



UNIVERSIDAD DE CHILE  
FACULTAD DE CIENCIAS FÍSICAS Y MATEMÁTICAS  
DEPARTAMENTO DE INGENIERÍA CIVIL

METODOLOGÍA PARA EL RELLENO DE REGISTROS FLUVIOMÉTRICOS DIARIOS  
EN CUENCAS CHILENAS CERCANAS AL RÉGIMEN NATURAL

MEMORIA PARA OPTAR AL TÍTULO DE  
INGENIERO CIVIL

MARIO ALBERTO ISAAC QUEVEDO LEIVA

PROFESOR GUÍA:  
PABLO MENDOZA ZÚÑIGA

MIEMBROS DE LA COMISIÓN:  
NICOLÁS VÁSQUEZ PLACENCIA  
XIMENA VARGAS MESA

SANTIAGO DE CHILE  
2021

RESUMEN DE LA MEMORIA PARA OPTAR  
AL TÍTULO DE INGENIERO CIVIL  
POR: MARIO ALBERTO ISAAC QUEVEDO LEIVA  
FECHA: MARZO 2021  
PROF. GUÍA: PABLO MENDOZA ZÚÑIGA

## METODOLOGÍA PARA EL RELLENO DE REGISTROS FLUVIOMÉTRICOS DIARIOS EN CUENCAS CHILENAS CERCANAS AL RÉGIMEN NATURAL

La disponibilidad de series continuas de caudal a lo largo del territorio nacional es esencial para cualquier caracterización hidrológica que busque sustentar la toma de decisiones respecto al uso del recurso hídrico. El objetivo general de esta memoria es desarrollar una metodología que permita rellenar vacíos en los registros fluviométricos del país, a diferentes escalas temporales, para estaciones fluviométricas con baja intervención humana.

En primera instancia, se entrega un diagnóstico general de la cobertura temporal de 158 estaciones fluviométricas en Chile. Posteriormente, se rellenan datos faltantes de caudales medios anuales, estacionales y mensuales mediante árboles de decisión y regresiones lineales con estaciones vecinas, probando siete criterios diferentes para establecer similitud y escoger una cuenca donante. En relación con los caudales medios anuales, el criterio óptimo para rellenar datos faltantes en las zonas norte y sur del país incorpora sólo atributos físicos, tales como elevación media, pendiente media, fracción de bosque, fracción de pasto, fracción de arbusto, clase geológica más común, tierras de cultivo, falta de cobertura terrestre y porcentaje de área nival. Las métricas resultantes de la validación cruzada son,  $R^2 = 0,97$  y  $\text{NRMSE} = 0,16$ . La estimación de caudales medios estacionales muestra que para los periodos de enero a marzo y de abril a junio, el criterio más utilizado comprende sólo similitud hidrológica, definida como la distancia euclidiana entre coeficientes de Pardé. Para el periodo julio a septiembre, el criterio más utilizado incorpora sólo atributos climáticos (índice de aridez, estacionalidad de la precipitación y fracción de eventos de precipitación que caen como nieve). Para el periodo octubre a diciembre, el criterio más utilizado considera sólo atributos físicos. Para los periodos enero a marzo y julio a septiembre se obtienen las mejores (e iguales) métricas de validación cruzada ( $R^2 = 0,98$  y  $\text{NRMSE} = 0,11$ ). En relación con los caudales medios mensuales, el mes para el cual se obtienen los mejores resultados es enero ( $R^2 = 0,98$ , en validación cruzada), mientras que mayo es el mes con los peores resultados ( $R^2 = 0,90$ ).

Finalmente, se utiliza el método no paramétrico *K- nearest neighbors* para estimar caudales medios diarios en cuatro cuencas: Estero Yerba Loca antes junta San Francisco, Río Aconcagua en Chacabuquito, Río Andalién camino a Penco y Río Lirquén en Cerro el Padre. Los resultados de la validación cruzada indican que al condicionar el método utilizando el caudal medio mensual como variable independiente, se obtienen buenos resultados para las estaciones Estero Yerba Loca antes junta San Francisco ( $R^2 = 0,85$ ) y Río Aconcagua en Chacabuquito ( $R^2 = 0,87$ ). Para las estaciones Río Andalién camino a Penco y Río Lirquén en Cerro el Padre, se incorporan además la precipitación y temperatura, y los resultados son regulares ( $R^2 = 0,38$  y  $R^2 = 0,44$  respectivamente). En las cuatro estaciones utilizadas, se obtienen mejores resultados (mayor  $R^2$ ) en el periodo octubre a marzo, debido a que, durante estos meses, los caudales no tienen cambios abruptos. Se concluye que la metodología propuesta tiene un mejor rendimiento entre estos meses, independiente del régimen hidrológico de la cuenca.



*A toda mi familia por su apoyo incondicional.*

# Agradecimientos

En primer lugar, quiero agradecer a mis padres Ginette y Saúl, por quererme, cuidarme y educarme durante todos los años de mi vida, en otras palabras, por formarme como persona y entregarme los valores y herramientas necesarias para desenvolverme adecuadamente en la sociedad. En segundo lugar, me gustaría agradecer a mis hermanas, Katherine, Gina y Lisette por ser parte de mi familia, por apoyarme en todo, y ayudarme durante todos mis años de estudio. A mis sobrinos queridos, Martín, Agustín e Isidora, por alegrar mis días. También a Camila, mi pareja, mi compañera de vida, mi “oveja”, por amarme y quererme incondicionalmente, en particular, agradecer por toda la ayuda que me ha brindado en estos últimos meses en los que he trabajado fervientemente en esta memoria. A mis abuelos, mami y tata, por estar siempre presente en los momentos trascendentes de mi vida. A mis tíos Lily y Juan por ser los mejores padrinos que pude haber tenido. A mis primas Geraldine y Constanza, por ser parte importante de mi familia.

Adicionalmente, quiero agradecer a los miembros de la comisión por sus correcciones, comentarios y críticas constructivas durante el desarrollo del presente trabajo. En particular, al profesor guía, Pablo Mendoza, quien ha sido un mentor en las ciencias hidrológicas, por su positivismo y dedicación con el trabajo, por enseñanzas como, “la gente no te recordará por lo bueno que fuiste técnicamente sino por tu trato con ellos” recalcando innumerables veces la importancia de las habilidades blandas. También agradecer al Profesor co-guía Nicolás Vásquez, quien siempre estuvo presente en todas las reuniones del lunes del “team hidrología” para aportar con su conocimiento y sus comentarios. Por último, a la profesora Ximena Vargas quien siempre ha estado pendiente de sus estudiantes, además agradecerle por darme la oportunidad, junto con el profesor guía, de realizar mi primera (y mejor) práctica profesional en el departamento de la facultad.

A todos mis amigos que sin duda han formado parte importante de este proceso. Amigos del colegio, como lo son, Claudia, Francisca, Juan soto y Diego. Amigos de primer año de universidad como, Ítalo, Oscar (sí, fue un honor haber aprendido a integrar contigo) y Ricardo. Amigos de Civil como, Daniel, Daniela, Montserrat y tantos otros nombres que no estoy mencionando, pero que de una u otra forma se quedan en mi corazón.

Finalmente, a mi gata Catalina que me acompañó en mis largas noches de estudio durante toda mi enseñanza media y superior.

# Tabla de Contenido

<b>1. Introducción</b>	<b>1</b>
1.1. Motivación . . . . .	1
1.2. Objetivos . . . . .	3
1.2.1. Objetivo general . . . . .	3
1.2.2. Objetivos específicos . . . . .	3
1.2.3. Organización del informe . . . . .	3
<b>2. Revisión Bibliográfica</b>	<b>4</b>
2.1. Similitud y clasificación de cuencas . . . . .	5
2.2. Métodos de relleno de datos . . . . .	7
2.2.1. Métodos paramétricos . . . . .	8
2.2.2. Métodos no paramétricos . . . . .	8
2.3. Árboles de decisión . . . . .	9
2.4. Hipótesis de trabajo . . . . .	10
<b>3. Metodología</b>	<b>11</b>
3.1. Datos . . . . .	12
3.2. Estimación de caudales medios anuales, estacionales y mensuales . . . . .	12
3.2.1. Elección de atributos . . . . .	12
3.2.2. Definición de criterios de similitud . . . . .	14
3.2.3. Relleno de registros fluviométricos . . . . .	14
3.2.4. Elección del mejor criterio para cada cuenca . . . . .	15
3.3. Estimación de caudales medios diarios . . . . .	16
3.3.1. Variables que condicionan el método K-NN . . . . .	17
3.3.2. Elección de la ventana temporal de comparación . . . . .	18
3.3.3. Estimación de caudales medios diarios . . . . .	18
<b>4. Resultados</b>	<b>20</b>
4.1. Caudal medio anual . . . . .	20
4.2. Caudal medio estacional . . . . .	25
4.3. Caudal medio mensual . . . . .	32
4.4. Caudal medio diario . . . . .	37
<b>5. Conclusiones</b>	<b>43</b>
<b>Bibliografía</b>	<b>45</b>

Anexos	50
A. Árboles de decisión	51
B. Ventanas de comparación para el relleno de caudales medios diarios	64
C. Resultados numéricos	68

# Índice de Tablas

3.1. Atributos finales seleccionados para caracterizar similitud física, climática e hidrológica . . . . .	13
3.2. Formas de condicionar el método K-NN . . . . .	18
3.3. Características de las estaciones fluviométricas utilizadas para validar el método propuesto de relleno de caudales medios diarios. . . . .	19
4.1. Resumen de los atributos más importantes por periodo para estimar caudales medios estacionales. . . . .	30
4.2. Resumen de los atributos más importantes por mes . . . . .	33



# Índice de Figuras

1.1.	Mapa de estaciones fluviométricas correspondiente a 158 cuencas definidas con régimen hidrológico cercano al natural. La escala de colores refleja la cantidad de años hidrológicos con registro completo. . . . .	2
2.1.	Descripción general de los valores promedio de índices climáticos calculados para 1984–2014. (a) Leyenda de índices climáticos para interpretar la Figura 2.1b, los tres ejes muestran como varía el color RGB final. (b) Mapa mundial con resolución de 0,5°, la escala de colores refleja el índice climático correspondiente. (c) Gráfico de aridez promedio anual (rojo), (d) gráfico del cambio estacional de la aridez (verde), y (e) gráfico de fracción de la precipitación que cae en forma de nieve (azul). Obtenido de Knoben et al. (2018). . . . .	7
3.1.	Estructura metodológica general . . . . .	11
3.2.	Ejemplo de relleno de caudales medios anuales. (a) Ubicación geográfica de una estación objetivo en color negro (Estero Yerba Loca antes junta San Francisco) y la estación candidata a transferir información hidrológica en color rojo (Río Aconcagua en Chacabuquito); (b) regresión lineal entre caudales medios anuales, normalizados por las áreas de las cuencas respectivas; (c) serie temporal con años observados (triángulos negros) y rellenados (círculos rojos); (d) validación cruzada, obtenida al eliminar, uno por uno, los puntos existentes de la estación objetivo y estimarlos mediante una regresión lineal con la estación candidata; y (e) curva doble acumulada, obtenida a partir de un patrón que promedia cuatro estaciones con un régimen de escorrentía similar y con un registro extenso (mayor o igual a 20 años completos). . . . .	15
3.3.	Esquema ilustrativo de la metodología propuesta para el relleno de caudales medios diarios. Los cuadrados representan los días del año; aquellos que tienen un color azul representan días con registro de caudal, mientras que los cuadrados blancos son los que no tienen registros. Se considera una ventana temporal de cinco días, donde el año $j$ (que minimiza la distancia en la ventana de comparación) dona información al año $i$ . . . . .	17
3.4.	Ubicación geográfica, registros de precipitación y temperatura para las cuencas seleccionadas para validar el método de relleno propuesto para caudales medios diarios. . . . .	19

4.1. Resultados del proceso de validación cruzada en el relleno de caudales medios anuales: (a) coeficiente de correlación de Pearson; (b) NRMSE; (c) porcentaje de sesgo en función de los criterios utilizados; y (d) mejor criterio de relleno obtenido por estación fluviométrica según NRMSE. . . . .	21
4.2. Árbol de decisión que entrega el mejor criterio para rellenar caudales medios anuales, teniendo como entrada un set de predictores que incluye atributos físicos y climáticos. . . . .	22
4.3. Resultados del relleno de caudales medios anuales: (a) mejor criterio obtenido para cada estación del set de entrenamiento, (b) gráfico final con el mejor criterio para las 158 estaciones; (c) histograma que indica la frecuencia con la cual un criterio es identificado como el mejor; (d) validación cruzada para la escala anual con todas las estaciones y años hidrológicos utilizados. . . . .	23
4.4. Valores de NRMSE obtenidos del proceso de validación cruzada en el relleno de caudales medios estacionales. Cada caja comprime el resultado de 70 estaciones, las cuales poseen más de nueve puntos en validación cruzada. . . . .	25
4.5. Árbol de decisión que entrega el mejor criterio para rellenar caudales medios estacionales de enero - marzo, utilizando como entrada un set de atributos físicos y climáticos. . . . .	26
4.6. Árbol de decisión que entrega el mejor criterio para rellenar caudales medios estacionales de abril - junio, utilizando como entrada un set de atributos físicos y climáticos. . . . .	27
4.7. Árbol de decisión que entrega el mejor criterio para rellenar caudales medios estacionales de julio - septiembre, utilizando como entrada un set de atributos físicos y climáticos. . . . .	28
4.8. Árbol de decisión que entrega el mejor criterio para rellenar caudales medios estacionales de octubre - diciembre, utilizando como entrada un set de atributos físicos y climáticos. . . . .	29
4.9. Resultados del relleno de caudales medios estacionales: (a) gráfico final con el mejor criterio para las 158 estaciones, (b) histograma que indica la frecuencia con la cual un criterio es identificado como el mejor, (c) validación cruzada para la escala estacional con todas las estaciones y años hidrológicos utilizados, para los cuatro periodos definidos. . . . .	31
4.10. Resultados del proceso de validación cruzada en el relleno de caudales medios mensuales. . . . .	32
4.11. Resultados del relleno de caudales medios mensuales desde enero a junio: (a) ubicación geográfica de estaciones, (b) histograma de mejores criterios, y (c) validación cruzada. . . . .	35
4.12. Resultados del relleno de caudales medios mensuales desde julio a diciembre: (a) ubicación geográfica de estaciones, (b) histograma de mejores criterios, y (c) validación cruzada. . . . .	36
4.13. Relleno de caudales medios diarios para la estación Estero Yerba Loca antes junta San Francisco: (a) serie temporal en un periodo de referencia 2011-2015; (b) validación cruzada para el periodo abril-septiembre; (c) validación cruzada para el periodo octubre-marzo; (d) validación cruzada para el año hidrológico completo. . . . .	38

4.14. Relleno de caudales medios diarios para la estación Río Aconcagua en Chacabucuito: (a) serie temporal en un periodo de referencia 2011-2015; (b) validación cruzada para el periodo abril-septiembre; (c) validación cruzada para el periodo octubre-marzo; (d) validación cruzada para el año hidrológico completo. . . . .	39
4.15. Relleno de caudales medios diarios para la estación Río Andalién Camino a Penco: (a) serie temporal en un periodo de referencia 2011-2015; (b) validación cruzada para el periodo abril-septiembre; (c) validación cruzada para el periodo octubre-marzo; (d) validación cruzada para el año hidrológico completo. . . . .	40
4.16. Relleno de caudales medios diarios para la estación Río Lirquén en Cerro el Padre: (a) serie temporal en un periodo de referencia 2011-2015; (b) validación cruzada para el periodo abril-septiembre; (c) validación cruzada para el periodo octubre-marzo; (d) validación cruzada para el año hidrológico completo. . . . .	41
A.1. Árbol de decisión que entrega el mejor criterio para rellenar caudales medios mensuales en enero, utilizando como entrada un set de atributos físicos y climáticos. . . . .	52
A.2. Árbol de decisión que entrega el mejor criterio para rellenar caudales medios mensuales en febrero, utilizando como entrada un set de atributos físicos y climáticos. . . . .	53
A.3. Árbol de decisión que entrega el mejor criterio para rellenar caudales medios mensuales en marzo, utilizando como entrada un set de atributos físicos y climáticos. . . . .	54
A.4. Árbol de decisión que entrega el mejor criterio para rellenar caudales medios mensuales en abril, utilizando como entrada un set de atributos físicos y climáticos. . . . .	55
A.5. Árbol de decisión que entrega el mejor criterio para rellenar caudales medios mensuales en mayo, utilizando como entrada un set de atributos físicos y climáticos. . . . .	56
A.6. Árbol de decisión que entrega el mejor criterio para rellenar caudales medios mensuales en junio, utilizando como entrada un set de atributos físicos y climáticos. . . . .	57
A.7. Árbol de decisión que entrega el mejor criterio para rellenar caudales medios mensuales en julio, utilizando como entrada un set de atributos físicos y climáticos. . . . .	58
A.8. Árbol de decisión que entrega el mejor criterio para rellenar caudales medios mensuales en agosto, utilizando como entrada un set de atributos físicos y climáticos. . . . .	59
A.9. Árbol de decisión que entrega el mejor criterio para rellenar caudales medios mensuales en septiembre, utilizando como entrada un set de atributos físicos y climáticos. . . . .	60
A.10. Árbol de decisión que entrega el mejor criterio para rellenar caudales medios mensuales en octubre, utilizando como entrada un set de atributos físicos y climáticos. . . . .	61
A.11. Árbol de decisión que entrega el mejor criterio para rellenar caudales medios mensuales en noviembre, utilizando como entrada un set de atributos físicos y climáticos. . . . .	62

A.12.Árbol de decisión que entrega el mejor criterio para rellenar caudales medios mensuales en diciembre, utilizando como entrada un set de atributos físicos y climáticos. . . . .	63
B.1. Ventanas temporales comparativas para las diferentes formas de condicionar el método K-NN, en la estación Estero Yerba Loca antes junta San Francisco	64
B.2. Ventanas temporales comparativas para las diferentes formas de condicionar el método K-NN, en la estación Río Aconcagua en Chacabuquito . . . . .	65
B.3. Ventanas temporales comparativas para las diferentes formas de condicionar el método K-NN, en la estación Río Andalién camino a Penco . . . . .	66
B.4. Ventanas temporales comparativas para las diferentes formas de condicionar el método K-NN, en la estación Río Lirruquén en Cerro el Padre . . . . .	67

# Capítulo 1

## Introducción

### 1.1. Motivación

El caudal es una variable clave para entender el comportamiento hidrológico de las cuencas (e.g., Gupta et al., 2008; Westerberg & McMillan, 2015; Addor et al., 2018). Su uso involucra aplicaciones tan variadas como, estudios de línea base, estudios de crecidas, caracterización y predicción de sequías hidrológicas. Algunas de estas aplicaciones no toleran registros con brechas temporales sin información para su desarrollo. Por lo tanto, es de suma importancia contar con datos históricos de caudales medidos en las estaciones fluviométricas, para que cualquier estudio hidrológico se sustente con un registro confiable y dentro de lo posible, sin interrupciones en el tiempo. En Chile esto no se cumple para varias estaciones fluviométricas (Figura 1.1), lo que dificulta cualquier tipo de caracterización hidrológica a la escala de cuenca, tanto para condiciones climáticas históricas como futuras (e.g., Vásquez et al., 2021). Este último punto requiere de una comprensión de los cambios temporales en variables hidrocimáticas en el presente antes de efectuar extrapolaciones – típicamente mediante modelos hidrológicos – en escenarios climáticos futuros (e.g., Wilby and Harris, 2006; Chen et al., 2011; Chegwiddden et al., 2019).

Durante las últimas décadas, se ha masificado el uso de principios de similitud para la estimación de caudales en cuencas no controladas (e.g., Oudin et al., 2008, 2010; Hrachowitz et al., 2013; Parajka et al., 2013; Beck et al., 2016), debido, en parte, a la aparición de nuevas bases de datos integradas de cuencas (e.g., Addor et al., 2017; Newman et al., 2015; Álvarez-Garretón et al., 2018; Chagas et al., 2020; Coxon et al., 2020). En aplicaciones de ingeniería, la estimación de caudales medios anuales para años sin información es típicamente efectuada construyendo modelos de regresión lineal simple con estaciones vecinas que presenten un régimen de escorrentía similar. Sin embargo, se desconoce si la inclusión de cuencas físicamente (e.g., área, pendiente, tipo de suelo) o climáticamente (e.g., temperatura media anual, índice de aridez) similares entre el grupo de candidatas para extrapolar información, podría mejorar las estimaciones en cuencas no controladas. Además, la complejidad del problema se acentúa al aumentar la resolución temporal, debido a que no es directo que la mejor cuenca donante para estimar caudales medios anuales sea también la más apropiada para realizar estimaciones a la escala diaria.

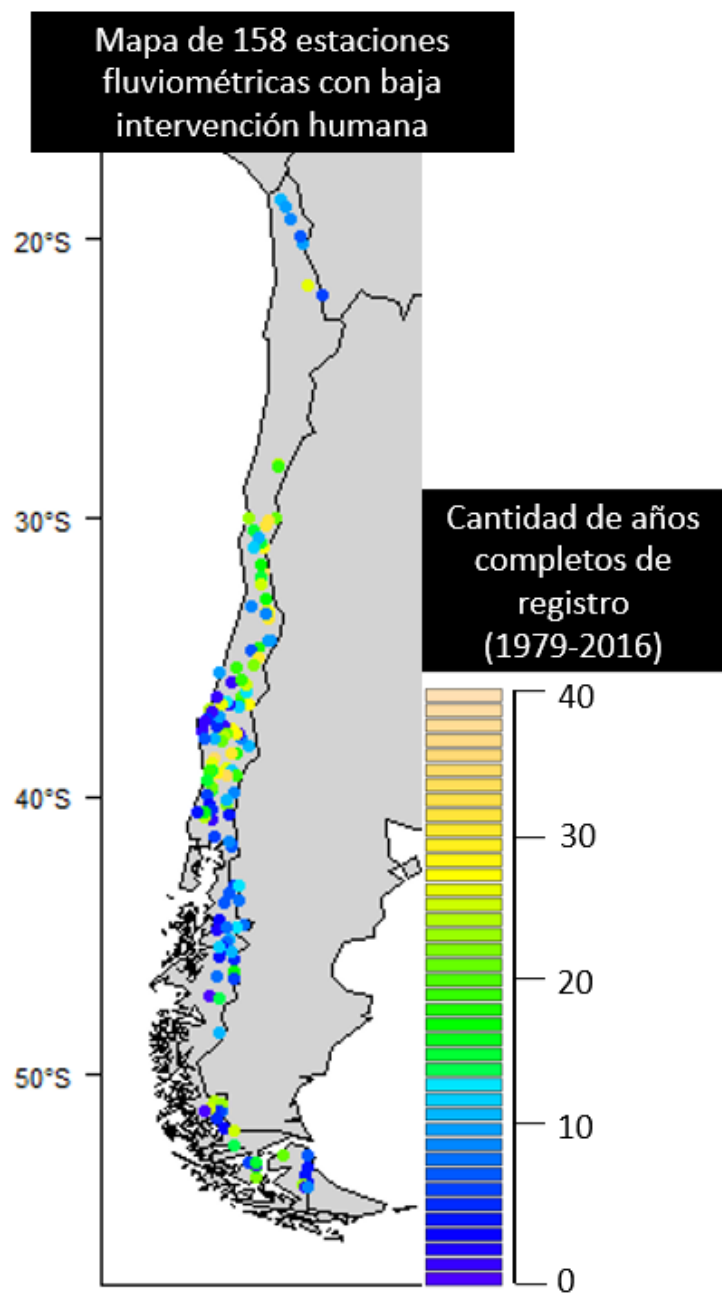


Figura 1.1: Mapa de estaciones fluviométricas correspondiente a 158 cuencas definidas con régimen hidrológico cercano al natural. La escala de colores refleja la cantidad de años hidrológicos con registro completo.

## 1.2. Objetivos

### 1.2.1. Objetivo general

Desarrollar una metodología que permita rellenar vacíos temporales en los registros fluviométricos (anuales, estacionales, mensuales y diarios), en estaciones con baja intervención humana, y determinar su habilidad predictiva mediante validación cruzada.

### 1.2.2. Objetivos específicos

- Comparar diferentes criterios de similitud para el relleno de datos fluviométricos.
- Proponer y evaluar una metodología de relleno de datos fluviométricos diarios en cuatro cuencas piloto.

### 1.2.3. Organización del informe

El informe se organiza de la siguiente forma:

- El capítulo 2 presenta una revisión bibliográfica sobre similitud de cuencas y procedimientos de clasificación de estas, también se abordan métodos tradicionales utilizados para el relleno de datos, adicionalmente, se introducen las bases de los árboles de decisión.
- El capítulo 3 presenta la metodología de trabajo para rellenar datos en los registros de 158 estaciones fluviométricas con baja intervención humana, a diferentes escalas temporales.
- El capítulo 4 contiene la presentación y discusión de los resultados obtenidos utilizando la metodología propuesta en el capítulo anterior.
- Finalmente, el capítulo 5 presenta las principales conclusiones de este estudio.

# Capítulo 2

## Revisión Bibliográfica

Para estudiar la escorrentía de las cuencas chilenas, es necesario contar con series de información completas que, en caso de tener un registro extenso, permitirían además detectar y comprender posibles cambios en patrones temporales. Típicamente, se cuenta, en primera instancia, con datos de caudales a escala temporal horaria o diaria en las cuencas de interés, que posteriormente se procesan para analizarlos en función del objetivo deseado.

Históricamente, se han utilizado principios de similitud para estimar variables hidrológicas en cuencas no controladas (Parajka et al., 2013). Si bien es posible cuantificar el grado de disimilitud entre dos cuencas mediante atributos físicos y climáticos (e.g., Beck et al., 2016), aún no se ha resuelto cómo escoger estos para identificar cuencas hidrológicamente similares de manera efectiva. Siguiendo esta línea, Knoben et al. (2018) afirman que aún no existe una clasificación formal y universalmente aceptada para agrupar cuencas hidrológicamente similares. Sin embargo, también existen otros métodos para estimar caudales en cuencas no controladas, y que se basan en estadísticos, por ejemplo, media aritmética (Pappas et al., 2014). Si se considera además que, al aumentar la resolución temporal, aumenta la incertidumbre en la estimación de variables hidrológicas, el problema se vuelve más complejo, ya que no es claro si es conveniente cambiar la forma de rellenar los registros fluviométricos al pasar de una escala temporal a otra.

En este contexto, la base de datos CAMELS-CL (Álvarez-Garretón et al., 2018), que incluye información de 516 cuencas a lo largo del territorio nacional y que se basa en múltiples fuentes de información, ofrece una gran oportunidad para abordar este desafío mediante la caracterización de condiciones hidroclimáticas y del paisaje a lo largo de Chile continental. Gupta et al. (2014) señalaron que la hidrología actual tiene como desafío mejorar la accesibilidad de los datos, lo cual cobra mayor relevancia en las zonas con escasez de registros. Sin duda, CAMELS-CL aborda esta temática, facilitando además un explorador en línea ([www.camels.cr2.cl](http://www.camels.cr2.cl)) para la visualización de los atributos que caracterizan una cuenca.

En este capítulo, se resume el estado del arte en cuanto a similitud y clasificación de cuencas, además de métodos de relleno de datos y árboles de decisión.



## 2.1. Similitud y clasificación de cuencas

Teniendo una base de datos con información fisiográfica, climática e hidrológica a la escala de cuenca, es posible establecer relaciones de similitud que permitan transferir información hidrológica de una zona a otra.

