



Universidad de Buenos Aires

Facultad de Ciencias Exactas y Naturales

Maestría en Explotación de Datos y Descubrimiento del
Conocimiento

***Incidencia de condiciones hidro-
meteorológicas sobre calidad del agua del
Embalse San Roque, mediante aplicación de
Minería de Datos con series temporales***

Tesis presentada para optar al título de Magíster de la Universidad
de Buenos Aires en Explotación de Datos y Descubrimiento del
Conocimiento

Ing. Pablo Facundo Andreoni

Director: Dra. Ana Laura Ruibal Conti

Lugar de trabajo: Buenos Aires

Fecha de presentación del ejemplar: 30 de agosto de 2019

Fecha de defensa: 2 de noviembre de 2020

Incidencia de condiciones hidro-meteorológicas sobre calidad del agua del Embalse San Roque, mediante aplicación de Minería de Datos con series temporales

Resumen

La importancia que los cuerpos de agua representan para la vida humana, conlleva un interés creciente en analizar la calidad del agua y comprender la naturaleza de sus cambios. El Embalse San Roque en la provincia de Córdoba presenta una condición eutrófica que se ha sostenido en el tiempo. A lo largo de casi 20 años, se han recolectado datos de calidad del agua, hidrológicos y meteorológicos de distintas fuentes, cuyo volumen y complejidad requieren la aplicación de técnicas de minería de datos para su explotación. El objetivo de este trabajo es emplear métodos basados en series de tiempo para determinar el efecto de las condiciones hidrológicas y meteorológicas (hidro-meteorológicas) sobre la calidad del agua a largo plazo. Se presentan tres aproximaciones: regresión dinámica, espacio de estados multivariado y regresión bayesiana jerárquica, todos ellos modelos aplicados sobre series de tiempo a distintas profundidades. La influencia de las predictoras ambientales sobre el estado trófico (calidad del agua) como respuesta, se estimó considerando distintos retardos (*alcances*). Se observa una fuerte incidencia positiva de la temperatura y la precipitación, particularmente en profundidades superficiales, y una incidencia negativa del nivel del embalse. Además, el fenómeno se explica mejor con alcances distintos, mayores a mayor profundidad. Por otro lado, resulta relevante el fenómeno autorregresivo con hasta 2 meses de *retardo*. Los modelos permitieron cuantificar la incidencia significativa específica de las variables hidro-meteorológicas a largo plazo sobre la calidad del agua, desde tres perspectivas estadísticas diferentes. La aplicación de los mismos en el análisis predictivo resultaría muy valiosa para la gestión del cuerpo de agua.

Palabras clave: calidad de agua, condiciones hidro-meteorológicas, embalse San Roque, series de tiempo, regresión dinámica, espacio de estados multivariado, regresión bayesiana jerárquica, autorregresión

Influence of hydro-meteorological conditions on water quality in San Roque reservoir, through Data Mining application with time series

Abstract

The importance of water bodies for human life carries an increasing interest in analyzing water quality and the nature of its changes. The San Roque reservoir, located in Cordoba province, presents a eutrophic condition that has been sustained over time. Water quality, hydrological and meteorological data have been collected for almost 20 years from different sources. The volume and complexity of the databases require the application of data mining techniques for exploring it. The goal of this thesis is to apply time series methods to assess the long-term effect of hydrological and meteorological (hydro-meteorological) conditions on water quality. Three approaches are presented: dynamic regression model, multivariate state-space model and Bayesian hierarchical regression model, all of them applied on time series associated to different levels of depth. The influence of environmental predictors on trophic state (water quality) as response is assessed at different lags (*scopes*). Temperature and precipitation have a strong positive influence, especially at surface depths, while the reservoir water level has a negative influence. Also, the effect is better explained with different scopes, larger scope at greater depth. It is also remarkable the importance of the autoregressive component with up to 2 months of lag. The models allowed to quantify, from three different statistical perspectives, the long-term specific significative influence of each of the hydro-meteorological variables on the water quality. The application of these models in predictive analysis would be very valuable for the management of the water body.

Key words: water quality, hydro-meteorological conditions, San Roque reservoir, time series, dynamic regression, multivariate state-space, Bayesian hierarchical regression, autoregression

Agradecimientos

Quisiera agradecer en primer lugar a tres personas sin las que no podría haber logrado concretar esta tesis. A Marcia Ruiz por aquel primer mail de respuesta a mi consulta para sumarme al proyecto de calidad de agua allá por mayo de 2016, sus muchos aportes en la cuestión bioquímica y su constante apoyo. A María Inés Rodríguez, por toda la ayuda para la interpretación de las cuestiones biológicas, por brindarme su experiencia y por invitarme a participar del monitoreo mensual para comprender mejor los datos. A Ana Laura Ruibal Conti, mi directora, por todas las valiosísimas observaciones y sugerencias, por guiarme siempre en el camino científico, por su paciencia y por confiar en mí para encarar esta tarea larga y apasionante que fue la aventura de conocer un poco más del San Roque.

Asimismo, quisiera agradecer a Laura Colladón por sus valiosos aportes en materia de procesamiento de datos hidrológicos y meteorológicos; a Clarita Dasso y Leticia Vicario, por su contribución teórica en lo concerniente a hidrología; y a mi colega Alejo Keuroghlanián por brindarme su visión en muchas cuestiones referidas a minería de datos.

Índice

1	Introducción	7
1.1	Marco teórico y estado del arte	7
1.2	Conceptos de meteorología	9
1.3	Conceptos de limnología	11
2	Materiales y métodos	13
2.1	Sitio de estudio: Embalse San Roque	13
2.2	Monitoreo de variables de calidad del agua	14
2.2.1	Sitios de muestreo, instrumental y técnicas de análisis	14
2.2.2	Datos de calidad de agua del lago	17
2.3	Monitoreo de variables hidrológicas y meteorológicas	19
2.3.1	Estaciones meteorológicas y sensores	19
2.3.2	Datos meteorológicos e hidrológicos relevados por sensores	22
2.4	Preprocesamiento de datos	25
2.4.1	Datos de calidad del agua	25
2.4.2	Datos meteorológicos e hidrológicos	26
2.4.2.1	Caso general	26
2.4.2.1.1	Tratamiento de inconsistencias	27
2.4.2.1.2	Agregación	30
2.4.2.2	Precipitación	33
2.4.2.2.1	Tratamiento de inconsistencias	34
2.4.2.2.2	Agregación	34
2.4.2.2.3	Rectificación	34
2.4.2.2.4	Imputación	36
2.4.2.3	Análisis de valores extremos y anomalías de las series de tiempo diarias	37
2.4.3	Integración de datos	38
2.4.3.1	Integración temporal	41
2.4.3.2	Integración espacial	43
2.5	Series de tiempo	46
2.5.1	Modelo de serie de tiempo	46
2.5.2	Conceptos principales	46
2.6	Análisis exploratorio	48
2.6.1	Descomposición de la serie de tiempo	48
2.6.2	Autocorrelación	48
2.6.3	Autocorrelación parcial	49

2.6.4	Interpolación	49
2.6.5	Índice de estado tráfico	50
2.7	Modelado	50
2.7.1	Regresión dinámica	50
2.7.1.1	Especificación del modelo	51
2.7.1.2	Estimación de los parámetros	53
2.7.1.3	Evaluación de los modelos de regresión dinámica	53
2.7.1.4	Metodología de modelado	54
2.7.1.5	Significatividad de los parámetros	54
2.7.2	Modelo de espacio de estados multivariado	55
2.7.2.1	Especificación del modelo	55
2.7.2.2	Estimación de los parámetros	56
2.7.2.3	Evaluación de los modelos MARSS	57
2.7.2.4	Metodología de modelado	57
2.7.2.5	Significatividad de los parámetros	57
2.7.3	Modelo de regresión bayesiano jerárquico	58
2.7.3.1	Especificación del modelo	58
2.7.3.2	Estimación de los parámetros	61
2.7.3.3	Evaluación de los modelos de regresión bayesianos jerárquicos	61
2.7.3.4	Metodología de modelado	62
2.7.3.5	Significatividad de los parámetros	63
2.7.4	Predicción	63
2.7.4.1	Regresión dinámica	63
2.7.4.2	Modelo de espacio de estados multivariado	64
2.7.4.3	Modelo de regresión bayesiano jerárquico	65
3	Resultados	66
3.1	Análisis exploratorio de series temporales	66
3.1.1	Análisis de las series de tiempo del lago	67
3.1.2	Análisis de las series de tiempo hidrológicas y meteorológicas	82
3.2	Modelos	102
3.2.1	Resultados del modelo de regresión dinámica	104
3.2.2	Resultados del modelo de espacio de estados multivariado	113
3.2.3	Resultados del modelo de regresión bayesiano jerárquico	126
4	Discusión y conclusiones	139
I	Anexo: Monitoreo de variables de calidad del agua extendido	149
II	Anexo: Preprocesamiento de datos extendido	153

III	Anexo: Análisis exploratorio extendido	158
IV	Anexo: Software	171
5	Referencias	172

1 Introducción

El objetivo de este capítulo es, por un lado, presentar la problemática de estudio y los trabajos realizados hasta el momento; y por otro lado exponer los conceptos principales que se aplicarán a lo largo del presente trabajo.

1.1 Marco teórico y estado del arte

Los lagos son cuerpos de agua muy importantes por sus variadas funciones y múltiples usos tales como albergar vida acuática, provisión de agua para consumo humano y regulación de flujos hídricos, entre otros (Jorgensen & Vollenweider, 1989). La calidad del agua de los lagos depende de múltiples factores tanto intrínsecos (procesos físico-químicos y biológicos dentro del ecosistema acuático), como extrínsecos (condiciones meteorológicas, climáticas, tipo de suelo, usos del cuerpo de agua, etc.). Los procesos relacionados a la calidad del agua pueden variar en períodos cortos (horarios, diarios o estacionales) y en períodos o ciclos más largos (quinquenios, décadas, etc.).

Los procesos que ocurren en períodos más largos suelen estar asociados a la variabilidad climática de la región. La evaluación de los mismos requiere de mediciones por largos períodos de tiempo generando extensas bases de datos que comprenden múltiples variables de calidad del agua, hidrológicas y meteorológicas. Dada la magnitud de estos repositorios, la manipulación, integración y análisis de los datos resulta complejo. Surge entonces la necesidad de utilizar nuevas herramientas que permitan explotar la información en todo su potencial y en tiempos razonables.

El campo de *Data Mining* o Minería de Datos (MD), entendida como disciplina, provee un marco metodológico (proceso de descubrimiento del conocimiento) que permite por un lado la estructuración y adecuación de grandes bases de datos (limpieza, integración, selección y transformación de los datos), y por otro la determinación de patrones interesantes en los datos que hasta el momento se encontraban ocultos (MD en sentido estricto). Así, por medio de la aplicación de métodos inteligentes a los datos preprocesados, se puede extraer conocimiento que finalmente sirve para apoyar la toma de decisiones (Han & Kamber, 2006).

La aplicación de MD en el campo ambiental y en particular en series temporales de calidad del agua en ecosistemas lacustres ha sido limitada. Shoba (2014) hace una extensa revisión de métodos de MD que se han aplicado en el ámbito de calidad del agua, en particular orientado al establecimiento de estándares, donde lo más frecuente es el uso de redes neuronales de variada arquitectura y lógica difusa, predominantemente para resolver problemas de segmentación y clasificación. Por otro lado, Liao et al. (2012) hacen una revisión de las técnicas de MD empleadas a lo largo de una década y sus ámbitos de aplicación, donde el autor destaca el uso de lógica difusa en el campo de la hidrología. Finalmente, Yang et al. (2018) desarrollan un modelo híbrido basado en máquina de soporte vector, análisis de componentes principales y redes neuronales artificiales tipo *back propagation*¹, para predecir calidad del agua. En la

¹ Se refiere al método de entrenamiento más común de redes neuronales, basado en la actualización de los pesos de la red neuronal desde la salida en dirección a la entrada, minimizando el error (Han & Kamber, 2006).

misma línea otros autores introdujeron el uso de modelos de redes neuronales para predecir floraciones algales en distintos cuerpos de agua (Recknagel et al., 1997), (Abunaka et al., 1997); modelado evolutivo, para analizar la dinámica de poblaciones de algas en lagos hipertróficos (Bobbin & Recknagel, 2001); y lógica difusa para predecir biomasa algal en un lago eutrófico (Chen & Mynett, 2003).

El Embalse San Roque (ESR), ubicado en la Provincia de Córdoba es un lago artificial construido en 1888 con el fin de proveer agua para consumo humano, riego y energía hidroeléctrica. Actualmente es uno de los centros turísticos más importantes del país con un extenso uso recreativo, recibiendo la localidad de Villa Carlos Paz aproximadamente unos 370.000 visitantes por quincena en verano según estimación oficial (Secretaría de Turismo de Villa Carlos Paz, 2015). La calidad de su agua se ha deteriorado con el tiempo y actualmente presenta un estado eutrófico avanzado (Instituto Nacional del Agua, 2019) generando serios problemas ambientales (mortalidad de peces y malos olores, entre otros), económicos (inversión en infraestructura sofisticada para potabilización del agua, remediación del lago y disminución del turismo, entre otros) y sociales (efectos sobre la salud pública y cierre de escuelas, entre otros).

Se sabe que el enriquecimiento por nutrientes inorgánicos (en particular nitrógeno y fósforo) favorece la eutrofización de los lagos (Sigg, 2005); aunque al mismo tiempo, se sabe que las condiciones ambientales generan tanto un efecto directo sobre los procesos biológicos, tal es el caso de la fotosíntesis, sensible a la radiación solar y la temperatura (Amé et al., 2017), como indirecto, por ejemplo la dinámica y magnitud de la liberación de fósforo de los sedimentos, afectada también por la temperatura y los vientos (Rodríguez et al., 2000). Esos procesos, sumado a los fenómenos estacionales asociados, la dinámica hidrológica del embalse y el cambio climático a largo plazo, dan indicio de que estas condiciones ambientales inducen un efecto sobre el estado trófico.

En el presente trabajo se estudia la aplicación de tres modelos de MD, con enfoques variados, tanto del paradigma estadístico frecuentista como *bayesiano*, con el propósito de responder a la pregunta: *¿cuál es la incidencia de las condiciones hidro-meteorológicas sobre la calidad del agua en el Embalse San Roque?*

Numerosos estudios en el ESR han aplicado modelos para comprender los fenómenos que gobiernan la calidad de sus aguas. Un primer modelo fue el aplicado por Gavilán (1981), quien utilizó el modelo de eutrofización empírico de Vollenweider. Posteriormente, se aplicaron modelos de balance de masa para estudiar el efecto de concentraciones de fósforo (Bechtel et al., 1992). Más recientemente, se han aplicado modelos más complejos tales como:

1. Modelado FLUX, PROFILE y BATHTUB orientado al monitoreo de calidad del agua y a la predicción de los aportes externos de nutrientes (Rodríguez et al., 2000).
2. Modelado numérico semiparamétrico para la obtención de mapas de clorofila-a y temperatura del embalse a partir de datos satelitales y mediciones de campo (Amé et al., 2017).
3. Finalmente, se aplicaron varios métodos de MD sobre datos de la cuenca del río Suquía, efluente del ESR, a los efectos de analizar la calidad de aguas de las subcuencas: análisis de *cluster* sobre parámetros químicos y microbiológicos, análisis de factor/análisis de componentes principales, análisis discriminante y análisis de *Procrustes* generalizado² (Baroni & Wunderlin, 2018).

² Método estadístico para analizar datos de perfilamiento de elección libre (Gower, 1975).

Este trabajo se distingue por la aplicación de modelado basado en series de tiempo, que hasta el momento no se había aplicado de manera extensiva en el ESR. A nivel internacional, si bien en otros trabajos se han abordado las series de tiempo para predecir la calidad del agua, como lo hicieron Zhang et al. (2016) y Tizro et al. (2014), no se han tenido en cuenta variables meteorológicas e hidrológicas ni se han empleado enfoques múltiples basados en esta técnica. Incluso el modelado de series de tiempo no constituye una técnica considerada entre las más comúnmente utilizadas en general (Liao et al., 2012).

Otro aporte de este trabajo, lo constituye el uso de series de tiempo consideradas *extensas* en el contexto de aplicación, 19 años en este caso. Trabajos previos de calidad de agua en el ESR consideraron series cortas (Rodríguez et al., 2000), (Amé et al., 2017). A nivel internacional, para analizar las aguas del lago Dianchi (Yang et al., 2018), se utilizan series de 12 años. Estudios de más largo alcance se realizaron sobre el lago Washington (Francis et al., 2014) pero sin abordar específicamente la problemática de calidad del agua en términos integrales ni el efecto de la profundidad. En relación a este último punto, en el presente trabajo se aborda el análisis y modelado de la calidad a distintos niveles de profundidad en la columna de agua.

De modo similar, en otros trabajos sobre calidad del agua se han empleado variables meteorológicas, como temperatura (Yang et al., 2018), (Weiyang et al., 2013), y precipitación (Barbosa Chaves et al., 2014) de manera individual. En tanto, en este trabajo se realiza un abordaje completo e integral del estado del tiempo considerando varios parámetros más, además del comportamiento dinámico del agua. Asimismo, un aspecto novedoso de este estudio, es el análisis de la influencia retardada de los procesos climatológicos e hidrológicos, es decir el efecto de las variables ambientales sobre la calidad del agua considerando no sólo el día inmediato anterior, sino la acumulación a lo largo de series de días consecutivos.

Este trabajo se organiza en 3 capítulos más: *Materiales y métodos*, *Resultados*, y *Discusión y conclusiones*, y completando el presente capítulo introductorio se proporciona al lector conceptos básicos de limnología y meteorología. En el capítulo de *Materiales y métodos*, se detallan las características del objeto de estudio: el ESR, las fuentes de datos tanto de calidad de agua como hidrológicas y meteorológicas, y los procedimientos empleados para la depuración, transformación, selección e integración de los datos. Asimismo, se explicitan las estrategias seguidas en cuanto al modelado de los datos como series de tiempo y el análisis exploratorio de dichas series de tiempo, y por último se menciona el procedimiento de modelado predictivo incluyendo la presentación de los algoritmos, la estrategia de modelado, la evaluación individual y comparativa de los modelos, y finalmente el método de predicción que se siguió en cada caso. En el capítulo de *Resultados* se exponen los resultados obtenidos tanto a partir del análisis exploratorio como del ajuste de los modelos de series de tiempo. Finalmente, en el capítulo de *Discusión y conclusiones* se analizan los resultados obtenidos y se presentan las conclusiones del estudio.

1.2 Conceptos de meteorología

La meteorología es el estudio de la atmósfera y sus fenómenos. El clima, en tanto, es la condición de la atmósfera en un momento y lugar particular (Ahrens & Henson, 2017).

Algunas de las variables más importantes que caracterizan el clima se detallan a continuación, siguiendo los conceptos de Ahrens y Henson (2017), y resultando en **negrita** aquellas que tienen un rol principal en este estudio:

- **Temperatura:** “es una medida de la velocidad promedio (movimiento promedio) de los átomos y moléculas, donde altas temperaturas se corresponden con velocidades promedio más altas y viceversa”. En particular se analizará la temperatura del aire, que remite al grado de calor o frío de los gases que lo componen.
- **Radiación solar:** “es la energía transferida desde el sol a un objeto. Se puede pensar la radiación como un flujo continuo de partículas o fotones, que son paquetes discretos de energía”.
- **Humedad:** “es una medida de la cantidad de vapor de agua en el aire”. En particular, la humedad relativa, variable de estudio en este trabajo es “el cociente entre la cantidad real de vapor de agua en el aire y la máxima cantidad de vapor de agua requerida para la saturación a una temperatura (y presión) particular”.
- **Evaporación:** es el cambio de estado de líquido a gaseoso del agua.
- **Precipitación:** “es cualquier cantidad de agua (líquida o sólida) que cae desde una nube y alcanza el suelo”.
- **Presión atmosférica:** “es la presión ejercida por una masa de aire sobre una región”. Es igual a la temperatura por la densidad del aire por una constante. A mayor altitud la presión atmosférica disminuye.
- **Viento:** es el movimiento en masa del aire por diferencia de presión atmosférica, “la compensación de las diferencias de presión atmosférica entre dos puntos” (Roth, 2003). En particular, el viento está definido por una velocidad, o distancia recorrida por unidad de tiempo; y también una dirección, determinada por el lugar desde y hacia donde sopla. Ambas variables se analizarán en este trabajo.

Estas variables meteorológicas se encuentran íntimamente relacionadas. A continuación, en la Fig. 1, se muestra un diagrama simplificado donde se establecen algunas de las principales relaciones entre las variables mencionadas anteriormente.

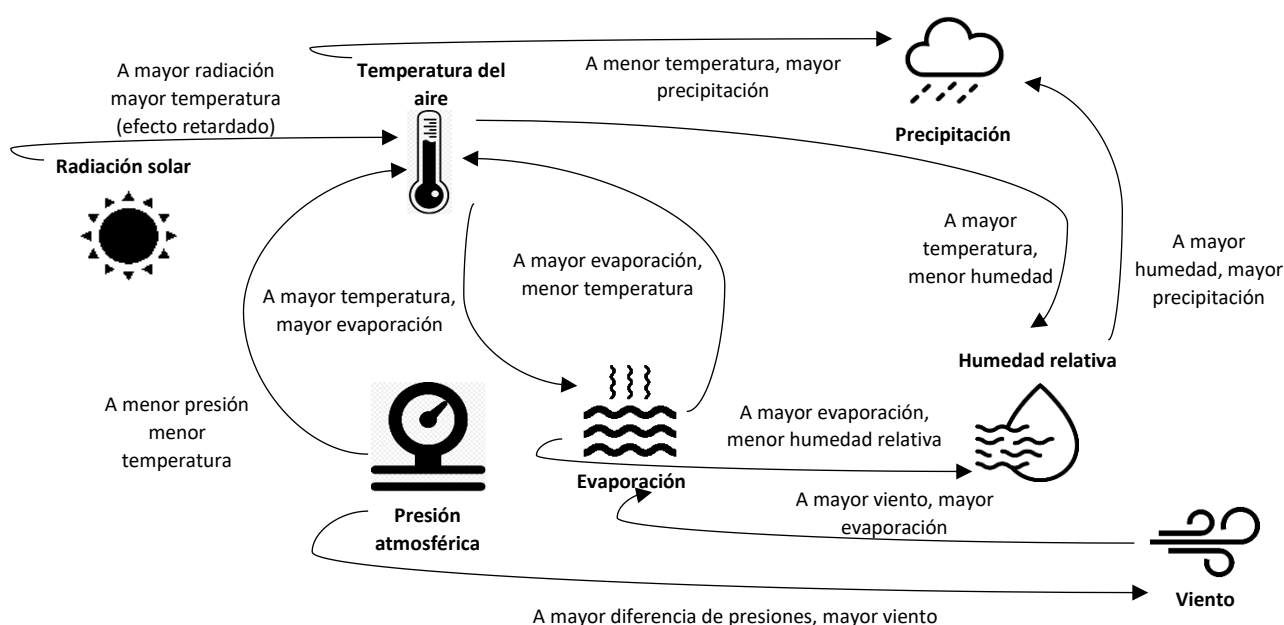


Fig. 1 Diagrama de algunas de las principales relaciones entre las variables meteorológicas. Fuente: elaboración propia

1.3 Conceptos de limnología

El estudio de las particularidades de la calidad del agua del ESR y sus ríos tributarios, requiere un abordaje profundo desde la limnología. Esta ciencia es una rama de la ecología que se dedica al estudio de los ecosistemas acuáticos continentales, tal es el caso del agua dulce (en el embalse y en los ríos). En términos más formales, es el estudio de las relaciones funcionales y de productividad de las comunidades de agua dulce y la manera como las afecta el ambiente químico, físico y biológico (Roldán Pérez & Ramírez Restrepo, 2008).

El objetivo de esta sección es dar un marco a la posterior interpretación de los resultados en términos de las variables más comunes que definen el estado de un cuerpo de agua, desde el punto de vista de la biología y las principales interacciones que se producen entre ellas. Algunas de estas variables se detallan a continuación, siguiendo distintos autores, y resaltando en negrita aquellas que tienen un rol principal en este estudio:

- **pH**: “es una abreviatura de potencial de hidrogeniones (H^+), e indica la concentración de estos en el agua”. Esta variable indica la intensidad de la condición ácida o básica del agua (Roldán Pérez & Ramírez Restrepo, 2008).
- **Oxígeno disuelto**: se refiere a la cantidad de O_2 disuelto en el agua.
- **Porcentaje de saturación de oxígeno disuelto**: “es el porcentaje máximo de oxígeno que puede disolverse en el agua a una presión y temperaturas determinadas” (Roldán Pérez & Ramírez Restrepo, 2008).
- **Conductividad eléctrica**: se refiere a la capacidad de una solución para conducir una corriente eléctrica en función de la concentración de iones presentes en ella (Roldán Pérez & Ramírez Restrepo, 2008).
- **Sólidos disueltos totales**: se refiere a la concentración total de sustancias o minerales disueltos (Roldán Pérez & Ramírez Restrepo, 2008).
- **Temperatura**: es la temperatura (tal y como se definió en la sección 1.2) del agua.
- **Potencial óxido-reducción**: químicamente hablando, se refiere a las reacciones redox, “aquellas en que cambia el estado o grado de oxidación de las especies reaccionantes, se produce un intercambio de electrones entre los reactivos” (Burriel Martí et al., 1985).
- **Turbiedad**: “es el grado en que el agua interfiere con la transmisión de luz a través de ella”, por este motivo la luz se reemite y no se transmite a través de la suspensión. La turbiedad originada en el agua por el aporte de materiales externos (por ejemplo, de creciente en el embalse) se denomina alóctona y la producida dentro del mismo cuerpo de agua, autóctona. Incide directamente en la producción primaria y en el flujo de energía en el ecosistema, asociado con la luz (Roldán Pérez & Ramírez Restrepo, 2008).
- **Color del agua**: “está constituido por la luz no absorbida” (Roldán Pérez & Ramírez Restrepo, 2008).
- **Condición de estratificación**: se refiere a la formación de capas en el agua con propiedades distintas según algún criterio. La estratificación térmica se refiere a la formación de dos capas bien definidas: en regiones de clima templado, se tiene una caliente superficial y otra fría profunda (en regiones frías es a la inversa), divididas por una zona de descenso brusco llamada termoclina. La zona superficial recibe el nombre de epilimnio y la profunda hipolimnio (Roldán Pérez & Ramírez Restrepo, 2008). Otro tipo de estratificación es la lumínica, que se da por las características de turbiedad del agua, donde asimismo se tienen capas diferenciadas en términos de la disponibilidad de

luz. En ausencia de estratificación el cuerpo de agua se dice que está en condición de mezcla. La variable definida en este estudio en el Anexo II combina ambos criterios.

- **Nitrógeno y sus variantes.** La importancia de este elemento en el agua radica en que es el componente fundamental de las proteínas, las cuales constituyen la base estructural de los seres vivos. El nitrógeno puede ser tomado por algas y plantas en tres formas: como nitratos (NO_3^-), como ion *amonio* (NH_4^+) y como nitrógeno molecular (N_2) (Roldán Pérez & Ramírez Restrepo, 2008).
- **Fósforo y sus variantes.** Se trata del elemento biológico que desempeña el papel más importante en el metabolismo biológico. Es el menos abundante de los elementos y el que desempeña un rol limitante en la productividad primaria (Roldán Pérez & Ramírez Restrepo, 2008). Existen distintas formas de medir el fósforo en el agua: fósforo reactivo soluble y *fósforo total*, que se obtiene por distintos pretratamientos de la muestra, entre otros. Para mayor detalle al respecto se sugiere remitirse a American Public Health Association (2017).
- **Clorofila- a:** es uno de los pigmentos que caracterizan a algas y cianobacterias. Hay varios tipos de clorofila, siendo la clorofila-a la que domina como pigmento fotosintético primario y es a menudo la utilizada para estimar biomasa (Wetzel & Likens, 2000).
- **Carbono orgánico total (COT):** se refiere a todo el material orgánico contenido en el agua (Roldán Pérez & Ramírez Restrepo, 2008).
- **Fitoplancton:** “son los productores primarios de plancton” (Roldán Pérez & Ramírez Restrepo, 2008). El fitoplancton de agua dulce se compone de bacterias fotosintéticas y algas, que varían en tamaño y forma de los organismos (Sigee, 2005).

Estas variables limnológicas (físicas, químicas y biológicas) se encuentran relacionadas y el diagrama que se observa en la Fig. 2 resume algunas de las principales asociaciones.

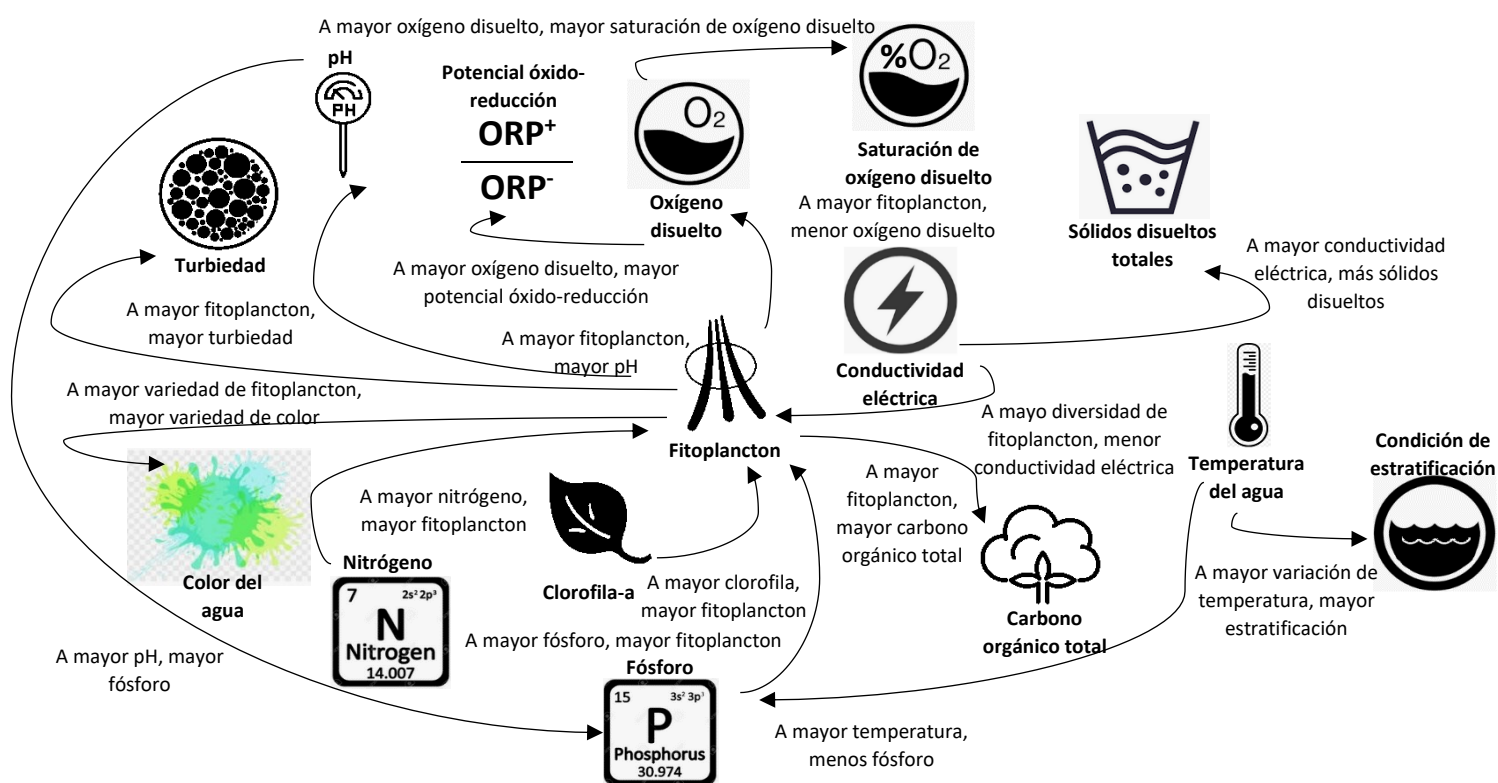


Fig. 2 Diagrama de algunas de las principales relaciones entre las variables limnológicas. Fuente: elaboración propia

2 Materiales y métodos

En este capítulo se detallan todos los materiales, tanto los de utilización directa para el análisis como son los datos, como los instrumentos utilizados para la medición de las variables y las características geográficas del entorno relevado.

2.1 Sitio de estudio: Embalse San Roque

El Embalse San Roque (ESR) se encuentra ubicado en la provincia de Córdoba, dentro del departamento de Punilla. Su ubicación exacta es 31° 22' 36" (S) - 64° 27' 54,6" (O) y se encuentra a una altitud de 608 msnm (metros sobre el nivel del mar). Se denomina embalse, y no lago en tanto se trata de un cuerpo de agua artificial, resultante del embalsamiento de sus actuales ríos tributarios. El dique o represa que lo forma, es una estructura de hormigón que embalsa 210 Hm³ de agua, ocupando una superficie de 1.630 Ha, siendo su profundidad media de 13 m. El tiempo de residencia, que es el tiempo promedio que toma rellenar la cuenca del embalse con agua si se vaciara, proporciona una medida de la circulación de agua dentro de los ecosistemas acuáticos (Sigee, 2005), es de 0,6 años. La cota del vertedero alcanza una altura de 35,5 m (Instituto Nacional del Agua, 2014).

El dique se construyó dos veces, la primera en 1888 (referencias posteriores a *paredón viejo*, se refieren a parte de la estructura que contenía al embalse en esta primera construcción) y la segunda en 1944. El embalse abastece de agua potable a un total de 940.000 personas en la Ciudad de Córdoba, representando más del 90% del caudal total (el restante es aportado por el dique Los Molinos). El caudal remanente permite el riego por medio de canales. Existe una central de energía eléctrica que aprovecha el flujo de agua a través del dique, denominada central San Roque, ubicada a 11km aguas abajo sobre el curso del río Suquía (construida en 1959), genera 26MW/h.

La red hidrográfica está conformada por los ríos Cosquín y San Antonio con módulos de aproximadamente 5,5 m³/s y 3,5 m³/s respectivamente, por un lado, y los arroyos Las Mojaras y Los Chorrillos con cerca de 0,3 m³/s cada uno (Rodríguez et al., 2000).

La Fig. 3 muestra una imagen del ESR.



Fig. 3 Panorámica del Embalse San Roque al amanecer, desde la margen sur próxima al muelle de embarque.
Fuente: elaboración propia

2.2 Monitoreo de variables de calidad del agua

Desde el año 1998, el INA-CIRSA realiza el monitoreo de la calidad del agua del ESR y de sus ríos tributarios. El mismo se realiza mensualmente y se registran variables físicas, químicas y biológicas. Los objetivos de este programa de monitoreo son realizar estudios limnológicos y hacer el seguimiento del proceso de eutrofización del lago (Instituto Nacional del Agua, 2014). En las secciones siguientes se describen los detalles al respecto.

2.2.1 Sitios de muestreo, instrumental y técnicas de análisis

En lo que respecta al monitoreo del lago, se realiza en los cuatro puntos de desembocadura de los ríos afluentes: San Antonio (DSA), Cosquín (DCQ), Las Mojaras (DLM) y Los Chorrillos (DLC), otros dos puntos ubicados dentro del lago: el centro (C) y la presa (TAC). Cabe mencionar además que otros dos puntos se muestrearon en el pasado: Paredón Viejo (PV) y Club Náutico (CN), pero no se analizarán en este estudio. La Fig. 4 muestra los puntos de muestreo mencionados, mientras que la Tabla 1 muestra el detalle de las distintas ubicaciones físicas asociadas a los puntos de muestreo en el lago y sus respectivas denominaciones (a lo largo del tiempo han variado, de ahí que se tenga más de una entrada por cada caso en la tabla).

En lo que se refiere al monitoreo de los ríos tributarios, el mismo se realiza en puntos específicos del recorrido de cada uno, tal y como se detalla en la Tabla 2. Existe un punto de muestreo particular, que corresponde al Río Suquía, efluente del agua del lago, que atraviesa la presa, el cual no es tenido en cuenta en el presente estudio.

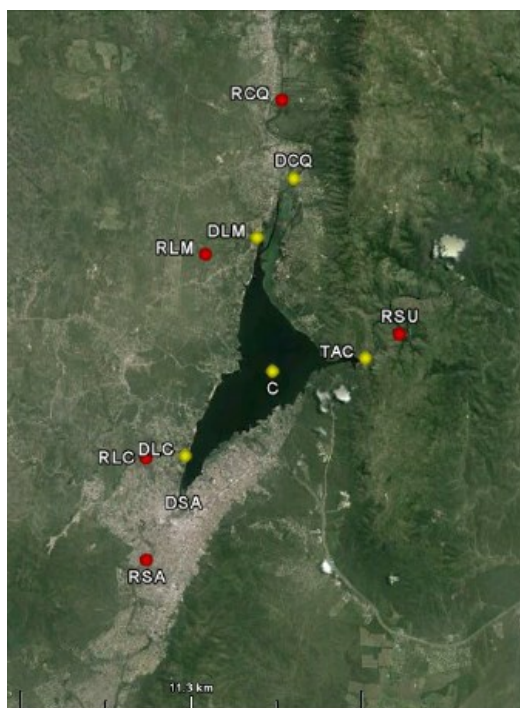


Fig. 4 Puntos de muestro del programa de monitoreo del ESR y su cuenca. Los puntos en color rojo indican los ríos (RSA: Río San Antonio, RCQ: Río Cosquín, RLC: Río Los Chorrillos, RLM: Río Las Mojaras, RSU: Río Suquía) y los puntos en color amarillo los sitios de muestreo en el lago propiamente dichos (DSA: Desembocadura San Antonio, DCQ: Desembocadura Cosquín, DLC: Desembocadura Los Chorrillos, DLM: Desembocadura Las Mojaras, C: Centro, TAC: Toma de Aguas Cordobesas, o presa). Fuente: *Instituto Nacional del Agua (2014)*

Tabla 1 Detalle de la denominación y ubicación de los puntos de muestreo en el lago

Código	Nombre	Latitud (S)	Longitud (O)
C	Centro del lago	31° 22' 36"	64° 27' 54.6"
TAC	Área de presa	31° 22' 24.27"	64° 25' 58.25"
DLM	Desembocadura Arroyo las Mojaras.	31° 20' 13.77"	64° 28' 8.38"
DCQ	Desembocadura del Río Cosquín próxima al puente	31° 19' 11.34"	64° 27' 21.88"
	Desembocadura del Río Cosquín en Plaza Federal	31° 20' 37.61"	64° 28' 9.48"
	Desembocadura del Río Cosquín conjunta al Arrollo las Mojaras	31° 20' 23.30"	64° 28' 78"
DLC	Desembocadura del Arroyo Los Chorrillos.	31° 24' 1"	64° 29' 52.34"
DSA	Desembocadura del Río San Antonio próxima al puente	31° 24' 56.06"	64° 29' 49.03"
	Desembocadura del Río San Antonio próxima al hotel Hippocampus	31° 24' 35.24"	64° 29' 48.36"

Tabla 2 Detalle de la denominación y ubicación de los puntos de muestreo en los ríos

Código	Nombre	Latitud (S)	Longitud (O)
RLM	Arrollo Las Mojaras a la altura de Azud	31° 20' 29.87"	64° 29' 14.84"
RLC	Arrollo Los Chorrillos a la altura del balneario El Diquecito	31° 24' 5.39"	64° 30' 34.89"
RSA	Río San Antonio a la altura del balneario Fantasio	31° 25' 53.88"	64° 30' 37.26"
RCQ	Río Cosquín a la altura de Villa Caeiro	31° 17' 47.32"	64° 27' 35.58"

A efectos de llevar a cabo el monitoreo de calidad de agua, se cuenta con varios instrumentos. Algunas variables físicas y químicas de interés son relevadas *in-situ*, mientras que otras variables químicas y biológicas son determinadas en laboratorio, tanto en lo que respecta al agua del lago, como de los ríos.

En cuanto a las variables que se determinan *in-situ*, el procedimiento consiste en realizar mediciones en el agua por medio de una sonda multiparamétrica, que consiste en un equipo que consta de diversos sensores (parte sumergible), unido por un cable a una pantalla que permite realizar la lectura de los parámetros. La sonda se sumerge a distintas profundidades en los diferentes puntos de muestreo mencionados previamente. La variable principal en este sentido, en cuanto a que da contexto a todas las demás es la **profundidad**. El instrumento de medición indica la profundidad en metros (m) a la que se encuentra la sonda, calculada en función de la presión de la columna de agua. Para el detalle de las demás variables remitirse al Anexo I.

Al mismo tiempo se mide también la transparencia del agua en el lago, utilizando un **disco de Secchi**. La transparencia, es una característica del agua que varía con los efectos del color y la turbiedad. Su medición es de forma muy simple, con un disco de 25 cm de diámetro, con fondo blanco y pintado de modo tal que un cuadrante blanco alterne con uno negro (disco de Secchi). El mismo cuenta con una cuerda graduada en cm con el disco ubicado en cero. El procedimiento consiste en introducir el disco desde la superficie y observar a la profundidad a la cual deja de verse. La medición se realiza en un área de la superficie del agua a la sombra.

Por último, se releva también la **cota del lago**, a partir de una escala en metros que se encuentra sujeta a una de las paredes de la presa.

Por otra parte, el equipo que realiza la campaña de monitoreo, toma muestras de agua en los distintos puntos de muestreo mencionados y a profundidades determinadas, para ser luego analizadas en el laboratorio. A tal efecto se utiliza un muestreador vertical u horizontal de tipo Niskin, que permite, mediante el sistema de un mensajero, tomar la muestra de agua a las distintas profundidades de muestreo cuando se encuentra sumergido.

A continuación, se detallan las variables químicas y biológicas principales para este estudio, que se miden en laboratorio³:

- **Fósforo Total (PT)**: esta variable se mide en $\mu\text{g/l}$ por medio de dos métodos: A – espectrofotometría SM 4500 P y B – Flujo segmentado SM 4120.
- **Clorofila-a**: esta variable se mide en $\mu\text{g/L}$ por medio de los siguientes métodos: extracción con acetona, espectrofotometría, método clorofila-a con corrección de feofitina y clorofila total SM 10200-H.

Las restantes variables, se encuentran detalladas en el Anexo I.

La Tabla 3 muestra el detalle de los períodos relevados de datos por tipo de monitoreo (lago y ríos) y el medio utilizado (modelo de sonda; laboratorio a cargo de los análisis; u otros medios, como ser determinación por simple observación, Secchi o escala de cota).

³ A lo largo de los años, los laboratorios que realizan las mediciones cambiaron. En esta lista se detallan las técnicas de determinación que utiliza el laboratorio de Aguas Cordobesas, que actualmente realiza los análisis.

Tabla 3 Detalle de los períodos relevados por tipo de monitoreo y medio.

U-10: Horiba® Water Quality Chequer U-10.

U-23: Horiba® Multiparameter Meter W-23XD.

WTW: WTW® Multi 350i.

YSI: YSI® Multiparameter System 556.

UCC: Laboratorio de la Universidad Católica de Córdoba.

AC: Laboratorio de Aguas Cordobesas S.A.

Tipo de monitoreo	Sonda				Laboratorio			Otros
	U-10	U-23	WTW	YSI	UCC	INA-CIRSA	AC	
Lago	1999-2000, 2004,2006, 2008-2011	2000-2008	2011-2014	2014-2017	1999-2000		1999-2017	1999-2017
Ríos	1999-2016	2007 (eventual)		2016-2017	1999-2003	2003-2011	2011-2017	

2.2.2 Datos de calidad de agua del lago

Los datos de calidad de agua que se relevan a partir del programa de monitoreo en el ESR son, en parte, registrados de manera manual en una planilla papel que posteriormente se transcribe en formato digital y se revisa (datos relevados *in-situ*) y en parte informados por el laboratorio que realiza las determinaciones restantes (datos de laboratorio), en forma digital. Posteriormente, se conforma una nueva versión de una planilla MS Excel®, actualizada con los nuevos datos revisados. Esto se realiza una vez por año y está a cargo del equipo de calidad de agua de INA-CIRSA en la sede de Córdoba capital.

Los datos del lago se estructuran a partir de los siguientes variables de referencia: número de monitoreo (secuencial), registro (identificador del registro, secuencial), muestra (identificador que asigna el laboratorio a cada muestra que se analiza), fecha, sitio (correspondiente a uno de los códigos de los puntos de muestreo que se presentan en la Tabla 1), código (es un indicador del perfil de agua que se detallará enseguida) y Z (profundidad de la medición). Por cada combinación de estos valores, se tienen una serie de variables medidas. La profundidad, se mide en general metro a metro, y 1m por encima del fondo. El diagrama de la Fig. 5 ilustra el alcance del muestreo, en términos de los principales conceptos involucrados.

La Fig. 6 muestra un ejemplo de la estructura de las variables de referencia del lago, mientras que la Tabla 4 muestra el detalle de las principales variables medidas, su alcance y determinación. Las variables se pueden clasificar según su alcance sea: por profundidad, donde se tiene una medición por cada fecha, sitio, código y Z (nivel de profundidad); por perfil, donde se tiene una única medición por fecha, sitio y código; por sitio, donde la medición es única por fecha y sitio; o general, donde se tiene una única medición por fecha. A su vez, las mediciones de ciertas variables, de acuerdo al método de medición pueden presentar datos censurados si se alcanza el límite de sensibilidad. Tal es el caso de las variables químicas que se miden en laboratorio (donde el instrumento no permite determinar concentraciones menores a un cierto límite, hay una pérdida por defecto, y se presume menor) y la variable de transparencia o Secchi (donde el instrumento no permite determinar transparencia a mayor profundidad de un cierto límite, hay una pérdida por exceso, se presume mayor). Por último, vale mencionar que los monitoreos que contemplan la medición de todas las variables mencionadas con anterioridad se han efectuado desde septiembre de 1999 (anteriormente algunas de las variables de

laboratorio aún no se medían), empleando distintas sondas y laboratorios (en ocasiones de manera alternada) según estuvieron disponibles (los períodos de tiempo correspondientes se observan en la Tabla 3).

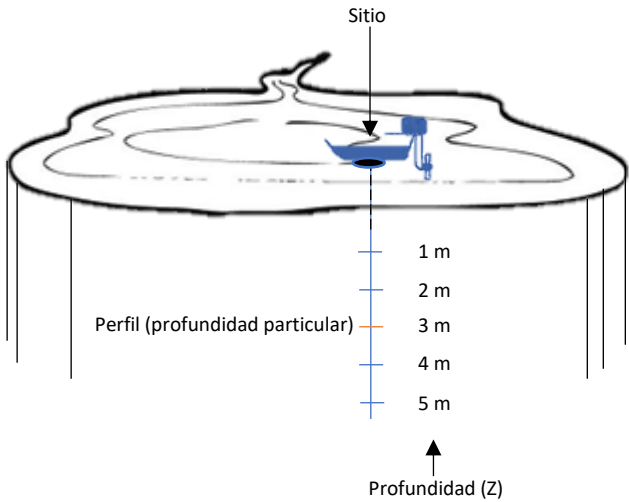


Fig. 5 Diagrama del alcance del muestreo. Se observa el sitio puntual sobre el plano horizontal del lago, las profundidades de toma sobre el eje vertical y un caso de una profundidad particular que reviste interés, denominada perfil. Fuente: elaboración propia

N° de Monitoreo	Registro	MUESTRA	Fecha	Sitio	Codigo	Z (m)
91	4518	125165	25-Sep-2002	C	C1	0.2
91	4519		25-Sep-2002	C		1.0
91	4520		25-Sep-2002	C		2.0
91	4521		25-Sep-2002	C		3.0
91	4522		25-Sep-2002	C		4.0
91	4523		25-Sep-2002	C		5.0
91	4524		25-Sep-2002	C		6.0
91	4525		25-Sep-2002	C		7.0
91	4526		25-Sep-2002	C		8.0
91	4527		25-Sep-2002	C		9.0
91	4528	125166	25-Sep-2002	C	C3	10.0
91	4529	125167	25-Sep-2002	C	C5	14.0

Fig. 6 Ejemplo de las variables de referencia de los datos del lago

Tabla 4 Detalle de las principales variables medidas en el lago, su alcance y determinación

Variable	Alcance	Determinación
Cota	General	Medición
Secchi	Por sitio	Medición
> Secchi	Por sitio	Indicación de censura
PT	Por perfil	Medición
< PT	Por perfil	Indicación de censura
Clorofila-a	Por perfil	Medición

Se menciona, además, que existen otras variables disponibles en la base, pero dado que se han comenzado a medir recientemente en relación a las mencionadas, o bien de forma irregular a lo

largo del tiempo, y al presentar una proporción considerable de valores faltantes para el período a analizar se descartan a los efectos de este estudio.

