variable respuesta contenidos en él son similares entre sí, es decir, cuando presentan una baja dispersión. Por el contrario, un nodo es más impuro cuando dichos valores son muy distintos. En este sentido, la impureza se asocia a la variabilidad interna de la variable que se desea predecir (Breiman et al., 1984; Hastie et al., 2009).

Sea R un nodo que contiene N_R observaciones de la variable respuesta y_i. La media del nodo se define como:

$$\overline{y}_R = \frac{1}{N_R} \sum_{i \in R} y_i \quad ,$$

y la impureza del nodo se mide mediante la varianza de la variable respuesta dentro del

nodo, calculada a partir de las observaciones que contiene:

$$I(R) = \frac{1}{N_R} \sum_{i \in R} (y_i - \bar{y_R})^2.$$

Esta expresión corresponde a una medida de dispersión de la variable respuesta dentro del nodo y permite cuantificar la variabilidad interna de los valores contenidos en él. En árboles de regresión del tipo CART (Classification and Regression Trees), esta cantidad se utiliza como criterio estándar para evaluar la homogeneidad de un nodo (Breiman et al., 1984). En este proceso, la predicción asociada a cada nodo corresponde siempre al valor promedio de la variable respuesta, mientras que las variables explicativas se utilizan únicamente para definir las divisiones que agrupan las observaciones. El criterio de división de un nodo consiste en seleccionar la variable predictora y el punto de corte que produzcan la mayor reducción de impureza. Si un nodo R se divide en dos nodos hijos, R_{izq} y R_{der}, la reducción de impureza asociada a dicha división se define como:

$$\Delta I = I(R) - [I(R_{izq}) + I(R_{der})].$$

El algoritmo evalúa todas las divisiones candidatas definidas por la elección de una variable predictora y un punto de corte numérico, y selecciona aquella que maximiza ΔI, es decir, la que produce nodos hijos con menor variabilidad interna que el nodo original. Este criterio de selección puede observarse de manera concreta en el ejemplo del árbol de regresión para turbiedad presentado en la figura 2.2. En dicho caso, cada partición del árbol se define a partir de la elección de una variable predictora y un punto de corte que permiten reducir la variabilidad de la variable respuesta dentro de los nodos resultantes.

En particular, la primera división del nodo raíz se realiza utilizando la variable Caudal R-F3 lag0, con un punto de corte cercano a 89.347. Esta partición separa el conjunto inicial de observaciones en dos subconjuntos cuyos valores de turbiedad presentan una menor dispersión interna en comparación con el nodo original, lo que implica una reducción de la impureza según la definición de ΔI.

De manera análoga, en los niveles siguientes del árbol, las particiones basadas en la turbiedad y en sus rezagos temporales corresponden a divisiones que continúan disminuyendo la variabilidad de la variable respuesta dentro de cada nodo. Así, el crecimiento del árbol refleja la aplicación recursiva del criterio de reducción de impureza, mediante el cual el algoritmo selecciona en cada nodo la división que maximiza la disminución

de la varianza interna de la turbiedad.

Este procedimiento se aplica de manera recursiva en cada nodo del árbol, permitiendo estructurar el espacio de predictores en regiones donde la variable respuesta presenta comportamientos más uniformes (Hastie et al., 2009).

Tal como se observa en el ejemplo de la figura 2.2, el crecimiento del árbol continúa mientras existan divisiones que permitan reducir la impureza del nodo. En este estudio, el criterio de detención se alcanza cuando no es posible encontrar una partición adicional que produzca una disminución de la impureza. En ese momento, el nodo se considera terminal y la predicción asociada corresponde al valor medio de la variable respuesta calculado sobre las observaciones contenidas en dicho nodo.

Las medidas de impureza y los criterios de división determinan la estructura de los árboles individuales y constituyen un componente fundamental del modelo Random Forest. En particular, la reducción de impureza generada por cada variable a lo largo de las divisiones se utiliza posteriormente para cuantificar su importancia relativa dentro del modelo, permitiendo identificar qué predictores contribuyen en mayor medida a explicar la variabilidad de la variable de interés (Breiman, 2001).

En conjunto, la noción de impureza y los criterios de división constituyen el mecanismo fundamental mediante el cual los árboles de regresión organizan la información contenida en los datos, y representan un componente esencial para comprender tanto el funcionamiento de los árboles individuales como su integración posterior en el modelo Random Forest utilizado en este trabajo.

2.4. Bagging

Los árboles de decisión son modelos flexibles que permiten capturar relaciones complejas entre las variables, pero presentan una alta sensibilidad al conjunto de datos utilizado para su entrenamiento. Esta característica implica que pequeñas variaciones en los datos pueden producir estructuras de árbol distintas y, en consecuencia, predicciones con alta variabilidad. Para abordar este problema, Breiman (1996) propuso la técnica conocida como bagging (Bootstrap Aggregating), cuyo objetivo es reducir la varianza del modelo mediante la combinación de múltiples árboles de decisión.

El principio fundamental del bagging consiste en entrenar varios árboles de decisión a partir de distintas muestras bootstrap generadas desde el conjunto de datos original. Una muestra bootstrap se obtiene seleccionando observaciones de manera aleatoria

con reemplazo, hasta conformar un conjunto de datos en general del mismo tamaño que el conjunto original. Este procedimiento implica que algunas observaciones pueden aparecer más de una vez dentro de una misma muestra, mientras que otras pueden no ser seleccionadas. Como resultado, cada árbol se entrena con un conjunto de datos ligeramente distinto, lo que introduce diversidad entre los modelos individuales.

A partir de cada muestra bootstrap se construye un árbol de decisión independiente, utilizando los mismos criterios de partición y de impureza descritos en la sección 2.3. En el caso de problemas de regresión, cada árbol entrega una predicción para una observación dada, y la predicción final del modelo bagging se obtiene promediando las predicciones individuales de todos los árboles construidos. Si se consideran B árboles y se denota por $\widehat{f}^{(b)}(x)$ la predicción del árbol b para una observación x, la predicción agregada se expresa como:

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

Este promedio entre árboles permite reducir la varianza del estimador final, ya que al combinar varios árboles se obtiene una predicción más estable que la entregada por un único modelo.

Esta reducción de la varianza puede explicarse observando cómo se comporta la variabilidad de una predicción cuando se promedian varios modelos construidos a partir de distintas muestras bootstrap del mismo conjunto de datos. Bajo el supuesto de que cada árbol individual entrega una predicción con varianza σ^2 y que existe una correlación ρ entre las predicciones de dos árboles distintos, la varianza del estimador obtenido mediante bagging está dada por:

$$\mathrm{Var}\left(\hat{f}_{\mathrm{bag}}(x)\right) = \frac{\sigma^2}{B}\Big[1 + (B-1)\rho\Big]. \tag{2.1}$$

Esta expresión muestra que, siempre que la correlación entre las predicciones de los árboles sea menor que uno, la varianza del promedio disminuye al aumentar el número de árboles. De este modo, el bagging permite disminuir la variabilidad asociada a cada árbol individual y obtener estimaciones más estables (Breiman, 1996).

Cabe destacar que, en el esquema de bagging, cada árbol se construye de forma independiente y utiliza el conjunto completo de variables predictoras disponibles para realizar las divisiones. La diversidad entre los árboles proviene de las diferencias entre las muestras bootstrap, lo que ya permite una mejora significativa en la estabilidad del modelo respecto de un único árbol de decisión.

En síntesis, el bagging constituye un mecanismo eficaz para reducir la varianza a los árboles de decisión, mejorando su desempeño predictivo y su capacidad de generalización. Esta técnica representa la base del modelo Random Forest, el cual extiende el enfoque de bagging incorporando un nivel adicional de aleatoriedad en la selección de las variables predictoras (Breiman, 2001).

2.5. Modelo Random Forest

El modelo Random Forest es un método de aprendizaje automático basado en la combinación de múltiples árboles de decisión. Este enfoque puede entenderse como una extensión del bagging, en la cual, además de utilizar muestras bootstrap para generar varios árboles, se introduce un mecanismo adicional de aleatoriedad en la selección de las variables predictoras.

Al igual que en bagging, cada árbol del Random Forest se construye a partir de una muestra bootstrap del conjunto de datos original. Sin embargo, durante la construcción del árbol, en cada nodo no se consideran todas las variables predictoras disponibles, sino solo un subconjunto seleccionado de manera aleatoria. A partir de ese subconjunto se elige la variable y el punto de corte que generan la mayor reducción de impureza. Este procedimiento evita que siempre se utilicen las mismas variables dominantes en las primeras divisiones.

Como resultado, los árboles que componen el bosque capturan distintos patrones presentes en los datos. La predicción final del modelo se obtiene promediando las predicciones individuales de todos los árboles, lo que permite reducir la variabilidad del modelo y obtener estimaciones más estables y consistentes.

Denotando por $\hat{f}_{\mathrm{RF}}^{(b)}(x)$ la predicción entregada por el árbol b del bosque, la predicción final del modelo Random Forest se expresa como:

$$\hat{f}_{\mathrm{forest}}(x) = \frac{1}{B}\sum_{b=1}^{B} \hat{f}_{\mathrm{RF}}^{(b)}(x).$$

A diferencia del estimador bagging, en el modelo Random Forest cada $\hat{f}_{\mathrm{RF}}^{(b)}$ incorpora una fuente adicional de aleatoriedad asociada a la selección de un subconjunto de variables predictoras en cada división del árbol. Este mecanismo tiene un efecto directo sobre la correlación entre las predicciones de los árboles individuales.

En efecto, como se observa en la expresión de la varianza del estimador bagging dada

en la ecuación (2.1), la varianza del estimador agregado depende, entre otros factores, de la correlación ρ entre las predicciones de los árboles. En el caso de Random Forest, la selección aleatoria de variables en cada nodo contribuye a disminuir dicha correlación, ya que fuerza a que los árboles utilicen distintos predictores y estructuras de partición.

Como consecuencia, al reducirse el valor de ρ, el término $(B-1)\rho$ disminuye, lo que se traduce en una reducción adicional de la varianza del estimador final respecto del esquema de bagging.

En este estudio, el modelo Random Forest se aplica en un contexto de series temporales, incorporando rezagos de la variable respuesta y de las variables externas como predictores. De esta manera, el modelo es capaz de capturar dependencias temporales, efectos estacionales y respuestas diferidas propias de los procesos hidrológicos y de calidad del agua analizados.

Tras el proceso de entrenamiento, el modelo permite reproducir el comportamiento histórico de la serie y generar proyecciones futuras de la variable de interés. El ajuste histórico entrega una estimación del desempeño del modelo sobre el período observado, permitiendo evaluar su capacidad para capturar la dinámica temporal y la relación con las variables explicativas consideradas.

En el caso de las proyecciones, el procedimiento se realiza de forma secuencial en el tiempo. En cada instante futuro, los valores estimados previamente por el modelo se utilizan para construir los rezagos necesarios de la variable respuesta, los cuales se incorporan nuevamente como predictores. De manera complementaria, se consideran las variables externas proyectadas, tales como variables climáticas o hidrológicas, asegurando la coherencia entre la información disponible y el horizonte de proyección.

Este esquema de proyección iterativa permite mantener una continuidad temporal entre el período observado y el período proyectado. De esta forma, el modelo Random Forest es capaz de generar trayectorias futuras que respetan la estructura temporal de los datos y reflejan tanto la influencia de las condiciones externas como las dependencias propias de la serie analizada.

En conjunto, el modelo Random Forest combina la flexibilidad de los árboles de decisión, la reducción de varianza obtenida mediante el bagging y la diversidad generada por la selección aleatoria de variables. Estas características lo convierten en una herramienta adecuada para modelar sistemas complejos y no lineales, como los procesos hidrológicos y de calidad del agua considerados en este trabajo.

2.6. Importancia de variables

Una de las ventajas del modelo Random Forest es que permite evaluar la relevancia relativa de cada variable predictora en la construcción del modelo. Esta información permite complementar el análisis predictivo con una interpretación del rol que desempeñan las distintas variables explicativas en el comportamiento de la variable de interés.

En el contexto de Random Forest, la importancia de una variable se define a partir de su contribución a la reducción de la impureza a lo largo del conjunto de árboles que componen el bosque. En particular, cada vez que una variable es utilizada para realizar una división en un nodo de un árbol, se produce una disminución de la impureza asociada a esa partición, donde, dicha reducción se asocia directamente a la variable que define la división del nodo.

La importancia de una variable se define a partir de la suma de las reducciones de impureza que genera en las distintas divisiones de todos los árboles del bosque. Posteriormente, esta suma puede normalizarse para expresar la importancia relativa de cada predictor en términos porcentuales. De este modo, una variable será considerada más importante si, en promedio, contribuye en mayor medida a disminuir la variabilidad de la variable respuesta dentro de los nodos del árbol.

Este criterio de importancia se encuentra directamente unido a las medidas de impureza descritas anteriormente en la sección 2.3. En el caso de árboles de regresión, la reducción de impureza se basa en la disminución de la suma de los errores cuadrados dentro de los nodos, por lo que una variable será relevante en la medida en que permita generar particiones que agrupen observaciones con valores más homogéneos de la variable respuesta (Breiman, 2001).

En este estudio, la importancia de variables se utiliza como una herramienta de análisis complementaria, que permite identificar cuáles predictores (variables externas y rezagos temporales) tienen un mayor impacto en la explicación de la dinámica del caudal y de los contaminantes analizados. Este análisis resulta clave para interpretar el comportamiento del modelo, evaluar la pertinencia de las variables incluidas y apoyar la toma de decisiones desde una perspectiva operativa y ambiental.

En conjunto, la medida de importancia de variables proporciona una visión interpretativa del modelo Random Forest, permitiendo ir más allá de la predicción y comprender el rol que desempeña cada predictor en la estructura del bosque.

2.7. MAE (Mean Absolute Error)

Para evaluar el comportamiento del estimador Random Forest $\hat{f}_{forest}(x)$ frente a distintas configuraciones, en este estudio se utiliza el error absoluto medio, conocido como MAE por sus siglas en inglés (*Mean Absolute Error*). Esta métrica permite cuantificar, en promedio, qué tan alejadas se encuentran las predicciones del modelo respecto de los valores observados.
Sea y_t el valor observado de la variable respuesta en el instante t e $\hat{y}_t = \hat{f}(x_t)$ la predicción entregada por el modelo en ese mismo instante. El MAE se define como:

$$MAE = \frac{1}{T}\sum_{t=1}^{T}|y_t - \hat{y}_t|,$$

donde T corresponde al número total de observaciones consideradas en la evaluación.

Una de las principales ventajas del MAE es que se expresa en las mismas unidades de la variable analizada, lo que permite interpretar el error de manera directa y consistente al comparar distintas configuraciones del modelo.

En este trabajo, el MAE se utiliza como criterio central para la selección de los principales parámetros del modelo Random Forest. En particular, esta métrica se emplea para analizar cómo varía el error al modificar el número de variables predictoras incluidas en el modelo. A partir de este análisis, es posible identificar configuraciones en las que el MAE se estabiliza, lo que permite seleccionar un conjunto de predictores que capture adecuadamente la dinámica del sistema sin incorporar complejidad innecesaria.

De manera análoga, el MAE se utiliza para definir el número óptimo de árboles del bosque. Al evaluar la evolución del error en función de la cantidad de árboles, se puede identificar un rango a partir del cual el MAE presenta variaciones marginales. Este criterio permite seleccionar un número de árboles que entregue un desempeño adecuado, evitando incrementos innecesarios en el costo computacional.

En conjunto, el MAE cumple un rol fundamental en la etapa de calibración del modelo Random Forest, ya que proporciona una medida objetiva y consistente para comparar distintas configuraciones. Su utilización permite fundamentar la elección del número de variables predictoras y del número de árboles sobre la base del comportamiento del error, asegurando un equilibrio entre desempeño y eficiencia del modelo.

2.8. MAPE (Mean Absolute Percentage Error)

Para evaluar el desempeño de las proyecciones generadas por el estimador Random Forest, en este estudio se utiliza el error absoluto medio porcentual, conocido como MAPE por sus siglas en inglés (*Mean Absolute Percentage Error*). Esta métrica permite cuantificar el error de predicción en términos relativos, expresándolo como un porcentaje respecto de los valores reales observados.

Sea y_t el valor real de la variable respuesta en el instante t e $\hat{y}_t$ el valor proyectado por el modelo en ese mismo instante. Considerando un horizonte de proyección de H instantes, el MAPE se define como:

$$\text{MAPE} = \frac{100}{H} \sum_{t=T+1}^{T+H} \frac{|\hat{y}_t - y_t|}{|y_t|},$$

donde T corresponde al último instante del período utilizado para el entrenamiento del modelo.

El MAPE se construye a partir del cálculo del error porcentual absoluto en cada instante del período de proyección, el cual mide cuánto difiere la proyección del modelo respecto del valor real en términos relativos. Posteriormente, estos errores porcentuales individuales se promedian a lo largo del horizonte de proyección, obteniéndose una medida global del desempeño predictivo.

Una de las principales ventajas del MAPE es que entrega un resultado expresado en porcentaje, lo que facilita la interpretación y comparación del error entre distintas variables o escenarios de proyección. En particular, esta métrica permite evaluar el desempeño del modelo de forma independiente de la escala de la variable analizada, lo que resulta especialmente útil al trabajar con series temporales ambientales que pueden presentar magnitudes y variabilidad distintas.

En este trabajo, el MAPE se utiliza exclusivamente para evaluar la calidad de las proyecciones generadas por el estimador Random Forest, diferenciándose del MAE, que se emplea en la etapa de calibración del modelo. De este modo, el MAPE permite cuantificar el grado de precisión de las proyecciones y evaluar de qué medida el modelo es capaz de reproducir la evolución observada de las variables estudiadas durante el período de proyección.

Capítulo 3

Metodología y descripción de los datos

3.1. Objetivo metodológico

El objetivo metodológico de este estudio es desarrollar un modelo predictivo confiable y explicativo que permita proyectar, a horizontes de 1 año y 5 años, el comportamiento de tres contaminantes monitoreados en la Planta San Juan: Sólidos Disueltos Totales (SDT), nitratos (NO_3^-) y turbiedad. Este modelo tiene por finalidad apoyar la toma de decisiones del Departamento de Producción de Esval contribuyendo a la gestión operativa, la anticipación de riesgos y la planificación de escenarios futuros.

Para alcanzar este objetivo, se propone una metodología basada en la integración de información histórica, hidrológica y climática, incorporando tanto los contaminantes medidos en la planta como un conjunto de variables externas consideradas relevantes para explicar la dinámica de la calidad del agua. Estas variables incluyen: caudal, temperatura, precipitaciones y nieve acumulada, las cuales permiten capturar procesos ambientales asociados a estacionalidad, escorrentía, eventos climáticos intensos y ciclos de deshielo.

Con el fin de situar geográficamente el área de estudio y comprender la relación entre las distintas fuentes de datos utilizadas, se presenta a continuación un mapa georreferencial donde se indican:

- La ubicación de la Planta San Juan.
- La estación hidrométrica Río Maipo en Cabimbao, donde se obtuvo el caudal.

- Las comunas utilizadas para la extracción de variables climáticas (San Antonio) y de niveles de nieve acumulada (San José de Maipo).

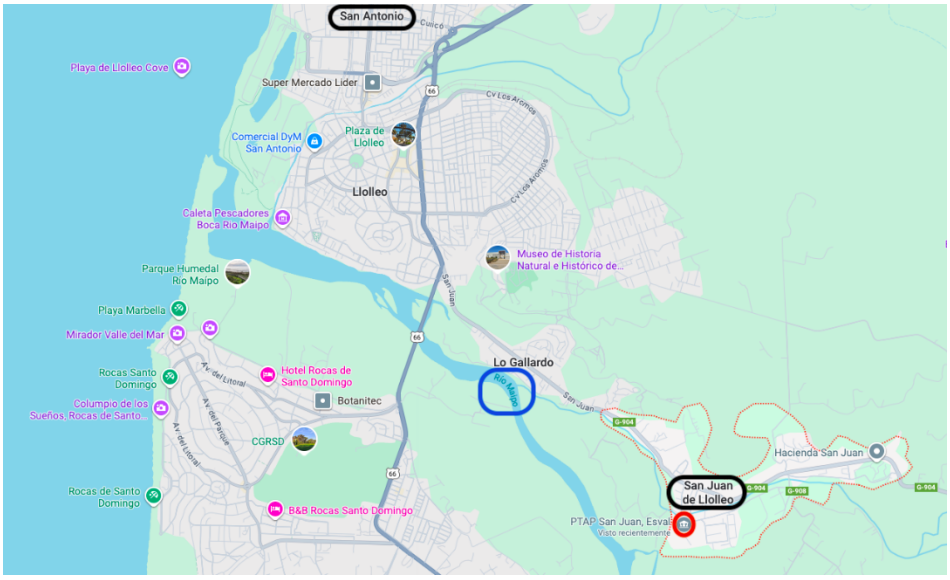

Figura 3.1: Mapa georreferencial del área de estudio.

En la figura 3.1 se aprecia la localización territorial del área de estudio. La Planta San Juan se encuentra ubicada en la ciudad de San Juan de Llolleo, perteneciente a la comuna de San Antonio. En el mapa, el sector urbano de San Juan de Llolleo aparece demarcado mediante una línea segmentada de color naranja, lo que permite identificar el entorno inmediato donde se sitúa la captación de agua cruda operada por Esval. Adicionalmente, la ubicación exacta de la PTAP San Juan se destaca mediante un círculo rojo, lo que facilita reconocer con claridad el punto específico desde el cual se obtienen los datos de calidad del agua utilizados en este estudio.

Asimismo, el mapa permite identificar el curso del río Maipo, principal fuente hídrica asociada a la planta, cuyo nombre se resalta mediante un recuadro azul con el fin de facilitar su reconocimiento dentro del contexto territorial. Por otra parte, los nombres de las localidades de San Antonio y San Juan de Llolleo se destacan mediante recuadros negros, lo que contribuye a contextualizar territorialmente la infraestructura analizada y su relación con el entorno urbano considerado en la modelación.

La metodología general contempla las siguientes etapas principales:

1. Recolección y depuración de datos provenientes de fuentes internas (Esval) y externas (MAPA S2, ARCLIM).

2. Análisis exploratorio de los datos, incluyendo el estudio de correlaciones y la evaluación de rezagos temporales.

3. Selección y justificación de las variables predictoras.

4. Implementación del modelo Random Forest para la estimación de caudal y contaminantes.

5. Evaluación del desempeño del modelo mediante métricas de error en el período histórico (validación).

6. Generación de proyecciones futuras de caudal y posterior estimación de los contaminantes a partir de dichas proyecciones.

Con ello, se establece un proceso metodológico sistemático y reproducible, orientado a la construcción de un modelo predictivo con fundamento estadístico y utilidad operativa.

3.2. Fuentes de datos

La construcción del modelo predictivo requiere integrar información proveniente tanto de registros internos de Esval como de diversas plataformas externas de carácter hidrológico y climático. Dado que la calidad del agua captada por la PTAP San Juan se encuentra influenciada por distintos procesos ambientales, resulta fundamental contar con datos confiables, consistentes y comparables en el tiempo.

En este apartado se describen las fuentes de datos utilizadas y las características principales de cada conjunto de información. La descripción incluye variables internas, correspondientes a los contaminantes monitoreados en la planta San Juan y las variables externas obtenidas desde plataformas, tales como caudal del río Maipo, temperatura del aire, precipitaciones y nieve acumulada.

Cada subsección detalla la plataforma desde donde se obtuvieron los datos, el período disponible y su resolución temporal.

3.2.1. Contaminantes

Los datos correspondientes a los contaminantes analizados en este estudio fueron proporcionados directamente por el Departamento de Producción de Esval, en el marco

del trabajo desarrollado en la PTAP San Juan. Estos registros corresponden a mediciones de agua cruda captada antes de su ingreso al sistema de tratamiento, lo cual permite evaluar la calidad del recurso hídrico en su estado original y comprender las variaciones que afectan al proceso operativo.

Las variables consideradas son:

- Sólidos Disueltos Totales (SDT).
- Nitratos (NO_3^-).
- Turbiedad.

Es importante señalar que los datos entregados por Esval no se encontraban originalmente en formato mensual. Las mediciones se realizaron de manera irregular, existiendo meses con pocas observaciones y otros con una mayor cantidad de registros, dependiendo del contaminante.

Debido a la limitación en la cantidad total de datos disponibles y con el fin de integrar adecuadamente esta información con las demás variables externas, se optó por mensualizar las series de contaminantes. Este procedimiento consistió en agrupar las mediciones por mes y obtener un promedio mensual, garantizando así la coherencia temporal entre todos los conjuntos de datos utilizados en el modelo.

Las series mensuales resultantes abarcan desde enero de 2021 hasta mayo de 2025, constituyendo la base histórica empleada tanto para el análisis exploratorio como para el entrenamiento y validación de los modelos predictivos.

Estas series mensualizadas representan las variables objetivo en los modelos construidos y permiten estudiar la evolución temporal de la calidad del agua cruda en la PTAP San Juan en relación con factores hidrológicos y climáticos.

3.2.2. Caudal

La información de caudal utilizada en este estudio fue obtenida desde la plataforma del Ministerio de Obras Públicas de Chile (https://mapas2.mop.gob.cl). Esta plataforma reúne registros hidrométricos oficiales de diversas estaciones a nivel nacional y constituye una fuente confiable y ampliamente utilizada para estudios ambientales e hidrológicos.

Para el presente análisis se seleccionó la estación Río Maipo en Cabimbao, ya que corresponde al punto de monitoreo más cercano y hidrológicamente representativo del

aporte de agua que posteriormente es captada por la Planta San Juan. El río Maipo constituye la principal fuente que alimenta el sistema, por lo que su caudal influye directamente en la calidad y disponibilidad del recurso hídrico que ingresa a la planta. Adicionalmente, esta estación cuenta con un historial de mediciones extenso y actualizado, lo que garantiza la disponibilidad de información confiable para el análisis y la modelación desarrollada en este estudio.

Los registros descargados corresponden a caudales medios mensuales, con un período disponible que abarca desde enero de 2000 hasta mayo de 2025. Esta amplia serie temporal permite analizar tendencias de largo plazo, variabilidad interanual y efectos asociados a fenómenos climáticos como sequías o eventos extremos.

La elección de esta variable se fundamenta en su relevancia hidrológica, ya que variaciones en el caudal pueden modificar la concentración de contaminantes, generar procesos de dilución o arrastre de sedimentos y afectar directamente la calidad del agua cruda ingresada a la planta. Además, este conjunto de datos se integró de manera mensual con el resto de las variables para construir la base de predictores utilizada en el modelo.

3.2.3. Temperatura y precipitaciones

Los datos de temperatura del aire y precipitaciones utilizados en este estudio fueron obtenidos desde la plataforma oficial ARCLIM (https://arclim.mma.gob.cl), desarrollada por el Ministerio del Medio Ambiente de Chile. Esta plataforma integra información climática histórica y proyectada para todas las comunas del país, y constituye una fuente confiable y ampliamente utilizada en estudios hidrológicos, ambientales y de cambio climático.