La similitud puede definirse como el grado de cercanía que existe entre dos o más unidades estructurales (por ejemplo, personas, objetos, lugares, etc.) en base a características, propiedades o atributos en común. En particular, la similitud hidrológica sirve para abordar problemas de predicción en cuencas no controladas, incluyendo variables como el caudal medio anual, estimación de curvas de duración de caudales medios diarios, caudales de crecidas, etc (e.g., Beck et al., 2016). Para resolver estos problemas, Beck et al. (2016) plantean el índice de disimilitud  $S_{ij}$ :

$$S_{ij} = \sum \frac{|A_i - A_j|}{IQR(A)} \quad (2.1)$$

Donde:

$S_{ij}$ : disimilitud entre la cuenca  $i$  y la cuenca  $j$ .

$A_i$ : atributo de la cuenca  $i$ .

$A_j$ : atributo de la cuenca  $j$ .

$IQR(A)$ : rango inter-cuartil de la serie de atributos  $A$ .

La elección de atributos a incluir en la ecuación (2.1) depende del problema particular que se quiera resolver. La transferencia de información en base a criterios de similitud puede ser facilitada con el uso de un enfoque hidrológico Darwiniano, que consiste en relacionar los procesos que ocurren en diferentes cuencas e interpretarlo como información global, en vez, de tratar cada cuenca de forma aislada (Mendoza et al., 2020). Dicho enfoque ha contribuido enormemente al desarrollo de la Hidrología a nivel mundial, y en particular a la Actualización del Balance Hídrico Nacional (DGA, 2017, 2018, 2019) mediante la incorporación de nuevos registros a nivel nacional para cuantificar la disponibilidad del recurso hídrico.

Un desafío de la Hidrología es comprender los procesos del ciclo hídrico para generar modelos que produzcan simulaciones detalladas y confiables en diferentes entornos a múltiples escalas espacio-temporales y bajo condiciones ambientales cambiantes (e.g, Michel et al., 2006; Gupta et al., 2014). Sin embargo, no siempre es posible efectuar estudios detallados a nivel de cuenca, incluso si ésta se encuentra altamente instrumentalizada. Con la finalidad de reducir la incertidumbre propia de las estimaciones de los modelos hidrológicos, y de generar conocimiento desde la diversidad de cuencas, la iniciativa Predicción en Cuencas no Controladas (PUB, por sus siglas en inglés) ha optado por alejarse del uso de un modelo universal calibrado en cada ubicación (Hrachowitz et al., 2013). Este movimiento tiene sus orígenes décadas atrás, pero ha tomado fuerza desde la reunión de la IAHS de 1999 en Birmingham, donde se discutió exhaustivamente el problema (Gupta et al., 2014) y, a lo largo de los años, ha ayudado a impulsar la búsqueda de una mejor comprensión de posibles nexos entre las

características fisiográficas, climáticas y la respuesta hidrológica de las cuencas. El desarrollo de estudios en el contexto PUB ha revelado la necesidad de aprovechar los conjuntos de datos disponibles para hacer común un enfoque de gran muestra para la investigación hidrológica (Gupta et al., 2014), lo que también se conoce como hidrología comparativa (e.g., Thompson et al., 2011). La ciencia hidrológica no cuenta con una metodología universalmente aceptada para clasificar y estructurar la respuesta hidrológica frente a características climáticas y atributos físicos (e.g., Wagener et al., 2007; Knoben et al., 2018). Actualmente, se trabaja en este problema con el objetivo de generar un lenguaje común para la interpretación y predicción de las condiciones y relaciones entre cuencas no controladas. En este marco, se pueden consultar varios sistemas de clasificación derivados de otras disciplinas científicas y que no necesariamente aseguran una clasificación coherente en la ciencia que estudia el ciclo del agua. Para realizar una clasificación de cuencas, se debe considerar un mapeo del paisaje, geomorfología y condiciones hidroclimáticas dentro del área de drenaje de una cuenca (Wagener et al., 2007).

La biología es una de las ciencias primordiales al hablar de clasificación: recurriendo al concepto de taxonomía para la clasificación de organismos, se busca establecer un orden jerarquizado de los elementos que componen un sistema. En la actualidad, el concepto se utiliza de forma más amplia para referirse a los principios que subyacen a la clasificación en otras áreas de estudio. Ejemplos de esto incluyen a la geología con la clasificación de los minerales y los tipos de suelos en relación con la columna estratigráfica; y la astronomía con su clasificación de cuerpos celestes en función del tamaño y las propiedades del objeto de estudio. La ciencia hidrológica tiene éxito localmente en sus diversas ramificaciones, pero surgen problemas al abordar temáticas de investigación globales (Wagener et al., 2007), lo que fue una declaración análoga a la biología desde 1830 (Woods, 2004). Con esto, Wagener et al. (2007) mencionan que “la biología también tuvo una fase de preclasificación similar a la situación actual de la hidrología”.

La cuenca es el objeto fundamental de estudio en la hidrología (Wagener et al., 2007), formando también un elemento paisajístico que integra todos los aspectos del ciclo del agua dentro de un área determinada (Wagener et al., 2004). En general, las cuencas son sistemas abiertos en su entrada y salida de flujos de agua (e.g, Wagener et al., 2007), por lo que se pueden entender como sistemas ambientales complejos con cierto grado de organización (e.g., Sivapalan, 2005). La agrupación de cuencas hidrográficas bajo criterios específicos, basados en distintos tipos de atributos, ayudaría a comprender relaciones entre respuestas hidrológicas y descriptores físicos y climáticos.

Al contrastar cuencas en diferentes lugares mediante el clima o la geología, se reconocen variaciones a través de escalas espaciales y temporales (Wagener et al., 2007). Por lo tanto, un esquema de clasificación general de cuencas requiere consideraciones explícitas de dichas variables.

Para desarrollar esquemas de clasificación que agrupen unidades similares, se han explorado métodos basados en factores físicos, climáticos e hidrológicos (e.g., Sawicz et al., 2011; Kuentz et al., 2017; Jehn et al., 2020). Por ejemplo, Knoben et al. (2018) diseñaron un método de clasificación del clima global en base a tres índices adimensionales calculados con una resolución de  $0,5^\circ \times 0,5^\circ$  en toda el área terrestre. Primero, se encuentra el índice  $l_m$

calculado en base a la aridez promedio anual, luego el índice  $l_{m,r}$  calculado en base al cambio estacional de la aridez, y por último, el índice  $f_s$  que se calcula en base a la fracción de precipitación que cae en forma de nieve. Estos índices climáticos se agrupan mediante el algoritmo de agrupamiento difuso c-means (Schwämmle y Jensen, 2010) con la finalidad de definir un espectro de climas representativos de la superficie terrestre. Posteriormente, los autores evaluaron la correspondencia de los grupos o conjuntos climáticos con la respuesta global de la escorrentía. La figura 2.1 muestra el espectro de climas definidos mediante los índices climáticos calculados por Knoben et al. (2018) en toda la superficie terrestre.

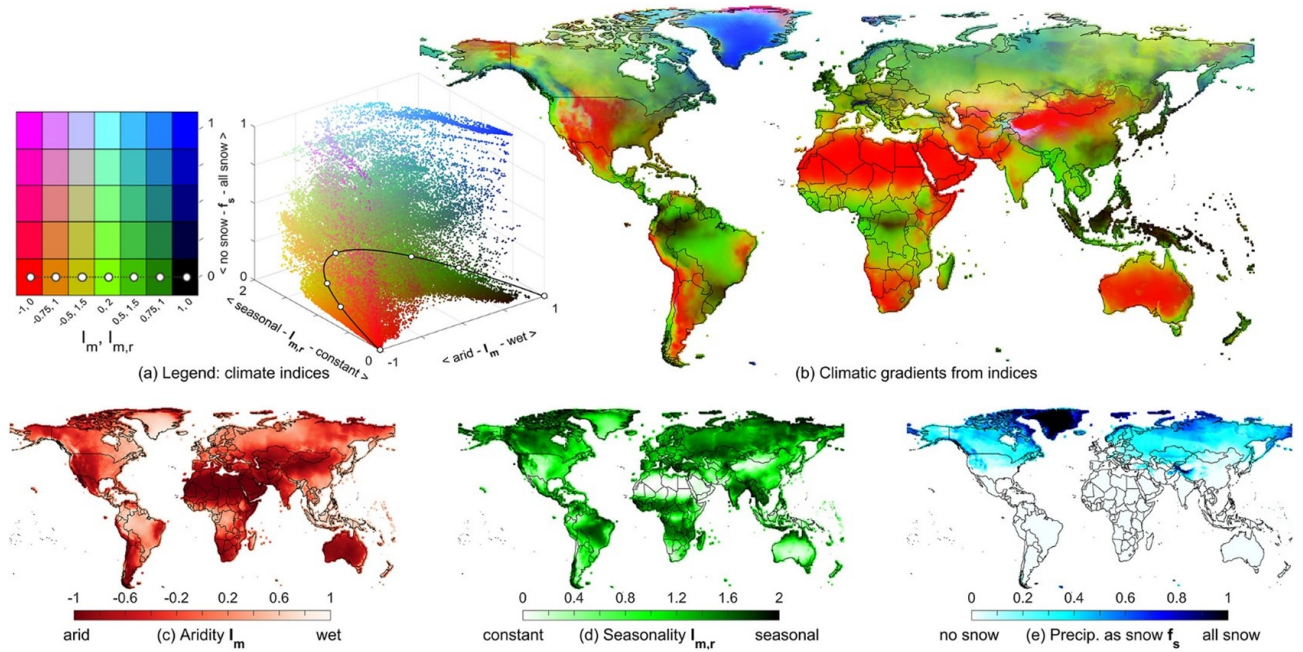


Figura 2.1: Descripción general de los valores promedio de índices climáticos calculados para 1984–2014. (a) Leyenda de índices climáticos para interpretar la Figura 2.1b, los tres ejes muestran como varía el color RGB final. (b) Mapa mundial con resolución de  $0,5^\circ$ , la escala de colores refleja el índice climático correspondiente. (c) Gráfico de aridez promedio anual (rojo), (d) gráfico del cambio estacional de la aridez (verde), y (e) gráfico de fracción de la precipitación que cae en forma de nieve (azul). Obtenido de Knoben et al. (2018).

## 2.2. Métodos de relleno de datos

Una vez establecidas relaciones cuantitativas de similitud entre cuencas, es posible definir las formas de transferencia de información hidrológica. En este contexto, existe una diversidad de técnicas para distribuir uno o más valores de escorrentía a diferentes escalas, ya sea espacial y/o temporalmente, conservando estadísticos como la media aritmética, varianza, valores máximos y mínimos, de los registros existentes.

Para comprender y simular la variabilidad espacial de procesos en sistemas naturales, se requiere de una red de monitoreo densa, que no sólo sea extensa y confiable, sino que también continua en el tiempo (e.g., Butler., 2014; Pappas et al., 2014). Sin embargo, cualquier falla instrumental y la ocurrencia de eventos meteorológicos extremos puede generar vacíos esporádicos y/o continuos en los registros. Además, la existencia de brechas en las series

temporales se exagera en países en desarrollo, donde los recursos para la recopilación de datos, la evaluación de la calidad y el mantenimiento de equipos son limitados (Dembélé et al., 2019). En hidrología, muchas aplicaciones - por ejemplo, análisis de valores extremos, modelación hidrológica, análisis espectral, etc. - requieren de series continuas, por lo que el problema anterior debe ser resuelto adecuadamente. Hay diferentes formas de lograr esto, empezando por los métodos paramétricos, que requieren ajustar un modelo que considere explícitamente una ecuación de la forma:

$$y = f(x) + \varepsilon \quad (2.2)$$

Por otra parte, se encuentran los métodos no paramétricos, que no requieren del ajuste de una ecuación en particular, sino más bien de un muestreo que establezca relaciones para el relleno de datos faltantes. En este análisis, también hay que considerar que el método de relleno no trasciende la escala temporal, ya que muchos de los métodos paramétricos utilizados funcionan correctamente a escala anual, pero no tienen la misma eficacia a la escala diaria (Nowak et al., 2010).

### 2.2.1. Métodos paramétricos

A lo largo del tiempo, se han desarrollado varios métodos de relleno de datos con diferentes niveles de complejidad y limitaciones. Di Piazza et al. (2011), probaron técnicas de relleno de datos aplicadas a los vacíos temporales en los registros de precipitación en Italia, concluyendo que, para completar datos de precipitación a nivel medio mensual y anual, el mejor método correspondía a Kriging ordinario complementado con una regresión lineal simple entre precipitación y elevación topográfica. También indicaron que las técnicas de interpolación para el relleno de registros hidrológicos proveen una representación limitada del espacio-tiempo, por lo que no son eficientes para representar periodos con inundaciones, eventos extremos o eventos de larga duración. Por otra parte, la regresión lineal es uno de los métodos más utilizados debido a su fácil aplicación, aunque la linealidad entre las variables implicadas no siempre se cumple (Mwale et al., 2012). Henn et al. (2013) compararon cinco formas de relleno de datos en los registros temporales de temperatura, utilizando correlaciones espaciotemporales. Cabe destacar que el objetivo del trabajo de Henn et al. (2013) contemplaba una escala temporal horaria, lo cual no garantiza la misma eficacia de relleno en escalas temporales más gruesas. Además, Pappas et al. (2014) indicaron que métodos convencionales como el promedio de la muestra completa para el manejo de datos faltantes cuenta con serias limitaciones al analizar la serie temporal hidrometeorológica, y que para rellenar brechas esporádicas (un sólo valor) es mejor utilizar una suma de valores ponderados y estrictamente escalados, por lo que constituye una herramienta valiosa para rellenar datos faltantes esporádicos de forma rápida, sin necesidad de una estación vecina en el proceso.

### 2.2.2. Métodos no paramétricos

Dawson et al. (2002) tomaron otro enfoque, utilizando redes neuronales artificiales para pronósticos hidrológicos en el Río Yangtze en China. Este método en general es preciso en sus estimaciones, pero tiene un alto costo computacional, lo que dificulta su uso en aplicaciones cotidianas. Otros autores han utilizado un método no paramétrico incipiente llamado

Muestreo Directo (DS, por sus siglas en inglés) para resolver el problema de los vacíos temporales (Dembélé et al., 2019). Este método corresponde a un algoritmo estadístico de puntos múltiples, el cual está condicionado para la reproducción de ciertos patrones. Dembélé et al. (2019) señalaron que la principal ventaja de este método es la capacidad de proporcionar estimaciones probabilísticas de los vacíos en los registros, lo que permite cuantificar la incertidumbre.

También se ha utilizado el método no paramétrico *K Nearest Neighbors* (K-NN, Lall y Sharma, 1996) para diferentes aplicaciones. Por ejemplo, Nowak et al. (2010) lo utilizaron para simular directamente datos de caudales a una escala temporal diaria en múltiples ubicaciones, basado en la desagregación de un único valor de caudal a escala temporal anual. La implementación del método es sencilla, con bajo costo computacional, y puede considerar explícitamente la desagregación de caudal medio anual a caudal medio diario en caso que esta última sea la variable de interés. Para este objetivo, los valores de caudal medio diario observados se convierten a una proporción del caudal medio anual, dando como resultado una matriz P de dimensiones  $n \times 365$ , siendo n el número de años completos con datos observados. Luego, se denota el caudal medio anual como Z, el cual posteriormente se desglosará en caudal medio diario. Posteriormente, se buscan los K vecinos más cercanos a Z, a partir de los registros históricos de caudal medio anual. Los vecinos más cercanos a Z se encuentran mediante una distancia, definida por la diferencia entre los valores de caudales medios anuales. Por otra parte, el número K de vecinos más cercanos está dado por la raíz cuadrada de n (Lall y Sharma, 1996) y a cada vecino se le debe calcular un peso, según la siguiente ecuación (Lall y Sharma, 1996).

$$W(i) = \frac{\frac{1}{i}}{\sum_i^k \frac{1}{i}} \quad (2.3)$$

Dónde k corresponde al número de vecinos más cercanos al caudal Z e i corresponde al índice de cada vecino. Luego, se construye una función de probabilidad acumulada, considerando los pesos calculados anteriormente, después se muestrean números aleatorios entre 0 y 1 y se elige el vecino correspondiente mediante la intersección de estos números con la curva generada mediante la función de probabilidad acumulada, este proceso puede repetirse las veces deseadas. Se selecciona el vector de caudales medios diarios correspondiente al año del vecino escogido, llamado  $P_y$ . Por último, se proyecta el caudal medio anual Z en la distribución diaria del vecino más cercano, resultando el vector de caudales medios diarios (d) del año de interés.

$$d = Z \cdot P_y \quad (2.4)$$

## 2.3. Árboles de decisión

Un árbol de decisión puede ser expresado como una forma gráfica o analítica donde se representan eventos que pueden surgir a partir de una decisión. La predicción de un árbol es realizada en base a una serie de condiciones umbrales en los predictores, además, el esquema de predicción se inicia en la parte superior del árbol. Esto es relevante, ya que se puede conocer

la importancia de un predictor en el resultado final, mediante la posición de este en el árbol, siendo los de más arriba los más importantes. Finalmente, Addor et al. (2018) mencionan que la predicción resultante de los árboles de decisión corresponde a valores discretos, ya que estos resultan del valor medio del predictor para todos los elementos que se encuentran en una clase.

Los árboles de decisión pueden ser separados en dos grupos:

- Árboles de regresión: producen una respuesta cuantitativa
- Árboles de clasificación: producen una respuesta cualitativa

James et al. (2013) señalan las ventajas y desventajas de utilizar este modelo.

Ventajas:

- Los árboles pueden ser mostrados gráficamente, y, de esta forma ser interpretados por personas no expertas en la materia.
- Los árboles de decisión reflejan de forma más realista la toma de decisiones que los enfoques de regresión lineal.
- Los árboles pueden manejar predictores cualitativos sin la necesidad de variables ficticias.

Desventajas:

- En general, los árboles de decisión no tienen la misma precisión en sus predicciones que otros enfoques de regresión (por ejemplo: bosques aleatorios).

## 2.4. Hipótesis de trabajo

La revisión de la literatura revela puntos de convergencia que hay que considerar para comenzar el estudio del relleno de datos en series temporales, aplicado a los registros fluvio-métricos de cuencas chilenas cercanas al régimen natural. Las conclusiones de este capítulo se exponen a continuación:

- Para transferir información hidrológica de una cuenca a otra, siguiendo un enfoque Darwiniano, hay que establecer criterios de similitud.
- La literatura no provee de una guía definitiva sobre qué atributos hay que considerar para cuantificar cada tipo de similitud (hidrológica, física y climática).
- La técnica de relleno de datos más adecuada depende de varios factores, incluyendo la longitud de los vacíos, el régimen hidrológico, la longitud de los registros y la disponibilidad de datos complementarios en estaciones vecinas (Mwale et al., 2012).
- De acuerdo con Nowak et al. (2010), se espera que una metodología no paramétrica tenga mejores resultados que un método paramétrico para el relleno de datos a una escala temporal diaria.

# Capítulo 3

## Metodología

El problema de vacíos en los registros fluviométricos puede ser abordado mediante métodos estadísticos de relleno de datos. La metodología general para el relleno de caudales considera etapas que pasan por diferentes escalas temporales. Primero, se deben estimar los caudales de la escala temporal más gruesa, para luego aumentar la resolución temporal y rellenar los caudales para pasos temporales más finos. También se debe considerar un factor de corrección, para asegurar que exista consistencia entre las estimaciones efectuadas a distintas escalas temporales (Figura 3.1).

El primer paso de la metodología considera una recopilación de antecedentes, cuyo resultado es una lista de características de las cuencas de interés. Luego, se definen cuatro escalas temporales de relleno (anual, estacional, mensual y diaria), distinguiéndose las técnicas utilizadas para pasos temporales gruesos (anual a mensual) y para caudales medios diarios.

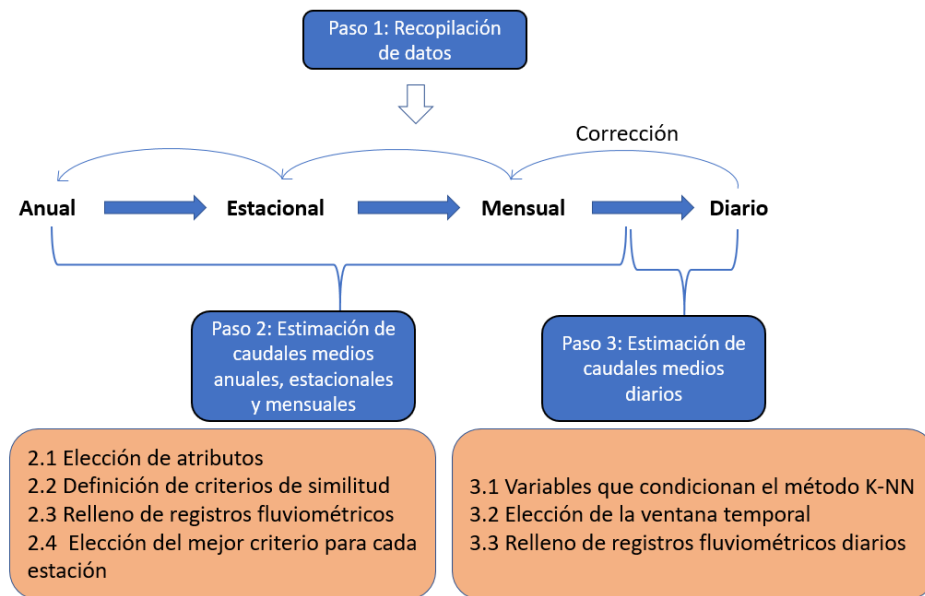


Figura 3.1: Estructura metodológica general

## 3.1. Datos

Para analizar los registros fluviométricos, se utiliza la base de datos CAMELS-CL (Álvarez-Garretón et al., 2018), la que contiene atributos que permiten cuantificar el grado de similitud entre cuencas (e.g, Wagener et al., 2007). Varios autores han destacado la importancia de las bases de datos para inferir leyes generalizables para otras cuencas (e.g, Gupta et al., 2014), lo que se complementa con la creación de CAMELS-CL para 516 cuencas en Chile. Específicamente, CAMELS-CL incluye series de tiempo para variables hidrometeorológicas y un conjunto de atributos basados en el clima, la hidrología, la geología, la cobertura del suelo y el uso del agua. En particular, se incluye un atributo que cuantifica el impacto de la intervención antrópica en cada cuenca, tomando la relación entre los derechos consuntivos y permanentes de agua asignados al año, dividido por el caudal medio anual de la cuenca correspondiente.

En el presente estudio, se utilizan las cuencas chilenas cercanas al régimen natural, las que deben cumplir con tres condiciones: (1) un índice de intervención humana (según CAMELS-CL) menor a 0,05, (2) que la cuenca no cuente con embalses de grandes dimensiones, y (3) que la estación asociada a la cuenca no esté intervenida por derechos de agua no consuntivos que devuelvan sus aguas abajo de la estación fluviométrica. Bajo estas condiciones, se escoge un total de 158 cuencas.

## 3.2. Estimación de caudales medios anuales, estacionales y mensuales

### 3.2.1. Elección de atributos

Siguiendo el flujo metodológico, se deben escoger los atributos asociados a las cuencas de interés que caracterizan los diferentes tipos de similitudes (física, climática e hidrológica). Inicialmente, se consideran todos los atributos presentes en CAMELS-CL, y se eliminan aquellos que tengan un coeficiente de correlación de Spearman mayor a 0,7 con otro atributo, con el objetivo de desechar información redundante. También se eliminan los atributos en los cuales no se tenga información para todas las cuencas, ya que no se podría utilizar la ecuación 2.1 para cuantificar el grado de similitud entre dos cuencas.

La tabla 3.1 muestra todos los atributos utilizados en el presente trabajo, en estos se incluyen los seleccionados del trabajo de Álvarez-Garretón et al. (2018), además de distancia euclidiana entre coeficientes de Pardé y porcentaje de área nival.



Tabla 3.1: Atributos finales seleccionados para caracterizar similitud física, climática e hidrológica

<b>Similitud</b>	<b>Atributos</b>	<b>Descripción</b>	<b>Unidades</b>
Física	Elevación media (elev_mean)	Elevación media de la cuenca.	m.s.n.m
	Pendiente media (slope_mean)	Pendiente media de la cuenca.	$m \cdot km^{-1}$
	Fracción de bosque (forest_frac)	Fracción de la cuenca cubierta por bosques, incluidos bosques nativos y plantaciones.	%
	Fracción de pasto (grass_frac)	Porcentaje de la cuenca cubierta por pastizales.	%
	Fracción de arbusto (shrub_frac)	Fracción de la cuenca cubierta por arbustos.	%
	Clase geológica más común (geol_class)	Fracción de la cuenca asociada con la clase geológica más común, dependiente del mapa litológico.	-
	Tierras de cultivo (crop_frac)	Porcentaje de la cuenca cubierta por tierras de cultivo.	%
	Falta de cobertura terrestre (land_cover_missing)	Porcentaje de la cuenca no cubierta por el mapa de cobertura terrestre.	%
	Porcentaje de área nival	Definida por la intersección de la curva hipsométrica con la línea de nieve promedio (Peña & Vidal, 1993).	%
Climática	Fracción de nieve (frac_snow)	Fracción de los eventos de precipitación que caen en forma de nieve.	-
	Estacionalidad de la precipitación (p_seasonality)	Estacionalidad y momento en que ocurre la precipitación respecto a la temperatura, valores cercanos a cero indican una evapotranspiración uniforme.	-
	Índice de aridez (aridity_i)	Relación entre la PET media diaria y la precipitación media diaria.	-
Hidrológica	Distancia entre Coeficientes de Pardé	Se calcula como la distancia euclidiana entre 12 puntos que se componen del caudal medio mensual dividido caudal medio anual.	-

### 3.2.2. Definición de criterios de similitud

Para rellenar datos faltantes en alguna estación de interés, se utilizan estaciones cercanas (menos de 200 km) y similares. Para determinar la medida de similitud más apropiada, se definen criterios de búsqueda en base a las siguientes combinaciones:

- Criterio 1: similitud hidrológica.
- Criterio 2: similitud climática.
- Criterio 3: similitud física.
- Criterio 4: similitud hidrológica y física.
- Criterio 5: similitud hidrológica y climática.
- Criterio 6: similitud física y climática.
- Criterio 7: similitud hidrológica, física y climática.

### 3.2.3. Relleno de registros fluviométricos

En esta etapa, se estiman los registros fluviométricos faltantes mediante el ajuste de una regresión lineal simple, utilizando como referencia las estaciones obtenidas de aplicar los criterios definidos en la etapa anterior. Los pasos para el relleno son:

1. Se elabora un ranking de similitud con las estaciones vecinas, considerando el set de atributos asociado al criterio de interés y la ecuación 2.1.
2. Se calcula el coeficiente de determinación entre los caudales de la cuenca objetivo y los caudales de las diez cuencas más similares según el ranking efectuado anteriormente.
3. De las diez cuencas encontradas, se escoge un máximo de cuatro cuencas candidatas que cuenten con coeficientes de determinación mayor a 0,5.
4. Se ajusta un modelo lineal entre la estación objetivo y la primera estación candidata, y se rellenan los caudales faltantes con dicho modelo. Se repite el procedimiento con las estaciones candidatas restantes (considerando sólo datos observados) y se estiman los datos que no pudieron ser rellenados con estaciones de mejor ranking.
5. Finalmente, se le aplica un factor de corrección a los valores estimados para que exista coherencia entre escalas temporales.

Hay que destacar que, para la escala temporal mensual, sólo se utilizan los criterios 1, 4, 5 y 7, ya que su estimación requiere que exista similitud hidrológica.