En relación al perfil de agua, en que se miden las variables de laboratorio, se trata de determinadas profundidades (en que también están medidas las variables *in-situ*) que se seleccionan siguiendo un criterio conjunto de luz y temperatura, determinado según la época (remitirse a la definición de condición de estratificación en la sección 1.3). Así en condición de estratificación, se tienen los siguientes: subsuperficial, fótica, epilimnio, hipolimnio y fondo; mientras que, en condición de mezcla, se tienen: subsuperficial, fótica, intermedia y fondo. La Tabla 5 muestra el detalle de los perfiles de agua. Dada la escasa cantidad de datos disponible para los códigos de perfil marcados en color gris, los mismos no serán tenidos en cuenta en el presente estudio.

Tabla 5 Detalle de los perfiles de agua

Sitio	Código del perfil	Mezcla		Estratificación	
		Criterio	Descripción del perfil	Criterio	Descripción del perfil
TAC	TAC1	0,20m	Subsuperficial	0,20m	Subsuperficial
	TAC2	2,5m*Secchi(m) (zona fótica)	Fótica	2,5m* Secchi(m) (zona fótica)	Fótica
	TAC3	NA		1m por encima de termoclina	Epilimnio
	TAC4	Altura de toma conductos turbina	Toma	Altura de toma conductos turbina	Toma
	TAC5	1m del fondo	Fondo	1 metro del fondo	Fondo
C	C1	0,20m	Subsuperficial	0,20m	Subsuperficial
	C2	2,5m* Secchi(m) (zona fótica)	Fótica	2,5m* Secchi(m) (zona fótica)	Fótica
	C3	NA		1m por encima de termoclina	Epilimnio
	C4	Entre 2 y 4m	Intermedia	1m por debajo de termoclina	Hipolimnio
	C5	1m del fondo	Fondo	1m del fondo	Fondo
DSA	DSA1	0,20m	Subsuperficial	0,20m	Subsuperficial
	DSA5	1m del fondo	Fondo	1m del fondo	Fondo
DCQ	DCQ1	0,20m	Subsuperficial	0,20m	Subsuperficial
	DCQ5	1m del fondo	Fondo	1m del fondo	Fondo
DLM	DLM1	0,20m	Subsuperficial	0,20m	Subsuperficial
DLC	DLC1	0,20m	Subsuperficial	0,20m	Subsuperficial

En lo que respecta a los datos de ríos, se detallan en el Anexo I.

2.3 Monitoreo de variables hidrológicas y meteorológicas

En esta sección se describen las características del procedimiento de relevamiento de datos hidrológicos y meteorológicos.

2.3.1 Estaciones meteorológicas y sensores

Los datos de las variables hidrológicas y meteorológicas son obtenidos a través del Sistema Telemétrico de transmisión de datos hidrológicos y meteorológicos a tiempo real (STC) del Centro de la Región Semiárida (CIRSA) perteneciente al Instituto Nacional del Agua (INA). El mismo se encuentra funcionando desde 1986. Este sistema se encuentra operado desde la

estación central de recepción ubicada en la sede Villa Carlos Paz del INA-CIRSA (Colladón & Vélez) y consta de una red de estaciones automáticas que están distribuidas a lo largo de las subcuencas que conforman la cuenca del ESR (ver sección 2.1). Entre las subcuencas más importantes se encuentran la del río San Antonio y la del río Cosquín, principales ríos afluentes. Además, existen cuencas menores como la cuenca de los arroyos Las Mojaras y Los Chorrillos, y una estación ubicada en la cuenca del río Suquía, único efluente del embalse.

A efectos de este estudio, se consideraron un total de 22 estaciones ubicadas en las subcuencas de los ríos afluentes, más 3 estaciones especiales, localizadas: en la sede Villa Carlos Paz del INA-CIRSA, en la zona del perilago, y en la cuenca del río Suquía. Existió adicionalmente una estación en el centro del lago, pero hoy en día no se encuentra operativa. Cada estación consta de un nombre y se encuentra identificada por un código numérico. La Tabla 6 indica código y nombre de la estación, cuenca a la que pertenece, coordenadas y altitud sobre el nivel del mar, mientras que la Fig. 7, muestra la ubicación geográfica de las estaciones consideradas para este análisis.

Tabla 6 Detalle de las estaciones analizadas

Cuenca	Código	Nombre	Latitud (S)	Longitud (O)	Altitud S.N.M (m)
San Antonio	500	El Galpón	31° 30' 29"	64° 49' 12"	2.380
	200	Puesto Pereyra (Los Gigantes)	31° 26' 56"	64° 48' 32"	2.292
	1200	Las Ensenadas (El Cóndor)	31° 36' 4"	64° 47' 28"	2.286
	1800	El Cajón	31° 26' 34"	64° 41' 31"	1.280
	400	Cañada Larga (Copina)	31° 34' 30"	64° 42' 42"	1.740
	900	Puesto Garay	31° 24' 55"	64° 44' 7"	1.625
	300	La Casita	31° 28' 60"	64° 44' 30"	1.555
	1000/1010	Confluencia Malambo	31° 29' 54"	64° 40' 45"	1.340
	2300	Observatorio Bosque Alegre	31° 35' 54"	64° 32' 59"	1.218
	2200	San Bernardo	31° 33' 48"	64° 32' 47"	965
	1700	La Quebrada	31° 29' 18"	64° 35' 59"	918
	700	Confluencia El Cajón	31° 29' 23"	64° 38' 20"	760
	600	Bo. El Canal	31° 26' 48"	64° 30' 52"	675
	1100	CIRSA - Villa Carlos Paz	31° 23' 57"	64° 28' 25"	660
	100	Santo Tomás	31° 33' 55"	64° 49' 35"	2.250
Cosquín	1400	La Hoyada	31° 21' 26"	64° 41' 39"	1.393
	2400	Pampa de Olaen	31° 11' 27"	64° 36' 26"	1.256
	2700	Villa Giardino	31° 3' 6"	64° 30' 6"	1.045
	2100	San José de los Ríos	31° 14' 54"	64° 36' 15"	967
	3900	Las Junturas Cosquín	31° 12' 31"	64° 29' 3"	716
	2900	Villa Caeiro	31° 17' 29"	64° 27' 37"	681
	4800	Camino El Cuadrado	31° 5' 52"	64° 26' 49"	1.378
	4050	Pan de Azúcar	31° 13' 60"	64° 25' 8"	1.100
Suquía	4400	El Diquecito	31° 21' 13"	64° 23' 36"	536
Perilago	3050/3051	Rep. Cerro Minero San Roque	31° 19' 47"	64° 25' 30"	1.100

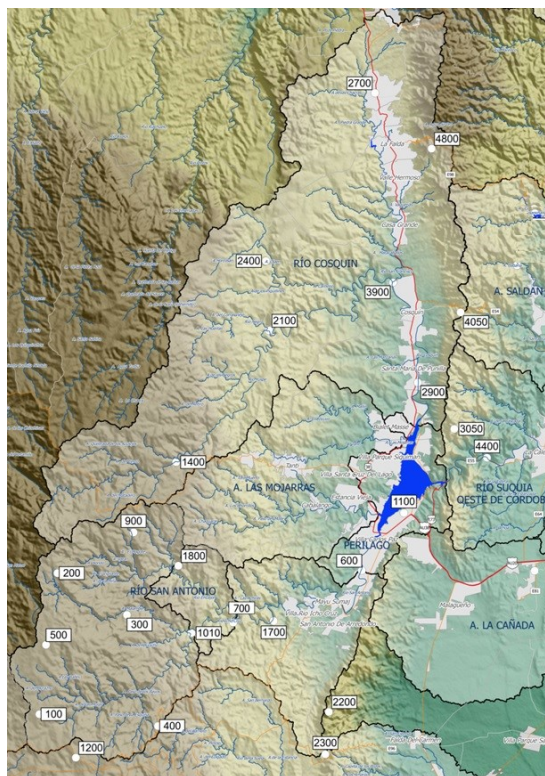


Fig. 7 Ubicación geográfica de las estaciones analizadas. Fuente: INA-CIRSA – cartografía: Ana L. Rydzewsky (geomorfología), equipo técnico: Andrea F. Rico y María E. Capone, diciembre de 2016.

Las estaciones se encuentran dotadas de instrumentos (sensores) que permiten realizar la medición de las principales variables meteorológicas y el nivel de los ríos. Cada puesto transmite en tiempo real, por banda radial *Very High Frequency* (VHF), en forma directa o a través de repetidoras, a la central de recepción y procesamiento ubicada en Villa Carlos Paz.

A continuación, se detallan los sensores instalados y su tipo, nuevamente en **negrita** las variables más relevantes en este estudio:

- 1) **Temperatura**: el sensor mide la temperatura del aire en grados centígrados (°C). En particular, el tipo de instrumento se denomina *termistor*, es un termómetro eléctrico.
- 2) **Humedad Relativa**: el sensor mide la humedad relativa del aire (representa una proporción, por ende, no tiene unidades, referirse a la sección 1.2), por medio de un dispositivo de tecnología propietaria denominado HUMICAP® (Vaisala, 2012).
- 3) **Precipitación**: el instrumento registra la lluvia caída en milímetros (mm). Para medir esta variable se utiliza un *pluviómetro* del tipo doble cubeta basculante. El dispositivo acumula el valor precipitado en un contador totalizador (con precisión de 1mm).
- 4) **Velocidad del viento**: el dispositivo mide la velocidad en kilómetros por hora (km/h), y se denomina *anemómetro con tazas*. El dispositivo acumula la distancia recorrida por los giros, y al completar los 5km (*step*) emite una señal, que, junto con el diferencial de tiempo transcurrido desde la última señal, permite calcular la velocidad.
- 5) **Dirección del viento**: en este caso se utiliza una *veleta*, que apunta en la dirección de destino del viento, entre 0° y 360°. Como este instrumento funciona en conjunto con el anterior, cada vez que se emite una señal para registrar la velocidad, a la vez se indica la dirección del viento en el momento del envío de la señal.
- 6) **Presión atmosférica**: el instrumento utilizado registra la presión en hectopascales (hPa), y se denomina *barómetro* del tipo *aneroide*.

- 7) **Radiación solar:** este sensor registra la radiación medida en langleys (Ly) se denomina piranómetro y es del tipo bimetálico fotovoltaico.

Todos los sensores detallados anteriormente son digitales y se encuentran ubicados a 4m de altura en los puestos.

Adicionalmente se mide la evaporación. En este caso existe un sensor que mide la cantidad de agua en milímetros (mm) que se transforma en vapor, pero por ser los datos de baja confianza no se consideran en este estudio.

Por último, se tienen sensores que miden el nivel de los ríos. Este tipo de instrumento se denomina *hidrómetro*, y es capaz de registrar el nivel medido en metros. En cada estación el sensor tiene un valor de altura que se denomina *nivel de referencia* del hidrómetro, que es el menor valor que es capaz de registrar. Este valor, es distinto del *cero*, que se refiere al punto físico en que se encuentra ubicado. Los sensores registran valores dentro de un rango predeterminado, pudiéndose establecer la sensibilidad del mismo para detectar la variación en el nivel del río (*step*) (Colladón, 2018). Análogamente, se tiene un sensor que media el **nivel** del lago.

2.3.2 Datos meteorológicos e hidrológicos relevados por sensores

Los datos relevados por los sensores ubicados en las estaciones automáticas de monitoreo, una vez transmitidos a la sede central de procesamiento son almacenados en los servidores en archivos de texto plano, mediante la intervención de software. Posteriormente, los archivos son revisados y comentados por un experto para garantizar la calidad de los datos. La Fig. 8 muestra el diagrama del proceso.

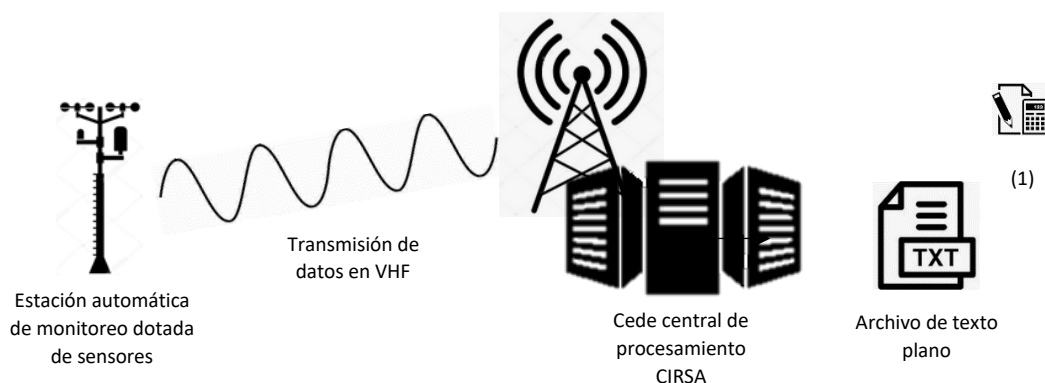


Fig. 8 Proceso de transmisión y almacenamiento de datos. (1) Revisión a cargo de un experto. Fuente: elaboración propia

Se genera un archivo por año, identificado por el año y el código del sensor. La estructura del archivo, consta de: encabezado con el código del sensor, nombre de la estación y tipo de sensor (variable que mide); y datos organizados en columna, donde por cada medición se registra fecha, hora y valor. En el caso particular del viento, en un mismo archivo se registra velocidad y dirección del viento (se tienen dos columnas de valor). La Fig. 9 muestra un ejemplo de archivo para la variable temperatura.

Sensor # 302 La Casita T. Temperature Sensor		
DATE	TIME	degrees C
12/31/2009	23:30:29	12.8
12/31/2009	22:28:35	14.5
12/31/2009	21:26:40	14.5
12/31/2009	20:24:46	14.5
12/31/2009	19:22:52	15.1
12/31/2009	18:20:58	15.6

Fig. 9 Extracto de un archivo típico que registra información de un sensor (temperatura en este caso).

En el caso de los datos de relevamiento manual de evaporación, los mismos se almacenaron en un archivo MS Excel®, donde consta fecha y valor.

Dadas las características geomorfológicas de cada una de las cuencas, y teniendo en cuenta las recomendaciones de la Organización Meteorológica Mundial, las estaciones automáticas se hallan estratégicamente ubicadas para censar la realidad que las rodea. La mayor parte de las mismas está dotada de sensores de precipitación y nivel de los ríos, puesto que el fenómeno de crecidas producto de precipitación intensa es quizás el fenómeno más importante que se busca monitorear en términos de la seguridad de los habitantes de estas regiones. En cuanto a los demás tipos de sensores, los más comunes son los correspondientes a temperatura y humedad relativa en la cuenca del río San Antonio; el resto se ubican en muy pocas estaciones, ya que son de más largo alcance.

Debido a diferentes factores, entre ellos la gradual instalación de las estaciones, dificultades técnicas en los procesos de medición y transmisión, rotura del instrumental o imposibilidad de mantenimiento o calibración, los datos no han estado siempre disponibles durante el período completo de medición (1986 al 31 de agosto de 2017), pudiendo faltar un año completo, meses de determinados años, fechas específicas o determinadas horas para una misma fecha. La Tabla 7 resume los períodos con datos para cada una de las estaciones y sensores que miden variables en términos de años (se consideró que exista al menos un mes con dato por año para la elaboración de la misma).

En lo que respecta a la frecuencia de medición, varía entre estaciones, sensores y fecha porque depende de la calibración del instrumento, pero considerando las medianas por estación, podría decirse en general, que es de entre 18 y 62 minutos para todas las variables que se consideran de medición *programada*: temperatura, nivel, presión atmosférica, radiación velocidad y dirección del viento. Este término hace referencia a que cada registro es el resultado de una medición instantánea que se realiza a plazos de tiempo preestablecidos, y dicho registro es suficiente para determinar el valor de la variable en el momento en que ocurre la medición.

En lo que se refiere a los sensores de la serie 2000 que miden el nivel o cota del lago, algunos han operado de forma instantánea como los demás y otros (de más largo alcance) lo han hecho a partir de tomas diarias. Cabe mencionar además que los mismos midieron nivel del lago de manera constante hasta el 14 de noviembre de 2016, con posterioridad el registro fue inestable, habiendo solo algunos casos en el otoño-invierno de 2017. Por tal motivo, se incorporaron datos al análisis que provienen de la lectura manual de la cota del nivel del lago, a partir del 15 de noviembre de 2016. Estos datos fueron aportados por el Ministerio de Agua y Ambiente y Servicios Públicos de la Provincia de Córdoba (MAAySP).

Tabla 7 Períodos con dato para cada estación y sensor

Código	Precipitación	Temperatura	Nivel	Humedad Relativa	Velocidad del Viento	Dirección del Viento	Radiación Solar	Presión Atmosférica	Evaporación
500	1986-2017								
200	1992-2017	1994-1997		1987,1989-1996					
1200	1990-2017	1994-2006		1986-2006					
1800	1989-2017		1991-2017						
400	1992-2017	1994-2017		1991-2017					
900	1992-2017	1994-2017		1990-2017	1994-1997, 2000-2017	1994-1997, 2000-2017	1986,1989-2017	1986, 1989-2017	
300	1991-2017	1994-2017		1990-2017	1994-2017	1994-2017	1989-2017	1989-2009, 2011-2017	
1000/1010	1989-2017	1994-2016	1991-2017(*)						
2300	1994,2009, 2011-2017								
2200	2008-2009, 2011-2017								
1700	2006-2017								
700	1988-2017	1994-2013	1992-2017(*)	1992-2000, 2011	1994-2014	1994-2014	1989, 1991-2010, 2014-2015	1991-2009	
600	1989-2017	1994-2008	1987-2017						
1100	1986-2017	1994-2017		1989-2017	1994-2017	1994-2002, 2005-2017	1993-2017	1990, 1992-2017	1994-2004
100	2006-2017								
1400	2011-2017		2011-2012, 2014-2017						
2400	1992-2017								
2700	1991-2017								
2100	1993-1997, 2010-2017		1991-1992, 1994-1995, 1997, 2010-2017						
3900	2010-2017		2010-2012, 2014-2017						
2900	2010-2017		2010-2017						
4800	2011-2017								
4050	2011-2017								
4400	2011-2017		2011-2017						
3050/3051	2011-2017								
2000	1989-2000		1994-2017(*)						

Nota: una celda vacía indica que no existe sensor para la variable y estación correspondientes.

(*) Reúne un conjunto de sensores.

En el caso de la precipitación, la medición es *no programada*, ya que se realiza cuando se produce 1 milímetro de precipitación y la naturaleza del registro es acumulativa. Esto quiere decir, que la determinación del valor de la variable es *diferencial*, así, por ejemplo, a efectos de establecer la lluvia precipitada en un día, debe calcularse la diferencia entre el primer y el último registro para esa fecha. El dispositivo de medición en este caso emite una señal con el valor actual del acumulador dos veces al día (cada doce horas) en caso que no ocurra precipitación, como señal de latido, indicando que se encuentra operativo. Otro aspecto importante respecto de la lógica de funcionamiento del sensor de precipitación, que afecta la interpretación de los datos, es que al producirse el volcado de uno de los recipientes por alcanzar el peso suficiente, el otro recipiente entra en acción generando un

reinicio del acumulador a cero. Esto significa, que, a partir de ese momento, el primer registro es cero, y el registro siguiente marca uno, para indicar el primer milímetro medido desde el volcado.

2.4 Preprocesamiento de datos

En esta sección se describe la etapa de preprocesamiento consistente en la depuración, cálculo y estructuración de los datos crudos tendiente a generar el conjunto de datos apto para la aplicación de técnicas y modelos de Minería de Datos.

2.4.1 Datos de calidad del agua

En relación a los datos del agua del lago y los ríos, los mismos presentan una buena organización y no se detectaron errores en la registración. La estrategia de preprocesamiento tuvo dos objetivos:

- 1) Por un lado, completar los registros en situaciones donde los datos se encontraban disponibles solo a determinado perfil (*imputación por extensión*) o bien se podían calcular a partir de otros datos (*imputación teórica por cálculo*).
- 2) Por otra parte, calcular variables derivadas de las variables originales (secundarias), ya sea para estructurar texto libre (*estructuración de variables categóricas*) o bien para enriquecer la base con variables que resumen aspectos interesantes desde el punto de vista del dominio.

En este sentido, como primer paso se realizó la carga de los datos en formato MS Excel® en tablas relacionales (una para los datos del lago y otra para los datos del río).

A continuación, se describe el proceso aplicado a los datos correspondientes a las variables principales del lago, donde se siguió la estrategia de imputación por extensión. Este tipo de imputación aplica a casos donde los datos se han registrado en un determinado nivel, y corresponde *extenderlos* a otros, vale decir, la medición atañe a más registros de los que se ha imputado.

- Cota: representa una condición general de estado del lago, y se registra en el perfil subsuperficial por convención, correspondiendo a todos los sitios la misma medición. A efectos de completar los registros que no tienen valor, se extendió dicha medición de cota desde el corte subsuperficial a los demás cortes de profundidad y sitios en caso que faltaban (desde el sitio en que esté completa). Si existía discrepancia entre los sitios se corrigió de acuerdo al registro escrito del día. En este caso fue posible completar la totalidad de los registros.
- Secchi: esta variable representa asimismo una condición de estado, pero referida al punto de muestreo, lo que significa que difiere de sitio a sitio, pero asimismo (como en el caso anterior) tiende a registrarse en perfil subsuperficial. Así, se extendió el valor registrado por sitio y corte al mismo sitio para los demás cortes, completándose un total

de 1518 registros sobre 1766 sin valor. Los restantes 248 no pudieron completarse porque se trata de verdaderos faltantes de la variable.

En cuanto a las demás variables del lago y las de ríos, se tiene el detalle en el Anexo II.

2.4.2 Datos meteorológicos e hidrológicos

En lo que respecta al preprocesamiento de los datos meteorológicos e hidrológicos, los objetivos fueron, por un lado, depurar los errores propios del funcionamiento de los sensores y/o la transmisión de los datos (en adelante denominados *inconsistencias*), y por otro, agregar los datos instantáneos a nivel horario o diario.

Como primer paso tendiente a lograr los objetivos mencionados, los datos disponibles en formato de archivo de texto plano se estructuraron en una tabla relacional (una por variable). Esta tabla se definió con los siguientes campos: id (identificador único del registro y clave primaria de la tabla), estación, fecha, hora, medición (por ejemplo, temperatura, presión atmosférica, etc.) y observaciones (para volcar los comentarios vertidos por el experto revisor del archivo). Este paso se denominó carga inicial.

Con posterioridad, el proceso se dividió en dos etapas principales: por un lado, la identificación de inconsistencias, y por otro lado la agregación de los datos filtrando los registros aislados en el paso anterior. En el caso particular de la variable precipitación, la naturaleza de la medición es acumulativa, a diferencia del resto de las variables donde es instantánea. Por tal motivo, las etapas de procesamiento para esta variable son distintas. A continuación, se describen las etapas para el caso general y de precipitación.

2.4.2.1 Caso general

La Fig. 10 muestra un diagrama general de la fase de preprocesamiento de los datos, y a continuación se describe cada etapa (en cada punto, los números indicados entre paréntesis corresponden a las actividades descritas en el diagrama).

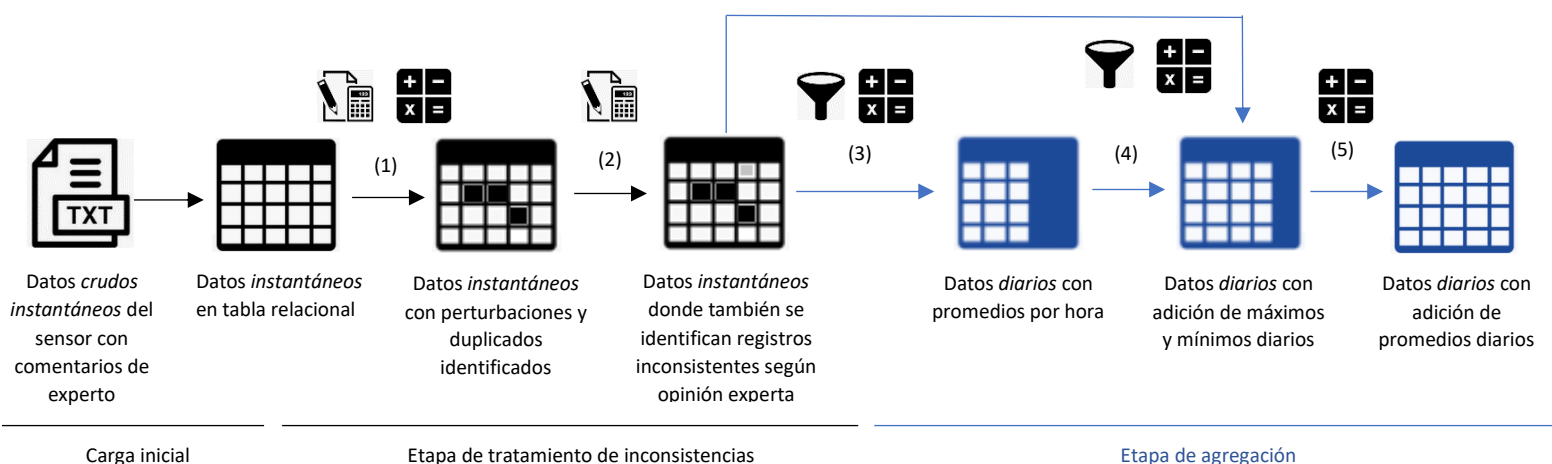


Fig. 10 Diagrama del flujo de pre procesamiento de datos meteorológicos e hidrológicos. Referencia de íconos: papel y lápiz corresponde a tarea que requiere intervención manual, cuadrante de operaciones indica cálculo automático y embudo indica operación de filtrado. (1) Detección de perturbaciones. (2) Corrección por comentarios del experto. (3) Cálculo de promedios por hora. (4) Cálculo de máximo y mínimo diario. (5) Cálculo del promedio diario y casos especiales. Fuente: elaboración propia

2.4.2.1.1 Tratamiento de inconsistencias

Esta etapa consiste en identificar diferentes tipos de anomalías en el registro de los datos. Es una etapa fundamental para poder realizar luego la fase de agregación, ya que cualquier irregularidad o error en los datos puede eventualmente distorsionar las métricas de cálculo que se aplicarán posteriormente.

(1) Detección de perturbaciones: es preciso identificar irregularidades que pudieran eventualmente distorsionar las métricas de cálculo, alteraciones que se producen en el desarrollo normal de un proceso en adelante *denominadas perturbaciones*. Las mismas pueden agruparse en dos clases, tal y como se menciona a continuación:

a. Perturbaciones en sentido estricto: entendidas como uno o más valores consecutivos que en el contexto de la serie temporal resultan atípicos y deben ignorarse para evitar sesgar los promedios. Estos valores, una vez identificados se marcan en el campo de observaciones de la tabla, para permitir su filtrado con posterioridad. Las hay de tres tipos:

i. Perturbaciones puntuales: las que implican un incremento o un descenso de un único valor en el contexto de la serie. Este es el caso más frecuente, y para su corrección se considera el cumplimiento de tres condiciones, a saber:

1. Perturbación. Sean tres registros A, B y C, tal que A se produce primero, B después que A y C después que B, tal que son consecutivos en ese orden; entonces la medición A es menor que B y C es menor que B (B es un pico) o A es mayor que B y C es mayor que B (B es una depresión). Por ejemplo, considerando la variable temperatura, la siguiente secuencia representa la primera condición (pico): A=10°C, B=20°C, C=11°C; mientras que la siguiente secuencia cumpliría con la segunda (depresión): A=30°C, B=24°C, C=29°C.

2. Límite temporal. Responde a la pregunta: ¿hasta cuándo se considera una perturbación? Vale decir la tolerancia en términos de tiempo transcurrido entre dos registros consecutivos, uno de los cuales es candidato a ser perturbación puntual. En este sentido la estrategia utilizada consistió en calcular el valor absoluto de la diferencia de tiempo entre dos registros consecutivos, y posteriormente un determinado percentil de la distribución de las diferencias de tiempo según sus características (ver Tabla 8 para el detalle), sin por ello seleccionar valores demasiado altos que pudieran hacer ver como perturbaciones fenómenos de más largo alcance por falta de datos (crecidas de los ríos, descenso brusco de temperatura por avance de un frente frío, entre otros).

Un caso especial en cuanto al límite temporal constituye la variable nivel, donde, cuando se mide para el lago, las estaciones 2003 y 2004 lo hacen de manera *instantánea* (como se definió antes), mientras que la estación 2020 lo hace de manera diaria. A su vez esta estación finaliza su actividad el 28 de agosto de 2017 y desde entonces los datos que se consideran en este análisis son los correspondientes a los datos de recolección manual (desde noviembre de 2016 se consideran ambos), también de forma diaria. Dicho esto, es preciso establecer límites temporales por separado: para las dos estaciones que miden nivel varias veces al día y para la estación restante y los datos de recolección manual donde se tiene la información una vez al día. En la Tabla 8, se pueden observar dos filas para el nivel del lago, una para cada caso de los expuestos.

3. Límite de medición. Responde a la pregunta: ¿hasta cuánta variación comprenderá una perturbación? Se refiere a la tolerancia en términos de la diferencia en la magnitud de dos mediciones consecutivas, una de ellas candidata a ser considerada perturbación puntual. En este caso se recurre a una estrategia similar a la del punto anterior, calculando el valor absoluto de las variaciones de magnitud entre registros consecutivos, y posteriormente estableciendo un umbral superior (tolerancia) por medio de la definición de valor extremo severo: 3 veces la diferencia intercuartílica más el cuartil 75.

Un caso particular constituye la dirección del viento, que sigue una distribución circular (0° es igual a 360°), por ende, se calculan cuartiles circulares.

Otro caso especial resulta la variable nivel que, al medirse para ríos y lagos, la dinámica y características propias de cada uno hace que las distribuciones sean completamente distintas y por ende deban realizarse por separado las determinaciones: por un lado, estaciones que miden nivel de ríos y por otro las que hacen lo propio con el nivel del lago. Con respecto al caso del río, para la diferencia en tiempo establecida en el punto anterior, la variación en magnitud es 0, lo cual da indicio de no existencia de perturbaciones puntuales en este subgrupo.

- ii. Perturbaciones no puntuales o extendidas: las que implican un alza o una baja de al menos dos valores, en el contexto de la secuencia. A tal efecto se identificarán como perturbaciones no puntuales los casos que cumplan las siguientes condiciones:

1. Perturbación. Sean cuatro registros: A, B, C y D, donde A se produce primero, B después que A, C después que B y D después que C. A es el primer registro anterior a B sin perturbación, mientras que D es el registro inmediato siguiente a C sin perturbación. Entre B y C puede haber cero o más registros, y estos dos son los que marcarán el comienzo y el final de la secuencia de valores atípicos, respectivamente. Por ejemplo, considerando la variable temperatura, la siguiente secuencia cumpliría con la condición: A=10°C, B=20°C, C=20°C, D=11°C (se trata de un pico extendido).
 2. Límite temporal. Entre cada uno de los registros consecutivos que intervienen debe haber una diferencia no mayor en tiempo que corresponde, siguiendo la estrategia indicada en el punto anterior, al percentil 75 como valor límite. Esto apunta a reducir aún más las posibilidades de falsos positivos en términos de perturbaciones.
 3. Límite de medición. Entre los registros externos (A y D) y los registros límite internos (B y C) debe haber una diferencia de magnitud mayor a la identificada en i.3 (marcando el inicio y el final de la secuencia). En cambio, entre los registros límite internos la diferencia debe ser menor a ese mismo valor (se espera que la perturbación se mantenga en el tiempo).
Vale aclarar que en aquellos casos en que, por la naturaleza de la variable se consideró que la perturbación continua podía ser consistente con fenómenos perfectamente explicables (como los antes mencionados), no se aplicó filtrado por perturbación continua (ver Tabla 8 para el detalle).
- iii. *Perturbaciones excepcionales o valores extremos*: constituyen mediciones minoritarias extremadamente altas o bajas en el contexto de todas las series, que escapan a las definiciones en i y ii (alguna de las condiciones antes explicitadas no se cumple), pero que sin embargo no hacen sentido en términos de la naturaleza de la variable o bien no se tiene evidencia histórica disponible de que tales valores se hubieran producido alguna vez. A efectos de realizar esta corrección, se consideró la definición de valor extremo severo para determinar los umbrales de tolerancia (ver Tabla 8), utilizando diferentes percentiles según las características de la distribución.
- En el caso de las variables dirección del viento y humedad relativa, que se encuentran acotadas a un rango, se refiere a los valores fuera de dicho rango. En particular, la humedad relativa tiene un límite inferior de 0 y un límite superior que se definió en 1.01 (101%) para preservar la condición excepcional de sobresaturación que podría presentarse. Para el detalle remitirse a Ahrens y Henson (2017), pág. 169 para la elección del límite y pág. 97 para la definición correspondiente.
- Un caso particular constituye en este sentido el nivel de los ríos, donde el mínimo nivel viene dado por la altura de instalación del sensor (variable de acuerdo a la estación y algunos casos el período de tiempo) determinando el menor valor que es capaz de registrar. En este caso, se siguió la documentación técnica referida al respecto para filtrar valores por debajo del límite. Para mayor detalle al respecto referirse a Colladón (2018).

- b. Valores duplicados: se trata de dos registros que, o bien tienen la misma fecha y hora, o bien tienen la misma fecha y una diferencia muy pequeña en la hora, y no constituyen perturbaciones. A partir de las observaciones indicadas por el experto al respecto de duplicados en los rangos de fecha en que efectivamente se observan duplicados (dos mediciones iguales) en algunas estaciones, se establece un orden de frecuencia, donde esa diferencia pequeña entre las horas puede ser de hasta 5 segundos siendo este el tiempo máximo menor a un minuto (entendiendo que el sensor pasado este umbral es capaz de emitir una nueva señal que no constituiría un duplicado sino una nueva medición, con mayor frecuencia de lo normal) con mayor ocurrencia entre todas las variables en que se pudo determinar. De esta manera, los registros que cumplan dicha condición y además el valor de la medición sea el mismo, entonces arbitrariamente todos excepto uno, tendrán un comentario en el campo de observaciones que los identifique como duplicado, para que se puedan filtrar con posterioridad. En el caso en que la medición no coincida (caso minoritario), todos los registros intervienen del cálculo posterior (quedan sin comentario, no hay razón para invalidar alguno, se interpreta que son registros independientes, que se han generado con una frecuencia más alta por alguna razón de configuración momentánea del dispositivo de medición). Referirse a la Tabla 9 para el detalle de las observaciones afectadas por este paso.

- (2) Corrección por comentarios del experto: se procede a revisar las observaciones formuladas por el experto para identificar en la tabla como inválidos aquellos registros que presenten alguna irregularidad. Referirse a la Tabla 9 para el detalle de las observaciones afectadas por este paso.

2.4.2.1.2 Agregación

Como paso previo al cálculo de los valores agregados se crearon las correspondientes tablas con los siguientes campos: id (identificador único del registro), estación, promedio hora 0, ..., promedio hora 23 (24 campos de promedio, uno por hora), máximo diario, mínimo diario y promedio diario y observaciones.

- (3) Cálculo de promedios por hora: se realiza el promedio de todas las mediciones instantáneas para cada estación y fecha, para cada una de las 24 horas (desde las 00 hasta las 23), desestimando todos aquellos registros marcados como perturbación o como duplicado.
- (4) Cálculo de máximo y mínimo diario: se calcula el máximo y el mínimo partiendo de las mediciones instantáneas por estación y fecha. A tal efecto, es necesario previamente determinar que condición se exigirá para establecer dichos valores, teniendo en cuenta que puede que existan valores faltantes para una o más horas y el promedio se vea sensiblemente distorsionado. En este sentido, la metodología empleada fue conformar la distribución circular de las horas de los máximos y mínimos diarios (frecuencia) y posteriormente tomar los cuartiles 25 y 75 como hora de inicio y fin, rango en el cual deben existir mediciones para calcular máximo y mínimo respectivamente (independientemente de que el mínimo o máximo pueda producirse en esos casos en una hora distinta, se garantiza que pertenezca el grueso de la distribución, donde es más probable que ocurra). Caso contrario no se calculan. Por ejemplo, en el caso de temperatura se espera que el máximo se produzca después del mediodía y el mínimo al comienzo del día.

Un caso particular, a este respecto, constituye la variable dirección del viento, donde no aplica el concepto de mínimo o máximo. Otro caso especial es la velocidad del viento, donde la ubicación de las estaciones que relevan la información hace que la distribución por hora sea marcadamente diferente y por ende las determinaciones se realizan de manera individual por estación.

(5) Cálculo del promedio diario: se realiza basado en los promedios por hora⁴, a excepción del nivel cuya dinámica no responde a fenómenos del día (en este caso el promedio diario es el promedio de todos los registros del día correspondiente). Este cálculo se lleva a cabo por tandas y se indica en la columna de observaciones cuál fue el procedimiento empleado en cada caso (ver Tabla 10 para el detalle de los casos en que aplica cada paso):

a. Casos completos. En primer término, se calculan todos los casos que tienen promedios para todas las horas.

Un caso especial en este sentido, constituye la variable Nivel, donde no hay diferencias marcadas en lo que respecta a la distribución por hora de las mediciones y por ende el promedio aplica a todos los casos independientemente de los faltantes. El otro caso es la variable dirección del viento, donde por naturaleza de la variable, como se mencionó no hay máximo ni mínimo (necesario para el procedimiento que se describe) para alcanzar a cubrir la mayoría de casos como ocurre con las restantes variables⁵.

b. Casos semicompletos. En segundo lugar, de los casos remanentes, se calculan los promedios donde se tienen 2 o más promedios por hora cada 3 horas, agrupando desde las 00 horas en lo sucesivo (por ejemplo: al menos dos de entre las 0, 1 y 2 horas).

c. Casos en horas centrales. Seguidamente, de entre los casos remanentes, se calculan aquellos promedios donde exista al menos un promedio por hora de los siguientes grupos con centro en las horas 9, 15 y 21: 8, 9 y 10; 14, 15 y 16; y 20, 21 y 22. El promedio es ponderado hacia los extremos, de manera que los grupos laterales pesan 9 y el grupo central pesa 6. Primeramente, se calculan los promedios de cada grupo, y luego el promedio ponderado final. Cabe aclarar que este procedimiento está basado en la naturaleza de la variable temperatura, y por ende se aplica únicamente en los casos de variables que tienen un comportamiento similar o consistente con ella. Ver Tabla 10 para el detalle.

d. Casos por mínimo y máximo. Por último, se recurre a la condición más laxa posible para los casos remanentes, calculado el promedio a partir del máximo y mínimo que deben estar simultáneamente presentes.

Otros cálculos especiales

Por último, cabe aclarar que para algunas magnitudes se calcularon variables adicionales:

a. Nivel: marca de censura, identificando aquellos registros en que al menos uno de los promedios horarios coincidió con el mínimo nivel registrable por el sensor, como una aproximación a la posible falta de datos por no alcanzar el río la altura mínima medible.

⁴ A efectos de hacer la serie resultante menos susceptible a variaciones espurias no capturadas.

⁵ Solo podrían aplicarse los pasos a. y b., obteniendo un 28% de casos a nivel diario.

Velocidad del viento: intensidad en la escala de Beaufort como una manera de clasificar los vientos en categorías según rangos de velocidad.

- b. Dirección del viento: orientación de la intensidad del viento considerando 8 puntos cardinales: N, NE, E, SE, S, SO, O y NO según rangos medidos en grados.

La Tabla 10 resume los procedimientos aplicados en cada caso junto con el porcentaje de casos afectados para las variantes del cálculo de promedio diario.

Tabla 8 Resumen de técnicas y parámetros aplicados en la etapa de tratamiento de inconsistencias: detección de perturbaciones.

Variable	Perturbación puntual			Perturbación continua	Valores extremos		
	Ventana de tiempo (minutos)	Criterio (percentil)	Variación tolerada	Ventana de tiempo (minutos)	Límite inferior	Límite superior	Criterio (percentiles)
Temperatura	62	90	4,4°C	33	-19,9°C	47,3°C	25-75
Nivel de lago	182	90	0,04m	NA	(*)	55,16m	5-95
Nivel de lago (diario)	1440	90	0,13m	NA	(*)	55,16m	5-95
Nivel de río	22	90	0m	NA	(*)	5m	5-95
Humedad relativa	62	90	0,148	62	0	1,01	NA
Velocidad del viento	73	90	10,4km/h	NA	0km/h(**)	147,70 km/h	10-90
Dirección del viento	37	75	0,173682º	NA	0º	360º	NA
Radiación solar	62	90	0,448ly	NA	0ly	2,43ly	25-75
Presión atmosférica	62	90	1,6hPa	62	669.6hPa	1092,40hPa	25-75

Nota: () El límite inferior o mínimo de nivel depende del punto de muestreo. NA: No Aplica.*

*(**) Cabe aclarar que, en el caso de la velocidad del viento, los valores con velocidad 0 corresponden a los casos en que justo se completó el step de 5km y se decide considerarlos para el cálculo (ver 2.3.1).*

Tabla 9 Resumen de proporción de registros afectados por el tratamiento de inconsistencias: otras inconsistencias

Variable	Duplicados		Comentarios
	Igual registro	Distinto registro	
Temperatura	0,39%	<0,01%	<0,01%
Nivel	1,77%	0,05%	0%
Humedad relativa	0,22%	<0,01%	0,12%
Velocidad del viento	0,05%	0%	<0,01%
Dirección del viento	0,06%	0%	3,39%
Radiación solar	0,48%	<0,01%	0%
Presión atmosférica	0,85%	<0,01%	1,24%

Tabla 10 Detalle de cálculo de promedio diario.

Variable	Mínimo	Máximo	Promedio				Otros
			Completo	2/3hs	Hs centrales	Min-Max	
Temperatura	0-6hs	12-15hs	58%	34%	6%	1%	NA
Nivel	0-23hs	0-23hs	100%	NA	NA	NA	Censurado
Humedad relativa	11-16hs	23-4hs	40%	43%	12%	3%	NA
Velocidad del viento	22-7hs (300,900) 22-8hs (700) 8-13hs (1100)	2-13hs (300) 12-16hs (700) 1-13hs (900) 12-18hs (1100)	9%	19%	NA	64%	Intensidad Beaufort
Dirección del viento	NA	NA	100%	NA	NA	NA	Orientación media
Radiación solar	0-7hs	12-13hs	14%	22%	15%	29%	NA
Presión atmosférica	21-4hs	9-13hs	50%	43%	NA	6%	NA

Nota: NA indica que No Aplica

2.4.2.2 Precipitación

En el caso de precipitación, como se mencionó anteriormente el tratamiento es distinto. La Fig. 11 muestra el diagrama de la fase preprocesamiento de datos para el caso particular de la variable precipitación.

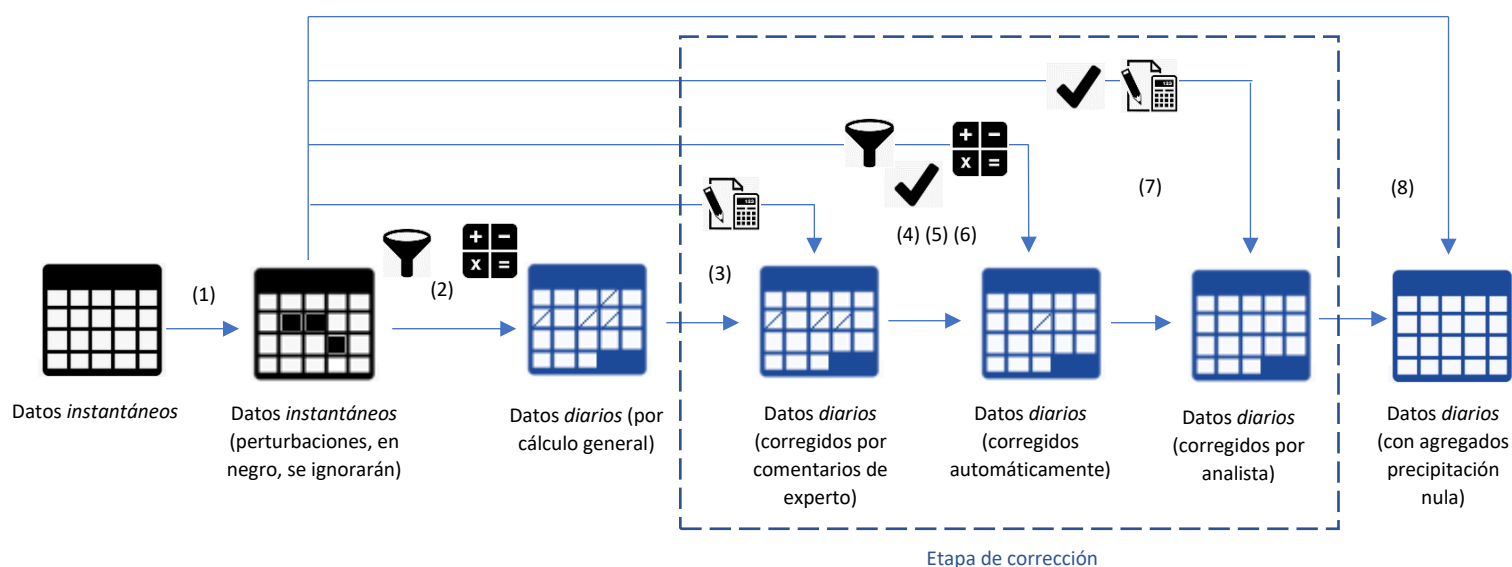


Fig. 11 Diagrama del flujo de preprocesamiento de datos de precipitación. Referencia de íconos: papel y lápiz corresponde a tarea que requiere intervención manual, cuadrante de operaciones indica cálculo automático y embudo indica operación de filtrado. (1) Detección de perturbaciones. (2) Cálculo. (3) Corrección por comentario del experto. (4) Corrección por reinicio del acumulador. (5) Corrección por atribución de precipitación. (6) Corrección por repartición de precipitación. (7) Corrección por valores extremos. (8) Imputación. Fuente: elaboración propia

2.4.2.2.1 Tratamiento de inconsistencias

(1) Detección de perturbaciones (considerando la misma definición del apartado anterior):

- a. Descenso abrupto en la serie temporal: sean A, B, C tres registros de precipitación, donde existe precedencia en el orden de las precipitaciones: A primero, luego B y luego C, y además B es menor que A, y B es menor que C, y A es menor o igual que C. Asimismo, A se produce un día antes que B y las fechas de B y C coinciden; o C se produce un día después que B, y las fechas de A y B coinciden; o la precipitación B y C coinciden y las fechas son todas distintas.
- b. Caída aislada a 0 dentro del mismo día (representa un falso reinicio del acumulador en los términos en que se definió en la sección 2.3.2): sean A, B, C tres registros de precipitación, donde existe precedencia en el orden de las precipitaciones: primero A, segundo B y tercero C, y además B es menor que A, y B es menor que C, y A es menor o igual que C. Asimismo, las fechas de A, B y C coinciden (sucieron el mismo día), y B es igual a cero.
- c. Incremento aislado en la serie temporal: sean A, B, C tres registros de precipitación, donde existe precedencia en el orden de las precipitaciones: A primero, luego B y luego C, y además B es mayor que A, y B es mayor que C, y A es menor o igual que C. Asimismo, A se produce un día antes que B y las fechas de B y C coinciden; o C se produce un día después que B, y las fechas de A y B coinciden; o la precipitación B y C coinciden y las fechas son todas distintas.