En el contexto de este trabajo, se seleccionó la comuna de San Antonio, ya que corresponde al sector geográfico donde se ubica la PTAP San Juan. Esto permite obtener una caracterización climática representativa del entorno de la planta que pueden afectar la calidad del agua cruda, ya sea a través de procesos de escorrentía, arrastre de sedimentos o variabilidad estacional.

Los datos descargados desde ARCLIM corresponden a series mensuales de temperatura del aire y precipitaciones. Si bien las proyecciones climáticas disponibles abarcan un período más extenso, en este estudio se consideraron registros comprendidos entre enero de 2000 y junio de 2030, incluyendo tanto información histórica como proyecciones climáticas coherentes con el horizonte temporal definido para el análisis. Esta

extensión temporal resulta especialmente útil, ya que permite integrar valores futuros directamente en el modelo predictivo sin necesidad de generar proyecciones adicionales para estas variables.

3.2.4. Nieve acumulada

La información correspondiente a la nieve acumulada fue obtenida desde la plataforma ARCLIM (https://arclim.mma.gob.cl), la misma fuente empleada para las variables de temperatura del aire y precipitaciones. Sin embargo, en este caso la selección territorial difiere, ya que se utilizó la comuna de San José de Maipo, ubicada en la zona precordillerana de la Región Metropolitana y caracterizada por su sistema nivoso estacional.

La elección de esta comuna se justifica por motivos hidrológicos, dado que la nieve acumulada en sectores cordilleranos constituye una de las principales fuentes de aporte al río Maipo, especialmente durante los períodos de deshielo. Dado que este río alimenta el sistema desde el cual se abastece la PTAP San Juan, las variaciones en la acumulación y derretimiento de nieve pueden influir indirectamente en el caudal y, por consiguiente, en la calidad del agua cruda que ingresa a la planta.

Los datos descargados corresponden a una serie mensual de nieve acumulada. Si bien las proyecciones climáticas disponibles abarcan un período más extenso, en este estudio se consideraron únicamente los registros comprendidos entre enero de 2000 y junio de 2030, correspondiente a un horizonte de proyección de cinco años a partir de junio de 2025. Esta resolución temporal mensual permite integrar esta variable de manera consistente con el resto de las fuentes utilizadas en el modelo.

3.3. Justificación de variables externas

La incorporación de variables externas en el modelo predictivo responde a la necesidad de capturar los factores ambientales que influyen en la calidad del agua cruda que ingresa a la PTAP San Juan.

En este sentido:

- El caudal del río Maipo refleja la dinámica hidrológica de la cuenca y puede influir en fenómenos de dilución, arrastre de sedimentos y variaciones en la concentración de sólidos y nutrientes.

- La temperatura del aire está asociada a procesos físico-químicos y estacionales que modifican el comportamiento de los compuestos en el agua.
- Las precipitaciones pueden generar escorrentía y aportar un incremento significativo de material particulado o disuelto hacia el agua.
- La nieve acumulada constituye una fuente de aporte diferido al caudal a través del deshielo, afectando tanto los volúmenes como la continuidad temporal del flujo del río.

Estas variables permiten consolidar el análisis al incorporar información exógena relevante para explicar la variación de los contaminantes y contribuyen a un modelo predictivo con mayor capacidad para representar estacionalidad, efectos climáticos y variaciones ambientales relevantes.

3.4. Análisis exploratorio: correlaciones y rezagos

Antes de la construcción del modelo predictivo, se realizó un análisis exploratorio con el fin de examinar la relación entre los contaminantes y las variables externas seleccionadas. Este estudio preliminar permite identificar patrones generales, posibles asociaciones y comportamientos comunes entre las series temporales, proporcionando una primera aproximación a los factores ambientales que pueden influir en la calidad del agua cruda captada por la PTAP San Juan.

En una primera instancia, se elaboró un mapa de calor considerando las correlaciones contemporáneas entre todas las variables, es decir, utilizando los valores correspondientes al mismo instante temporal (tiempo t). Este análisis permite observar asociaciones directas e inmediatas entre los contaminantes (SDT, nitratos y turbiedad) y las variables externas (caudal, temperatura, precipitaciones y nieve acumulada).

Para su correcta interpretación, es importante señalar que el mapa de calor se construye a partir del coeficiente de correlación de Pearson. En este contexto, valores cercanos a 1 indican una relación directa fuerte entre dos variables, es decir, que tienden a aumentar o disminuir en el mismo sentido (cuando una variable aumenta, la otra también tiende a aumentar). Por el contrario, valores cercanos a -1 reflejan una relación inversa fuerte, lo que significa que se mueven en sentido contrario (cuando una variable aumenta, la otra tiende a disminuir). Finalmente, valores próximos a 0 sugieren una asociación débil o inexistente.

Sin embargo, dado que los procesos hidrológicos y climáticos pueden presentar efectos diferidos en el tiempo, como por ejemplo, impactos que aparecen uno o más meses después de ocurrido un evento de lluvia, variaciones en la temperatura o cambios en el caudal, se decidió complementar este análisis con el estudio de correlaciones entre las variables en el tiempo t y sus valores rezagados en $t-1$, $t-2$, $t-3$ y $t-12$. Estos rezagos permiten capturar:

- Relaciones de corto plazo ($t-1$, $t-2$, $t-3$), asociadas a persistencia temporal o efectos acumulativos.
- Relaciones de largo plazo ($t-12$), asociadas a estacionalidad anual o patrones que se repiten cíclicamente.

Para ello, se generaron mapas de calor independientes para cada caso, permitiendo examinar cómo varían las correlaciones cuando se consideran desfases temporales entre las variables. Este análisis resulta especialmente útil, ya que algunas asociaciones que no aparecen en el tiempo contemporáneo pueden volverse más evidentes al considerar rezagos, y viceversa. Además, estos resultados ayudan a justificar la inclusión de ciertos rezagos dentro del conjunto de predictores utilizados posteriormente en el modelo Random Forest.

En conjunto, los mapas de calor contemporáneos y rezagados permiten obtener una visión más completa del comportamiento conjunto de las variables, identificar posibles relaciones diferidas y comprender mejor la dinámica temporal del sistema, todo lo cual constituye una base fundamental para la etapa de modelación predictiva.

A continuación, se presenta el primer mapa de calor correspondiente a las correlaciones contemporáneas (tiempo t–t), donde todas las variables se analizan en el mismo instante temporal.

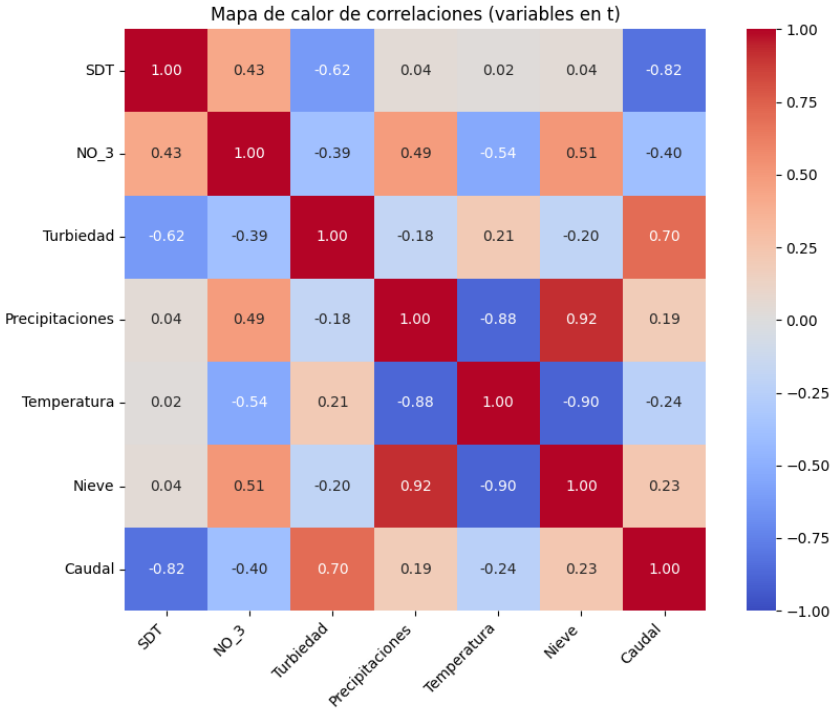

Figura 3.2: Correlaciones entre contaminantes y variables externas (tiempo t).

En la figura 3.2 se presenta el mapa de calor correspondiente al tiempo t, donde permite identificar las asociaciones contemporáneas entre los contaminantes y las variables externas. En primer lugar, se observa que SDT presenta una correlación negativa fuerte con el caudal (-0.82), lo que indica que en períodos de mayor caudal la concentración de sólidos disueltos disminuye, probablemente debido a un efecto de dilución del río. En contraste, la turbiedad muestra una correlación positiva considerable con el caudal (0.70), lo que sugiere que los incrementos en el flujo pueden generar arrastre de sedimentos y material particulado hacia la captación.

Los nitratos presentan correlaciones moderadas con varias variables externas, destacando su relación con precipitaciones (0.49) y nieve acumulada (0.51). Estos valores sugieren que eventos de lluvia o deshielo podrían influir en el aporte de nitratos hacia el sistema. Asimismo, se observa una correlación negativa con la temperatura (-0.54), lo que podría estar asociado a variaciones estacionales o procesos que afectan la disponibilidad de nitratos en distintos períodos del año.

Por otro lado, las variables climáticas muestran entre sí correlaciones muy altas, como precipitaciones con nieve acumulada (0.92) y precipitaciones con temperatura (-0.88), lo cual refleja una estructura climática coherente, marcada por una fuerte es-

tacionalidad. Esto confirma que el comportamiento de estas variables está fuertemente relacionado entre sí.

Finalmente, los contaminantes presentan correlaciones moderadas entre ellos, como SDT–NO_3 (0.43) y turbiedad–SDT (-0.62), lo que indica que, si bien comparten ciertas dinámicas, no están completamente determinados unos por otros. Esto es consistente con su decisión metodológica de no utilizarlos como predictores cruzados dentro de los modelos.

En conjunto, estas correlaciones contemporáneas ofrecen una visión inicial del comportamiento conjunto de las variables y permiten identificar relaciones relevantes que podrían influir en la modelación posterior.

A continuación, se presenta el mapa de calor correspondiente a las correlaciones entre las variables en el tiempo t y sus valores rezagados en $t-1$. Este análisis permite identificar si las condiciones del mes anterior muestran influencia sobre los valores actuales, lo cual es relevante para determinar la presencia de efectos de corto plazo en las series.

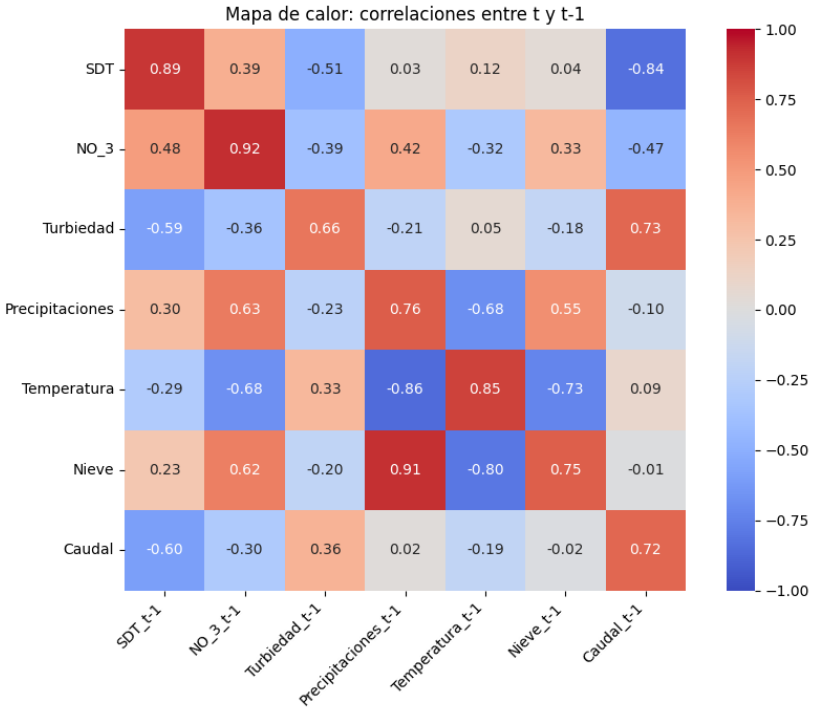

Figura 3.3: Correlaciones entre variables con rezago de un mes (t vs. t–1).

En la figura 3.3 el mapa de correlaciones t–$(t-1)$ evidencia una marcada presencia de persistencia temporal en todas las series, destacando las autocorrelaciones altas

observadas en SDT (0.89), nitratos (0.92), turbiedad (0.66), precipitaciones (0.76), temperatura (0.85), nieve acumulada (0.75) y caudal (0.72). Esto confirma que gran parte del comportamiento actual de estas variables depende de sus valores del mes anterior.

En cuanto a las relaciones entre contaminantes y variables externas, la turbiedad mantiene una correlación positiva considerable con el caudal rezagado (0.73), lo que sugiere que los incrementos en el flujo del río pueden manifestarse en aumentos de turbiedad con un desfase de un mes, posiblemente asociados a procesos de arrastre de sedimentos.

En el caso de los nitratos, el mapa muestra una asociación positiva moderada entre NO_3^-(t) y las precipitaciones del mes anterior (0.42), así como una asociación positiva más débil con la nieve acumulada en t–1 (0.33). Esto sugiere que las condiciones hidrometeorológicas del mes previo podrían relacionarse con la concentración actual de nitratos, aunque con una intensidad moderada. Por otra parte, destaca una correlación negativa moderada entre NO_3^-(t) y el caudal rezagado (–0,47), consistente con un posible efecto de dilución o con cambios en la dinámica de aporte cuando el caudal aumenta.

En conjunto, los resultados muestran que los valores del mes anterior influyen significativamente tanto en los contaminantes como en las variables ambientales, justificando plenamente la incorporación del rezago $t-1$ como predictor dentro del modelo Random Forest.

A continuación, se presenta el mapa de calor correspondiente a las correlaciones entre las variables en el tiempo t y sus valores rezagados en $t-2$. Este análisis permite evaluar si los efectos de ciertas condiciones hidrológicas o climáticas se manifiestan con un desfase de dos meses, lo cual puede ser relevante para identificar patrones de influencia de corto a mediano plazo dentro del sistema.

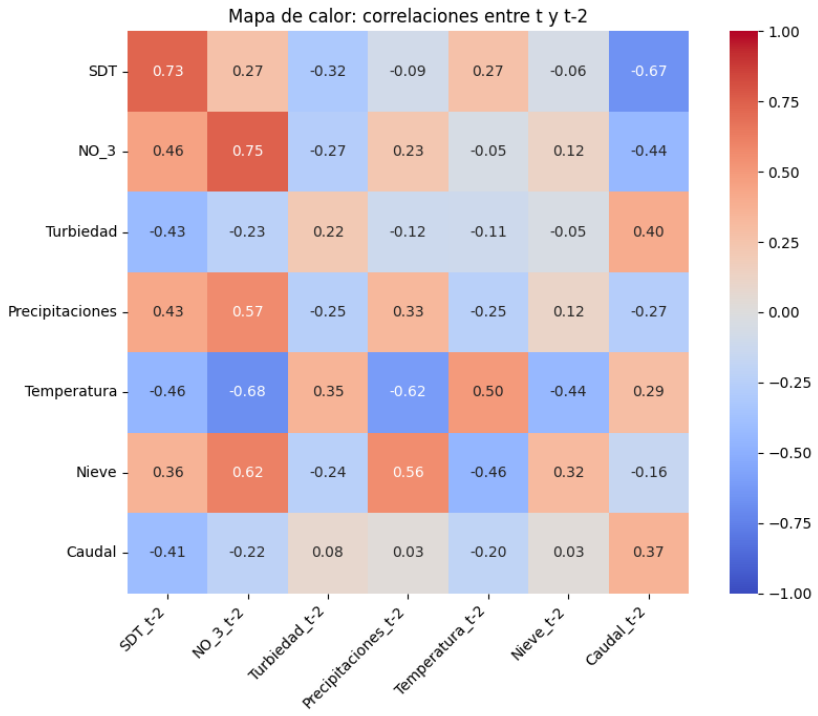

Figura 3.4: Correlaciones entre variables con rezago de dos meses (t vs. t–2)

El mapa de correlaciones t–$(t-2)$ en la figura 3.4 muestra que, aunque la fuerza de asociación disminuye respecto de los rezagos de un mes, aún se observan dependencias temporales relevantes en las series. En particular, los contaminantes mantienen autocorrelaciones moderadas, destacando SDT (0.73), nitratos (0.75) y turbiedad (0.22), lo que indica que sus valores actuales siguen influenciados, en menor medida, por los valores registrados dos meses antes.

En cuanto a las relaciones con variables externas, $NO_3^-(t)$ presenta correlaciones bajas con las precipitaciones rezagadas en dos meses (0.23) y con la nieve acumulada en $t-2$ (0.12), lo que sugiere que, en este desfase, la asociación directa con estas variables es débil. En contraste, se observa una correlación negativa moderada entre $NO_3^-(t)$ y el caudal rezagado en dos meses ($-0{,}44$), consistente con un posible efecto de dilución o con cambios en la dinámica de aporte asociados a mayores caudales.

La turbiedad mantiene una correlación positiva moderada con el caudal rezagado en dos meses (0.40), lo cual refuerza la idea de que episodios de mayor flujo pueden generar efectos de arrastre que persisten más allá del mes inmediatamente posterior.

Finalmente, si bien las correlaciones son más bajas que en el rezago $t-1$, el conjunto de resultados indica que algunas variables externas mantienen influencia sobre

los contaminantes incluso dos meses después, lo que justifica considerar el rezago $t-2$ como predictor potencial dentro del modelo.

A continuación, se presenta el mapa de calor correspondiente a las correlaciones entre las variables en el tiempo t y sus valores rezagados en $t-3$. Este análisis permite evaluar si ciertas relaciones entre los contaminantes y las variables externas se manifiestan con un desfase de tres meses, lo cual resulta útil para detectar influencias de corto a mediano plazo que no son visibles en rezagos más inmediatos.

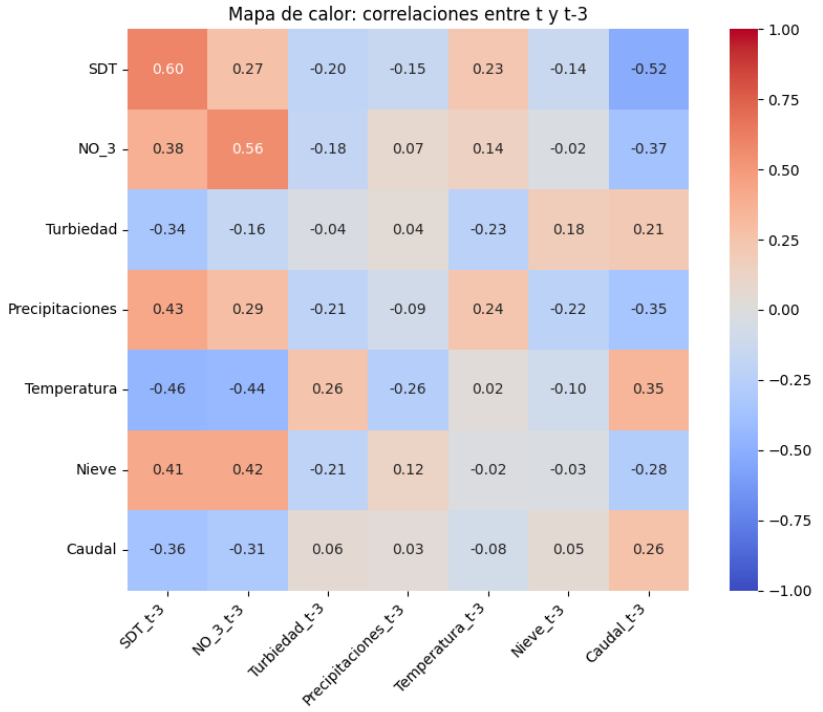

Figura 3.5: Correlaciones entre variables con rezago de tres meses (t vs. t–3).

El mapa de correlaciones t–$(t-3)$ de la figura 3.5 muestra que, aunque las asociaciones disminuyen respecto de los rezagos de uno y dos meses, todavía persisten dependencias temporales relevantes en algunas series. En particular, los contaminantes presentan autocorrelaciones moderadas, destacando SDT (0.60) y nitratos (0.56), lo que indica que sus valores actuales conservan cierta influencia de las mediciones realizadas tres meses antes.

En cuanto a las variables externas, las correlaciones con los contaminantes son en general bajas en este rezago. Para SDT, las asociaciones con precipitaciones y nieve acumulada rezagadas en tres meses presentan valores pequeños y de signo negativo, lo que sugiere que, a este nivel de desfase temporal, no se observa una relación directa

relevante con estas variables climáticas.

La turbiedad muestra correlaciones externas de baja magnitud en este rezago, lo que refuerza la idea de que su respuesta frente a variaciones hidrológicas y climáticas tiende a manifestarse principalmente en horizontes temporales más cortos.

En el caso de los nitratos, además de su autocorrelación moderada, las asociaciones con las variables climáticas rezagadas en tres meses presentan valores cercanos a cero, lo que indica que la influencia directa de las condiciones ambientales se reduce a medida que aumenta el desfase temporal.

En conjunto, este mapa indica que algunas relaciones entre contaminantes y variables externas aún pueden presentar efectos a tres meses, aunque con menor intensidad, lo que confirma la conveniencia de considerar el rezago $t-3$ dentro del conjunto exploratorio de predictores del modelo.

A continuación, se presenta el mapa de calor correspondiente a las correlaciones entre las variables en el tiempo t y sus valores rezagados en $t-12$, es decir, un desfase de un año. Este análisis permite identificar posibles patrones de estacionalidad anual, donde ciertas condiciones ambientales o niveles de contaminantes puedan repetirse cíclicamente con intervalos de doce meses.

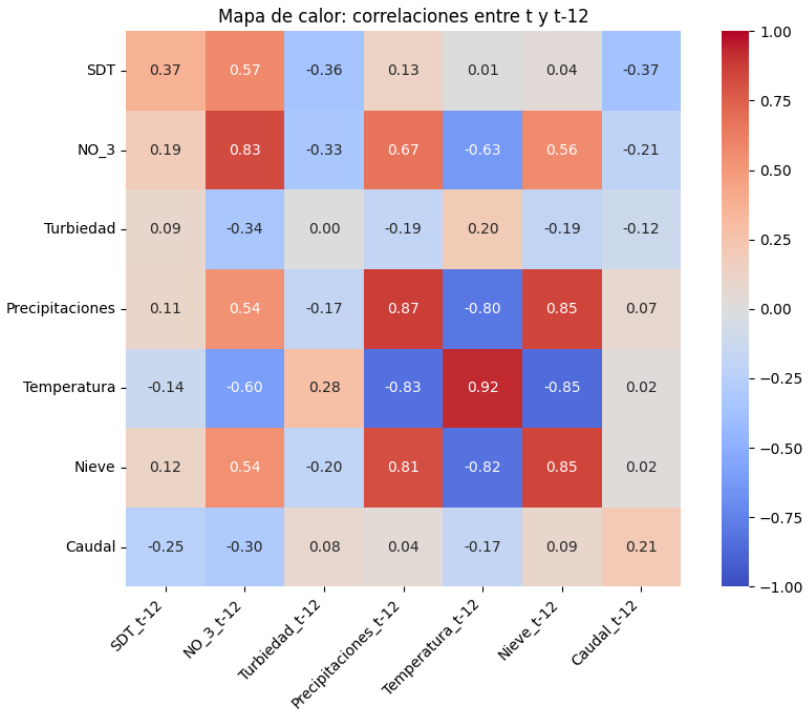

Figura 3.6: Correlaciones entre variables con rezago anual (t vs. t–12).

El mapa de correlaciones t–$(t-12)$ de la figura 3.6 revela que varias variables mantienen asociaciones significativas con sus valores del año anterior, lo que evidencia la presencia de patrones estacionales dentro del sistema. En particular, las variables climáticas muestran correlaciones muy altas entre t y $t-12$, como precipitaciones (0.87), temperatura (0.92) y nieve acumulada (0.85), lo cual es consistente con la marcada estacionalidad del régimen hidrometeorológico de la cuenca.

En cuanto a los contaminantes, se observan autocorrelaciones moderadas en SDT (0.37) y nitratos (0.83), lo que indica la presencia de recurrencia anual, especialmente marcada en NO_3^-. La turbiedad, en cambio, presenta valores muy bajos en este rezago, lo que sugiere que su comportamiento es menos estacional y está más influido por variaciones de corto plazo.

También se observan asociaciones entre NO_3^-(t) y variables climáticas rezagadas en un año, destacando la correlación positiva con precipitaciones $t-12$ (0.67) y con nieve acumulada $t-12$ (0.56), junto con una correlación negativa con la temperatura $t-12$ (-0.63). Estas relaciones reflejan principalmente la estructura estacional del sistema hidrometeorológico, indicando que condiciones típicas de un mismo período anual tienden a repetirse y a asociarse con la variabilidad anual de los nitratos.

En conjunto, el análisis de correlaciones y rezagos permite obtener una visión integral del comportamiento conjunto de los contaminantes y de las variables externas en distintos horizontes temporales. Los mapas de calor mostraron que algunas relaciones se manifiestan de manera inmediata (rezago 0), mientras que otras se fortalecen cuando se consideran desfases temporales, revelando dinámicas propias de corto, mediano y largo plazo. En particular, los rezagos anuales evidenciaron una fuerte estacionalidad en las variables climáticas, mientras que los rezagos de uno a tres meses permitieron identificar dependencias relevantes en SDT, nitratos y turbiedad asociadas a procesos de escorrentía, deshielo o variaciones hidrológicas.

Este diagnóstico preliminar permite identificar los patrones observados que justifican la inclusión de rezagos en los modelos finales y ofrecen fundamentos para interpretar, con mayor claridad, los mecanismos que pueden estar detrás de los cambios en la calidad del agua cruda captada por la PTAP San Juan. De esta forma, el análisis exploratorio constituye un paso esencial para asegurar coherencia y solidez en las etapas posteriores de modelación y proyección.

3.5. Proceso metodológico en el modelo Random Forest

El proceso de modelación desarrollado en este estudio se basa en la aplicación del algoritmo Random Forest Regressor, adaptado a un contexto de series temporales mediante la incorporación explícita de rezagos en los predictores. La decisión de incluir rezagos proviene directamente de los resultados del análisis exploratorio previo, donde se identificaron relaciones relevantes entre las variables en distintos horizontes temporales, especialmente en los rezagos de uno, dos y tres meses, así como en el rezago anual de doce meses. Esta estructura permite capturar dependencias de corto plazo, efectos estacionales y respuestas diferidas que son características de sistemas hidrológicos y ambientales.