La Figura 3.2 ilustra un ejemplo de relleno de caudal medio anual, en la estación objetivo Estero Yerba Loca antes junta San Francisco, utilizando el criterio 1 (similitud hidrológica).

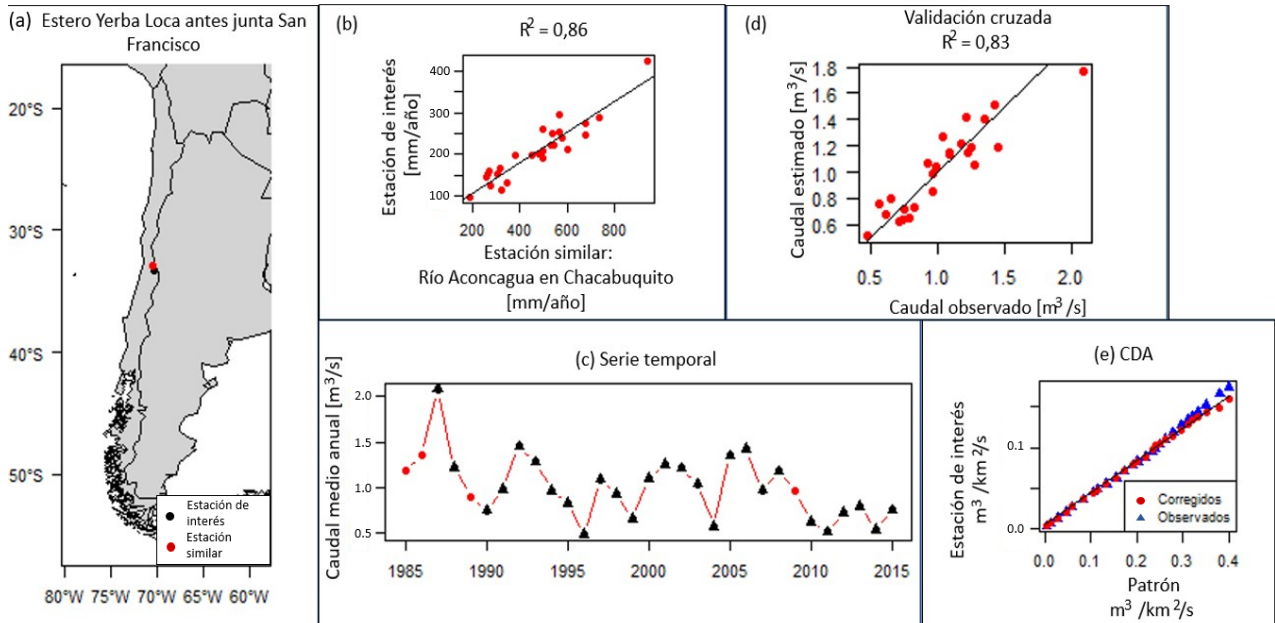


Figura 3.2: Ejemplo de relleno de caudales medios anuales. (a) Ubicación geográfica de una estación objetivo en color negro (Estero Yerba Loca antes junta San Francisco) y la estación candidata a transferir información hidrológica en color rojo (Río Aconcagua en Chacabucuito); (b) regresión lineal entre caudales medios anuales, normalizados por las áreas de las cuencas respectivas; (c) serie temporal con años observados (triángulos negros) y rellenos (círculos rojos); (d) validación cruzada, obtenida al eliminar, uno por uno, los puntos existentes de la estación objetivo y estimarlos mediante una regresión lineal con la estación candidata; y (e) curva doble acumulada, obtenida a partir de un patrón que promedia cuatro estaciones con un régimen de escorrentía similar y con un registro extenso (mayor o igual a 20 años completos).

### 3.2.4. Elección del mejor criterio para cada cuenca

Esta etapa considera repetir, en cada cuenca, el paso anterior para los siete criterios definidos, escogiéndose aquel que entregue el menor valor del error cuadrático medio normalizado (NRMSE, por sus siglas en inglés) obtenido de la validación cruzada. Este procedimiento es aplicado sólo en estaciones que tengan al menos diez puntos estimados en validación cruzada. Luego, se entrena un árbol de decisión utilizando sólo dichas estaciones con el paquete “rpart” (Therneau et al., 2019) del programa computacional R, para luego determinar con este modelo el mejor criterio de relleno para las estaciones restantes en función de los atributos físicos y climáticos de cada cuenca. El árbol resultante tiene una condición en cada nodo, si se cumple la condición se pasa al nodo izquierdo, en caso contrario, se pasa al nodo derecho, este procedimiento se repite hasta llegar al final del árbol. Finalmente, se obtiene el criterio de relleno óptimo para cada cuenca en cada escala temporal (anual, estacional y mensual).

### 3.3. Estimación de caudales medios diarios

Para rellenar caudales medios diarios, se utiliza un enfoque no paramétrico basado en el método *K Nearest Neighbors* (K-NN; Lall y Sharma, 1996). En particular, se adapta la metodología propuesta por Nowak et al. (2010), y originalmente utilizada para desagregar caudales medios anuales a nivel diario. El método se basa en adoptar el valor en un sitio de interés del vecino más cercano, teniendo como referencia una medida de distancia.

En primer lugar, se genera una matriz de caudales medios diarios con dimensiones  $n \times 365$ , en donde  $n$  corresponde al número de años observados en una estación de interés. Luego, se define la ventana temporal en el año objetivo a rellenar, la cual es comparada con todos los años del periodo histórico en los cuales se tengan datos de caudales medios diarios, y se escoge la distribución diaria del año más similar a la ventana objetivo de acuerdo con algún criterio de similitud. Los datos estimados de caudal medio diario utilizando el método K-NN son multiplicados por un factor de corrección, para que exista coherencia entre los caudales medios mensuales y los datos a escala diaria.

La Figura 3.3 ilustra un ejemplo de relleno de caudales medios diarios. Primero, se tiene un año objetivo, denotado con la letra  $i$ , el cual se descompone en 365 días con o sin registros de caudal. Luego, se define una ventana temporal cinco días, la cual compara el año objetivo ( $i$ ) con todos los otros años que tengan registros ( $j$ ) en esa misma brecha temporal, utilizando una distancia euclidiana en base a valores de precipitación diaria y temperatura media diaria (para el mismo rango de fechas). El objetivo es encontrar un año  $j$  que minimice dicha distancia y que, por lo tanto, pueda donar información de caudales medios diarios. En el cálculo de las distancias, tanto la precipitación ( $P$ ) como la temperatura diaria ( $T$ ) son normalizadas utilizando la media aritmética y la desviación estándar de la ventana temporal (Figura 3.3). También hay que notar que el número  $K$  de vecinos más cercanos para la simulación de caudales corresponde a la raíz cuadrada de la cantidad de años completos de registro.

Nota: Todas las figuras generadas con software (por ejemplo: Figura 3.2) tienen un punto como separador de decimales, en el resto del informe se utiliza la coma para este efecto.

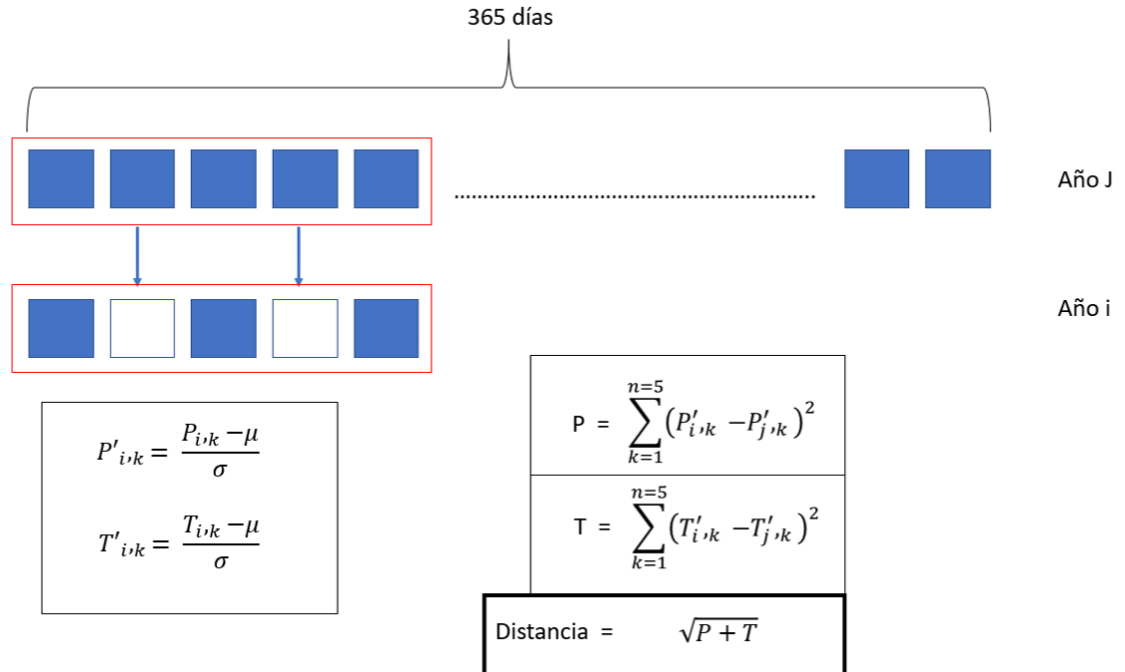


Figura 3.3: Esquema ilustrativo de la metodología propuesta para el relleno de caudales medios diarios. Los cuadrados representan los días del año; aquellos que tienen un color azul representan días con registro de caudal, mientras que los cuadrados blancos son los que no tienen registros. Se considera una ventana temporal de cinco días, donde el año j (que minimiza la distancia en la ventana de comparación) dona información al año i.

Como se puede ver en la Figura 3.3, existen dos decisiones metodológicas que inciden en los resultados del relleno de caudales medios diarios. Por un lado, se encuentran las variables independientes que condicionan el método y que, por lo tanto, son responsables de las distancias calculadas, y, por otro lado, se encuentra la ventana temporal de comparación, ya que la situación cambia radicalmente al comparar dos días de ventana temporal o un mes de esta.

### 3.3.1. Variables que condicionan el método K-NN

En el presente trabajo, se proponen cuatro formas de condicionar el método K-NN en base al cálculo de distancia euclidiana y la comparación de brechas temporales.

- Caudal medio mensual ( $Q_m$ ): se propone utilizar valores observados o estimados (siguiendo los pasos previos de la metodología) de caudal medio mensual para el cálculo de la distancia euclidiana, y posteriormente escoger los años similares que puedan donar información de caudales medios diarios.
- Precipitación y temperatura ( $P, T$ ): se propone utilizar datos de precipitación y temperatura media diaria para calcular la distancia euclidiana con el objetivo de encontrar años similares, tal como se muestra en el ejemplo de la Figura 3.3.
- Caudal medio mensual, precipitación y temperatura ( $Q_m, P, T$ ): se propone combinar

las dos formas anteriores de condicionar el método, e incluir las tres variables ya mencionadas en el cálculo de distancia. Además, para esta forma de condicionar el método, se proponen dos formas diferentes de calcular la distancia variando el peso del caudal medio mensual.

La tabla 3.2 muestra las cuatro formas propuestas para condicionar el método K-NN, cada variable corresponde a la diferencia entre el valor registrado o estimado en el año objetivo y el año con registro en esa misma ventana temporal.

Tabla 3.2: Formas de condicionar el método K-NN

<b>VARIABLES</b>	<b>ABREVIACIÓN</b>	<b>CÁLCULO DE LA DISTANCIA</b>
Caudal medio mensual	Qm	$\sqrt{Qm^2}$
Precipitación y temperatura	P, T	$\sqrt{P^2 + T^2}$
Caudal medio mensual, precipitación y temperatura	Qm, P, T	$\sqrt{Qm^2 + P^2 + T^2}$
Caudal medio mensual, precipitación y temperatura	2xQm, P, T	$\sqrt{2Qm^2 + P^2 + T^2}$

### 3.3.2. Elección de la ventana temporal de comparación

Para escoger la ventana temporal de comparación, se aplica el método para distintas ventanas temporales y se escoge aquella que cuente con las mejores métricas en validación cruzada. Es decir, se estiman los caudales medios diarios para ventanas temporales de comparación de 5, 10, 15, 20, 25 y 30 días, obteniéndose tanto el coeficiente de determinación como el error cuadrático medio normalizado. Finalmente, se escoge la ventana temporal que entregue los mejores indicadores.

### 3.3.3. Estimación de caudales medios diarios

El relleno final de caudales medios diarios viene dado por la mejor combinación entre las cuatro variables consideradas para estimar la distancia y el cálculo de la mejor ventana temporal de comparación. Finalmente, el caudal medio diario estimado proviene del año j, el cual corresponde al año más similar al año objetivo i, sin embargo, este dato debe ser multiplicado por un factor de escala, el cual corresponde al caudal medio anual del año i, dividido por el caudal medio anual del año j.

Para evaluar la efectividad de la metodología de relleno de caudales medios diarios, se escogen cuatro cuencas piloto - dos de régimen pluvial y dos de régimen nival - que además cuentan con al menos 20 años de registros.

La Tabla 3.3 y la Figura 3.4 muestran las principales características de las estaciones utilizadas, además de los registros de precipitación y temperatura para el relleno de caudales medios diarios. Las estaciones fluviométricas Estero Yerba Loca antes junta San Francisco y

Río Aconcagua en Chacabucuito se asocian con un régimen nival, mientras que las estaciones Río Andalién camino a Penco y Río Lirquén en Cerro el Padre tienen un régimen pluvial.

Tabla 3.3: Características de las estaciones fluviométricas utilizadas para validar el método propuesto de relleno de caudales medios diarios.

Estaciones	Régimen	Área [ $km^2$ ]	Elevación media [m.s.n.m]	Pendiente Media [m/km]	Índice de aridez []
Estero Yerba loca antes junta San Francisco	Nival	110	3424	295,5	1,7
Río Aconcagua en Chacabucuito	Nival	2113,4	3178	279,9	2
Río Andalién Camino a Penco	Pluvial	751,7	210	112,6	1,1
Río Lirquén en Cerro el Padre	Pluvial	103,4	668	164	0,5

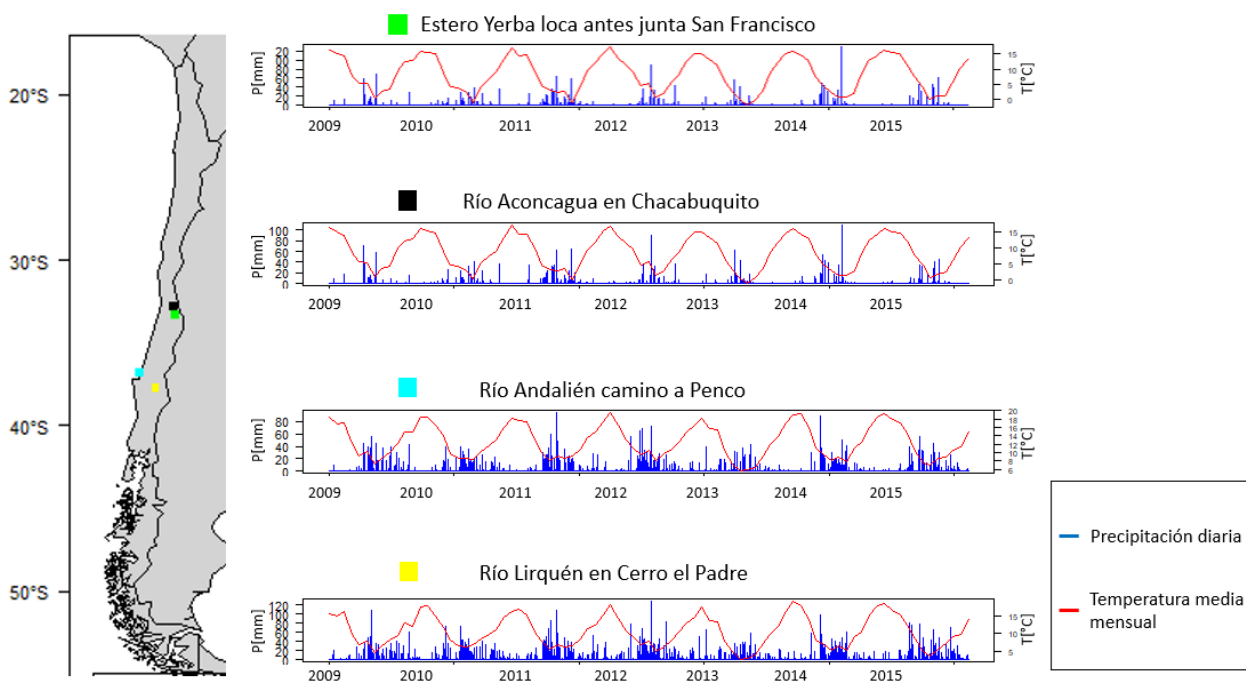


Figura 3.4: Ubicación geográfica, registros de precipitación y temperatura para las cuencas seleccionadas para validar el método de relleno propuesto para caudales medios diarios.

# Capítulo 4

## Resultados

### 4.1. Caudal medio anual

La Figura 4.1 contiene gráficos de cajas con las métricas resultantes de la validación cruzada, incluyendo el coeficiente de correlación de Pearson, el error cuadrático medio normalizado (NRMSE) y el sesgo, para cada criterio de relleno. Se puede notar que el criterio 1 es el que cuenta con los mejores resultados de forma global, puesto que tiene la mayor correlación de Pearson, el menor NRMSE, y el porcentaje de sesgo más cercano a cero.

Las Figuras 4.1a, 4.1b y 4.1c muestran los resultados de aplicar el mismo criterio a todas las estaciones para las estimaciones de caudal medio anual. Sin embargo, también se puede variar el enfoque y rellenar datos faltantes en cada estación con el mejor criterio que le corresponda. La Figura 4.1d muestra el resultado de este ejercicio, utilizando NRMSE como función objetivo (i.e., el color indicado corresponde a aquel criterio que entrega el NRMSE más bajo).



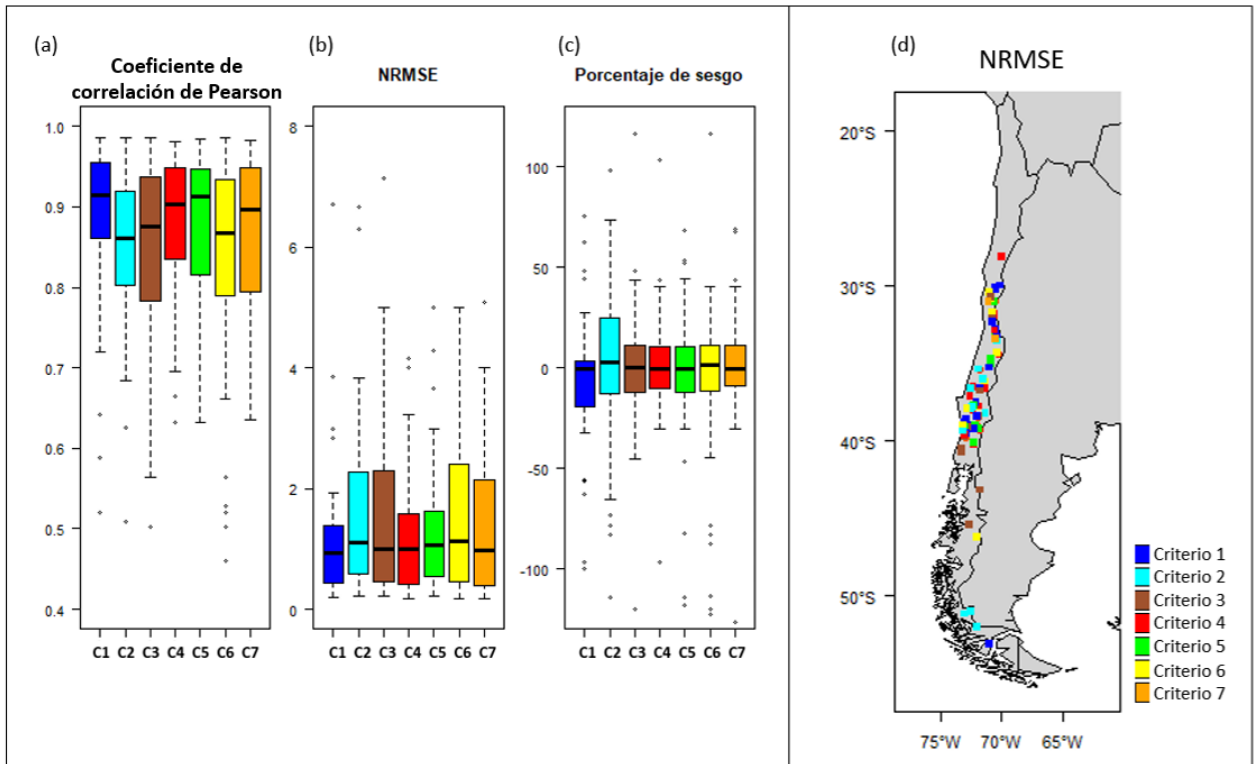


Figura 4.1: Resultados del proceso de validación cruzada en el relleno de caudales medios anuales: (a) coeficiente de correlación de Pearson; (b) NRMSE; (c) porcentaje de sesgo en función de los criterios utilizados; y (d) mejor criterio de relleno obtenido por estación fluviométrica según NRMSE.

Dado que no todas las estaciones fluviométricas a rellenar cuentan con la longitud de registro necesaria para escoger un criterio, se decide ajustar un modelo de árbol de decisión (Figura 4.2), mediante el cual se obtiene que la estacionalidad de la precipitación, la clase geológica más común y la fracción de arbusto tienen la mayor preponderancia en la estimación de caudales medios anuales. Hay que notar que, las unidades de medida de los predictores son las mencionadas en la Tabla 3.1. El modelo anterior permite inferir el criterio óptimo de relleno en aquellas cuencas donde no existe suficiente información para realizar una validación cruzada. La Figura 4.3 muestra el resultado de este proceso para las 158 estaciones fluviométricas incluidas en este estudio.

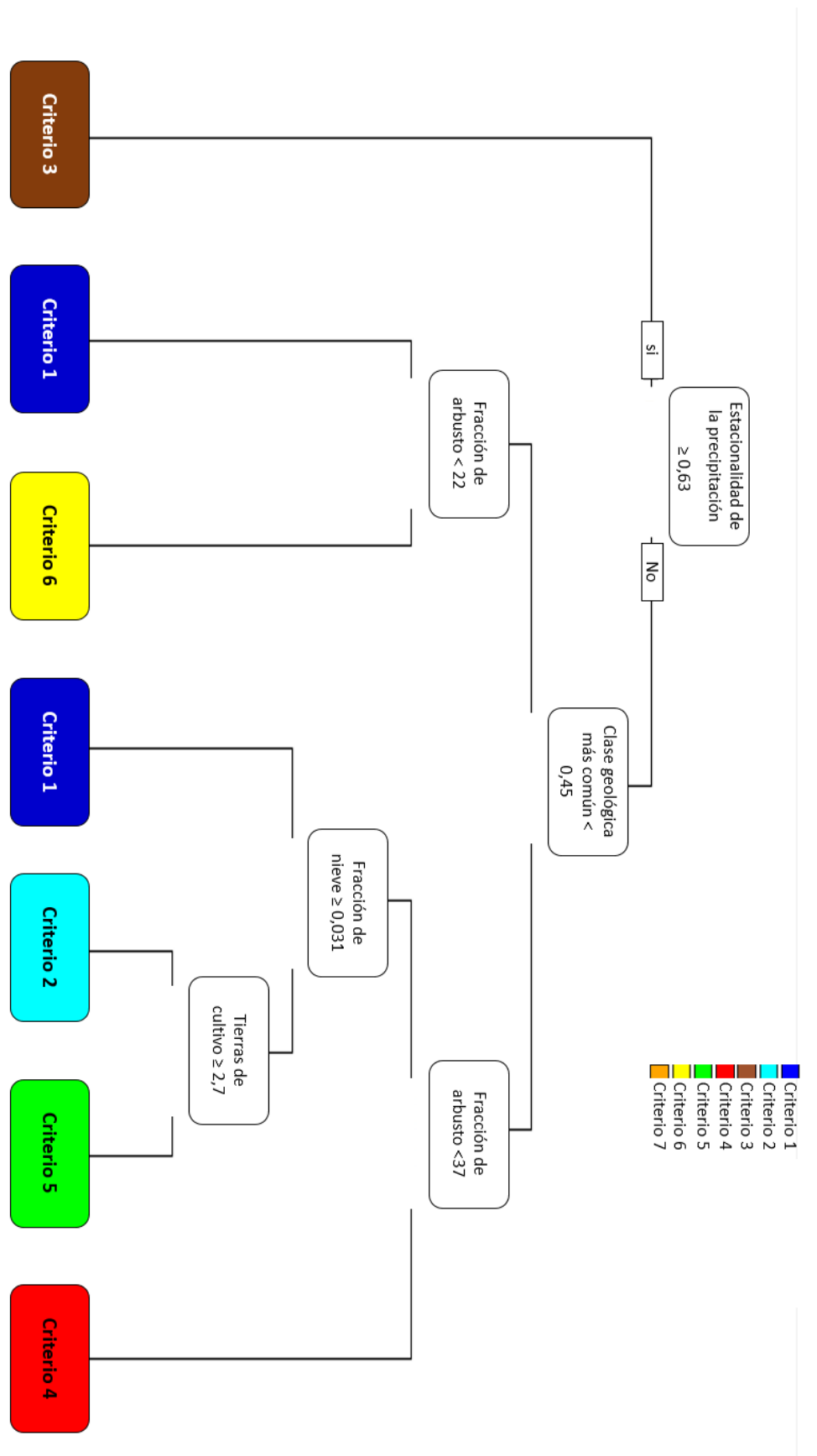


Figura 4.2: Árbol de decisión que entrega el mejor criterio para rellenar caudales medios anuales, teniendo como entrada un set de predictores que incluye atributos físicos y climáticos.