Cabe aclarar que en este caso la noción de valor duplicado no aplica.

2.4.2.2.2 Agregación

- (2) Cálculo: se procedió al cálculo de precipitación diaria, en general, como diferencia entre el primer y el último registro que se disponga por día, en todos los casos en que dicho registro no corresponda a una perturbación, tal cual se definió en el punto anterior. En el caso particular de un único registro diario (ver sección 2.3.2), atribuible a una falla del sensor en la transmisión del dato, la diferencia se resuelve cero⁶.

2.4.2.2.3 Rectificación

Esta etapa, que no está presente en el procedimiento general, corresponde a la corrección o modificación de valores calculados en base a premisas.

- (3) Corrección por comentario del experto. En este sentido caben dos posibilidades, según el comentario en el registro instantáneo (el valor de precipitación acumulada reportado por el

⁶ El mismo registro es a la vez el primero y el último por ser único.

sensor) afecte el cálculo de días adyacentes o no: en un caso simplemente se corrige el cálculo del día puntual y en el otro caso además se corrigen los días afectados, en los dos casos indicando en el campo de observaciones el tipo de corrección efectuada.

- (4) Corrección por reinicio del acumulador. Denominada por diferencia negativa: sean A y B dos precipitaciones consecutivas que se producen el mismo día y en ese orden, si la diferencia de B y A es negativa, y siendo que el primer ni el último registro diario corresponden a perturbaciones, y la diferencia entre el último y el primero es menor que cero o bien B es cero, y siendo por último que ese día no fue objeto de corrección manual identificada como perturbación (segundo caso en (3)), entonces se procede al cálculo por partes. Es decir, por un lado, se calcula la diferencia entre A y el primer registro diario, y a eso se suma la diferencia entre el último registro diario y B.
- (5) Corrección por atribución de precipitación. Una particularidad que puede darse es que se haya registrado precipitación inter diaria, por la madrugada. Vale decir, cuando el último registro de un día, no coincide con el primero del día siguiente, se asume que llovió entre medio. Debido al principio de funcionamiento del sensor, si la diferencia entre uno y otro de los registros mencionados es de un milímetro y además el tiempo transcurrido entre los dos es igual o inferior a las doce horas (como se indicó anteriormente el sensor debe transmitir al menos cada doce horas), entonces se está en condiciones normales de operación y el milímetro llovido corresponde al segundo día⁷. Debe cumplirse además que ninguno de los registros intervinientes sea considerado una perturbación ni tampoco se haya hecho una corrección manual por perturbación para el día en cuestión, según comentario. El milímetro se adiciona siempre al valor calculado para el día en cuestión.
- (6) Corrección por repartición de precipitación. En caso que no se cumpla alguna de las dos condiciones principales mencionadas, se procede al cálculo por repartición. Es decir, en situación donde la diferencia de precipitaciones es mayor a un milímetro o bien han transcurrido más de doce horas, se asume mal funcionamiento del sensor o pérdida de información. En estas circunstancias no es posible determinar inequívocamente a qué día corresponde sumar la diferencia, por lo que se reparte entre el primer y el segundo día. Nuevamente, los registros intervinientes no deben estar considerados perturbaciones ni tampoco ninguno de los días haber sido objeto de cálculo manual por perturbación atribuible a un comentario. Asimismo, se tiene en cuenta que el registro siguiente al primero del segundo día contenga una precipitación distinta que la del último registro del primer día (en tal caso se estaría ante una irregularidad, vale decir perturbación, de la serie y se ignora).
- (7) Corrección por valores extremos
 - a. Cuando sean negativos, se los considera inválidos y se realiza la corrección manual. Estos casos constituyen situaciones excepcionales que no fueron contempladas en las reglas precedentes.

⁷ Se debe tener presente que el sensor registra lluvia en el momento en que sucede, por lo tanto, la variación es atribuible, siempre bajo las condiciones mencionadas, *a posteriori* desde el punto de vista del límite diario.

- b. Cuando superen los cien milímetros⁸, puede que el cálculo sea correcto y en tal caso no se interviene o bien puede suceder que, como en el punto anterior, sea producto de un caso no previsto y se corrija manualmente.

2.4.2.2.4 Imputación

- (8) Rellenado de días sin lluvia. Un párrafo aparte corresponde a la revisión de los días con valores faltantes. En aquellos casos en que, el último registro de precipitación previo al/los día/días faltante/s, coincide con el correspondiente al primer registro siguiente, entonces significa que no se registró lluvia en ese período (siempre y cuando no exista un comentario que indique lo contrario), y por ende corresponde imputar cero.

A tal efecto, se realizó un análisis por período de días faltantes consecutivos respecto de los períodos de seca observados para verificar consistencia. En la mayor parte de los casos la imputación fue factible. Se remite al lector al trabajo de Andreoni et al. (2018) para tener el detalle de este procedimiento y profundizar sobre aspectos adicionales relativos al procesamiento de precipitación. La Tabla 11, muestra los resultados acordes a las fechas procesadas (el documento referenciado anteriormente abarca hasta diciembre de 2015). Se indica como *patrón*, el resultado del análisis de los meses completos sin dato, donde la mayoría de las estaciones (todas a excepción de la estación 3051, la que cuenta con menor cantidad de datos) presentan al menos un mes sin dato entre mayo y septiembre, donde los meses con mayor frecuencia corresponden a la temporada invernal, y en menor medida fin del otoño (distribución normal estadísticamente significativa y media próxima al 5% de meses faltantes).

⁸ A modo heurístico se fija este umbral de revisión, considerándose “muy poco factibles” los casos en que se tengan más de cien milímetros de precipitación diaria.

Tabla 11 Cantidad de días sin datos, frecuencia de aparición y determinación respecto de la presunción de no lluvia

N.º de días sin dato	Frecuencia	Determinación
1	1129	Aplica, se observa en la mayoría de las estaciones para todos los meses
2	325	
3	94	
4	44	
5	28	
6	13	
7	5	Aplica según <i>patrón</i> con extensión hasta los meses de abril y octubre
8	6	
9	4	
10	6	
11	4	
13	3	
14	4	Aplica según <i>patrón</i>
15	2	
17	1	
21	1	
22	1	
23	1	
27	1	Aplica según <i>patrón</i> , corresponde a la temporada invernal
30	1	
31	8	Aplica en 7 de los casos, en el restante no existe evidencia en la temporada estival
33	1	No aplica, corresponde a la temporada estival y no se tiene evidencia empírica que lo sustente
61	1	Aplica, corresponde mayormente a la temporada invernal y existe evidencia en la mayoría de las estaciones
66	1	
89	1	No aplica, no existe evidencia empírica
92	1	
356	1	

2.4.2.3 Análisis de valores extremos y anomalías de las series de tiempo diarias

Como paso final en el proceso de obtención de las series meteorológicas e hidrológicas por estación, se realizó un análisis de irregularidades de las series diarias resultantes de los procesos descriptos en las secciones anteriores. El mismo estuvo centrado en los tres aspectos que se mencionan a continuación:

1. Identificar anomalías en secciones de las series por comparación entre las estaciones que miden la misma variable. Los períodos irregulares fueron filtrados y en ciertos casos se desestimó la serie completa de una estación cuando por distintas razones no era posible conciliarla con las demás series de la misma variable.
2. Aislar valores atípicos puntuales en el contexto de las series diarias (anteriormente se hizo en el contexto de las series instantáneas), por medio del filtrado basado en límites para la magnitud de la variable. Tales valores fueron removidos de la serie.
3. Reconocer cambios en el comportamiento de las series diarias que obedecen a una razón instrumental de efecto permanente, determinadas por la sensibilidad del sensor. Se reemplazaron los valores que superen el límite de censura por el mismo límite.

La Tabla 12 muestra el detalle de los casos hallados para cada variable.

Tabla 12 Detalle del resultado del análisis de anomalías y valores extremos sobre las series diarias de variables meteorológicas e hidrológicas

Variable	Estación	Anomalías de sección	Valores atípicos	Corrección instrumental
Temperatura	1100	<i>Incremento espurio:</i> desde 2017-03-01 hasta 2017-12-31 <i>Descenso espurio:</i> desde 2014-01-01 hasta 2016-12-31	NA	NA
	400	<i>Incremento espurio:</i> desde 2013-06-01 a 2013-09-30 <i>Descenso espurio:</i> desde 2014-08-01		
	300	<i>Descenso espurio:</i> desde 2006-06-01		
	900	<i>Descenso espurio:</i> desde 2004-04-01 hasta 2005-10-31		
Radiación Solar	900	<i>Descenso espurio:</i> Desde 2000-09-01 hasta 2007-03-31	>=0.8ly	NA
	300	<i>Incremento espurio:</i> Desde 2011-11-24		
	1100, 700	<i>Desestimadas por irregularidades generales</i>	NA	
Humedad Relativa	300, 1200	NA	NA	Sensibilidad límite 97% (valores mayores se cambian por 97%)
	400	<i>Desestimada por presentar múltiples niveles máximos</i>	NA	NA
	700, 900, 1100	<i>Desestimadas, presentan irregularidades generales</i>	NA	
Presión Atmosférica	1100	NA	<= 930hPA y >= 970hPA	NA
	700, 300, 900	<i>Desestimadas, presentan irregularidades generales</i>	NA	
Velocidad del Viento	300, 900	NA	>= 60k/h	NA
	700, 1100	<i>Desestimadas, presentan un régimen de vientos marcadamente distinto que las anteriores</i>	NA	
Dirección del Viento	700, 1100	<i>Desestimadas, presentan un régimen de vientos marcadamente distinto que las anteriores</i>	NA	NA
Nivel de Lago	NA	NA	NA	NA
Nivel de Ríos	NA	NA	NA	NA
Precipitación	NA	NA	NA	NA

Nota: NA indica que no se aplicó

Cabe aclarar que las series de nivel del lago en 2016 presentan discrepancias con la serie diaria de relevamiento manual a cargo de la provincia de Córdoba. Sin embargo, no existe evidencia de mal funcionamiento de los sensores de acuerdo a la metodología desarrollada en el presente trabajo (ni comentarios de experto ni irregularidades, anomalías o valores atípicos) por lo que los datos de sensores de este período se deciden emplear.

2.4.3 Integración de datos

Una vez realizados todos los procedimientos de depuración, estructuración, cálculo y análisis de las series resultantes sobre los datos de las tres bases: lago, ríos, e hidrológicos y

meteorológicos por separado, en esta sección se define la arquitectura del *dataset* o conjunto de datos que se empleará en adelante para la aplicación de modelos de series temporales. Este *dataset* consiste, en primera instancia, de la unión (correspondencia, combinación) a nivel registro entre las tres bases de datos que se mencionaron con anterioridad: base de datos del lago, base de datos de los ríos y base de datos hidrológicos y meteorológicos. Esta integración, se realiza en dos dimensiones: temporal y espacial, como se ilustra en el diagrama de la Fig. 12. El resultante, es por ende una tabla que tiene todas las variables seleccionadas de cada una de las tres bases de datos, constituidas en series de tiempo.

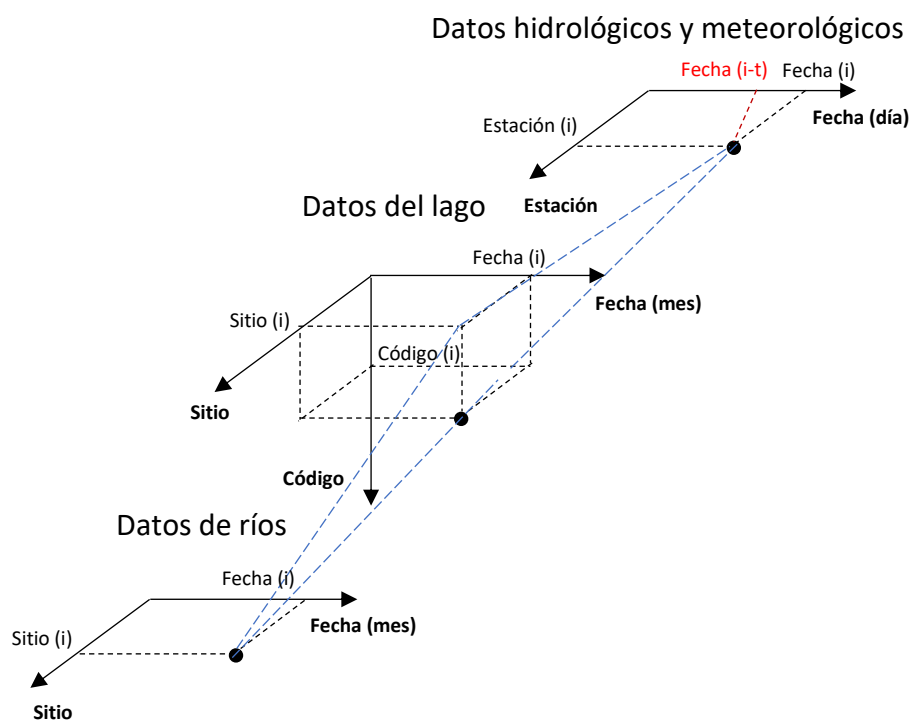


Fig. 12 Diagrama de integración temporal y espacial de los datos de lago, ríos e hidrológicos y meteorológicos. Fuente: elaboración propia

Retornando al origen, el objetivo de este trabajo consiste en determinar la incidencia de las condiciones meteorológicas e hidrológicas sobre la calidad del agua en el embalse, por ende, el eje central de las series serán los datos del lago. En este sentido, se opta por tener una serie por cada sitio y perfil con frecuencia mensual, conteniendo las variables del lago junto con las correspondientes a ríos y estaciones meteorológicas. La selección de los sitios y códigos se realizó considerando la menor cantidad de datos faltantes en término de monitoreos.

La Fig. 13 representa el calendario de cada una de las series seleccionadas, indicando la cantidad de monitoreos realizados en cada mes desde septiembre de 1999 a agosto de 2017. Las demás sitios-perfiles presentan una cantidad mayor de faltantes y se prefiere no analizarlos debido a la escasa información que aportarían. A continuación, la lista de las 10 series seleccionadas:

- a) Serie Centro-subsuperficial (en adelante C1)
- b) Serie Centro-fótica (en adelante C2)
- c) Serie Centro-intermedia/hipolimnion (en adelante C4)
- d) Serie Centro-fondo (en adelante C5)
- e) Serie TAC-subsuperficial (en adelante TAC1)
- f) Serie TAC-fótica (en adelante TAC2)
- g) Serie TAC-toma (en adelante TAC4)
- h) Serie TAC-fondo (en adelante TAC5)

- i) Serie DCQ-subsuperficial (en adelante DCQ1)
- j) Serie DSA-subsuperficial (en adelante DSA1)

C1 (10%)

AÑO/MES	1	2	3	4	5	6	7	8	9	10	11	12
1999									1	1	3	2
2000	5	4	4	2	2	2	1	3	2	2	1	2
2001	2	0	0	1	1	1	1	1	1	3	1	
2002	1	0	1	0	1	1	1	1	1	1	1	1
2003	1	1	1	1	1	1	1	1	0	1	1	1
2004	1	1	1	1	1	0	1	1	1	1	1	1
2005	1	1	1	1	1	1	1	1	1	1	1	1
2006	1	1	1	0	1	0	1	1	0	1	1	1
2007	1	1	1	1	1	1	1	1	1	1	0	2
2008	1	1	1	1	1	1	1	1	1	1	1	1
2009	1	1	1	1	1	1	0	1	1	1	1	1
2010	1	1	1	1	1	1	0	1	1	1	1	1
2011	1	1	1	1	1	0	1	1	1	1	1	1
2012	1	0	1	1	1	0	1	1	1	1	1	0
2013	1	1	0	1	1	1	1	1	1	1	1	1
2014	1	1	1	1	1	1	1	1	1	1	1	1
2015	1	0	1	1	1	1	0	1	1	1	1	1
2016	1	1	1	1	1	1	1	1	0	1	1	1
2017	1	1	1	0	1	1	1	1				

C5 (12%)

AÑO/MES	1	2	3	4	5	6	7	8	9	10	11	12
1999									1	1	2	2
2000	5	4	4	2	2	2	1	3	2	2	1	2
2001	2	0	0	1	1	1	1	0	1	3	1	
2002	1	0	1	0	1	1	1	1	1	0	1	1
2003	1	1	1	1	1	1	1	1	0	1	1	1
2004	1	1	1	1	1	0	1	1	1	1	1	1
2005	1	1	1	1	1	1	1	1	1	1	1	1
2006	1	1	1	0	1	0	1	1	0	1	1	1
2007	1	0	1	1	1	1	1	1	1	1	0	2
2008	1	1	1	1	1	1	1	1	1	1	1	1
2009	1	1	1	1	1	1	0	1	1	1	1	1
2010	1	1	1	1	1	1	0	1	1	1	1	1
2011	1	1	1	1	1	0	1	1	1	1	1	1
2012	1	0	1	1	1	0	1	1	1	1	1	0
2013	1	1	0	1	1	1	1	1	1	1	1	1
2014	1	1	1	1	1	1	1	1	1	1	1	1
2015	1	0	1	1	1	1	0	0	1	1	1	1
2016	1	1	1	1	1	1	1	1	0	1	1	1
2017	1	1	1	0	1	1	1	1				

TAC4 (14%)

AÑO/MES	1	2	3	4	5	6	7	8	9	10	11	12
1999									0	0	0	0
2000	0	0	0	0	0	1	1	3	2	2	1	2
2001	2	0	0	1	1	1	1	0	0	1	1	1
2002	1	0	1	0	1	1	1	1	1	1	1	1
2003	1	1	1	1	1	1	1	1	0	1	1	1
2004	1	1	1	1	1	0	1	1	1	1	1	1
2005	1	1	1	1	1	1	1	1	1	1	1	1
2006	1	1	1	0	1	1	1	0	0	1	1	1
2007	1	1	1	1	1	1	1	1	1	1	0	2
2008	1	1	1	1	1	1	1	1	1	1	1	1
2009	1	1	1	1	1	1	0	1	1	1	1	1
2010	1	1	1	1	1	1	1	1	1	1	1	1
2011	1	1	1	1	1	0	1	1	1	1	1	1
2012	1	0	1	1	1	1	0	1	1	1	1	0
2013	0	1	0	1	1	1	1	1	1	1	1	1
2014	1	1	1	1	1	1	1	1	1	1	1	1
2015	1	0	1	1	1	1	0	1	1	1	1	1
2016	1	1	1	1	1	1	1	1	0	1	1	1
2017	1	1	1	0	1	1	1	1				

C2 (16%)

AÑO/MES	1	2	3	4	5	6	7	8	9	10	11	12
1999									6		2	
2000	4	4	4	2	2	2	1	3	2	2	1	2
2001	2	0	0	1	1	1	1	0	0	1	0	
2002	0	0	0	0	1	1	0	0	0	1	1	0
2003	1	1	1	1	1	1	1	1	0	1	1	1
2004	1	1	1	1	1	0	1	1	1	1	1	1
2005	1	1	1	1	1	1	1	1	1	1	1	1
2006	1	1	1	0	0	0	0	1	0	1	1	1
2007	1	0	1	1	1	1	1	1	1	1	0	2
2008	1	1	1	1	1	1	1	1	1	1	1	1
2009	1	1	1	1	1	1	0	1	1	1	1	1
2010	1	1	1	1	1	1	0	1	1	1	1	1
2011	1	1	1	1	1	0	1	1	1	1	1	1
2012	1	0	1	1	1	0	1	1	1	1	1	0
2013	1	1	0	1	1	1	1	1	1	1	1	1
2014	1	1	1	1	1	1	1	1	1	1	1	1
2015	1	0	1	1	1	1	0	0	1	1	1	1
2016	1	1	1	1	1	1	1	1	0	1	1	1
2017	1	1	1	0	1	1	1	1				

TAC1 (5%)

AÑO/MES	1	2	3	4	5	6	7	8	9	10	11	12
1999									1	1	3	2
2000	4	4	4	2	2	2	1	3	2	2	1	2
2001	2	0	0	1	1	1	1	1	1	1	1	1
2002	2	0	1	0	1	1	1	1	1	1	1	1
2003	1	1	1	1	1	1	1	1	1	1	1	1
2004	1	1	1	1	1	1	1	1	1	1	1	1
2005	1	1	1	1	1	1	1	1	1	1	1	1
2006	1	1	1	0	1	1	1	1	1	1	1	1
2007	1	1	1	1	1	1	1	1	1	1	0	2
2008	1	1	1	1	1	1	1	1	1	1	1	1
2009	1	1	1	1	1	1	0	1	1	1	1	1
2010	1	1	1	1	1	1	1	1	1	1	1	1
2011	1	1	1	1	1	0	1	1	1	1	1	1
2012	1	0	1	1	1	1	1	1	1	1	1	0
2013	1	1	1	1	1	1	1	1	1	1	1	1
2014	1	1	1	1	1	1	1	1	1	1	1	1
2015	1	1	1	1	1	1	1	1	1	1	1	1
2016	1	1	1	1	1	1	1	1	1	1	1	1
2017	1	1	1	0	1	1	1	1				

TAC5 (11%)

AÑO/MES	1	2	3	4	5	6	7	8	9	10	11	12
1999									1		3	2
2000	4	4	4	2	2	2	1	3	2	2	1	2
2001	2	0	0	1	1	1	1	0	0	1	1	1
2002	2	0	1	0	1	1	1	1	1	1	1	1
2003	2	1	1	1	1	1	1	1	0	1	1	1
2004	1	1	1	1	1	0	1	0	1	1	1	1
2005	1	1	1	1	1	1	1	1	1	1	1	1
2006	1	1	1	0	1	1	1	1	0	1	1	1
2007	1	1	1	1	1	1	1	1	1	1	0	2
2008	1	1	1	1	1	1	1	1	1	1	1	1
2009	1	1	1	1	1	1	0	1	1	1	1	1
2010	1	1	1	1	1	1	1	1	1	1	1	1
2011	1	1	1	1	1	0	1	1	1	1	1	1
2012	1	0	1	1	1	0	1	1	1	1	1	0
2013	0	1	0	1	1	1	1	1	1	1	1	1
2014	1	1	1	1	1	1	1	1	1	1	1	1
2015	1	0	1	1	1	1	0	1	1	1	1	1
2016	1	1	1	1	1	1	1	1	0	1	1	1
2017	1	1	1	0	1	1	1	1				

DSA1 (7%)

AÑO/MES	1	2	3	4	5	6	7	8	9	10	11	12
1999									1	1	2	2
2000	4	4	4	2	2	2	1	3	2	2	1	2
2001	2	0	0	1	1	1	1	1	1	1	1	1
2002	1	0	1	0	1	1	1	1	1	1	1	1
2003	1	1	1	1	1	1	1	1	1	1	1	1
2004	1	1	1	1	1	1	1	1	1	1	1	1
2005	1	1	1	1	1	1	1	1	1	1	1	1
2006	1	1	1	0	1	1	1	1	0	1	1	1
2007	1	0	1	1	1	1	1	1	1	1	0	2
2008	1	0	1	1	1	1	1	1	1	1	1	1
2009	0	1	1	1	1	1	0	1	1	1	1	1
2010	1	1	1	1	1	1	0	1	1	1	1	1
2011	1	1	1	1	1	0	1	1	1	1	1	1
2012	1	0	1	1	1	1	1	1	1	1	1	0
2013	1	1	1	1	1	1	1	1	1	1	1	1
2014	1	1	1	1	1	1	1	1	1	1	1	1
2015	1	1	1	1	1	1	1	1	0	1	1	1
2016	1	1	1	1	1	1	1	1	1	1	1	1
2017	1	1	1	1	1	1	1	1				

C4 (22%)

AÑO/MES	1	2	3	4	5	6	7	8	9	10	11	12
1999									0	0	0	0
2000	0	0	0	0	0	0	0	0	1	1	1	1
2001	1	0	0	0	1	1	1	0	0	1	1	1
2002	1	0	1	0	1	1	1	1	1	1	1	1
2003	1	1	1	1	1	1	1	1	0	1	1	1
2004	1	1	1	1	1	0	1	1	1	1	1	1
2005	1	1	1	1	1	1	1	1	1	1	1	1
2006	1	1	1	0	0	0	0	0	1	0	1	1
2007	1	0	0	1	1	0	0	1	1	1	0	2
2008	1	1	1	1	1	1	1	1	1	1	1	1
2009	1	1	1	1	1	1	0	1	0	1	1	0
2010	1	1	1	1	1	0	0	1	1	1	1	1
2011	1	1	1	1	1	0	1	1	1	1	1	1
2012	1	0	1	1	1	1	0	1	1	1	1	0
2013	1	1	0	1	1	1	1	1	1	1	1	1
2014	0	1	1	1	1	1	1	1	1	1	1	1
2015	1	0	1	1	1	1	0	0	1	1	0	1
2016	1	1	1	1	1	1	1	1	0	1	1	1
2017	1	1	1	0	1	1	1	1				

TAC2 (13%)

AÑO/MES	1	2	3	4	5	6	7	8	9	10	11	12
1999									1	1	2	2
2000	4	4	4	2	2	2	1	3	2	2	1	2
2001	2	0	0	0	1	1	1	0	1	1	1	0
2002	0	0	0	0	0	1	0	0	1	1	1	0
2003	1	1	1	1	1	1	1	0	1	1	1	1
2004	1	1	1	1	1	0	1	1	1	1	1	1
2005	1	1	1	1	1	1	1	1	1	1	1	1
2006	1	1	1	0	1	1	1	1	0	1	1	1
2007	1	1	1	1	1	1	1	1	1	1	0	2
2008	1	1	1	1	1	1	1	1	1	1	1	1
2009	1	1	1	1	1	1	0	1	1	1	1	1
2010	1	1	1	1	1	1	1	1	1	1	1	1
2011	1	1	1	1	1	0	1	1	1	1	1	1
2012	1	0	1	1	1	1	0	1	1	1	1	0
2013	0	1	0	1	1	1	1	1	1	1	1	1
2014	1	1	1	1	1	1	1	1	1	1	1	1
2015	1	0	1	1	1	1	0	1	1	1	1	1
2016	1	1	1	1	1	1	1	1	0	1	1	1
2017	1	1	1	0	1	1	1	1				

DCQ1 (12%)

AÑO/MES	1	2	3	4	5	6	7	8	9	10	11	12
1999									1	1	2	2
2000	5	4	4	2	2	2	1	3	2	2	1	2
2001	2	0	0	1	1	1	1	1	1	1	1	1
2002	1	0	1	0	1	1	1	1	0	1	1	1
2003	1	1	1	1	1	1	1	1	1	1	1	1
2004	1	1	1	1	1	1	1	1	1	1	1	1
2005	1	1	1	1	1	1	1	1	1	1	1	1
2006	1	1	1	0	0	0	0	1	0	1	1	1
2007	1	1	1	1	1	1	1	1	1	0	0	2
2008	1	1	1	1	1	1	1	1	1	1	1	1
2009	1	1	1	1	1	1	0	1	1	1	1	1
2010	1	1	1	1	1	1	0	1	1	1	1	1
2011	1	1	1	0	1	0	1	1	1	1	1	0
2012	1	0	1	1	1	1	1	1	1	1	1	0
2013	1	1	1	1	1	1	1	0	1	1	1	1
2014	1	1	1	1	1	1	0	1	1	1	1	1
2015	1	1	1	1	1	0	1	0	0	1	1	1
2016	1	1	1	1	0	1	1	1	1	1	1	1
2017	1	1	1	0	1	1	1	0				

Referencia de colores:

- naranja pálido: representa falta de monitoreo en general
- naranja: representa falta de monitoreo para el sitio
- marrón representa falta de monitoreo para el perfil

Fig. 13 Calendario de monitoreo de cada sitio y perfil seleccionado (entre paréntesis porcentaje de monitoreos faltantes)

Teniendo en cuenta el marco de referencia de cada serie, a continuación, se detallan las pautas de integración en términos de las dimensiones mencionadas más arriba aplicadas a cada una de las ellas.

2.4.3.1 Integración temporal

En primera instancia el objetivo es conformar series de tiempo a partir de los datos provenientes de las tres fuentes, todos los cuales tienen fecha; por lo que esta dimensión constituye el eje central del proceso. En lo que se refiere al nivel de agregación, los datos de lago y ríos se encuentran a nivel mensual, mientras que los datos hidrológicos y meteorológicos procesados se encuentran a nivel diario.

Como ya se comentó en la introducción de esta sección, los datos del lago ocupan el lugar central, y las series se estructurarán en base a las fechas de los monitoreos correspondientes. Como se pudo observar en la Fig. 13, para ciertos meses hubo más de un monitoreo mensual y a la vez para algunos meses no hubo ninguno en absoluto. Para resolver esta situación, teniendo en cuenta que se busca una serie mensual de valores, la estrategia fue la siguiente:

1. Definición de fechas de referencia. Teniendo en cuenta que los monitoreos se realizan todos los últimos martes de cada mes, salvo excepciones (clima muy severo o adverso, o festividades como Navidad o Año Nuevo), es posible establecer para cada mes una fecha de referencia de monitoreo. Representa *la fecha en que el monitoreo se hubiera realizado siguiendo la regla acordada*.
2. Composición de la serie temporal de datos del lago (para cada serie). Para cada mes, se aplica el siguiente procedimiento:
 - a. Si existe una fecha de monitoreo para ese mes, sitio y perfil, entonces el mes estará asociado a esa fecha. Si existe más de una fecha de monitoreo para ese mes, sitio y perfil, entonces se calcula la diferencia en días a la fecha de referencia, y se elige la fecha que tenga la menor diferencia (en caso de haber dos diferencias iguales, se elige la fecha mayor).
 - b. Si no existe una fecha de monitoreo para ese mes, sitio y perfil, pero sí para el mes y el sitio, es decir, se ha realizado el monitoreo en al menos alguno de los otros perfiles, entonces se asocia al mes en cuestión esta fecha. En caso que exista más de una fecha que cumpla la condición, se aplica el procedimiento descrito en (a).
 - c. Si no existe una fecha de monitoreo para ese mes, sitio y perfil, pero sí para el mes (se ha realizado el monitoreo en al menos alguno de los otros sitios), entonces se asocia al mes en cuestión esta fecha. En caso que exista más de una fecha que cumpla la condición, se aplica el procedimiento descrito en (a).
 - d. Si no existe fecha alguna de monitoreo para el mes, entonces se toma como fecha para ese mes la fecha de referencia en (1). Esta pauta tiende a preservar la equidistancia de las observaciones mensuales.
3. Asociación de todas las variables medidas en el lago para las fechas del punto (a), vale decir variables *in-situ* y de laboratorio.

La necesidad de este procedimiento es contar con un calendario completo para cada una de las series. Es claro a partir de lo indicado en el punto (3), que algunas fechas no tendrán valor en

ninguna de las variables de lago, pero permitirán actuar de nexo para integrar datos de las otras dos bases.

Cabe mencionar que, al momento de conformar las series del lago, se completaron los casos de las variables a nivel general y por sitio (ver Tabla 4), en base a los valores presentes en los demás sitios o en los demás códigos respectivamente utilizando la moda. Por ejemplo, si un cierto mes no se realizó el monitoreo para TAC5, pero sí se lo hizo para el resto de los perfiles en el sitio TAC, entonces por la transformación en 2.4.1, el valor correspondiente de Secchi, imputado por convención en TAC1, se trasladó también a los demás puntos de TAC, pero no a TAC5, porque para TAC5 no se tiene registro en la base para esa fecha. Ahora bien, al momento de conformar las series de tiempo, TAC5 tendrá para el mes en cuestión muchas de las variables sin dato, pero correspondería que Secchi, además de Cota, tenga el valor correspondiente a ese monitoreo, presente en los demás perfiles de TAC. En general, los valores de estas dos variables para todos los perfiles de un mismo sitio son iguales, excepto que se hubieran muestreado en fechas distintas del mismo mes. Previendo tal caso se aplica la moda, es decir que se considera el valor más frecuente entre todos los códigos del sitio en cuestión.

La correspondencia temporal de datos de lagos y ríos es teóricamente directa, es uno a uno, es decir, a un registro de lago (resultado de un monitoreo particular), le corresponde uno y solo un registro de río (resultado del correspondiente monitoreo del río llevado a cabo por otro equipo de trabajo) cada mes. En este sentido, se presentan varias situaciones particulares a considerar. Por un lado, pese a que se realiza un único monitoreo mensual, las fechas de lago y ríos pueden ser distintas dentro de un mismo mes, vale decir que primero se haya realizado la visita al lago y luego a los ríos o viceversa. Incluso ha sucedido que el muestreo de ríos se ha repartido en dos días (un día se monitorean algunos y al día siguiente los restantes), y lo mismo ha sucedido con el lago (un día ciertos sitios y otro día los restantes). Por último, así como ocurre en el caso del lago, también ha habido meses en que se ha realizado más de un monitoreo para un determinado mes. A efectos de establecer la correspondencia temporal, entre los datos de lago y los datos de ríos, atento a estas particularidades, se considera que, dos registros, uno correspondiente al lago y otro al río se integran, siguiendo la estrategia a continuación, para cada mes y cada uno de los puntos donde se muestrean los ríos principales (RSA, RCQ):

1. Composición de la serie temporal de datos del río:
 - a. Si existe una fecha de monitoreo para el mes y el sitio, se selecciona esta fecha para asociarla al mes en cuestión. Al igual que para datos del lago, si existe más de una fecha, entonces se considera la diferencia mínima a la fecha correspondiente al mismo mes y año de la serie de datos del lago descripta anteriormente, ésta será la fecha de referencia (en caso de diferencias iguales, se selecciona la mayor fecha). De esta manera queda hecha la asociación de fechas entre lago y río, porque para un mes en particular, ya ha sido elegida la fecha de nexo del lago, y a ese mismo mes se ha asociado una fecha de río (en su mayoría las fechas son coincidentes).
 - b. En caso que no exista una fecha para el mes y el sitio, se considera la fecha de referencia del lago.
2. Asociación de las variables medidas en el río para las fechas del punto (a).
3. Integración de la serie de ríos con la serie del lago, por coincidencia de la fecha del lago en la serie de datos del lago, con la fecha de referencia del lago en la serie de datos del río.

En lo que se refiere a los datos de las estaciones meteorológicas, al presentar los datos un nivel de granularidad mayor (día en vez de mes), se requiere un paso previo de agregación. Es decir, los datos hidrológicos y meteorológicos se precisan resumir, de manera tal de tener un único valor por mes, para así integrarlo con la serie mensual del lago. La estrategia considerada fue la siguiente:

1. Agregación de los de n días previos, siendo:
 - $n=1$: es el caso trivial, representa el pasado reciente, y de acción casi inmediata, son las condiciones meteorológicas del día anterior.
 - $n=3$: es el efecto de la tendencia meteorológica a muy corto plazo.
 - $n=7$: es el efecto de la tendencia meteorológica a corto plazo.
 - $n=14$: es el efecto de la tendencia meteorológica a mediano plazo.
 - $n=21$: es el efecto de la tendencia meteorológica a más largo plazo.

La función de agregación empleada es la suma para la precipitación (que siempre debe calcularse acumulada) y el promedio para el resto de las variables.

Cada uno de estos valores agregados, se calcula para cada una de las variables hidrológicas y meteorológicas, tomando como fecha base para cada mes, la fecha de la serie de datos del lago. Así, por ejemplo, si el monitoreo del mes de septiembre de 2006 para C1 se realizó el día 26 (o así se fijó en caso de ausencia), entonces el promedio a muy corto plazo de temperatura, considera los días 25, 24 y 23.

2. Integración de la serie de datos hidrológicos y meteorológicos con la serie de datos del lago por coincidencia de fechas.

2.4.3.2 Integración espacial

Simultáneamente a la integración de los datos a nivel temporal es preciso asociarlos a nivel geográfico. La lógica detrás de esta estrategia recae en el supuesto de que, dado un punto cualquiera del lago considerando la dimensión horizontal por razones de simplicidad (no así la profundidad⁹), las condiciones que reflejan las variables hidrológicas y meteorológicas, y de los ríos, tendrán más relevancia para el sitio en cuestión en la medida en que más se aproximen a él, y menos con respecto a terceros puntos en la medida en que más alejados se encuentren de ellas. Por ejemplo, refiriéndose al centro del lago, seguramente la relevancia de la temperatura que pueda medirse en la estación 1100 (ubicada a aproximadamente 2,6km lineales) tendrá más relevancia que la medida en la estación 1800 (ubicada aproximadamente a 22,8km lineales).

En base a lo dicho anteriormente, se procede a la asociación de los puntos geográficos en que se encuentran estaciones meteorológicas, con los sitios del lago. La estrategia seguida en este caso se presenta a continuación, y se basa en el promedio ponderado para la determinación de los efectos:

1. Cálculo de la distancia entre el sitio del lago y la estación meteorológica. Para la conversión de coordenadas polares a coordenadas cartesianas se utilizó como radio de

⁹ Es despreciable la profundidad en relación a las distancias entre puntos sitio-perfil del lago y puntos de monitoreo de los ríos o las estaciones meteorológicas.

la tierra 6378137 m. La Fig. 14 muestra un diagrama de las distancias entre la estación y el punto de muestreo.

2. Cálculo del peso en base a la inversa de la distancia, según la Ecuación (1).

$$P_e = \frac{1/d_e}{\sum_{i=1}^n 1/d_i} \quad (1)$$

Donde:

P_e es el peso que corresponde a la estación meteorológica e

d_e es la distancia desde la estación meteorológica e al sitio para el cual se está calculando la variable

n es la cantidad de estaciones que intervienen en el cálculo

3. Cálculo de la variable meteorológica para el sitio del lago que se trate en base a la Ecuación (2).

$$Va_s = \sum_{i=1}^n P_i \cdot Va_i \quad (2)$$

Donde:

Va_s es el valor de la variable meteorológica Va que corresponde al sitio s

Va_i es el valor de la variable meteorológica Va en la estación i

Cabe aclarar que en caso en que alguno de los valores Va sea nulo, los pesos se re-escalán para sumar 1.

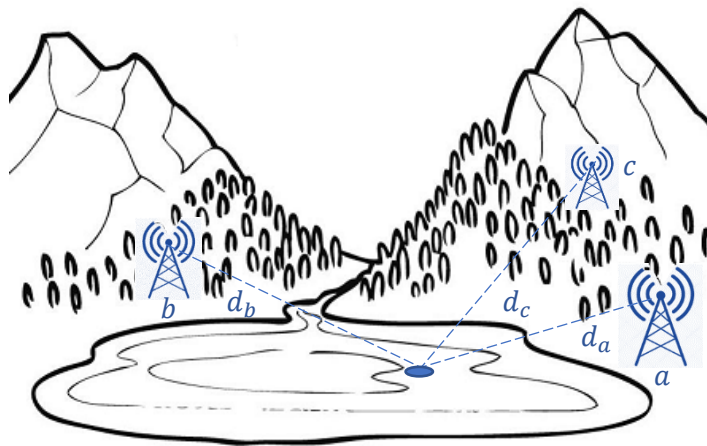


Fig. 14 Diagrama de las distancias entre las estaciones y el punto de muestreo (sitio). Donde

$d_a \ll d_b \ll d_c$. Fuente: elaboración propia

De esta manera, en términos generales es posible calcular un valor para cada una de las variables meteorológicas cuando se busca asociarlas a un punto horizontal de muestreo en el lago. Existen excepciones que se mencionan a continuación:

1. Las variables temperatura, y por consiguiente presión atmosférica, se ven afectadas por la altura (remitirse a la sección 1.2 para los conceptos meteorológicos y Tabla 6 para la altitud de las estaciones meteorológicas), por ende se requiere de un ajuste para llevar presión y temperatura a la altitud del lago, para así poder aplicarles promedio ponderado sin incrementar artificialmente el efecto de estaciones más próximas a la altitud del lago y a la inversa. En este sentido, se consideró lo siguiente, tomado de Ahrens y Henson (2017), pág. 203:
 - a. Por cada 1000 m de incremento en la altitud, la temperatura disminuye 6,5 °C.
 - b. Por cada 100 m de incremento en la altitud, la presión atmosférica disminuye 10 hPa. En este caso particular, solo se tienen datos provenientes de una estación (1100), pero igualmente se aplica la corrección para llevar el dato de presión atmosférica al valor correspondiente para la altitud del lago.Así las variables mencionadas se ajustaron de acuerdo a la altitud de las estaciones previo al cálculo del promedio ponderado.
2. En el caso de la variable hidrológica nivel, se presenta una particularidad, en el sentido que se tienen dos clases de nivel, el correspondiente al lago y el propio de los ríos San Antonio y Cosquín, principales tributarios del mismo. En lo que respecta al primero, puesto que representa una cota del lago medida en tres estaciones a lo largo del tiempo, que afecta todos los puntos sobre el plano horizontal del mismo de igual manera, se aplicó el promedio simple. En lo que se refiere a los ríos, las estaciones que miden nivel se encuentran ubicadas a la vera de los mismos, en distintos puntos de su curso, a distintas altitudes. Puesto que la geomorfología impide asumir un efecto de distancia en este caso no se aplica promedio ponderado. Mas aún, dado que ha habido inconvenientes en la medición de los niveles que sólo permite comparar los valores medidos “pico a pico”, quiere decir en términos de incrementos relativos a la serie de una misma estación, pero no en términos absolutos entre estaciones; entonces se opta por considerar la estación más próxima a cada punto de muestreo de los ríos, y a su vez, en el caso del Río San Antonio sólo para períodos de tiempo donde se considera que la medición fue correcta según informe interno de INA-CIRSA.
3. En lo que respecta a la dirección del viento, se empleó la variable secundaria “orientación” (remitirse a la sección 2.4.2.1.2) y por ser una variable categórica se aplicó la moda de los valores de las cuatro estaciones que miden dirección.
4. En lo que se refiere a precipitación, cabe aclarar que esta estrategia asume contribuciones lineales por distancia euclídea, y en rigor de verdad la precipitación de una cuenca se estima en base a polígonos. En vistas a que la aplicación del procedimiento podría requerir un estudio mucho mayor, se prefiere simplificar en base al promedio, con la intuición de que se tratará de una buena aproximación de mejores resultados que podrían obtenerse con estrategias propias de la hidrología.

En base a lo descripto anteriormente en mayor o menor medida las estaciones meteorológicas contribuyen a la determinación de las variables para los sitios del lago. En lo que respecta a la asociación espacial de los puntos de muestreo de ríos, se asume que es válida únicamente en relación a los puntos de desembocadura (uno a uno) que, bajo la misma pauta mencionada al comienzo, tienen una distancia menor al sitio que a cualquier otro punto del lago. Conforme la

distancia aumenta, el efecto resultaría muy débil para un fenómeno de tan corto alcance (a diferencia de los efectos meteorológicos).

La Tabla 13 resume los métodos de asociación descriptos anteriormente para cada caso.

Tabla 13 Resumen de los métodos de asociación espacial entre los sitios del lago, los de los ríos y las estaciones meteorológicas por variable.

Sitio (lago)	Sitio (río)	Temp.	Presión Atm.	Rad. Solar	Humedad Relativa	Vel. Viento	Dir. Viento	Precip.	Nivel (lago)	ID Est. Met. Nivel (río)
C	NA	PPa	Ua	PP	U	PP	M	PP	P	NA
TAC	NA	PPa	Ua	PP	U	PP	M	PP	P	NA
DSA	RSA	PPa	Ua	PP	U	PP	M	PP	P	600
DCQ	RCQ	PPa	Ua	PP	U	PP	M	PP	P	2900

Nota:

- PP: Promedio Ponderado de todas las estaciones que miden la variable.
- PPa: Promedio Ponderado de todas las estaciones que miden la variable después de ajuste.
- M: Moda de todas las estaciones que miden la variable.
- P: Promedio de todas las estaciones que miden la variable.
- U: Estación única
- Ua: Estación única ajustada
- NA: No Aplica.
- ID Est. Met: Número de Identificación de la Estación Meteorológica. Para el detalle de las estaciones meteorológicas que miden cada variable remitirse a la Tabla 7.

2.5 Series de tiempo

La forma más adecuada para analizar los datos con que se cuenta y así dar respuesta a la pregunta inicial, resulta ser considerarlos en términos de *series de tiempo*. Una serie de tiempo, es “un conjunto de observaciones x_t en un período t específico” (Brockwell & Davis, 2006), donde los intervalos de tiempo son fijos.

2.5.1 Modelo de serie de tiempo

A efectos de poder analizar las series de tiempo, es preciso seleccionar un modelo adecuado. Formalmente, un “modelo de serie de tiempo para datos observados $\{x_t\}$ es una especificación de las distribuciones conjuntas (o posiblemente la media y la covarianza únicamente) de una secuencia de variables aleatorias $\{X_t\}$ de la cual $\{x_t\}$ se postula como una realización” (Brockwell & Davis, 2002). La serie de tiempo se caracteriza por tener: un momento de inicio en el tiempo (a partir de cuándo los valores se comenzaron a medir) y una frecuencia (que es la periodicidad con que se mide el fenómeno: día, mes, año, etc.).

2.5.2 Conceptos principales

Una serie de tiempo puede expresarse como se muestra en la Ecuación (3), como la suma de varias componentes (este es el caso aditivo, existiendo su contraparte multiplicativa donde las componentes se multiplican en lugar de sumarse). Por un lado, se tiene una componente de

tendencia que determina la pendiente de la serie temporal a lo largo del tiempo, es decir existe cuando se produce un incremento o decremento de los datos a lo largo del tiempo (por ejemplo, a partir de la abundancia de nutrientes puede dispararse el crecimiento de algas a lo largo del tiempo). Por otro lado, se tiene un componente de *estacionalidad*, cuando la serie presenta patrones repetitivos en relación con períodos de tiempo (por ejemplo, la serie de tiempo de temperatura responde a patrones de estaciones del año donde en verano los valores tienden a ser más altos y en invierno más bajos). Por último, se tiene un componente denominado ruido blanco, de carácter puramente aleatorio.

$$X_t = T_t + S_t + W_t \quad (3)$$

Donde:

X_t es la serie temporal compuesta

T_t es el componente de tendencia

S_t es el componente de estacionalidad

W_t es el ruido blanco

La importancia de la determinación y análisis de estos componentes, estriba en el hecho de que muchos de los modelos que pueden proponerse para representar el proceso aleatorio que define una serie de tiempo particular, incluyen supuestos al respecto de la *estacionariedad* de la serie. Una serie se dice estacionaria cuando carece de componentes de tendencia y estacionalidad. Por este motivo, en caso de existir los componentes mencionados, se neutralizan con el objetivo de obtener ruido blanco estacionario. Formalmente, una serie se dice estacionaria en sentido estricto cuando el comportamiento probabilístico de cada colección de valores $\{x_{t1}, x_{t2}, \dots, x_{tk}\}$ es idéntico al de la colección de valores desfasados h períodos en el tiempo $\{x_{t1+h}, x_{t2+h}, \dots, x_{tk+h}\}$ (Shumway & Stoffer, 2017).

Otro concepto relevante, relacionado con lo anterior es el de *diferenciación*. Dada una serie X_t , puede obtenerse una serie $X'_t = X_t - X_{t-1}$, tal que X'_t es la serie de las diferencias entre las observaciones consecutivas de X_t (nótese que, si la longitud de la serie original es k , luego la longitud de la serie diferenciada es $k-1$). Una serie puede diferenciarse sucesivas veces y cada paso de diferenciación se conoce como orden. Este procedimiento es relevante en términos de que se emplea para conseguir que una serie no estacionaria se vuelve estacionaria. El mismo concepto puede aplicarse en términos de la estacionalidad. Es decir, la diferenciación estacional, es la diferencia entre una observación y la misma observación anterior de la misma estación, es decir $X'_t = X_t - X_{t-m}$, donde m es la cantidad de estaciones. Por ejemplo, si la estacionalidad de una serie de tiempo es mensual, entonces al valor de la serie en enero de un determinado año, se le resta el valor de la serie en enero del año anterior.