Antes de modelar los contaminantes, fue necesario desarrollar una etapa preliminar destinada a proyectar el caudal del río Maipo, ya que esta variable finaliza en mayo de 2025 y se utiliza como predictor en todos los modelos posteriores. Para ello, se entrenó un modelo Random Forest específico para el caudal utilizando su serie histórica completa, disponible desde enero de 2000 hasta mayo de 2025. Este modelo incorporó los mismos rezagos temporales identificados previamente de la variable caudal, junto con los valores contemporáneos y rezagados de las variables externas de temperatura, precipitaciones y nieve acumulada. El resultado es un ajuste histórico continuo del caudal, que reproduce el comportamiento pasado de la serie a partir de los patrones aprendidos por el modelo. Ese ajuste funciona como punto de partida para generar su proyección futura mediante un esquema mes a mes, lo que permite disponer de una serie de caudal proyectada coherente y necesaria para la modelación de los contaminantes.

Una vez obtenida la proyección del caudal, se construyeron los modelos predictivos para sólidos disueltos totales, nitratos y turbiedad, cada uno con su correspondiente conjunto de predictores. En estos casos, el período histórico disponible abarca desde enero de 2021 hasta mayo de 2025. El modelo se entrena utilizando los valores mensualizados de cada contaminante y sus rezagos temporales, junto con las variables externas proyectadas (incluyendo el caudal previamente estimado) y los rezagos asociados a estas. De este modo, cada modelo integra información instantánea y diferida, respetando la estructura temporal evidenciada en la fase exploratoria y permitiendo representar la influencia hidrológica y climática sobre la calidad del agua cruda.

El primer resultado que generan estos modelos es un ajuste histórico para cada contaminante dentro del período observado. Este ajuste permite evaluar si el modelo es capaz de reproducir adecuadamente la dinámica real registrada y, al mismo tiempo, constituye el punto de partida para elaborar las proyecciones futuras. A partir del último dato ajustado, el modelo estima de manera secuencial los valores del mes siguiente, utilizando exclusivamente las predicciones generadas por él mismo y las variables externas proyectadas. Esto asegura consistencia entre el tramo ajustado y el tramo proyectado, dado que toda la serie futura se construye siguiendo el mismo esquema temporal, con los rezagos requeridos en cada paso.

Posteriormente, una vez verificado el ajuste histórico, se generan las proyecciones de 1 y 5 años para cada contaminante. Estas proyecciones se elaboran mediante una actualización mensual, donde cada nuevo valor proyectado sirve para calcular los rezagos del mes siguiente, permitiendo así construir una evolución temporal continua y coherente con la dinámica estimada por el modelo.

En conjunto, la metodología aplicada permite integrar de manera consistente los patrones temporales identificados, la proyección inicial del caudal y la construcción de modelos Random Forest adaptados a series temporales ambientales. Este enfoque combina información histórica, hidrológica y climática para generar proyecciones continuas y coherentes de la calidad del agua cruda captada por la PTAP San Juan, proporcionando una base sólida para el análisis y la interpretación de los resultados obtenidos.

3.6. Validación del modelo

La validación del modelo tiene por objetivo evaluar la capacidad del estimador Random Forest para generar proyecciones consistentes con los valores reales en un período reciente, de modo que las proyecciones de 1 y 5 años se sustenten sobre evidencia cuantitativa. Dado que el enfoque propuesto trabaja con series temporales y utiliza rezagos como predictores, la validación se realizó mediante un esquema de proyección fuera de muestra, comparando valores proyectados con observaciones reales disponibles.

En particular, para cada variable modelada (caudal y contaminantes), el estimador fue entrenado utilizando únicamente la información disponible hasta diciembre de 2024. A partir de ese punto, se generó una proyección secuencial mes a mes para el período enero–mayo de 2025, utilizando en cada paso los rezagos requeridos y las variables externas correspondientes al mismo horizonte temporal. Este procedimiento reproduce

el escenario real de uso del modelo, ya que los valores futuros se construyen de manera iterativa a partir de las predicciones previamente generadas por el propio estimador.

La comparación entre las proyecciones obtenidas y los valores observados en enero–mayo de 2025 se cuantificó mediante el error absoluto medio porcentual (MAPE), el cual permite medir el error en términos relativos y expresarlo como un porcentaje respecto de los valores reales. Considerando un horizonte de validación de $H = 5$ meses, el MAPE se calcula como:

$$\text{MAPE} = \frac{100}{5} \sum_{t=T+1}^{T+5} \frac{|\hat{y}_t - y_t|}{|y_t|} \tag{3.1}$$

donde T corresponde al último mes del período de entrenamiento (diciembre de 2024), y_t es el valor observado e $\hat{y}_t$ la proyección entregada por el modelo en el mes t.De este modo, el MAPE entrega un indicador global del error relativo obtenido en el tramo proyectado, permitiendo evaluar la calidad de la proyección aun cuando las variables analizadas presentan escalas distintas.

De manera complementaria, se consideró la comparación gráfica entre valores observados y valores ajustados dentro del período de entrenamiento, lo que permite verificar visualmente la coherencia del estimador al reproducir patrones de tendencia, variabilidad y estacionalidad presentes en los datos históricos.

Es importante distinguir que, si bien el MAPE se utiliza como métrica principal para validar el desempeño de las proyecciones en el período enero–mayo de 2025, el error absoluto medio (MAE) se emplea en una etapa diferente del proceso metodológico, correspondiente a la calibración del modelo. En particular, el MAE se utiliza para analizar cómo varía el error al modificar el número de árboles del bosque y el número de variables predictoras incluidas en el estimador, identificando configuraciones en las que el error se estabiliza. Este criterio permite seleccionar un conjunto de parámetros eficiente, evitando complejidad innecesaria y controlando el costo computacional, sin afectar de manera significativa el desempeño predictivo.

En conjunto, la validación realizada mediante MAPE entrega evidencia cuantitativa sobre la capacidad del modelo para anticipar la evolución reciente de las series en un escenario de proyección realista. Por su parte, la calibración basada en MAE permite justificar la configuración final del estimador utilizada en los análisis posteriores. Con estos elementos, se establece una base metodológica consistente para presentar e interpretar las proyecciones a 1 y 5 años desarrolladas en la sección siguiente.

Finalmente, con el fin de asegurar la reproducibilidad del proceso metodológico, en los anexos se incluye el código en Python utilizado en las distintas etapas del estudio. En particular, el Anexo A presenta el código empleado para el análisis exploratorio de correlaciones y la generación de mapas de calor, orientado a examinar la relación entre los contaminantes y las variables externas, así como la identificación de rezagos temporales relevantes. Por otra parte, el Anexo B contiene el código correspondiente a la implementación del modelo Random Forest, incluyendo la selección del número de variables predictoras y de árboles, el análisis de importancia por reducción de impureza y la generación de proyecciones futuras. Esta organización permite distinguir claramente entre la etapa exploratoria y la etapa de modelación predictiva, facilitando la comprensión y replicabilidad del estudio.

Capítulo 4

Resultados para PTAP San Juan

En este capítulo se presentan los resultados obtenidos a partir de la aplicación del modelo Random Forest al sistema de captación San Juan. El análisis se enfoca en evaluar el desempeño del modelo para la estimación y proyección del caudal, así como de los principales parámetros de calidad del agua considerados en este estudio, los cuales son sólidos disueltos totales (SDT), nitratos (NO_3^-)y turbiedad.

Para cada variable se sigue una estructura común que permite analizar de manera sistemática el comportamiento del modelo. En primer lugar, se aborda la selección del número óptimo de variables predictoras, con el objetivo de identificar configuraciones que equilibren capacidad predictiva y simplicidad del modelo. Esta selección se realiza a partir del análisis del error absoluto medio (MAE), el cual permite evaluar el desempeño del modelo al incorporar progresivamente distintos conjuntos de variables explicativas. Posteriormente, se presenta la selección del número óptimo de árboles del bosque, evaluando su influencia en el desempeño del modelo y en la estabilidad de las predicciones, nuevamente utilizando el MAE como criterio de comparación entre las distintas configuraciones del modelo.

A continuación, se analiza la importancia relativa de las variables seleccionadas, lo que permite interpretar el rol de los distintos predictores en la estimación de cada variable de interés, incluyendo tanto variables externas de carácter climático como rezagos temporales de las propias series.

Finalmente, se presentan las proyecciones generadas por el modelo Random Forest, considerando horizontes de uno y cinco años, lo que permite analizar la evolución esperada de cada variable en distintos plazos temporales. Para cada caso, se incorpora además una etapa de validación de las proyecciones, en la cual se compara el valor pro-

yectado con el valor observado utilizando el error porcentual absoluto medio (MAPE), con el fin de evaluar el desempeño del modelo en términos relativos y cuantificar su capacidad predictiva sobre datos no utilizados en el entrenamiento.

El capítulo concluye con una comparación general de los resultados obtenidos para las distintas variables analizadas, con el fin de identificar patrones comunes, diferencias en el comportamiento predictivo y el aporte del modelo Random Forest en el análisis del comportamiento hidrológico y de la calidad del agua de la PTAP San Juan.

4.1. Resultados para caudal

En esta sección se presentan los resultados obtenidos para la modelación del caudal en la PTAP San Juan utilizando el modelo Random Forest. Dado el rol central del caudal en la disponibilidad del recurso hídrico y su influencia sobre la calidad del agua, se analizan las configuraciones del modelo que permiten representar de mejor manera su comportamiento temporal.

De manera preliminar, se presenta el criterio de selección y ordenamiento de las variables predictoras, con el fin de establecer una referencia inicial sobre la relevancia relativa de los predictores considerados. Este criterio constituye la base para la construcción de los distintos modelos evaluados en esta sección.

Los resultados incluyen el análisis de las variables predictoras consideradas, la interpretación de su importancia relativa, la evaluación del número de árboles utilizados en el modelo y las proyecciones de caudal a distintos horizontes temporales, con el objetivo de caracterizar la evolución esperada del caudal en la PTAP San Juan.

4.1.1. Criterio de selección y ordenamiento de variables predictoras

Antes de analizar el desempeño del modelo en función del número de predictores, es importante describir cómo se construyen los subconjuntos de variables utilizados. En este capítulo, se utilizará la notación $RF\text{-}k$ para referirse a un modelo Random Forest entrenado con k variables predictoras.

En este estudio, el ordenamiento de las variables predictoras se obtiene a partir de la estimación de su importancia relativa calculada por el modelo Random Forest durante el proceso de entrenamiento, utilizando como criterio la reducción promedio de impureza.

En primer lugar, se entrena un modelo utilizando el conjunto completo de predictores disponibles para caudal, compuesto por 19 variables (rezagos del caudal y variables climáticas externas con distintos rezagos). A partir de este modelo se calcula la importancia de cada predictor en función de la reducción de impureza generada a lo largo de los árboles que componen el bosque, obteniendo así un ranking de relevancia de las variables dentro del ensamble. La figura 4.1 presenta dicho ranking para el modelo entrenado con las 19 variables (modelo RF-19).

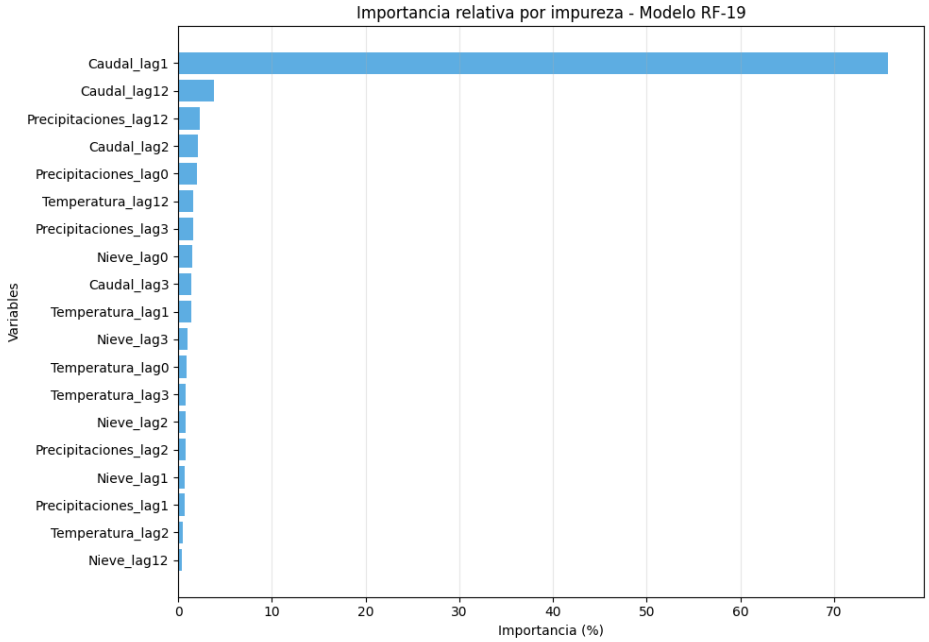

Figura 4.1: Importancia relativa por impureza para el modelo de caudal utilizando las 19 variables predictoras disponibles (modelo RF-19).

A partir de este ranking, para construir un modelo con k variables, se seleccionan las k variables con mayor importancia relativa y se entrena nuevamente un Random Forest utilizando únicamente dicho subconjunto. Por ejemplo, un modelo RF-8 corresponde a un Random Forest entrenado con las 8 variables que presentan mayor contribución en términos de reducción de impureza según este criterio.

No obstante, al modificarse el número de variables consideradas, el modelo es reentrenado. En este proceso, el subconjunto de k variables seleccionadas corresponde siempre a las variables de mayor importancia identificadas a partir del conjunto completo de 19 predictores. Sin embargo, dado que el modelo se ajusta nuevamente al utilizar un subconjunto reducido, el orden relativo de estas variables puede variar entre

distintas configuraciones *RF-k*. En consecuencia, el ranking mostrado en la figura 4.1 debe interpretarse como una referencia asociada al modelo con 19 variables, y no como un orden fijo e invariante.

4.1.2. Selección del número óptimo de variables predictoras

Con el objetivo de determinar una configuración adecuada del modelo Random Forest para la estimación de caudal, se analizó el comportamiento del error absoluto medio (MAE) al variar el número de variables predictoras incluidas en el modelo. Este análisis se realizó considerando distintas configuraciones del número de árboles del bosque, específicamente 50, 100, 200 y 500 árboles, lo que permite evaluar de manera conjunta la influencia del número de predictores y del tamaño del ensamble en el desempeño del modelo.

En el caso del caudal, el conjunto total de variables predictoras disponibles estuvo compuesto por 19 variables. Estas incluyen rezagos temporales del propio caudal en los instantes $t-1$, $t-2$, $t-3$ y $t-12$, junto con variables climáticas externas correspondientes a temperatura, precipitaciones y nieve acumulada, consideradas en el tiempo contemporáneo t como en los rezagos $t-1$, $t-2$, $t-3$ y $t-12$. De este modo, el conjunto de predictores incorpora información hidrológica y climática relevante, así como dependencias temporales de corto y mediano plazo.

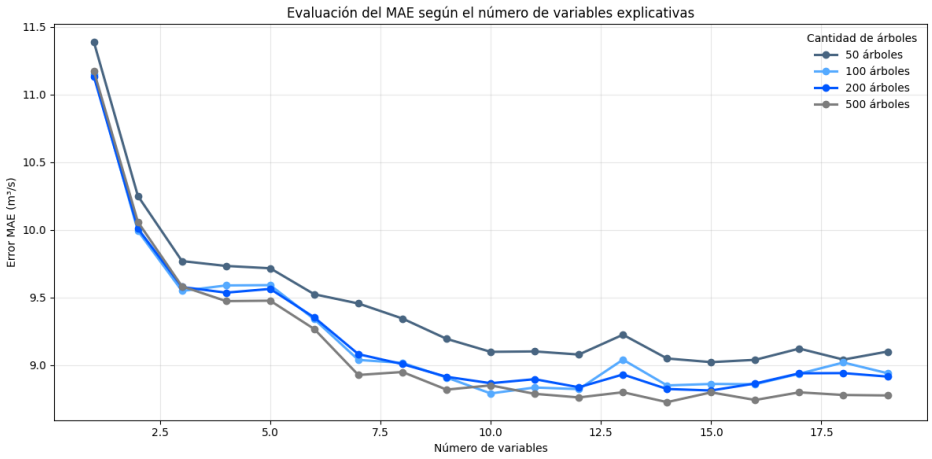

Figura 4.2: Evolución del MAE en función del número de variables predictoras para el caudal.

En la figura 4.2 se presenta la evolución del MAE en función del número de variables explicativas para cada una de las configuraciones de árboles analizadas. En términos

generales, se observa una disminución significativa del MAE al aumentar el número de variables desde valores bajos, lo que indica que la incorporación progresiva de información relevante mejora la capacidad del modelo para representar la dinámica del caudal. Sin embargo, esta reducción del error se vuelve progresivamente menor a medida que se incorporan más variables.

En particular, a partir de aproximadamente 8 variables predictoras, el MAE tiende a estabilizarse para los casos de 100, 200 y 500 árboles, mostrando valores muy similares entre estas configuraciones. Esto sugiere que, desde ese punto, la inclusión de variables adicionales no produce mejoras relevantes en el desempeño del modelo. En el caso de 50 árboles, el MAE se mantiene levemente por sobre las demás configuraciones; no obstante, la diferencia observada es pequeña y no modifica de gran manera el comportamiento general del error.

Considerando estos resultados, se optó por utilizar un conjunto de 8 variables predictoras para el modelo de caudal. Esta selección permite obtener un desempeño predictivo estable y consistente para configuraciones con 100, 200 y 500 árboles, evitando al mismo tiempo la incorporación de predictores adicionales que no aportan mejoras significativas. De este modo, se logra una configuración que equilibra adecuadamente precisión y simplicidad, y que constituye una base sólida para los análisis posteriores.

4.1.3. Selección del número óptimo de árboles

Este análisis tiene como objetivo identificar un número de árboles que permita obtener predicciones estables y precisas, evitando al mismo tiempo un aumento innecesario en el costo computacional. Para ello, se consideró un modelo Random Forest construido a partir del conjunto de ocho variables predictoras seleccionado previamente, el cual se denota en adelante como modelo $RF - 8$.

Para ello, se entrenaron modelos considerando distintos números de árboles, específicamente 10, 20, 50, 100, 200, 300, 400, 500, 600 y 1000 árboles, manteniendo fijo el conjunto de ocho variables predictoras seleccionado previamente. En cada configuración se evaluó el error absoluto medio (MAE), lo que permitió comparar de manera directa el efecto del tamaño del ensamble sobre el desempeño del modelo.

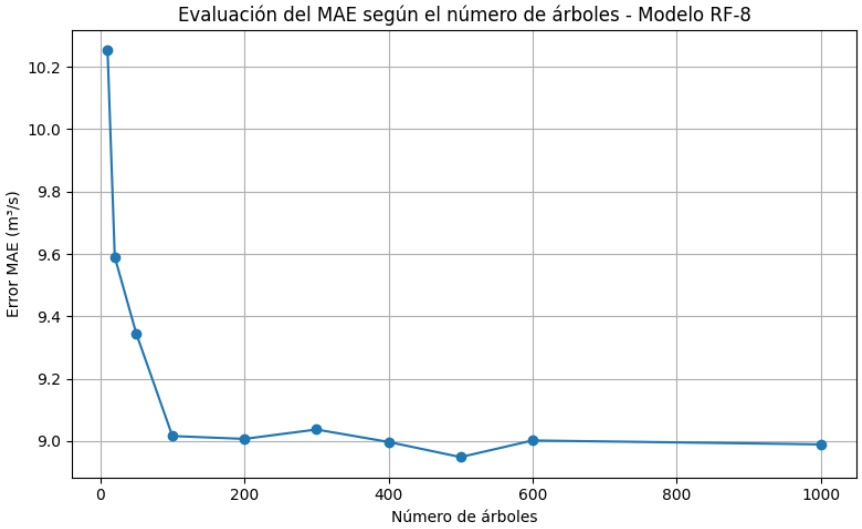

Figura 4.3: Comportamiento del MAE en función del número de árboles para el caudal.

En la figura 4.3 se presenta la evaluación del error absoluto medio (MAE) en función del número de árboles del modelo, considerando la configuración RF-8 (Random Forest con 8 variables predictoras). Se observa que el MAE disminuye de forma marcada al aumentar el número de árboles desde valores bajos, particularmente entre 10 y 100 árboles. Este comportamiento refleja una mejora progresiva en la estabilidad del modelo a medida que se incorporan más árboles al bosque. Sin embargo, a partir de aproximadamente 200 árboles, el MAE comienza a estabilizarse, mostrando variaciones leves y sin una tendencia clara a una reducción adicional del error.

Este patrón se mantiene al considerar configuraciones con un mayor número de árboles, como 300, 400, 500, 600 o 1000, lo que indica que incrementar el tamaño del ensamble más allá de ese punto no produce mejoras significativas en el desempeño predictivo del modelo.

En base a estos resultados, se optó por utilizar un total de 200 árboles en el modelo de caudal. Esta elección permite alcanzar un MAE bajo y estable, representativo del desempeño óptimo del modelo, sin generar un aumento innecesario del costo computacional asociado a un número mayor de árboles. De este modo, la configuración seleccionada equilibra adecuadamente precisión y eficiencia, y constituye la base para la generación de las proyecciones de caudal presentadas en los apartados siguientes.

4.1.4. Importancia de las variables

Una vez definida la configuración del modelo de caudal, correspondiente a un Random Forest entrenado con ocho variables predictoras y 200 árboles (modelo RF-8), se analizó la importancia relativa de cada una de ellas con el objetivo de identificar cuáles contribuyen en mayor medida a explicar la variabilidad del caudal estimado. La importancia de las variables se calcula a partir de la reducción de impureza generada por cada predictor a lo largo de los árboles que componen el bosque.

Tal como se describió en el apartado 4.1.1. el criterio de selección y ordenamiento de variables predictoras, el conjunto de ocho variables utilizadas en el modelo RF-8 corresponde a las ocho variables de mayor importancia identificadas a partir del conjunto completo de 19 predictores. No obstante, dado que el modelo se re-entrena al trabajar con este subconjunto reducido, el orden relativo de las variables puede diferir respecto del ranking obtenido para el modelo con 19 variables.

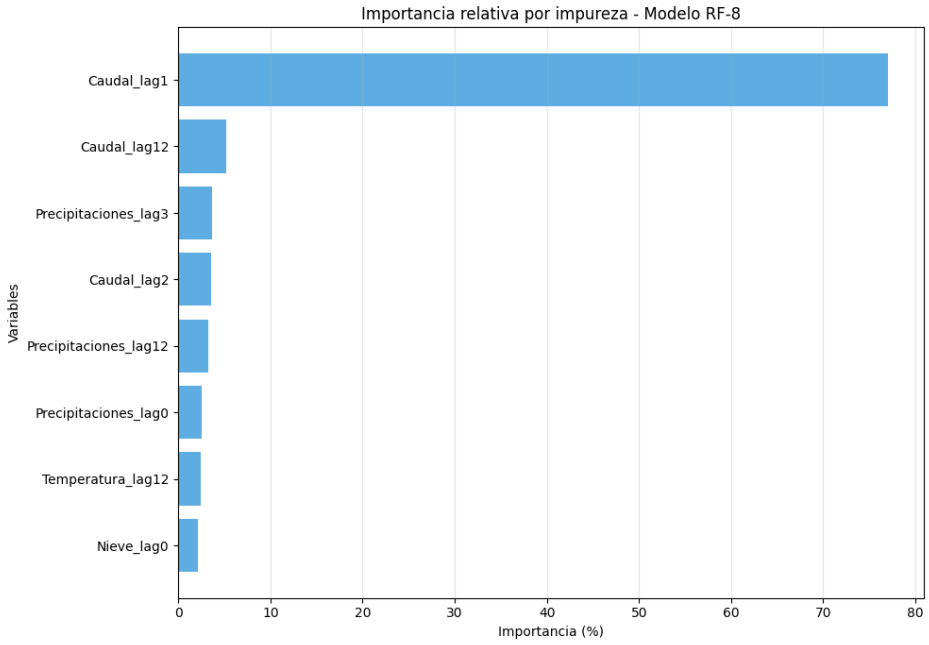

Figura 4.4: Importancia relativa de las variables predictoras en la estimación del caudal.

La figura 4.4 presenta la importancia relativa de las variables consideradas en el modelo RF-8. Se observa de manera clara que el rezago del caudal en el tiempo $t-1$ es la variable dominante, concentrando una proporción significativamente mayor de la importancia total en comparación con el resto de los predictores. Este resultado evidencia una fuerte dependencia temporal de corto plazo en la serie de caudal, donde

el valor del mes inmediatamente anterior aporta la mayor parte de la información para la estimación del caudal actual.

En segundo lugar, aparecen otros rezagos del propio caudal, en particular en los tiempos $t-12$ y $t-2$, los cuales presentan una contribución menor pero relevante. La presencia del rezago anual sugiere la existencia de patrones estacionales en el comportamiento del caudal, mientras que el rezago de corto plazo refuerza la persistencia temporal de la serie.

Las variables climáticas externas, correspondientes a precipitaciones, temperatura y nieve acumulada en distintos rezagos, presentan una importancia relativa considerablemente menor en comparación con los rezagos del caudal. No obstante, su inclusión resulta pertinente, ya que permite capturar efectos hidrológicos que influyen indirectamente en la dinámica del caudal.

En conjunto, la estructura de importancias obtenida indica que el caudal en la PTAP San Juan está fuertemente determinado por su comportamiento pasado, complementado por la influencia de variables climáticas externas. Estos resultados son coherentes con la naturaleza hidrológica del sistema y confirman que el modelo Random Forest identifica de manera adecuada los principales factores que explican la variabilidad del caudal, respaldando la selección final de variables utilizada en el modelo RF-8.

4.1.5. Proyecciones de caudal

A continuación, se presentan las proyecciones de caudal obtenidas mediante el modelo Random Forest definido previamente, el cual considera 8 variables predictoras y un total de 200 árboles. Esta configuración fue seleccionada por ofrecer un buen equilibrio entre desempeño predictivo y estabilidad del modelo. Las proyecciones se realizan a distintos horizontes temporales, permitiendo analizar la evolución esperada del caudal en el corto y mediano plazo. En particular, se estudian proyecciones a uno y cinco años, las cuales resultan relevantes para apoyar el análisis y la planificación operativa asociada al comportamiento del caudal en la PTAP San Juan.

Proyecciones a 1 año de caudal

A continuación, se presenta la proyección de caudal a un año obtenida mediante el modelo Random Forest seleccionado para este estudio, con el fin de analizar la evolución esperada del caudal en el corto plazo. Para el caso del caudal, el período histórico disponible abarca desde enero del año 2000 hasta mayo de 2025, mientras que la proyección

a un año se extiende desde junio de 2025 hasta junio de 2026.

Con el objetivo de facilitar la visualización de la transición entre el período histórico y el período proyectado, en la figura se muestra únicamente un tramo reciente de la serie histórica, correspondiente al período comprendido entre junio de 2023 y mayo de 2025, a partir del cual se inicia la proyección. Esta decisión tiene un carácter exclusivamente gráfico y no implica una reducción del conjunto de datos utilizados en el entrenamiento del modelo, el cual considera la totalidad del período histórico disponible.