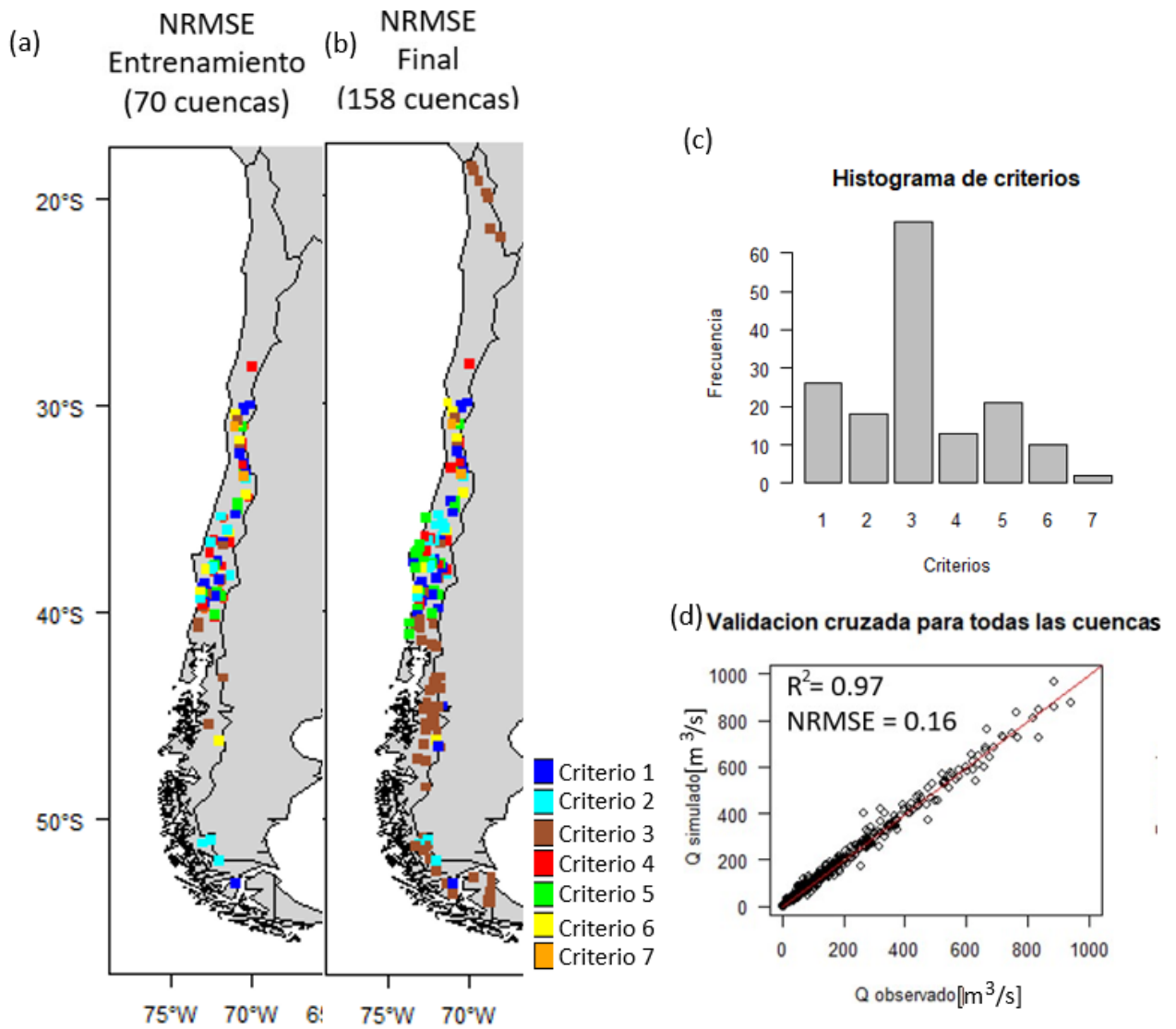


Figura 4.3: Resultados del relleno de caudales medios anuales: (a) mejor criterio obtenido para cada estación del set de entrenamiento, (b) gráfico final con el mejor criterio para las 158 estaciones; (c) histograma que indica la frecuencia con la cual un criterio es identificado como el mejor; (d) validación cruzada para la escala anual con todas las estaciones y años hidrológicos utilizados.

La Figura 4.3b muestra que la mejor forma de estimar caudales medios anuales al norte de los 30° y al sur de los 40° de latitud es utilizando el criterio 3 (similitud física), el cual incluye los siguientes atributos: elevación media, pendiente media, fracción de bosque, fracción de pasto, fracción de arbusto, clase geológica más común, tierras de cultivo, falta de cobertura terrestre y porcentaje de área nival. La figura 4.3c indica que aproximadamente la mitad de las estaciones incluidas en este estudio tienen al criterio 3 como óptimo. Lo anterior, tiene sentido debido a que el caudal medio anual de estaciones con régimen diferente puede ser el mismo. Finalmente, la figura 4.3d muestra los resultados obtenidos de la validación cruzada al aplicar el criterio óptimo de relleno a cada estación fluviométrica, con un coeficiente de determinación de 0,97 y un NRMSE de 0,16.

## 4.2. Caudal medio estacional

Para el relleno de caudales medios estacionales, se divide el año hidrológico en periodos de tres meses: enero a marzo, abril a junio, julio a septiembre y el último de octubre a diciembre. Luego, se sigue el mismo procedimiento descrito para el relleno anual considerando el paso 2 de la metodología.

La Figura 4.4 muestra los resultados del NRMSE al rellenar los caudales medios estacionales con cada uno de los criterios. Se observa que, a diferencia de la escala temporal anual el NRMSE tiene una menor dispersión.

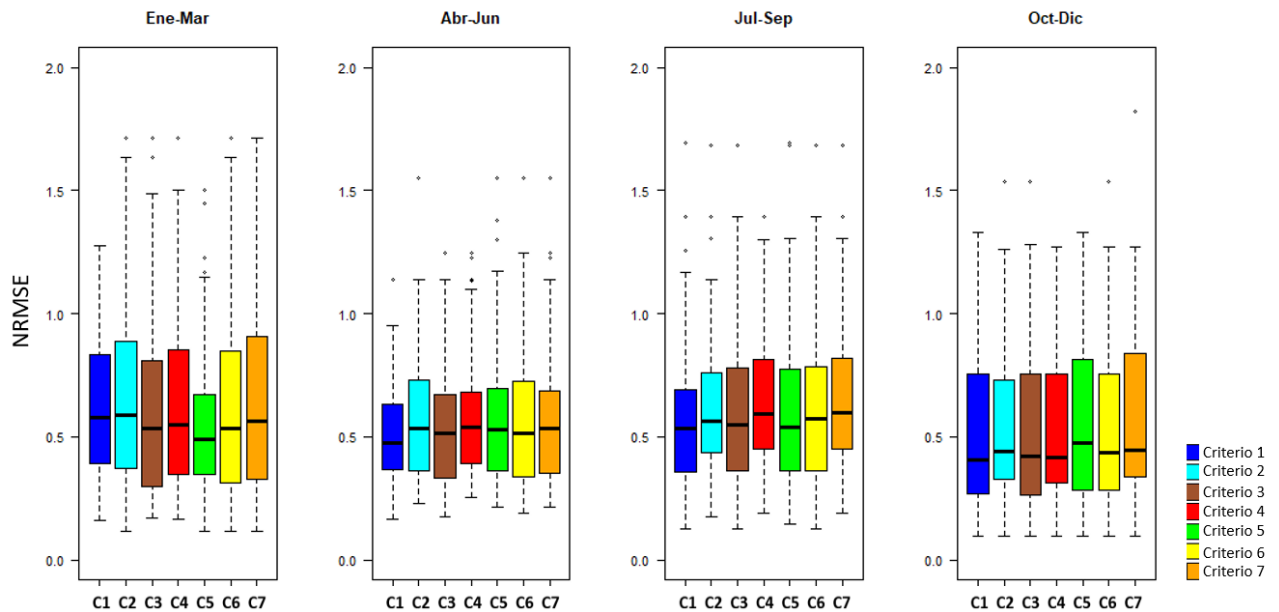


Figura 4.4: Valores de NRMSE obtenidos del proceso de validación cruzada en el relleno de caudales medios estacionales. Cada caja comprime el resultado de 70 estaciones, las cuales poseen más de nueve puntos en validación cruzada.

Las Figuras 4.5, 4.6, 4.7 y 4.8 muestran los árboles obtenidos para los periodos de enero a marzo, abril a junio, julio a septiembre y octubre a diciembre respectivamente.

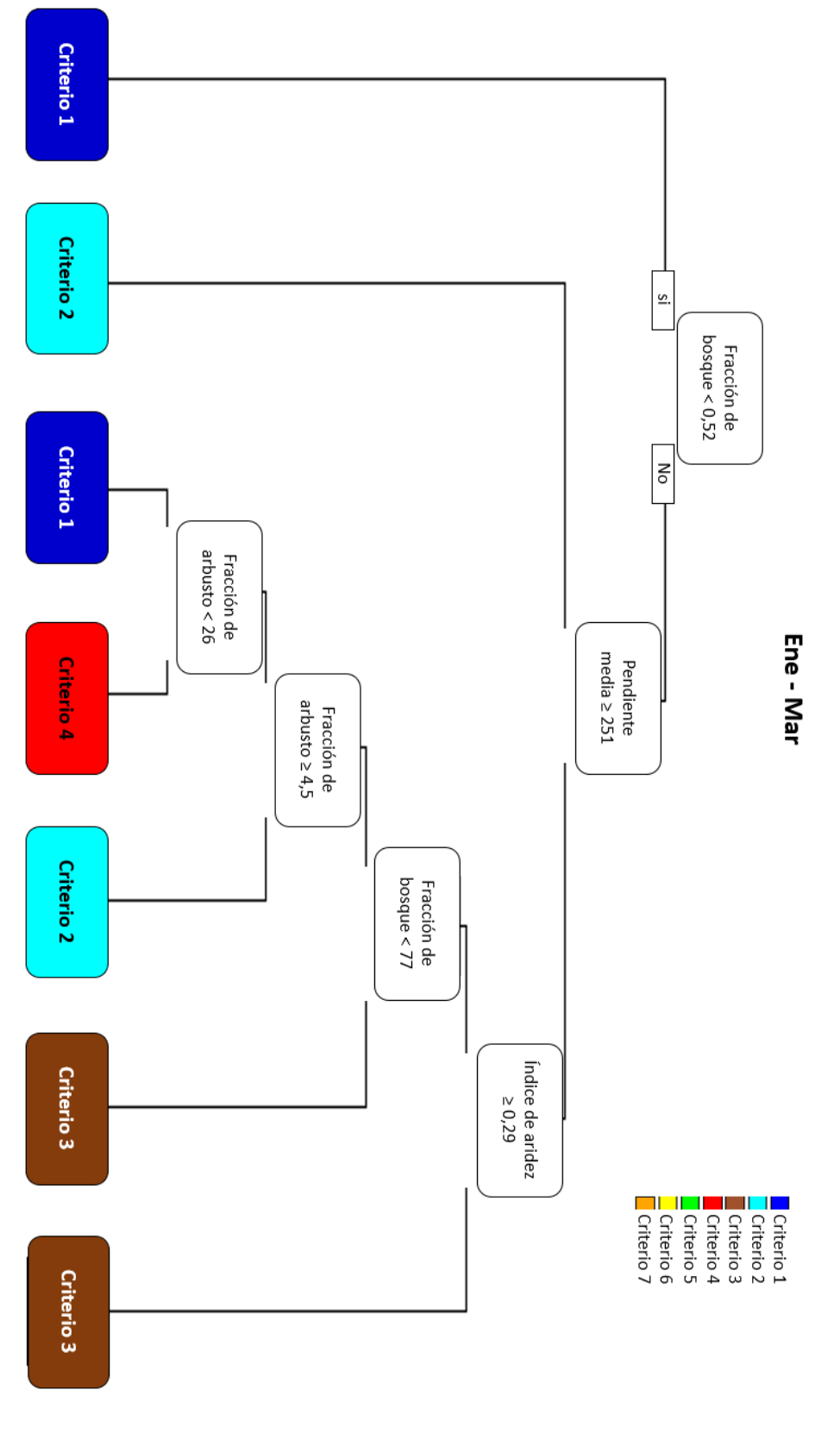


Figura 4.5: Árbol de decisión que entrega el mejor criterio para rellenar caudales medios estacionales de enero - marzo, utilizando como entrada un set de atributos físicos y climáticos.

Abr - Jun

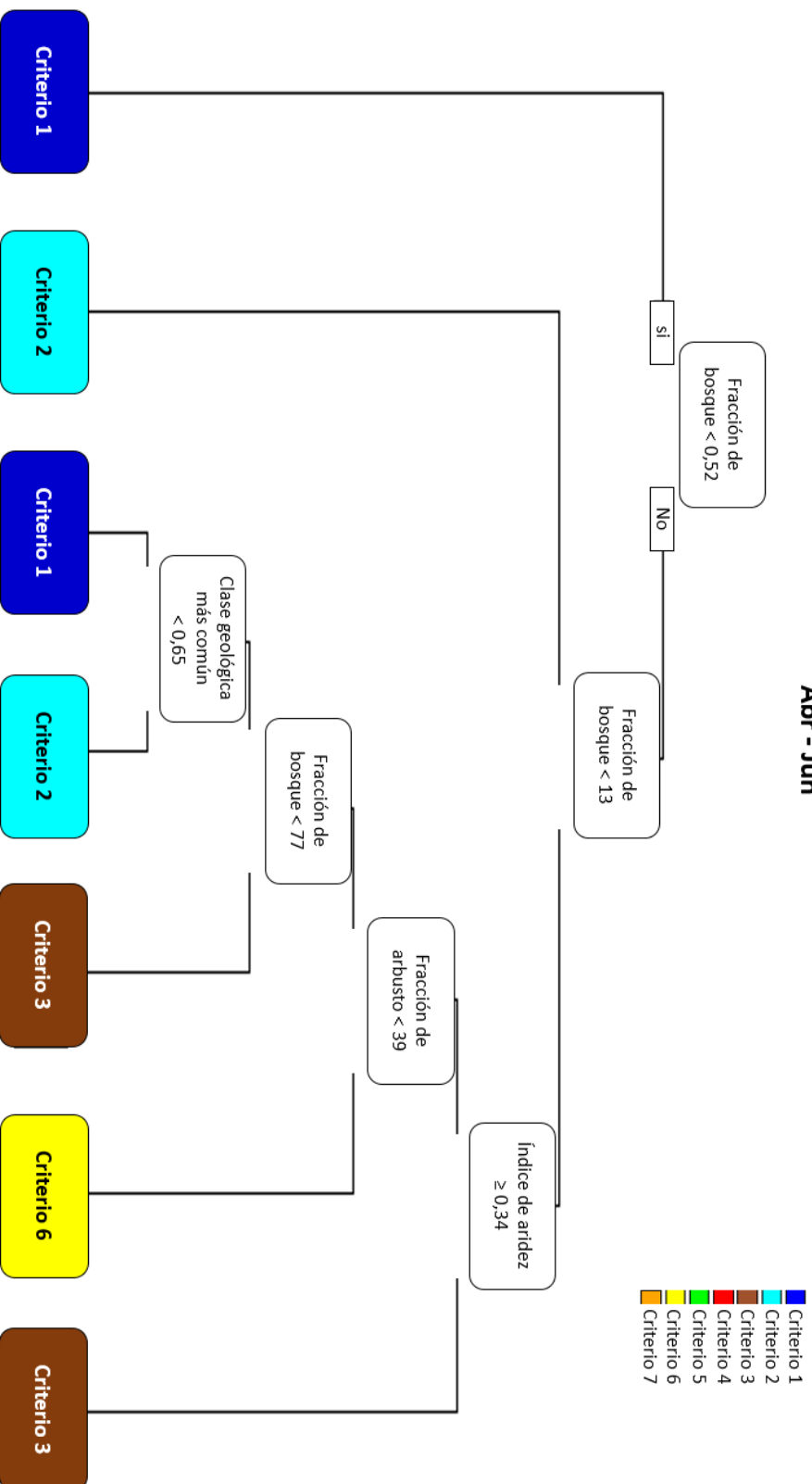


Figura 4.6: Árbol de decisión que entrega el mejor criterio para rellenar caudales medios estacionales de abril - junio, utilizando como entrada un set de atributos físicos y climáticos.

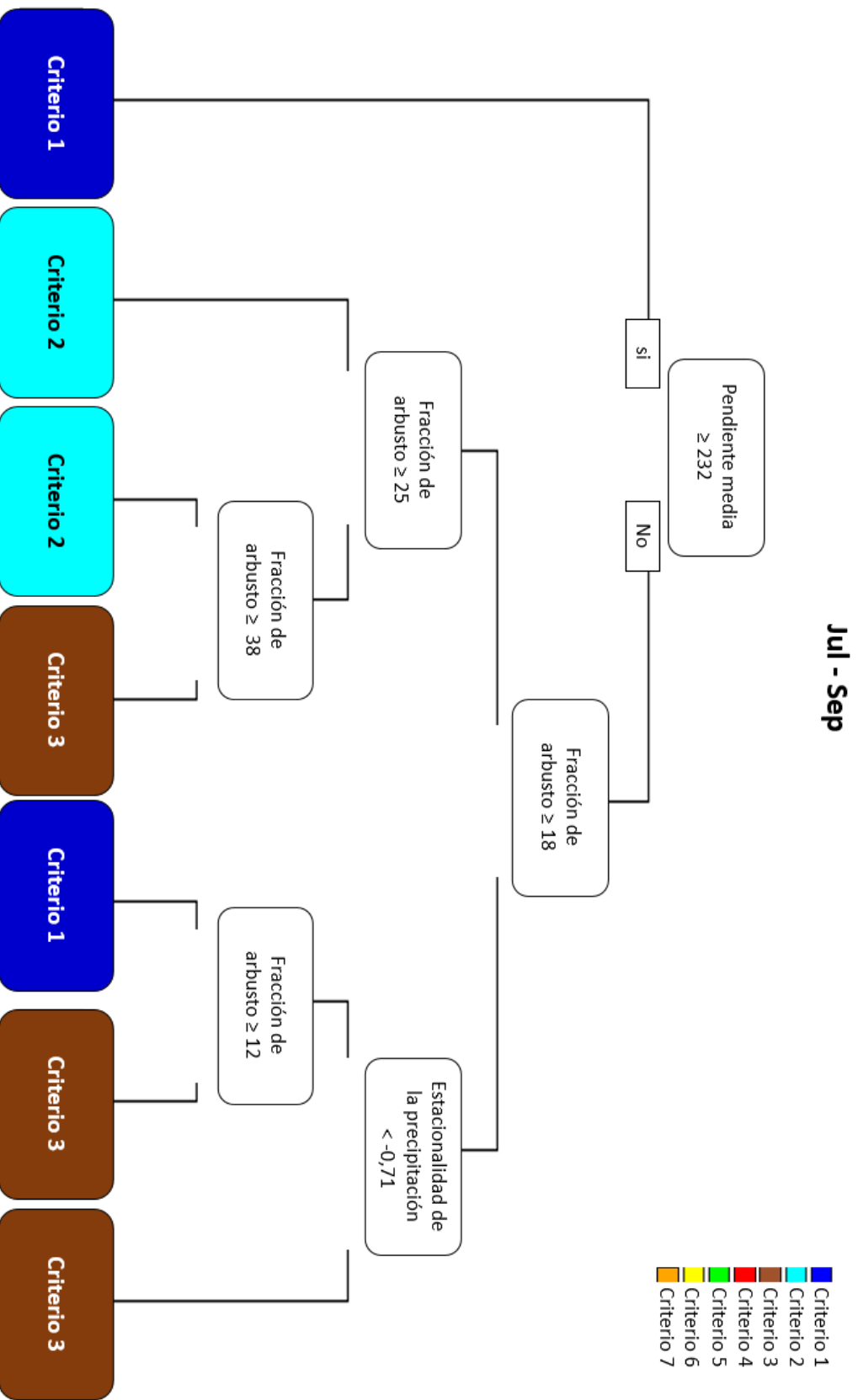


Figura 4.7: Árbol de decisión que entrega el mejor criterio para rellenar caudales medios estacionales de julio - septiembre, utilizando como entrada un set de atributos físicos y climáticos.



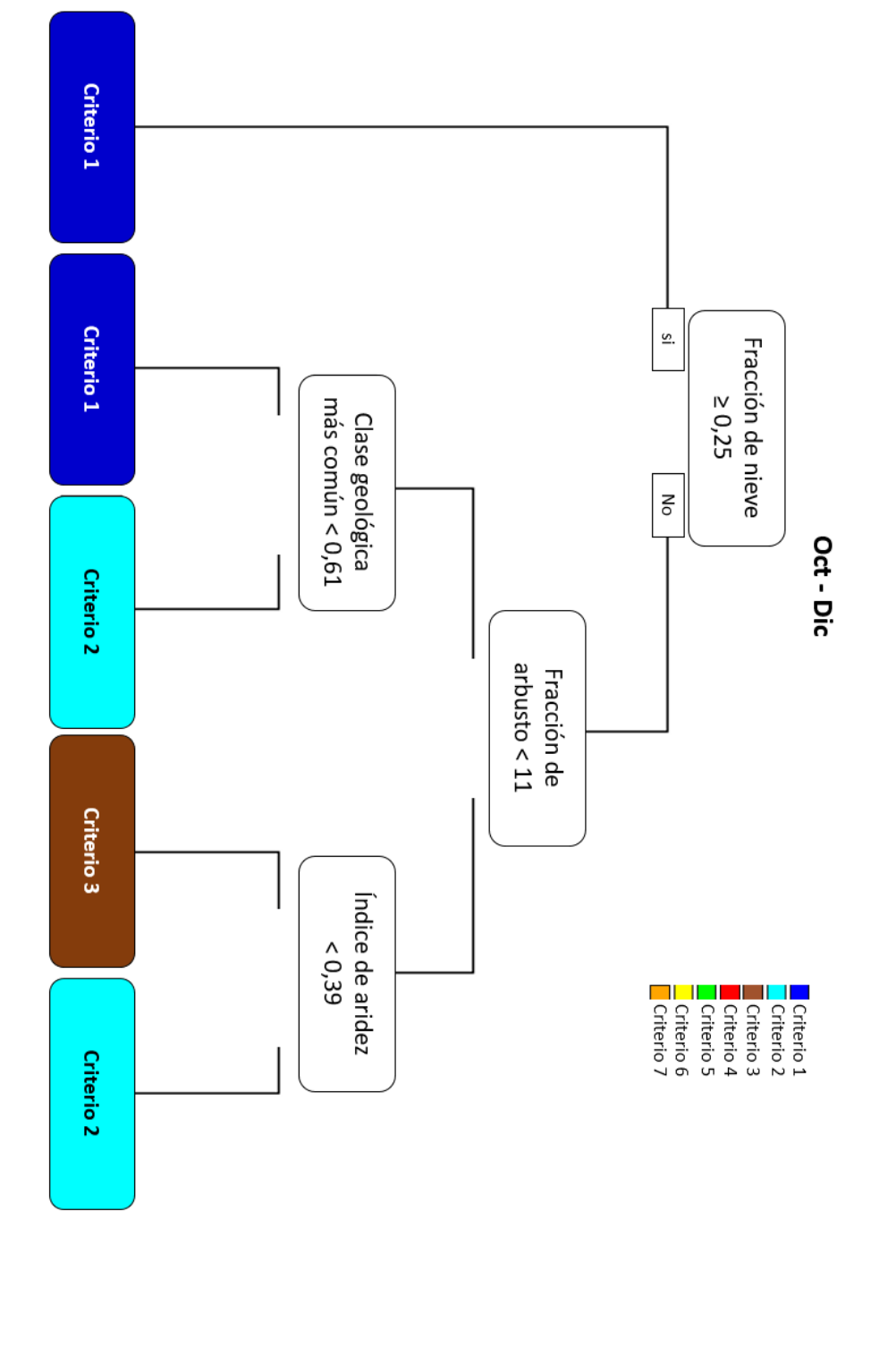


Figura 4.8: Árbol de decisión que entrega el mejor criterio para rellenar caudales medios estacionales de octubre - diciembre, utilizando como entrada un set de atributos físicos y climáticos.

Teniendo como referencia los árboles de decisión, se expone una tabla resumen con los tres atributos más importantes para cada periodo.

Tabla 4.1: Resumen de los atributos más importantes por periodo para estimar caudales medios estacionales.

Atributos	Periodo			
	Enero- Marzo	Abril- Junio	Julio- Septiembre	Octubre- Diciembre
1	Fracción de bosque	Fracción de bosque	Pendiente media	Fracción de nieve
2	Pendiente media	Índice de aridez	Fracción de arbusto	Fracción de arbusto
3	Índice de aridez	Fracción de arbusto	Estacionalidad de la precipitación	Índice de aridez

El atributo más importante para todos los periodos corresponde a uno proveniente de la similitud física, siendo la fracción de arbusto el más utilizado en los cuatro periodos, esto tiene sentido como un primer filtro, luego se observa que todos los árboles incluyen un atributo climático entre los tres más importantes. Esto puede interpretarse como la forma de crear un nexo entre cuencas que tengan una condición geomorfológica y una aridez similar.

Finalmente, se rellena cada estación para cada intervalo temporal según el mejor criterio encontrado, definido mediante análisis estadístico o a través del árbol de decisión ajustado, en caso de tener pocos registros. La Figura 4.9 muestra los resultados del proceso de relleno de caudales medios estacionales con las correcciones definidas para que exista coherencia con la escala temporal anual, donde cada color indica aquel criterio que minimiza el NRMSE. El histograma indica la distribución de mejores criterios por cuenca, seguido por un gráfico de validación cruzada con la información de todas las cuencas para cada intervalo de tiempo.

Se observa que las validaciones cruzadas entregan un coeficiente de determinación mayor a 0,93 y un NRMSE menor a 0,3, siendo los periodos de enero a marzo y julio a septiembre los que poseen mejores e iguales resultados ( $R^2 = 0,98$  y  $\text{NRMSE} = 0,11$ ). Las métricas reflejan un mejor desempeño de la metodología que en la escala temporal anual. En los histogramas se observa que el criterio 1 es el más utilizado desde enero a junio, pero luego de esto los criterios 2 y 3 cobran mayor relevancia. Lo anterior tiene sentido, puesto que al dividir el año cada tres meses, el régimen hidrológico de las cuencas es más importante para estimar caudales que en una escala temporal más gruesa.

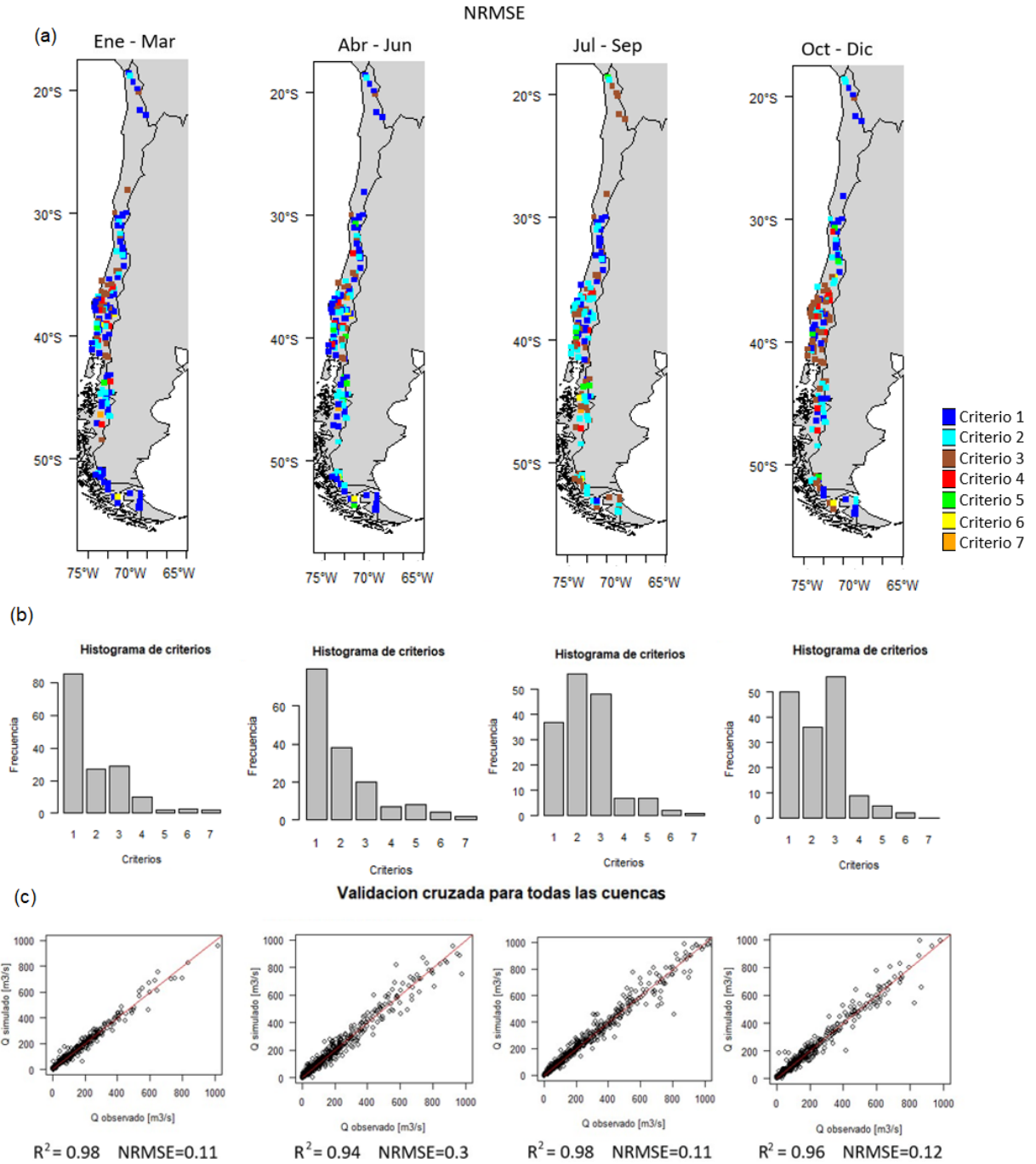


Figura 4.9: Resultados del relleno de caudales medios estacionales: (a) gráfico final con el mejor criterio para las 158 estaciones, (b) histograma que indica la frecuencia con la cual un criterio es identificado como el mejor, (c) validación cruzada para la escala estacional con todas las estaciones y años hidrológicos utilizados, para los cuatro periodos definidos.