El modelado por medio de series de tiempo, suple una falencia de los modelos tradicionales, que, como en el caso de la regresión lineal, asumen independencia de las observaciones. En el caso de una serie de tiempo, la observación en el tiempo t , podría estar relacionada con la observación en el tiempo $t-1, t-2, \dots, t_0$, desfases en el tiempo que se denominan *retardos*. Esto se debe a que los fenómenos de la naturaleza, los procesos físicos, exhiben inercia, lo que implica que no cambian tan rápidamente, o presentan efecto retardado. Este hecho, combinado con la frecuencia de muestreo, hace que las observaciones consecutivas estén correlacionadas, lo que se denomina *autocorrelación* (Bisgaard & Kulahci, 2011).

2.6 Análisis exploratorio

A los efectos de describir las series temporales y sus características, se pueden emplear ciertos procedimientos que se detallan a continuación.

2.6.1 Descomposición de la serie de tiempo

A efectos de poder caracterizar la serie de tiempo en función de sus componentes constitutivas, como se mencionó en la sección 2.5.2, es preciso realizar una descomposición de la misma. El método utilizado en este análisis es *STL* (siglas en inglés de “Seasonal and Trend Decomposition using Loess” o Descomposición de Estacionalidad y Tendencia usando Loess) (Cleveland et al., 1990). Se trata de un método de descomposición versátil y robusto basado en Loess, un procedimiento para estimar relaciones no lineales. El método STL parte de la definición de una ventana de tiempo para el cálculo de la tendencia y otra para el cálculo de la estacionalidad. Estos dos parámetros (que deben ser impares) permiten controlar la cantidad de variación que se admite en cada caso.

El resultado es una componente de tendencia de la serie original, una componente de estacionalidad y una componente de restos (ruido blanco), utilizadas para realizar la caracterización de la serie de tiempo. Cabe aclarar, además, que algunos autores se refieren a la componente de tendencia como de tendencia-ciclos, ya que permite además analizar los ciclos de la serie, entendidos como las oscilaciones de la serie que no se producen a una frecuencia fija como ocurre con la estacionalidad, y que normalmente tienen una longitud mayor que el patrón de estacionalidad (por lo general varios años) y presentan mayor variabilidad comparativamente.

En la misma línea, el test de Mann-Kendall permite determinar de forma analítica si existe evidencia suficiente para aseverar que existe tendencia monotónica en la serie. En particular, en este análisis se empleó la versión del test que contempla estacionalidad, propuesto por Hirsch et al. (1982). Se trata de un test no paramétrico y que admite valores faltantes de la serie. La hipótesis nula es que, para cada una de las m estaciones (en este caso de análisis 12 meses), las n observaciones son independientes y se distribuyen idénticamente, mientras que, la hipótesis alternativa es que existe tendencia monotónica.

Por último, y yendo un paso más allá para completar el análisis, en caso que exista evidencia suficiente para afirmar la presencia de tendencia monotónica, es posible calcular la pendiente de la recta que la interpola, denominada pendiente de Sen. Propuesto también en Hirsch et al. (1982), constituye una extensión al caso estacional del método de cálculo propuesto por Theil y Sen. Este estimador expresa la magnitud de una tendencia como una pendiente, o lo que es lo mismo, el cambio de la serie por unidad de tiempo.

2.6.2 Autocorrelación

La *autocorrelación*, es una función que mide la correlación entre valores desfasados, con retardo, de una serie de tiempo (Hyndman & Athanasopoulos, 2018). Así, es posible calcular los

coeficientes de correlación para una serie de retardos, así como su intervalo de confianza de la estimación puntual. Cuando alguno de los coeficientes presenta un valor por fuera del intervalo se interpreta que la condición de independencia fue violada y por lo tanto existe *dependencia* para el retardo correspondiente. Los coeficientes obtenidos para distintos retardos, junto con los intervalos de confianza se pueden representar en un gráfico conocido como *correlograma*.

La importancia de analizar la autocorrelación de una serie estriba en determinar el grado de dependencia que existe entre una serie y esa misma serie desfasada en el tiempo. En términos predictivos, puede servir para determinar si tiene sentido inferir el valor de una variable en el período $t+k$ (k es el retardo) a partir del período t , o más bien tiene un comportamiento eminentemente aleatorio. En términos de análisis de residuos de un modelo regresivo, puede servir para determinar si se trata de ruido blanco o más bien existe información que no ha sido capturada por el modelo en caso de autocorrelación significativa.

2.6.3 Autocorrelación parcial

Esta técnica permite analizar la correlación entre x_t y x_{t-k} después de quitar el efecto de los retardos 1, 2, ... $k-1$. Así, la primera autocorrelación parcial es idéntica a la primera autocorrelación (no hay retardos intermedios), y cada autocorrelación parcial subsiguiente se puede estimar como el último coeficiente en un modelo autorregresivo (Hyndman & Athanasopoulos, 2018). El intervalo de confianza es coincidente con el de la función de autocorrelación.

La autocorrelación parcial es importante para analizar dependencias en procesos de tiempo autorregresivos, donde la autocorrelación por sí sola no aporta suficiente información al respecto.

2.6.4 Interpolación

A efectos de subsanar el faltante de valores en las series de tiempo, ya sea para la aplicación de los métodos de análisis como para la posterior construcción de las variables predictoras del modelo, es posible aplicar un procedimiento clásico en el análisis de las series temporales denominado interpolación. En su enfoque más sencillo, consiste en la estimación del valor faltante como el punto medio de los valores adyacentes, lo que se conoce como interpolación lineal.

El método empleado en este análisis, siguiendo a Hyndman y Khandakar (2008), tiene un nivel más de complejidad: si la serie presenta estacionalidad, la interpolación del valor faltante se aplica como sigue:

1. Se realiza la descomposición STL robusta de la serie (como se explicó en la sección 2.6.1), para obtener la componente estacional.
2. Se elimina la componente de estacionalidad de la serie original.
3. Se calcula la interpolación lineal del punto faltante usando los valores adyacentes de la serie ajustada por estacionalidad (2).

4. Se suma el valor de la componente estacional al valor obtenido en (3), para obtener el valor final interpolado.

2.6.5 Índice de estado trófico

Por otro lado, como parte del análisis exploratorio, y tendiente a lograr el objetivo de este estudio, se calcula un índice asociado a la calidad del agua.

El *Índice de Estado Trófico (IET)* o *índice de Carlson* (Carlson, 1976), es un indicador que mide, en una escala de 0 a 100, el grado de eutrofización de los lagos, donde a mayor grado corresponde peor calidad de agua y viceversa. Este índice se puede calcular para Secchi, fósforo total y clorofila-a, como se muestra en la Ecuación (4).

$$\begin{aligned}
 IET(\text{Secchi}) &= 10 \left(6 - \frac{\ln \text{Secchi}}{\ln 2} \right) \\
 IET(\text{Clorofila}) &= 10 \left(6 - \frac{2.04 - 0.68 \ln \text{Clorifila}}{\ln 2} \right) \\
 IET(PT) &= 10 \left(6 - \frac{\ln \frac{48}{TP}}{\ln 2} \right)
 \end{aligned} \tag{4}$$

Cabe aclarar que, en el caso de la variable de transparencia de Secchi, se observan valores 0 (no así en las restantes dos). Puesto que el logaritmo natural no está definido en 0, se reemplaza este valor por 0,05 (5cm), arrojando un IET levemente por encima de 100 para hacer notar que se trata de un valor fuera de lo normal.

2.7 Modelado

En esta sección se detallan los modelos predictivos de series temporales que se emplearon para este estudio. Se consideran tres tipos de modelos que se han puesto a prueba bajo distintas configuraciones de parámetros. En cada caso se describen los fundamentos del modelo y la metodología de trabajo.

2.7.1 Regresión dinámica

Es el primero de los modelos probados, que admite sólo una serie temporal respuesta. Se basa en el modelo de regresión lineal con una modificación conveniente para el caso de series temporales.

2.7.1.1 Especificación del modelo

Un modelo de regresión dinámica, “establece cómo una respuesta (Y_t), se encuentra linealmente relacionada a los valores actuales y pasados de una o más entradas ($X_1, X_2, \dots, X_M$)” (Pankratz, 1994). Es decir, en términos de series temporales, se relaciona la serie temporal de una variable que se interesa predecir, con una o más series temporales que funcionan como *regresoras*. En este sentido es importante destacar un supuesto fundamental del modelo, y es que la respuesta no afecta a las entradas, la Fig. 15 ilustra este concepto.

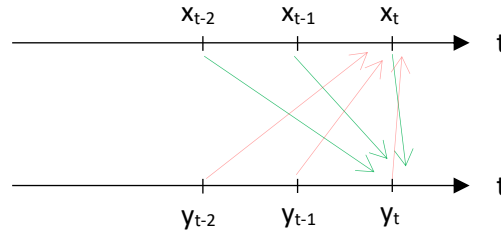


Fig. 15 Relaciones entre Y_t y X_t . Las flechas verdes representan el sentido aceptable de la relación, mientras que las rojas, el sentido no apropiado para un modelo de regresión dinámica.

Fuente: adaptado de Pankratz (1994).

Un modelo de regresión dinámica es, dicho de manera técnica, un modelo de regresión lineal donde el término correspondiente al error se asume que no sigue una distribución normal, sino que, los errores se encuentran correlacionados en el tiempo. De esta manera, el modelo tiene en cuenta dos aspectos para describir la relación lineal entre la respuesta y las entradas: la relación posiblemente desfasada en el tiempo entre una y otras, y la posible estructura del tiempo (patrón de autocorrelación) de la serie del error; son precisamente los aspectos que le dan su carácter dinámico (en términos del tiempo). La primera parte, se modela como una *función de transferencia* y el error como un modelo ARIMA (por sus siglas en inglés, modelo *Autorregresivo Integrado de Media Móvil*). El modelo ARIMA es un modelo compuesto que captura dos aspectos de una serie integrada, es decir una serie originalmente no estacionaria que se vuelve estacionaria a partir de diferenciación (ver sección 2.5.2 para definiciones al respecto):

- Modelo autorregresivo: es una regresión lineal de la serie de tiempo en términos de sí misma. Es decir, la variable en un tiempo t , depende de una combinación lineal de esa misma variable observada en un tiempo $t-1, t-2, \dots, t-p$, siendo p el orden del modelo.
- Modelo de media móvil: es una regresión lineal de los errores de la serie de tiempo en términos de sí misma. Es decir, la variable en un tiempo t , depende de una combinación lineal de los errores en ese momento t más en los tiempos $t-1, t-2, \dots, t-q$, siendo q el orden del modelo. Los términos de error corresponden a los errores de las regresiones individuales de cada observación en relación a la observación anterior.

La Ecuación (5) muestra la formulación de un modelo dinámico, y la Ecuación (6), el detalle del error de regresión multivariada como ARIMA. La función de transferencia en sentido estricto no abarca el término de error que precisamente se detalla por separado.

$$Y_t = C + \sum_{i=1}^M v_i(B) X_{i,t} + N_t \quad (5)$$

$$\varphi(B^s)\varphi(B)\nabla_s^D\nabla^d N_t = \theta(B^s)\theta(B)\varepsilon_t \quad (6)$$

Donde:

Y_t es la respuesta (serie de tiempo respuesta).

$X_{i,t}$ es la i -ésima entrada (serie de tiempo regresora).

M es el número de entradas (series de tiempo regresoras).

C es el término constante (independiente) de la regresión.

B es el operador de desfasaje, por ejemplo: $BX_t = X_{t-1}$.

$v_i(B)$ es la serie de los coeficientes de la entrada $X_{i,t}$, contemplando los efectos retardados sobre la respuesta Y_t .

N_t es el error de la regresión que se modela ARIMA(p,d,q)(P,D,Q), donde p es el orden del modelo autorregresivo no estacionario, d es el orden de diferenciación no estacionario, q es el orden del modelo de media móvil no estacionario, P es el orden del modelo autorregresivo estacionario, D es el orden de diferenciación estacionario, y Q es el orden del modelo de media móvil estacionario.

∇^d es el operador de diferenciación de orden d. $\nabla^d = (1 - B)^d$.

$\varphi(B)$ es el operador autorregresivo de orden p. $\varphi(B) = (1 - \varphi_1 B - \varphi_2 B^2 - \dots - \varphi_p B^p)$.

$\theta(B)$ es el operador de media móvil de orden q. $\theta(B) = (1 - \theta_1 B - \theta_2 B^2 - \dots - \theta_q B^q)$.

∇_S^D es el operador de diferenciación estacional de orden D, con S estaciones. $\nabla_S^D = (1 - B^S)^D$.

$\varphi(B^S)$ es el operador autorregresivo estacional de orden P. $\varphi(B^S) = (1 - \varphi_S B^S - \varphi_{2S} B^{2S} \dots - \varphi_{PS} B^{PS})$.

$\theta(B^S)$ es el operador de media móvil estacional de orden Q. $\theta(B^S) = (1 - \theta_S B^S - \theta_{2S} B^{2S} \dots - \theta_{QS} B^{QS})$.

ε_t es el error del modelo completo, asumiendo $\varepsilon_t \sim N(0, \sigma^2)$.

En concreto, el modelo de regresión dinámica requiere definir:

1. Una serie de tiempo respuesta (Y_t).
2. M series de tiempo regresoras $X_{i,t}$ ($i = 1, 2, \dots, M$), las cuales pueden constituir distintas combinaciones desfasadas en el tiempo de las variables X_i .
3. Los parámetros del modelo ARIMA: p, d, q, P, D, Q para un número S de estaciones.

En base a los cuales el modelo estimará los siguientes parámetros:

1. Término independiente C
2. Coeficientes v_i de las series de tiempo regresoras.
3. Coeficientes φ del modelo autorregresivo sobre los errores de la regresión.
4. Coeficientes θ del modelo de media móvil sobre los errores de la regresión.
5. Varianza σ^2 del error final del modelo.

2.7.1.2 Estimación de los parámetros

En lo que respecta a la estimación de los parámetros de un modelo de regresión dinámica no se utiliza el método de mínimos cuadrados ordinarios como se hace comúnmente en una regresión lineal estándar, debido a una serie de inconvenientes que se mencionan en Hyndman y Athanasopoulos (2018), por lo se realiza por el método de máxima verosimilitud (Hipel & McLeod, 1994).

2.7.1.3 Evaluación de los modelos de regresión dinámica

A efectos de seleccionar p , q , P y Q para cada uno de los modelos generados bajo distintas combinaciones, se utiliza el Criterio de Información de Akaike corregido (AICc) cuya fórmula se muestra en la Ecuación (7) (Bisgaard & Kulahci, 2011).

$$AICc = -2 \ln(L) + \frac{2rn}{n - r - 1} \quad (7)$$

Donde:

L es la máxima verosimilitud del modelo.

r es el número de parámetros del modelo.

n es la longitud de las series de tiempo.

En lo que respecta a las condiciones necesarias para la constitución de un modelo como candidato, se considera además el diagnóstico de residuos, siguiendo a Hyndman y Athanasopoulos (2018), donde se evalúan los siguientes aspectos:

1. No autocorrelación de los residuos. Se evalúa de forma gráfica que la autocorrelación tal cual se describió en 2.6.2 aplicada a la serie de tiempo de los residuos del modelo no presente rezagos con correlación significativa. En términos analíticos se aplica el test de Ljung-Box que contempla un *grupo* de autocorrelaciones de la serie de tiempo de residuos resultantes de un modelo ARIMA, siendo la hipótesis nula que los residuos son aleatorios (ruido blanco) y la hipótesis alternativa que no lo son (para detalles sobre el estadístico, remitirse a Hyndman y Athanasopoulos (2018)). A mayor valor del estadístico de Ljung-Box, mayores chances de autocorrelación en los residuos. Si este es el caso se asume que el modelo no ha capturado toda la información relevante para explicar los datos y por ende es débil.
2. Media cero. Si la media no es cero el modelo tiene un sesgo. Este aspecto se evalúa gráficamente.
3. Varianza constante. Se evalúa de forma gráfica.
4. Distribución normal. Se evalúa a partir del histograma. Este aspecto junto con el anterior, influyen en el cálculo de los intervalos de confianza.

2.7.1.4 Metodología de modelado

El modelo de regresión dinámica permite determinar los efectos de varias variables en el tiempo sobre otra. Asumiendo que existe independencia entre los procesos estocásticos subyacentes en cada sitio-perfil, es posible plantear diez modelos de regresión dinámica (uno por cada combinación sitio-perfil) que describan cada uno, de la mejor manera posible, la naturaleza de cambio de la variable respuesta en función de las regresoras en cada punto.

A efectos de aplicar el modelado de regresión dinámica, para cada uno de los diez casos, se siguió el siguiente procedimiento, adaptado de Hyndman y Athanasopoulos (2018):

1. Identificar grupos de variables regresoras que se modelan a la vez.
2. Aplicar interpolación como se describió en 2.6.4 a las variables regresoras seleccionadas (es requisito del modelo que no existan valores faltantes en las regresoras).
3. Normalizar las series para que tengan media cero y desvío estándar unitario, de esa manera se pueden comparar los coeficientes con posterioridad.
4. Aplicar el algoritmo de Hyndman-Khandakar para modelado ARIMA automático (Hyndman & Khandakar, 2008).
5. Realizar el diagnóstico del modelo.
6. Seleccionar, para cada grupo de modelos, el mejor teniendo en cuenta RMSE (ver 2.7.4 para la definición) y diagnóstico de residuos. Los autores proponen el uso de AICc para comparar los modelos siempre y cuando no exista distinto orden de diferenciación entre ellos, en cuyo caso proponen el uso de RMSE. Como en este caso se tienen variedad de combinaciones, donde a pesar de que en un punto no se dé el caso, para que el análisis sea justo en términos de los puntos entre sí, se adopta el criterio de RMSE.

2.7.1.5 Significatividad de los parámetros

A efectos de determinar el grado de significatividad de los parámetros se aplicó la prueba t fijando el nivel de confianza al 95%. A tal efecto se utilizó la fórmula en la Ecuación (8) para calcular el estadístico.

$$T_{0i} = \frac{\hat{\beta}_i - \beta_{0i}}{\sqrt{S(\hat{\beta}_i)}} \quad (8)$$
$$S(\hat{\beta}) = \sqrt{\text{diag}(\text{Var}(\hat{\beta}))}$$

Donde:

$\hat{\beta}_i$ es el i-ésimo coeficiente estimado por el modelo, con $i=1,...,k$

β_{0i} es el valor real del coeficiente, en este caso se fija en 0, porque la prueba consiste en conocer si el valor es significativamente distinto de 0.

$S(\hat{\beta}_i)$ es el error (o desvío) estándar estimado del i-ésimo parámetro del modelo. El mismo se calcula aplicando raíz cuadrada a la diagonal de la matriz de varianza-covarianza de los coeficientes estimados $\text{Var}(\hat{\beta})$.

T_{0i} es el i-ésimo estadístico t estimado.

A partir del test descrito en la Ecuación (9) se calcula el p-valor para determinar la significatividad de cada parámetro.

$$\begin{aligned} H_0: \beta_{0i} = 0, |T_0| &< \frac{t_{0,05}}{2}, n-k \\ H_1: \beta_{0i} \neq 0, |T_0| &> \frac{t_{0,05}}{2}, n-k \end{aligned} \quad (9)$$

Donde:

n es la longitud de la serie

Vale observar que la probabilidad de que T_0 sea menor que el valor de la función de distribución t de student para el percentil 2,5% se multiplica por 2 para contemplar los dos lados de la distribución. Si el valor es menor a 0,05 se considera que existe evidencia suficiente para descartar la hipótesis nula, y el parámetro estimado por el modelo se considera significativamente distinto de cero.

2.7.2 Modelo de espacio de estados multivariado

Se trata del segundo tipo de modelo empleado, que admite respuesta múltiple y consta de dos ecuaciones que cumplen funciones distintas.

2.7.2.1 Especificación del modelo

En su concepción general, “se asume que el desarrollo del sistema en estudio, a lo largo del tiempo, está determinado por una serie de vectores no observados, $x_1, x_2, \dots, x_n$, asociados a una serie de observaciones $y_1, y_2, \dots, y_n$; donde la relación entre unos y otros está especificada por un modelo de espacio de estados” (Durbin & Koopman, 2012). El objetivo del análisis de espacio de estados es, en este sentido, precisamente inferir las características más relevantes del proceso no observado basado en las observaciones disponibles. Este tipo de modelo también se conoce con el nombre de modelo lineal dinámico.

En términos más específicos y haciendo una natural extensión de la definición anterior a la inclusión dentro del modelo de variables exógenas, se tiene el modelo de espacio de estados multivariado. En este sentido, a los efectos de este estudio se sigue a Holmes et al. (2018a), quienes lo definen por medio de las Ecuaciones (10), (11) y (12).

$$x_t = B_t x_{t-1} + u_t + C_t c_t + G_t w_t, \quad w_t \sim NM(0, Q_t) \quad (10)$$

$$y_t = Z_t x_t + a_t + D_t d_t + H_t v_t, \quad v_t \sim NM(0, R_t) \quad (11)$$

$$x_1 = NM(\pi, \Lambda) \text{ ó } x_0 = NM(\pi, \Lambda) \quad (12)$$

La Ecuación (10) representa el proceso de estado, la Ecuación (11) el proceso observacional y la Ecuación (12) los estados iniciales, donde:

x_t es una realización de la variable aleatoria X_t en el momento t que representa el proceso de estados. Es una matriz de $m \times T$, siendo m la cantidad de estados que se definen y T la longitud

de las series de tiempo. El primer estado se calcula en base a una distribución y puede formar parte de los resultados (x_1) o bien ser el paso previo para el cálculo del primer estado (x_0).

w_t representa el proceso de error en el momento t , que se asume sigue una distribución normal multivariada (NM) de media 0 y covarianza Q_t . Es una matriz de $m \times T$.

y_t es una realización del proceso de observaciones en el momento t . Es una matriz de $n \times T$, siendo n la cantidad de variables observacionales y T la longitud de las series de tiempo.

v_t representa el error no debido al proceso en el momento t , que se asume sigue una distribución normal multivariada de media 0 y covarianza R_t .

B_t y Z_t son parámetros, matrices de $m \times m$ y $n \times m$ respectivamente.

u_t y a_t son parámetros, matrices de $m \times 1$ y $n \times 1$ respectivamente.

Q_t y R_t son parámetros, matrices de $g \times g$ (normalmente $m \times m$) y $h \times h$ (normalmente $n \times n$).

π es un parámetro o una distribución a priori fija. Es una matriz de $m \times 1$.

Λ es un parámetro o una distribución a priori fija. Es una matriz de covarianza de $m \times m$.

C_t y D_t son parámetros, matrices de $m \times p$ y $n \times q$ respectivamente.

c_t y d_t son valores fijos de entrada, matrices de $p \times T$ y $q \times T$ respectivamente.

G_t y H_t son valores fijos de entrada, matrices de $m \times g$ y $n \times h$ respectivamente.

Como x_t depende de x_{t-1} este modelo es además autorregresivo, y en adelante se denominará por las siglas de su nombre completo en inglés: MARSS (Espacio de Estados AutoRegresivo Multivariado).

En concreto el modelo MARSS, permite estimar los parámetros mencionados, aplicando restricciones o haciendo cero algunos de ellos, de manera tal de especificar un modelo con una única solución. De esta manera también es posible adaptarlo para responder a distintas preguntas, como se mostrará en este estudio.

2.7.2.2 Estimación de los parámetros

A efectos de estimar los múltiples parámetros del modelo se utiliza también (como en 2.7.1.2) el método de máxima verosimilitud, pero basado en un algoritmo de Esperanza Maximización (EM por sus siglas en inglés). Este algoritmo permite extender el procedimiento de estimación de máxima verosimilitud cuando existen variables aleatorias ocultas en el modelo (X_t). Los autores de MARSS, como paquete estadístico (Holmes et al., 2018c), mejoran el algoritmo clásico EM (Holmes, 2012) para permitir restricciones lineales genéricas entre elementos de matrices (las que se especificaron antes). La maximización de la verosimilitud se computa mediante suavizado de Kalman.

2.7.2.3 Evaluación de los modelos MARSS

A efectos de seleccionar el mejor modelo bajo la metodología de MARSS se empleó AICc, tal cual se describe en la Ecuación (7). A pesar de que los autores proponen una variante mejorada de AICb basado en *Bootstrap*, que se indica especialmente efectiva para prevenir la selección de modelos extremadamente complejo con series cortas, su cálculo resulta prohibitivo en términos de tiempo de procesamiento para una evaluación exhaustiva de parámetros como la propuesta en este trabajo.

2.7.2.4 Metodología de modelado

A diferencia de la estrategia planteada con anterioridad (ver 2.7.1.4), donde se modela por separado cada una de las diez series temporales como procesos *independientes*, el objetivo en este caso es integrar procesos en distintos modelos, de manera de capturar las interacciones que existen entre los diferentes puntos de muestreo. En este sentido, se consideran dos estrategias a los efectos de conocer aún más acerca de la influencia de las condiciones meteorológicas en la calidad del agua:

1. Modelo de los procesos tróficos en profundidad.
2. Modelo de los procesos tróficos a un mismo nivel.

En ambos casos el procedimiento seguido es el siguiente:

1. Definir el modelo a ajustar en términos de los parámetros necesarios, restricciones y valores de entrada.
2. Calcular la bondad de ajuste del modelo según el indicador AICc e interpretar los resultados.
3. Repetir 1 para otra parametrización que tenga sentido en el contexto del problema.
4. Comparar todos los modelos propuestos y seleccionar aquel cuyo valor de AICc sea el menor, cumpliéndose los requisitos de convergencia y residuos aceptables (no autocorrelación y normalidad).

Los modelos candidatos, serán todos aquellos que hayan alcanzado la convergencia de todos sus parámetros, y no presenten inconvenientes en el diagnóstico de los residuos. En este caso, siguiendo a los autores de MARSS, en la aplicación del modelo en casos prácticos (Holmes et al., 2019), se contemplan de forma gráfica: aspecto de la serie de residuos y autocorrelación. Adicionalmente, se agrega, también el gráfico Q-Q para analizar normalidad.

El tratamiento de las variables es el mismo que el indicado en los puntos 1, 2 y 3 en la sección 2.7.1.4.

2.7.2.5 Significatividad de los parámetros

A efectos de evaluar la significatividad de los parámetros estimados, en el caso de este modelo se procedió a calcular los intervalos de confianza en cada caso, utilizando un nivel de confianza del 95%. En el caso en que el intervalo no contenga el 0, se considera que el parámetro estimado

es significativamente distinto de 0. Estos intervalos se calcularon utilizando el método de *bootstrapping* paramétrico, fijando n en 10 (referirse a Holmes et al. (2018a) para el detalle de implementación y a Efron y Tibshirani (1994) para el concepto estadístico).

2.7.3 Modelo de regresión bayesiano jerárquico

El tercer y último tipo de modelo, pertenece a una categoría distinta de los anteriores en el plano estadístico. A nivel estructural, es un modelo que admite respuestas múltiples con parámetros compartidos.

2.7.3.1 Especificación del modelo

A diferencia de los modelos anteriores, basados en estadística frecuentista, este modelo se basa en estadística bayesiana. Siguiendo a Gelman et al. (2014) la inferencia bayesiana “es el proceso de ajustar un modelo de probabilidad a un conjunto de datos y resumir el resultado por medio de una distribución de probabilidad sobre los parámetros del modelo y sobre las cantidades no observadas, tales como predicciones de nuevas observaciones”.

En este caso, también el objetivo es expresar un modelo en términos de parámetros que describan la naturaleza dinámica de los procesos tróficos del lago, pero la diferencia está que la descripción se hace en términos probabilísticos. En su expresión más sencilla, un parámetro θ (o varios) se puede estimar basado en observaciones y , para lo cual se precisa establecer una distribución de probabilidad conjunta de ambos. El teorema de Bayes, plantea una solución a ese requerimiento, como se observa en la Ecuación (13).

$$p(\theta|y) = \frac{p(\theta)p(y|\theta)}{p(y)} \quad (13)$$

Donde

$p(\theta|y)$ es la distribución a posteriori del parámetro θ , dadas las observaciones y .

$p(\theta)$ es la distribución a priori del parámetro θ .

$p(y|\theta)$ es la distribución muestral o verosimilitud, o probabilidad de obtener las observaciones y , dado el parámetro θ .

$p(y)$ es la distribución marginal de y , cuya fórmula (en el caso continuo) se muestra en la ecuación (14). Se denomina también distribución predictiva a priori (no está condicionada a observaciones previas).

$$p(y) = \int p(\theta)p(y|\theta)d\theta \quad (14)$$

Por otro lado, para obtener predicciones de observaciones futuras, $\hat{y}$ (una vez que la variable y ha sido observada), se tiene análogamente una fórmula que se describe en la Ecuación (15).

$$p(\hat{y}|y) = \int p(\hat{y}|\theta)p(\theta|y)d\theta \quad (15)$$

Donde:

$p(\hat{y}|y)$ es la distribución predictiva a posteriori, la probabilidad de obtener observaciones estimadas a partir de las observaciones actuales.

$p(\hat{y}|\theta)$ es la probabilidad de obtener observaciones estimadas basadas en el parámetro estimado.

Aplicando estos conceptos es posible plantear un modelo de regresión, donde los parámetros se estimen de manera bayesiana, cumpliendo el mismo las siguientes características:

1. Es un modelo de regresión (no necesariamente lineal) que permite predecir una variable de calidad del agua basada en un conjunto de variables predictoras. Desde este punto de vista, se estiman los parámetros correspondientes a las distribuciones de los coeficientes de la regresión (distribuciones a priori) y también los parámetros de la función de verosimilitud.
2. Es un modelo jerárquico, en el sentido que se pueden estimar varios parámetros a la vez, algunos de los cuales están definidos por otros, potencialmente compartidos. Los parámetros están conectados a la estructura del problema. En este caso se dispone de información correspondiente a distintos puntos de muestreo, a diferentes profundidades y ubicaciones en el plano horizontal. Las respuestas de los procesos en cada punto se pueden modelar condicionadas a ciertos parámetros, los cuales en sí mismos están dados por una especificación de probabilidad en términos de otros parámetros, los que se denominan *hiperparámetros*. Esta dependencia de unos parámetros sobre otros genera la jerarquía del modelo. Por ejemplo, podría asumirse que se tienen varias respuestas $\hat{y}_j$ que se modelan como una regresión, donde para obtener las distribuciones, en lugar de estimar j varianzas distintas, se asume una única varianza σ^2 , que a su vez sigue una distribución determinada. Siguiendo el ejemplo, en el caso de los coeficientes de la regresión, se asume que, si bien son distintos, para cada media de la respuesta (μ_j), se distribuyen en base a una función de probabilidad de parámetros comunes. La Fig. 16 ilustra este caso.

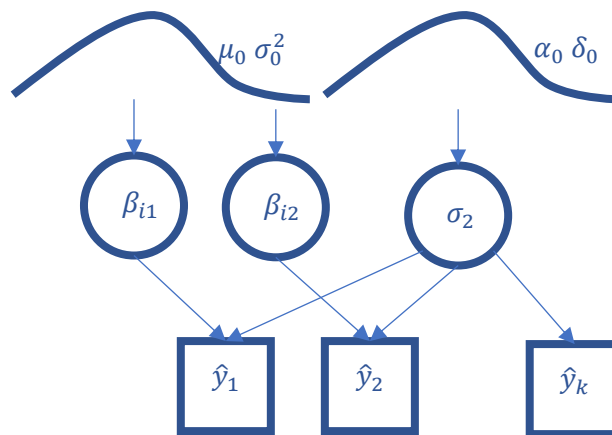


Fig. 16 Representación esquemática de parte de un modelo bayesiano de regresión jerárquico. Fuente: elaboración propia

3. Es un modelo de series de tiempo, en cuanto debe reflejar la dinámica de los procesos a lo largo del tiempo. A tales efectos, el modelo contempla términos autorregresivos y de media móvil.

Como resultado integral de las condiciones mencionadas con anterioridad, el modelo propuesto es ARIMAX. Se trata de un modelo ARIMA, como se describió en 2.7.1.1, pero extendido con el agregado de covariables como términos adicionales. El mismo, en su formulación bayesiana, adopta la forma que se muestra en la Ecuación (16)¹⁰.

$$\begin{aligned}
\hat{y}_{jt} &\sim N(\mu_{jt}, \sigma^2), & j = 1..k \\
\mu_{jt} &= R_{jt} + AR_{jt} + MA_{jt} \\
R_{jt} &= \beta_{j0} + \sum_i^n \beta_{ji} x_{ijt} \\
\beta_{ji} &\sim N(\mu_{\beta i}, \sigma_{\beta i}^2) \\
\beta_{j0} &\sim N(\mu_0, \sigma_0^2) \\
\varphi_j^m &\sim N(\mu_{\varphi_j^m}, \sigma_{\varphi_j^m}^2), & m = 1..p \\
\theta_j^n &\sim N(\mu_{\theta_j^n}, \sigma_{\theta_j^n}^2), & n = 1..p \\
\sigma &\sim U(l_i, l_s) \\
\varepsilon &\sim N(0, \sigma_\varepsilon^2)
\end{aligned} \tag{16}$$

Donde

$\hat{y}_{jt}$ es la respuesta predicha (distribución predictiva a posteriori) j-ésima (de un total de k respuestas). Se distribuye normal con media μ_{jt} y varianza σ^2 .

μ_{jt} es la media j-ésima.

σ^2 es la varianza del proceso integral. Este es uno de los parámetros compartidos del modelo jerárquico. La raíz cuadrada (desvío estándar) se sume uniforme con límites determinados.

R_{jt} es la componente de regresión lineal multivariada j-ésima.

β_{ji} es el coeficiente asociado a la i-ésima variable (de un total de n variables), correspondiente a la media de la componente de regresión j-ésima, distribuido normalmente con media $\mu_{\beta i}$ y varianza $\sigma_{\beta i}^2$. Estos últimos son parámetros compartidos entre los puntos de muestreo.

β_{j0} es el término independiente correspondiente a la media de la componente de regresión j-ésima, distribuido normalmente con media μ_0 y varianza σ_0^2 . Estos últimos son parámetros compartidos entre los puntos de muestreo.

AR_{jt} es la componente autorregresiva j-ésima, definida como en el lado izquierdo de la Ecuación (6), omitiendo el factor N (específico del error de la regresión dinámica). Los coeficientes φ_{jt} se distribuyen normales con una media y varianza determinadas para cada uno de los p términos (orden). Este podría ser uno de los parámetros compartidos en el modelo jerárquico.

¹⁰ La elección de las distribuciones de los parámetros no forma parte de la especificación del modelo ARIMAX, sino de su formulación en términos bayesianos, se eligieron a modo de ejemplo.

MA_{jt} es la componente de media móvil j -ésima, definida como en el lado derecho de la Ecuación (6). Los coeficientes θ_{jt} se distribuyen normales con una media y varianza determinadas para cada uno de los q términos (orden). El error necesario para calcular MA en cada paso, se determina como diferencia entre y_{jt} (respuesta real) y μ_{jt} . Este podría ser uno de los parámetros compartidos en el modelo jerárquico.

ε es el error o residuos de la estimación, entendido como diferencia entre y_{jt} y la media de la distribución predictiva a posteriori de $\hat{y}_{jt}$, que se asume sigue una distribución normal con media 0 y varianza determinada.

De esta manera, se integran en un único modelo las series temporales que son objeto de estudio principal de este análisis.

2.7.3.2 Estimación de los parámetros

Recapitulando, la idea central del modelado bayesiano es la obtención de una distribución a posteriori de los parámetros de interés que permite realizar una estimación de la respuesta (es en sí misma una distribución). Una vez desarrollado el modelo, como se planteó en la sección anterior, el trabajo restante es el cálculo de las distribuciones del mismo. A efectos de llevar a cabo este cometido, es indispensable remitirse a las Ecuaciones (13), (14) y (15). En particular, la Ecuación (14) plantea un desafío, puesto que según la complejidad que adopte la forma de la integral, podría no tener una resolución analítica. Por tal motivo, es necesario recurrir a una familia de métodos numéricos de estimación, denominados genéricamente *Cadenas de Markov Monte Carlo* o MCMC (por sus siglas en inglés).

MCMC permite producir una aproximación de la distribución a posteriori, en la forma de un gran número de valores del parámetro θ muestreados de esa distribución, pero sin calcular la misma explícitamente (remitirse a Gelman et al. (2014) para una explicación de la familia de métodos MCMC). Este conjunto de valores, se puede emplear luego para calcular la tendencia central de la distribución y su intervalo de más alta densidad. Teniendo cadenas de Markov lo suficientemente largas, las probabilidades dan lugar a una distribución estacionaria, condición deseable para garantizar estabilidad en las estimaciones.

Existe un amplio espectro de algoritmos para muestrear distribuciones a posteriori basados en las propiedades de las cadenas de Markov. Para este estudio se empleó *Gibbs* (ver Gelman et al. (2014)), implementado en el paquete de R: rJAGS (Plummer, 2018).

2.7.3.3 Evaluación de los modelos de regresión bayesianos jerárquicos

A los efectos de determinar la calidad de los modelos en términos de convergencia de las cadenas en distribuciones estacionarias se emplean los criterios que se mencionan a continuación:

- Autocorrelación de los valores de la cadena: analizando el correlograma correspondiente a medida que aumentan los retardos, la correlación debería tender a

cero como evidencia de convergencia. Cuanta mayor independencia entre los valores de la cadena, mayor información contenida dada una cantidad de muestras.

- Tamaño de muestra efectivo: es un concepto relacionado a lo anterior, que se refiere al número de muestras independientes de la distribución estacionaria que deberían muestrearse para tener información equivalente en la cadena de Markov. Es decir que, para una cantidad de muestras fijada en n , el tamaño de muestra efectivo m , es una medida de cuán bien *rindieron* las n muestras. El modelo es mejor en la medida en que m se aproxime a n .
- Calentamiento (*burn-in*): otro parámetro a observar es el número de muestras a descartar hasta que la cadena se encuentra lo suficientemente estabilizada. Es un parámetro que remite a la rapidez de convergencia.
- Diagnóstico de Gelman-Rubin: es un estadístico que constituye una prueba fuerte de convergencia cuando se dispone de varias cadenas a comparar (cada una con un valor de comienzo distinto). En concreto se calcula la variabilidad entre las cadenas, por lo que, en caso de convergencia, debiera ser pequeño y el estadístico próximo a 1.
- Distribución de las cadenas: inspeccionando las cadenas a partir de un gráfico, no debieran observarse patrones diferenciados entre las mismas, caso contrario no se ha conseguido convergencia.

En lo que respecta a bondad de ajuste se empleará un criterio análogo al el Criterio de Información de Akaike ajustado (AICc) que se describió en la sección 2.7.1.3, pero adecuado al contexto del paradigma bayesiano: Criterio de Información de Deviancia ajustado (DICc) (Spiegelhalter et al., 2002). Esencialmente es la suma de la estimación de la deviancia bayesiana del parámetro más un término de penalización por complejidad del modelo en cuanto a cantidad de parámetros. Así cuanto más grande el valor de DICc, peor es el modelo.

A partir de los conceptos mencionados anteriormente se procederá a evaluar los modelos generados. Adicionalmente, se contempla no autocorrelación y normalidad de los residuos. En este sentido vale la pena recordar que, como se mencionó antes, se consideran los residuos que se obtienen al calcular la media de la distribución predictiva a posteriori y restarla a la respuesta real.

2.7.3.4 Metodología de modelado

Siguiendo la propuesta de Gelman et al. (2014), se aplicará el siguiente procedimiento:

1. Configurar un modelo probabilístico completo, basado en las consideraciones de la sección 2.7.3.1. Para lo cual se contemplan distintas distribuciones a priori de los parámetros, así como también de la función de verosimilitud.
2. Calcular e interpretar la distribución a posteriori.
3. Evaluar la bondad de ajuste del modelo y las implicancias de la distribución a posteriori resultante.

El proceso será iterativo hasta encontrar el mejor modelo posible en base a distintas configuraciones iniciales, evaluados por los criterios mencionados en la sección 2.7.3.3, dando lugar así a un modelo propuesto para explicar la dinámica del sistema bajo consideración.

Por otra parte, el tratamiento de las variables es el mismo que el indicado en los puntos 1, 2 y 3 en la sección 2.7.1.4.

2.7.3.5 Significatividad de los parámetros

La evaluación de que los parámetros estimados sean significativamente distintos de cero, se realiza considerando el intervalo de credibilidad (concepto que podría equipararse al de intervalo de confianza en el paradigma frecuentista¹¹) de cada uno. En caso en que dicho intervalo no contenga el cero, se considera que el parámetro es significativamente distinto de cero.

2.7.4 Predicción

Luego de obtenidos los modelos bajo los distintos esquemas mencionados, es posible realizar la estimación, predicción o *forecast* hacia adelante. En este caso, se siguió la estrategia de reservar un año para comparar el pronóstico versus la respuesta real observada, abarcando dicho período desde septiembre de 2017 hasta agosto de 2018. En este sentido, se utilizó como medida de error en la estimación la raíz cuadrada del error cuadrático medio o *RMSE* por sus siglas en inglés. La Ecuación (17), muestra la fórmula de este indicador de error.

$$RMSE = \sqrt{\frac{\sum_{t=1}^T (RR_t - RE_t)^2}{T}} \quad (17)$$

Donde:

T es la longitud de las series a comparar

RR_t es la serie respuesta real en el tiempo t

RE_t es la serie respuesta estimada por el modelo en el tiempo t

El procedimiento de predicción, en términos generales consiste en la aplicación del modelo a las covariables del período de prueba para tener una respuesta estimada. En los apartados que siguen a continuación se explican las particularidades de este procedimiento según el tipo de modelo.

2.7.4.1 Regresión dinámica

¹¹ Teniendo en cuenta que el intervalo de credibilidad indica, para un cierto nivel prefijado, la probabilidad de que el parámetro (que sigue una distribución de probabilidad) asuma un valor entre los límites del intervalo. En tanto, el intervalo de confianza, para el mismo nivel prefijado, representa la probabilidad de que el intervalo contenga al parámetro (que se asume tiene un único valor verdadero).

A los efectos de generar las predicciones del modelo se siguió a Hyndman y Athanasopoulos (2018), aplicando la función *forecast* del paquete ffp2 (Hyndman, 2018). Como resultado se obtiene la serie respuesta, junto con los intervalos de confianza correspondientes a 80 y 95% (éste último es el estándar en este estudio).

2.7.4.2 Modelo de espacio de estados multivariado

Para generar las predicciones del modelo de espacio de estados multivariado, fue necesario programar una función *forecast*, ya que no estaba disponible en el paquete MARSS (Holmes et al., 2018c). A tal efecto se siguieron las fórmulas originales del filtro de Kalman enunciadas en Shumway y Stoffer (2017), como se muestran en la Ecuación (18), así como también el ejemplo práctico de predicción. Vale aclarar que la nomenclatura se modificó para ser consistente con la utilizada por los autores del paquete MARSS para el modelo general en las Ecuaciones (10), (11) y (12).

$$\begin{aligned}
 x_t^{t-1} &= Bx_{t-1}^{t-1} + Cc_t \\
 P_t^{t-1} &= BP_{t-1}^{t-1}B' + Q \\
 x_t^t &= x_t^{t-1} + K_t\varepsilon_t \\
 P_t^t &= (I - K_tZ_t)P_t^{t-1} \\
 \varepsilon_t &= y_t - Z_tx_t^{t-1} - Dd_t \\
 \sigma_t^2 &= Z_tP_t^{t-1}Z_t' + R \\
 K_t &= P_t^{t-1}Z_t'(\sigma_t^2)^{-1}
 \end{aligned} \tag{18}$$

Donde:

t es el tiempo de 1,...,n siendo n la cantidad total de observaciones.

x_t^{t-1} es el valor del proceso de estados en el tiempo t , condicionado a los datos hasta $t-1$ (ecuación de predicción).

P_t^{t-1} es la varianza de x_t condicionado a los datos hasta $t-1$.

x_t^t es el valor del proceso de estados en el tiempo t , condicionado a los datos hasta t (filtro de Kalman propiamente dicho).

P_t^t es la varianza de x_t condicionado a los datos hasta t .

B, C, c_t, Q, Z, D y R como se describieron en 2.7.2.1.

ε_t es el error de predicción o innovación.

y_t es el valor del proceso observacional en el tiempo t o respuesta del modelo.

σ_t^2 es la varianza-covarianza del error de predicción.

K_t es la ganancia de Kalman en el tiempo t . Para el cálculo de la inversa de σ_t^2 , se empleó, siguiendo a los autores del paquete MARSS, la función *pcholinv*, que calcula la inversa de una

matriz basada en la descomposición de Cholesky, pero modificada para permitir ceros en la diagonal de la matriz, teniendo la inversa resultante filas o columnas cero (Holmes et al., 2018b).

Las condiciones iniciales son x_0^0 descripto como x_0 o x_1 en 2.7.2.1 y P_0^0 que es nula (para la parametrización que se usa en este trabajo).

La predicción, cuando $t > n$, se realiza por medio de x_t^{t-1} para calcular la respuesta estimada $\hat{y}_t$ y P_t^{t-1} que se emplea para calcular el desvío estándar estimado $\hat{\sigma}_t$, que a su vez permite calcular los intervalos de confianza (95%). En la Ecuación (19) se tienen las fórmulas correspondientes.

$$\begin{aligned}\hat{y}_t &= Zx_t^{t-1} + Dd_t \\ \hat{\sigma}_t &= \sqrt{\text{diag}(\sigma_t^2)}\end{aligned}\tag{19}$$

Donde:

$\text{diag}(\sigma_t^2)$ es la varianza de la matriz de varianza-covarianza del error de predicción.

A efectos de predicción, $\hat{x}_t^t = x_t^{t-1}$.

Las condiciones iniciales de la predicción son x_n^n y P_n^n , vale decir el último valor de la serie de filtro de Kalman y su varianza asociada respectivamente.

En lo que respecta a los intervalos de confianza se calculan como se indica en la Ecuación (20).

$$IC(95\%) = \hat{y}_t \pm 1,96\hat{\sigma}_t\tag{20}$$

2.7.4.3 Modelo de regresión bayesiano jerárquico

En el caso del modelo bayesiano la predicción se realiza simplemente agregando 12 casos nulos (filas sin valor) a la estructura de datos que corresponde a la respuesta del modelo, completando las covariables que se emplean para la construcción del mismo con los correspondientes 12 meses de prueba y volviendo a ejecutarlo con estos datos. De esta manera se estiman las respuestas de los 12 meses finales para el parámetro respuesta y a su vez se generan los intervalos de credibilidad como cuartiles 2,5 y 97,5 en el mismo paso.

3 Resultados

En este capítulo se presentarán los resultados obtenidos tanto del análisis exploratorio como en términos de modelado, a partir de las pautas indicadas en las secciones 2.6 y 2.7 respectivamente.

3.1 Análisis exploratorio de series temporales

El objetivo de esta sección es la descripción de las series conformadas en la sección 2.4.3. En principio las diez series de tiempo multivariadas resultantes de la integración de datos podrían caracterizarse como series heterogéneas en su composición, ya que se tiene por un lado ocho series que permiten evaluar el estado trófico en términos de sitio y profundidad: C1, C2, C4, C5, TAC1, TAC2, TAC4 y TAC5; y por otro lado, se tienen otras dos series que permiten evaluar el estado trófico únicamente a nivel subsuperficial, pero contienen información del aporte de los ríos tributarios: DCQ1 y DSA1.

Cada una de las variables que constituyen las series se analizarán, siguiendo los conceptos vertidos en la sección 2.6, a saber:

- Definición de la serie en términos de tendencia, ciclos y estacionalidad.
- Análisis de la serie en términos de autocorrelación y autocorrelación parcial.

La Tabla 14 presenta el detalle de valores faltantes en variables meteorológicas e hidrológicas seleccionadas (omitiendo nivel de ríos cuyos datos resultan escasos por lo mencionado en 2.4.2.2.1 y 2.4.3.2), mientras que la Tabla 15 hace lo propio con las variables seleccionadas de calidad del agua, las que están directamente vinculadas al Índice de Estado Trófico (IET). Para cumplir con los requisitos de los métodos de análisis exploratorio, es necesario contar con series sin faltantes, motivo por el cual se aplicó interpolación, como se detalla en 2.6.4, para completarlas (a afectos del análisis).