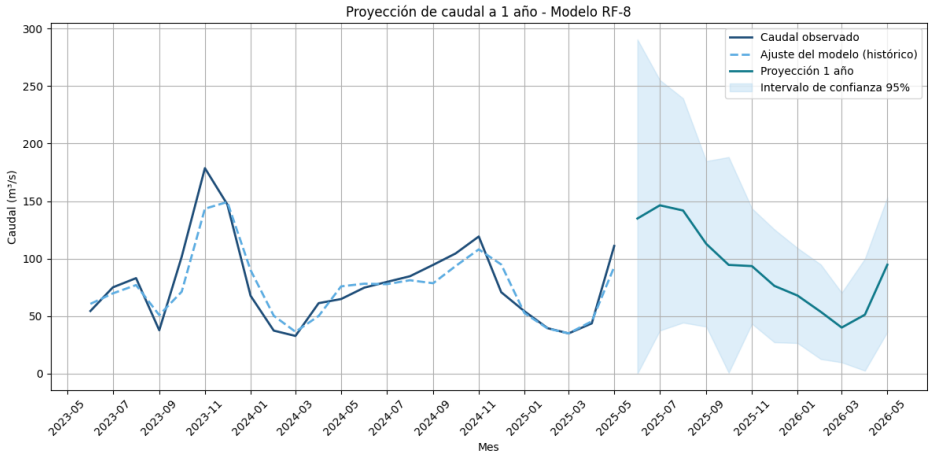

Figura 4.5: Ajuste histórico y proyección a 1 año del caudal.

En la figura 4.5 se observa el comportamiento del caudal histórico reciente, correspondiente al período desde el año 2023, junto con la proyección a un año generada por el modelo Random Forest. Es importante destacar que la visualización del período histórico ha sido acotada con fines exclusivamente gráficos, con el objetivo de resaltar la continuidad entre los valores observados más recientes y el tramo proyectado.

El modelo reproduce de manera adecuada la dinámica del caudal en el período histórico mostrado y entrega una proyección coherente con los valores recientes. En el tramo proyectado se aprecia un comportamiento estacional, caracterizado por una disminución del caudal durante los meses de verano y un aumento progresivo hacia el período de primavera, en concordancia con el comportamiento esperado del sistema hidrológico.

El intervalo de confianza asociado a la proyección se amplía progresivamente hacia el horizonte futuro, reflejando la incertidumbre a las estimaciones. En el contexto del modelo Random Forest, dichos intervalos se construyen a partir de la variabilidad de las predicciones generadas por los distintos árboles que componen el bosque, manteniéndose

dentro de rangos consistentes con la variabilidad observada en el período histórico reciente.

En conjunto, la proyección a un año entrega una estimación razonable del comportamiento esperado del caudal, constituyendo una base adecuada para el análisis de escenarios de corto plazo en la PTAP San Juan.

Proyecciones a 5 años de caudal

A continuación, se presenta la proyección de caudal a cinco años obtenida mediante el modelo Random Forest seleccionado. A diferencia del caso anterior, en esta figura se incluye la totalidad del período histórico disponible, lo que permite contextualizar las proyecciones dentro de la evolución de mediano plazo del caudal y evaluar su coherencia respecto de los patrones observados históricamente.

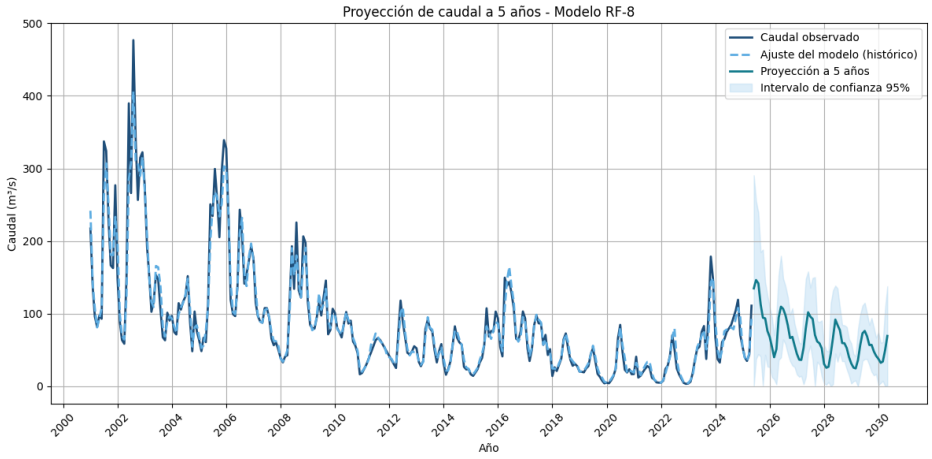

Figura 4.6: Ajuste histórico y proyección a 5 años del caudal.

En la figura 4.6 se observa el ajuste histórico del modelo sobre la serie completa de caudal, junto con la proyección a cinco años generada a partir del último valor ajustado. La inclusión de todo el período histórico permite apreciar la capacidad del modelo para reproducir distintas fases del comportamiento del caudal, incluyendo períodos de alta variabilidad en los primeros años de la serie y etapas posteriores con menores niveles y oscilaciones más acotadas.

La proyección a cinco años mantiene una dinámica coherente con la estructura temporal observada en la serie histórica, reproduciendo oscilaciones estacionales y niveles acordes al comportamiento reciente del caudal. El intervalo de confianza asociado a la

proyección refleja el incremento natural de la incertidumbre en proyecciones de mediano plazo y la mayor dispersión esperada en los valores estimados.

En conjunto, estos resultados indican que el modelo Random Forest logra generar proyecciones de caudal que se integran de manera coherente con la evolución histórica del sistema, proporcionando una base sólida para el análisis de escenarios futuros y su posterior utilización como variable predictora en la modelación de los contaminantes.

4.1.6. Validación del modelo

Con el objetivo de evaluar el desempeño predictivo del modelo Random Forest en la estimación del caudal en la PTAP San Juan, se llevó a cabo una etapa de validación basada en el uso del error porcentual absoluto medio (MAPE), definido en la ecuación 3.1. Esta métrica permite cuantificar el error relativo entre valores proyectados y valores observados, expresándolo en términos porcentuales. En este estudio, el MAPE se calculó considerando el período de validación comprendido entre enero y mayo de 2025, comparando las proyecciones generadas por el modelo con los valores reales observados en dicho intervalo.

La validación se abordó desde dos perspectivas complementarias. En primer lugar, se analizó el comportamiento del MAPE en función del número de variables predictoras consideradas, con el fin de evaluar cómo la complejidad del modelo influye en su desempeño predictivo. Posteriormente, se realizó una validación directa de la proyección de caudal generada por el modelo seleccionado, comparando los valores proyectados con los valores observados para un período de validación específico.

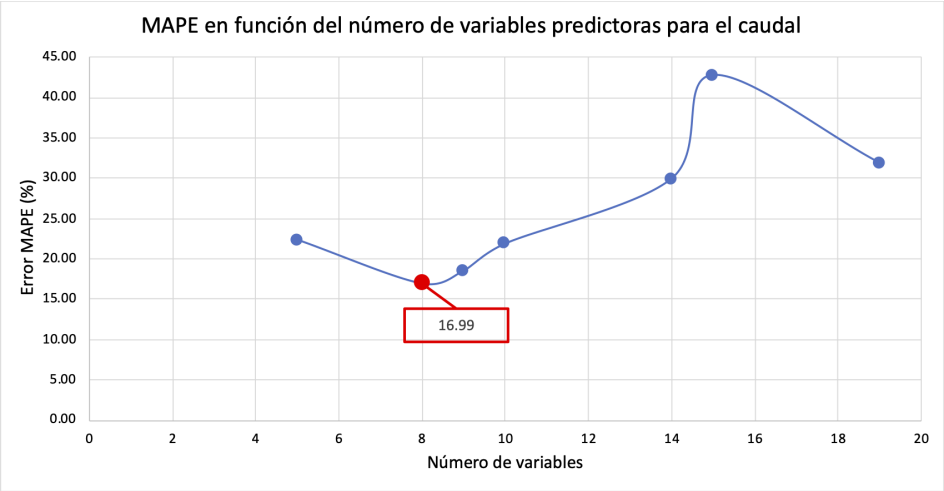

Figura 4.7: Comportamiento del MAPE para distintas configuraciones del modelo de caudal.

En primer lugar, la validación se empleó como criterio para analizar el efecto del número de variables predictoras sobre el desempeño del modelo. La figura 4.7 presenta la evolución del MAPE en función del número de variables consideradas, evaluándose específicamente las configuraciones de 5, 8, 9, 10, 14, 15 y 19 variables predictoras. Estas configuraciones fueron seleccionadas a partir de los análisis previos de selección del número óptimo de variables realizados mediante el error absoluto medio (MAE), los cuales permitieron identificar subconjuntos representativos del comportamiento del error al variar la complejidad del modelo.

Adicionalmente, la elección de estas configuraciones se realizó considerando el rol central del caudal dentro del proceso de modelación de los parámetros de calidad del agua abordados en este estudio. En particular, el caudal constituye una de las principales variables explicativas en las proyecciones de los contaminantes analizados en los apartados siguientes, por lo que se consideró relevante evaluar el desempeño del modelo de caudal bajo distintas configuraciones representativas de complejidad, asegurando así la coherencia y consistencia del enfoque predictivo.

A partir de la figura 4.7, se observa que el error porcentual absoluto medio alcanza su valor mínimo al considerar un conjunto de ocho variables predictoras. En particular, el modelo Random Forest con ocho variables (RF-8) presenta un valor de MAPE igual a 16,99 %, correspondiente al menor error obtenido entre todas las configuraciones evaluadas. Al considerar un número mayor de variables predictoras, el MAPE aumenta, lo que indica que un incremento en la complejidad del modelo no se traduce en una

mejora del desempeño predictivo y, por el contrario, puede afectar negativamente su capacidad de generalización.

En segundo lugar, una vez definida la configuración final del modelo, se procedió a validar directamente la proyección de caudal generada por el modelo Random Forest con ocho variables predictoras. Para ello, el modelo fue entrenado considerando información disponible hasta diciembre de 2024, y posteriormente se generó una proyección para el período comprendido entre enero y mayo de 2025. La figura 4.8 muestra la comparación entre los valores proyectados y los valores observados correspondientes a dicho período de validación.

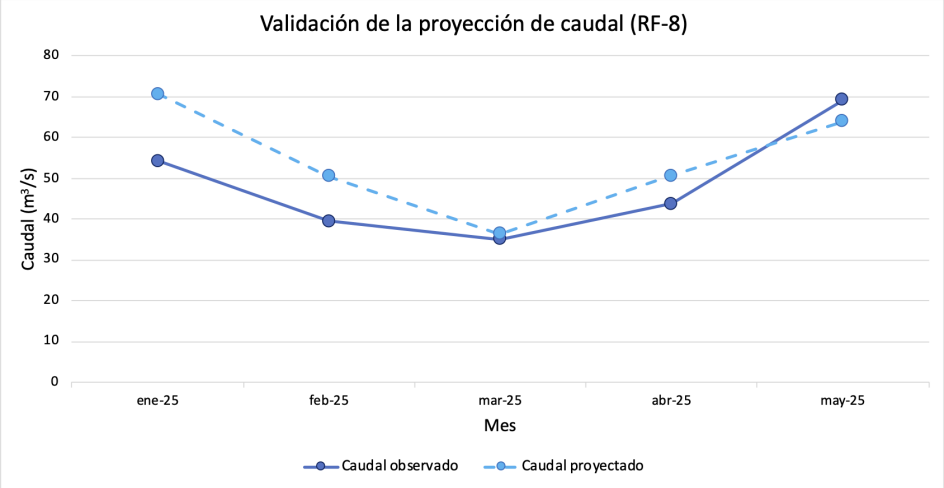

Figura 4.8: Comparación entre valores observados y proyectados del caudal (enero–mayo 2025).

A partir de esta comparación, se observa que el modelo reproduce adecuadamente la tendencia temporal del caudal observado, capturando el descenso hacia el mes de marzo y el posterior aumento hacia el mes de mayo. Si bien se presentan diferencias puntuales entre valores proyectados y reales, el modelo logra representar correctamente el orden de magnitud y la dinámica general del caudal durante el período analizado.

En conjunto, los resultados de validación indican que, tanto desde el punto de vista del MAE utilizado en las etapas de selección del modelo como del MAPE empleado en la validación de las proyecciones, la configuración RF-8 constituye la alternativa más adecuada para la modelación del caudal en la PTAP San Juan. Esto confirma que el uso de ocho variables predictoras ofrece el mejor equilibrio entre precisión predictiva y complejidad del modelo, respaldando su utilización para el análisis de escenarios futuros y su posterior incorporación como variable predictora en la modelación de los

parámetros de calidad del agua considerados en este estudio.

4.2. Resultados para sólidos disueltos totales (SDT)

En esta sección se presentan los resultados obtenidos para la modelación de los sólidos disueltos totales (SDT) en la PTAP San Juan mediante el modelo Random Forest. Los SDT constituyen un parámetro relevante de la calidad del agua, ya que reflejan la concentración de sales y minerales disueltos y su comportamiento se encuentra estrechamente vinculado a la dinámica hidrológica del sistema. Siguiendo el mismo enfoque metodológico utilizado para el caudal, se analizan distintas configuraciones del modelo con el fin de caracterizar el comportamiento de los SDT, evaluar el desempeño predictivo del modelo y generar proyecciones a uno y cinco años, junto con su correspondiente validación.

4.2.1. Selección del número óptimo de variables predictoras

Siguiendo el mismo procedimiento de selección y ordenamiento de variables predictoras descrito en la Sección 4.1.1 para el caso del caudal, el análisis para los sólidos disueltos totales se desarrolló de manera análoga, utilizando rankings de importancia obtenidos a partir de la reducción promedio de impureza estimada por el modelo Random Forest.

Con el objetivo de determinar una configuración adecuada del modelo Random Forest para la estimación de los sólidos disueltos totales (SDT), se analizó el comportamiento del error absoluto medio (MAE) al variar el número de variables predictoras incluidas en el modelo. Este análisis se realizó considerando distintas configuraciones del número de árboles del bosque, específicamente 50, 100, 200 y 500 árboles, lo que permitió evaluar de manera conjunta la influencia del número de variables explicativas y del tamaño del ensamble en el desempeño del modelo.

Para el caso de los SDT, el conjunto total de variables predictoras disponibles estuvo compuesto por 24 variables. Estas incluyen rezagos temporales de los propios SDT en los instantes $t-1$, $t-2$, $t-3$ y $t-12$, así como variables externas correspondientes a caudal, temperatura, precipitaciones y nieve acumulada, consideradas tanto en el tiempo contemporáneo como en los rezagos $t-1$, $t-2$, $t-3$ y $t-12$. De este modo, el conjunto de predictores incorpora información asociada al comportamiento histórico de los SDT y a la dinámica hidrológica y climática del sistema.

En la figura 4.9 se presenta la evolución del MAE en función del número de variables predictoras para las distintas configuraciones de árboles analizadas.

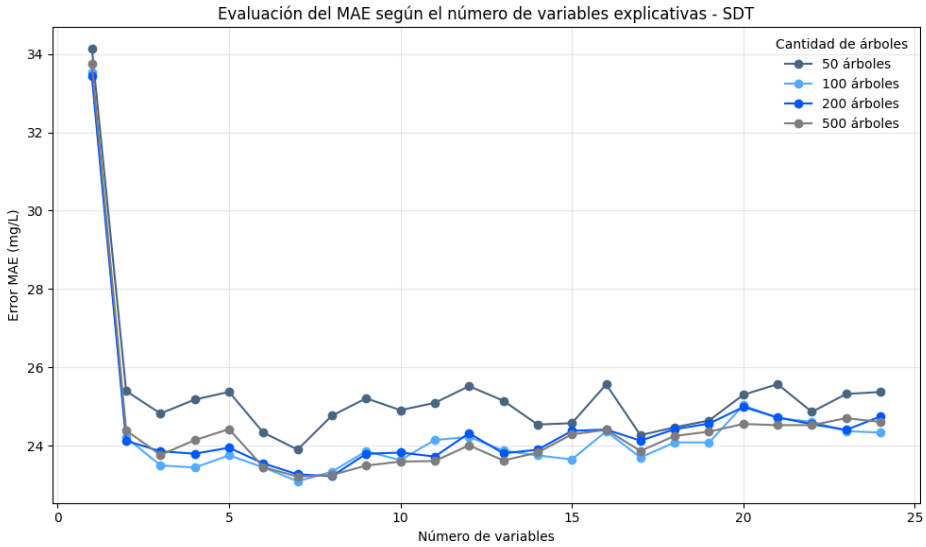

Figura 4.9: Evolución del MAE en función del número de variables predictoras para SDT.

A partir de la figura 4.9 se observa una disminución significativa del MAE al aumentar el número de variables predictoras desde valores bajos, lo que indica que la incorporación de información adicional resulta relevante para mejorar la capacidad del modelo de representar el comportamiento de los SDT. Sin embargo, a medida que se incorporan más variables, la reducción del error se vuelve menos pronunciada y el MAE tiende a estabilizarse para las distintas configuraciones de árboles consideradas.

En particular, se aprecia que a partir de un cierto número de variables predictoras las curvas asociadas a 100, 200 y 500 árboles presentan valores de MAE muy similares, lo que sugiere que la inclusión de variables adicionales no genera mejoras sustantivas en el desempeño del modelo. La configuración con 50 árboles muestra, en general, valores de MAE levemente superiores, aunque sin alterar de manera significativa el comportamiento global observado.

En base a estos resultados, se optó por seleccionar un subconjunto de 6 variables predictoras para la modelación de los SDT. Esta elección se fundamenta en que, a partir de dicho número de variables, el MAE presenta variaciones menores y no se observan mejoras significativas al incorporar variables adicionales. De este modo, se prioriza una configuración de baja complejidad que permite mantener un desempeño predictivo simi-

lar al obtenido con un mayor número de variables, evitando una sobrecarga innecesaria del modelo. La configuración seleccionada constituye la base para los análisis posteriores de ajuste del número de árboles, evaluación de la importancia de variables y generación de proyecciones de SDT.

4.2.2. Selección del número óptimo de árboles

Con el objetivo de determinar un número adecuado de árboles para la modelación de los sólidos disueltos totales (SDT), se analizó el comportamiento del error absoluto medio (MAE) al variar el tamaño del ensamble, considerando el modelo Random Forest con seis variables predictoras (RF-6). Para ello, se evaluaron distintas configuraciones del número de árboles, específicamente 50, 100, 200, 300, 400, 500, 600 y 1000 árboles, manteniendo fijo el conjunto de variables seleccionado en la etapa previa.

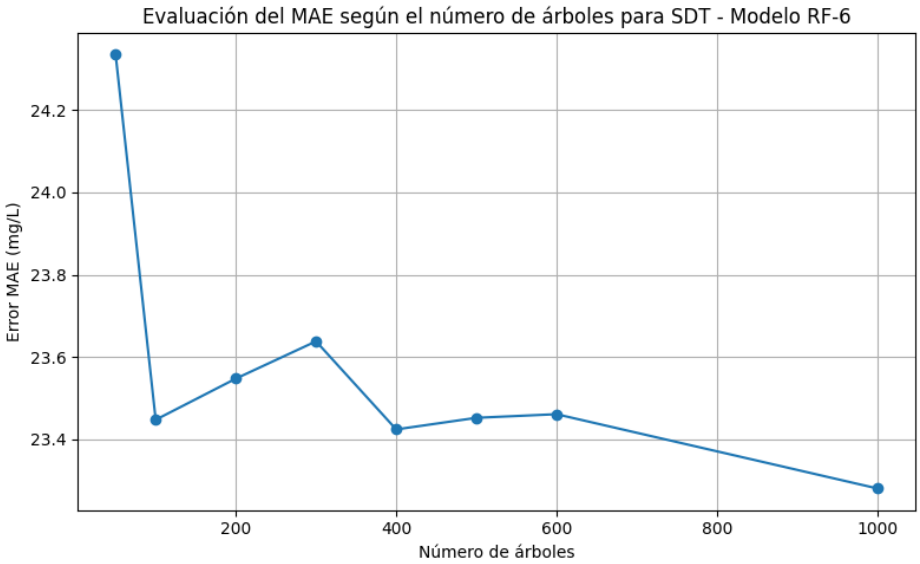

Figura 4.10: Comportamiento del MAE en función del número de árboles para SDT.

En la figura 4.10 se presenta la evolución del MAE en función del número de árboles del modelo. Se observa que el error disminuye de forma marcada al aumentar el número de árboles desde valores bajos, alcanzando valores reducidos ya a partir de configuraciones con 100 árboles. A partir de este punto, el MAE presenta variaciones menores y, en algunos casos, incluso se observa un leve aumento del error al considerar un mayor número de árboles intermedios.

Si bien al considerar un número elevado de árboles, como 1000, se obtiene un valor

de MAE ligeramente inferior, la diferencia respecto a configuraciones con un menor número de árboles resulta poco significativa. En este contexto, se optó por seleccionar un modelo con 100 árboles, ya que permite alcanzar un desempeño predictivo adecuado y estable, reduciendo al mismo tiempo el costo computacional asociado a un ensamble de mayor tamaño.

De este modo, la configuración seleccionada (6 variables predictoras y 100 árboles) equilibra de manera adecuada precisión y eficiencia computacional, y constituye la base para los análisis posteriores de importancia de variables y generación de proyecciones de SDT.

4.2.3. Importancia de las variables

Una vez definida la configuración final del modelo Random Forest para los sólidos disueltos totales, considerando seis variables predictoras y 100 árboles (modelo RF-6), se analizó la importancia relativa de las variables con el fin de identificar cuáles contribuyen en mayor medida a explicar la variabilidad de los SDT estimados por el modelo. Esta importancia se calcula a partir de la reducción de impureza generada por cada variable a lo largo de los árboles que componen el bosque, lo que permite obtener una medida del aporte relativo de cada predictor en el proceso de estimación.

Siguiendo el criterio de selección y ordenamiento de variables descrito previamente, el conjunto de seis variables consideradas en el modelo RF-6 corresponde a las variables de mayor importancia identificadas a partir del conjunto completo de predictores disponibles para los SDT. Al trabajar con este subconjunto reducido, el modelo se re-entrena, por lo que el orden relativo de las variables puede diferir respecto del ranking obtenido al considerar todas las variables, sin que ello implique un cambio en el conjunto de predictores utilizados.

La figura 4.11 presenta la importancia relativa de las variables seleccionadas para el modelo RF-6, expresada en términos porcentuales.

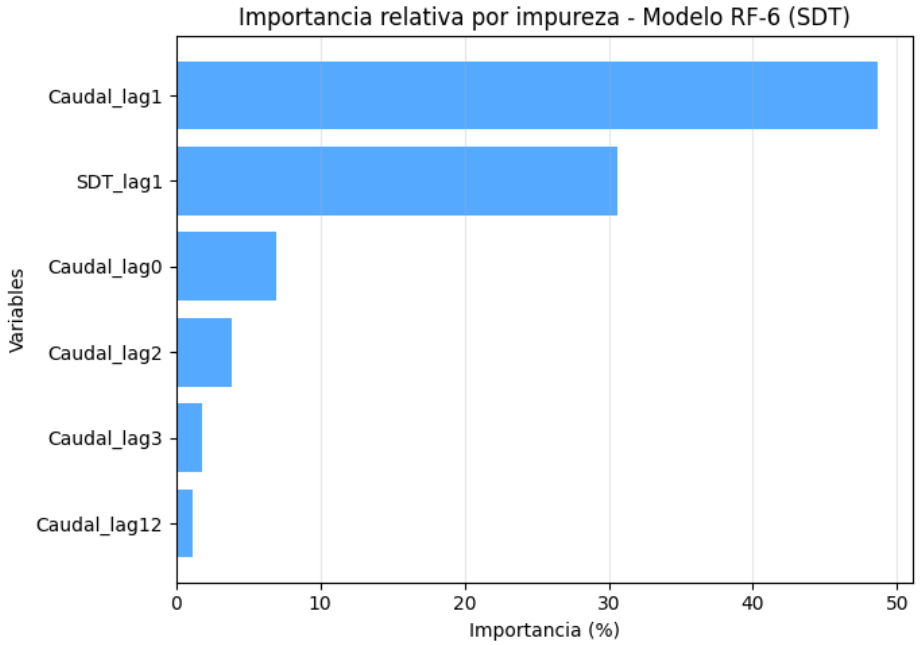

Figura 4.11: Importancia relativa de las variables predictoras en la estimación de SDT.

A partir de la figura 4.11 se observa que el rezago del caudal en el tiempo $t-1$ es la variable más influyente en la estimación de los SDT, concentrando una proporción significativa de la importancia total. Este resultado evidencia una relación significativa entre el comportamiento reciente del caudal y la concentración de sólidos disueltos, lo cual es consistente con la dinámica hidrológica del sistema, donde variaciones en el flujo pueden influir directamente en los procesos de dilución y arrastre de minerales disueltos.

En segundo lugar, destaca la importancia del rezago de los propios SDT en el tiempo $t-1$, lo que indica la presencia de una dependencia temporal de corto plazo en la serie de SDT. Las restantes variables, correspondientes a distintos rezagos del caudal, presentan una contribución menor, aunque no despreciable, lo que sugiere que el comportamiento de los SDT está determinado principalmente por la interacción entre su evolución reciente y la dinámica del caudal, más que por efectos climáticos directos en esta configuración particular del modelo.

En conjunto, la estructura de importancias obtenida respalda la selección de variables realizada previamente y confirma que el modelo Random Forest logra identificar de manera coherente los principales factores que explican la variabilidad de los sólidos disueltos totales en la PTAP San Juan.

4.2.4. Proyecciones de SDT

A continuación, se presentan las proyecciones de sólidos disueltos totales (SDT) obtenidas mediante el modelo Random Forest definido previamente para esta variable. En particular, se utiliza la configuración final del modelo, la cual considera 6 variables predictoras y un total de 100 árboles, seleccionada por ofrecer un buen equilibrio entre desempeño predictivo y complejidad del modelo.

Las proyecciones se realizan para horizontes temporales de uno y cinco años, lo que permite analizar la evolución esperada de los SDT tanto en el corto como en el mediano plazo, y evaluar la coherencia de las estimaciones generadas por el modelo en el contexto del comportamiento histórico del sistema San Juan.

Proyecciones a 1 año de SDT

A continuación, se presenta la proyección de sólidos disueltos totales (SDT) a un año obtenida mediante el modelo Random Forest seleccionado para esta variable. En este caso, el período histórico disponible para los SDT abarca desde enero de 2021 hasta mayo de 2025, mientras que la proyección a un año se extiende desde junio de 2025 hasta junio de 2026.

Con el objetivo de facilitar la visualización de la transición entre el período histórico y el período proyectado, en la figura se muestra únicamente un tramo reciente de la serie histórica, correspondiente al período comprendido entre junio de 2023 y mayo de 2025, junto con el ajuste histórico del modelo. Esta decisión tiene un carácter exclusivamente gráfico y no implica una reducción del conjunto de datos utilizados en el entrenamiento del modelo, el cual considera la totalidad del período histórico disponible.