### 4.3. Caudal medio mensual

La Figura 4.10 muestra los resultados de NRMSE obtenidos en la validación cruzada para cada criterio utilizado. El criterio 1 es superior al resto durante todo el año, exceptuando los meses de agosto y octubre, donde el criterio 5 tiene los mejores resultados. Si bien, en ambos meses los criterios 1 y 5 tienen una mediana de aproximadamente 0,5, la dispersión del criterio 1 es mayor.

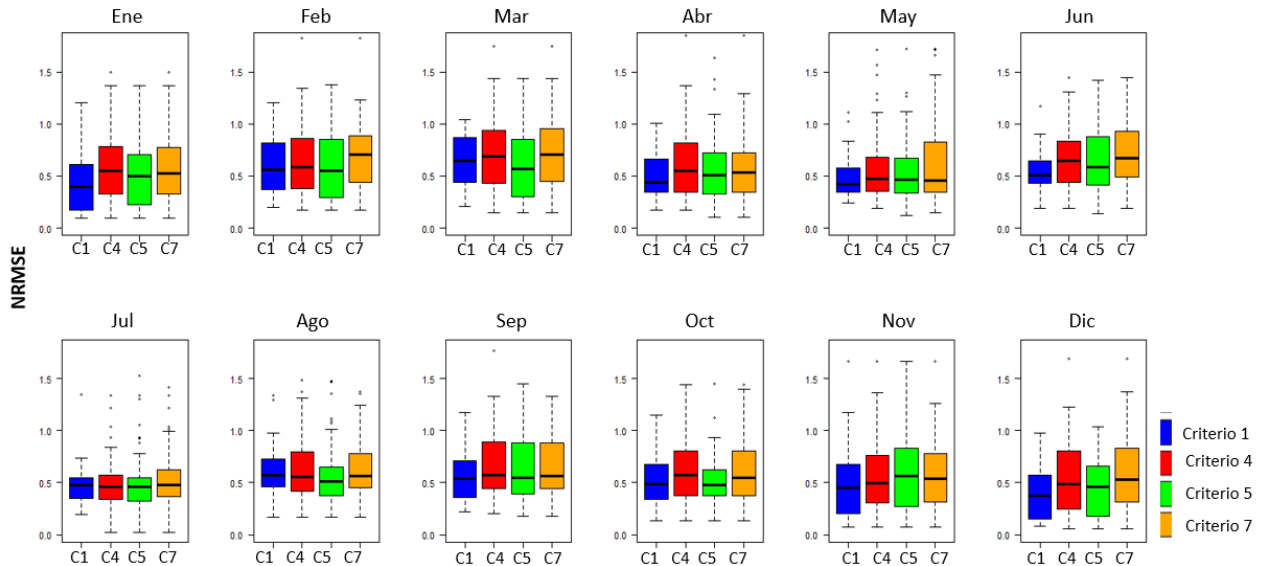


Figura 4.10: Resultados del proceso de validación cruzada en el relleno de caudales medios mensuales.

Siguiendo con la estructura metodológica, se ajustan 12 árboles de decisión que permitan predecir el mejor criterio de relleno, para cada mes, en cuencas con poca información (Anexo A). Además, se expone la Tabla 4.2 que resume los tres atributos más importante para cada mes según dichos árboles.

Tabla 4.2: Resumen de los atributos más importantes por mes

Meses	Atributos		
	1	2	3
Enero	Elevación media	Pendiente media	Fracción de arbusto
Febrero	Fracción de pasto	Tierras de cultivo	Fracción de arbusto
Marzo	Índice de aridez	Fracción de nieve	Clase geológica más común
Abril	Falta de cobertura terrestre	Fracción de arbusto	Estacionalidad de la precipitación
Mayo	Estacionalidad de la precipitación	Pendiente media	Fracción de nieve
Junio	Estacionalidad de la precipitación	Fracción de arbusto	Falta de cobertura terrestre
Julio	Elevación media	Tierras de cultivo	Fracción de pasto
Agosto	Falta de cobertura terrestre	Fracción de bosque	-
Septiembre	Tierras de cultivo	Porcentaje de área nival	Fracción de arbusto
Octubre	Estacionalidad de la precipitación	Fracción de arbusto	Clase geológica más común
Noviembre	Fracción de nieve	Estacionalidad de la precipitación	Pendiente media
Diciembre	Fracción de bosque	Porcentaje de área nival	Fracción de nieve

La Figura 4.11 muestra los resultados del proceso de relleno de caudales medios mensuales, donde cada color indica aquel criterio que minimiza el NRMSE. El histograma indica la distribución de mejores criterios por cuenca, seguido por un gráfico de validación cruzada con la información de todas las cuencas para cada mes. Se observa que el criterio más utilizado es el número 4, con excepción de los meses enero y marzo que tienen como criterio de mayor frecuencia el número 5. Esto indica que para el primer semestre del año la similitud hidrológica y física tienen un rol fundamental para estimar caudales medios mensuales. Por otra parte, el criterio 7 (combinación de las tres similitudes) es el que tiene la frecuencia más baja en todos los meses, por lo que no es recomendable estimar caudales medios mensuales utilizando las tres similitudes simultaneas como mecanismo de búsqueda para una cuenca donante. Enero es el mes que cuenta con los mejores resultados ( $R^2 = 0,98$ ), mientras que, la diferencia entre el coeficiente de determinación del mes con los mejores (enero) y peores (mayo) resultados es de 0,08.

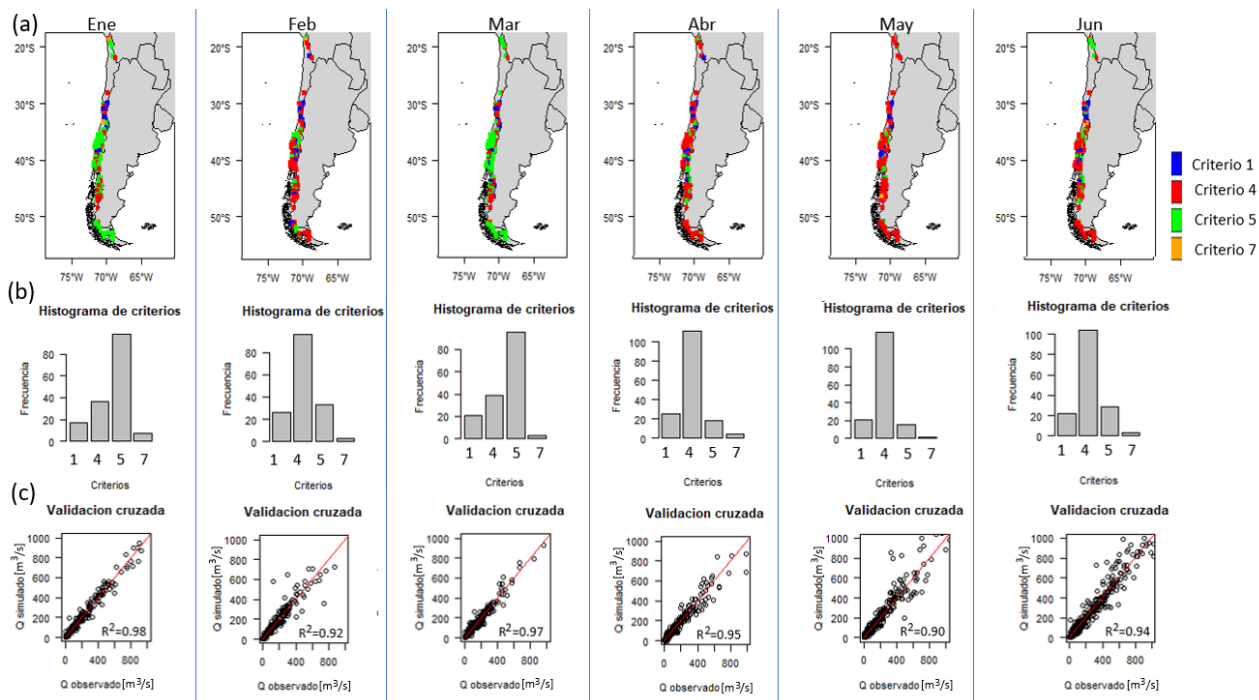


Figura 4.11: Resultados del relleno de caudales medios mensuales desde enero a junio: (a) ubicación geográfica de estaciones, (b) histograma de mejores criterios, y (c) validación cruzada.

La Figura 4.12 muestra los resultados del proceso de relleno de caudales medios mensuales para el segundo semestre del año. Se observa que el criterio más utilizado es el número 4, con excepción del mes de julio que tiene como criterio de mayor frecuencia el número 5. Esto indica que al igual que en el primer semestre, en el segundo semestre del año la similitud hidrológica y física tienen un rol fundamental para estimar caudales medios mensuales. Además, hay que destacar que el análisis realizado para el criterio 7 se repite, es decir, no se recomienda estimar caudales medios mensuales utilizando las tres similitudes como mecanismo de búsqueda para una cuenca donante. Julio y diciembre son los meses que cuentan con los mejores e iguales resultados ( $R^2 = 0,97$ ), mientras que, la diferencia entre el coeficiente de determinación del mes con los mejores y peores resultados es de 0,01, por lo que hay una dispersión menor que en el primer semestre.

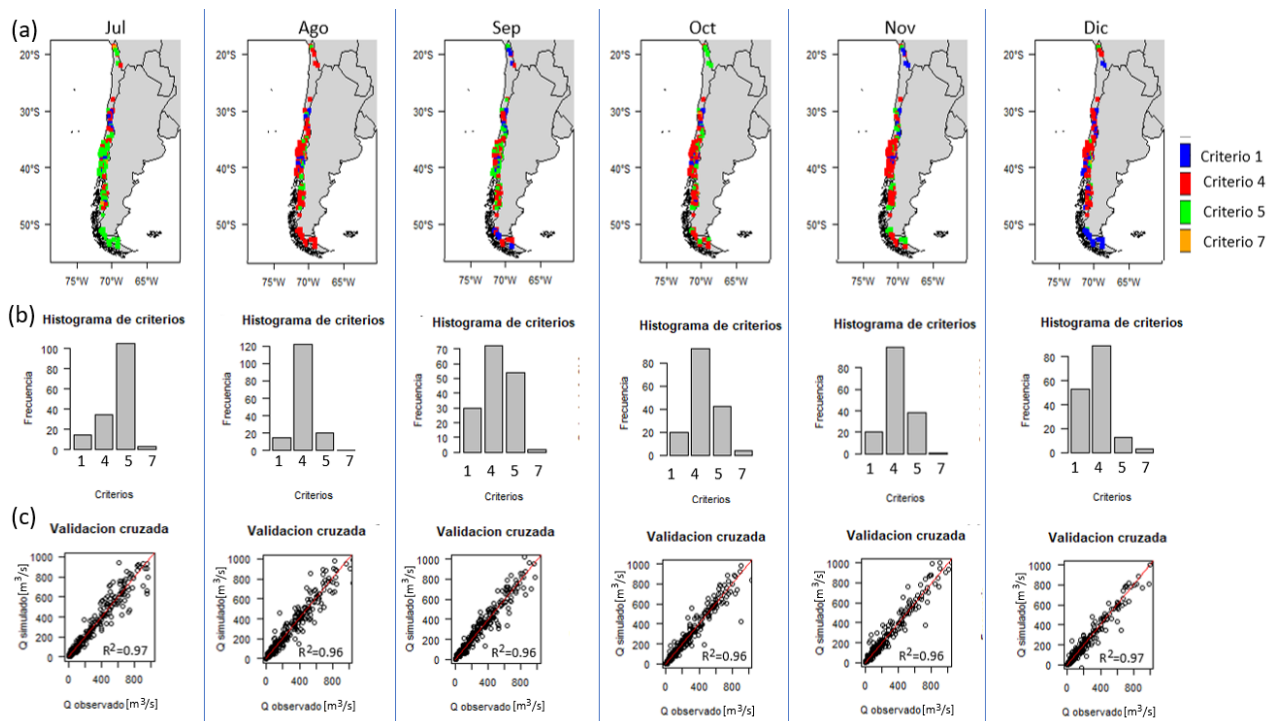


Figura 4.12: Resultados del relleno de caudales medios mensuales desde julio a diciembre: (a) ubicación geográfica de estaciones, (b) histograma de mejores criterios, y (c) validación cruzada.



## 4.4. Caudal medio diario

Las Figuras 4.13, 4.14, 4.15 y 4.16 muestran los resultados de ejecutar el método K-NN con las formas de condicionamiento propuestas en la metodología, considerando un número K de vecinos más cercanos igual a cinco. Si bien la metodología se aplica considerando todos los años desde 1979 a 2016, sólo se muestra la serie temporal para el periodo de referencia 2011-2015, para comprender visualmente la aplicación del método. En esta última, se muestran en color naranja los caudales estimados, que resultan de eliminar aleatoriamente brechas temporales de 60 días de la serie original. Adicionalmente, se expone la validación cruzada resultante de estimar todos los caudales medios diarios en el periodo abril – septiembre, octubre – marzo y del año completo. Cabe destacar que, cada forma de condicionar el método se realiza en su ventana óptima de comparación, la ejecución en distintas ventanas temporales de cada estación fluviométrica se encuentra en el anexo B del informe.

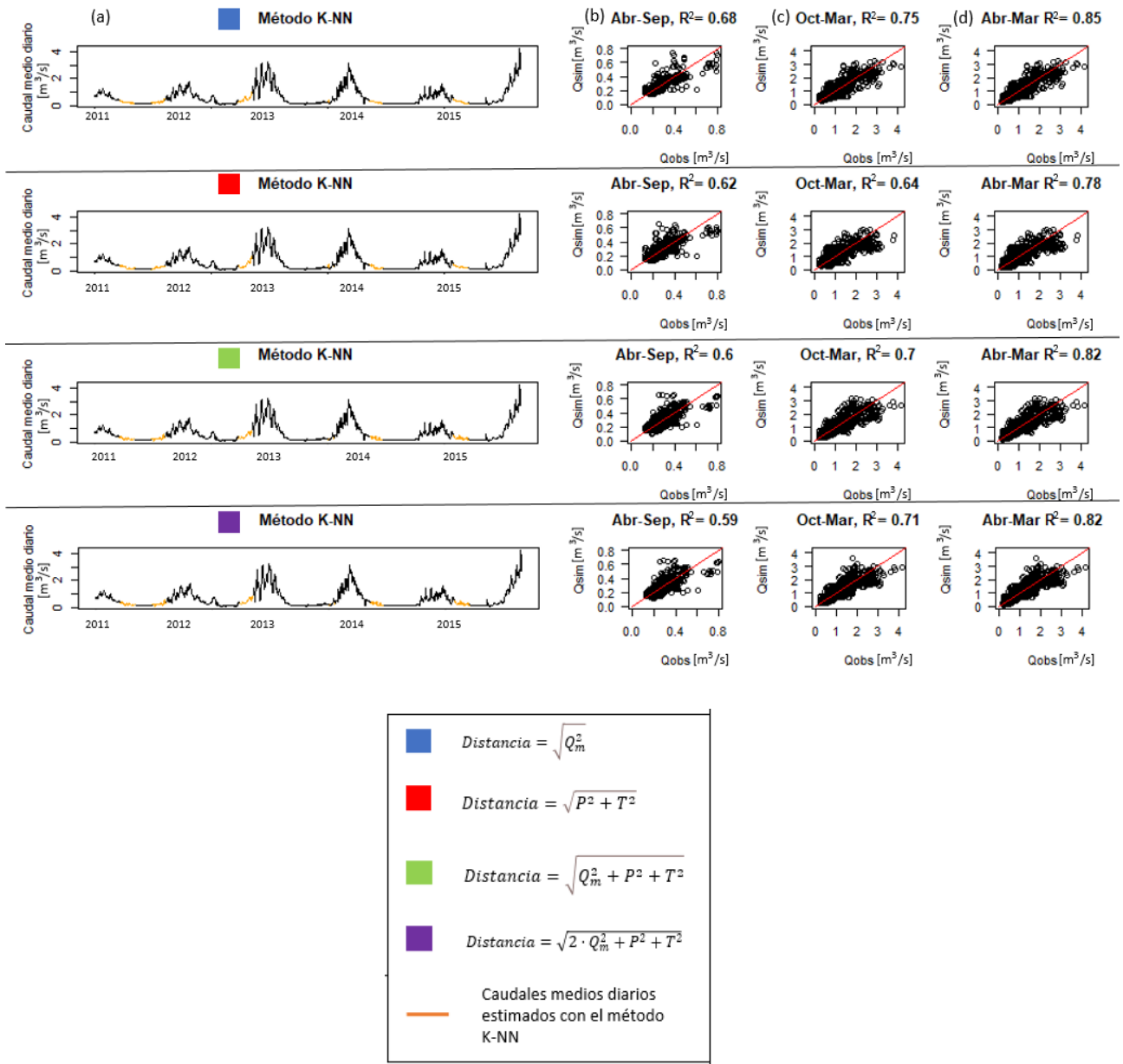


Figura 4.13: Relleno de caudales medios diarios para la estación Estero Yerba Loca antes junta San Francisco: (a) serie temporal en un periodo de referencia 2011-2015; (b) validación cruzada para el periodo abril-septiembre; (c) validación cruzada para el periodo octubre-marzo; (d) validación cruzada para el año hidrológico completo.

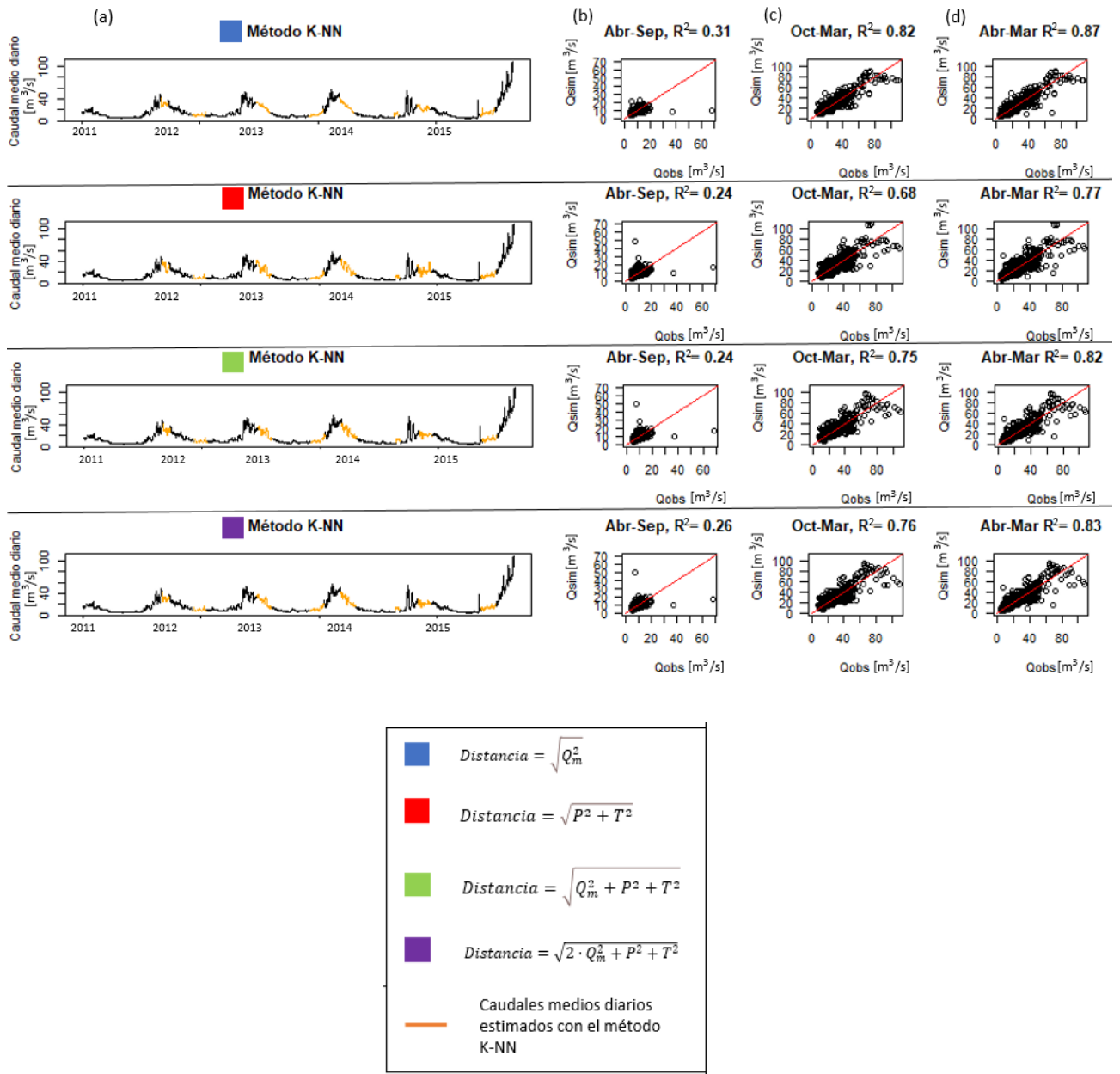


Figura 4.14: Relleno de caudales medios diarios para la estación Río Aconcagua en Chacabucuito: (a) serie temporal en un periodo de referencia 2011-2015; (b) validación cruzada para el periodo abril-septiembre; (c) validación cruzada para el periodo octubre-marzo; (d) validación cruzada para el año hidrológico completo.

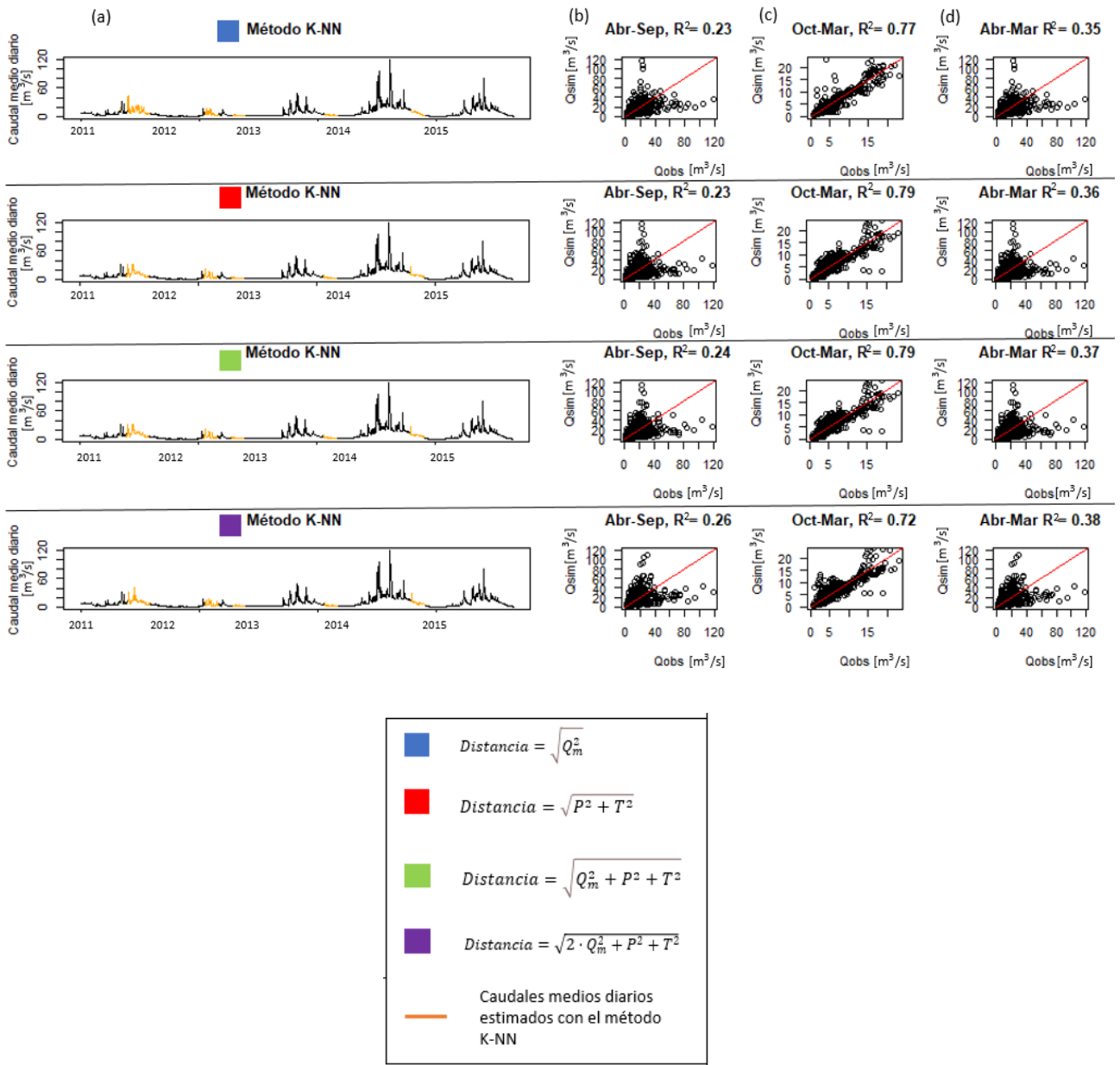


Figura 4.15: Relleno de caudales medios diarios para la estación Río Andalién Camino a Penco: (a) serie temporal en un periodo de referencia 2011-2015; (b) validación cruzada para el periodo abril-septiembre; (c) validación cruzada para el periodo octubre-marzo; (d) validación cruzada para el año hidrológico completo.