Tabla 14 Porcentaje de valores faltantes en las series mensuales del período de estudio (1999-2017) de las variables hidrológicas y meteorológicas

Variable	Porcentaje de faltantes de la serie				
	Alcance 1	Alcance 3	Alcance 7	Alcance 14	Alcance 21
Temperatura	0,46%	0,46%	0,46%	0,46%	0,00%
Radiación Solar	11,11%	8,33%	6,48%	5,55%	4,63%
Humedad Relativa	4,63%	3,70%	3,70%	3,70%	2,78%
Presión Atmosférica	10,65%	9,72%	9,26%	9,26%	8,33%
Velocidad del Viento	1,38%	0,93%	0,93%	0,93%	0,46%
Dirección del Viento	0,93%	0,93%	0,93%	0,93%	0,46%
Precipitación	0,00%	0,00%	0,00%	0,00%	0,00%
Nivel del lago	1,85%	1,39%	1,39%	1,39%	1,39%

Nota: todos los sitios presentan la misma proporción de faltantes

Tabla 15 Porcentaje de valores faltantes en las series mensuales del período de estudio (1999-2017) de las variables directamente relacionadas a calidad del agua

Sitio y perfil	Porcentaje de faltantes de la serie		
	Secchi	PT	Clorofila-a
C1	12,50%	11,57%	12,04%
C2	12,50%	17,13%	16,67%
C4	12,50%	23,15%	22,69%
C5	12,50%	13,43%	14,81%
TAC1	8,33%	6,48%	7,41%
TAC2	8,33%	15,74%	15,74%
TAC4	8,33%	15,74%	17,59%
TAC5	8,33%	12,96%	13,89%
DCQ1	19,91%	15,27%	16,20%
DSA1	16,20%	11,11%	12,04%

A efectos de caracterizar la serie en términos de tendencia, estacionalidad y ciclos, se definieron los parámetros del método STL (siglas en inglés del nombre completo en español *Descomposición de Estacionalidad y Tendencia usando Loess*, descrito en 2.6.1). Así se configuró una ventana de 13 meses para la estacionalidad y 25 para la tendencia. El nivel de significatividad en el test de tendencia de Mann-Kendall y la pendiente de Sen se fijó a 95%.

El análisis en esta sección, se remitirá exclusivamente a las variables directamente ligadas con la determinación de la calidad del agua, por un lado, y a las variables meteorológicas e hidrológicas, por otro, que se utilizarán para el modelado en la sección 3.2. Para un análisis general de las restantes variables se remite al lector al Anexo III. Vale aclarar que se encuentran allí representadas todas las variables que presentan menos de 20% de valores faltantes, no comportan irregularidades y aportan información discriminante.

Se aclara que, a lo largo de este análisis, el uso de la palabra “ciclo” se refiere exclusivamente a lo definición en la sección 2.5.2 que corresponde al concepto en el campo de las series de tiempo, distinta a la definición que pudiera tener en el ámbito de la hidrología.

3.1.1 Análisis de las series de tiempo del lago

En este apartado se analizarán las siguientes variables de observación directa: transparencia (Secchi), fósforo total y clorofila-a; así como las correspondientes variables calculadas siguiendo la fórmula del IET de Carlson (ver 2.6.5): IET a partir de Secchi, IET a partir de fósforo total e IET a partir de clorofila-a.

La Fig. 17 muestra las series de tiempo correspondientes a la variable transparencia para los cuatro sitios muestreados. Por tratarse de una variable que se determina en función de la profundidad, no hay en este caso dato a nivel perfil (a distintas profundidades), simplemente por conveniencia se asocia con el perfil (o profundidad) subsuperficial a la hora de graficar. Comparando las series se pueden observar mayor cantidad de picos en la primera mitad del período en DCQ, y en general valores más o menos constantes más altos en los sitios del centro que en los de desembocadura. TAC1 presenta valores más elevados de Secchi (es decir más transparencia, mejor calidad de agua), seguida de Centro y luego las desembocaduras (siendo DSA generalmente de menor calidad que DCQ).

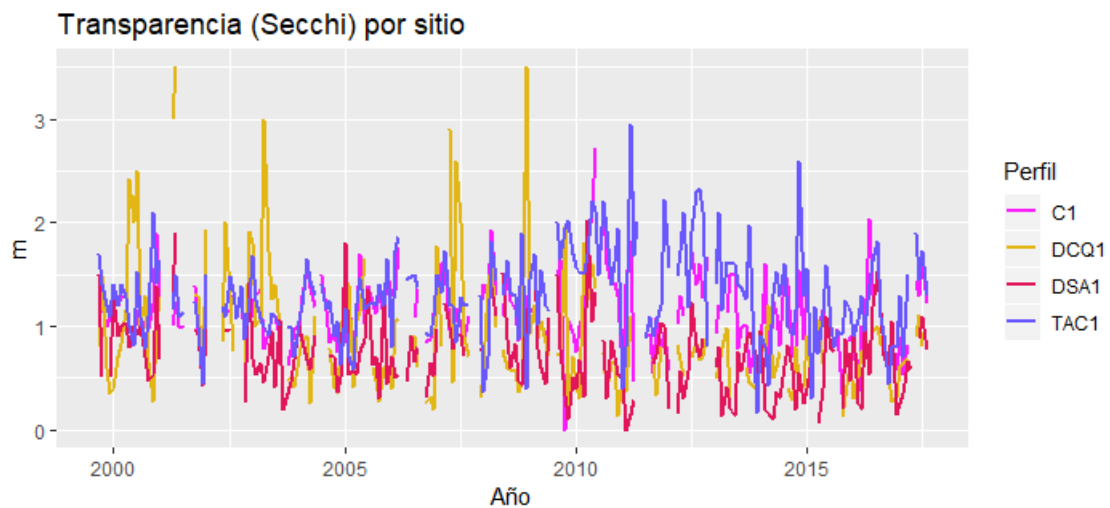


Fig. 17 Series de tiempo de transparencia por sitio (perfil subsuperficial)

Con respecto a la caracterización de la serie se observa lo siguiente, como puede apreciarse en la Fig. 18:

- Tendencia.** Comparando los cuatro sitios, se puede decir que en el caso del sitio Centro no se observa una tendencia clara en el gráfico, aunque sí podría hablarse de varios ciclos de corta duración (de forma aproximada): 2000-2003 (alza), 2005-2007 (alza), 2009-2011 (alza) y 2011-2013 (alza). En lo que se refiere a TAC, tampoco podría hablarse de tendencia, pero pareciera haber un ciclo más grande con pico en 2010 y que abarca prácticamente todo el período de estudio. En relación a las desembocaduras, se observa una tendencia hacia la baja a lo largo del período. En términos de ciclos, en DSA podrían mencionarse: 2000-2003 (alza), 2005-2010 (alza), 2011-2014 (alza) y 2014 en adelante (alza). Mientras que, en el caso de DCQ, únicamente uno: 1999-2003 (alza). Por medio del test de Mann-Kendal, se puede decir que existe evidencia suficiente para sostener que hay tendencia monotónica en los puntos de desembocadura, no así en los sitios centrales del lago. En lo que se refiere a la pendiente de Sen, la misma es de -0,0318 en el caso del sitio DCQ y -0,0165 en el caso de DSA. Es decir, se espera un descenso en la serie (o un empeoramiento de la calidad del agua) del orden de 3,18cm por año en un caso, y 1,65cm en el otro, a lo largo de los 18 años analizados.
- Estacionalidad.** En todos los sitios se observa estacionalidad, con periodicidad de aproximadamente 12 meses. En el caso de los sitios Centro y TAC, el pico más alto se produce a fines del verano con cambio a principios del invierno y mediados del otoño respectivamente en 2009 y 2010; mientras que la depresión más acentuada coincide con mediados de la primavera, con cambios a inicios del verano en 2009 en el caso del sitio Centro, y a mediados de la primavera en 2005 y mediados del verano en 2011 en TAC. En lo que respecta a la varianza hay un cambio a lo largo del período, al comienzo y hasta 2008/2009 es alta, luego se reduce y vuelve a crecer a partir de 2011 aproximadamente. En el caso de DSA la varianza de la estacionalidad se mantiene mas o menos constante a lo largo del período, pero presenta un cambio de patrón coincidente aproximadamente con el señalado para los sitios centrales, pero aquí se alcanzaba el pico a fines del otoño y la depresión a mediados de la primavera, para luego producirse a principios del invierno a partir de 2014 y a fines del verano a partir de 2009

respectivamente. En el caso de DCQ, por último se observa una reducción paulatina y constante de la varianza en la estacionalidad a lo largo del período, el patrón se presenta del siguiente modo: el pico a principios del invierno y la depresión a mediados de la primavera, para luego producirse a fines del verano a partir de 2008 y a fines de la primavera a partir de 2010 respectivamente.

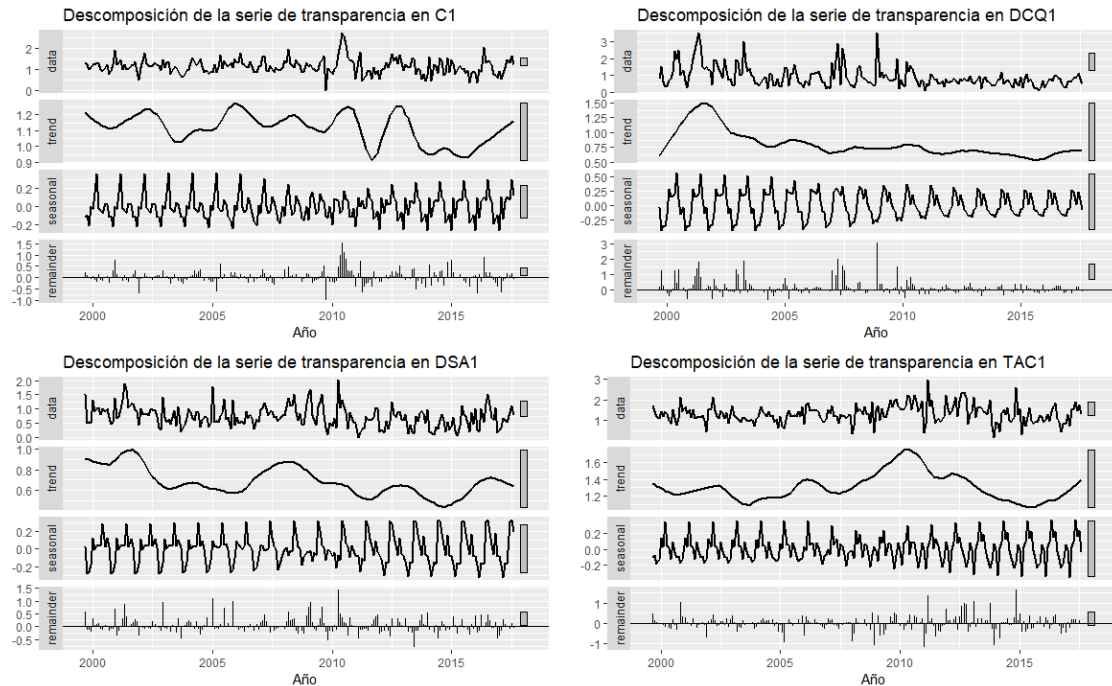


Fig. 18 Descomposición de las series de tiempo de transparencia por sitio: datos originales en *m* (*data*), parte de tendencia (*trend*), parte estacional (*seasonal*) y resto (*remainder*)

Asimismo, pueden analizarse las series en términos de correlación en el tiempo de la siguiente manera:

- **Autocorrelación.** En lo que respecta a la autocorrelación de las series, como se puede ver en la Fig. 19, no se observa en sitio Centro ni TAC autocorrelación significativa con ningún retardo; mientras que se observa correlación negativa significativa con retardos 4 y 5 en DSA1, y con retardo 5 en DCQ1, además de correlación positiva con retardo 1 en este último sitio.

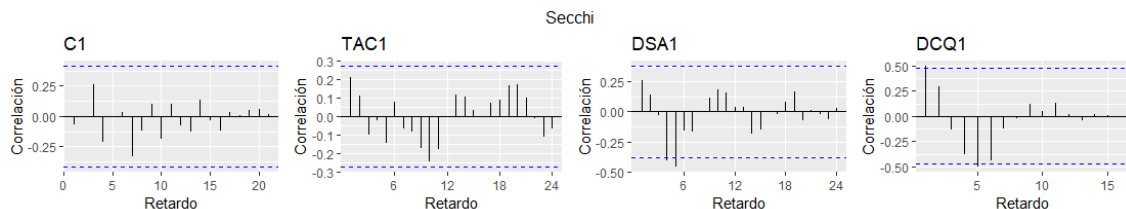


Fig. 19 Autocorrelación de las series de tiempo de transparencia del agua (Secchi)

- **Autocorrelación parcial.** Con respecto a la autocorrelación parcial, como se aprecia en la Fig. 20, nuevamente se observa significativa solo en las desembocaduras: negativa con retardo 4 en DSA y positiva con retardo 1 en DCQ.

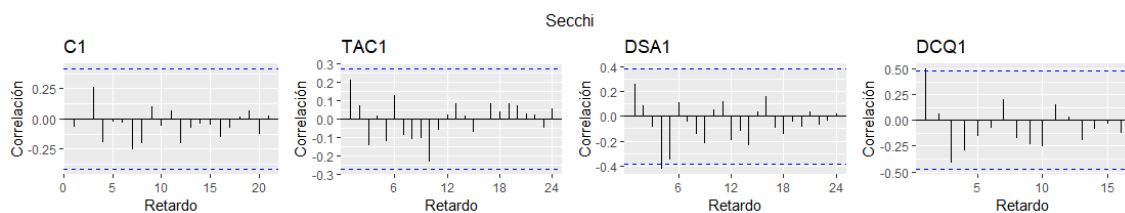


Fig. 20 Autocorrelación parcial de las series de tiempo de transparencia del agua (Secchi)

La Fig. 21 muestra las series de tiempo de fósforo total para cada sitio y perfil. Asimismo, se observan picos en DSA muy por encima de los que se producen en otros sitios. También se observan valores altos especialmente en TAC para varios perfiles. En general, en la segunda mitad del periodo de estudio se observan picos marcadamente más importantes respecto de la primera mitad (2000-2010) con valores que superan ampliamente y de manera periódica los 500ug/L.

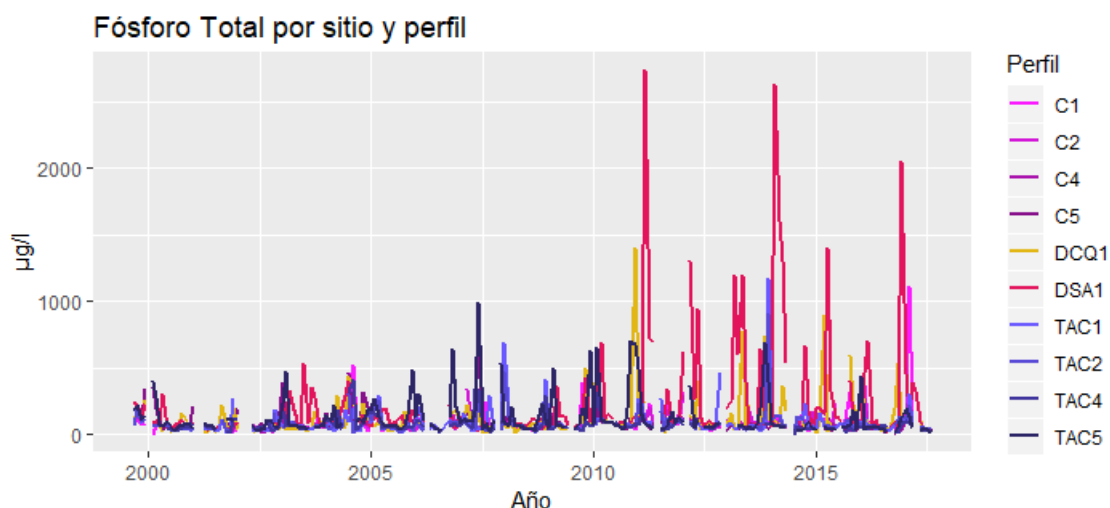


Fig. 21 Series de tiempo de fósforo total por sitio y perfil

En lo que respecta a la categorización de las series, a partir de la Fig. 22, Fig. 23 y Fig. 24, se tiene lo siguiente:

- **Tendencia.** En lo que respecta al sitio Centro puede observarse que, a nivel subsuperficial (C1) no pareciera haber tendencia a simple vista, pero si varios ciclos más o menos definidos: 1999-2005 (baja) y 2009-2012 (alza). En la zona fótica (C2) se observa una tendencia creciente, y los ciclos son menos definidos. En la zona intermedia (C4) nuevamente se tienen ciclos bien definidos como en el caso del nivel subsuperficial, sin tendencia definida: 1999-2004 (alza) y 2009-2013 (alza). Por último, en el caso del fondo del sitio Centro (C5) se observa una tendencia decreciente sin ciclos bien definidos. De acuerdo al test de Mann-Kendall, existe tendencia monótona en C1, C2 y C5, con pendientes de Sen respectivas: 1,8730, 2,1251 y -2,3210.

En lo que respecta a TAC, el patrón es similar al observado para sitio Centro. A nivel subsuperficial (TAC1) no se observa una tendencia clara, pero sí ciclos: 1999-2005 (baja), 2005-2008 (alza) y 2009-2012 (alza). En el caso de la zona fótica (TAC2), el patrón es similar a la zona subsuperficial, coincidiendo los ciclos (a excepción del último) y la

tendencia más o menos constante. En la zona intermedia (TAC4), tampoco se observa tendencia y se tienen varios ciclos: 2000-2002 (alza), 2003-2006 (alza), 2007-2009 (alza) y 2009-2011 (alza). Por último, en el fondo (TAC5) no se observan ciclos, pero sí tendencia decreciente. De acuerdo al test de Mann-Kendall, existe tendencia monótona en TAC5, siendo la pendiente de Sen -1,7450.

Por último, en las desembocaduras no se observa tendencia a simple vista. En el caso de la desembocadura del río Cosquín (DCQ1) se observan los siguientes ciclos: 1999-2004 (baja), 2005-2008 (alza), 2009-2013 (alza) y 2013-2015 (alza). En el caso de la desembocadura del río San Antonio, se observan solo dos ciclos: 1999-2004 (baja) y 2014-2016 (alza). De acuerdo al test de Mann-Kendall, existe tendencia en ambas desembocaduras, siendo las pendientes de Sen, en DCQ1 2,1711 y en DSA1 3.

- **Estacionalidad.** En lo que se refiere al sitio Centro, se observa que los picos se producen durante el verano: en C1 a mediados del verano con cambio a comienzos del verano a partir de 2006; en C2 a mediados del verano con cambio al comienzo en 2007 y a fines del verano en 2009; en C4 a comienzos del verano con cambio a fines del verano en 2008 y en C5 a mediados del verano. En lo que respecta a las depresiones, se producen durante el invierno: en C1 a mediados del invierno con cambio a fines del invierno en 2006; en C2 a mediados del invierno con cambio a fines del invierno en 2009, nuevamente a mediados en 2011 y a comienzos del invierno en 2016; en C4 a fines del invierno; y en C5 a mediados del invierno, con cambio a fines del invierno en 2006. En cuanto a la varianza, aumenta paulatinamente en C1 y C2, mientras que se reduce en C4 y C5 (más intensamente a partir de 2011). El cambio de patrón es más evidente en C1 y C2 como se observó según la variación del régimen de picos y depresiones.

En el caso de TAC, el patrón es similar al Centro, tanto para los picos como para las depresiones. Los picos ocurren en verano: en TAC1 a mediados del verano, en TAC2 es más variable, a fines de la primavera con cambio a mediados de la primavera en 2002, cambio a fines del verano en 2011 y desde el 2014 a comienzos del verano; en TAC4 a comienzos del verano con cambio a mediados del verano en 2003 y nuevamente cambio a fines del verano en 2008; mientras que en TAC5 ocurren a mediados del verano. En cuanto a las depresiones: en TAC1 suceden a fines del invierno, y a partir de 2011 a mediados del invierno; en TAC2 mayormente a mediados del invierno; en TAC4 a fines del invierno; y en TAC 5 a mediados del invierno, luego a fines del invierno desde 2006 y nuevamente a mediados del invierno cambiando en 2013. En lo que se refiere a la varianza, aumenta en TAC1, en TAC2 fluctúa y se reduce en TAC4 y TAC5 (más intensamente a partir de 2007, luego se estabiliza a partir de 2010). El cambio de patrón es más evidente en TAC1, TAC2 y TAC4 coincidente con lo señalado anteriormente.

En relación a las desembocaduras, en lo que se refiere a los picos, en DCQ1 hay más variabilidad y se producen a principios de la primavera, con cambio a mediados de la primavera en 2008 y cambio nuevamente a mediados del verano en 2010; mientras que en DSA1 se producen a mediados del verano y desde 2013 a fines del verano. En lo que respecta a las depresiones, en DCQ1 se producen a principios del invierno y en DSA1 a fines del otoño con cambio en 2010 a principios del invierno. Hay un cambio en la varianza en DSA1, acrecentándose a partir de 2013, y en DCQ1 crece desde el inicio y se reduce para estabilizarse a partir de 2010. Es interesante observar como el cambio en la varianza en DSA1 es casi inverso al que se produce en DCQ1.

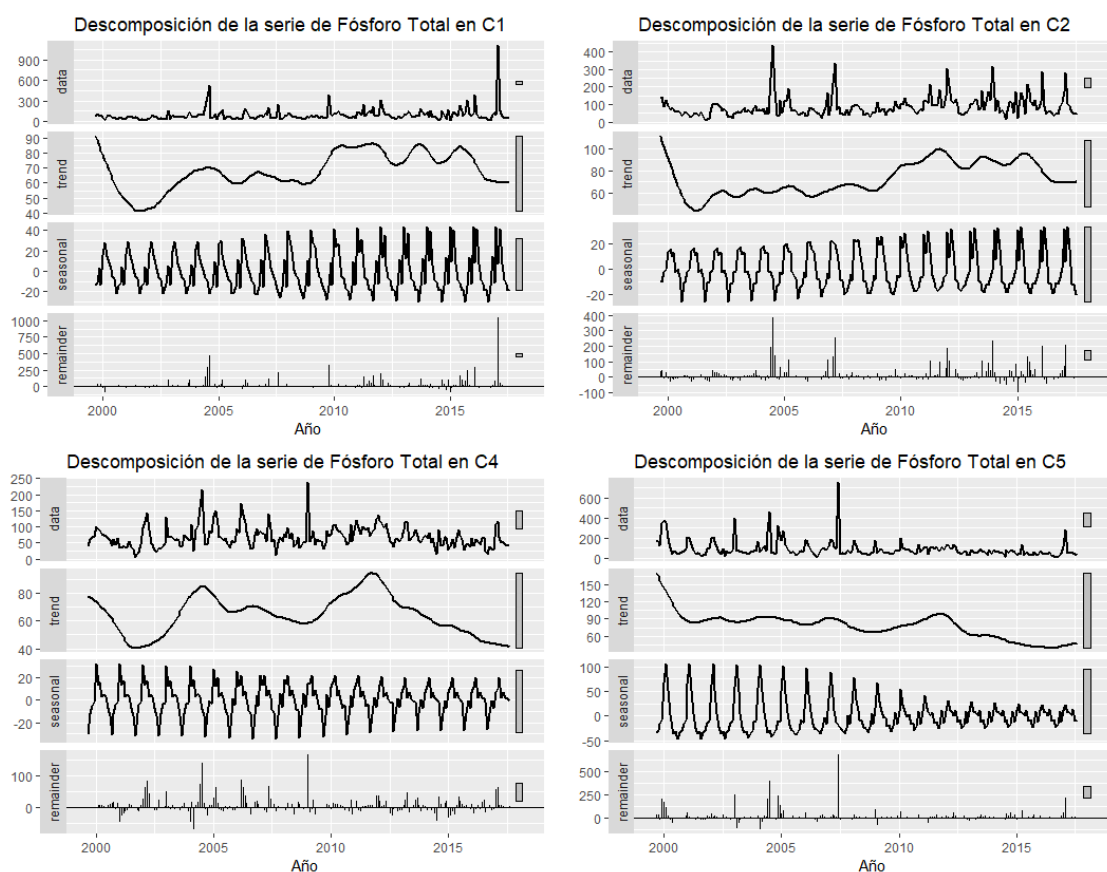


Fig. 22 Descomposición de las series de tiempo de fósforo total en sitio Centro: datos originales en $\mu\text{g/l}$ (*data*), parte de tendencia (*trend*), parte estacional (*seasonal*) y resto (*remainder*)

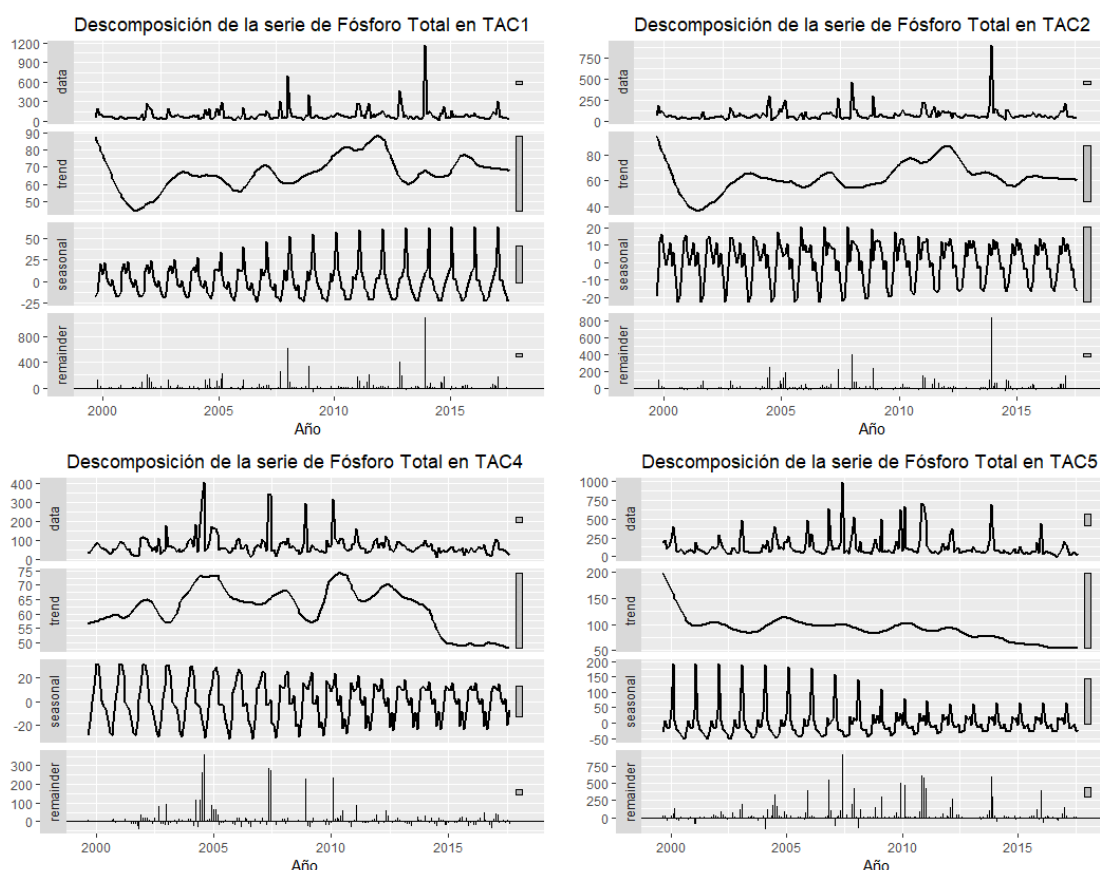


Fig. 23 Descomposición de las series de tiempo de fósforo total en sitio TAC: datos originales en $\mu\text{g/l}$ (*data*), parte de tendencia (*trend*), parte estacional (*seasonal*) y resto (*remainder*)

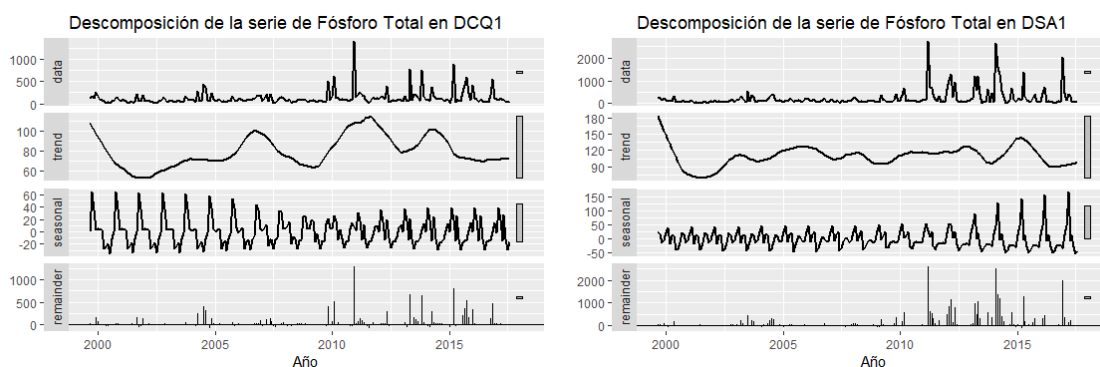


Fig. 24 Descomposición de las series de tiempo de fósforo total en desembocadura del Río Cosquín y en desembocadura del Río San Antonio: datos originales en $\mu\text{g/l}$ (*data*), parte de tendencia (*trend*), parte estacional (*seasonal*) y resto (*remainder*)

Con respecto a la autocorrelación y autocorrelación parcial, se observa lo siguiente:

- **Autocorrelación.** Se observa significativa positiva con retardo 4 en DCQ1, como se puede ver en la Fig. 25.

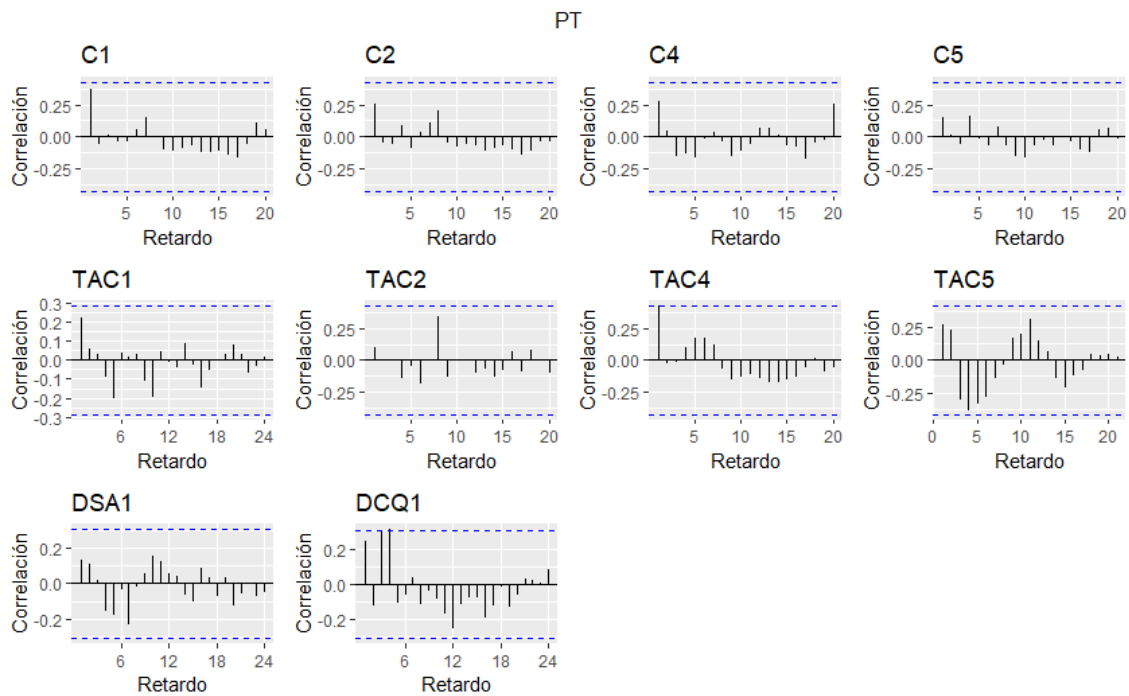


Fig. 25 Autocorrelación de las series de fósforo total por sitio y perfil

- **Autocorrelación parcial.** Como se muestra en la Fig. 26, se observa significativa negativa con retardo 3 en TAC5 y positiva con retardo 3 en DCQ1.

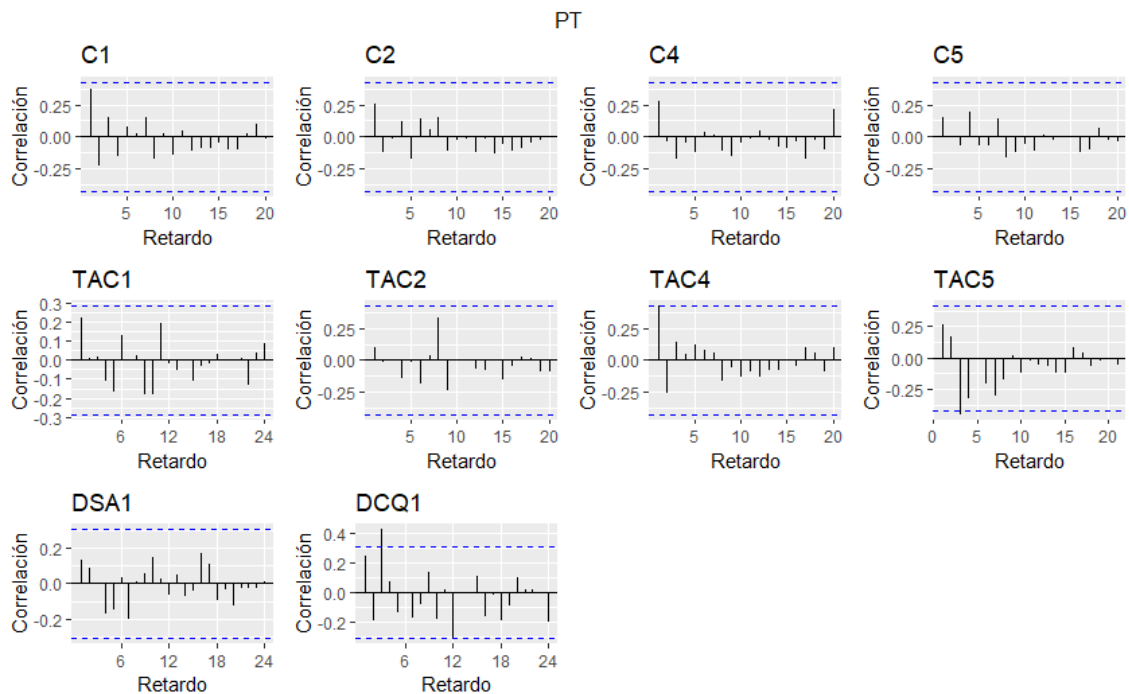


Fig. 26 Autocorrelación parcial de las series de fósforo total por sitio y por perfil

Las series correspondientes a la última de las variables de calidad del agua del lago, clorofila-a, se encuentran representadas para cada sitio y perfil en la Fig. 27. Como se puede notar, el

comportamiento es similar, a grandes rasgos, a las de fósforo total, en términos de grandes picos de concentración en la desembocadura del río San Antonio con posterioridad a 2010, presentando las demás series una dinámica más o menos regular (con algunos picos más altos en la desembocadura del río Cosquín y TAC a partir del mismo año).

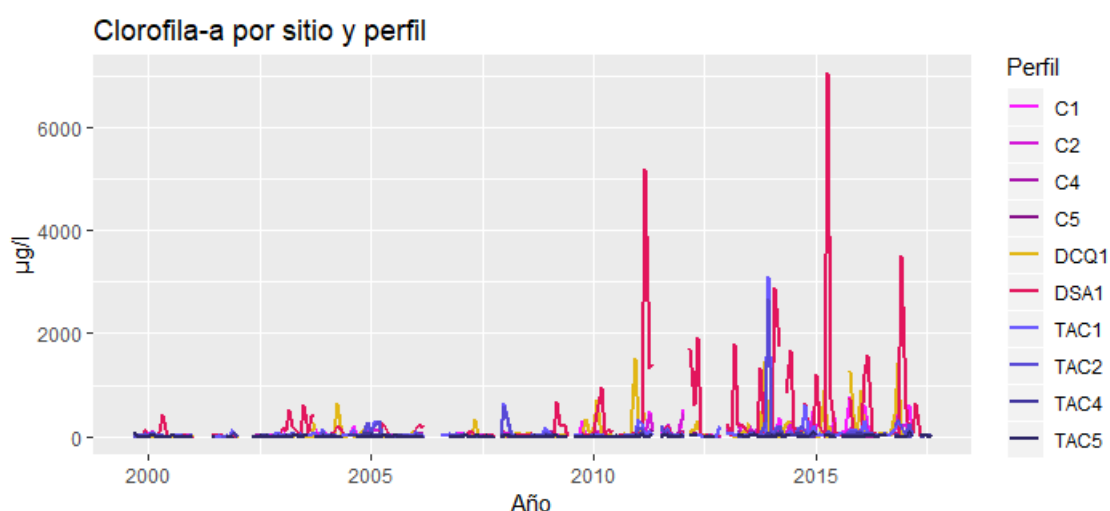


Fig. 27 Series de tiempo de clorofila-a por sitio y perfil

Con respecto a la caracterización de las series, la Fig. 28, Fig. 29 y Fig. 30 muestran lo siguiente:

- **Tendencia.** En el caso del sitio Centro, en C1 no se observa una tendencia creciente o decreciente a simple vista, pero sí un ciclo que comienza en 2011 (alza) y al cierre del período aún no finaliza. El mismo patrón se observa en C2. En el caso de C4, se observa el mismo fenómeno, pero desplazado en el tiempo, el ciclo parece comenzar en 2009 con tendencia a la finalización en 2017 (la serie se corta en agosto). Por último, en C5 vuelve a repetirse el patrón, pero se trata de un fenómeno de más corta duración, puesto que comienza aproximadamente en 2009 y finaliza en 2015. El test de Mann-Kendall indica que existe evidencia suficiente de que hay tendencia en los cuatro perfiles. El estimador de la pendiente de Sen es 3,3241 en C1, 3,1488 en C2, 1,2094 en C4 y 0,2634 en C5; es decir, la pendiente disminuye en profundidad.

En el caso del sitio TAC, el escenario es diferente. En TAC1, tampoco se observa una tendencia clara creciente o decreciente, pero aparecen dos ciclos, uno en el período 2000-2008 (alza) y otro 2010-2017 (alza, no terminado). En cuanto a TAC2, el patrón es similar a TAC1, pero los ciclos son algo más tenues: 2000-2005 (alza) y 2010-2017 (alza, no terminado), también se podría pensar otro ciclo como 1999-2003 (baja). Ya en TAC4 se observa únicamente el segundo de los ciclos informados en TAC1 y TAC2 y es coincidente con el observado en C1 y C2. Por último, en TAC5 aparecen nuevamente dos ciclos marcados, el primero 2000-2006 (alza) y el segundo 2012-2017 (alza, sin terminar); y al igual que en TAC2 se podría pensar un tercer ciclo como 1999-2003 (baja). Con respecto al test de Mann-Kendall, existe evidencia de tendencia en TAC1, TAC2 y TAC4, no así en TAC5. La pendiente de Sen, es respectivamente 1, 1 y 0,5355.

En las desembocaduras, no parece haber tendencia clara en DCQ1, pero sí en DSA1 (al aumento). En el primer caso se observa un ciclo bien definido en el período 2012-2017 (alza). En el segundo caso, también se observa un ciclo, pero es más difuso, en el período

2012-2016 (alza). Con respecto al test de Mann-Kendall, existe evidencia de tendencia en los dos sitios, siendo la pendiente de Sen en DCQ1 2,5 y en DSA1 3,2768.

- **Estacionalidad.** Observando el sitio Centro, se encuentra estacionalidad anual, con períodos más o menos bien definidos en C1 y C2, mientras que se vuelve más difuso en C4 y C5; por otra parte, el patrón general es distinto en C5 comparado con los demás, está invertido. Los picos se producen en verano en C1, C2 y C4, con el siguiente patrón: a principios del verano en C1; a mediados del verano en C2 con cambio a principios del verano en 2010; y a mediados del verano en C4. Mientras que las depresiones ocurren mayormente en otoño-invierno en la superficie y también en primavera a más profundidad, de la siguiente manera: en C1 a fines del otoño con cambio a principios del invierno en 2006; en C2 a fines del otoño con cambio a mediados de la primavera en 2005 y a mediados del otoño a partir de 2012; y en C4 a mediados de la primavera con cambio a mediados del otoño en 2010. En cuanto al caso de C5, los picos se producen a mediados del otoño con cambio a principios del invierno en 2014; mientras que las depresiones se producen a comienzos del verano. En lo que respecta a la variabilidad, se observa varianza creciente en C1, C2, C4 (en este caso a partir de 2007) y C5 (a partir de 2010). Los cambios de patrón son más bien suaves en este último sitio.

En lo que respecta a TAC, el patrón es más irregular en general que en el sitio Centro. Los picos se producen de la siguiente manera: en TAC1 a comienzos de la primavera con cambio a fines de la primavera en 2002, principios del verano en 2006 y mediados del verano en 2010, es decir existe una transición hacia épocas más cálidas; por otro lado, en TAC2 ocurren a principios del verano con cambio a mediados del verano en 2009; en TAC4 a mediados del invierno con cambio a mediados del verano en 2007; y en TAC5 a mediados del verano con cambio a fines del otoño en 2012. En cuanto a las depresiones se encuentra lo siguiente: en TAC1 a principios del invierno con cambio a mediados del otoño en 2007 y principios del otoño en 2011; en TAC2 a fines del otoño con cambio a mediados del otoño en 2011; en TAC4 a mediados de la primavera; y en TAC5 mayormente a mediados de la primavera. La varianza es creciente en TAC1 y TAC4 (a partir de 2008), en el caso de TAC2 es fluctuante pero aproximadamente constante, y en TAC5 decrece hasta 2011 y luego vuelve a crecer. Los cambios de patrón son más evidentes en TAC1, TAC2 y TAC5, consistente con lo observado en cuanto al régimen de picos y depresiones.

Por último, en lo que respecta a las desembocaduras, se observa cambio de varianza en DCQ1 y varianza constante de la estacionalidad en DSA1. En el primer caso, la varianza se incrementa levemente a partir de 2011 y el patrón cambia en 2008. En cuanto a los picos, en DCQ1 se producen mayormente a mediados del otoño, con cambio a mediados del verano en 2011; mientras que en DSA1 se producen a principios del otoño. Por otro lado, en cuanto a las depresiones es mucho más variable, en DCQ1 se producen a fines del invierno con cambio a fines del otoño en 2005 y fines del verano en 2008; mientras que en DSA1 se producen a fines del invierno con cambio en 2006 a principios del invierno, nuevamente cambio mayormente a fines del invierno en 2008 y por último cambio a fines del verano en 2011. El cambio de patrón en ambos es consistente con lo observado en cuanto a la irregularidad de picos y depresiones, siendo más evidente en DCQ1.

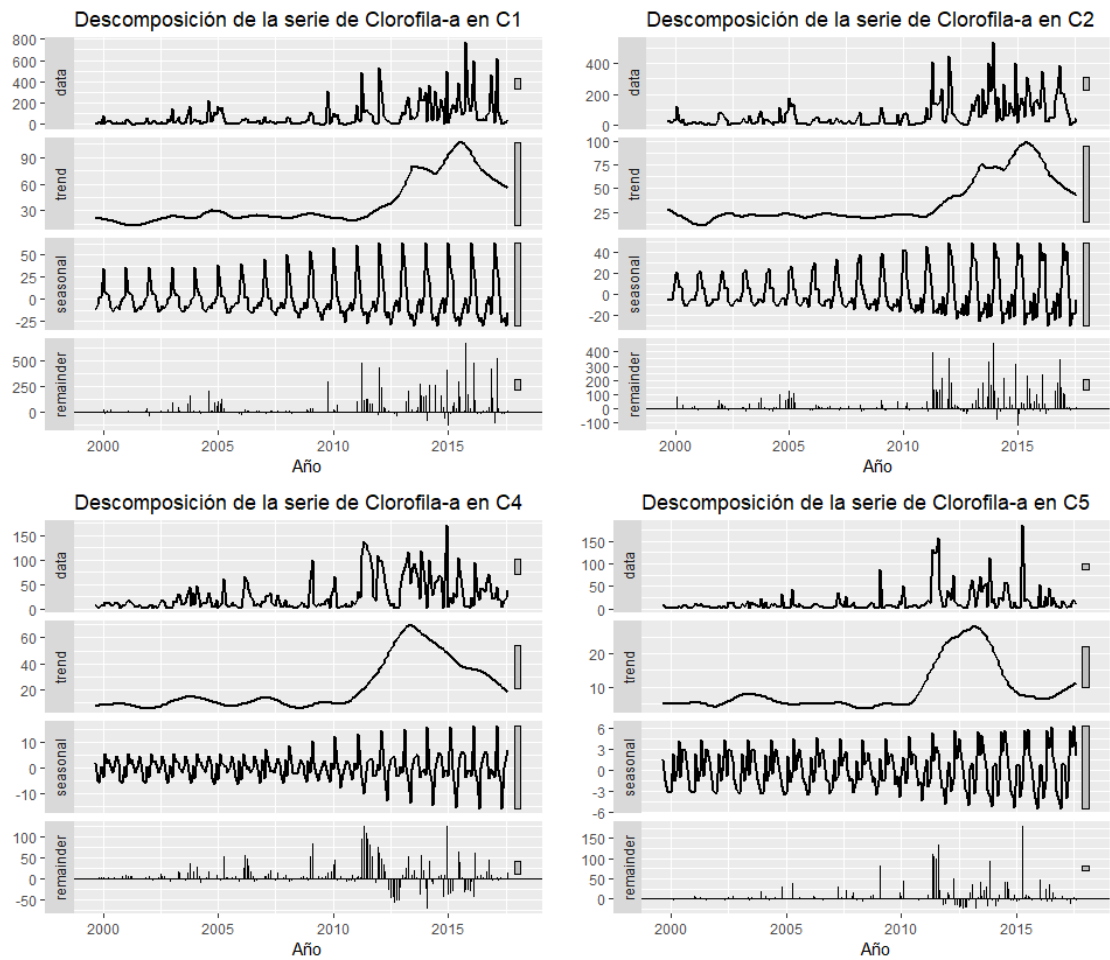


Fig. 28 Descomposición de las series de tiempo de clorofila-a en sitio Centro: datos originales en $\mu\text{g/l}$ (*data*), parte de tendencia (*trend*), parte estacional (*seasonal*) y resto (*remainder*)

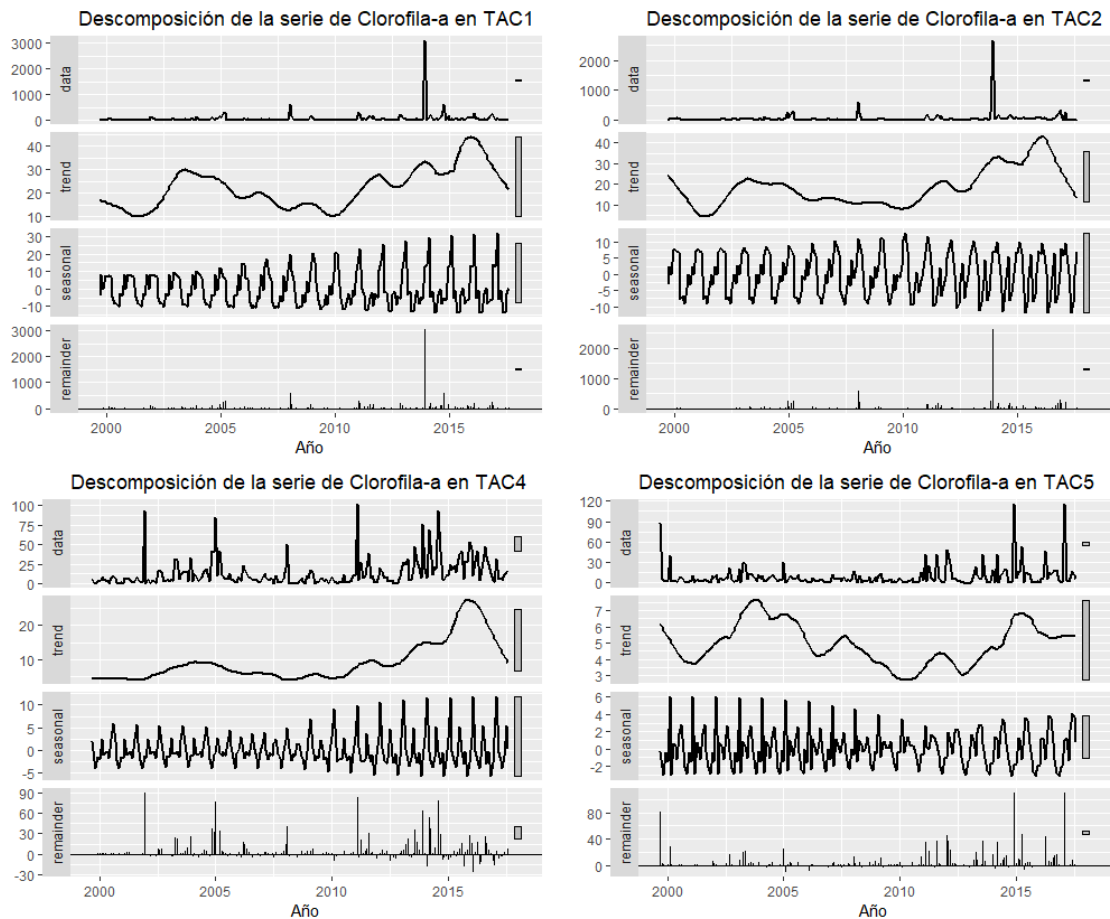


Fig. 29 Descomposición de las series de tiempo de clorofila-a en sitio TAC: datos originales en $\mu\text{g/l}$ (*data*), parte de tendencia (*trend*), parte estacional (*seasonal*) y resto (*remainder*)

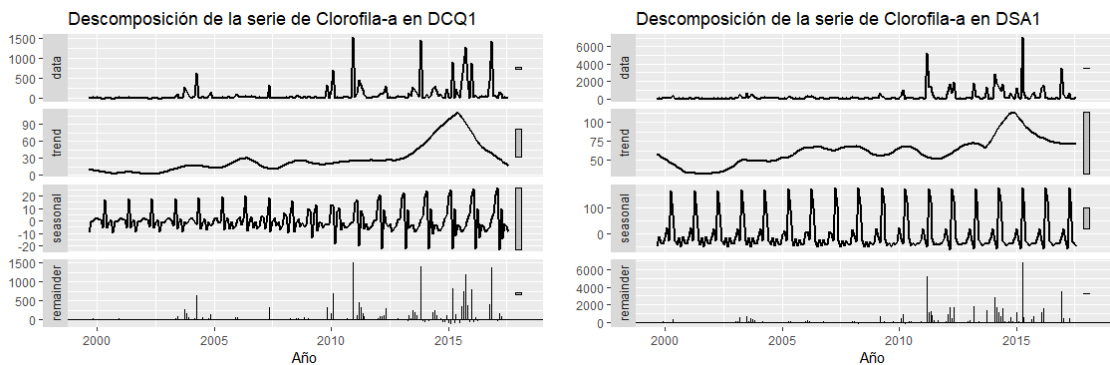


Fig. 30 Descomposición de las series de tiempo de clorofila-a en desembocadura del Río Cosquín y en desembocadura del Río San Antonio: datos originales en $\mu\text{g/l}$ (*data*), parte de tendencia (*trend*), parte estacional (*seasonal*) y resto (*remainder*)

En relación a la correlación de las series de tiempo, se tiene lo siguiente:

- **Autocorrelación.** Se observa autocorrelación significativa positiva con retardo 1 en TAC2 y DSA1, y con retardo 2 en TAC5, como se puede ver en la Fig. 31.