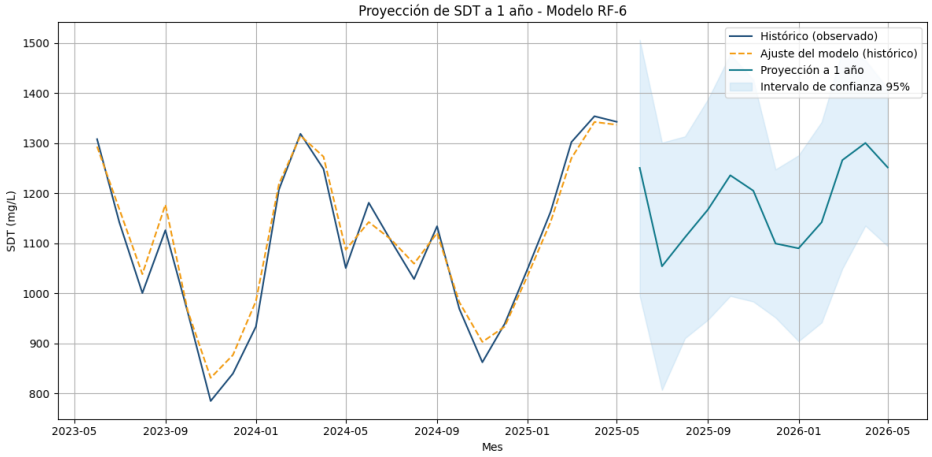

Figura 4.12: Ajuste histórico y proyección a un año de SDT.

En la figura 4.12 se observa que el modelo reproduce adecuadamente el comportamiento de los SDT en el período histórico mostrado, capturando tanto los niveles como las oscilaciones presentes en la serie. La proyección a un año presenta una evolución coherente con el comportamiento reciente, manteniendo valores dentro de rangos consistentes con la dinámica histórica de los SDT en el sistema San Juan.

En el tramo proyectado se aprecia una variabilidad moderada, con fluctuaciones que reflejan la influencia conjunta de las variables explicativas consideradas en el modelo, en particular el caudal y los rezagos temporales de los propios SDT. El intervalo de confianza asociado a la proyección refleja la incertidumbre de las estimaciones, y se construye a partir de la dispersión de las predicciones generadas por los distintos árboles que componen el bosque.

En conjunto, la proyección a un año de SDT entrega una estimación razonable del comportamiento esperado de este parámetro en el corto plazo, constituyendo una base adecuada para el análisis de escenarios operativos y para su comparación con las proyecciones a mayor horizonte temporal que se presentan en el apartado siguiente.

Proyecciones a 5 años de SDT

A continuación, se presenta la proyección de sólidos disueltos totales (SDT) a cinco años obtenida mediante el modelo Random Forest definido previamente. A diferencia del caso de proyección a un año, en esta figura se incluye la totalidad del período histórico disponible, correspondiente al intervalo comprendido entre enero de 2021 y mayo de 2025, junto con el ajuste histórico del modelo. Esto permite contextualizar la

proyección dentro de la evolución temporal completa de los SDT y evaluar su coherencia respecto de los patrones observados históricamente.

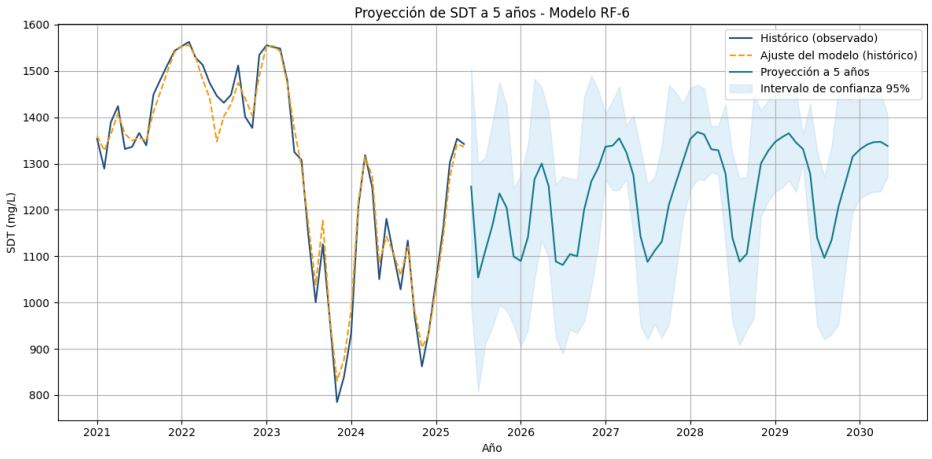

Figura 4.13: Ajuste histórico y proyección a cinco años de SDT.

En la figura 4.13 se observa que el modelo reproduce adecuadamente la dinámica histórica de los SDT, capturando tanto los niveles promedio como los episodios de mayor variabilidad presentes en la serie. La proyección a cinco años mantiene una estructura temporal coherente con el comportamiento histórico reciente, mostrando oscilaciones periódicas y niveles compatibles con los valores observados hacia el final del período histórico.

A medida que aumenta el horizonte temporal, el intervalo de confianza asociado a la proyección se amplía de manera progresiva, reflejando el incremento natural de la incertidumbre en proyecciones de mediano plazo. Estos intervalos se construyen a partir de la dispersión de las predicciones generadas por los distintos árboles del bosque, y permanecen dentro de rangos consistentes con la variabilidad histórica de los SDT en el sistema San Juan.

En conjunto, la proyección a cinco años de SDT proporciona una estimación coherente del comportamiento esperado de este parámetro en el mediano plazo, constituyendo un insumo relevante para el análisis de tendencias futuras y para su posterior utilización en la evaluación integrada de la calidad del agua de la PTAP San Juan.

4.2.5. Validación del modelo

Con el fin de evaluar el desempeño predictivo del modelo Random Forest en la estimación de los sólidos disueltos totales (SDT) en la PTAP San Juan, se realizó una

etapa de validación utilizando el error porcentual absoluto medio (MAPE), el cual permite cuantificar el error relativo entre valores proyectados y observados.

La validación se desarrolló en dos etapas. En primer lugar, se analizó el comportamiento del MAPE al variar el número de variables predictoras, con el objetivo de identificar la configuración que entrega el mejor equilibrio entre precisión y complejidad del modelo. Posteriormente, se evaluó la coherencia de las proyecciones generadas por el modelo seleccionado mediante la comparación directa entre valores proyectados y observados en un período de validación.

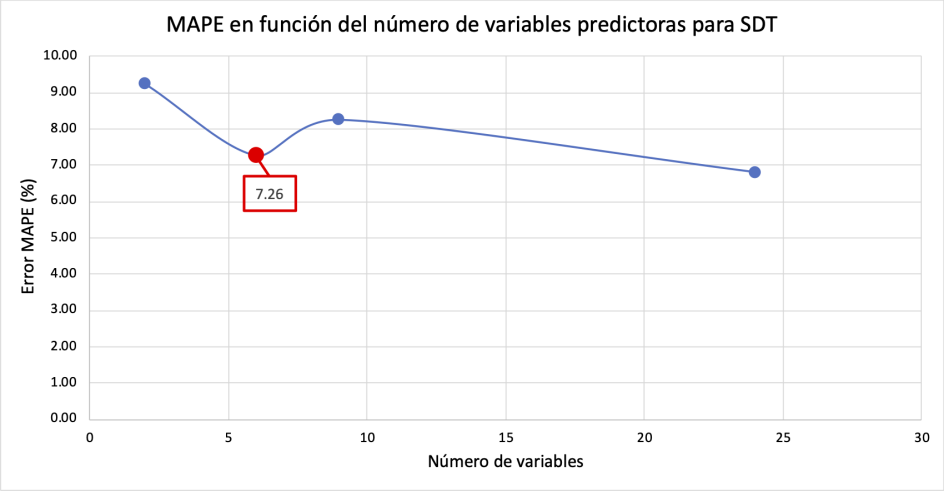

Figura 4.14: Comportamiento del MAPE para distintas configuraciones del modelo de SDT.

En primer lugar, la figura 4.14 presenta el comportamiento del MAPE en función del número de variables predictoras consideradas en el modelo, evaluándose las configuraciones de 2, 6, 9 y 24 variables. Estas configuraciones fueron seleccionadas a partir de los análisis previos de selección del número óptimo de variables mediante el error absoluto medio (MAE), los cuales permitieron identificar subconjuntos representativos del desempeño del modelo al variar su complejidad.

A partir de la figura, se observa que el error porcentual absoluto medio disminuye de manera significativa al pasar de 2 a 6 variables predictoras, alcanzando en esta configuración un valor de MAPE igual a 7,26 %. Al considerar 9 variables, el error aumenta levemente, mientras que al incorporar la totalidad de las 24 variables disponibles se observa una disminución del MAPE, alcanzando un valor comparable, aunque ligeramente inferior, al obtenido con seis variables.

No obstante, considerando que la reducción adicional del error al utilizar 24 variables

es baja, se optó por seleccionar la configuración con seis variables predictoras. Esta elección permite mantener un desempeño predictivo adecuado, concentrando el modelo en las variables más relevantes y evitando una complejidad innecesaria. De este modo, la configuración RF-6 ofrece un equilibrio adecuado entre precisión, interpretabilidad y simplicidad del modelo para la estimación de los sólidos disueltos totales.

Una vez definida la configuración final del modelo para los sólidos disueltos totales, se procedió a validar directamente la capacidad predictiva del modelo Random Forest con seis variables predictoras (RF-6). Para ello, el modelo fue entrenado utilizando la información histórica disponible hasta diciembre de 2024 y posteriormente se generó una proyección para el período comprendido entre enero y mayo de 2025. Esta etapa de validación permite evaluar el grado de concordancia entre los valores proyectados por el modelo y los valores observados.

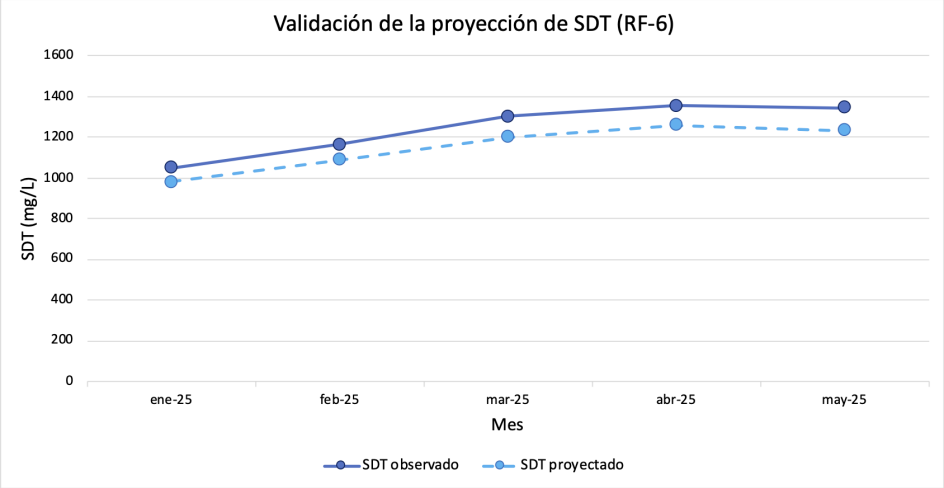

Figura 4.15: Comparación entre valores observados y proyectados de SDT (enero–mayo 2025).

La figura 4.15 muestra la comparación entre los valores observados y los valores proyectados de SDT para el período enero–mayo de 2025. A partir de esta comparación, se observa que el modelo RF-6 reproduce adecuadamente la tendencia creciente de los sólidos disueltos totales durante el período analizado, capturando tanto el aumento progresivo observado entre los meses de verano y otoño como el nivel general de la serie.

Si bien se aprecian diferencias puntuales entre los valores proyectados y los observados, estas se mantienen dentro de rangos acotados y no alteran la dinámica general de la serie. En conjunto, estos resultados indican que el modelo Random Forest con seis

variables predictoras presenta un desempeño adecuado en la proyección de SDT a corto plazo, respaldando su utilización para el análisis de escenarios futuros y su incorporación en la modelación integrada de los parámetros de calidad del agua abordados en este estudio.

4.3. Resultados para nitratos (NO_3^-)

En esta sección se presentan los resultados obtenidos para la modelación de nitratos (NO_3^- en la PTAP San Juan mediante el modelo Random Forest. Dado que esta variable puede responder tanto a la dinámica hidrológica (dilución/arrastre) como a factores externos asociados al sistema, se analizan distintas configuraciones del modelo con el fin de identificar un balance adecuado entre desempeño predictivo e interpretabilidad. En particular, se estudia la selección de variables predictoras y del número de árboles del bosque, se interpreta la importancia relativa de los predictores seleccionados y se presentan proyecciones a uno y cinco años, junto con su validación.

4.3.1. Selección del número óptimo de variables predictoras

Siguiendo el mismo procedimiento de selección y ordenamiento de variables predictoras descrito para el caso del caudal en la sección 4.1.1, el análisis para la estimación de nitratos se realizó utilizando rankings de importancia obtenidos a partir de la reducción promedio de impureza estimada por el modelo Random Forest.

Con el objetivo de determinar una configuración adecuada del modelo Random Forest para la estimación de nitratos, se analizó el comportamiento del error absoluto medio (MAE) al variar el número de variables predictoras incluidas en el modelo. Este análisis se realizó considerando distintas configuraciones del número de árboles del bosque, específicamente 50, 100, 200 y 500 árboles, con el fin de evaluar de manera conjunta la influencia del número de predictores y del tamaño del ensamble en el desempeño del modelo.

En la figura 4.16 se presenta la evolución del MAE en función del número de variables predictoras para cada una de las configuraciones de árboles analizadas.

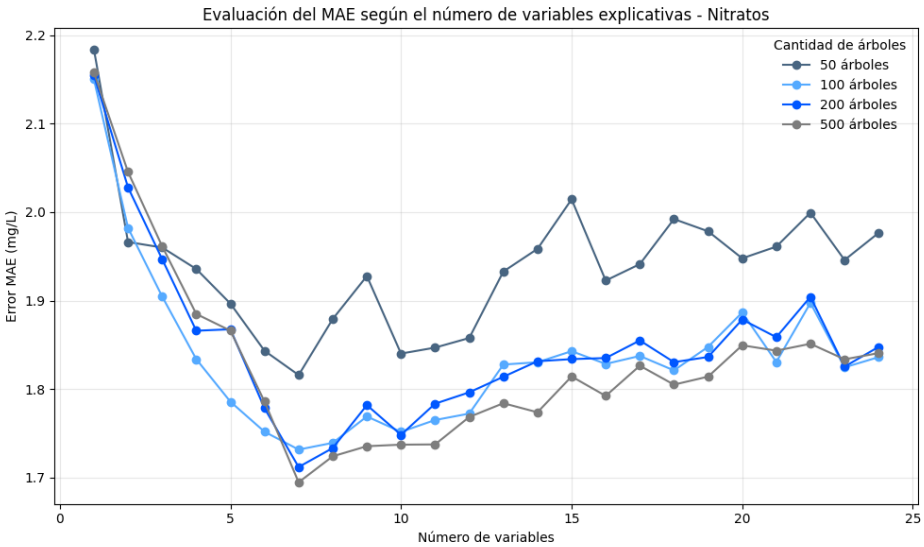

Figura 4.16: Evolución del MAE en función del número de variables predictoras para NO_3^-.

A partir de la figura 4.16 se observa una disminución marcada del MAE al incrementar el número de variables desde valores bajos, lo que indica que la incorporación de información adicional mejora la capacidad del modelo para representar la dinámica de NO_3^-. Sin embargo, a medida que aumenta el número de variables, la reducción del error se vuelve menos pronunciada y las curvas tienden a estabilizarse, especialmente para configuraciones con 100, 200 y 500 árboles, las cuales presentan valores de MAE similares entre sí a partir de un cierto umbral. En particular, se aprecia que alrededor de siete variables predictoras se alcanza un nivel de error bajo y relativamente estable, mientras que incorporar variables adicionales no genera mejoras sustantivas y, en algunos tramos, puede incluso introducir fluctuaciones del MAE. Considerando estos resultados, se optó por seleccionar un subconjunto de 7 variables predictoras para la modelación de NO_3^-, privilegiando una configuración de menor complejidad que mantiene un desempeño comparable al obtenido con un mayor número de variables. Esta configuración constituye la base para la selección del número óptimo de árboles, el análisis de importancia de variables y la generación de proyecciones presentadas en los apartados siguientes.

4.3.2. Selección del número óptimo de árboles

Una vez definido el conjunto de siete variables predictoras para la estimación de nitratos, se analizó la influencia del número de árboles del bosque sobre el desempeño

del modelo Random Forest. Para este análisis se consideró el modelo RF-7 y se evaluó el comportamiento del error absoluto medio (MAE) al variar el tamaño del ensamble, con el objetivo de identificar una configuración que entregara predicciones estables sin incurrir en un costo computacional innecesario.

En la figura 4.17 se presenta la evolución del MAE en función del número de árboles considerados en el modelo.

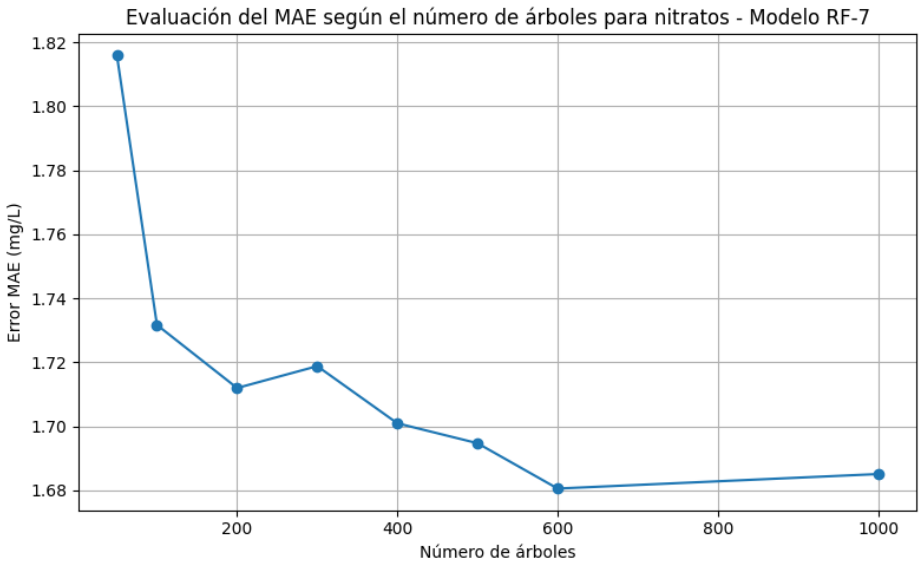

Figura 4.17: Comportamiento del MAE en función del número de árboles para NO_3^-.

A partir de la figura 4.17 se observa que el MAE disminuye de manera significativa al aumentar el número de árboles desde valores bajos, evidenciando una mejora en la estabilidad de las predicciones al incrementar el tamaño del bosque. No obstante, esta disminución del error se vuelve progresivamente más leve a partir de aproximadamente 200 árboles, a partir de los cuales las variaciones del MAE son reducidas y no muestran una tendencia clara a una mejora sustantiva.

Si bien al considerar un número mayor de árboles, como 600 o 1000, se observa un valor de MAE ligeramente inferior, la diferencia respecto a configuraciones con 200 árboles resulta marginal. En este contexto, se optó por seleccionar un modelo con 200 árboles y siete variables predictoras, ya que esta configuración permite alcanzar un desempeño predictivo adecuado y estable, manteniendo un balance razonable entre precisión y eficiencia computacional. La configuración seleccionada constituye la base para el análisis de importancia de variables y la generación de proyecciones de NO_3^- presentadas en los apartados siguientes.

4.3.3. Importancia de las variables

Una vez definida la configuración final del modelo Random Forest para nitratos, considerando siete variables predictoras y 200 árboles (modelo RF-7), se analizó la importancia relativa de las variables con el fin de identificar los principales factores que explican la variabilidad estimada de NO_3^-. La importancia de cada variable se calcula a partir de la reducción de impureza generada a lo largo de los árboles del bosque, lo que permite evaluar su aporte relativo dentro del modelo.

De acuerdo con el criterio de selección y ordenamiento de variables utilizado en este estudio, las siete variables incorporadas en el modelo RF-7 corresponden a los predictores con mayor importancia relativa identificados a partir del conjunto completo de variables disponibles para nitratos. Al entrenar el modelo utilizando únicamente este subconjunto, el Random Forest se ajusta nuevamente, por lo que el orden relativo de las variables puede diferir respecto del ranking obtenido al considerar todas las variables, sin modificar el conjunto de predictores empleados.

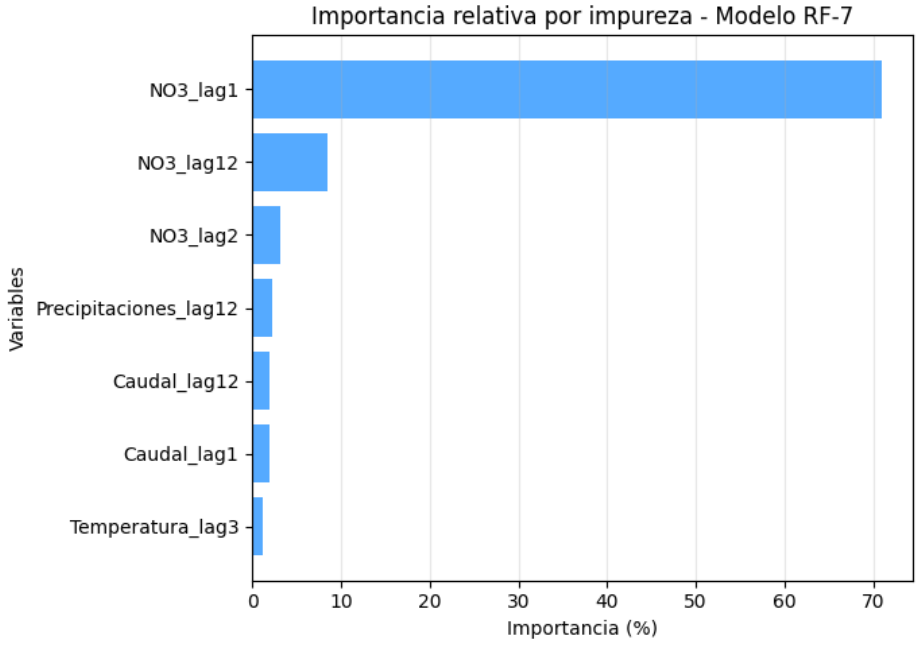

Figura 4.18: Importancia relativa de las variables predictoras en la estimación de NO_3^-.

A partir de la figura 4.18 se observa que el rezago de los nitratos en el tiempo $t-1$ es, con amplia diferencia, la variable más influyente en la estimación de NO_3^-, concentrando la mayor parte de la importancia total. Este resultado evidencia una fuerte dependencia

temporal de corto plazo en la serie de nitratos, donde el valor inmediatamente anterior entrega información clave para la estimación del valor actual.

En segundo lugar, destacan los rezagos de NO_3^- en el tiempo $t-12$ y $t-2$, lo que sugiere la presencia de efectos estacionales y dependencias temporales adicionales. Las variables externas, como precipitaciones, caudal y temperatura en ciertos rezagos, presentan una contribución menor, aunque no despreciable, indicando que su rol es principalmente complementario. En conjunto, la estructura de importancias confirma que el comportamiento de los nitratos en la PTAP San Juan está dominado por su propia dinámica temporal, con una influencia secundaria de variables hidrológicas y climáticas, lo cual respalda la selección de variables realizada para el modelo RF-7.

4.3.4. Proyecciones de NO_3^-

A continuación, se presentan las proyecciones de nitratos obtenidas mediante el modelo Random Forest definido previamente para esta variable. En particular, se utiliza la configuración final del modelo, la cual considera siete variables predictoras y un total de 200 árboles, seleccionada por ofrecer un desempeño predictivo adecuado y una complejidad controlada.

Las proyecciones se realizan para horizontes temporales de uno y cinco años, lo que permite analizar la evolución esperada de la concentración de nitratos tanto en el corto como en el mediano plazo. Estos resultados se presentan en conjunto con el ajuste histórico del modelo, con el fin de evaluar la coherencia de las estimaciones proyectadas respecto del comportamiento observado en la serie histórica.

Proyecciones a 1 año de NO_3^-

A continuación, se presenta la proyección de nitratos a un año, con el objetivo de analizar la evolución esperada de este parámetro en el corto plazo. El período histórico disponible para los nitratos abarca desde enero de 2021 hasta mayo de 2025, mientras que la proyección se extiende desde junio de 2025 hasta junio de 2026.

Con fines de visualización, en la figura se muestra únicamente el tramo más reciente de la serie histórica, correspondiente al período junio de 2023–mayo de 2025, junto con el ajuste histórico del modelo. Esta representación gráfica no implica una reducción de la información utilizada en el entrenamiento, el cual considera la totalidad del período histórico disponible.

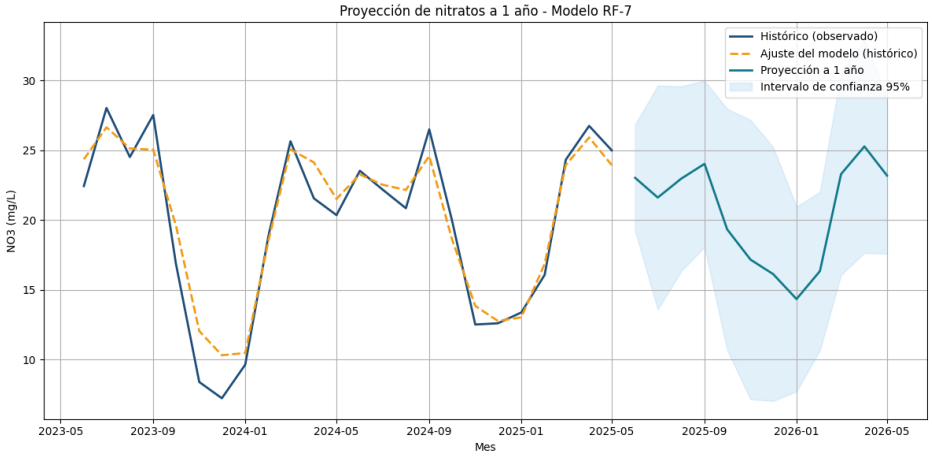

Figura 4.19: Ajuste histórico y proyección a un año de NO_3^-.

En la figura 4.19 se observa que el modelo reproduce adecuadamente el comportamiento histórico reciente de los nitratos, capturando tanto los niveles como las oscilaciones de la serie. La proyección a un año presenta una evolución coherente con la dinámica observada, manteniéndose dentro de rangos consistentes con el comportamiento histórico del sistema.

El intervalo de confianza asociado a la proyección se amplía progresivamente hacia el horizonte futuro, reflejando la incertidumbre inherente a las estimaciones de corto plazo. En conjunto, estos resultados indican que el modelo entrega una estimación razonable del comportamiento esperado de los nitratos, constituyendo una base adecuada para el análisis de escenarios futuros.