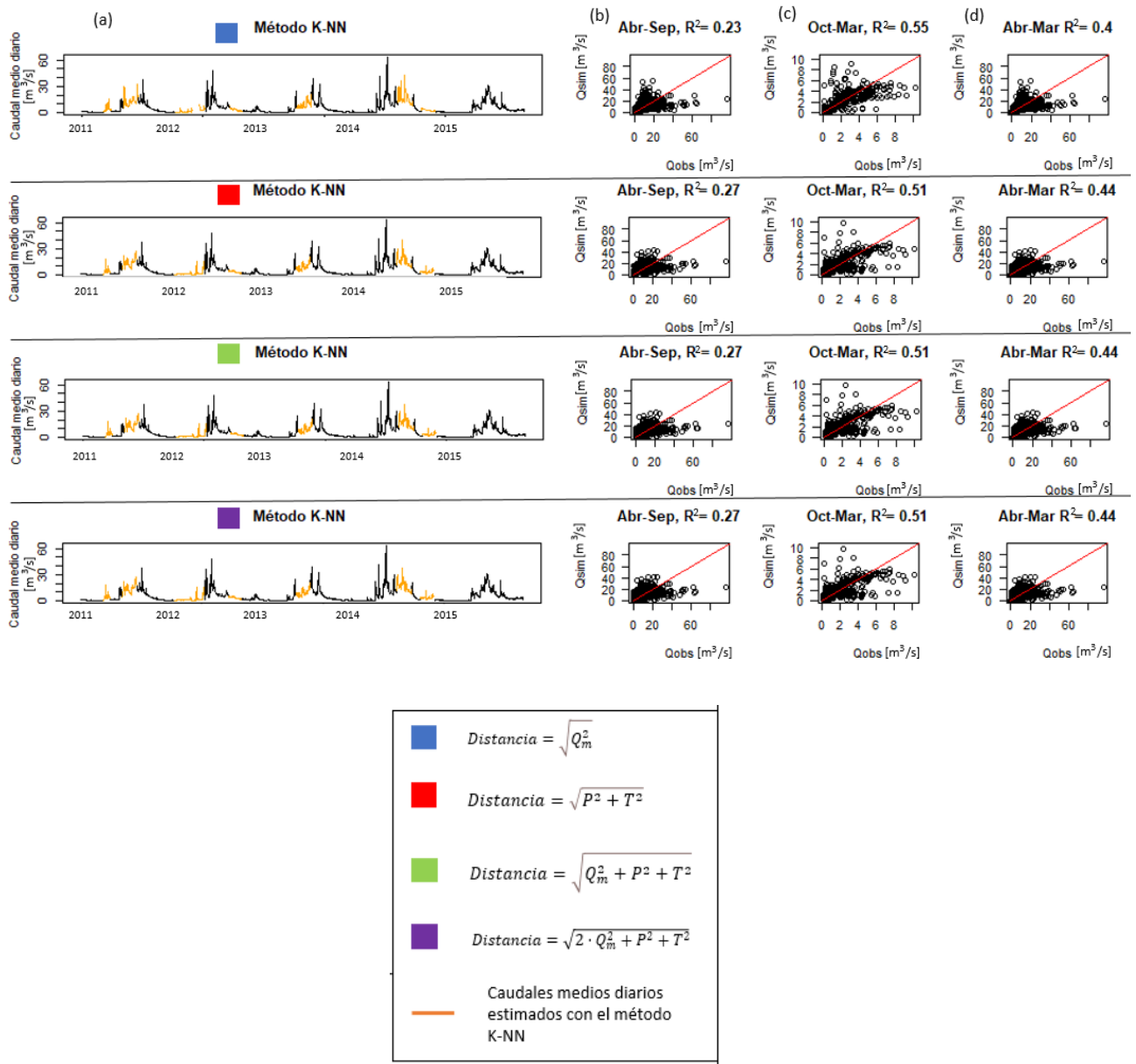


Figura 4.16: Relleno de caudales medios diarios para la estación Río Lirquén en Cerro el Padre: (a) serie temporal en un periodo de referencia 2011-2015; (b) validación cruzada para el periodo abril-septiembre; (c) validación cruzada para el periodo octubre-marzo; (d) validación cruzada para el año hidrológico completo.

De las figuras anteriores, se puede observar que las estaciones Estero Yerba Loca antes junta San Francisco y Río Aconcagua en Chacabuquito, de régimen nival, tienen mejores resultados a nivel anual ( $R^2 = 0,85$  y  $R^2 = 0,87$  respectivamente) considerando el caudal medio mensual como variable para condicionar el método (Figura 4.13a y 4.14a). Por otra parte, para las estaciones Río Andalién Camino a Penco y Río Lirquén en Cerro el Padre, de régimen pluvial, se obtienen mejores resultados a nivel anual ( $R^2 = 0,38$  y  $R^2 = 0,44$  respectivamente) al considerar el caudal medio mensual, la precipitación y la temperatura como variables que condicionan el método.

Si bien los resultados en validación cruzada para el año completo difieren significativamente para las cuencas en régimen pluvial y nival, esta diferencia es mínima en las estimaciones realizadas entre octubre a marzo. Las cuatro estaciones piloto tienen un coeficiente de determinación mayor a 0,5 en validación cruzada entre octubre - marzo, esto indica que el rendimiento de la metodología funciona para este periodo. Lo anterior cobra sentido al observar las series temporales, en las cuales se muestra que en las cuencas nivales el caudal crece de forma suave entre octubre y marzo, manteniendo un patrón de crecida sin variaciones abruptas. Por otro lado, las cuencas en régimen pluvial poseen un caudal bajo (aproximadamente 10% del caudal de abril a septiembre) entre octubre y marzo, que consta de un flujo prácticamente constante.

# Capítulo 5

## Conclusiones

En este trabajo, se estiman datos faltantes en los registros fluviométricos de 158 estaciones asociadas a cuencas con un régimen hidrológico cercano al natural, a distintas escalas temporales, utilizando el NRMSE como indicador de mejor criterio. En este contexto, se extraen las siguientes conclusiones:

- En relación con los caudales medios anuales, el criterio más utilizado es aquel que incorpora sólo atributos físicos (criterio 3), tales como elevación media, pendiente media, fracción de bosque, fracción de pasto, fracción de arbusto, clase geológica más común, tierras de cultivo, falta de cobertura terrestre y porcentaje de área nival. Esto se acentúa al norte de los 30° y al sur de los 40° de latitud en Chile continental, donde se observa que para más del 90 % de las cuencas el criterio 3 resulta ser el mejor para ejecutar el relleno.
- Para la estimación de caudales medios estacionales, se corrigen los valores en base a los datos anuales, y se observa que no existe una estratificación geográfica para el criterio óptimo. Sin embargo, en los periodos de enero a marzo y de abril a junio, el criterio más utilizado comprende sólo la similitud hidrológica, definida como la distancia euclidiana entre Coeficientes de Pardé. Para el periodo de julio a septiembre, el criterio más utilizado incorpora sólo atributos climáticos (índice de aridez, estacionalidad de la precipitación y fracción de nieve) y, finalmente, para el periodo de octubre a diciembre, el criterio más utilizado posee sólo atributos físicos. Esto indica que, al analizar el conjunto completo de 158 cuencas, en el primer semestre del año se pueden estimar caudales medios estacionales sólo con similitud hidrológica, posteriormente, el clima y la geomorfología cobran relevancia para el desarrollo de las estimaciones.
- Para estimar caudales medios mensuales, el criterio más utilizado como mecanismo de búsqueda para una cuenca donante debe combinar similitud hidrológica y física, exceptuando los meses de enero, marzo y julio, en los cuales es mejor utilizar similitud hidrológica y climática.
- Enero presenta los mejores resultados en validación cruzada ( $R^2 = 0,98$  y  $\text{NRMSE} = 0,11$ ) para la estimación de caudales medios mensuales, siendo las métricas calculadas incluso mejores que para la escala anual ( $R^2 = 0,97$  y  $\text{NRMSE} = 0,16$ ). Lo anterior indica que la dispersión de los resultados al aumentar la resolución temporal es despreciable

debido al factor de corrección entre escalas temporales.

- En todas las escalas temporales, el criterio que combina las tres similitudes (física, climática e hidrológica) tiene los peores resultados, y es el menos utilizado a partir del análisis de los histogramas. Por lo tanto, no se recomienda utilizar este criterio como búsqueda de cuencas donantes en ninguna escala temporal.
- En relación con los caudales medios diarios, los resultados de validación cruzada indican que condicionar el método K-NN con el caudal medio mensual como variable independiente, entrega mejores resultados a nivel anual para las estaciones Estero Yerba Loca antes junta San Francisco y Río Aconcagua en Chacabuquito, con  $R^2 = 0,85$  y  $R^2 = 0,87$  respectivamente. Por otra parte, para las estaciones Río Andalién camino a Penco y Río Lirquén en Cerro el Padre, los mejores resultados a nivel anual se obtienen al incorporar el caudal medio mensual, precipitación y temperatura como variables del método de cálculo de distancia, aunque los valores del coeficiente de determinación ( $R^2 = 0,38$  y  $R^2 = 0,44$  respectivamente) no son aceptables.
- Para las cuatro estaciones fluviométricas piloto utilizadas en la estimación de caudales medios diarios, se obtienen mejores resultados en el periodo octubre a marzo. Lo anterior se debe a que, durante el periodo de deshielo, no hay crecidas significativas en las cuencas nivales mientras que, en las cuencas pluviales, en este periodo el caudal es bajo y prácticamente constante. Por lo tanto, la metodología propuesta tiene un mejor rendimiento entre estos meses independiente del régimen de la cuenca.

Como trabajo futuro, se propone condicionar el método K-NN comparando descriptores de cuencas vecinas, e incluir otros métodos paramétricos y no paramétricos en el análisis.



# Bibliografía

Addor, N., Nearing, G., Prieto, C., Newman, A. J., Le Vine, N., Clark, M. P. (2018). A Ranking of Hydrological Signatures Based on Their Predictability in Space. *Water Resources Research*, 54(11), 8792–8812. <https://doi.org/10.1029/2018WR022606>

Addor N, Newman AJ, Mizukami N, Clark MP. (2017). The CAMELS data set: Catchment attributes and meteorology for large-sample studies. *Hydrology and Earth System Sciences* <https://doi.org/10.5194/hess-21-5293-2017>

Álvarez-Garreton C, Mendoza PA, Boisier JP, Addor N, Galleguillos M, Zambrano-Bigiarini M, Lara A, Puelma C, Cortes G, Garreaud R, et al. (2018). The CAMELS-CL dataset: catchment attributes and meteorology for large sample studies – Chile dataset. *Hydrology and Earth System Sciences* 22 (11): 5817–5846 <https://doi.org/10.5194/hess-22-5817-2018>

Beck HE, van Dijk AIJM, de Roo A, Miralles DG, McVicar TR, Schellekens J, Bruijnzeel LA. (2016). Global-scale regionalization of hydrologic model parameters. *Water Resources Research* 52 (5): 3599–3622 <https://doi.org/10.1002/2015WR018247>

Butler, D. (2014). Earth observation enters next phase. *Nature*, 508(7495), 160–161. <https://doi.org/10.1038/508160a>

Chagas, V. B., Chaffe, P. L., Addor, N., Fan, F. M., Fleischmann, A. S., Paiva, R. C., Siqueira, V. A. (2020). CAMELS-BR: hydrometeorological time series and landscape attributes for 897 catchments in Brazil. *Earth System Science Data*, 12(3), 2075–2096.

Chegwidden OS, Nijssen B, Rupp DE, Arnold JR, Clark MP, Hamman JJ, Kao S, Mao Y, Mizukami N, Mote P, et al. (2019). How do modeling decisions affect the spread among hydrologic climate change projections? Exploring a large ensemble of simulations across a diversity of hydroclimates. *Earth’s Future* 7: 623–637 <https://doi.org/10.1029/2018EF001047>

Chen J, Brissette FP, Poulin A, Leconte R. (2011). Overall uncertainty study of the hydrological impacts of climate change for a Canadian watershed. *Water Resources Research* 47: W12509 <https://doi.org/10.1029/2011WR010602>

Cortés G, Vargas X, McPhee J. (2011). Climatic sensitivity of streamflow timing in the extratropical western Andes Cordillera. *Journal of Hydrology* 405 (1–2): 93–109 <https://doi.org/10.1016/j.jhydrol.2011.05.013>

Coxon, G., Addor, N., Bloomfield, J. P., Freer, J., Fry, M., Hannaford, J., ... Woods, R. (2020). CAMELS-GB: hydrometeorological time series and landscape attributes for 671 catchments in Great Britain. *Earth System Science Data*, 12(4), 2459-2483.

Dawson, C. W., Harpham, C., Wilby, R. L., Chen, Y. (2002). Evaluation of artificial neural network techniques for flow forecasting in the River Yangtze, China. *Hydrology and Earth System Sciences*, 6(4), 619–626. <https://doi.org/10.5194/hess-6-619-2002>

Dembélé, M., Oriani, F., Tumbulto, J., Mariéthoz, G., Schaeffli, B. (2019). Gap-filling of daily streamflow time series using Direct Sampling in various hydroclimatic settings. *Journal of Hydrology*, 569(December), 573–586. <https://doi.org/10.1016/j.jhydrol.2018.11.076>

DGA. (2017). Actualización del Balance Hídrico Nacional. SIT N° 417, Ministerio de Obras Públicas, Dirección General de Aguas, División de Estudios y Planificación, Santiago, Chile, Realizado Por: Universidad de Chile Pontificia Universidad Católica de Chile, 378. <https://snia.mop.gob.cl/sad/REH5796v1.pdf>

DGA. (2018). Aplicación de La Metodología de Actualización del Balance Hídrico Nacional en las Cuencas de la Macrozona Norte y Centro

DGA. (2019). Aplicación de La Metodología de Actualización del Balance Hídrico Nacional en las Cuencas de la Macrozona Sur y Parte de la Macrozona Austral

di Piazza, A., Conti, F. Lo, Noto, L. V., Viola, F., La Loggia, G. (2011). Comparative analysis of different techniques for spatial interpolation of rainfall data to create a serially complete monthly time series of precipitation for Sicily, Italy. *International Journal of Applied Earth Observation and Geoinformation*, 13(3), 396–408. <https://doi.org/10.1016/j.jag.2011.01.005>

Gupta, H. V., Perrin, C., Blöschl, G., Montanari, A., Kumar, R., Clark, M., Andréassian, V. (2014). Large-sample hydrology: A need to balance depth with breadth. *Hydrology and Earth System Sciences*, 18(2), 463–477. <https://doi.org/10.5194/hess-18-463-2014>

Gupta, H.V., Wagener, T. and Liu, Y. (2008), Reconciling theory with observations: elements of a diagnostic approach to model evaluation. *Hydrol. Process.*, 22: 3802-3813. <https://doi.org/10.1002/hyp.6989>

Henn, B., Raleigh, M. S., Fisher, A., Lundquist, J. D. (2013). A comparison of methods for filling gaps in hourly near-surface air temperature data. *Journal of Hydrometeorology*, 14(3), 929–945. <https://doi.org/10.1175/JHM-D-12-027.1>

Hrachowitz M, Savenije HHG, Blöschl G, McDonnell JJ, Sivapalan M, Pomeroy JW, Arheimer B, Blume T, Clark MP, Ehret U, et al. (2013). A decade of Predictions in Ungauged Basins (PUB)—a review. *Hydrological Sciences Journal* 58 (6): 1198–1255 <https://doi.org/10.1080/02626667.2013.803183>

James, G., Witten, D., Hastie, T., Tibshirani, R. (2013). An introduction to statical learning with application in R. In *Springer Texts in Statistics*. <https://doi.org/10.1007/978-1-4614-7138-7>

Jehn, F. U., Bestian, K., Breuer, L., Kraft, P., Houska, T. (2020). Using hydrological and climatic catchment clusters to explore drivers of catchment behavior. *Hydrology and Earth System Sciences*, 24(3), 1081-1100.

Knoben, W. J. M., Woods, R. A., Freer, J. E. (2018). A Quantitative Hydrological Climate Classification Evaluated With Independent Streamflow Data. *Water Resources Research*, 54(7), 5088–5109. <https://doi.org/10.1029/2018WR022913>

Kuentz, A., Arheimer, B., Hundecha, Y., Wagener, T. (2017). Understanding hydrologic variability across Europe through catchment classification. *Hydrology and Earth System Sciences*, 21(6), 2863-2879.

Lall, U., Sharma, A. (1996). A nearest neighbor bootstrap for resampling hydrologic time series. *Water Resources Research*, 32(3), 679–693. <https://doi.org/10.1029/95WR02966>

Mendoza, P. A., Muñoz, E., Vásquez, N., Lagos, M. A., Vargas, X. (2020). Uso de principios Darwinianos para caracterizar el balance hídrico natural en Chile. 1(1).

Michel, C., Perrin, C., Andréassian, V., Oudin, L., Mathevet, T. (2006). Has basin-scale modelling advanced beyond empiricism? IAHS-AISH Publication, 307, 108–116.

Mwale, F. D., Adeloye, A. J., Rustum, R. (2012). Infilling of missing rainfall and streamflow data in the Shire River basin, Malawi - A self organizing map approach. *Physics and Chemistry of the Earth*, 50–52, 34–43. <https://doi.org/10.1016/j.pce.2012.09.006>

Newman AJ, Clark MP, Sampson K, Wood A, Hay LE, Bock A, Viger RJ, Blodgett D, Brekke L, Arnold JR, et al. (2015). Development of a large-sample watershed-scale hydro-meteorological data set for the contiguous USA: data set characteristics and assessment of regional variability in hydrologic model performance. *Hydrology and Earth System Sciences* 19 (1): 209–223 <https://doi.org/10.5194/hess-19-209-2015>

Nowak, K., Prairie, J., Rajagopalan, B., Lall, U. (2010). A nonparametric stochastic approach for multisite disaggregation of annual to daily streamflow. *Water Resources Research*, 46(8). <https://doi.org/10.1029/2009WR008530>

Olden, J. D., and Poff, N. L. (2003). Redundancy and the choice of hydrologic indices for characterizing streamflow regimes. *River Research and Applications* 19, pp. 101–121.

Oudin L, Andréassian V, Perrin C, Michel C, Le Moine N. (2008). Spatial proximity, physical similarity, regression and ungaged catchments: A comparison of regionalization approaches based on 913 French catchments. *Water Resources Research* 44 (3): 1–15 <https://doi.org/10.1029/2007WR006240>

Oudin L, Kay A, Andréassian V, Perrin C. (2010). Are seemingly physically similar catchments truly hydrologically similar? *Water Resources Research* 46 (11): W11558 <https://doi.org/10.1029/2009WR008887>

Pappas, C., Papalexiou, S. M., Koutsoyiannis, D. (2014). *Journal of geophysical research. Nature*, 175(4449), 238. <https://doi.org/10.1038/175238c0>

Parajka J, Viglione A, Rogger M, Salinas JL, Sivapalan M, Blöschl G. (2013). Comparative assessment of predictions in ungauged basins - Part 1: Runoff-hydrograph studies. *Hydrology and Earth System Sciences* 17 (5): 1783–1795 <https://doi.org//hess-17-1783-2013>

Peña, H., Vidal, F. (1993). Estimación estadística de la Línea de Nieves durante los Eventos de Precipitación entre las latitudes 28 y 38 grados Sur. XI Congreso Chileno de ingeniería Hidráulica

Poff, N. L. R., Olden, J. D., Pepin, D. M., Bledsoe, B. P. (2006). Placing global stream flow variability in geographic and geomorphic contexts. *River Research and Applications*, 22(2), 149–166. <https://doi.org/10.1002/rra.902>

Sawicz, K., Wagener, T., Sivapalan, M., Troch, P. A., Carrillo, G. (2011). Catchment classification: empirical analysis of hydrologic similarity based on catchment function in the eastern USA. *Hydrology Earth System Sciences Discussions*, 8(3).

Schwämmle, V., Jensen, O. N. (2010). A simple and fast method to determine the parameters for fuzzy c-means cluster analysis. *Bioinformatics*, 26(22), 2841–2848. <https://doi.org/10.1093/bioinformatics/btq534>

Sivapalan, M. (2005). Pattern, Process and Function: Elements of a Unified Theory of Hydrology at the Catchment Scale. *Encyclopedia of Hydrological Sciences*. <https://doi.org/10.1002/0470848944.hsa012>

Stewart IT, Cayan DR, Dettinger MD. (2005). Changes toward earlier streamflow timing across western North America. *Journal of Climate* 18 (8): 1136–1155 <https://doi.org/10.1175/JCLI3321.1>

Therneau, T., Atkinson, B., Ripley, B. (2019). rpart: Recursive partitioning for classification, regression and survival trees. CRAN R Package Version 4.1-15. <https://cran.r-project.org/package=rpart>

Thompson, S. E., Harman, C. J., Konings, A. G., Sivapalan, M., Neal, A., Troch, P. A. (2011). Comparative hydrology across AmeriFlux sites: The variable roles of climate, vegetation, and groundwater. *Water Resources Research*, 47(7), 1–17. <https://doi.org/10.1029/2010WR009797>

Vásquez N. et al. (2021) Catchment-Scale Natural Water Balance in Chile. In: Fernández B., Gironás J. (eds) *Water Resources of Chile*. World Water Resources, vol 8. Springer, Cham. [https://doi.org/10.1007/978-3-030-56901-3\\_9](https://doi.org/10.1007/978-3-030-56901-3_9)

Wagener, T., Wheeler, H. S., and Gupta, H. V. (2004). *Rainfall-runoff modelling in gauged and ungauged catchments*. London, UK: Imperial College Press

Wagener, T., Sivapalan, M., Troch, P., Woods, R. (2007). Catchment Classification and Hydrologic Similarity. *Geography Compass*, 1(4), 901–931. <https://doi.org/10.1111/j.1749-8198.2007.00039.x>

Westerberg, I. K., McMillan, H. K. (2015). Uncertainty in hydrological signatures. Hy-

drology and Earth System Sciences, 19(9), 3951–3968. <https://doi.org/10.5194/hess-19-3951-2015>

Wilby RL, Harris I. (2006). A framework for assessing uncertainties in climate change impacts: Low-flow scenarios for the River Thames, UK. *Water Resources Research* 42 (2): W02419 <https://doi.org/10.1029/2005WR004065>

Woods, R. A. (2004). Role of catchment classification in PUB. Presented at the IAHS PUB Workshop 2004, Perth, Western Australia, 2–5 Febr

# Anexos

# Anexos A

## Árboles de decisión

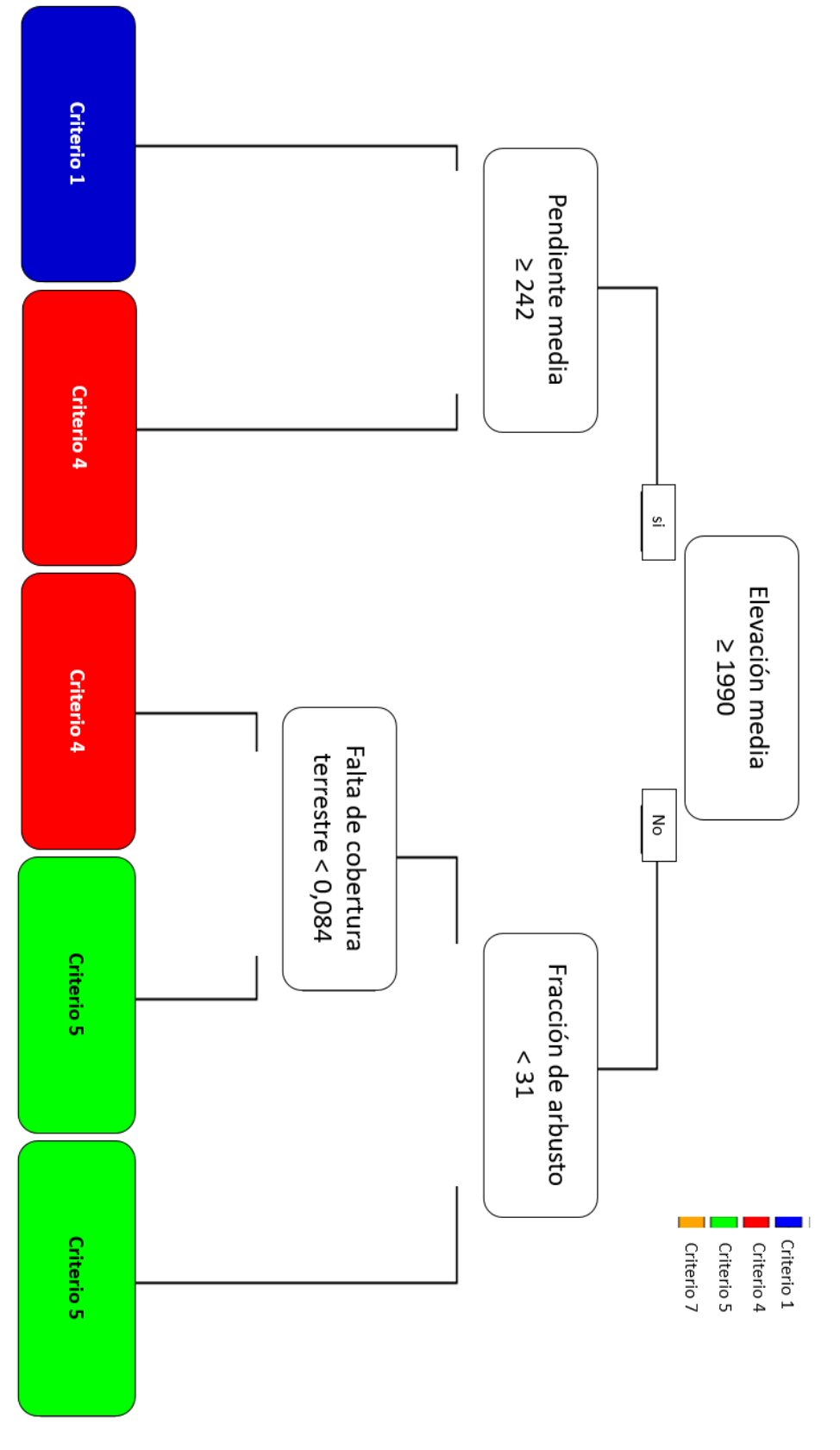


Figura A.1: Árbol de decisión que entrega el mejor criterio para rellenar caudales medios mensuales en enero, utilizando como entrada un set de atributos físicos y climáticos.



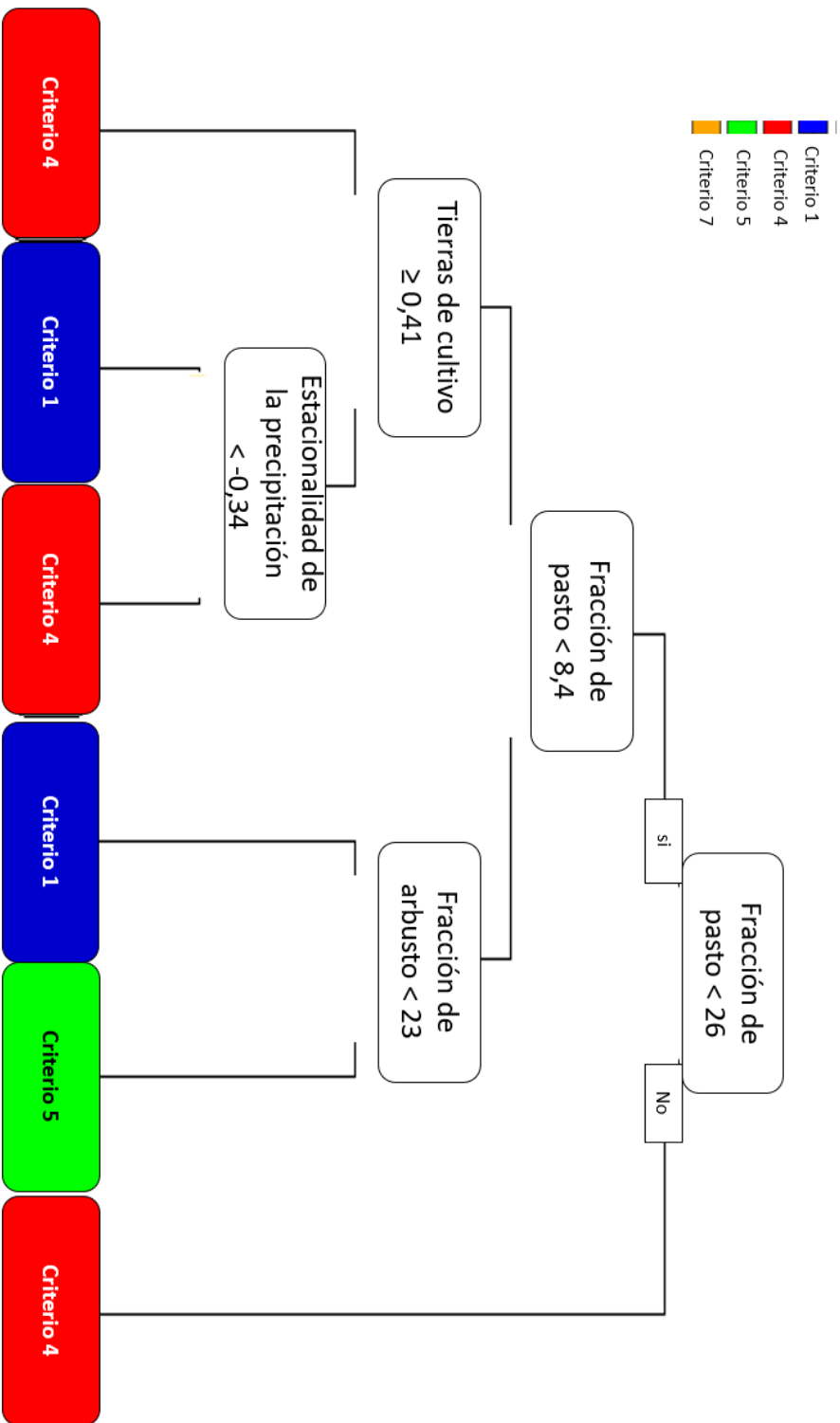


Figura A.2: Árbol de decisión que entrega el mejor criterio para rellenar caudales medios mensuales en febrero, utilizando como entrada un set de atributos físicos y climáticos.