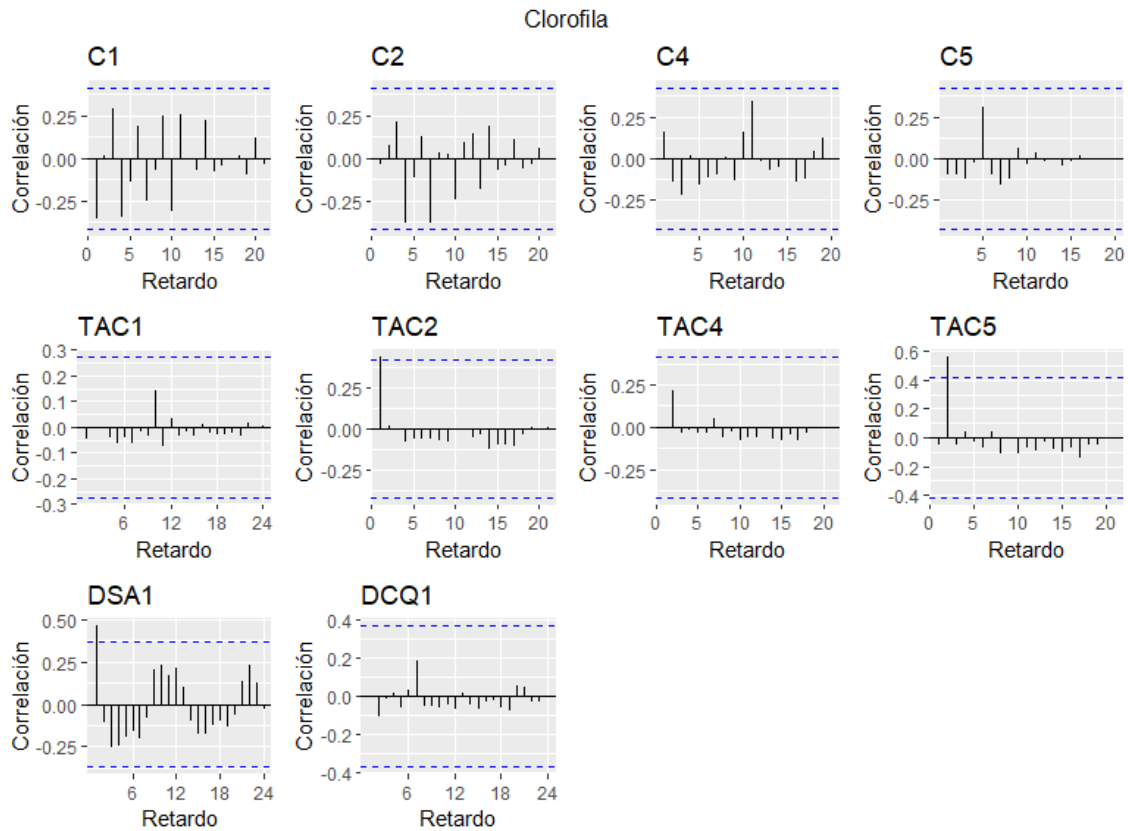


Fig. 31 Autocorrelación de las series de tiempo de clorofila-a por sitio y perfil

- **Autocorrelación parcial.** Como muestra la Fig. 32, se observa autocorrelación parcial significativa positiva con retardo 1 en TAC2 y DSA1, y con retardo 2 en TAC5. Por otro lado, se observa autocorrelación parcial significativa negativa con retardo 2 en DSA1.

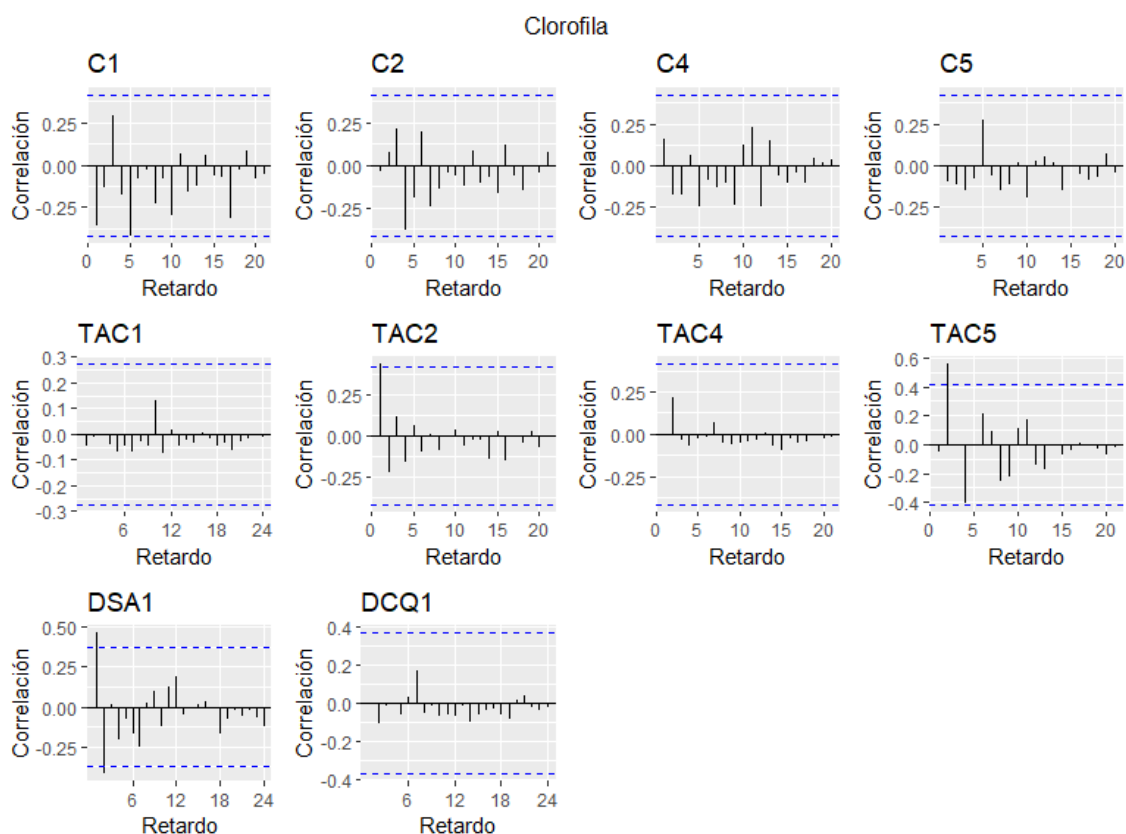


Fig. 32 Autocorrelación parcial de las series de tiempo de clorofila-a por sitio y perfil

Finalmente, se calculan las variables de IET basado en transparencia (Secchi), fósforo total y clorofila-a¹². En la Fig. 33, se pueden apreciar las series resultantes del cálculo del IET basado en la transparencia. Comparadas con las series originales, la varianza es mucho más conservada por el efecto del logaritmo.

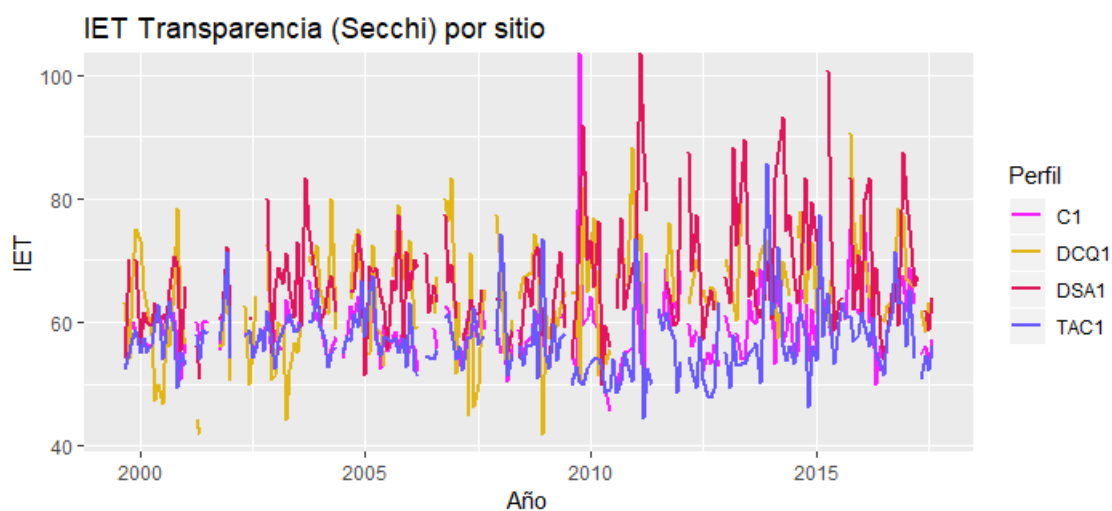


Fig. 33 Series de tiempo de IET basado en transparencia por sitio (perfil subsuperficial)

¹² Las series calculadas preservan las propiedades de interés para este análisis de cada una de las series observas, por lo que no se repite el análisis en estos casos.

Asimismo, en la Fig. 34, se observa el IET calculado a partir del fósforo total. Se puede apreciar como la varianza se reduce y se puede ver con mayor claridad como al comienzo del período no había demasiadas distinciones entre sitios y perfiles, pero sobre el final se observa como las desembocaduras presentan peores condiciones en términos de calidad del agua.

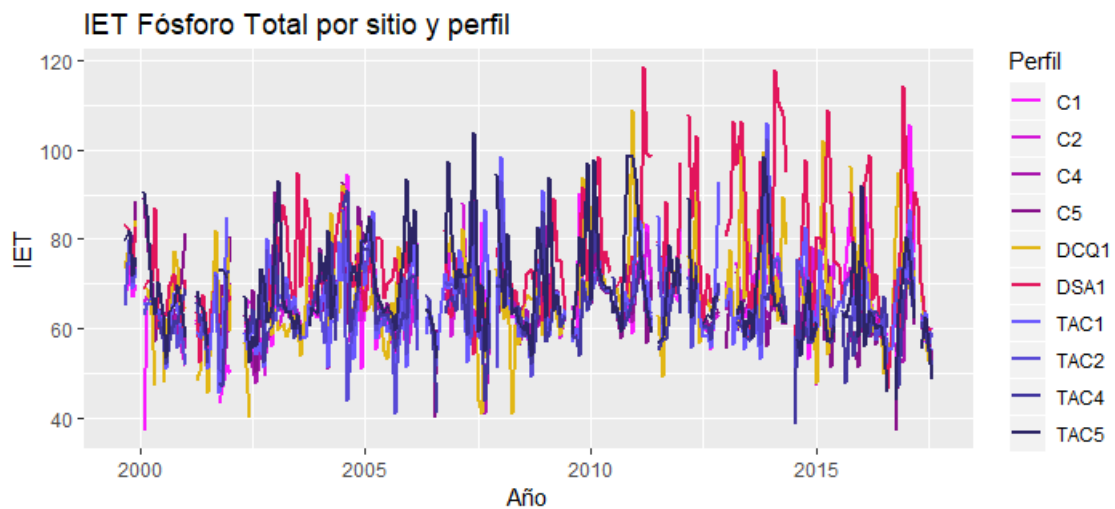


Fig. 34 Series de tiempo de IET basado en fósforo total por sitio y perfil

Por último, en la Fig. 35, se pueden observar las series resultantes del IET aplicado sobre la variable clorofila-a por sitio y perfil. Como cabía esperar por la aplicación de logaritmo la varianza está reducida respecto de la serie original y por otra parte se observa un claro fenómeno de valores más altos en los niveles subsuperficiales de cada sitio y más bajos próximo al fondo.

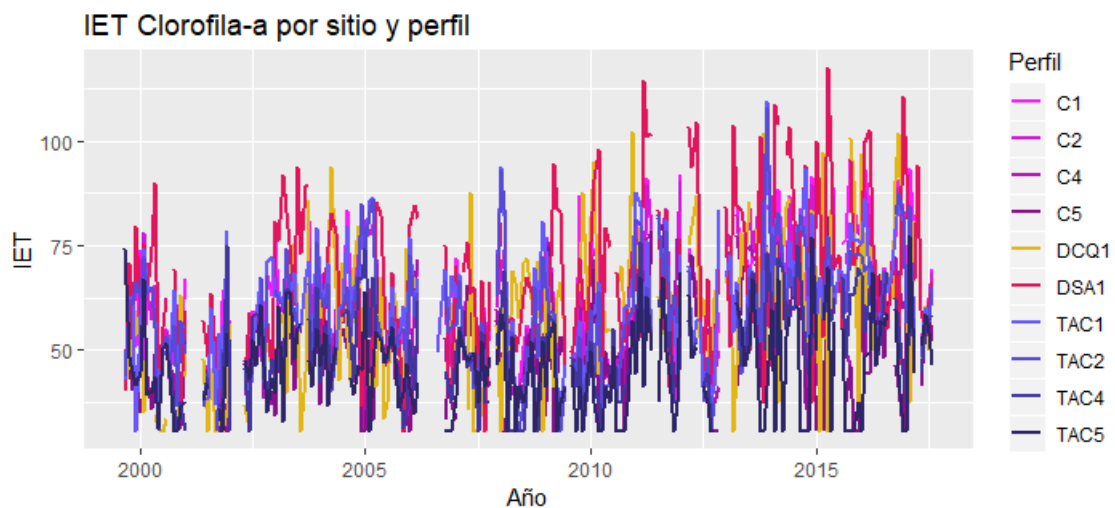


Fig. 35 Series de tiempo de IET basado en clorofila-a por sitio y perfil

En lo que respecta a estas variables, no se realiza el análisis de descomposición ni de autocorrelación y autocorrelación parcial, puesto que la transformación logarítmica que implica el IET preserva las propiedades de las series de tiempo de las variables originales en términos de la caracterización.

3.1.2 Análisis de las series de tiempo hidrológicas y meteorológicas

A continuación, se realiza el análisis de las series de tiempo de las variables hidrológicas y meteorológicas. En este caso, en general se observa poca variación entre los sitios de muestreo del lago, producto de la cercanía entre éstos y las estaciones meteorológicas. Los fenómenos climáticos son de largo alcance comparados con los fenómenos biológicos y químicos que se analizan en este estudio, de ahí la menor variabilidad comparando unos y otros. Las variables meteorológicas corresponden a los efectos en el medio ambiente en el sitio en cuestión, por lo que los valores por perfil son los mismos, y se asume que influyen de distinta manera en función de la profundidad, como se estudiará en lo sucesivo (ver sección 3.2). Por lo dicho anteriormente, los datos de variables meteorológicas se representan arbitrariamente para el perfil subsuperficial (común a todos los sitios) y el análisis se realiza para el sitio Centro. La cuestión relevante en este caso será el efecto en cada variable de los alcances definidos (1, 3, 7, 14 y 21 días, como se detalla en la sección 2.4.3.1).

Los aspectos más relevantes a considerar se resumen a continuación:

- Comparando las series de más corto alcance con las de más largo alcance se observa el efecto de suavizado esperado. A través del uso de estas series de distinto alcance, se cubre el espectro de posibilidades para tratar de abarcar desde fenómenos de acción inmediata a fenómenos de efecto retardado. En algunos casos como temperatura del aire el efecto del alcance es más marcado, mientras que en otros casos como humedad relativa es más atenuado.
- En el caso de radiación solar y presión atmosférica se dispone de una única estación que da origen a las series respectivas que se aplicarán a todos los sitios de muestreo. Las pocas discrepancias al inicio de la serie, para los sitios C1 y DCQ1 con respecto a TAC1, están relacionadas con las fechas de inicio elegidas (ver 2.4.3.1).
- Las series de tiempo de dirección del viento no tienen prácticamente variación de sitio a sitio, puesto que se ha empleado la moda para agregar los datos.

La Fig. 36 muestra las series de tiempo de temperatura por sitio para distintos alcances. No se observan irregularidades ni patrones globales.

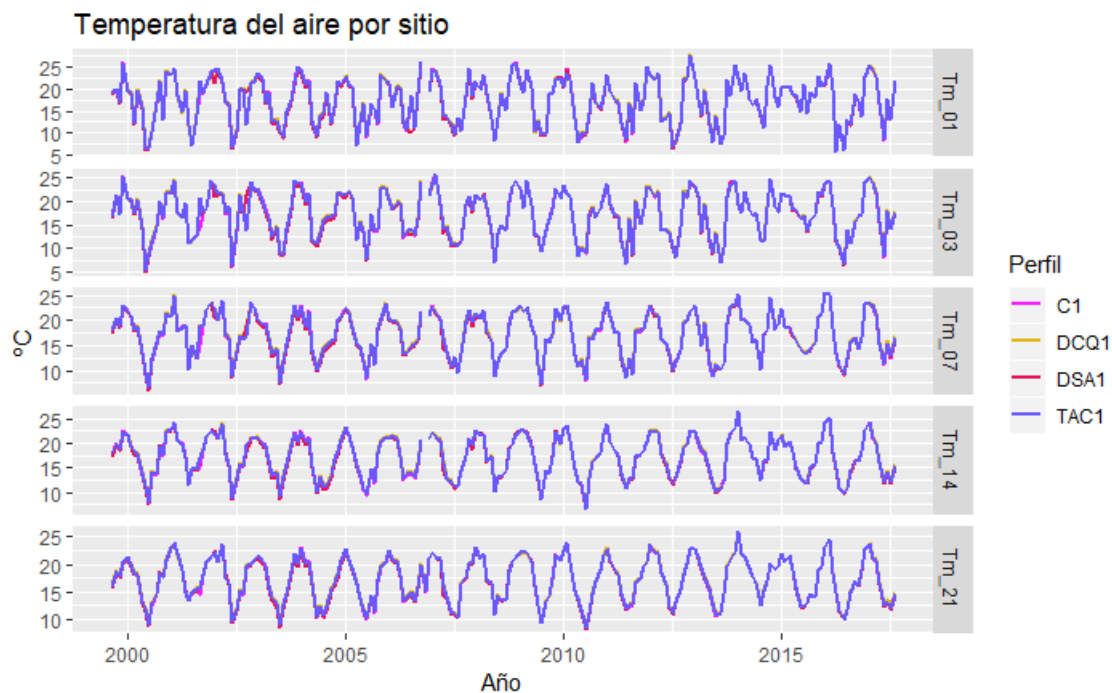


Fig. 36 Series de tiempo de temperatura del aire, para distintos alcances, por sitio

La Fig. 37, por otra parte, muestra los gráficos que permiten caracterizar las series de tiempo de temperatura para los distintos alcances en C1. Así se tiene lo siguiente:

- **Tendencia.** Para ningún alcance se observa una tendencia clara hacia el alza o la baja de temperatura. Sí se observan varios ciclos que cumplen el mismo patrón, pero son más intensos o más suaves y se desplazan levemente en el tiempo dependiendo del alcance (más pronto a mayor alcance). Así, podrían observarse tres ciclos: 2000-2004 (alza), 2007-2010 (alza) y 2013-2016 (alza). El test de Mann-Kendall confirma que no existe evidencia de tendencia para los cuatro primeros alcances. En el caso de 21 días sí, y la estimación de la pendiente de Sen es 0,0364.
- **Estacionalidad.** Confirmando lo que es conocido, en los gráficos de estacionalidad se observan picos coincidentes con los meses de verano y depresiones con los meses de invierno. A medida que se incrementa el alcance la estacionalidad está mejor definida, en los alcances 14 y 21 se producen los picos a principios del verano y las depresiones a principios del invierno, mientras que, en los alcances previos, es algo más variable, incluyendo mediados y fines de la primavera en el caso de los picos, y fines del otoño y mediados del invierno en cuanto a las depresiones. No se observan cambios relevantes en la varianza de la estacionalidad.

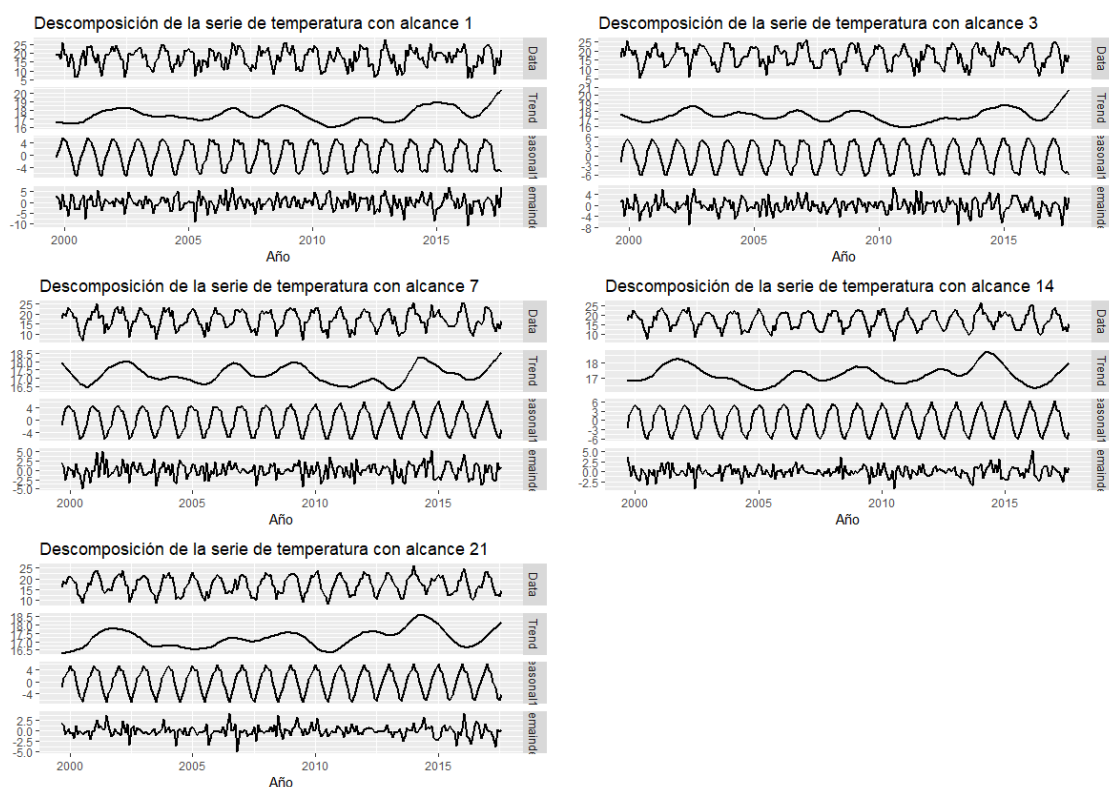


Fig. 37 Descomposición de las series de tiempo de temperatura en el sitio Centro para los distintos alcances: datos originales en °C (*data*), parte de tendencia (*trend*), parte estacional (*seasonal*) y resto (*remainder*)

En lo que respecta al análisis de correlación temporal de temperatura, se observa lo siguiente:

- **Autocorrelación.** Como se puede observar en la Fig. 38, la autocorrelación es significativa excepto en los retardos de transición (3, 10, 15 y 21) en todos los retardos analizados, siendo perfectamente estacional: máxima y positiva con retardos 1, 12 y 24 y mínima y negativa con retardos 6 y 18.

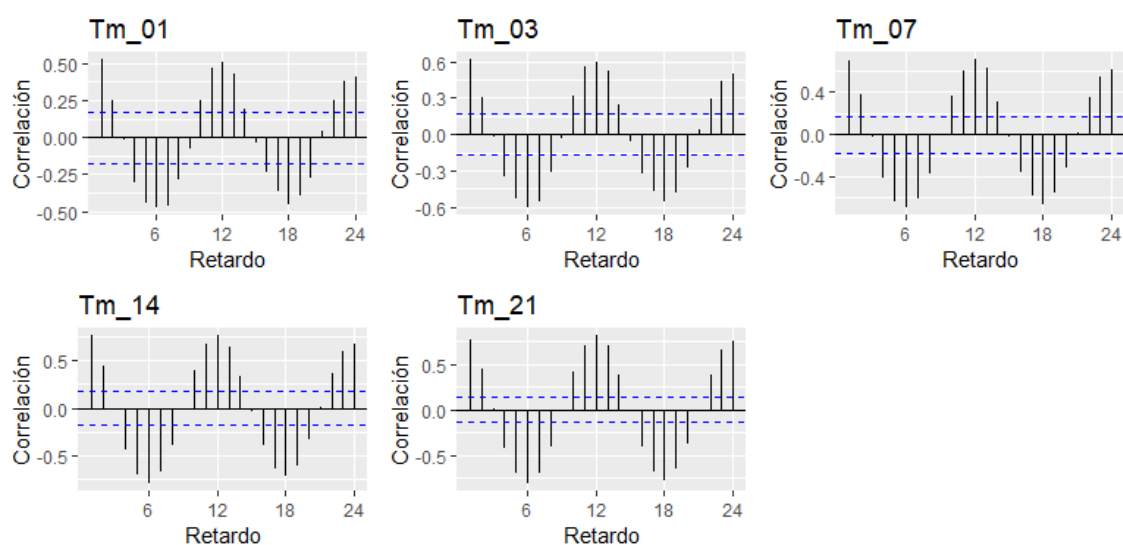


Fig. 38 Autocorrelación de las series de tiempo de temperatura en el sitio Centro por alcance

- **Autocorrelación parcial.** En este caso, como se ve en la Fig. 39, se observa autocorrelación parcial significativa y negativa con retardo 4 en todos los alcances, 7 en alcances 1, 3 y 7, y 2 a 4 en alcances 14 y 21; y significativa positiva con retardo 10 en los alcances 14 y 21; además de otras de menor relevancia.

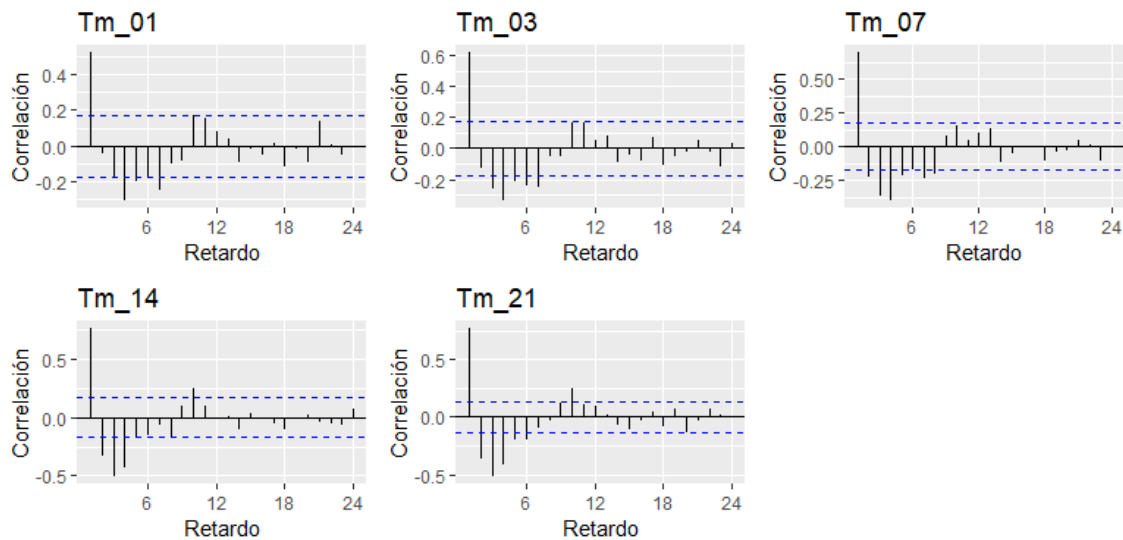


Fig. 39 Autocorrelación parcial de las series de tiempo de temperatura en el sitio Centro por alcance

En la Fig. 40 se observan las series temporales de radiación solar para distintos alcances. Comparadas con las series de temperatura muestran mayor variabilidad, volviéndose más regular a medida que el alcance es mayor.

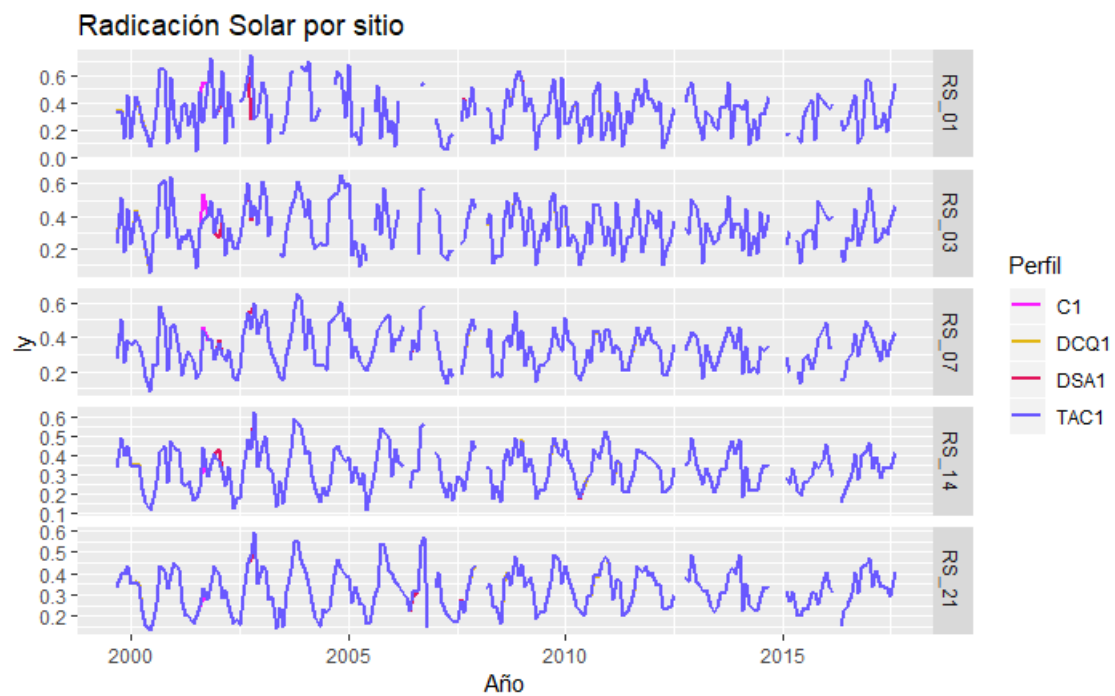


Fig. 40 Series de tiempo de radiación solar, para distintos alcances, por sitio

Al mismo tiempo, la Fig. 41 muestra la descomposición de las series en términos de tendencia y estacionalidad, lo que permite hacer el siguiente análisis:

- **Tendencia.** No se observa tendencia creciente ni decreciente evidente, en ningún alcance. En lo que respecta a ciclos, no pareciera haber para el primer alcance, mientras que para 3, 7, 14 y 21 se aprecia un ciclo 2004-2006 (alza) y otro 2005-2007 (alza). De acuerdo al test de Mann-Kendall no existe evidencia suficiente que respalde la existencia de tendencia en los cuatro primeros alcances, mientras que sí la hay para el alcance de 21 días. En este caso la estimación de la pendiente de Sen es 0,0017.
- **Estacionalidad.** La estacionalidad se muestra bien definida y constante. En cuanto a los períodos de picos y depresiones es algo más irregular que en el caso de la temperatura y se observan cambios de patrón, más evidentes en los alcances 1, 3 y 7. Así, en estos casos, los picos se presentan a mediados de la primavera, con cambios en 2008 (3) y 2009 (7), un poco antes o algo después; mientras que las depresiones se producen en el alcance 1 a principios del invierno, y en los restantes dos a fines del otoño, y en los tres casos con períodos de cambio hacia mediados del otoño, siendo los quiebres en 2007 (3 y 7), 2008 (1), 2014 (7) y 2015 (3). Para los alcances 14 y 21, los picos se producen a principios de la primavera, fines de la primavera y comienzos del verano, con quiebres 2009 (14 y 21), 2010 (14) y 2013 (21); mientras que las depresiones se producen a mediados del otoño, fines del invierno y principios del invierno con quiebres en 2007 y 2012 (21). En todas las series, a partir de 2009 aproximadamente comienza a reducirse la varianza y como se mencionó en algunos casos cambia el patrón.

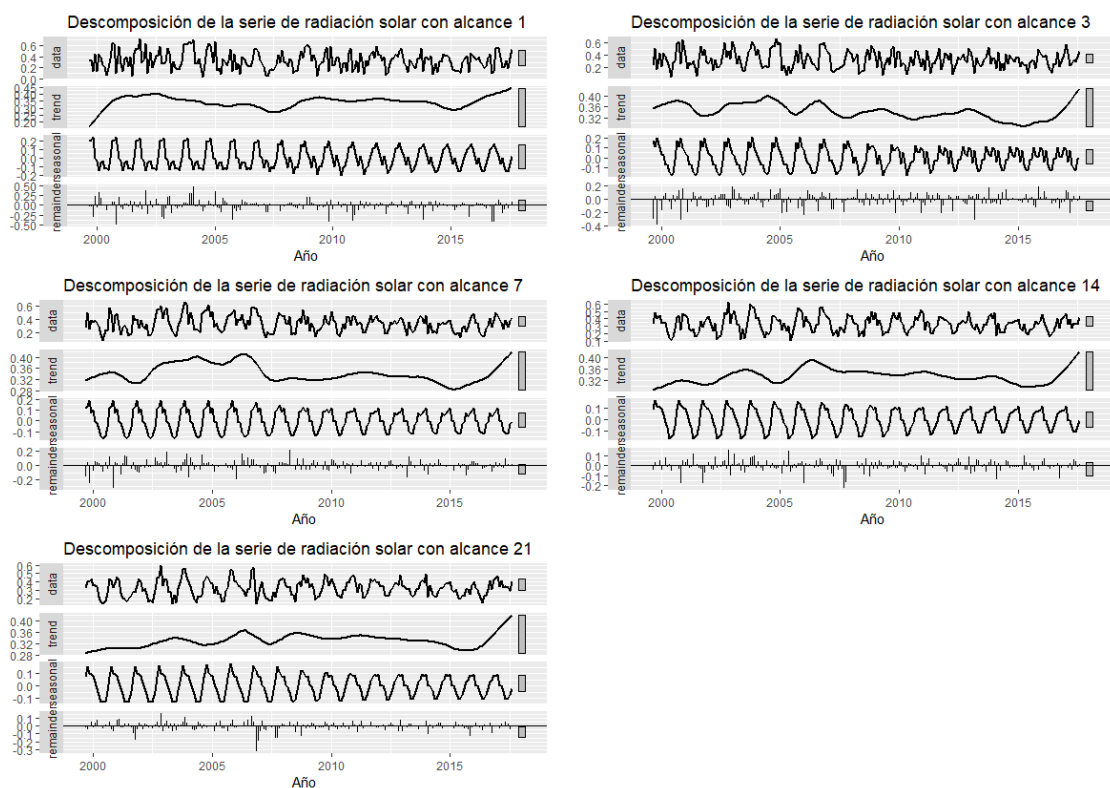


Fig. 41 Descomposición de las series de tiempo de radiación solar en el sitio Centro para los distintos alcances: datos originales en ly (*data*), parte de tendencia (*trend*), parte estacional (*seasonal*) y resto (*remainder*)

En lo que respecta al análisis de correlación temporal, se puede notar lo siguiente:

- **Autocorrelación.** Como se puede apreciar en la Fig. 42, la autocorrección es significativa y perfectamente estacional en los alcances 14 y 21, excepto en los retardos de transición coincidentes con los observados en el caso de la temperatura y con el mismo patrón que en este caso por lo demás. Yendo en sentido desde el retardo 7 al 1, la autocorrelación se hace más débil, reflejando aún estacionalidad, pero no siempre significativa en el alcance 7, y en el caso de los alcances 1 y 3, podría decirse que es significativa y alta sólo en el retardo 12.

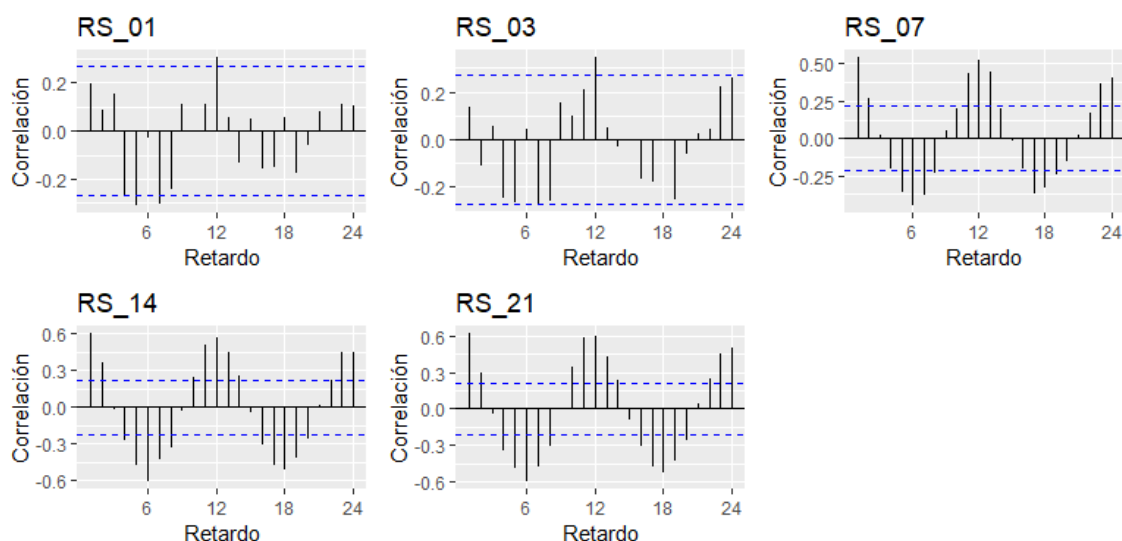


Fig. 42 Autocorrelación de las series de tiempo de radiación solar en el sitio Centro por alcance

- **Autocorrelación parcial.** En la Fig. 43, se puede observar que en el caso de los retardos 1 y 3 se observa autocorrelación parcial significativa negativa con retardo 4, y también 7 sólo en el alcance 3. En los alcances, 7, 14 y 21 se observa autocorrelación parcial significativa positiva con retardo 12 y negativa con retardo 6 y 3; y otras de menor relevancia.

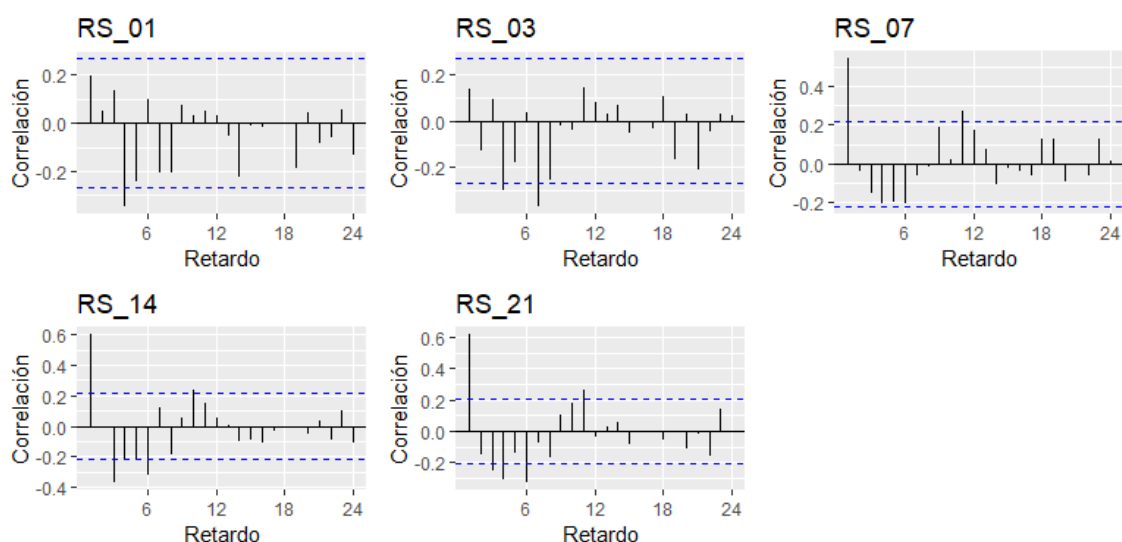


Fig. 43 Autocorrelación parcial de las series de tiempo de radiación solar en el sitio Centro por alcance

La Fig. 44 muestra las series de tiempo de humedad relativa por sitio y los destinos alcances. Se trata de series con abundante variabilidad, que no parece estabilizarse mucho conforme el alcance se incrementa.

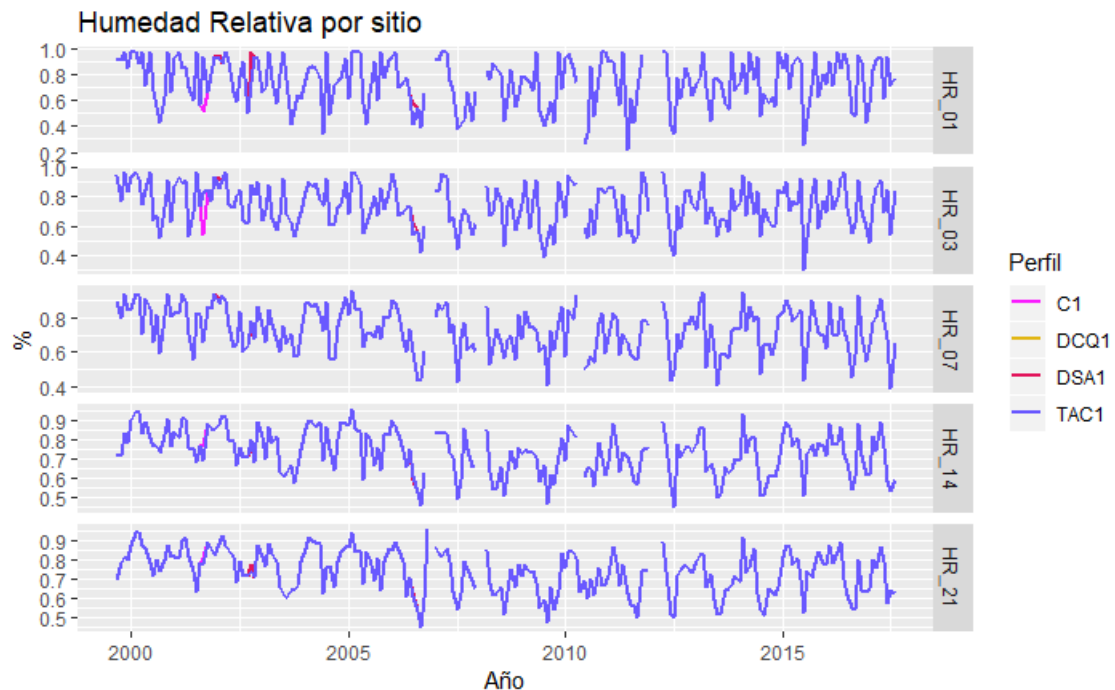


Fig. 44 Serie de tiempo de humedad relativa, para distintos alcances, por sitio

La Fig. 45 presenta la descomposición de las series, que puede entenderse como se comenta a continuación:

- **Tendencia.** En los 3 primeros alcances no se observa tendencia creciente ni decreciente, habiendo un ciclo 2009-2013 (alza) en el alcance 1 bastante acentuado. En el alcance 7, en tanto, sobresale un ciclo 2004-2006 (alza). Los dos ciclos mencionados se observan en el alcance 14 y 21, sumando otro 2003-2005 (baja) y 2013-2015 (baja), en ambos casos con tendencia que pareciera decreciente. De acuerdo al test de Mann-Kendall, existe evidencia de tendencia en todos los alcances excepto el 1. Las pendientes estimadas de Sen son para los alcances 3, 7, 14 y 21 respectivamente -0,0035, -0,0061, -0,0069 y -0,0070.
- **Estacionalidad.** Se observa un incremento de la varianza de la estacionalidad en todos los alcances desde el 3 a partir del 2006 aproximadamente. Los picos de humedad se producen a mediados del verano en alcances 1, 3 y 21, con cambios hacia fines del verano en los otros alcances por períodos cortos, con cortes en 2000 (7) y 2004 (14). Por otro lado, las depresiones se producen mayormente a inicios del invierno con cambio a fines del invierno hasta 2007 (1 y 3); y desde mediados del invierno con cambios a principios del invierno en 2010 (7), 2014 (14) y 2011 (21).