Proyecciones a 5 años de NO_3^-

A continuación, se presenta la proyección de nitratos a cinco años, con el objetivo de analizar su comportamiento esperado en el mediano plazo y evaluar la coherencia de las estimaciones respecto de la evolución histórica del sistema.

En la figura 4.20 se incluye la totalidad del período histórico disponible, correspondiente al intervalo enero de 2021–mayo de 2025, junto con el ajuste histórico del modelo. Esta representación permite contextualizar la proyección dentro de la dinámica completa de los nitratos y analizar su comportamiento a lo largo de distintos regímenes observados en la serie.

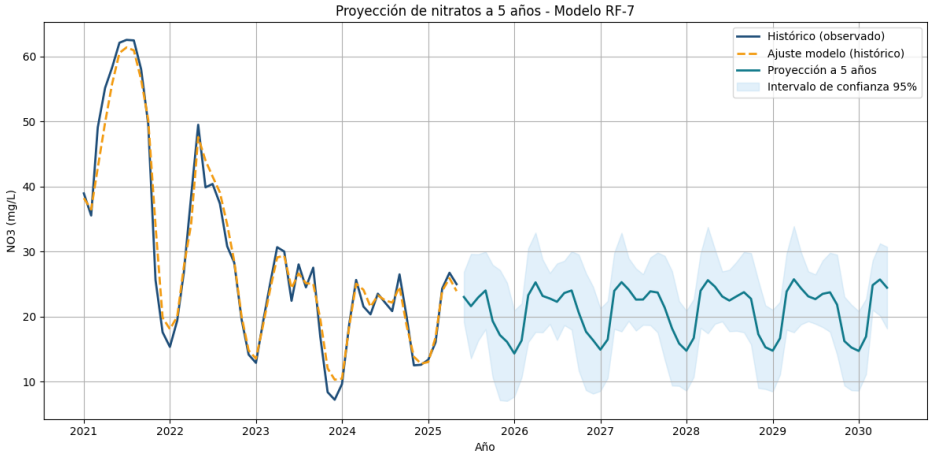

Figura 4.20: Ajuste histórico y proyección a cinco años de NO_3^-.

En la figura 4.20 se observa que la proyección mantiene una estructura temporal coherente con el comportamiento histórico de los nitratos, reproduciendo oscilaciones periódicas y niveles acordes a los valores observados en los últimos años de la serie. El modelo proyecta una dinámica relativamente estable, sin tendencias abruptas de aumento o disminución sostenida, lo que sugiere una persistencia del patrón observado recientemente.

A medida que se extiende el horizonte temporal, el intervalo de confianza asociado a la proyección se amplía de manera progresiva, reflejando el incremento natural de la incertidumbre en estimaciones de mediano plazo. No obstante, los rangos proyectados se mantienen dentro de valores compatibles con la variabilidad histórica de los nitratos en la PTAP San Juan.

En conjunto, la proyección a cinco años de NO_3^- entrega una estimación consistente del comportamiento esperado de este parámetro en el mediano plazo, aportando información relevante para el análisis de tendencias futuras y la evaluación integrada de la calidad del agua del sistema San Juan.

4.3.5. Validación del modelo

Con el fin de evaluar el desempeño predictivo del modelo Random Forest en la estimación de nitratos, se analizó el comportamiento del error porcentual absoluto medio (MAPE) al variar el número de variables predictoras consideradas. En particular, se evaluaron las configuraciones de 4, 7 y 24 variables, las cuales fueron definidas a partir de los análisis previos de selección del modelo mediante el error absoluto medio (MAE).

Este análisis permite complementar los criterios utilizados en etapas anteriores, incorporando una medida de error relativa que resulta especialmente útil para comparar configuraciones con distinta complejidad en un mismo horizonte de validación.

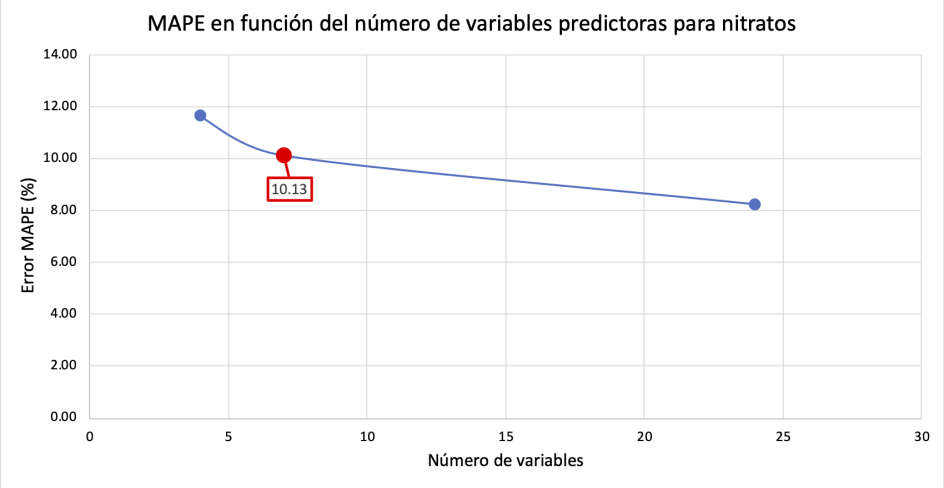

Figura 4.21: Comportamiento del MAPE para distintas configuraciones del modelo de NO_3^-.

La figura 4.21 muestra el comportamiento del MAPE para distintas configuraciones del modelo de nitratos, considerando conjuntos reducidos y ampliados de variables predictoras. Se observa que el error porcentual disminuye de manera relevante al pasar desde configuraciones muy simples hacia un conjunto intermedio de variables, alcanzando un valor de MAPE igual a 10,13 % al considerar siete variables predictoras. Si bien al incorporar un número mayor de variables se observa una reducción adicional del error, esta mejora resulta menor en relación con el aumento de la complejidad del modelo.

En este contexto, se optó por mantener la configuración con siete variables predictoras, priorizando un modelo más simple que conserve un desempeño predictivo adecuado. Esta decisión permite concentrar el análisis en los predictores más relevantes para la dinámica de los nitratos, manteniendo coherencia con las etapas previas de selección del modelo y favoreciendo su interpretabilidad y estabilidad.

Luego, una vez definida la configuración final del modelo para nitratos, se procedió a realizar una validación directa de las proyecciones generadas por el modelo Random Forest. Para ello, el modelo fue entrenado utilizando la información histórica disponible hasta diciembre de 2024 y posteriormente se generó una proyección para el período comprendido entre enero y mayo de 2025. Esta etapa permite evaluar de manera directa

la capacidad del modelo para reproducir valores no utilizados durante el entrenamiento, mediante la comparación entre los valores proyectados y los valores observados.

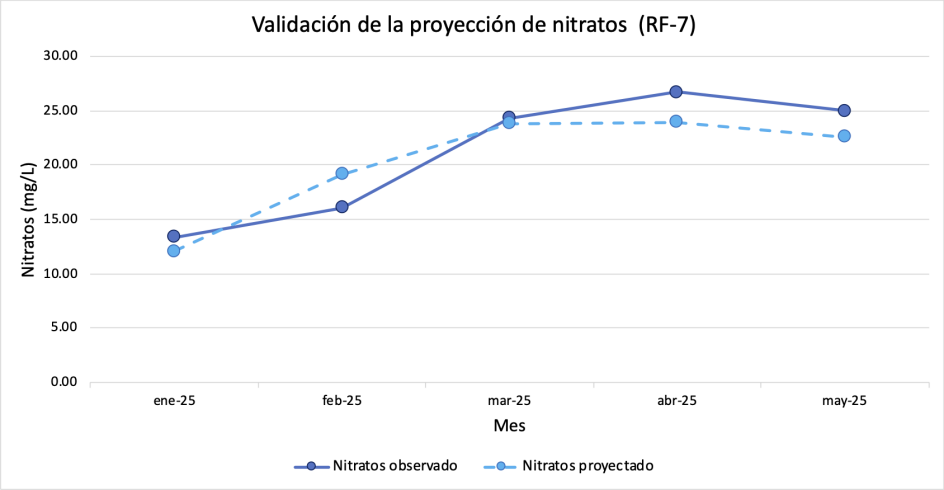

Figura 4.22: Comparación entre valores observados y proyectados de NO_3^- (enero–mayo 2025).

La figura 4.22 presenta la comparación entre los valores observados y proyectados de nitratos para el período enero–mayo de 2025. A partir de esta comparación, se observa que el modelo RF-7 logra reproducir adecuadamente la tendencia general de la serie, capturando el aumento progresivo de las concentraciones desde los primeros meses del año hasta el mes de abril, así como la leve disminución observada en mayo.

Si bien se identifican diferencias puntuales entre los valores proyectados y observados, estas se mantienen acotadas y no alteran el comportamiento global de la serie. En términos generales, el modelo reproduce correctamente el orden de magnitud y la dinámica temporal de los nitratos durante el período de validación, lo que respalda su capacidad predictiva a corto plazo y confirma su adecuación para el análisis de escenarios futuros en la PTAP San Juan.

4.4. Resultados para turbiedad

En esta sección se presentan los resultados obtenidos para la modelación de la turbiedad en la PTAP San Juan mediante el modelo Random Forest. La turbiedad constituye un parámetro clave de la calidad del agua, ya que está asociada a la presencia de partículas en suspensión y puede verse influenciada por procesos hidrológicos, climáticos

y operacionales, especialmente durante eventos de aumento de caudal o precipitaciones intensas.

Siguiendo la misma metodología aplicada a las variables analizadas previamente, se evalúan distintas configuraciones del modelo con el fin de caracterizar el comportamiento de la turbiedad, identificar los predictores más relevantes y analizar su desempeño predictivo. En particular, se aborda la selección del número óptimo de variables predictoras y del número de árboles del bosque, el análisis de la importancia relativa de las variables seleccionadas y la generación de proyecciones a horizontes de uno y cinco años, junto con su correspondiente validación.

Este enfoque permite evaluar la capacidad del modelo Random Forest para representar la dinámica temporal de la turbiedad y generar estimaciones coherentes con el comportamiento histórico del sistema de captación San Juan.

4.4.1. Selección del número óptimo de variables predictoras

Siguiendo el mismo procedimiento de selección y ordenamiento de variables predictoras descrito para el caso del caudal en la sección 4.1.1, el análisis para la estimación de la turbiedad se realizó utilizando rankings de importancia obtenidos a partir de la reducción promedio de impureza estimada por el modelo Random Forest.

Con el objetivo de definir una configuración adecuada del modelo Random Forest para la estimación de la turbiedad en la PTAP San Juan, se analizó el comportamiento del error absoluto medio (MAE) al variar el número de variables predictoras incluidas en el modelo. Este análisis se realizó considerando distintas configuraciones del número de árboles del bosque, específicamente 50, 100, 200 y 500 árboles, lo que permite evaluar de manera conjunta la influencia del número de predictores y del tamaño del ensamble en el desempeño del modelo.

Para el caso de la turbiedad, el conjunto total de variables predictoras disponibles estuvo compuesto por 24 variables, las cuales incluyen rezagos temporales de la propia turbiedad, así como variables externas asociadas al caudal y a condiciones climáticas (temperatura, precipitaciones y nieve acumulada), consideradas tanto en el tiempo contemporáneo como en distintos rezagos temporales. Este conjunto permite capturar tanto la dependencia temporal de la turbiedad como su relación con la dinámica hidrológica del sistema.

En la figura 4.23 se presenta la evolución del MAE en función del número de variables predictoras para las distintas configuraciones de árboles analizadas.

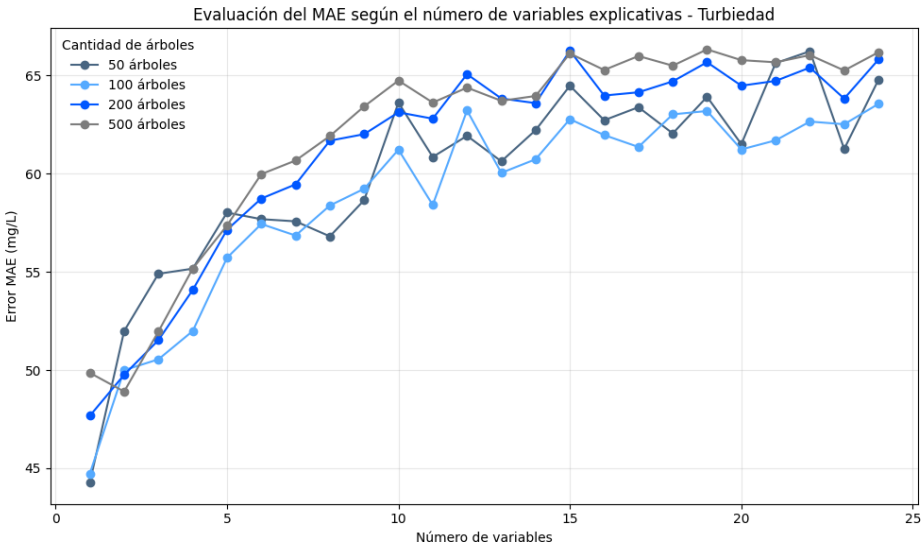

Figura 4.23: Evolución del MAE en función del número de variables predictoras para turbiedad.

A partir de la figura 4.23 se observa que el MAE aumenta de manera progresiva al incorporar un mayor número de variables predictoras, comportamiento que se mantiene de forma consistente para las distintas configuraciones de árboles consideradas. En particular, el error presenta valores relativamente más bajos al considerar un número reducido de variables, mientras que al aumentar la complejidad del modelo mediante la inclusión de predictores adicionales no se observan mejoras en el desempeño predictivo, sino más bien una mayor variabilidad del error.

Este comportamiento sugiere que, para la turbiedad, la incorporación de un número elevado de variables no aporta información adicional relevante y puede introducir ruido en el proceso de estimación. En este contexto, se optó por seleccionar un subconjunto de 7 variables predictoras, ya que esta configuración permite mantener un valor de MAE relativamente bajo y estable, evitando una complejidad innecesaria del modelo.

La selección de siete variables constituye un compromiso adecuado entre simplicidad y capacidad predictiva. Si bien el mínimo valor del MAE se obtiene al considerar una única variable predictora, esta configuración no fue seleccionada como modelo final, ya que el análisis basado en MAE corresponde a una etapa de calibración interna del modelo y no refleja necesariamente su desempeño en un escenario de proyección fuera de muestra. Como se analizará en la etapa de validación, el modelo con una sola variable presenta un mayor error porcentual absoluto medio (MAPE) al comparar valores proyectados con observados, mientras que la configuración con siete variables muestra

un comportamiento más estable en términos predictivos, proporcionando una base consistente para los análisis posteriores de selección del número de árboles, evaluación de la importancia de las variables y generación de proyecciones de turbiedad.

4.4.2. Selección del número óptimo de árboles

Con el fin de definir un número adecuado de árboles para la modelación de la turbiedad, se analizó el comportamiento del error absoluto medio (MAE) al variar el tamaño del ensamble del modelo Random Forest, manteniendo fijo el conjunto de siete variables predictoras seleccionado previamente. Este análisis permite evaluar la estabilidad del modelo y su desempeño predictivo frente a distintas configuraciones del número de árboles.

Para este estudio se consideraron modelos con 50, 100, 150, 200, 300, 400, 500, 600 y 1000 árboles, evaluando en cada caso el valor del MAE asociado a la estimación de la turbiedad.

La figura 4.24 presenta la evolución del error en función del número de árboles considerados.

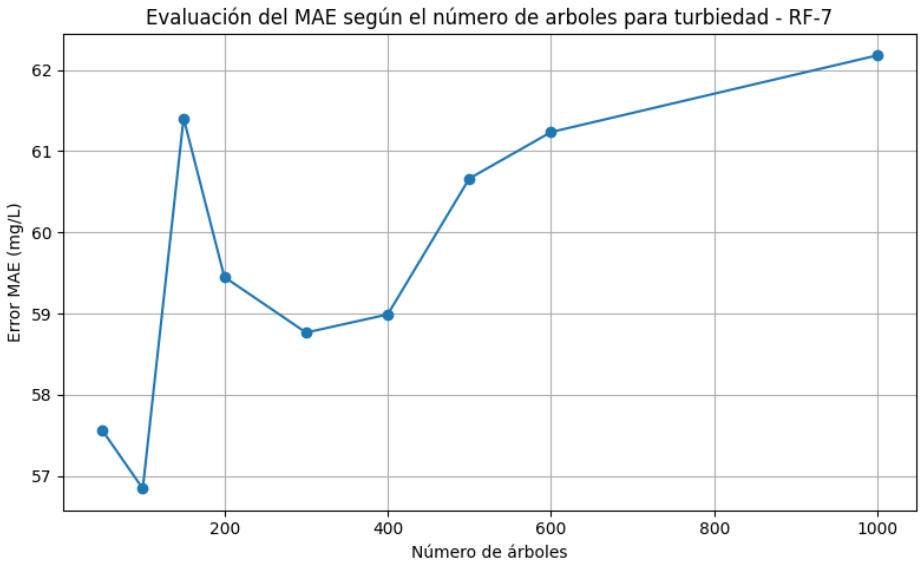

Figura 4.24: Comportamiento del MAE en función del número de árboles para turbiedad.

En la figura 4.24 se observa que el MAE alcanza valores bajos para configuraciones con un número reducido de árboles, destacando particularmente el caso de 100 árboles,

donde el error presenta uno de sus valores mínimos. Al aumentar el número de árboles por sobre este valor, el MAE no muestra una disminución, sino que presenta un incremento progresivo, alcanzando valores más altos para configuraciones con un mayor tamaño del ensamble.

Este comportamiento sugiere que, en el caso de la turbiedad, el incremento del número de árboles no se traduce en una mejora del desempeño predictivo del modelo y, por el contrario, puede introducir mayor variabilidad en las estimaciones. En este contexto, se optó por seleccionar un modelo con 100 árboles, ya que esta configuración permite obtener un MAE bajo y estable, evitando un aumento innecesario del costo computacional asociado a ensambles más grandes.

De este modo, la configuración final del modelo para la turbiedad, compuesta por siete variables predictoras y 100 árboles, ofrece un equilibrio adecuado entre precisión y eficiencia, y constituye la base para el análisis de la importancia de las variables y la generación de proyecciones presentadas en los apartados siguientes.

4.4.3. Importancia de las variables

Una vez definida la configuración final del modelo Random Forest para la turbiedad, considerando siete variables predictoras y 100 árboles (modelo RF-7), se analizó la importancia relativa de las variables con el objetivo de identificar los principales factores que influyen en la estimación de este parámetro. La importancia se calculó a partir de la reducción de impureza generada por cada variable a lo largo de los árboles que componen el bosque.

En coherencia con el criterio de selección y ordenamiento de variables aplicado en este estudio, el conjunto de siete variables consideradas en el modelo RF-7 corresponde a los predictores con mayor importancia relativa identificados a partir del conjunto completo de variables disponibles para turbiedad. Al entrenar el modelo utilizando este subconjunto, el Random Forest se ajusta nuevamente, lo que puede generar variaciones en el orden relativo de las variables respecto del ranking obtenido al considerar todas las variables, sin alterar el conjunto de predictores empleados.

La figura 4.25 presenta la importancia relativa de las variables predictoras incluidas en el modelo RF-7.

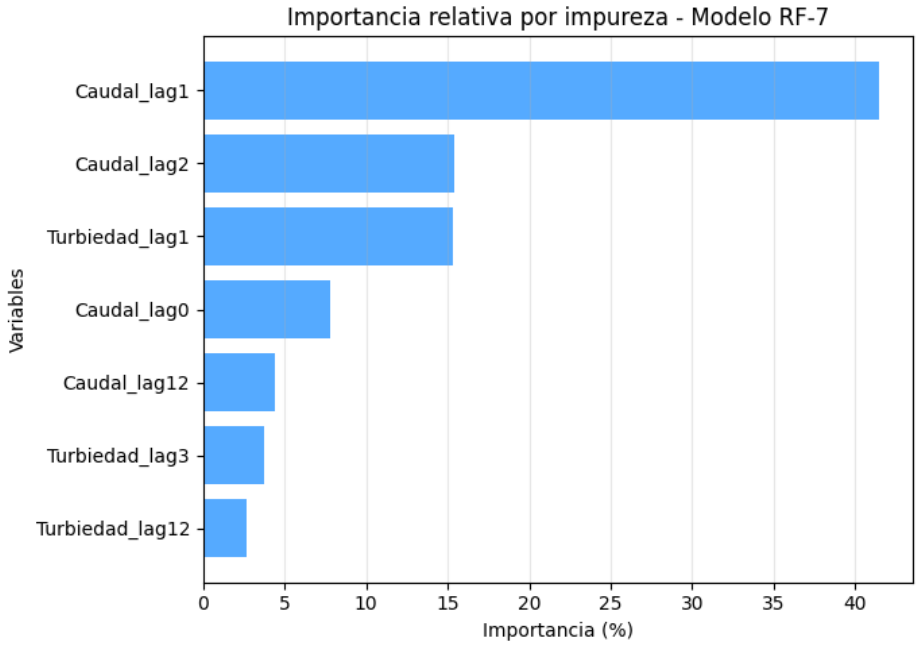

Figura 4.25: Importancia relativa de las variables predictoras en la estimación de turbiedad.

A partir de la figura 4.25, se observa que el rezago del caudal en el tiempo $t-1$ es la variable más influyente en la estimación de la turbiedad, concentrando una proporción significativa de la importancia total. Este resultado evidencia la fuerte relación entre el comportamiento reciente del caudal y los niveles de turbiedad, lo cual es coherente con los procesos físicos asociados al arrastre de material particulado durante variaciones en el flujo. En segundo lugar, destacan los rezagos del caudal en los tiempos $t-2$ y t, los cuales también presentan una contribución relevante en la explicación de la variabilidad de la turbiedad. La presencia de estos rezagos refuerza la idea de que la turbiedad responde de manera directa a cambios en la dinámica hidrológica del sistema.

Las variables asociadas a los rezagos de la propia turbiedad, en particular en los tiempos $t-1$, $t-3$ y $t-12$, presentan una importancia menor en comparación con los rezagos del caudal, aunque su inclusión resulta relevante para capturar efectos de persistencia y memoria temporal en la serie.

En conjunto, la estructura de importancias obtenida indica que la turbiedad en la PTAP San Juan está dominada principalmente por la dinámica del caudal, complementada por la influencia del comportamiento pasado de la propia turbiedad. Estos resultados son consistentes con la naturaleza del parámetro analizado y respaldan la

selección de variables realizada previamente, confirmando que el modelo Random Forest identifica de manera adecuada los factores clave que controlan la variabilidad de la turbiedad en el sistema.

4.4.4. Proyecciones de turbiedad

A continuación, se presentan las proyecciones de turbiedad obtenidas mediante el modelo Random Forest definido para esta variable. En particular, se utiliza la configuración final seleccionada en los apartados anteriores, la cual considera siete variables predictoras y un total de 100 árboles, al ofrecer un adecuado equilibrio entre desempeño predictivo y complejidad del modelo.

El análisis de proyecciones permite examinar la evolución esperada de la turbiedad en distintos horizontes temporales y evaluar la coherencia de las estimaciones generadas por el modelo en relación con el comportamiento histórico observado. Para ello, se consideran horizontes de proyección de uno y cinco años, lo que posibilita analizar tanto la evolución de corto plazo como las tendencias de mediano plazo de la turbiedad en la PTAP San Juan.

Proyecciones a 1 año de turbiedad

Luego, se presenta la proyección de turbiedad a un año obtenida mediante el modelo Random Forest seleccionado para esta variable. En el caso de la turbiedad, el período histórico disponible abarca desde enero de 2021 hasta mayo de 2025. No obstante, con el fin de facilitar la visualización de la transición entre el comportamiento histórico reciente y el período proyectado, en la figura se muestra únicamente el tramo comprendido entre junio de 2023 y mayo de 2025, junto con el ajuste histórico del modelo.

La proyección a un año se extiende desde junio de 2025 hasta junio de 2026, permitiendo analizar la evolución esperada de la turbiedad en el corto plazo y evaluar la coherencia de las estimaciones generadas por el modelo en relación con los valores observados más recientes.

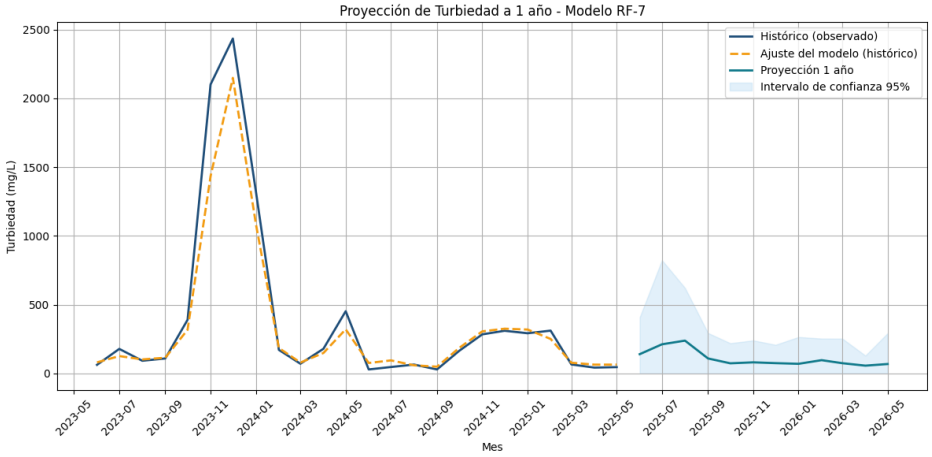

Figura 4.26: Ajuste histórico y proyección a un año de turbiedad.

En la figura 4.26 se observa que el modelo reproduce adecuadamente la dinámica general de la turbiedad en el período histórico mostrado, incluyendo episodios de alta variabilidad asociados a eventos puntuales, así como períodos de valores más bajos y estables. La proyección a un año presenta niveles de turbiedad relativamente acotados, manteniéndose dentro de rangos consistentes con el comportamiento observado hacia el final del período histórico.

El intervalo de confianza asociado a la proyección refleja la incertidumbre de las estimaciones de corto plazo, particularmente en una variable como la turbiedad, caracterizada por una alta variabilidad temporal y una fuerte sensibilidad a eventos hidrológicos extremos. En conjunto, la proyección a un año entrega una estimación razonable del comportamiento esperado de la turbiedad en el corto plazo, constituyendo un insumo relevante para el análisis de la calidad del agua en la PTAP San Juan.

Proyecciones a 5 años de turbiedad

A continuación, se presenta la proyección de turbiedad a 5 años obtenida mediante el modelo Random Forest definido para esta variable. En este caso, la figura incluye la totalidad del período histórico disponible, comprendido entre enero de 2021 y mayo de 2025, junto con el ajuste histórico del modelo. La proyección se extiende desde junio de 2025 hasta junio de 2030, lo que permite analizar la evolución esperada de la turbiedad en un horizonte de mediano plazo y evaluar la coherencia de las estimaciones en relación con el comportamiento histórico observado en el sistema San Juan.