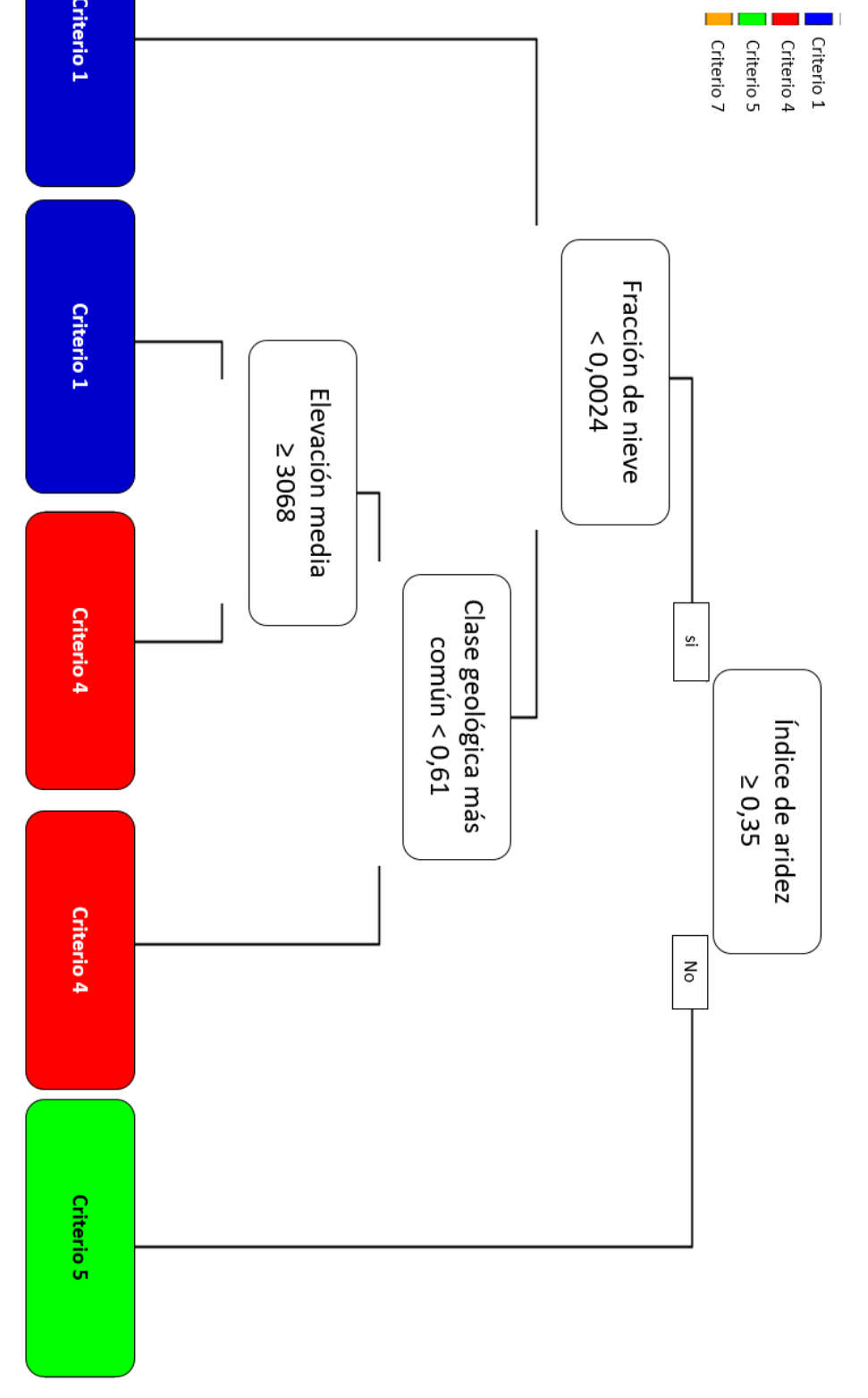


Figura A.3: Árbol de decisión que entrega el mejor criterio para rellenar caudales medios mensuales en marzo, utilizando como entrada un set de atributos físicos y climáticos.

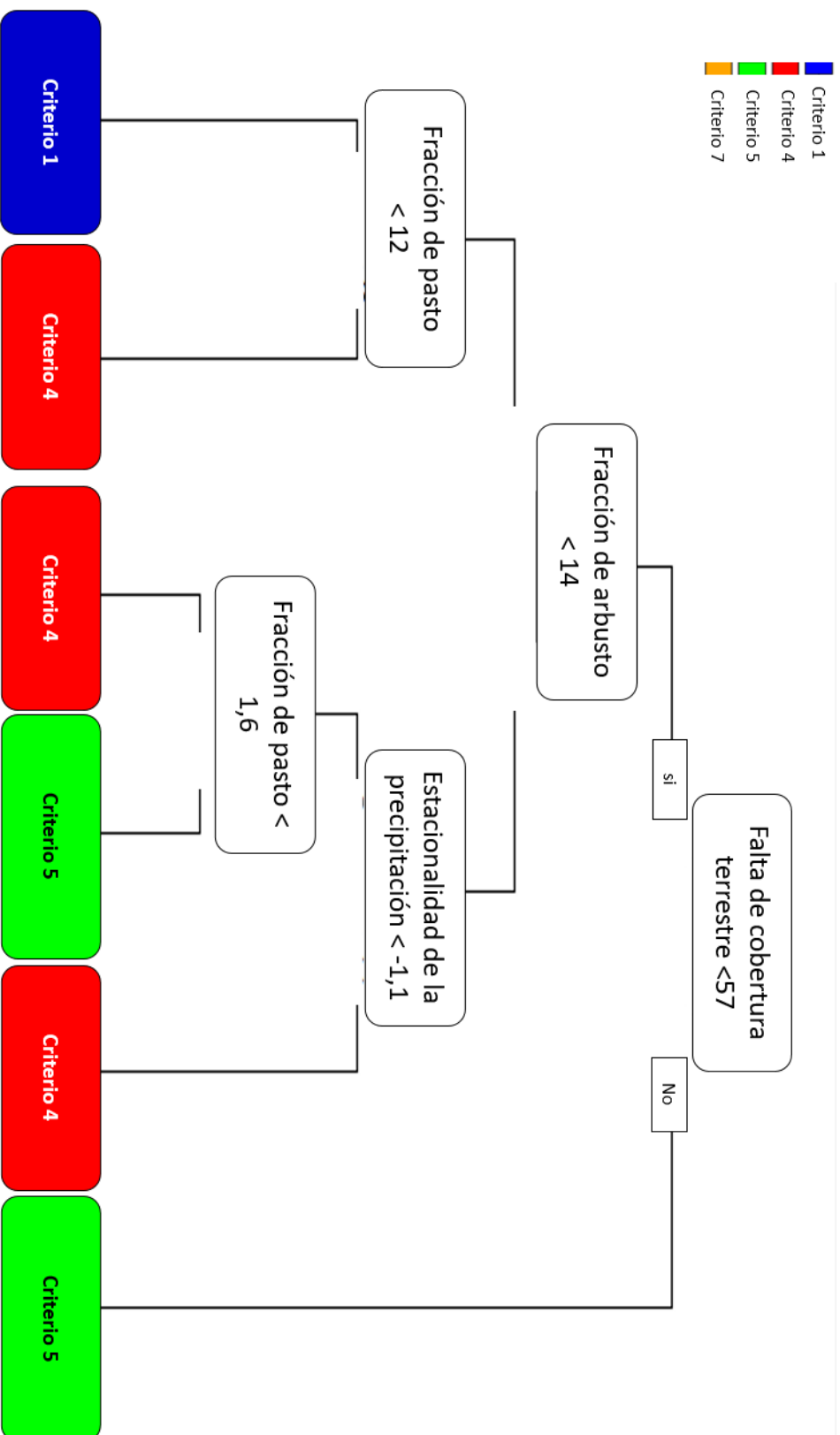


Figura A.4: Árbol de decisión que entrega el mejor criterio para rellenar caudales medios mensuales en abril, utilizando como entrada un set de atributos físicos y climáticos.

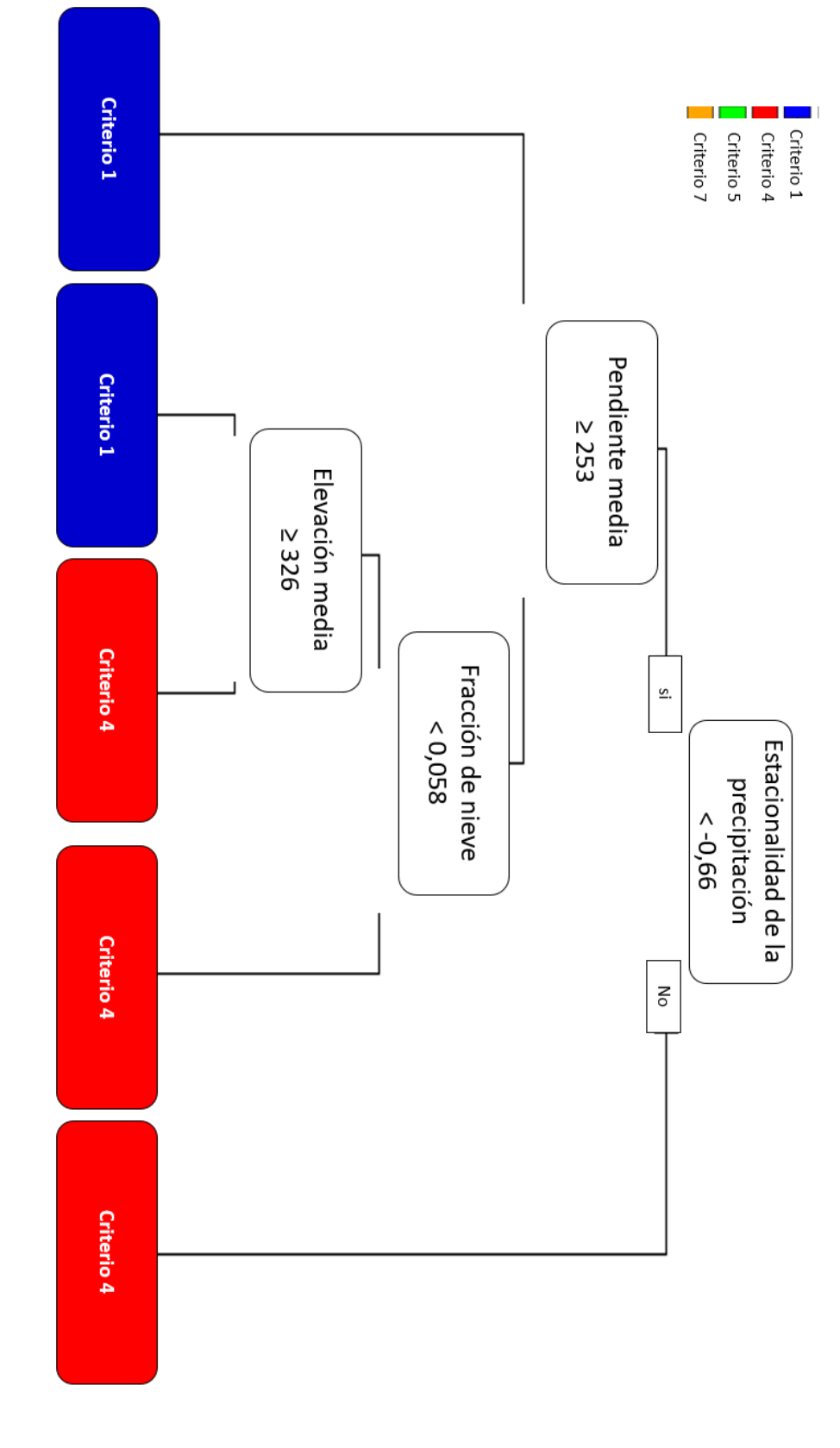


Figura A.5: Árbol de decisión que entrega el mejor criterio para rellenar caudales medios mensuales en mayo, utilizando como entrada un set de atributos físicos y climáticos.

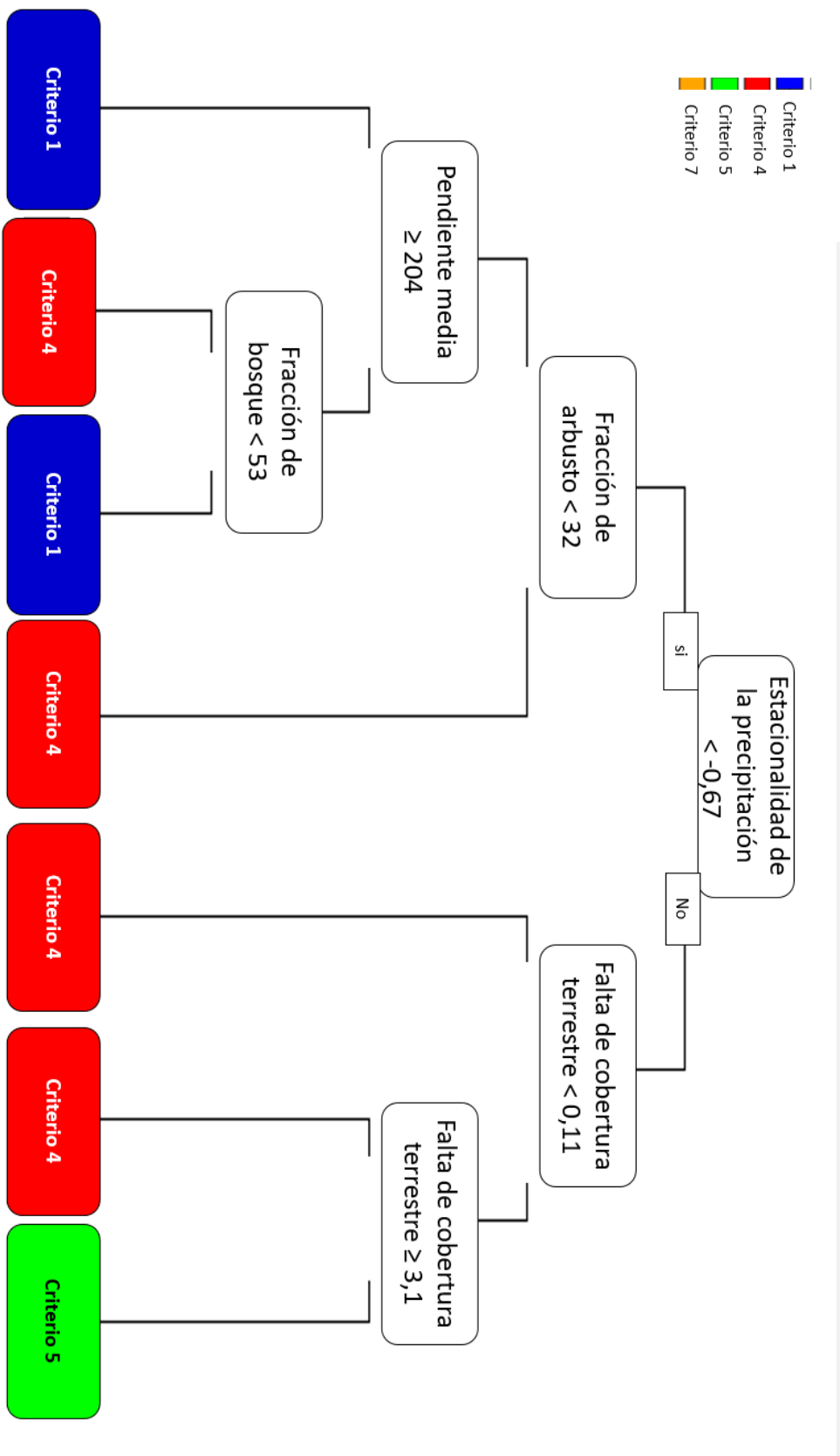


Figura A.6: Árbol de decisión que entrega el mejor criterio para rellenar caudales medios mensuales en junio, utilizando como entrada un set de atributos físicos y climáticos.

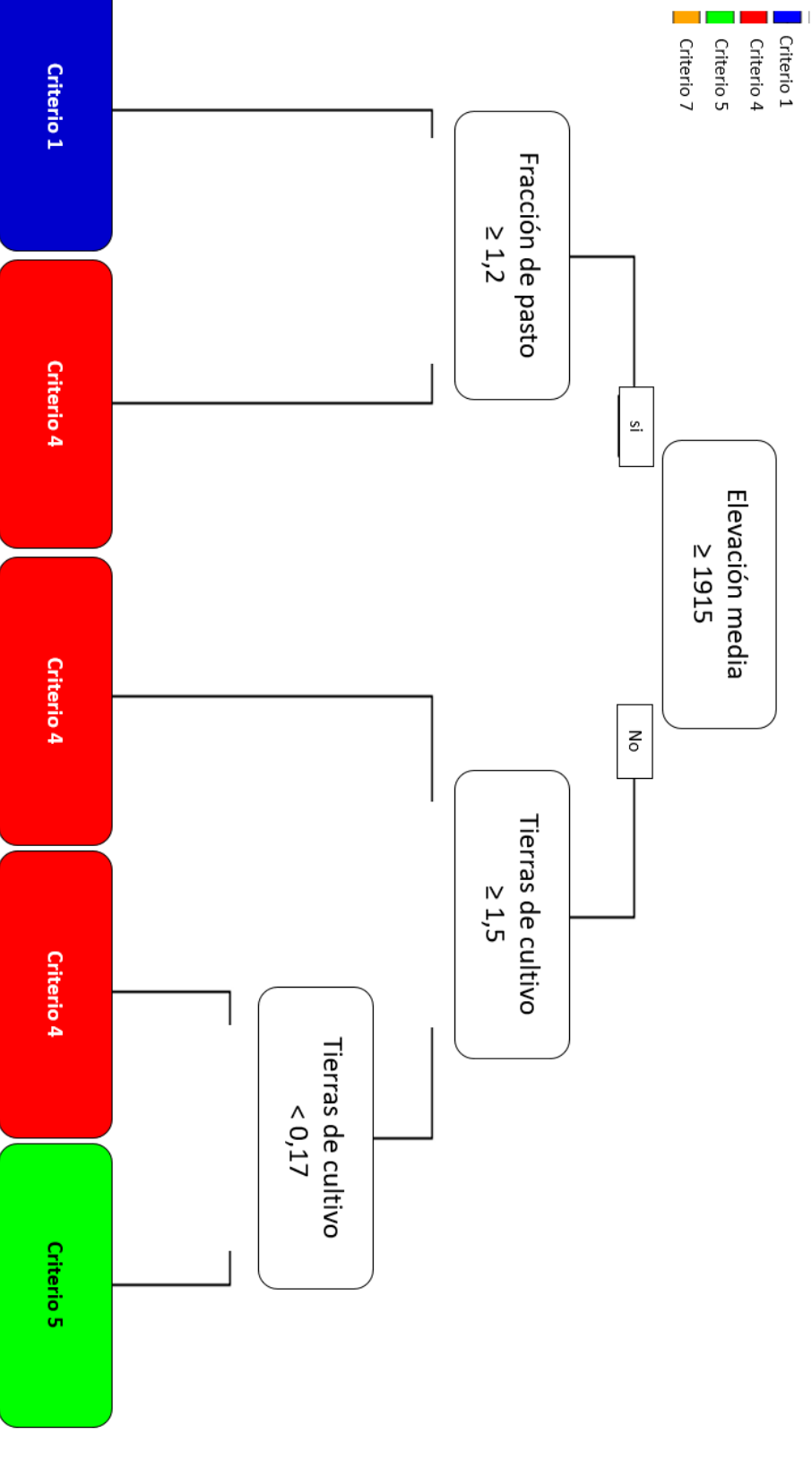


Figura A.7: Árbol de decisión que entrega el mejor criterio para rellenar caudales medios mensuales en julio, utilizando como entrada un set de atributos físicos y climáticos.

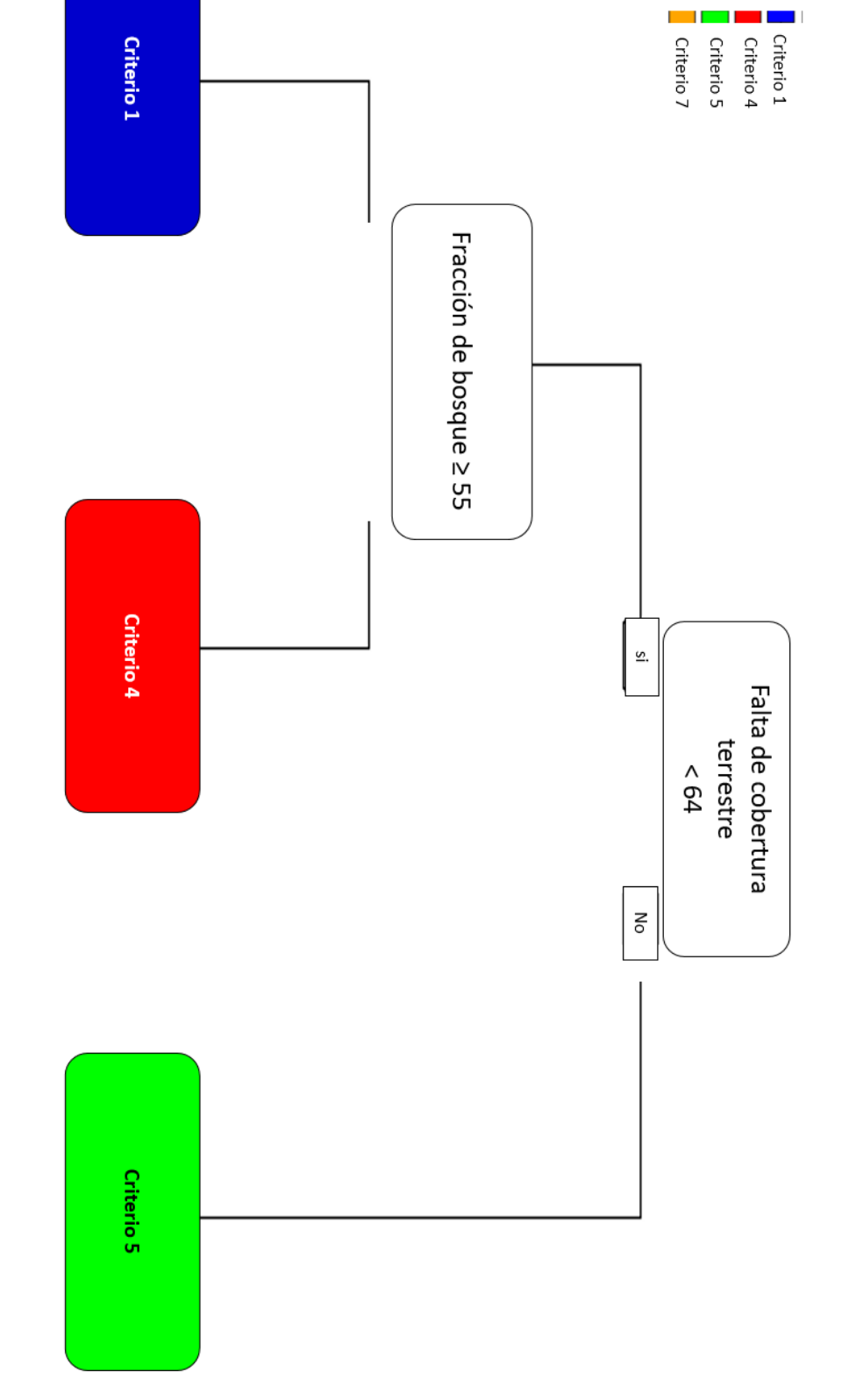


Figura A.8: Árbol de decisión que entrega el mejor criterio para rellenar caudales medios mensuales en agosto, utilizando como entrada un set de atributos físicos y climáticos.

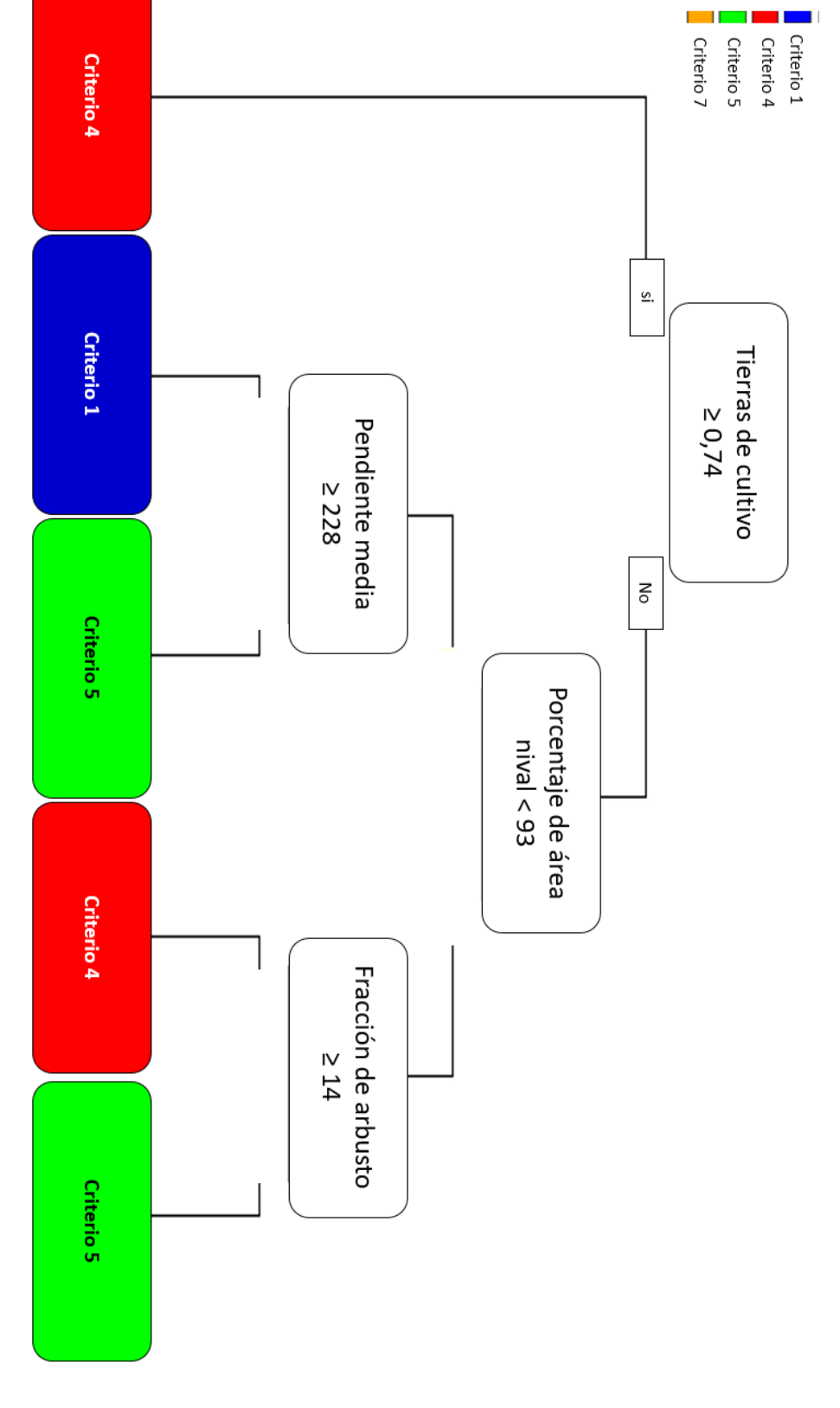


Figura A.9: Árbol de decisión que entrega el mejor criterio para rellenar caudales medios mensuales en septiembre, utilizando como entrada un set de atributos físicos y climáticos.