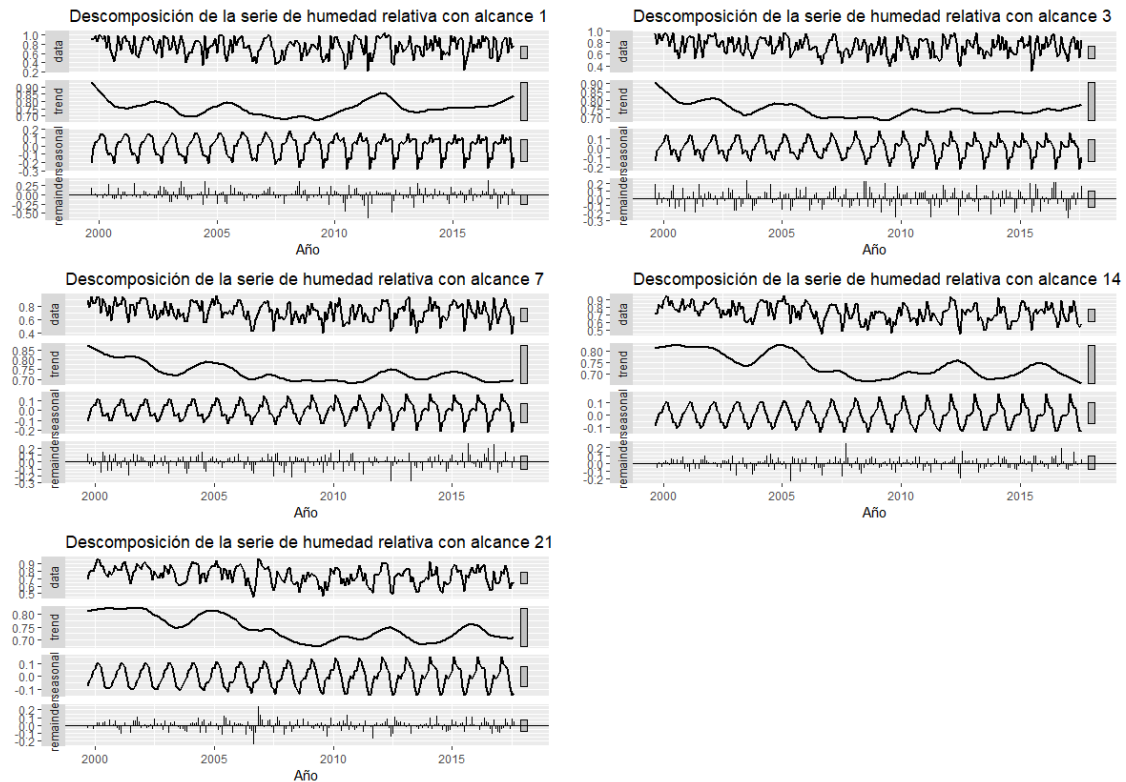


Fig. 45 Descomposición de las series de tiempo de humedad relativa en el sitio Centro para los distintos alcances: datos originales en % (*data*), parte de tendencia (*trend*), parte estacional (*seasonal*) y resto (*remainder*)

En lo que se refiere a la correlación temporal se observa lo siguiente:

- **Autocorrelación.** En la Fig. 46, se puede observar que en el alcance 21 existe autocorrelación significativa casi perfectamente coincidente con la estacionalidad, siendo más fuerte en los retardos 1, 6, 12 y 18 y más débil en los retardos inmediatos a los mencionados. En los alcances 14 y 7 aún se observa este patrón, pero es más débil desde el punto de vista de la significatividad en otros retardos distintos a los centrales. En el caso del alcance 1 y 3, la autocorrelación aún es significativa y positiva en el retardo 1 y significativa y positiva en el 6; además de en otros casos menos relevantes.

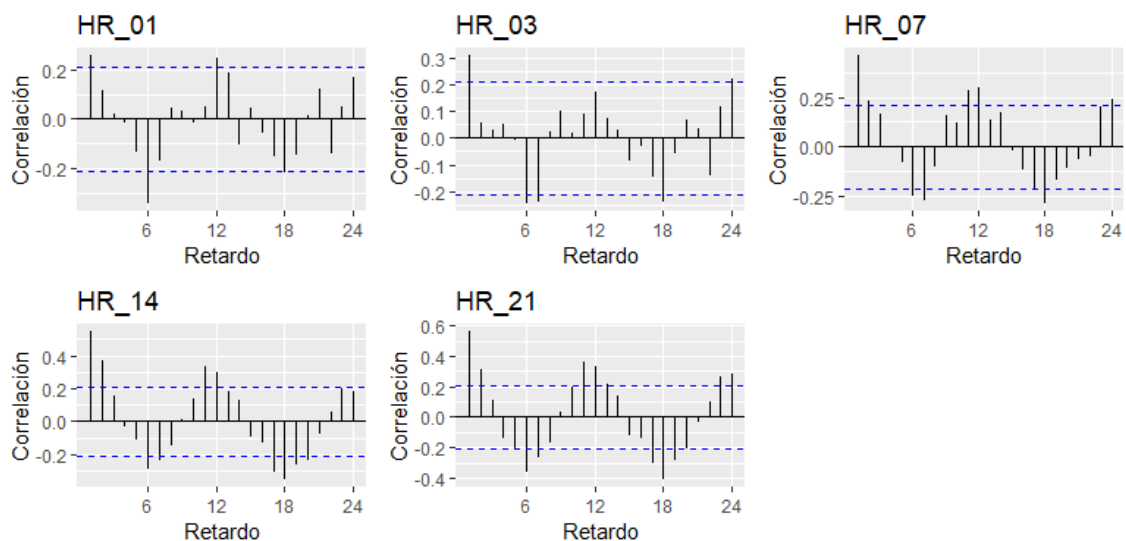


Fig. 46 Autocorrelación de las series de tiempo de humedad relativa en el sitio Centro por alcance

- **Autocorrelación parcial.** En lo que respecta a la autocorrelación parcial, en la Fig. 47, se observa que es significativa y negativa con retardo 6 en todos los alcances, y es significativa y positiva con retardo 11 en los alcances 7, 14 y 21, y con retardo 9 en el alcance 7; además de otros casos menos relevantes.

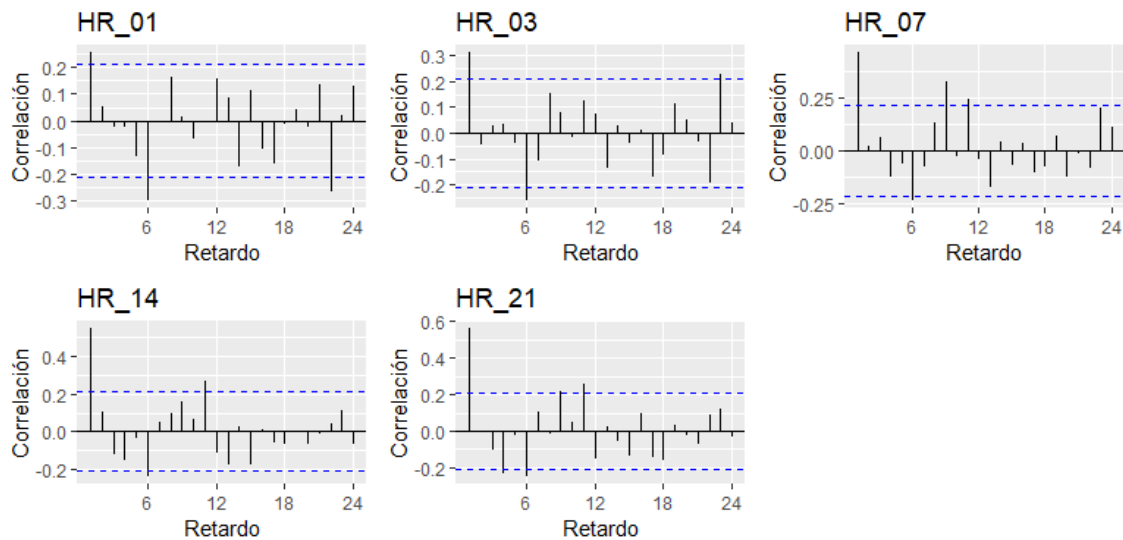


Fig. 47 Autocorrelación parcial de las series de tiempo de humedad relativa en el sitio Centro por alcance

La Fig. 48 muestra las series de tiempo de presión atmosférica para distintos alcances. Si bien la estacionalidad parece definirse mejor conforme aumenta el alcance se observa gran variabilidad.

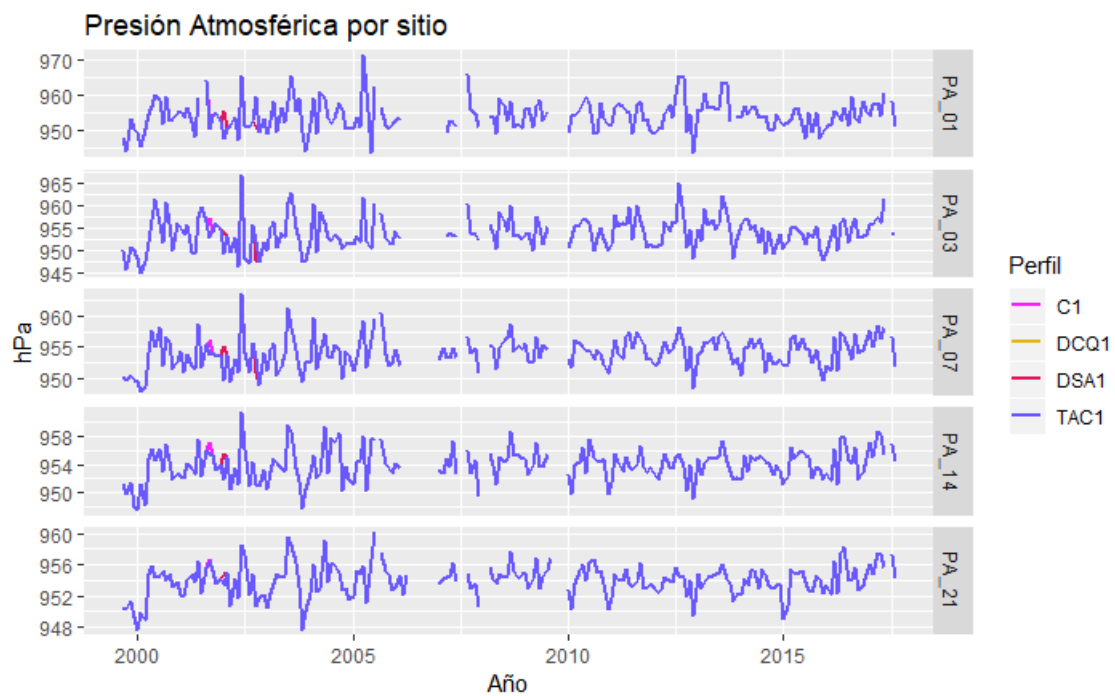


Fig. 48 Series de tiempo de presión atmosférica, para distintos alcances, por sitio

La Fig. 49 muestra la descomposición de las series que permiten realizar el siguiente análisis:

- **Tendencia.** No se observa tendencia creciente ni decreciente para ningún alcance. Tampoco se observan ciclos más que para los dos primeros alcances en el período 1999-2002 (alza). De acuerdo al test de Mann-Kendall existe tendencia únicamente en el caso del alcance de 7 días. La pendiente de Sen es 0,0999.
- **Estacionalidad.** La varianza de la estacionalidad se observa conservada, pero se aprecian varios cambios de patrón a lo largo del tiempo, en distintos alcances, es una serie con estacionalidad más bien irregular. En términos generales podría decirse que los picos máximos se producen en el período otoño-invierno, con mayor presencia de fin del invierno en todos los alcances. En lo que respecta a las depresiones, se producen en el período primavera-verano, con mayor presencia de fines de la primavera en todos los alcances.

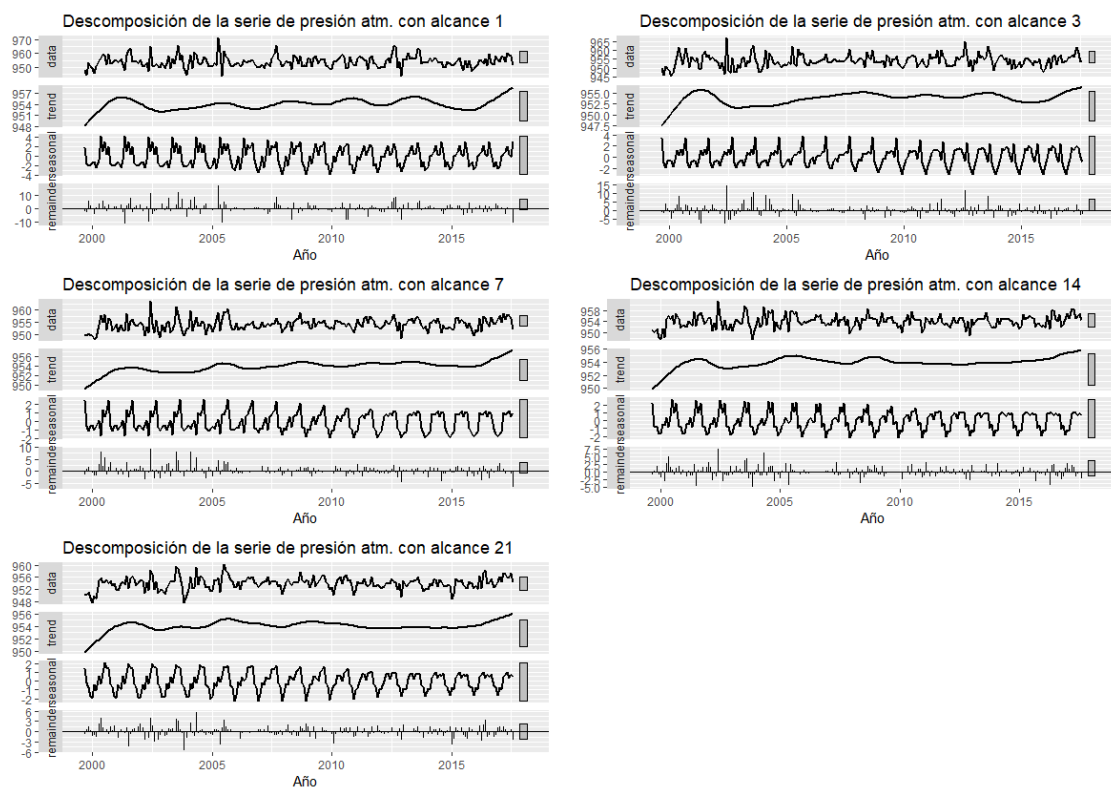


Fig. 49 Descomposición de las series de tiempo de presión atmosférica en el sitio Centro para los distintos alcances: datos originales en hPa (*data*), parte de tendencia (*trend*), parte estacional (*seasonal*) y resto (*remainder*)

En lo que respecta a la correlación temporal de las series de presión atmosférica por alcances, se observa lo siguiente:

- **Autocorrelación.** Como se puede observar en la Fig. 50, en el alcance 21 se aprecia un patrón que podría reflejar cierta estacionalidad, pero algo débil: la correlación es significativa y alta en los retardos 1 y 11, donde es positiva, y 5 donde es negativa. Un patrón similar se observa en los alcances 3, 7 y 14 con distinto grado de relevancia. En el caso del alcance 1 no se observa autocorrelación.

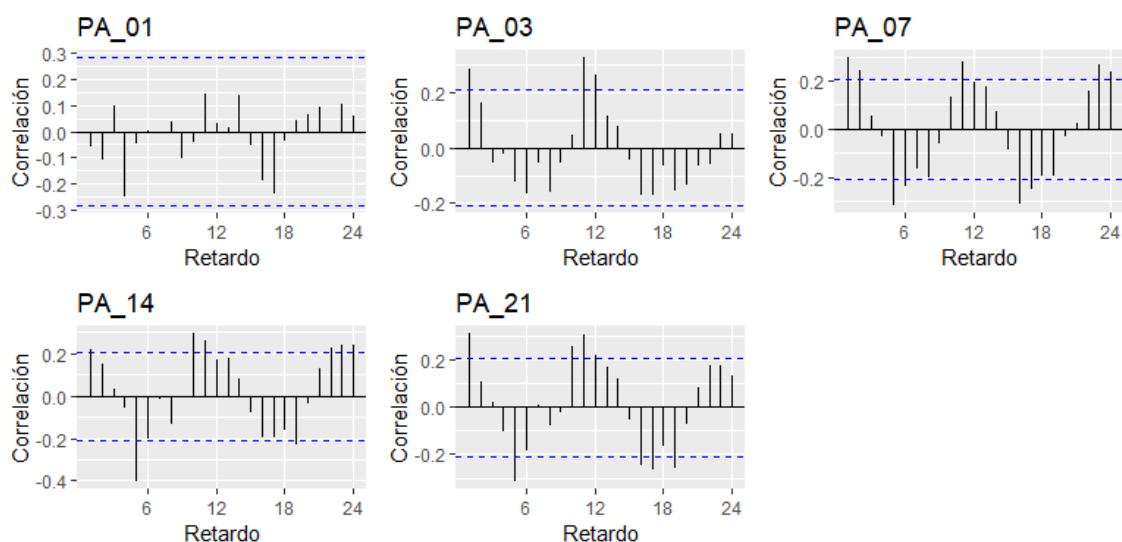


Fig. 50 Autocorrelación de las series de tiempo de presión atmosférica en el sitio Centro por alcance

- **Autocorrelación parcial.** Se observa en la Fig. 51, que la autocorrelación es significativa negativa con retardo 5 en los alcances 7, 14 y 21; mientras que en el alcance 3 se observa significativa y positiva con retardo 11; además existen otros casos de menor relevancia.

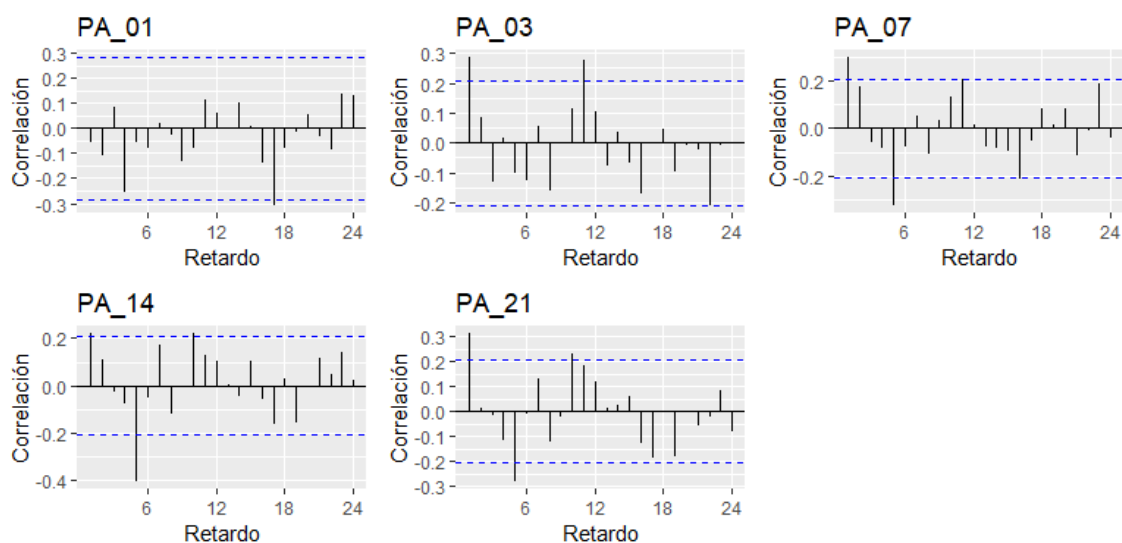


Fig. 51 Autocorrelación parcial de las series de tiempo de presión atmosférica en el sitio Centro por alcance

La Fig. 52 muestra las series de tiempo de velocidad del viento para los distintos alcances y sitios, observándose estabilización de la varianza conforme aumenta el alcance y un incremento de los picos de viento sobre el final del período.

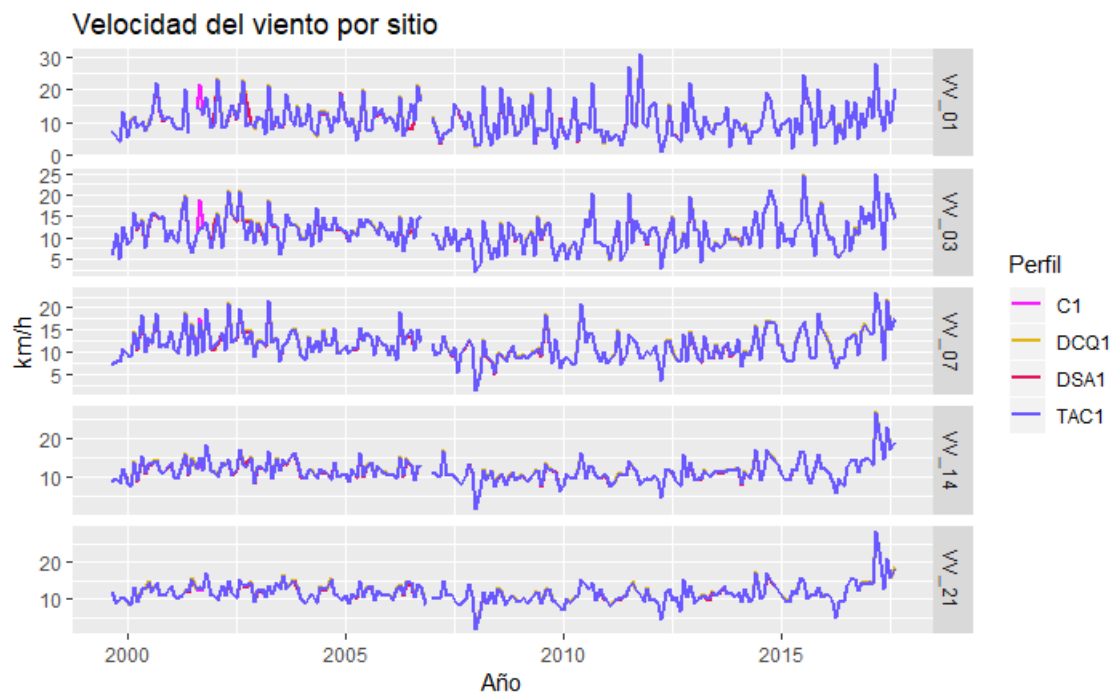


Fig. 52 Velocidad del viento, para distintos alcances, por sitio

Por otra parte, la Fig. 53 muestra el resultado de la descomposición de las series, que amerita el siguiente análisis:

- **Tendencia.** No se observa una tendencia creciente ni decreciente evidente en ningún caso y tampoco se observan ciclos. De acuerdo al test de Mann-Kendall no existe tendencia para ningún alcance.
- **Estacionalidad.** Se observa de varianza aproximadamente constante, con cambios de patrón en todos los alcances y mayormente definida en el alcance 21. Se trata de una serie irregular con recurrentes cambios en los patrones en los primeros tres alcances. En esos casos podría decirse que los picos se producen mayormente en invierno, primavera y otoño (en menor medida); mientras que las depresiones se producen mayormente en verano y otoño (en menor medida). En el caso de los alcances 14 y 21 las depresiones se producen a mediados del verano y los picos durante el invierno (eventualmente a mediados del otoño o mediados de la primavera).

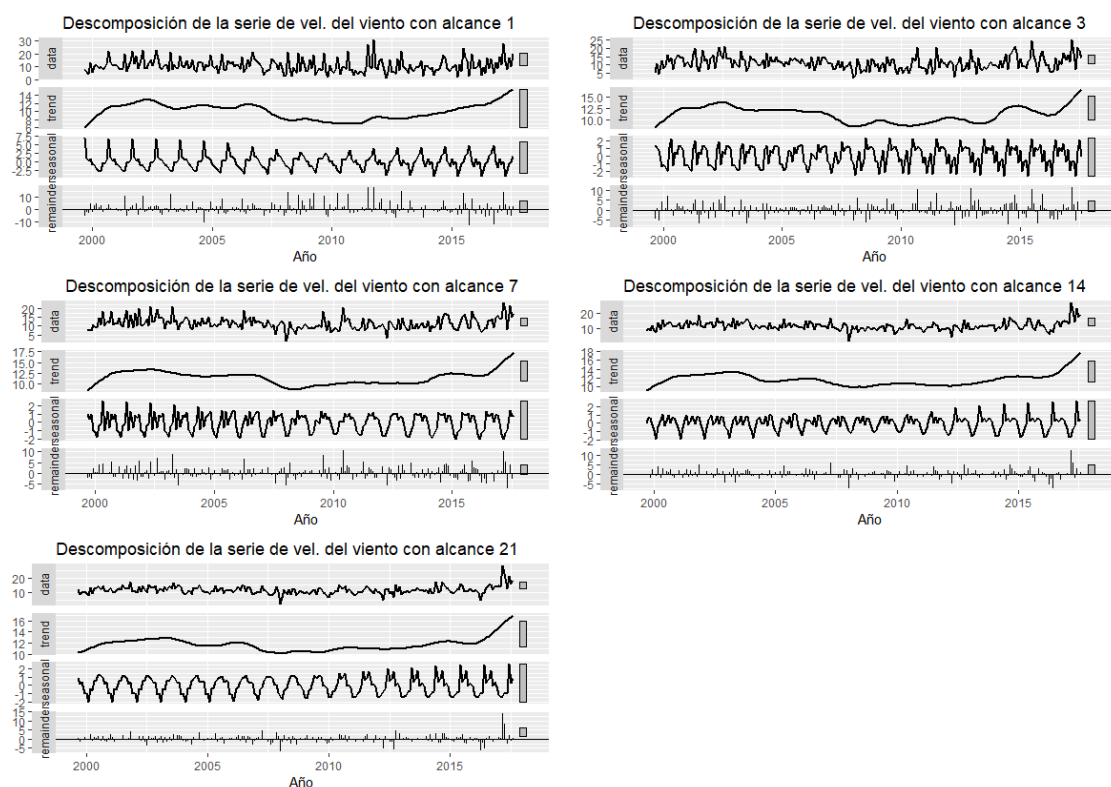


Fig. 53 Descomposición de las series de tiempo de velocidad del viento en el sitio Centro para los distintos alcances: datos originales en km/h (*data*), parte de tendencia (*trend*), parte estacional (*seasonal*) y resto (*remainder*)

En lo que respecta a la correlación temporal de la velocidad del viento, se observa lo siguiente:

- **Autocorrelación.** Como se puede apreciar en la Fig. 54, en este caso el patrón es marcadamente distinto, en los alcances 14 y 21 se observa una correlación que no parece ser estacional sino más bien decreciente en términos de los retardos. Así en estos dos casos se observa autocorrelación positiva significativa con retardos 1 a 5. En cuanto a los alcances 3 y 7 se observa autocorrelación significativa positiva en el retardo 1, además de otros menos relevantes. Por último, en el alcance 1, hay autocorrelación significativa negativa con alcance 9.

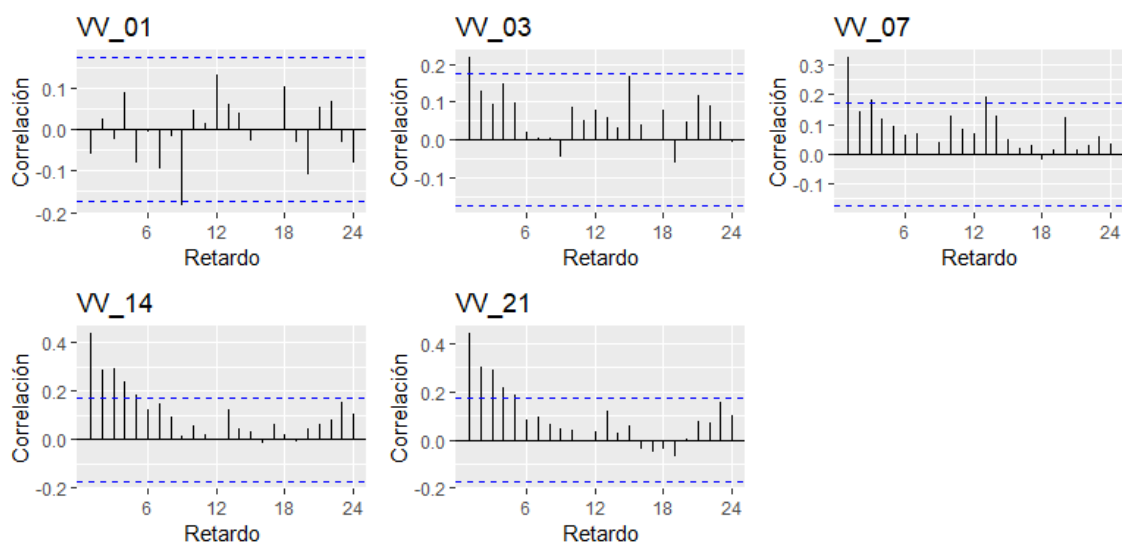


Fig. 54 Autocorrelación de las series de tiempo de velocidad del viento en el sitio Centro por alcance

- **Autocorrelación parcial.** Como se nota en la Fig. 55, no se observa autocorrelación parcial.

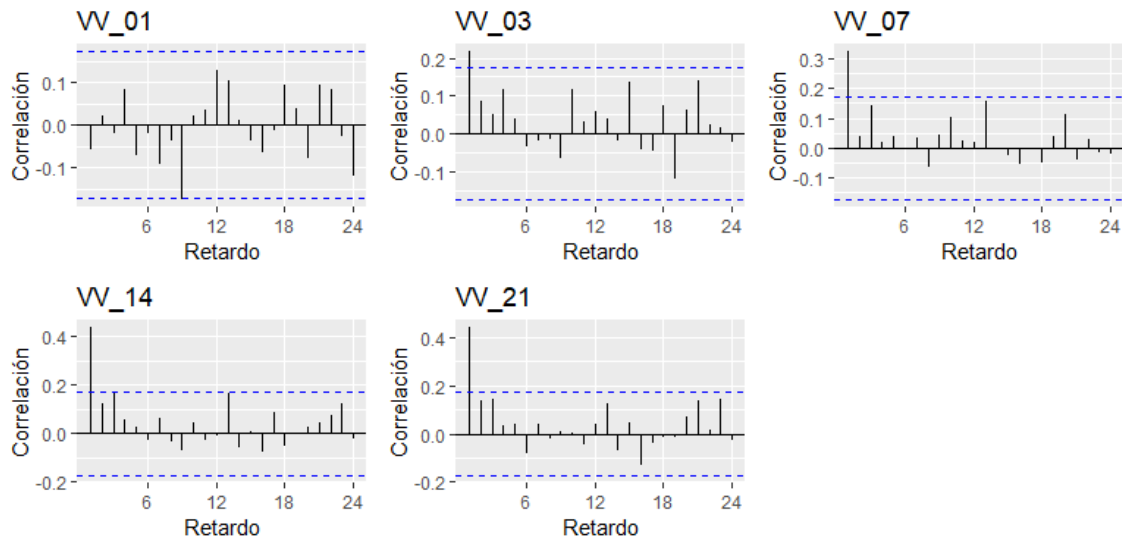


Fig. 55 Autocorrelación parcial de las series de tiempo de velocidad del viento en el sitio Centro por alcance

La Fig. 56 muestra las series de tiempo de dirección del viento (en términos de 8 puntos cardinales: N=Norte, S=Sur, E=Este, O=Oeste, NE=Noreste, NO=Noroeste, SE=Sureste, SO=Suroeste) por alcance. Vale aclarar que las rectas que unen los puntos se utilizaron únicamente a fines de presentar una visualización ordenada en el tiempo de las transiciones entre las orientaciones (no indican interpolación de otros puntos). El autor entiende que existen otras representaciones comúnmente usadas para este tipo de variables como las rosas de vientos. Sin embargo, la elección de la visualización que se presenta estuvo dirigida a mantener la consistencia con respecto a las representaciones en series de tiempo de las demás variables continuas, y así poder eventualmente comparar patrones por período de tiempo.

En base al análisis de la figura, se observa predominancia del viento NE, NO y E a partir del alcance 3 acentuado conforme se incrementa el alcance, lo que se haya indicado con recuadros de color (amarillo, rojo y verde respectivamente). En este caso, por tratarse de una variable de tipo categórica no es posible realizar la descomposición de las series ni llevar a cabo análisis de autocorrelación en el tiempo.

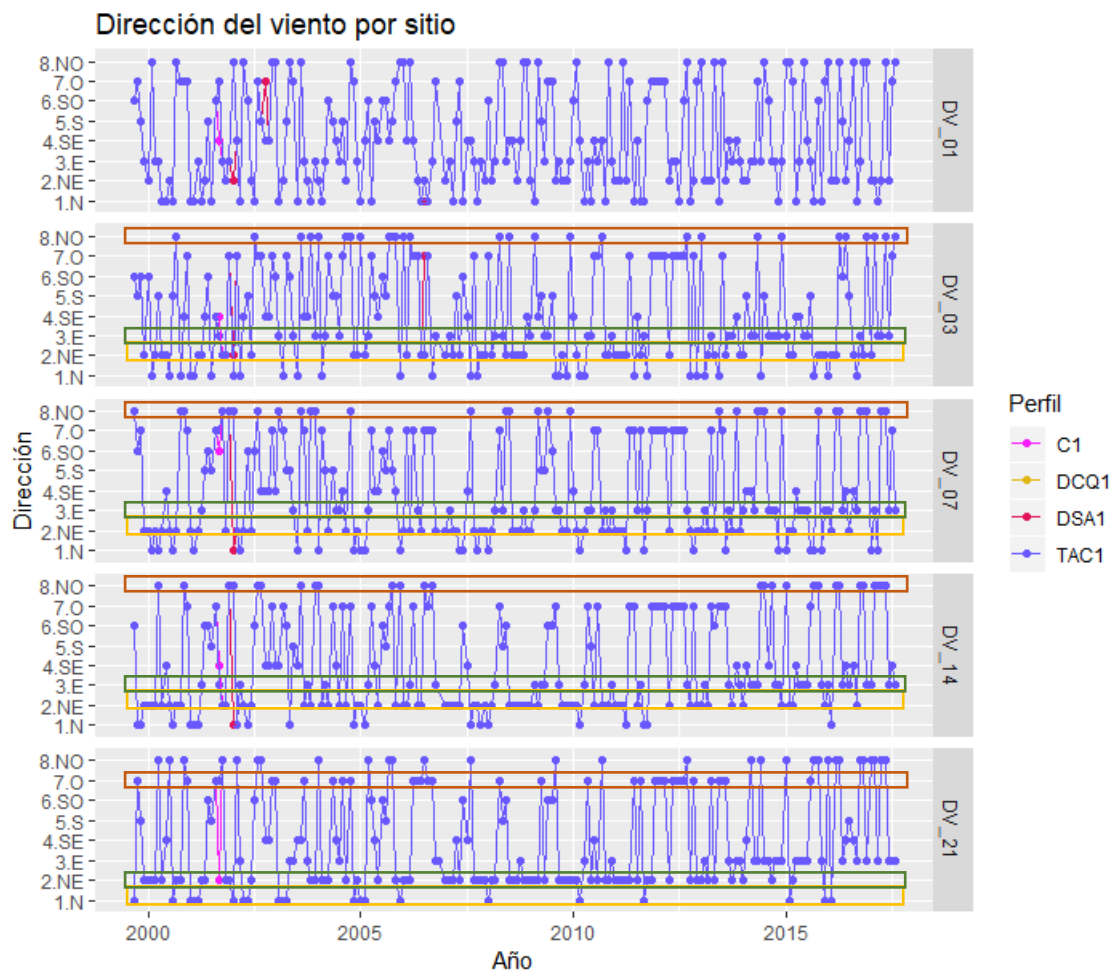


Fig. 56 Dirección del viento, para distintos alcances, por sitio

La variable precipitación se encuentra representada en la Fig. 57 en distintas series para cada sitio y alcance. Vale la pena remarcar que en el caso de esta variable se calcula la lluvia acumulada para los distintos alcances y no el promedio como en las demás variables continuas. Se observan picos eventuales de alta precipitación y a medida que aumenta el alcance se define con mayor precisión la estacionalidad, y la varianza aparece más conservada.

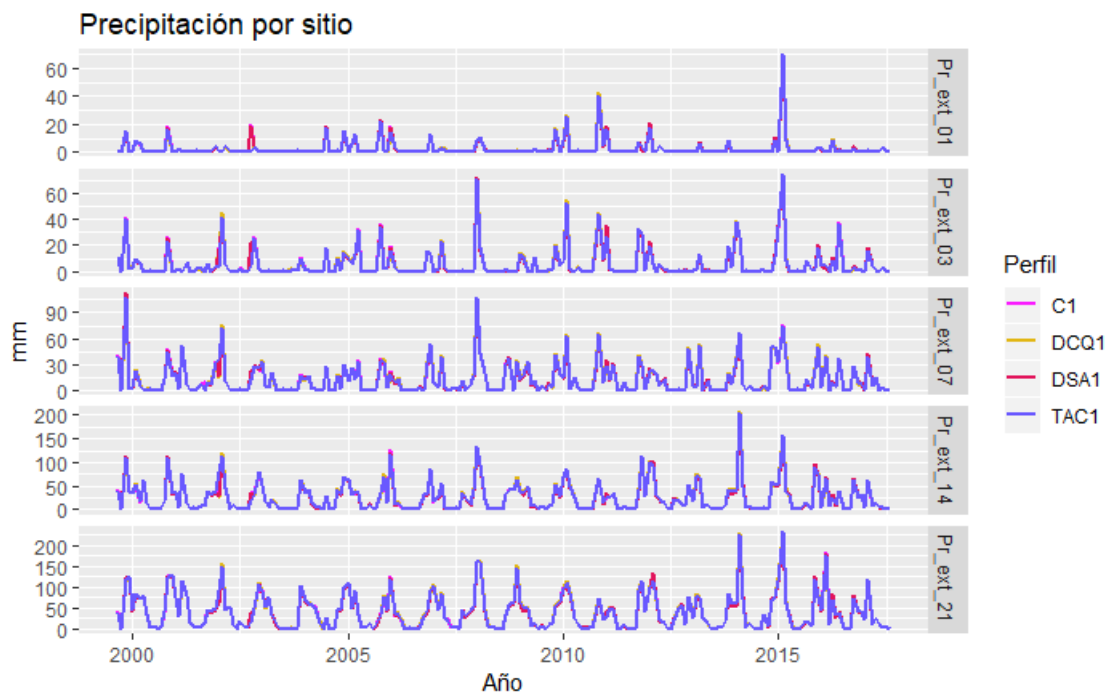


Fig. 57 Series de tiempo de precipitación, para distintos alcances, por sitio

La Fig. 58 muestra el resultado de la descomposición de las series, pudiendo realizarse el siguiente análisis:

- Tendencia.** No se observa tendencia creciente ni decreciente en ningún alcance. En el alcance 3 se observan cuatro ciclos: 2002-2005 (alza), 2005-2007 (alza), 2008-2010 (alza) y 2015-2017 (alza). En el alcance 7 solo dos: 2000-2002 (alza) y 2008-2011 (alza). En el alcance 14 se observan también tres ciclos bien definidos: 2004-2006 (alza), 2008-2011 (alza) y 2014-2017 (alza). Por último, el alcance 21 muestra un ciclo 2011-2013 (alza). De acuerdo al test de Mann-Kendall no existe evidencia de la existencia de tendencia en ninguno de los alcances.
- Estacionalidad.** Se observa estacionalidad con varianza conservada, pero cambios de patrón evidentes en alcances 14 y 21. En los primeros tres alcances, los picos se producen a mediados del verano (1), comienzos del verano (3) y mediados de la primavera (7); mientras que las depresiones se producen en invierno y otoño, con cortes: inicios del invierno con cambios a fines del otoño en 2007 e inicios del otoño en 2010 (1), inicios del otoño con cambio a mediados del invierno en 2009 (3), y fines del otoño con cambios a mediados del invierno en 2006 (7). En cuanto a los alcances 14 y 21, los picos se producen desde fines de la primavera, con cambios a mediados de la primavera en 2009 y mediados del verano en 2014 (14) por un lado; y a inicios del verano en 2006 (21), por el otro.

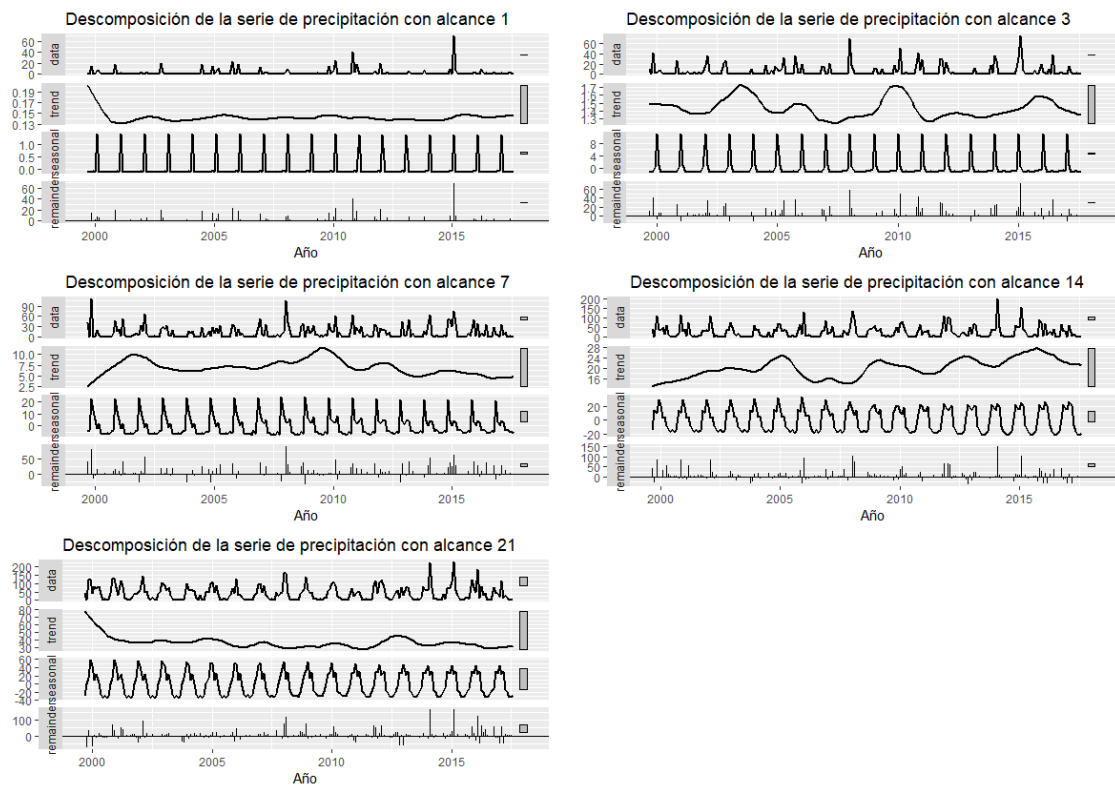


Fig. 58 Descomposición de las series de tiempo de precipitación en el sitio Centro para los distintos alcances: datos originales en mm (*data*), parte de tendencia (*trend*), parte estacional (*seasonal*) y resto (*remainder*)

En lo que respecta a la correlación temporal, se puede notar lo siguiente:

- **Autocorrelación.** Como se puede ver en la Fig. 59, la autocorrelación se observa significativa y siguiendo un patrón perfectamente estacional en los alcances 14 y 21, excepto en los retardos de transición observados en temperatura. En el caso de los alcances 3 y 7 se debilita el patrón, pero aún se mantienen ciertas autocorrelaciones significativas en los retardos 1 y 12 siendo positivas, y 5 y 7 siendo negativas (para el alcance 7 hay más casos aun siendo el debilitamiento del patrón gradual). Para el alcance 1 no existe autocorrelación significativa.

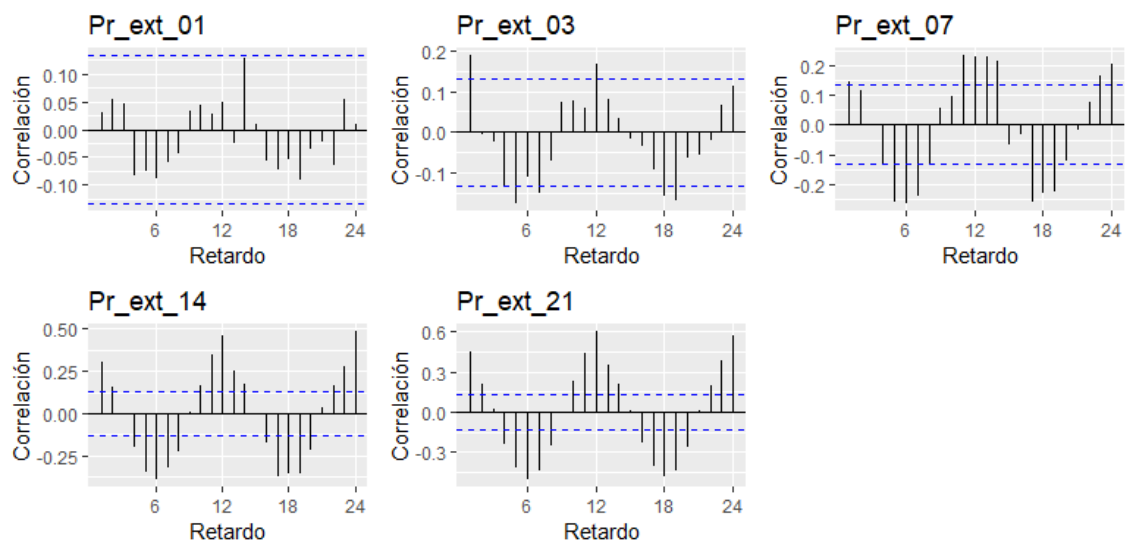


Fig. 59 Autocorrelación de las series de tiempo de precipitación en el sitio Centro por alcance

- **Autocorrelación parcial.** Se observa en la Fig. 60, que la autocorrelación parcial es más intensa en los alcances 7, 14 y 21, en particular la autocorrelación parcial significativa negativa en los retardos 4 a 7, y 8 sólo en el alcance 14; y la autocorrelación parcial significativa positiva en el retardo 12, sólo en los alcances 14 y 21, y en el retardo 14 para el alcance 7. En el alcance 3 hay autocorrelación significativa negativa en el retardo 7.

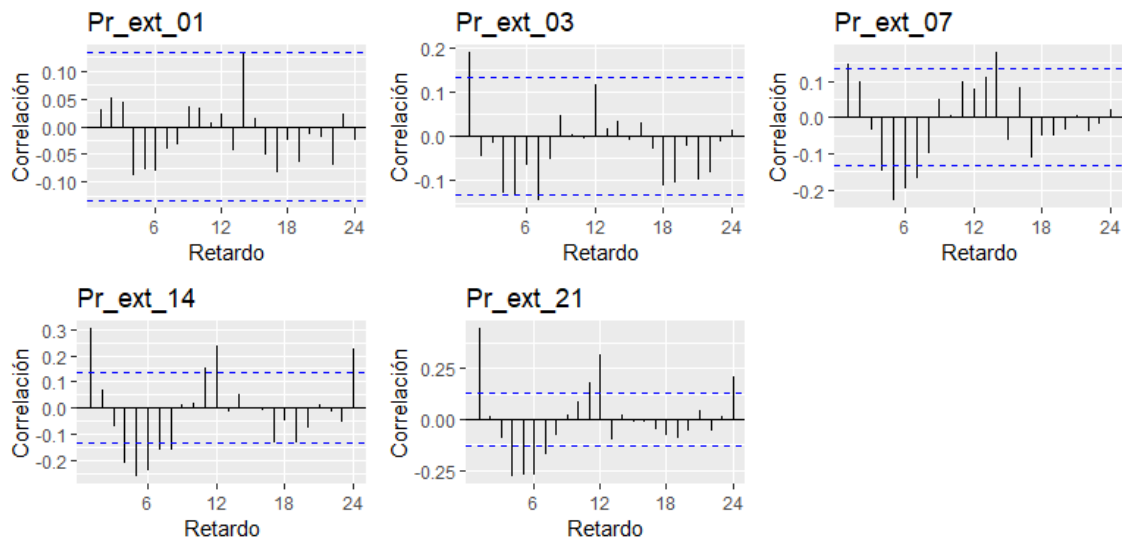


Fig. 60 Autocorrelación parcial de las series de tiempo de precipitación en el sitio Centro por alcance

La Fig. 61 muestra las series de tiempo de la cota o nivel del lago. En el panel superior, se puede observar la serie correspondiente a la medición mensual que se realiza como parte del monitoreo permanente que realiza el CIRSA, denominada cota. En los paneles inferiores siguientes, se muestran las series diarias de los datos de los sensores automáticos instalados en distintas estaciones junto con datos diarios complementarios partir de 2016, proporcionados por MMySP, cuando la operación de los sensores se tornó irregular. Es importante resaltar que no se observan discrepancias relevantes entre las series de cota del embalse tomadas por CIRSA y las series de nivel del embalse (Nv), registradas por los sensores y tomadas por MAAySP, y a la vez se destaca una marcada estacionalidad.