En la figura 4.27 se observa que el modelo logra reproducir adecuadamente los princi-

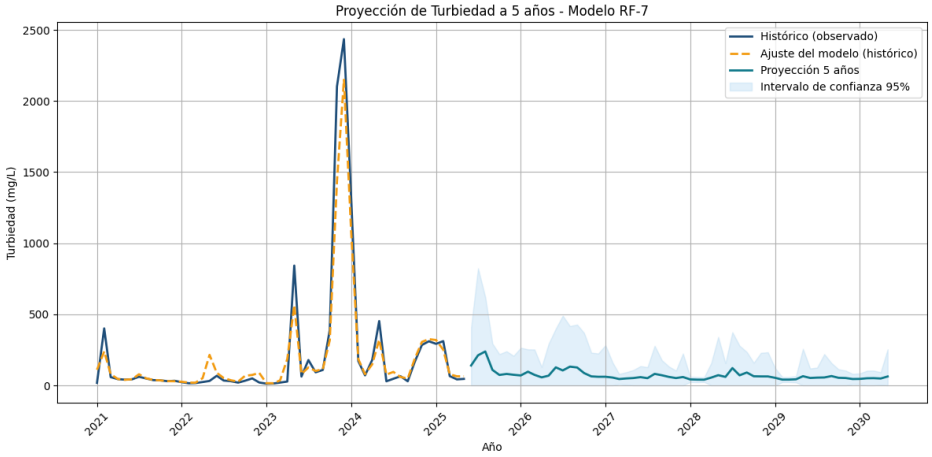

Figura 4.27: Ajuste histórico y proyección a cinco años de turbiedad.

pales rasgos del comportamiento histórico de la turbiedad, incluyendo episodios de alta variabilidad asociados a eventos puntuales, así como períodos prolongados con niveles bajos y relativamente estables. La proyección a cinco años muestra una dinámica más suavizada en comparación con los valores históricos extremos, manteniéndose dentro de rangos consistentes con el comportamiento reciente de la serie.

A medida que aumenta el horizonte temporal, el intervalo de confianza asociado a la proyección refleja el aumento natural de la incertidumbre en estimaciones de mediano plazo. No obstante, las trayectorias proyectadas permanecen dentro de rangos aceptables desde el punto de vista hidrológico, lo que sugiere que el modelo entrega una representación razonable de la evolución esperada de la turbiedad en el sistema.

En conjunto, la proyección a cinco años proporciona una visión integrada del comportamiento futuro de la turbiedad en la PTAP San Juan, constituyendo un insumo relevante para el análisis de tendencias de calidad del agua y para su consideración en estudios de planificación y gestión a mediano plazo.

4.4.5. Validación del modelo

Con el fin de evaluar el desempeño predictivo del modelo Random Forest en la estimación de la turbiedad, se analizó el comportamiento del error porcentual absoluto medio (MAPE) en función del número de variables predictoras consideradas. Este análisis permite evaluar cómo la complejidad del modelo influye en la precisión relativa de las proyecciones y complementa los resultados obtenidos previamente a partir del análisis del error absoluto medio (MAE).

En particular, se evaluaron tres configuraciones representativas del número de variables predictoras, correspondientes a 1, 7 y 24 variables, las cuales fueron seleccionadas a partir del análisis previo de selección del número óptimo de variables mediante el MAE. Estas configuraciones permiten comparar el desempeño del modelo bajo escenarios de baja complejidad, complejidad intermedia y máxima complejidad.

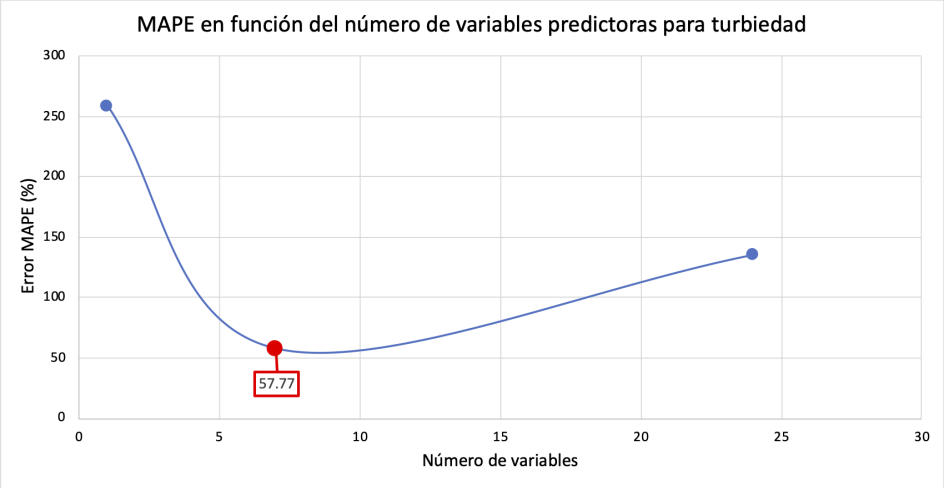

Figura 4.28: Comportamiento del MAPE para distintas configuraciones del modelo de turbiedad.

A partir de la figura 4.28 se observa que el MAPE disminuye de manera significativa al pasar de una a siete variables predictoras, alcanzando en esta última configuración su valor mínimo, correspondiente a un MAPE de $57,77\,\%$. Este resultado indica una mejora sustantiva en la capacidad predictiva del modelo al incorporar información adicional relevante.

Al considerar la configuración con 24 variables predictoras, el MAPE aumenta nuevamente, lo que sugiere que la inclusión de un mayor número de variables no aporta mejoras en términos de precisión relativa y puede introducir información redundante que afecta negativamente el desempeño del modelo.

En base a estos resultados, se confirma la selección del modelo con siete variables predictoras, ya que esta configuración entrega el mejor compromiso entre precisión predictiva y complejidad del modelo. Esta elección resulta además coherente con el análisis previo realizado mediante el MAE, reforzando la consistencia del proceso de selección del modelo para la estimación de la turbiedad.

Una vez definida la configuración final del modelo Random Forest para la turbiedad, se procedió a evaluar directamente la capacidad predictiva del modelo seleccionado

mediante la comparación entre los valores proyectados y los valores observados. Para esta etapa de validación, el modelo fue entrenado utilizando la información histórica disponible hasta diciembre de 2024 y posteriormente se generó una proyección para el período comprendido entre enero y mayo de 2025.

La figura 4.29 presenta la comparación entre los valores observados y proyectados de turbiedad para dicho período de validación, permitiendo analizar el grado de concordancia entre ambas series y la capacidad del modelo para reproducir la dinámica temporal de esta variable.

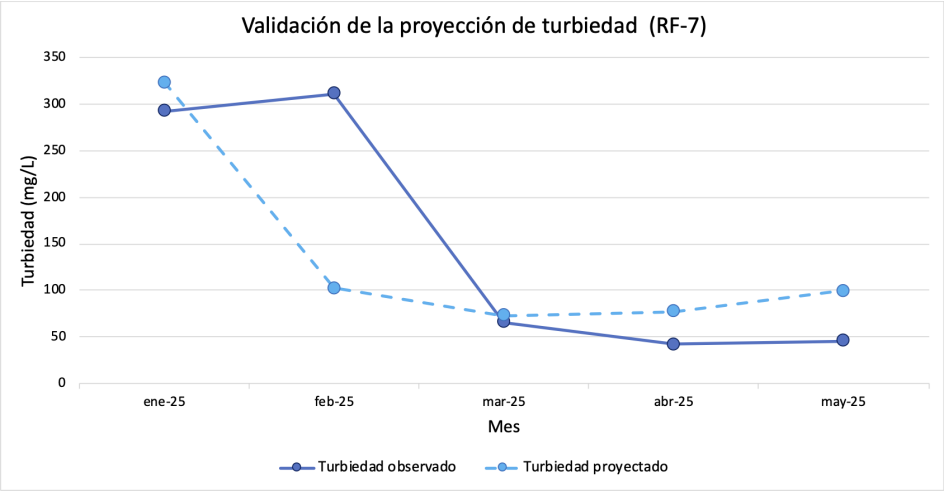

Figura 4.29: Comparación entre valores observados y proyectados de turbiedad (enero–mayo 2025).

A partir de la figura 4.29 se observa que el modelo RF-7 reproduce de manera adecuada la evolución general de la turbiedad durante el período de validación. En particular, se aprecia un buen ajuste entre los valores proyectados y observados en los meses de enero, marzo y abril, donde el modelo logra representar correctamente tanto el nivel como la tendencia de la serie. En el mes de mayo, si bien se observa una diferencia algo mayor entre ambos valores, esta se mantiene dentro de un rango acotado y no altera la dinámica general de la proyección.

La mayor discrepancia se presenta en el mes de febrero, donde el valor observado experimenta un aumento significativo que no es capturado por el modelo, el cual proyecta una disminución de la turbiedad. Este comportamiento sugiere la ocurrencia de un evento puntual o abrupto que no es explicado por las variables predictoras consideradas, lo cual es consistente con la naturaleza altamente variable de la turbiedad y su sensibilidad a fenómenos esporádicos, como episodios de arrastre de sedimentos o

condiciones hidrológicas excepcionales.

En conjunto, los resultados de validación indican que el modelo Random Forest con siete variables predictoras presenta un desempeño satisfactorio en la proyección de la turbiedad a corto plazo, capturando adecuadamente la tendencia general de la serie y los cambios más relevantes, con excepción de eventos puntuales de alta variabilidad. Esto respalda su utilización para el análisis de escenarios futuros y su incorporación en la evaluación integrada de la calidad del agua en la PTAP San Juan.

4.5. Comparación general

En esta sección se presenta una síntesis comparativa de los resultados obtenidos para caudal y los parámetros de calidad del agua modelados (SDT, NO_3^- y turbiedad) en la PTAP San Juan mediante Random Forest. El objetivo es resumir, de manera integrada, las configuraciones seleccionadas para cada variable y contrastar su desempeño predictivo, destacando similitudes y diferencias en cuanto a complejidad del modelo, variables relevantes y nivel de error alcanzado en las etapas de entrenamiento y validación.

En la tabla 4.1 se resumen las configuraciones finales adoptadas en cada caso, indicando el número de variables predictoras y el número de árboles del bosque, junto con las métricas MAE y MAPE empleadas como criterios de desempeño. Es importante señalar que el MAE corresponde a un error absoluto expresado en las mismas unidades de la variable modelada, por lo que se reporta en m^3/s para caudal y en mg/L para los contaminantes. En cambio, el MAPE se expresa en porcentaje (%), lo que permite evaluar el error relativo de forma comparable entre variables, particularmente en la validación con datos no utilizados durante el entrenamiento.

Tabla 4.1: Configuraciones seleccionadas del modelo Random Forest y errores asociados para cada variable analizada

Variable	**N° de variables predictoras**	**N° de árboles**	**MAE (entrenamiento)**	**MAPE % (validación)**
Caudal	8	200	9,0071	16,99
SDT	6	100	24,1277	7,26
NO_3^-	7	200	1,7691	10,13
Turbiedad	7	100	56,7102	57,77

A partir de los resultados en la tabla 4.1, se observa que el caudal fue modelado con una configuración de complejidad intermedia (8 variables, 200 árboles), coherente con su rol estructural dentro del sistema y con su utilización posterior como predictor en los modelos de contaminantes. En términos relativos, el MAPE del caudal (16,99 %) refleja una capacidad predictiva adecuada para una variable de alta variabilidad hidrológica, especialmente considerando que la validación se realiza fuera del conjunto de entrenamiento.

Para los contaminantes, se aprecian diferencias relevantes. En el caso de los SDT, se

obtiene el menor MAPE de validación (7,26 %), lo que indica un desempeño predictivo relativo superior en comparación con el resto de parámetros de calidad del agua. Este comportamiento es consistente con la mayor regularidad temporal de la serie y con la presencia de dependencia de corto plazo capturada por rezagos, además de la influencia del caudal como variable explicativa.

En el caso de los nitratos, la configuración seleccionada (7 variables, 200 árboles) presenta un error porcentual de validación intermedio (10,13 %). Este resultado sugiere que el modelo logra capturar adecuadamente la dinámica temporal predominante de la serie, la cual se encuentra fuertemente determinada por sus propios rezagos, complementada por aportes secundarios de variables hidrológicas y climáticas.

Por el contrario, la turbiedad presenta el mayor MAPE (57,77 %), lo cual evidencia una mayor dificultad del modelo para anticipar variaciones abruptas y episodios extremos, propios de una variable altamente sensible a eventos puntuales (por ejemplo, aumentos repentinos asociados a arrastre de sedimentos). En este sentido, aunque el modelo reproduce de manera satisfactoria el comportamiento en la mayor parte del período de validación, el desempeño relativo se ve afectado por meses con alta discrepancia, lo que incrementa el error porcentual.

En conjunto, la comparación general permite concluir que Random Forest entrega resultados consistentes para las variables analizadas, con configuraciones finales que equilibran complejidad e interpretabilidad. El desempeño predictivo relativo es más favorable para SDT y nitratos, mientras que la turbiedad presenta mayores desafíos debido a su naturaleza episódica y altamente variable. Estos resultados respaldan la utilización del enfoque propuesto como herramienta de apoyo para el análisis y la proyección de variables hidrológicas y de calidad del agua en el sistema San Juan, y proporcionan una base clara para la discusión final y las conclusiones del estudio.

Capítulo 5

Conclusiones generales

El presente trabajo tuvo como objetivo desarrollar un modelo predictivo basado en Random Forest que permitiera analizar y proyectar el comportamiento del caudal y de distintos parámetros de calidad del agua, sólidos disueltos totales (SDT), nitratos (NO_3^-) y turbiedad en la PTAP San Juan, integrando información histórica, hidrológica y climática relevante. A través de una metodología sistemática y consistente, se logró construir modelos capaces de reproducir adecuadamente la dinámica observada en las series temporales y generar proyecciones a horizontes de uno y cinco años, aportando información útil para el análisis y la planificación operativa.

Desde un punto de vista metodológico y teórico, los resultados obtenidos son coherentes con los fundamentos del modelo Random Forest presentados en el marco teórico. En particular, se evidenció su capacidad para capturar relaciones no lineales y dependencias complejas entre variables sin imponer una estructura funcional previa, lo que resulta especialmente adecuado para sistemas ambientales e hidrológicos. La incorporación explícita de rezagos temporales permitió adaptar el algoritmo a un contexto de series de tiempo, aprovechando la persistencia, estacionalidad y memoria temporal presentes tanto en el caudal como en los contaminantes analizados. Asimismo, el análisis de importancia de variables entregó un componente interpretativo relevante, permitiendo identificar los predictores con mayor contribución en cada modelo y respaldando empíricamente las relaciones observadas entre las variables.

En relación con los resultados obtenidos, el caudal fue modelado con una configuración de complejidad intermedia, lo que permitió reproducir adecuadamente su comportamiento histórico y generar proyecciones coherentes con la dinámica hidrológica del sistema. En el caso de los contaminantes, se observaron diferencias relevantes en

el desempeño predictivo del modelo. Los sólidos disueltos totales presentaron el mejor desempeño relativo, reflejado en un bajo error porcentual de validación, lo que es consistente con una dinámica más regular y una fuerte dependencia de corto plazo. Los nitratos mostraron un desempeño intermedio, dominado principalmente por su propia dinámica temporal, mientras que la turbiedad evidenció mayores dificultades de predicción, asociadas a su alta variabilidad y sensibilidad a eventos puntuales, los cuales no siempre pueden ser capturados por variables explicativas de carácter mensual.

Cabe destacar que el análisis se desarrolló utilizando series temporales mensualizadas, decisión que estuvo determinada por la disponibilidad y continuidad de los datos históricos, tanto para los contaminantes como para las variables externas consideradas. La mensualización permitió trabajar con series completas y coherentes en el tiempo, asegurando consistencia entre las distintas fuentes de información y evitando problemas asociados a datos faltantes. Esta elección metodológica representa un compromiso entre resolución temporal y calidad de los datos disponibles, privilegiando la estabilidad del proceso de modelación.

A partir del trabajo realizado, se identifican diversos desafíos y líneas de trabajo futuro que permitirían fortalecer y ampliar el alcance de este estudio:

- Actualizar y extender periódicamente las bases de datos de contaminantes, con el fin de ampliar la serie histórica disponible y mejorar la precisión de los modelos entrenados.

- Realizar procesos de validación más extensos a medida que se disponga de nuevos datos, evaluando el desempeño del modelo en distintos períodos y condiciones hidrológicas.

- Incorporar nuevas variables explicativas, tanto de carácter hidrológico, climático u operacional, que permitan capturar de mejor manera eventos extremos o procesos específicos no completamente representados en el modelo actual.

- Extender la aplicación de la metodología desarrollada a otras plantas de tratamiento o a otros parámetros de calidad del agua, evaluando la transferibilidad del enfoque y su utilidad en distintos contextos operacionales.

- Desarrollar modelos basados en series temporales diarias, una vez que se disponga de información suficiente, orientados a proyecciones de corto plazo con mayor resolución temporal y capacidad de anticipación frente a eventos críticos.

Finalmente, desde un punto de vista personal y profesional, el desarrollo de esta memoria constituyó una experiencia altamente enriquecedora. Este trabajo me permitió integrarme a un equipo de trabajo en conjunto con la empresa Esval, enfrentando un problema real de relevancia operativa y social, y aplicando de manera concreta los conocimientos adquiridos a lo largo de la formación universitaria en Ingeniería Civil Matemática. La posibilidad de trabajar con datos reales, comprender las necesidades de la empresa y contribuir al análisis de la calidad del agua como un recurso fundamental para la comunidad representó una instancia formativa valiosa, que fortaleció tanto mis competencias técnicas como mi capacidad de vinculación entre el ámbito académico y el profesional.

Anexos

Anexo A: Código para el análisis de correlaciones y rezagos

En este anexo se presenta el código desarrollado en Python para la realización del análisis exploratorio de correlaciones y rezagos temporales entre los contaminantes y las variables externas consideradas en el estudio. Este análisis constituye una etapa previa a la modelación predictiva y tiene por objetivo identificar asociaciones contemporáneas y diferidas en el tiempo, así como evidenciar patrones de persistencia temporal y estacionalidad en las series analizadas.

El código incluido permite generar los mapas de calor de correlación basados en el coeficiente de Pearson para el tiempo contemporáneo (t) y para los rezagos $t-1$, $t-2$, $t-3$ y $t-12$, los cuales son utilizados para fundamentar la selección de variables predictoras incorporadas en los modelos Random Forest. Este procedimiento se aplica de manera consistente a los distintos contaminantes y variables externas, por lo que el código presentado resulta representativo del análisis exploratorio realizado en el estudio.

```
1
2  # ====================================================
3  # MAPAS DE CALOR DE CORRELACIÓN
4  # Variables de calidad del agua y variables externas
5  # ====================================================
6
7  import os
8  import pandas as pd
9  import matplotlib.pyplot as plt
10 import seaborn as sns
```

```
11
12 # ======================================================
13 # 0) Configuración general
14 # ======================================================
15
16 # Ruta al escritorio
17 ruta_escritorio = os.path.join(os.path.expanduser("~"), "Desktop")
18
19 # Archivo de datos
20 archivo_datos = os.path.join(ruta_escritorio, "Datos para proyeccion.xlsx")
21
22 # Estilo de gráficos
23 sns.set(style="white")
24
25 # ======================================================
26 # 1) Carga y preprocesamiento de datos
27 # ======================================================
28
29 datos = pd.read_excel(archivo_datos)
30
31 if "Fecha" in datos.columns:
32     datos["Fecha"] = pd.to_datetime(datos["Fecha"], errors="coerce")
33     datos = datos.sort_values("Fecha")
34
35 # ======================================================
36 # 2) Selección y renombrado de variables
37 # ======================================================
38
39 mapa_columnas = {
40     "SDT": "SDT",
41     "NO3": "NO3",
42     "Turbiedad": "Turbiedad",
43     "Precipitaciones Arclim": "Precipitaciones",
44     "Temperatura Arclim": "Temperatura",
45     "Nieve Acumulada Arclim": "Nieve",
46     "Caudal real": "Caudal"
47 }
```

```
48
49 datos_variables =
   ↪ datos[list(mapa_columnas.keys())].rename(columns=mapa_columnas)
50
51 # ======================================================
52 # 3) Mapa de calor de correlaciones en t (sin rezagos)
53 # ======================================================
54
55 correlacion_t = datos_variables.corr(method="pearson")
56
57 plt.figure(figsize=(9, 7))
58 sns.heatmap(
59     correlacion_t,
60     annot=True,
61     fmt=".2f",
62     vmin=-1,
63     vmax=1,
64     cmap="coolwarm",
65     square=True
66 )
67 plt.title("Mapa de calor de correlaciones (variables en t)", fontsize=12)
68 plt.xticks(rotation=45, ha="right")
69 plt.yticks(rotation=0)
70 plt.tight_layout()
71 plt.show()
72
73 # ======================================================
74 # 4) Función: mapa de calor entre t y t-k
75 # ======================================================
76
77 def graficar_correlacion_con_rezago(datos_vars, rezago):
78     """
79     Genera un mapa de calor de correlaciones entre:
80     - Filas: variables en el tiempo t
81     - Columnas: variables en el tiempo t-rezago
82     """
83
```

```
    # Variables en t
    datos_actuales = datos_vars.copy()

    # Variables rezagadas
    datos_rezagados = datos_vars.shift(rezago)
    datos_rezagados.columns = [f"{col}_t-{rezago}" for col in
    ↪  datos_rezagados.columns]

    # Unión de ambas matrices
    datos_combinados = pd.concat(
        [datos_actuales, datos_rezagados],
        axis=1
    ).dropna()

    # Matriz de correlación cruzada
    correlacion_rezago = datos_combinados.corr().loc[
        datos_vars.columns,
        datos_rezagados.columns
    ]

    # Gráfico
    plt.figure(figsize=(9, 7))
    sns.heatmap(
        correlacion_rezago,
        annot=True,
        fmt=".2f",
        vmin=-1,
        vmax=1,
        cmap="coolwarm",
        square=True
    )
    plt.title(
        f"Mapa de calor de correlaciones entre t y t-{rezago}",
        fontsize=12
    )
    plt.xticks(rotation=45, ha="right")
    plt.yticks(rotation=0)
```

```
    plt.tight_layout()
    plt.show()

# ====================================================
# 5) Mapas de calor para distintos rezagos
# ====================================================

for k in [1, 2, 3, 12]:
    graficar_correlacion_con_rezago(datos_variables, rezago=k)


```

Anexo B: Código de implementación del modelo Random Forest

En este anexo se presenta el código desarrollado en Python para la implementación del modelo Random Forest aplicado al caso de los nitratos (NO_3^-), el cual se incluye a modo de referencia metodológica. Este código contempla las principales etapas del proceso de modelación predictiva, incluyendo la construcción de las variables predictoras con rezagos temporales, la selección del número óptimo de variables y de árboles del bosque, el análisis de importancia de variables mediante reducción de impureza, la validación del modelo a través del error porcentual absoluto medio (MAPE) y la generación de proyecciones futuras.