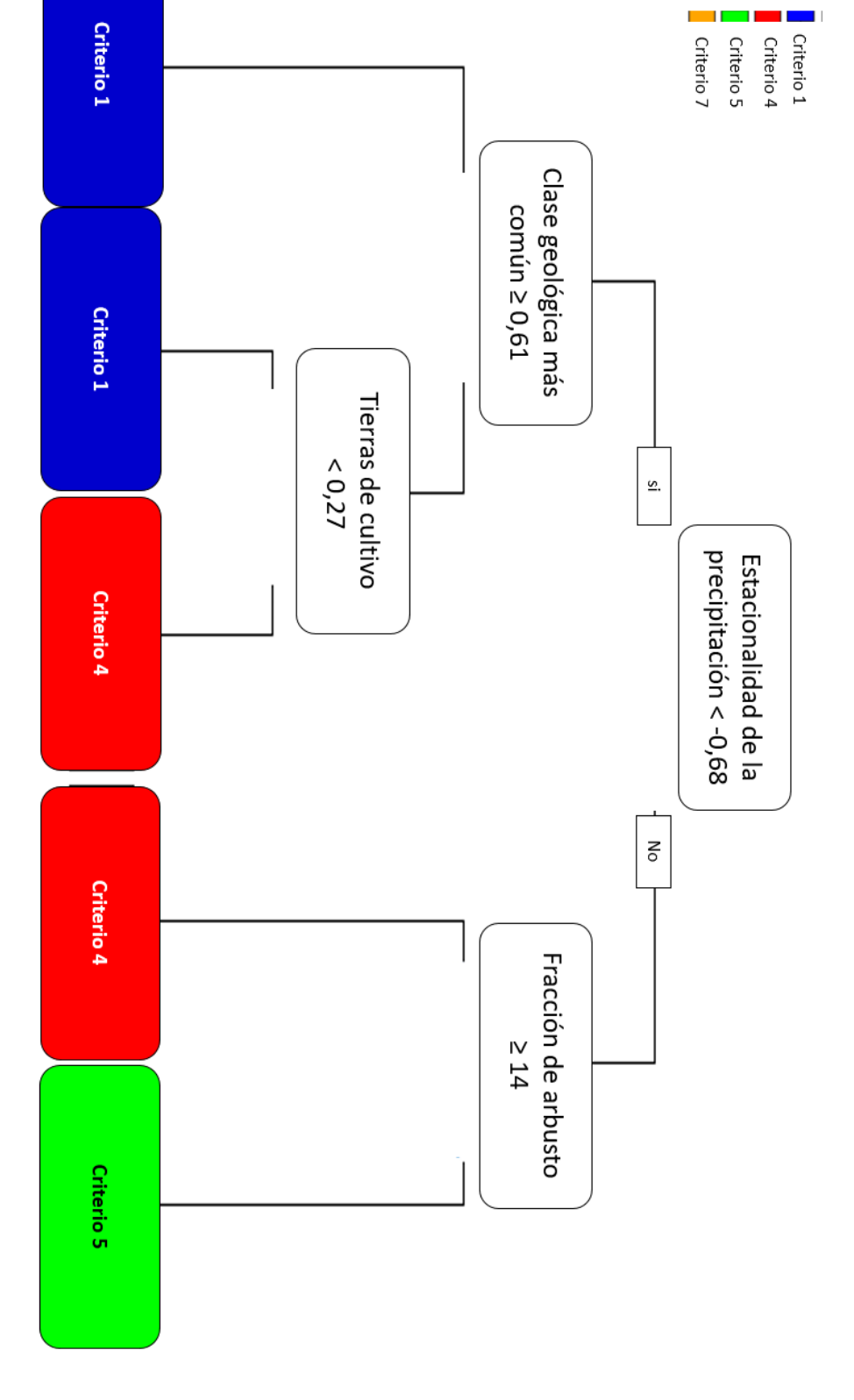


Figura A.10: Árbol de decisión que entrega el mejor criterio para rellenar caudales medios mensuales en octubre, utilizando como entrada un set de atributos físicos y climáticos.

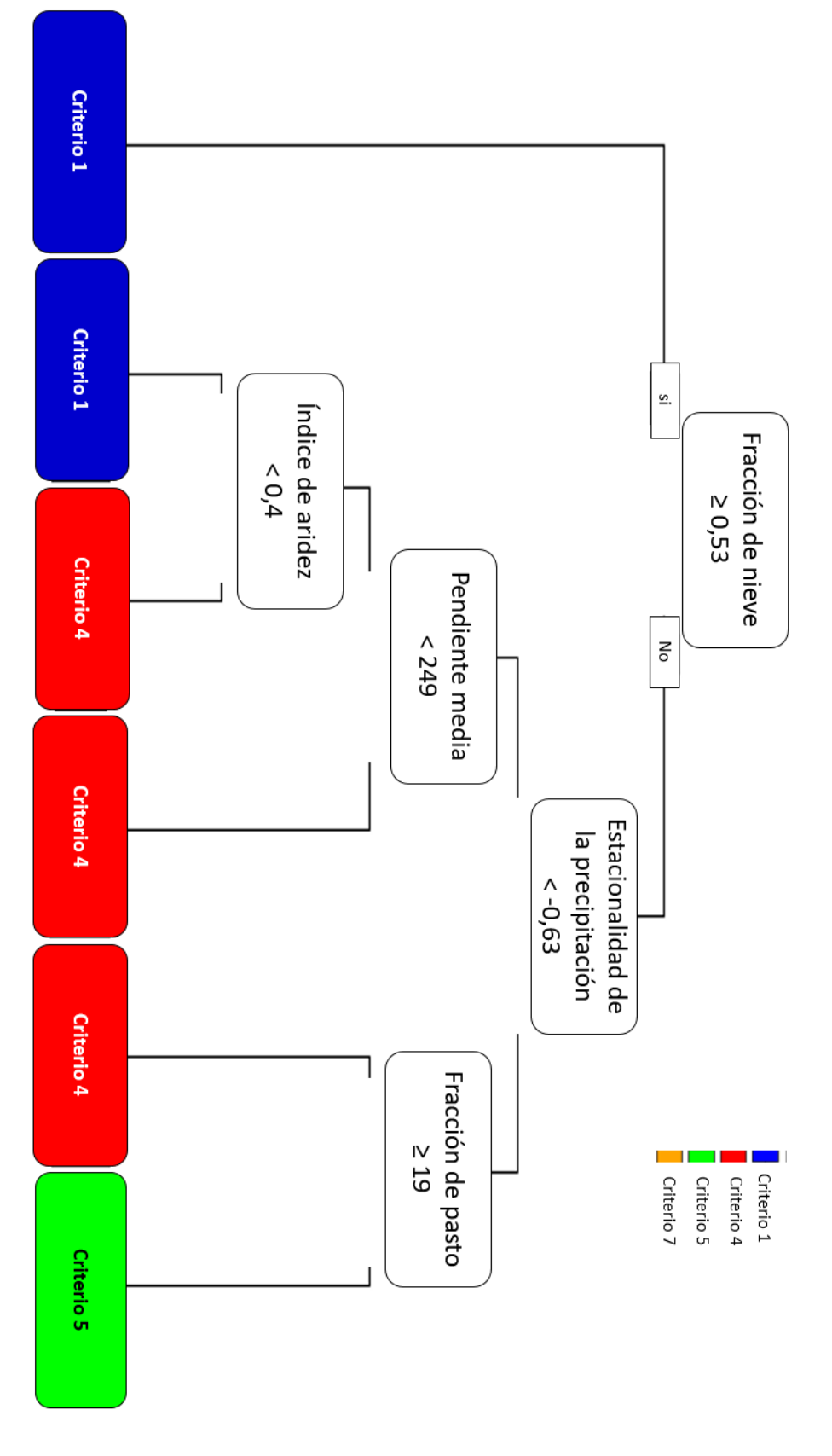


Figura A.11: Árbol de decisión que entrega el mejor criterio para rellenar caudales medios mensuales en noviembre, utilizando como entrada un set de atributos físicos y climáticos.

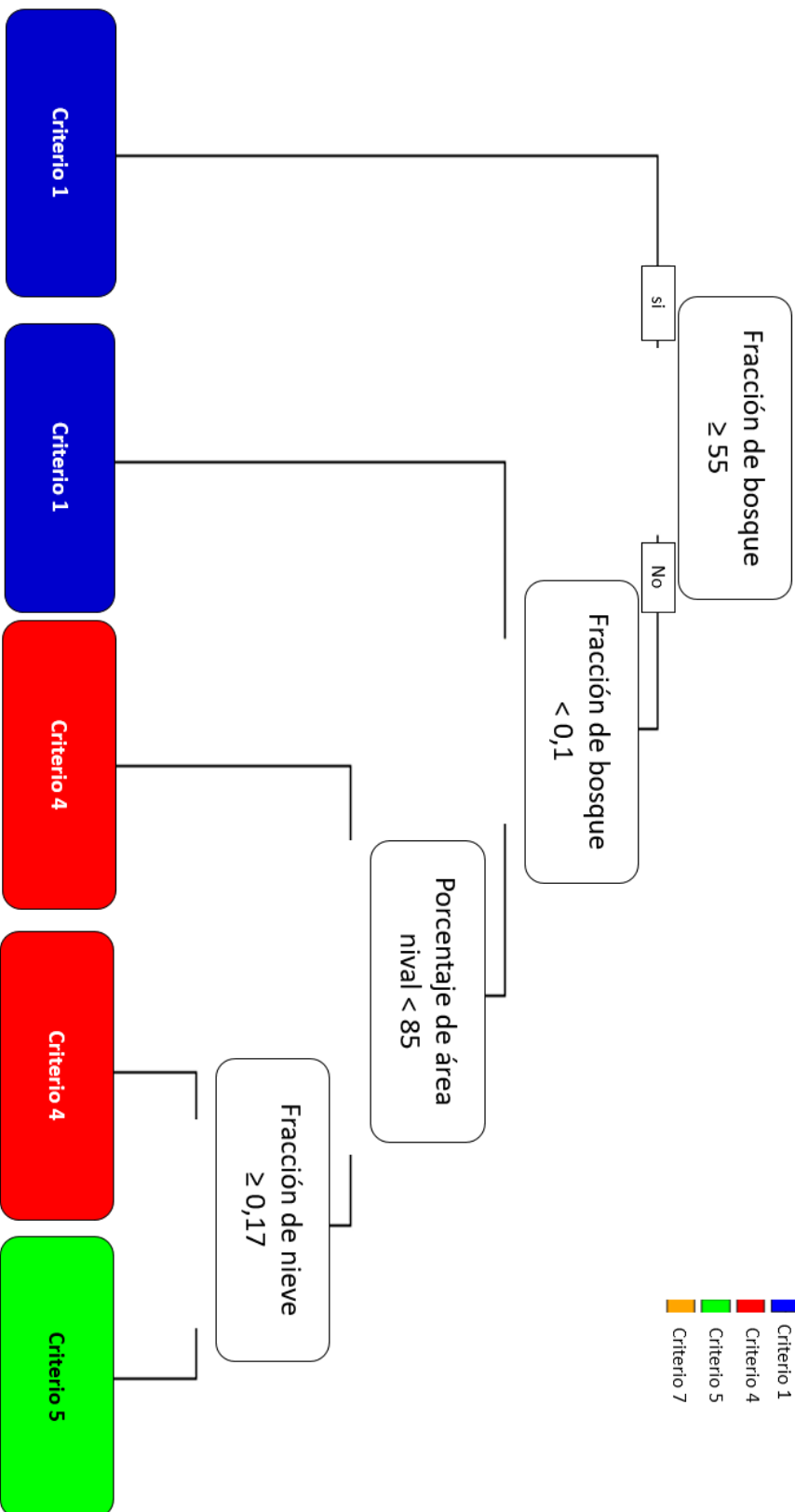


Figura A.12: Árbol de decisión que entrega el mejor criterio para rellenar caudales medios mensuales en diciembre, utilizando como entrada un set de atributos físicos y climáticos.

# Anexos B

## Ventanas de comparación para el relleno de caudales medios diarios

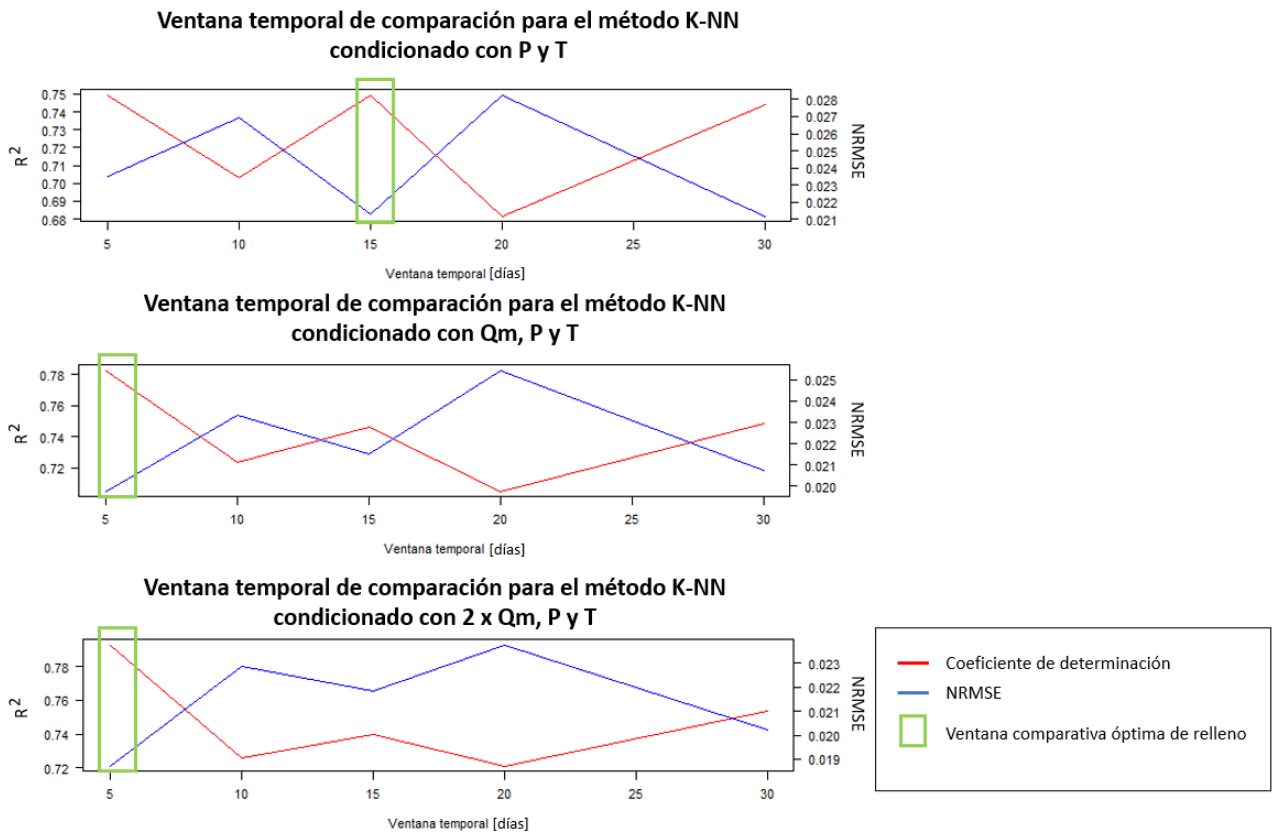


Figura B.1: Ventanas temporales comparativas para las diferentes formas de condicionar el método K-NN, en la estación Estero Yerba Loca antes junta San Francisco

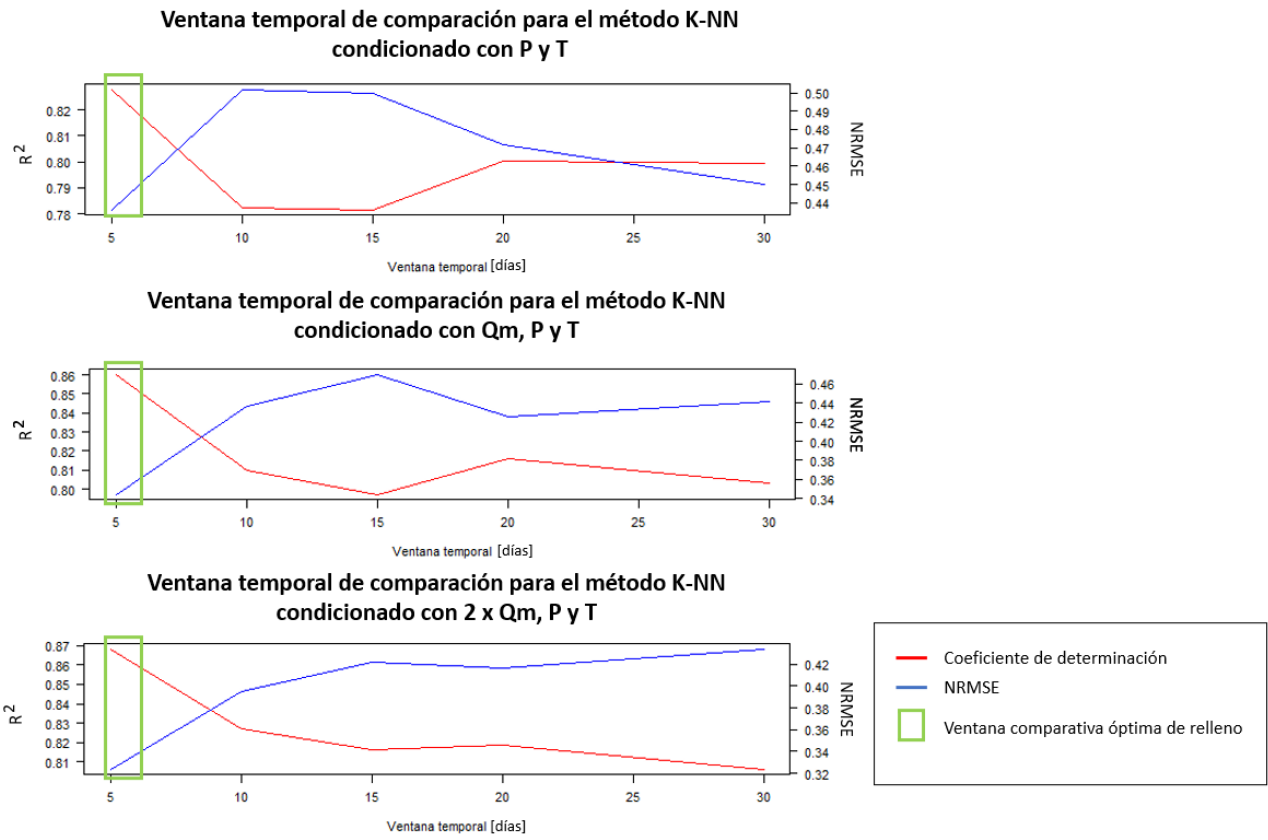


Figura B.2: Ventanas temporales comparativas para las diferentes formas de condicionar el método K-NN, en la estación Río Aconcagua en Chacabucquito

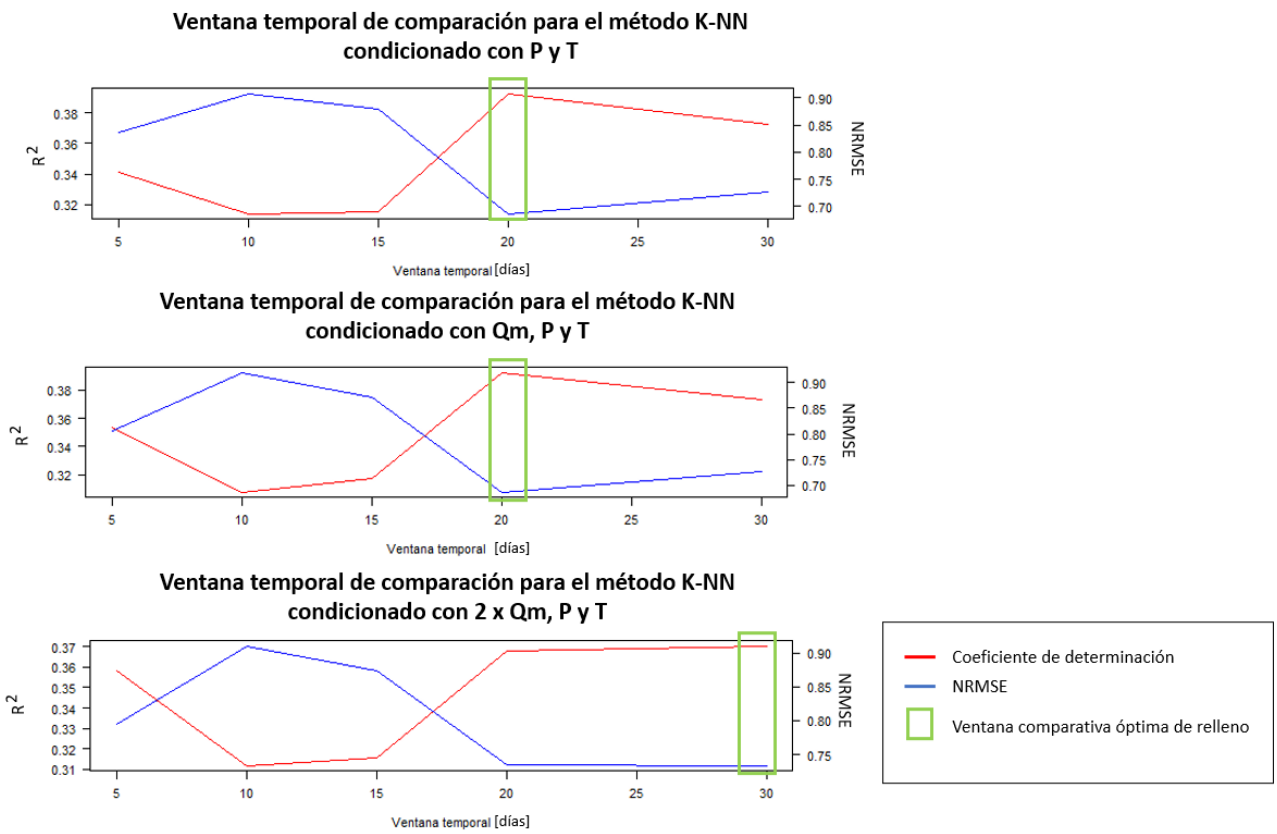


Figura B.3: Ventanas temporales comparativas para las diferentes formas de condicionar el método K-NN, en la estación Río Andalién camino a Penco

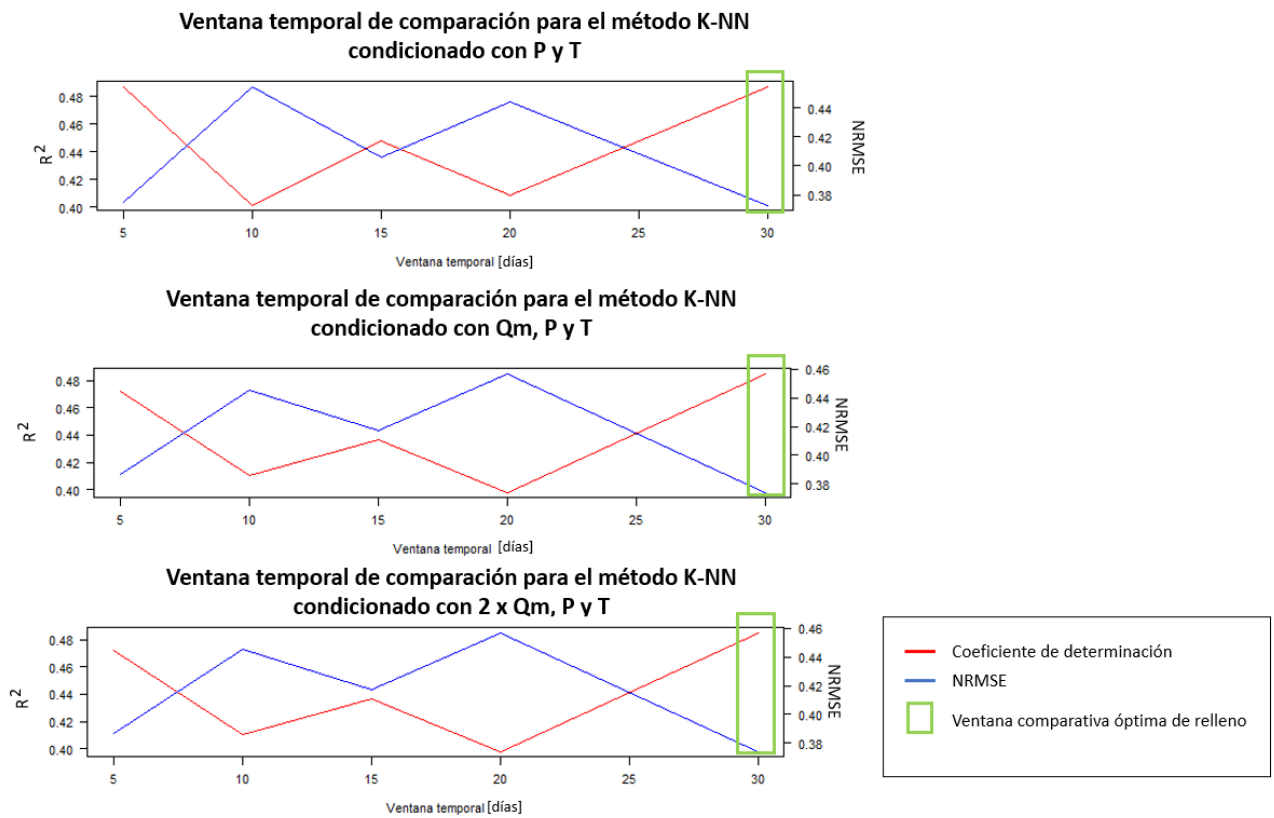


Figura B.4: Ventanas temporales comparativas para las diferentes formas de condicionar el método K-NN, en la estación Río Lirriquén en Cerro el Padre

# Anexos C

## Resultados numéricos

A continuación, se expone un enlace para descargar los resultados numéricos obtenidos en el presente trabajo. En particular, el archivo tiene los caudales medios mensuales observados y estimados desde enero de 1979 a diciembre de 2018, para las 158 estaciones fluviométricas con régimen hidrológico cercano al natural.

<https://drive.google.com/drive/folders/1xGpB5TDE-1mM3JUp2RHg1icIEXkl7NkN?usp=sharing>