Cabe destacar que el procedimiento de modelación aplicado para el caudal, los sólidos disueltos totales (SDT) y la turbiedad sigue exactamente la misma estructura metodológica que la presentada en este anexo, diferenciándose únicamente en la variable objetivo considerada y en el conjunto específico de predictores asociados a cada caso. Por esta razón, y con el fin de evitar redundancias innecesarias, se optó por incluir únicamente el código correspondiente a nitratos, el cual resulta representativo del enfoque general utilizado en el estudio.

```

# ======================================================
# MODELO RANDOM FOREST PARA NITRATOS (NO3)
# ======================================================

import os
import time
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import matplotlib.dates as mdates

from sklearn.ensemble import RandomForestRegressor
from sklearn.metrics import mean_absolute_error


```

```python
# ============================================================
# 0) Configuración general
# ============================================================

# Ruta de archivos
escritorio = os.path.join(os.path.expanduser("~"), "Desktop")
ruta_excel = os.path.join(escritorio, "Datos para proyeccion.xlsx")

# Fechas (histórico y corte)
fecha_inicio_historico = pd.Timestamp("2021-01-01")
fecha_corte            = pd.Timestamp("2025-05-31")

# Unidades y horizonte futuro
unidad_no3 = "mg/L"
horizonte_futuro_anios = 5
horizonte_futuro_meses = 12 * horizonte_futuro_anios

# Colores para series
COLOR_OBS    = "#1f4e79"   # Histórico (observado)
COLOR_AJUSTE = "#f39c12"   # Ajuste RF
COLOR_PROY   = "#117a8b"   # Proyección
COLOR_BANDA  = "#aed6f1"   # Banda IC 95%
ALPHA_BANDA  = 0.35

# Color barras importancia
COLOR_IMP = "#55aaff"

# Paleta para curva MAE vs número de variables
COLORES_CURVAS = [
    "#486581",
    "#55aaff",
    "#0057ff",
    "#7d7d7d"
]

# Configuración de IC 95%
z_95 = 1.96
```

```
54
55  # ============================================================
56  # 1) Carga y limpieza de datos
57  # ============================================================
58
59  datos = pd.read_excel(ruta_excel)
60
61  datos["Fecha"] = pd.to_datetime(datos["Fecha"], format="%d-%m-%y",
    ↪  errors="coerce")
62  datos.loc[datos["Fecha"].dt.year < 2000, "Fecha"] = datos["Fecha"] +
    ↪  pd.DateOffset(years=100)
63  datos = datos.dropna(subset=["Fecha"]).sort_values("Fecha")
64
65  datos = datos.rename(columns={
66      "Precipitaciones Arclim": "Precipitaciones",
67      "Temperatura Arclim":     "Temperatura",
68      "Nieve Acumulada Arclim": "Nieve",
69      "Caudal (8 var ext)":     "Caudal"
70  })
71
72  # Convertir a numérico
73  columnas_numericas = ["NO3", "Caudal", "Precipitaciones", "Temperatura",
    ↪  "Nieve"]
74  for col in columnas_numericas:
75      datos[col] = (
76          pd.to_numeric(datos[col], errors="coerce")
77            .interpolate("linear")
78            .ffill()
79            .bfill()
80      )
81
82  # ============================================================
83  # 2) Mensualización (MS) + interpolación
84  # ============================================================
85
86  datos_mensuales = (
```

```
87      datos.groupby("Fecha")[["NO3", "Caudal", "Precipitaciones",
        ↪  "Temperatura", "Nieve"]]
88           .mean()
89           .sort_index()
90  )
91
92  # Frecuencia mensual (inicio de mes) y completar faltantes con interpolación
93  datos_mensuales = datos_mensuales.asfreq("MS").interpolate("linear")
94
95  # ============================================================
96  # 3) Creación de rezagos (lags)
97  # ============================================================
98
99  def agregar_lags(tabla: pd.DataFrame, columna_base: str, rezagos: tuple) ->
    ↪  pd.DataFrame:
100     """
101     Agrega columnas rezagadas del tipo: <columna_base>_lagL
102     """
103     for L in rezagos:
104         tabla[f"{columna_base}_lag{L}"] = tabla[columna_base].shift(L)
105     return tabla
106
107 # Variables externas: t, t-1, t-2, t-3, t-12
108 for var_base in ["Precipitaciones", "Temperatura", "Nieve", "Caudal"]:
109     datos_mensuales = agregar_lags(datos_mensuales, var_base, rezagos=(0, 1,
        ↪  2, 3, 12))
110
111 # NO3 autoregresivo: t-1, t-2, t-3, t-12
112 datos_mensuales = agregar_lags(datos_mensuales, "NO3", rezagos=(1, 2, 3, 12))
113
114 # ============================================================
115 # 4) Definición de features y conjunto histórico
116 # ============================================================
117
118 columnas_X = [
119     # Precipitaciones
```

```
    "Precipitaciones_lag0", "Precipitaciones_lag1", "Precipitaciones_lag2",
    ↪ "Precipitaciones_lag3", "Precipitaciones_lag12",
    # Temperatura
    "Temperatura_lag0", "Temperatura_lag1", "Temperatura_lag2",
    ↪ "Temperatura_lag3", "Temperatura_lag12",
    # Nieve
    "Nieve_lag0", "Nieve_lag1", "Nieve_lag2", "Nieve_lag3", "Nieve_lag12",
    # Caudal
    "Caudal_lag0", "Caudal_lag1", "Caudal_lag2", "Caudal_lag3",
    ↪ "Caudal_lag12",
    # NO3 autoregresivo
    "NO3_lag1", "NO3_lag2", "NO3_lag3", "NO3_lag12"
]

mascara_historico = (datos_mensuales.index >= fecha_inicio_historico) &
↪ (datos_mensuales.index <= fecha_corte)

X_historico = datos_mensuales.loc[mascara_historico, columnas_X].dropna()
y_historico = datos_mensuales.loc[X_historico.index, "NO3"]

# Índice futuro (hasta 5 años post-corte)
indice_futuro = datos_mensuales.index[
    (datos_mensuales.index > fecha_corte) &
    (datos_mensuales.index <= fecha_corte +
    ↪ pd.DateOffset(years=horizonte_futuro_anios))
]

# ============================================================
# 5) RF-24: MAE vs Número de árboles
# ============================================================

print("\n=== Evaluación del MAE según el número de árboles para NO3 - Modelo
↪ RF-24 ===")

lista_arboles = [50, 100, 200, 300, 400, 500, 600, 1000]
lista_mae_rf24 = []

```

```python
for n_arboles in lista_arboles:
    modelo_tmp = RandomForestRegressor(
        n_estimators=n_arboles,
        max_depth=None,
        min_samples_leaf=1,
        random_state=42,
        n_jobs=-1
    )
    modelo_tmp.fit(X_historico, y_historico)
    pred_tmp = modelo_tmp.predict(X_historico)
    mae_tmp = mean_absolute_error(y_historico, pred_tmp)

    lista_mae_rf24.append(mae_tmp)
    print(f"n_estimators = {n_arboles:4d} -> MAE = {mae_tmp:0.4f} {unidad_no3}")

idx_mejor_rf24 = int(np.argmin(lista_mae_rf24))
mejor_n_rf24 = lista_arboles[idx_mejor_rf24]
print(f"\n Mejor número de árboles (Modelo RF-24): {mejor_n_rf24} "
      f"(MAE = {lista_mae_rf24[idx_mejor_rf24]:0.4f} {unidad_no3})")

plt.figure(figsize=(8, 5))
plt.plot(lista_arboles, lista_mae_rf24, marker="o")
plt.xlabel("Número de árboles")
plt.ylabel(f"Error MAE ({unidad_no3})")
plt.title("Evaluación del MAE según el número de árboles para nitratos - Modelo RF-24")
plt.grid(True)
plt.tight_layout()
plt.show()

cantidad_arboles = 200

# ============================================================
# 6) RF-24: Entrenamiento + Ajuste histórico + Proyección a 5 años
# ============================================================

```

```
modelo_rf24 = RandomForestRegressor(
    n_estimators=cantidad_arboles,
    max_depth=None,
    min_samples_leaf=1,
    random_state=42,
    n_jobs=-1
)
modelo_rf24.fit(X_historico, y_historico)
arboles_rf24 = modelo_rf24.estimators_

# Ajuste histórico
preds_hist_arboles = np.stack([arb.predict(X_historico) for arb in
↪  arboles_rf24])
media_hist = preds_hist_arboles.mean(axis=0)
std_hist   = preds_hist_arboles.std(axis=0)

ajuste_hist_rf24 = pd.Series(np.clip(media_hist, a_min=0, a_max=None),
↪  index=X_historico.index)
mae_rf24 = mean_absolute_error(y_historico, ajuste_hist_rf24)

# ============================================================
# 7) RF-24: Importancia por impureza
# ============================================================

importancia_rf24 = pd.Series(modelo_rf24.feature_importances_,
↪  index=columnas_X).sort_values(ascending=False)
importancia_pct_rf24 = (100 * importancia_rf24 /
↪  importancia_rf24.sum()).round(2)

variable_top1 = importancia_pct_rf24.index[0]
print(f"\n Variable más importante (RF-24): {variable_top1}")

print("\n=== Importancia relativa por impureza (%) - MODELO RF-24 (Top-10)
↪  ===")
print(importancia_pct_rf24.head(10).to_string())

p = importancia_pct_rf24.head(23)[::-1]
```

```
plt.figure(figsize=(10, 7))
plt.barh(p.index, p.values, color=COLOR_IMP)
plt.title("Importancia relativa por impureza - Modelo RF-24")
plt.xlabel("Importancia (%)")
plt.ylabel("Variables")
plt.grid(axis="x", alpha=0.3)
plt.tight_layout()
plt.show()

# ============================================================
# 8) RF-24: MAE vs Número de variables explicativas
# ============================================================

print("\n=== Curva MAE vs número de variables y número de árboles - NO3
↪  (RF-24) ===")

features_ordenadas = list(importancia_pct_rf24.index)   # de mayor a menor
lista_arboles_curva = [50, 100, 200, 500]

resultados_mae_k = []

for ntrees in lista_arboles_curva:
    print(f"\n>>> Probando n_estimators = {ntrees}\n")
    for k in range(1, len(features_ordenadas) + 1):
        feats_k = features_ordenadas[:k]
        X_k = datos_mensuales.loc[X_historico.index, feats_k]

        modelo_k = RandomForestRegressor(
            n_estimators=ntrees,
            max_depth=None,
            min_samples_leaf=1,
            random_state=42,
            n_jobs=-1
        )
        modelo_k.fit(X_k, y_historico)
        pred_k = modelo_k.predict(X_k)
        mae_k = mean_absolute_error(y_historico, pred_k)
```

```
254
255         resultados_mae_k.append({
256             "variables": k,
257             "n_estimators": ntrees,
258             "MAE": mae_k
259         })
260
261         print(f"[ntrees={ntrees:3d}] TOP-{k:2d} vars -> MAE = {mae_k:0.4f}
        ↪  {unidad_no3}")
262
263 tabla_mae_k = pd.DataFrame(resultados_mae_k)
264 print("\n=== TABLA COMPLETA MAE(k, árboles) ===")
265 print(tabla_mae_k)
266
267 plt.figure(figsize=(10, 6))
268 for i, ntrees in enumerate(lista_arboles_curva):
269     mask = tabla_mae_k.n_estimators == ntrees
270     plt.plot(
271         tabla_mae_k[mask].variables,
272         tabla_mae_k[mask].MAE,
273         marker="o",
274         color=COLORES_CURVAS[i % len(COLORES_CURVAS)],
275         label=f"{ntrees} árboles"
276     )
277
278 plt.xlabel("Número de variables")
279 plt.ylabel(f"Error MAE ({unidad_no3})")
280 plt.title("Evaluación del MAE según el número de variables explicativas -
    ↪  Nitratos")
281 plt.legend(title="Cantidad de árboles", frameon=False)
282 plt.grid(True, alpha=0.3)
283 plt.tight_layout()
284 plt.show()
285
286 # ============================================================
287 # 9) RF-24: Proyección futura (recursiva) + IC 95%
288 # ============================================================
```

```
289
290 # Serie extendida para NO3 (reemplaza histórico por ajuste para que sea
    ↪ consistente recursivamente)
291 serie_no3_ext = datos_mensuales["NO3"].copy()
292 serie_no3_ext.loc[ajuste_hist_rf24.index] = ajuste_hist_rf24.values
293
294 lista_media_fut, lista_ic_inf, lista_ic_sup = [], [], []
295
296 for fecha in indice_futuro:
297     fila = {
298         # NO3 autoregresivo
299         "NO3_lag1":  serie_no3_ext.get(fecha - pd.DateOffset(months=1),
            ↪ np.nan),
300         "NO3_lag2":  serie_no3_ext.get(fecha - pd.DateOffset(months=2),
            ↪ np.nan),
301         "NO3_lag3":  serie_no3_ext.get(fecha - pd.DateOffset(months=3),
            ↪ np.nan),
302         "NO3_lag12": serie_no3_ext.get(fecha - pd.DateOffset(months=12),
            ↪ np.nan),
303
304         # Precipitaciones
305         "Precipitaciones_lag0":  datos_mensuales.at[fecha,
            ↪ "Precipitaciones"],
306         "Precipitaciones_lag1":
            ↪ datos_mensuales["Precipitaciones"].shift(1).get(fecha,  np.nan),
307         "Precipitaciones_lag2":
            ↪ datos_mensuales["Precipitaciones"].shift(2).get(fecha,  np.nan),
308         "Precipitaciones_lag3":
            ↪ datos_mensuales["Precipitaciones"].shift(3).get(fecha,  np.nan),
309         "Precipitaciones_lag12":
            ↪ datos_mensuales["Precipitaciones"].shift(12).get(fecha, np.nan),
310
311         # Temperatura
312         "Temperatura_lag0":  datos_mensuales.at[fecha, "Temperatura"],
313         "Temperatura_lag1":
            ↪ datos_mensuales["Temperatura"].shift(1).get(fecha,  np.nan),
```

```
314         "Temperatura_lag2":
            ↪ datos_mensuales["Temperatura"].shift(2).get(fecha,  np.nan),
315         "Temperatura_lag3":
            ↪ datos_mensuales["Temperatura"].shift(3).get(fecha,  np.nan),
316         "Temperatura_lag12":
            ↪ datos_mensuales["Temperatura"].shift(12).get(fecha, np.nan),
317
318         # Nieve
319         "Nieve_lag0":  datos_mensuales.at[fecha, "Nieve"],
320         "Nieve_lag1":  datos_mensuales["Nieve"].shift(1).get(fecha,  np.nan),
321         "Nieve_lag2":  datos_mensuales["Nieve"].shift(2).get(fecha,  np.nan),
322         "Nieve_lag3":  datos_mensuales["Nieve"].shift(3).get(fecha,  np.nan),
323         "Nieve_lag12": datos_mensuales["Nieve"].shift(12).get(fecha, np.nan),
324
325         # Caudal
326         "Caudal_lag0":  datos_mensuales.at[fecha, "Caudal"],
327         "Caudal_lag1":  datos_mensuales["Caudal"].shift(1).get(fecha,
            ↪ np.nan),
328         "Caudal_lag2":  datos_mensuales["Caudal"].shift(2).get(fecha,
            ↪ np.nan),
329         "Caudal_lag3":  datos_mensuales["Caudal"].shift(3).get(fecha,
            ↪ np.nan),
330         "Caudal_lag12": datos_mensuales["Caudal"].shift(12).get(fecha,
            ↪ np.nan),
331     }
332
333     x_fila = pd.DataFrame([fila], columns=columnas_X)
334
335     pred_todos = np.array([arb.predict(x_fila)[0] for arb in arboles_rf24])
336     media = pred_todos.mean()
337     std   = pred_todos.std()
338
339     lista_media_fut.append(max(0.0, media))
340     lista_ic_inf.append(max(0.0, media - z_95 * std))
341     lista_ic_sup.append(media + z_95 * std)
342
343     # actualizar serie extendida para recursividad
```

```
344     serie_no3_ext.loc[fecha] = media
345
346 proy_media_rf24 = pd.Series(lista_media_fut, index=indice_futuro)
347 proy_ic_inf_rf24 = pd.Series(lista_ic_inf, index=indice_futuro)
348 proy_ic_sup_rf24 = pd.Series(lista_ic_sup, index=indice_futuro)
349
350 # ============================================================
351 # 10) Gráfica proyección 5 años - RF-24
352 # ============================================================
353
354 plt.figure(figsize=(12, 6))
355
356 serie_obs_hist = datos_mensuales.loc[
357     (datos_mensuales.index >= fecha_inicio_historico) &
        ↪  (datos_mensuales.index <= fecha_corte),
358     "NO3"
359 ]
360
361 plt.plot(serie_obs_hist.index, serie_obs_hist.values,
362          label="Histórico (observado)", color=COLOR_OBS, linewidth=2)
363
364 plt.plot(ajuste_hist_rf24.index, ajuste_hist_rf24.values,
365          label="Ajuste del modelo (histórico)", color=COLOR_AJUSTE,
           ↪  linestyle="--", linewidth=2)
366
367 plt.plot(proy_media_rf24.index, proy_media_rf24.values,
368          label="Proyección 5 años", color=COLOR_PROY, linewidth=2)
369
370 plt.fill_between(proy_media_rf24.index, proy_ic_inf_rf24.values,
    ↪  proy_ic_sup_rf24.values,
371                  color=COLOR_BANDA, alpha=ALPHA_BANDA, label="Intervalo de
                     ↪  confianza 95%")
372
373 plt.title("Proyección de nitratos a 5 años - Modelo RF-24")
374 plt.xlabel("Año")
375 plt.ylabel(f"NO3 ({unidad_no3})")
376 plt.grid(True)
```

```
plt.legend()

ax = plt.gca()
ax.xaxis.set_major_formatter(mdates.DateFormatter("%Y"))
ax.xaxis.set_major_locator(mdates.YearLocator(base=1))
plt.xticks(rotation=45)

plt.tight_layout()
plt.show()

# ============================================================
# 11) Gráfica proyección 1 año - RF-24
# ============================================================

ventana_contexto = 24
inicio_contexto = serie_obs_hist.index.max() -
↪  pd.DateOffset(months=ventana_contexto - 1)

obs_ctx = serie_obs_hist[serie_obs_hist.index >= inicio_contexto]
fit_ctx = ajuste_hist_rf24[ajuste_hist_rf24.index >= inicio_contexto]

horizonte_1_anio = 12
idx_1y = proy_media_rf24.index[:horizonte_1_anio]

proy_1y = proy_media_rf24.loc[idx_1y]
ic1y_inf = proy_ic_inf_rf24.loc[idx_1y]
ic1y_sup = proy_ic_sup_rf24.loc[idx_1y]

plt.figure(figsize=(12, 6))
plt.plot(obs_ctx.index, obs_ctx.values, label="Histórico (observado)",
↪  color=COLOR_OBS, linewidth=2)
plt.plot(fit_ctx.index, fit_ctx.values, label="Ajuste del modelo
↪  (histórico)", color=COLOR_AJUSTE, linestyle="--", linewidth=2)

plt.plot(proy_1y.index, proy_1y.values, label="Proyección 1 año",
↪  color=COLOR_PROY, linewidth=2)
```

```
plt.fill_between(proy_1y.index, ic1y_inf.values, ic1y_sup.values,
    color=COLOR_BANDA, alpha=ALPHA_BANDA, label="Intervalo de confianza 95%")

plt.title("Proyección de nitratos a 1 año - Modelo RF-24")
plt.xlabel("Mes")
plt.ylabel(f"NO3 ({unidad_no3})")
plt.grid(True)
plt.legend()

ax = plt.gca()
ax.xaxis.set_major_formatter(mdates.DateFormatter("%Y-%m"))
ax.xaxis.set_major_locator(mdates.MonthLocator(interval=2))
plt.xticks(rotation=45)

plt.tight_layout()
plt.show()

# ============================================================
# 12) Exportar a Excel (RF-24)
# ============================================================

no3_obs_ajuste = datos_mensuales.loc[ajuste_hist_rf24.index, "NO3"]

ruta_out_hist_rf24 = os.path.join(escritorio, "Ajuste_NO3_RF24.xlsx")
ruta_out_fut_rf24  = os.path.join(escritorio, "Proyeccion_NO3_RF24.xlsx")

pd.DataFrame({
    "Fecha": ajuste_hist_rf24.index,
    "NO3_Observado": no3_obs_ajuste.values,
    "NO3_Ajuste_RF24": ajuste_hist_rf24.values
}).to_excel(ruta_out_hist_rf24, index=False)

pd.DataFrame({
    "Fecha": proy_media_rf24.index,
    "NO3_Proyectado": proy_media_rf24.values,
    "IC95_inf": proy_ic_inf_rf24.values,
    "IC95_sup": proy_ic_sup_rf24.values
```

```
}).to_excel(ruta_out_fut_rf24, index=False)

print(f"\nExcel guardado (histórico ajustado NO3 - RF-24):
↪ {ruta_out_hist_rf24}")
print(f" Excel guardado (proyección 5 años NO3 - RF-24):
↪ {ruta_out_fut_rf24}")

# ============================================================
# 13) RF-4 (Top-4 variables según importancia RF-24)
# ============================================================

print("\n\n==============================")
print("     MODELO RF-4 (TOP-4)     ")
print("==============================")

top4_features = list(importancia_pct_rf24.index[:4])
print("\nVariables usadas en RF-4 (Top-4):")
for v in top4_features:
    print(" -", v)

imp_top4 = importancia_pct_rf24.loc[top4_features]
print("\nImportancia relativa por impureza (%) - MODELO RF-4 (según RF-24)")
print(imp_top4.to_string())

plt.figure(figsize=(7, 5))
plt.barh(imp_top4.index[::-1], imp_top4.values[::-1], color=COLOR_IMP)
plt.xlabel("Importancia (%)")
plt.ylabel("Variables")
plt.title("Importancia relativa por impureza - Modelo RF-4")
plt.grid(axis="x", alpha=0.3)
plt.tight_layout()
plt.show()

X_hist_rf4 = datos_mensuales.loc[X_historico.index, top4_features]

print("\n=== MAE vs número de árboles (Modelo RF-4) ===")
lista_mae_rf4 = []
```

```
mejor_n_rf4 = None
mejor_mae_rf4 = np.inf

for n_arboles in lista_arboles:
    modelo_tmp = RandomForestRegressor(
        n_estimators=n_arboles,
        max_depth=None,
        min_samples_leaf=1,
        random_state=42,
        n_jobs=-1
    )
    modelo_tmp.fit(X_hist_rf4, y_historico)
    pred_tmp = modelo_tmp.predict(X_hist_rf4)
    mae_tmp = mean_absolute_error(y_historico, pred_tmp)

    lista_mae_rf4.append(mae_tmp)

    if mae_tmp < mejor_mae_rf4:
        mejor_mae_rf4 = mae_tmp
        mejor_n_rf4 = n_arboles

    print(f"n_estimators = {n_arboles:4d} -> MAE (RF-4) = {mae_tmp:0.4f}
    ↪  {unidad_no3}")

print(f"\n Mejor número de árboles (Modelo RF-4): {mejor_n_rf4} "
      f"(MAE = {mejor_mae_rf4:0.4f} {unidad_no3})")

plt.figure(figsize=(8, 5))
plt.plot(lista_arboles, lista_mae_rf4, marker="o")
plt.xlabel("Número de árboles")
plt.ylabel(f"Error MAE ({unidad_no3})")
plt.title("Evaluación del MAE según el número de árboles para nitratos -
↪  Modelo RF-4")
plt.grid(True)
plt.tight_layout()
plt.show()
```

```
modelo_rf4 = RandomForestRegressor(
    n_estimators=cantidad_arboles,
    max_depth=None,
    min_samples_leaf=1,
    random_state=42,
    n_jobs=-1
)
modelo_rf4.fit(X_hist_rf4, y_historico)
arboles_rf4 = modelo_rf4.estimators_

preds_hist_rf4 = np.stack([arb.predict(X_hist_rf4) for arb in arboles_rf4])
media_hist_rf4 = preds_hist_rf4.mean(axis=0)
std_hist_rf4   = preds_hist_rf4.std(axis=0)

ajuste_hist_rf4 = pd.Series(np.clip(media_hist_rf4, a_min=0, a_max=None),
    index=X_hist_rf4.index)
mae_rf4 = mean_absolute_error(y_historico, ajuste_hist_rf4)

# Proyección recursiva RF-4
serie_ext_rf4 = datos_mensuales["NO3"].copy()
serie_ext_rf4.loc[ajuste_hist_rf4.index] = ajuste_hist_rf4.values

lista_media_fut_rf4, lista_ic_inf_rf4, lista_ic_sup_rf4 = [], [], []

for fecha in indice_futuro:
    fila = {}
    for feat in top4_features:
        base, lag_str = feat.split("_lag")
        lag = int(lag_str)

        if base == "NO3":
            val = serie_ext_rf4.shift(lag).get(fecha, np.nan)
        else:
            val = datos_mensuales[base].shift(lag).get(fecha, np.nan) if lag
            > 0 else datos_mensuales.at[fecha, base]

        fila[feat] = val
```

```
550
551     x_fila = pd.DataFrame([fila], columns=top4_features)
552
553     pred_todos = np.array([arb.predict(x_fila)[0] for arb in arboles_rf4])
554     media = pred_todos.mean()
555     std   = pred_todos.std()
556
557     lista_media_fut_rf4.append(max(0.0, media))
558     lista_ic_inf_rf4.append(max(0.0, media - z_95 * std))
559     lista_ic_sup_rf4.append(media + z_95 * std)
560
561     serie_ext_rf4.loc[fecha] = media
562
563 proy_media_rf4 = pd.Series(lista_media_fut_rf4, index=indice_futuro)
564 proy_ic_inf_rf4 = pd.Series(lista_ic_inf_rf4, index=indice_futuro)
565 proy_ic_sup_rf4 = pd.Series(lista_ic_sup_rf4, index=indice_futuro)
566
567 # Gráficas RF-4
568 serie_obs_hist_rf4 = datos_mensuales.loc[(datos_mensuales.index >=
    ↪  fecha_inicio_historico) & (datos_mensuales.index <= fecha_corte), "NO3"]
569
570 plt.figure(figsize=(12, 6))
571 plt.plot(serie_obs_hist_rf4.index, serie_obs_hist_rf4.values,
    ↪  label="Histórico (observado)", color=COLOR_OBS)
572 plt.plot(ajuste_hist_rf4.index, ajuste_hist_rf4.values, "--",
    ↪  color=COLOR_AJUSTE, label="Ajuste del modelo (histórico)")
573 plt.plot(proy_media_rf4.index, proy_media_rf4.values, label="Proyección a 5
    ↪  años", color=COLOR_PROY)
574 plt.fill_between(proy_media_rf4.index, proy_ic_inf_rf4.values,
    ↪  proy_ic_sup_rf4.values, color=COLOR_BANDA, alpha=ALPHA_BANDA,
    ↪  label="Intervalo de confianza 95%")
575 plt.title("Proyección de nitratos a 5 años - Modelo RF-4")
576 plt.grid()
577 plt.legend()
578 plt.tight_layout()
579 plt.show()
580
```

```
ventana_contexto_rf4 = 24
inicio_ctx_rf4 = serie_obs_hist_rf4.index.max() - pd.DateOffset(months=ventana_contexto_rf4 - 1)
obs_ctx_rf4 = serie_obs_hist_rf4[serie_obs_hist_rf4.index >= inicio_ctx_rf4]
fit_ctx_rf4 = ajuste_hist_rf4[ajuste_hist_rf4.index >= inicio_ctx_rf4]

plt.figure(figsize=(12, 6))
plt.plot(obs_ctx_rf4.index, obs_ctx_rf4.values, label="Histórico (observado)", color=COLOR_OBS)
plt.plot(fit_ctx_rf4.index, fit_ctx_rf4.values, "--", color=COLOR_AJUSTE, label="Ajuste del modelo (histórico)")
plt.plot(proy_media_rf4.index[:12], proy_media_rf4.values[:12], label="Proyección a 1 año", color=COLOR_PROY)
plt.fill_between(proy_media_rf4.index[:12], proy_ic_inf_rf4.values[:12], proy_ic_sup_rf4.values[:12], color=COLOR_BANDA, alpha=ALPHA_BANDA, label="Intervalo de confianza 95%")
plt.title("Proyección de nitratos a 1 año - Modelo RF-4")
plt.grid()
plt.legend()
plt.tight_layout()
plt.show()

# Exportar RF-4
ruta_out_hist_rf4 = os.path.join(escritorio, "Ajuste_NO3_RF4_Top4.xlsx")
ruta_out_fut_rf4  = os.path.join(escritorio, "Proyeccion_NO3_RF4_Top4.xlsx")

pd.DataFrame({
    "Fecha": ajuste_hist_rf4.index,
    "NO3_Observado": serie_obs_hist_rf4.loc[ajuste_hist_rf4.index].values,
    "NO3_Ajuste_RF4": ajuste_hist_rf4.values
}).to_excel(ruta_out_hist_rf4, index=False)

pd.DataFrame({
    "Fecha": proy_media_rf4.index,
    "NO3_Proyectado_RF4": proy_media_rf4.values,
    "IC95_inf_RF4": proy_ic_inf_rf4.values,
    "IC95_sup_RF4": proy_ic_sup_rf4.values
```

```
612 }).to_excel(ruta_out_fut_rf4, index=False)
613
614 print(f"\ Excel guardado RF-4 (histórico): {ruta_out_hist_rf4}")
615 print(f" Excel guardado RF-4 (proyección): {ruta_out_fut_rf4}")
616
617
```

Bibliografía

Bishop, C. M. (2006). *Pattern Recognition and Machine Learning*. Springer.

Breiman, L. (1996). Bagging predictors. *Machine Learning*, 24(2):123–140.

Breiman, L. (2001). Random forests. *Machine Learning*, 45(1):5–32.

Breiman, L., Friedman, J. H., Olshen, R. A., and Stone, C. J. (1984). *Classification and Regression Trees*. Chapman & Hall/CRC, Boca Raton.

De'ath, G. and Fabricius, K. E. (2000). Classification and regression trees: A powerful yet simple technique for ecological data analysis. *Ecology*, 81(11):3178–3192.

Desconocido, A. (s.f.). Árboles de decisión. Documento en formato PDF.

Hastie, T., Tibshirani, R., and Friedman, J. (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer, 2 edition.

Hill, J., Linero, A., and Murray, J. (2020). Bayesian additive regression trees: A review and look forward. *Annual Review of Statistics and Its Application*, 7:251–278.

James, G., Witten, D., Hastie, T., Tibshirani, R., and Taylor, J. (2023). *An Introduction to Statistical Learning: With Applications in Python*. Springer, Cham.

Murphy, K. P. (2012). *Machine Learning: A Probabilistic Perspective*. MIT Press.

Murphy, K. P. (2022). *Probabilistic Machine Learning: An Introduction*. MIT Press.