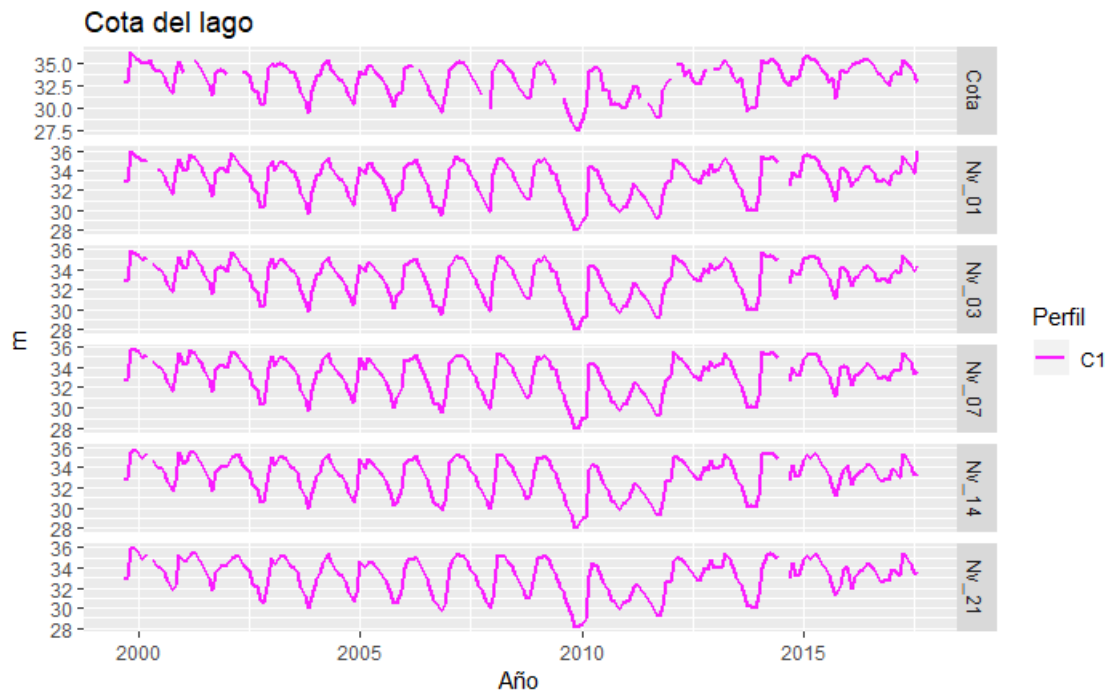


Fig. 61 Series de tiempo de nivel del lago, para distintos alcances y medición in situ (cota)

Por otra parte, la Fig. 62 muestra la descomposición de las series en los siguientes términos:

- **Tendencia.** A partir de los gráficos no podría aseverarse que exista tendencia creciente o decreciente. Sí aparece bien definido un ciclo principal en todas las series: 2007-2012 (baja). De acuerdo al test de Mann-Kendall existe tendencia únicamente en los alcances 7, 14 y 21, con las siguientes pendientes de acuerdo a la estimación de Sen, respectivamente: -0,0232, -0,0269 y -0,0287.
- **Estacionalidad.** En cuanto a la estacionalidad, se observa de varianza conservada y un mismo patrón de periodicidad. En todos los alcances las depresiones ocurren a mediados de la primavera y a partir de 2008 a principios de la primavera. En lo que respecta a los picos, se producen mayormente entre fines del verano y principios del otoño, con el siguiente patrón de cambios en cada alcance: principios del otoño, cambio en 2007 a fines del verano y nuevamente a principios de otoño a partir de 2013 (1); fines del verano, y a partir de 2016 a principios del otoño (3); fines del verano (7); fines del verano, a partir de 2010 a mediados y a partir de 2012 a principios del otoño (14); y principios del otoño (21).

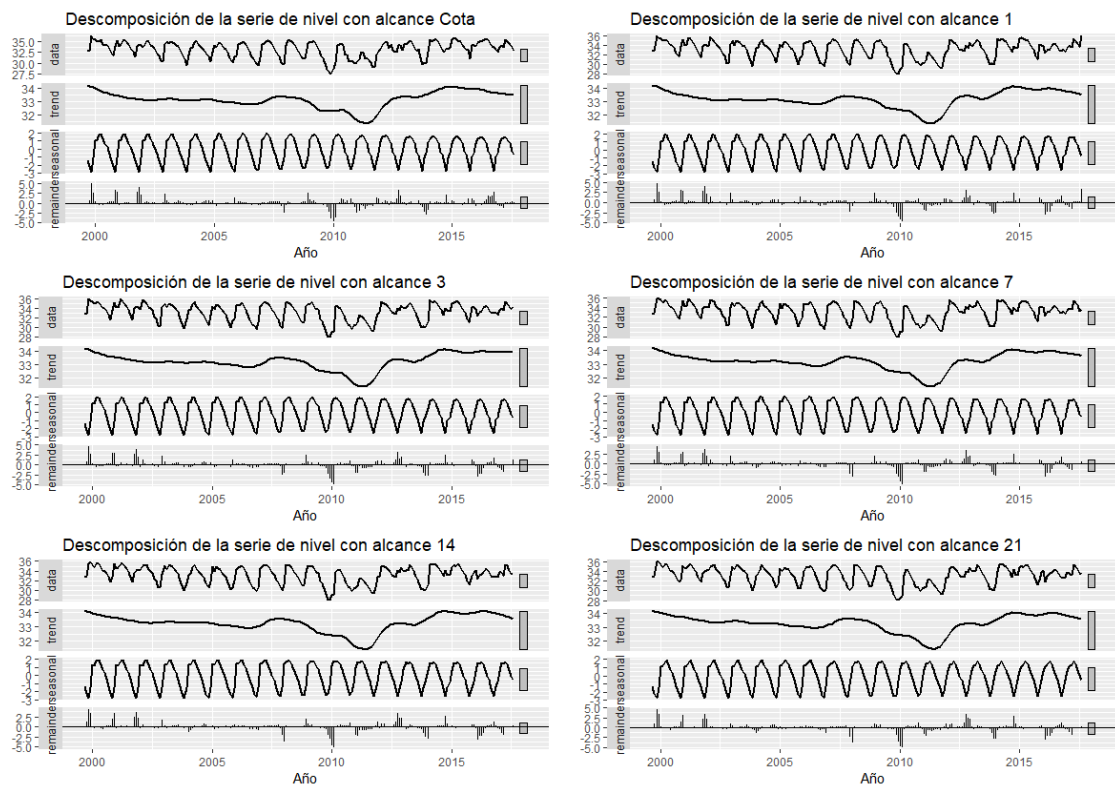


Fig. 62 Descomposición de las series de tiempo del nivel del lago en términos de cota (medida manual durante el monitoreo) y de nivel medido de forma diaria para los distintos alcances: datos originales en m (*data*), parte de tendencia (*trend*), parte estacional (*seasonal*) y resto (*remainder*)

En lo que respecta a la correlación temporal, se observa lo siguiente:

- **Autocorrelación.** De acuerdo a la Fig. 63, se aprecia que existe un patrón perfectamente estacional en la autocorrelación, siendo significativo en todos los alcances de nivel excepto en los retardos de transición mencionados con anterioridad. En el caso de cota, también se observa el patrón de estacionalidad, pero se hace débil a medida que se incrementan los retardos, siendo menos significativo en ciertos retardos.

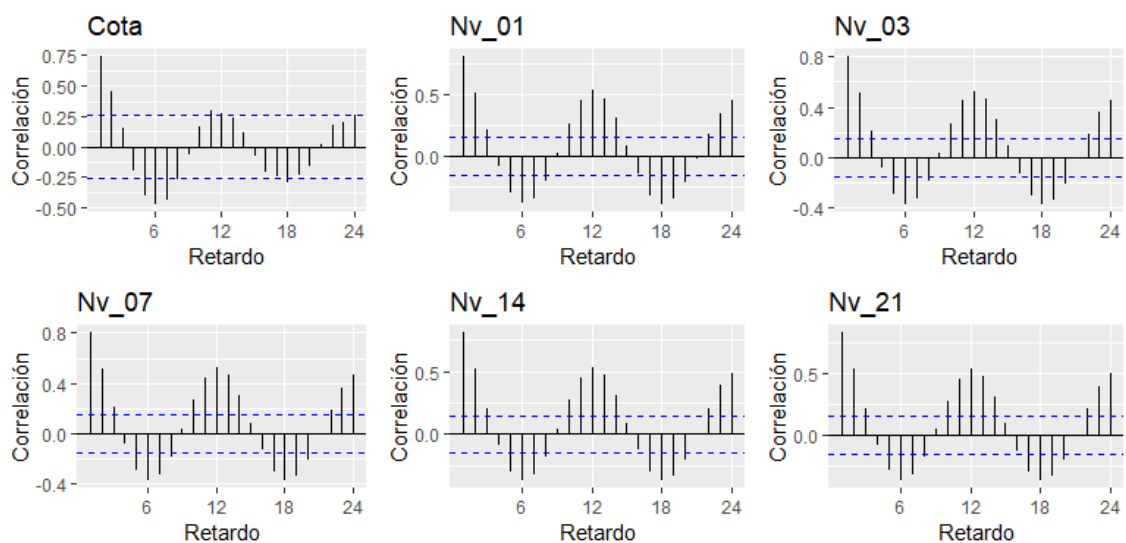


Fig. 63 Autocorrelación de las series de tiempo de nivel del lago: cota (registro de monitoreo) y los distintos alcances del nivel (medido mayormente por sensor)

- **Autocorrelación parcial.** Se observa en la Fig. 64, que la autocorrelación parcial es significativa y negativa con retardos 2, 3 y 4 en los alcances 3, 7 y 14, en alcance 1 sólo 2 y 4, y en alcance 21 solo 2. En todos los alcances se observa autocorrelación parcial significativa positiva con retardo 10. En lo que respecta a cota, se observa autocorrelación parcial significativa negativa con retardo 4 y positiva con retardo 21.

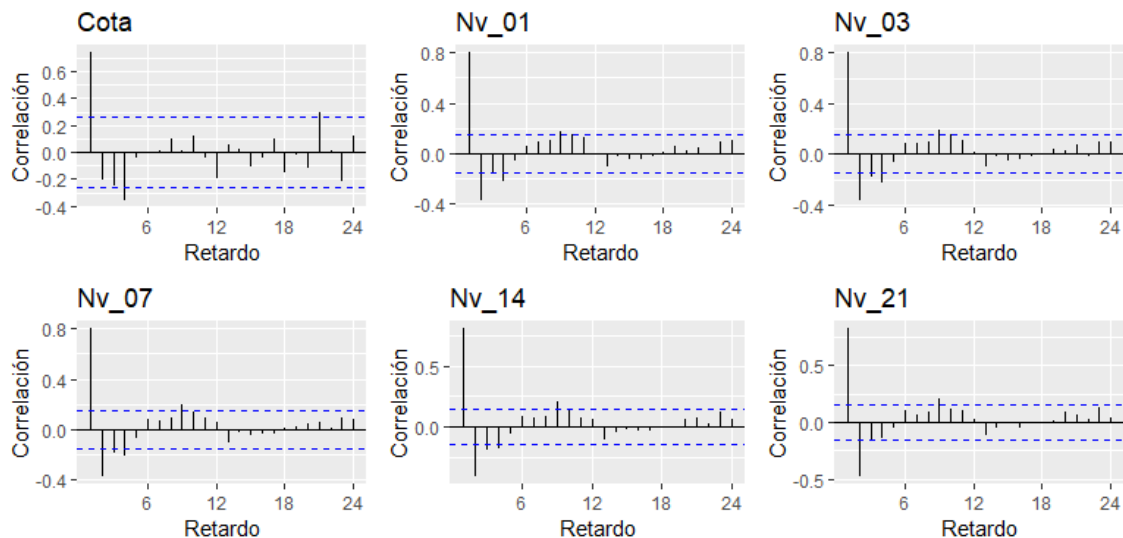


Fig. 64 Autocorrelación parcial de las series de tiempo de nivel del lago: cota (registro de monitoreo) y los distintos alcances del nivel (medido mayormente por sensor)

3.2 Modelos

A efectos de analizar la incidencia de las condiciones meteorológicas sobre la calidad del agua, se desarrollaron tres modelos que se detallan a continuación a partir de los lineamientos fijados en la sección 2.7. Concretamente, cada modelo, bajo distintos supuestos, predice el IET como respuesta. El mismo se calculó a partir de transparencia de Secchi, fósforo total y clorofila-a, series transformadas que se analizan individualmente en la sección 3.1.1. En la Fig. 65 se grafican las tres series de IET junto con el promedio de las mismas, por sitio y perfil, que admiten dos conclusiones. Por un lado, como se puede observar, en los *perfiles superficiales*: zona subsuperficial y zona fótica, los índices de estado trófico son consistentes, o lo que es lo mismo, la estimación del índice por cualquiera de las tres variables y con algún grado de variabilidad es aproximadamente coincidente. Esto concuerda con el hecho de que el IET de Carlson fue inicialmente concebido para determinar el estado trófico de los lagos analizando el agua superficial. Por lo tanto, y siguiendo a Kratzer y Brezonik (1981), y a Osgood (1982), en estos casos se utiliza el IET promedio como respuesta. Por otro lado, en el caso de la los *perfiles profundos o de fondo*: zona intermedia y zona fondo, se observa claramente la discrepancia entre el IET basado en clorofila-a y el basado en fósforo total. Este fenómeno se debe a dos factores: por un lado, la clorofila-a es escasa a ese nivel de profundidad porque la vida de las algas está sumamente limitada, siendo una de las principales razones la ausencia de luz; y, por otro lado, la presencia de fósforo total es mayor porque estas zonas son las más próximas al suelo donde se acumulan los sedimentos de esta sustancia. Concretamente, la clorofila-a deja

de ser un buen indicador del estado trófico a este nivel, por estar limitada, mientras que el fósforo total refleja fuertemente la condición del agua en esta zona por su abundancia. La tercera variable, transparencia de Secchi, no juega un rol relevante en estas zonas, por ser una variable común a todos los perfiles. Por los motivos mencionados, para estas dos zonas la variable utilizada como respuesta es el IET basado en fósforo total. Vale remarcar que se trata de una innovación, porque a la fecha no se tiene registro del uso del IET en profundidad.

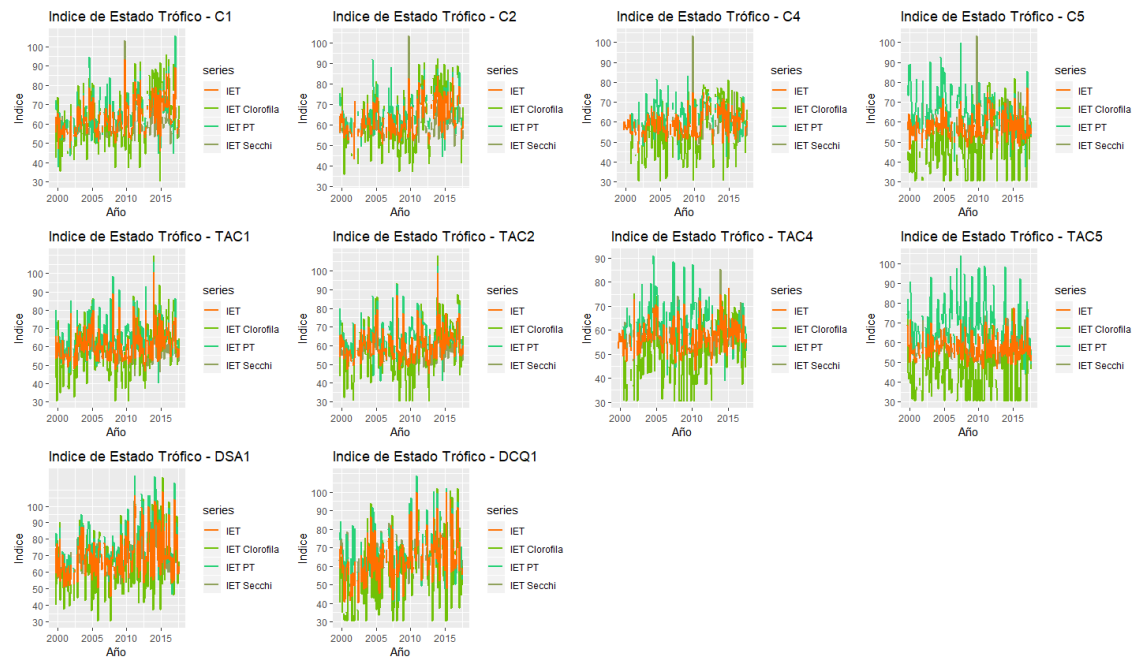


Fig. 65 Variable respuesta: Índice de Estado Trófico promedio (IET) junto con los IET basados en Secchi, fósforo total (PT) y clorofila-a por, sitio y perfil

Por otro lado, en los distintos modelos las variables predictoras utilizadas fueron las hidrológicas y meteorológicas numéricas continuas, a saber:

1. Temperatura
2. Presión Atmosférica
3. Humedad Relativa
4. Radiación Solar
5. Velocidad del Viento
6. Precipitación
7. Nivel

La variable de dirección del viento no fue considerada por ser una variable categórica que no podía ser incluida en todos los modelos, de manera que se podría considerar bajo otras aproximaciones en futuros estudios. Por lo demás se tomaron todas las variables de que se disponía de datos al no existir evidencia a priori para descartar ninguna, y en aras de conseguir el mayor aporte posible a la investigación del fenómeno.

En lo que respecta a correlación de las variables predictoras, ya interpoladas según procedimiento descrito en la sección 2.6.4, se realizó un análisis exploratorio preliminar para

descartar correlación fuerte. La Tabla 16 muestra los resultados¹³ correspondientes a los datos del sitio Centro, como caso de estudio (siguiendo el lineamiento de la sección 3.1.2), por alcance. Vale la pena mencionar que, si bien es cierto que como se señaló en la sección introductoria las variables meteorológicas se encuentran interrelacionadas por distintos fenómenos atmosféricos (ver Fig. 1), la correlación entre las distintas variables hidrológicas y meteorológicas es relativamente baja para los diferentes alcances, por tanto es válido hacer un análisis comparativo de sus contribuciones a la variable respuesta a partir de los coeficientes obtenidos de los modelos.

Tabla 16 Correlación de variables regresoras por alcance en sito Centro

	Alcance	Temperatura	Precipitación	Presión Atmosférica	Humedad Relativa	Radiación Solar	Nivel	Velocidad del Viento
Temperatura	1	1.000	0.075	-0.572	0.165	0.394	-0.055	0.268
	3	1.000	0.188	-0.568	0.181	0.481	-0.058	0.182
	7	1.000	0.364	-0.526	0.310	0.534	-0.050	0.040
	14	1.000	0.519	-0.537	0.402	0.597	-0.028	-0.075
	21	1.000	0.640	-0.545	0.463	0.591	-0.034	-0.192
Precipitación	1		1.000	-0.149	0.338	-0.235	-0.091	-0.081
	3		1.000	-0.162	0.396	-0.113	-0.099	-0.214
	7		1.000	-0.281	0.425	0.055	-0.032	-0.253
	14		1.000	-0.294	0.492	0.189	-0.004	-0.336
	21		1.000	-0.410	0.557	0.260	0.038	-0.353
Presión Atmosférica	1			1.000	-0.133	-0.113	0.049	-0.242
	3			1.000	-0.083	-0.181	0.032	-0.255
	7			1.000	-0.169	-0.246	0.022	-0.100
	14			1.000	-0.163	-0.370	0.082	0.022
	21			1.000	-0.214	-0.385	0.074	0.118
Humedad Relativa	1				1.000	-0.416	0.231	-0.312
	3				1.000	-0.406	0.279	-0.351
	7				1.000	-0.219	0.347	-0.259
	14				1.000	-0.154	0.342	-0.236
	21				1.000	-0.120	0.296	-0.297
Radiación Solar	1					1.000	-0.249	0.362
	3					1.000	-0.320	0.297
	7					1.000	-0.440	0.141
	14					1.000	-0.475	0.080
	21					1.000	-0.432	0.060
Nivel	1						1.000	-0.087
	3						1.000	0.022
	7						1.000	0.030
	14						1.000	0.061
	21						1.000	0.067
Velocidad del Viento	1							1.000
	3							1.000
	7							1.000
	14							1.000
	21							1.000

3.2.1 Resultados del modelo de regresión dinámica

A efectos de analizar la incidencia de las condiciones meteorológicas sobre la calidad del agua, se define en primera instancia un modelo de regresión dinámica. Con este modelo se busca poder comprender el aporte que el cambio climático a lo largo del tiempo produce en la variabilidad del IET.

En cuanto a la estrategia, en este caso se busca comprender la relación *individual*, del IET, como medida de la calidad del agua, con las variables predictoras hidrológicas y meteorológicas, para cada punto de muestreo. Vale decir que, en principio se generará un modelo por cada

¹³ Se utilizó el coeficiente de Pearson (el análisis de los diagramas de dispersión no revela relaciones no lineales evidentes entre los pares de variables).

combinación de la variable respuesta, punto de muestreo y alcance de las variables meteorológicas e hidrológicas (inmediato, muy corto, corto, mediano y largo plazo), de manera tal que sean comparables entre sí, y puedan evaluarse tanto bondad de ajuste como relación entre los coeficientes dentro del mismo modelo y entre los distintos modelos.

A continuación, se presentarán los resultados de la aplicación de la metodología descrita en 2.7.1.4 sobre las diez series multivariadas por sitio y perfil. En primera instancia se analizarán tanto los criterios de selección de modelos como el diagnóstico de residuos, indicando el modelo elegido. En segunda instancia, se analizará la definición del modelo para el caso seleccionado.

La Tabla 17, muestra los resultados en términos de bondad de ajuste por medio de AICc (criterio de información de Akaike corregido, como se describe en 2.7.1.3), RMSE (raíz cuadrada del error cuadrático medio, ver definición en 2.7.4) y autocorrelación de residuos obtenidos (p-valor del test de Ljung-Box, remitirse a 2.7.1.3) para los cuatros perfiles del sitio Centro. Como se puede observar comparando RMSE en el caso de C1 el alcance que permite el mejor ajuste es 3, mientras que en C2 es 1. Por otro lado, en C5, al alcance que permite el mejor ajuste es 21. En estos tres casos el test de autocorrelación de los residuos no es significativo con el 95% de confianza. En el caso de C4, el test es significativo para todos los alcances. Se realizó la prueba de remover variables gradualmente, hasta que contemplando únicamente dos variables se consigue el límite de significatividad del test para el alcance 1 con variables radiación solar y precipitación.

Tabla 17 Resumen de modelos de regresión dinámica ajustados para el sitio Centro

Sitio	Alcance	ARIMA (p,d,q)(P,D,Q)	AICc	Prueba de Ljung_Box	RMSE
C1	1	(3,1,1)(0,0,0)[12]	483.85	0.58	1.05
C1	3	(1,1,4)(0,0,0)[12]	493.29	0.74	1.03
C1	7	(1,1,4)(0,0,0)[12]	501.19	0.70	1.06
C1	14	(1,1,4)(0,0,0)[12]	500.16	0.63	1.26
C1	21	(1,1,4)(0,0,0)[12]	501.15	0.60	1.20
C2	1	(3,1,1)(0,0,0)[12]	469.17	0.24	0.64
C2	3	(0,1,5)(0,0,0)[12]	473.60	0.14	0.88
C2	7	(3,1,1)(0,0,0)[12]	477.66	0.17	0.79
C2	14	(3,1,1)(0,0,0)[12]	473.49	0.25	1.05
C2	21	(3,1,1)(0,0,0)[12]	477.70	0.28	0.90
C4	1	(2,0,2)(0,0,1)[12]	437.96	0.00	0.68
C4	3	(0,0,3)(0,0,2)[12]	445.20	0.00	0.68
C4	7	(0,0,3)(0,0,2)[12]	444.74	0.00	0.69
C4	14	(0,0,3)(0,0,2)[12]	445.02	0.00	0.70
C4	21	(0,0,3)(0,0,2)[12]	447.91	0.00	0.69
C4	1	(2,0,2)(0,0,1)[12]	433.33	0.05	0.82
C5	1	(0,1,3)(0,0,0)[12]	486.26	0.02	1.09
C5	3	(0,1,1)(0,0,0)[12]	490.39	0.01	1.19
C5	7	(0,1,1)(0,0,0)[12]	490.15	0.01	0.96
C5	14	(0,1,3)(0,0,0)[12]	491.71	0.07	0.91
C5	21	(0,1,1)(0,0,0)[12]	494.20	0.09	0.81

La Fig. 66 muestra el diagnóstico de residuos que complementa el test de Ljung-Box. En líneas generales no se observan patrones en la serie de errores, desarrollándose en la franja de 4 desvíos estándar, incluso menos en algunos casos (C4 y C2). El gráfico de autocorrelación no muestra retardos significativos. En lo que se refiere a la normalidad, podría decirse que la distribución de los errores es aproximadamente normal en C2, C5, mientras que en C1 y C4 se observa alguna irregularidad, aunque la distribución es aproximadamente simétrica.

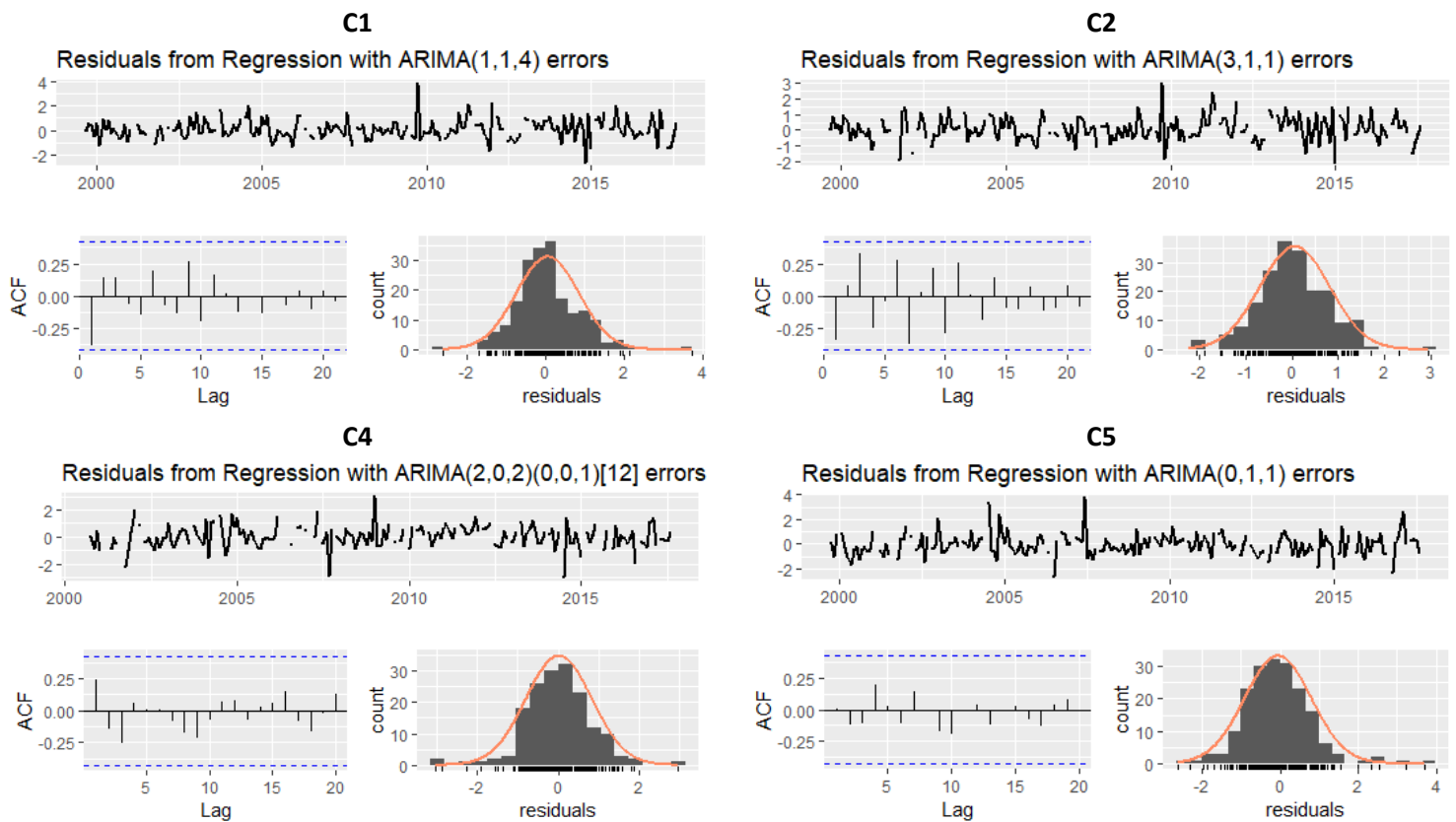


Fig. 66 Diagnóstico de residuos de los modelos seleccionados del sitio Centro

Por otra parte, la Fig. 67 muestra la curva de predicción de los modelos seleccionados del sitio Centro, junto con los intervalos de confianza del 80 y 95% y la curva real. Se puede notar que en todos los casos a excepción del pico en C1 la curva real queda contenida en los intervalos de confianza y la curva predicha sigue las variaciones de la curva real.

Análogamente, la tabla Tabla 18 muestra los resultados correspondientes a TAC. En este caso el modelo elegido considerando RMSE para TAC1 y TAC2 tiene alcance 1, mientras que para TAC4 y TAC5 tiene alcance 7. En todos los casos el test no es significativo con el 95% de confianza.

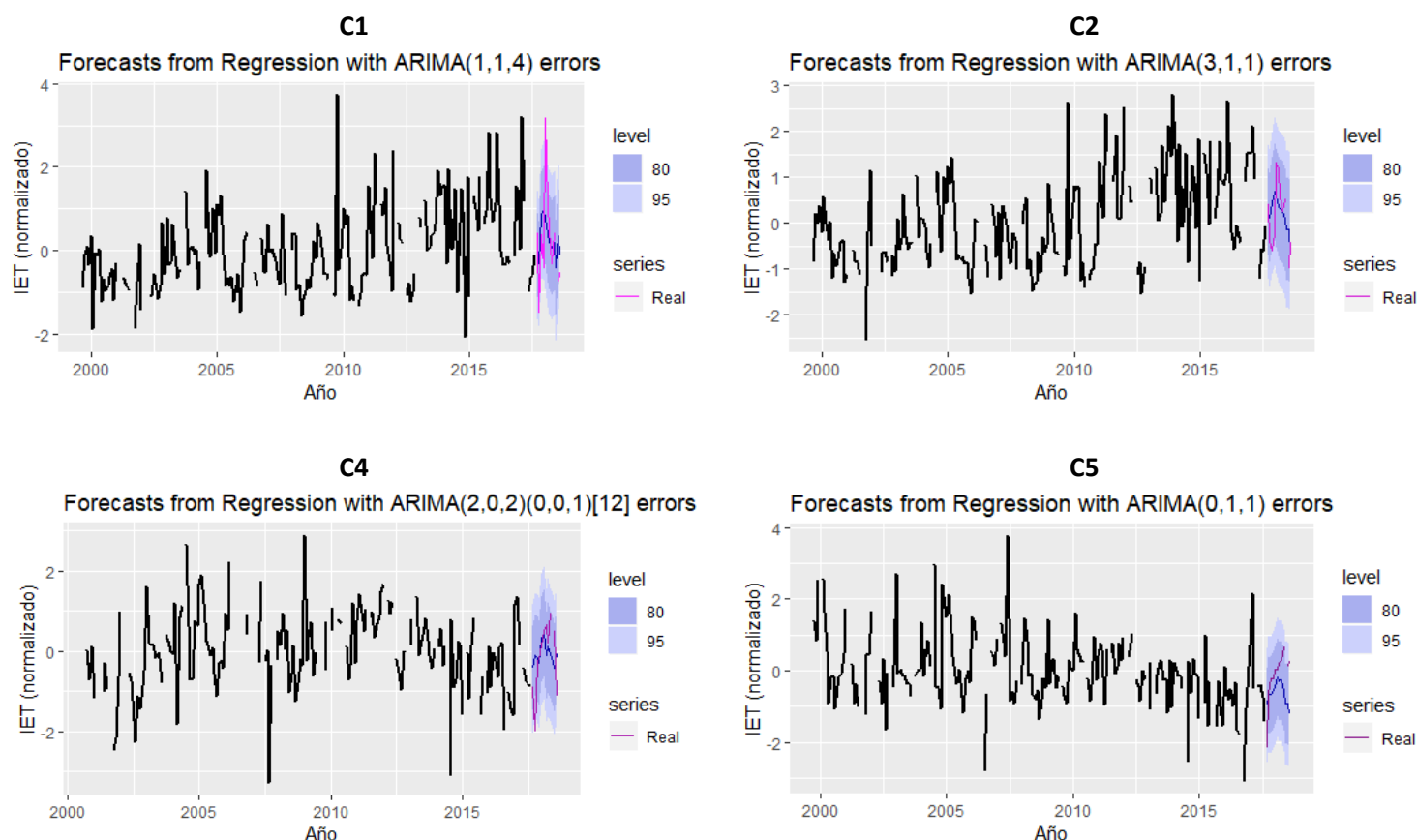


Fig. 67 Predicciones de los modelos seleccionados del sitio Centro

Tabla 18 Resumen de modelos de regresión dinámica ajustados para el sitio TAC

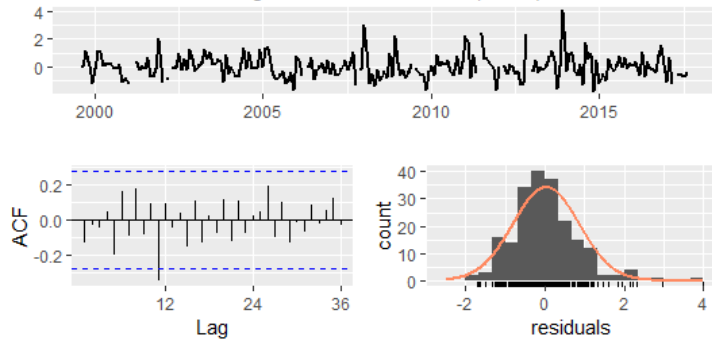
Sitio	Alcance	ARIMA	AICc	Prueba de Ljung_Box	RMSE
TAC1	1	(0,1,1)(0,0,0)[12]	528.62	0.06	0.88
TAC1	3	(0,1,1)(0,0,0)[12]	525.86	0.03	1.03
TAC1	7	(0,1,1)(0,0,0)[12]	536.38	0.26	0.97
TAC1	14	(0,1,1)(0,0,0)[12]	540.26	0.15	1.00
TAC1	21	(1,0,1)(0,0,0)[12]	536.22	0.07	0.99
TAC2	1	(1,1,2)(0,0,0)[12]	516.91	0.17	0.76
TAC2	3	(0,1,1)(0,0,0)[12]	515.21	0.05	0.95
TAC2	7	(0,1,1)(0,0,0)[12]	529.42	0.38	0.89
TAC2	14	(0,1,1)(0,0,1)[12]	527.50	0.37	0.86
TAC2	21	(0,1,1)(0,0,1)[12]	519.94	0.35	0.93
TAC4	1	(1,1,1)(0,0,2)[12]	476.32	0.17	0.82
TAC4	3	(0,1,2)(0,0,2)[12]	483.53	0.06	1.02
TAC4	7	(1,0,0)(0,0,1)[12]	487.51	0.06	0.58
TAC4	14	(1,0,0)(0,0,1)[12]	490.54	0.08	0.63
TAC4	21	(1,0,0)(0,0,1)[12]	490.47	0.09	0.62
TAC5	1	(0,1,3)(0,0,0)[12]	485.87	0.07	0.77
TAC5	3	(0,1,3)(1,0,1)[12]	479.99	0.02	0.84
TAC5	7	(0,0,2)(1,0,0)[12]	472.96	0.18	0.60
TAC5	14	(0,0,2)(1,0,0)[12]	473.27	0.10	0.71
TAC5	21	(0,0,2)(1,0,0)[12]	471.66	0.19	0.78

En lo que se refiere al diagnóstico complementario de residuos, la Fig. 68 muestra series de errores sin patrones evidentes, habiendo mayor variabilidad en TAC1 y TAC2 con hasta 4 desvíos

estándar, que en TAC4 y TAC5, que tienen residuos en la franja de más y menos 2 desvíos estándar. En TAC1 se observa autocorrelación significativa con retardo 11, pero no así a nivel global como indica el test. Con respecto a la normalidad, se observa cola a derecha en TAC1 y TAC2 mientras que en TAC4 y TAC5 la distribución de los errores es aproximadamente normal.

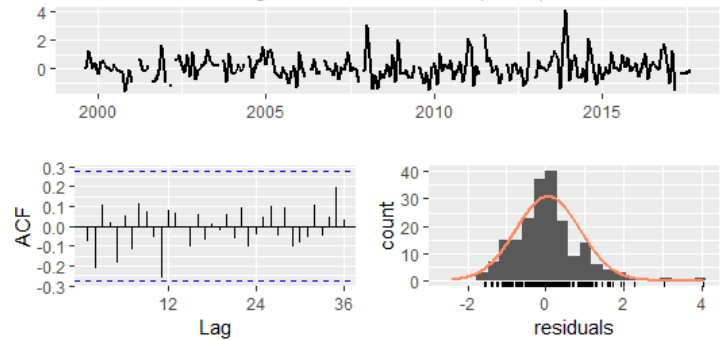
TAC1

Residuals from Regression with ARIMA(0,1,1) errors



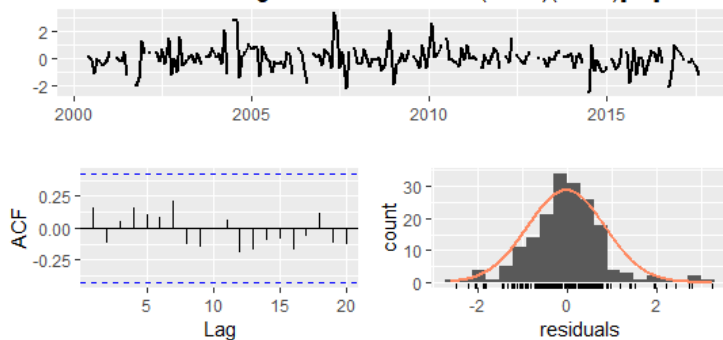
TAC2

Residuals from Regression with ARIMA(1,1,2) errors



TAC4

Residuals from Regression with ARIMA(1,0,0)(0,0,1)[12] errors



TAC5

Residuals from Regression with ARIMA(0,0,2)(1,0,0)[12] errors

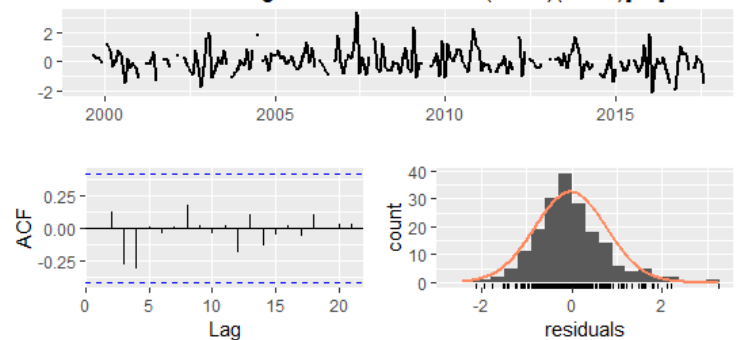


Fig. 68 Diagnóstico de residuos de los modelos seleccionados del sitio TAC

La Fig. 69 muestra la curva de predicción de los modelos seleccionados del sitio TAC junto con los intervalos de confianza correspondientes y la curva real. Como se puede observar en todos los casos la curva real queda contenida en los intervalos de confianza y mayormente la curva de predicción sigue a la curva real. En el caso de TAC4 y TAC5 se observa una buena aproximación sobre todo en los primeros meses de la estimación.

La Tabla 19 muestra los resultados para el sitio DCQ, donde se observa que el test de Ljung-Box es no significativo únicamente con alcance 1.

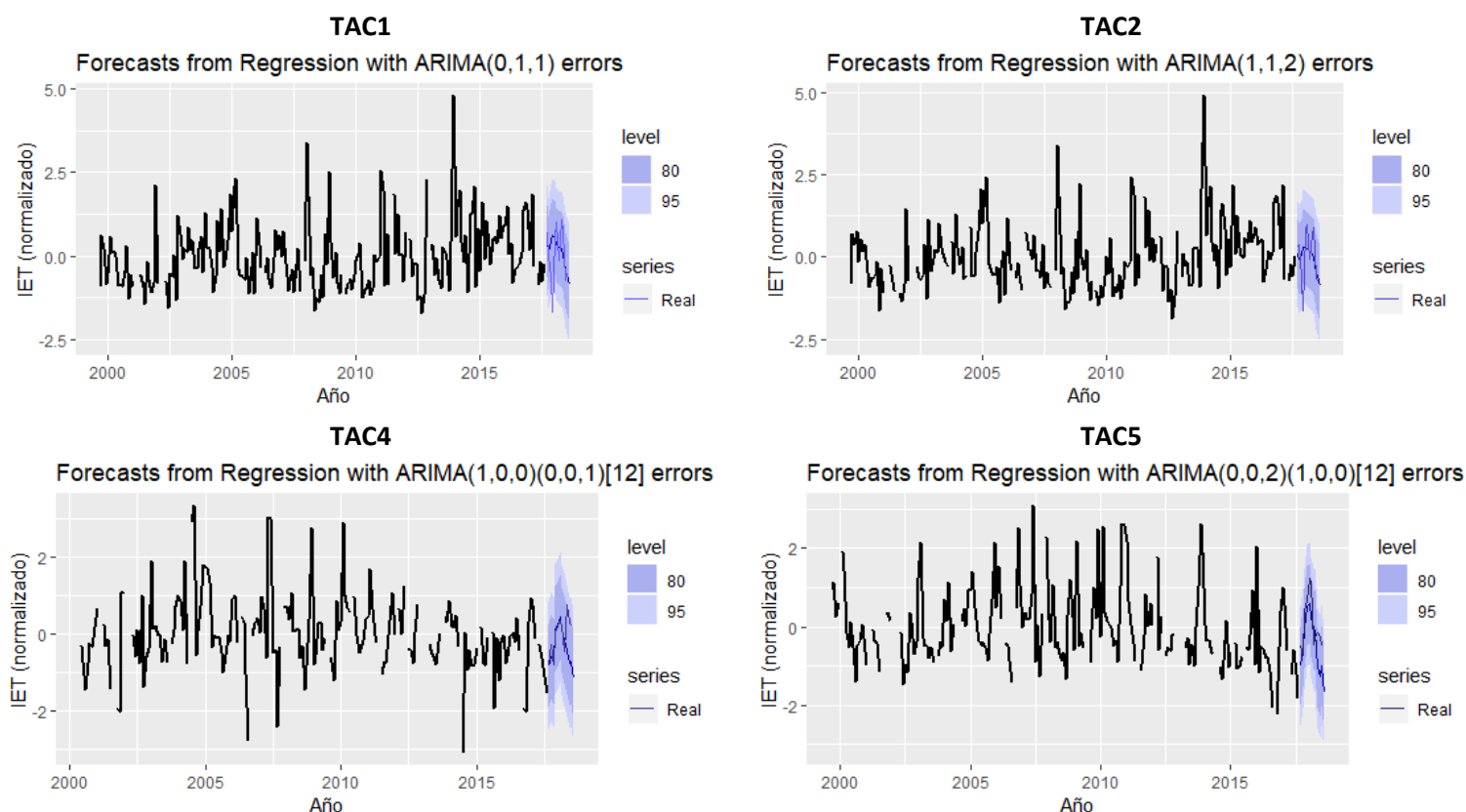


Fig. 69 Predicciones de los modelos seleccionados del sitio TAC

Tabla 19 Resumen de modelos de regresión dinámica ajustados para el sitio DCQ

Sitio	Alcance	ARIMA	AICc	Prueba de Ljung_Box	RMSE
DCQ1	1	(3,1,2)(0,0,0)[12]	485.25	0.06	0.90
DCQ1	3	(2,1,3)(0,0,0)[12]	487.05	0.03	1.06
DCQ1	7	(3,1,2)(0,0,0)[12]	498.86	0.02	0.85
DCQ1	14	(3,1,2)(0,0,0)[12]	498.42	0.04	0.80
DCQ1	21	(1,1,4)(0,0,0)[12]	493.93	0.04	0.80

En la Fig. 70 se observan los gráficos del diagnóstico de residuos. Con respecto a la serie de errores no se ven patrones y la varianza se observa estable. Existe únicamente correlación positiva significativa en el retardo 4, pero no a nivel global. Por último, la distribución se observa aproximadamente normal.

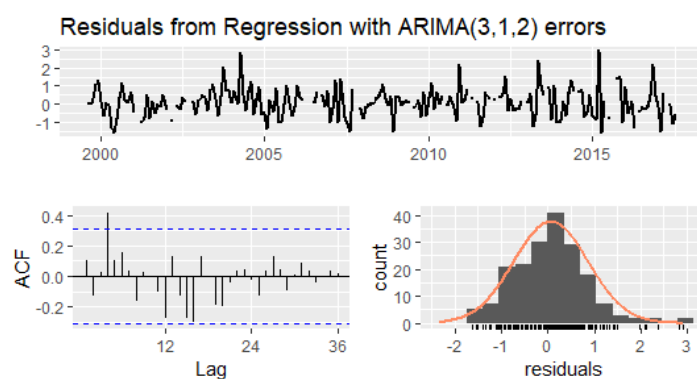


Fig. 70 Diagnóstico de residuos del modelo seleccionado del sitio DCQ (DCQ1)

La Fig. 71 muestra la curva de predicción del modelo seleccionado del sitio DCQ junto con los intervalos de confianza correspondientes y la curva real. La misma se observa casi enteramente contenida en los intervalos de confianza, siguiendo la curva predicha el patrón general de variación.

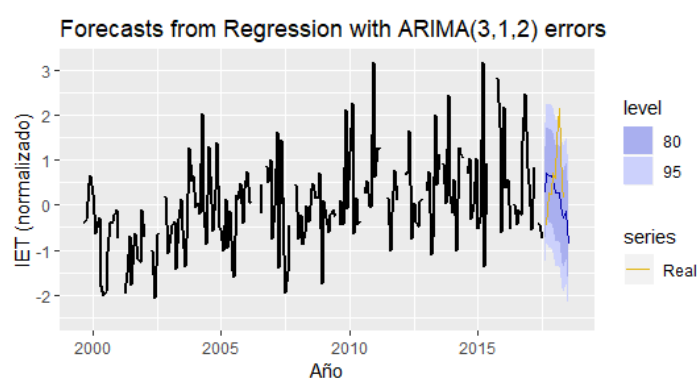


Fig. 71 Predicciones del modelo seleccionado del sitio DCQ (DCQ1)

Por último, para el sitio DSA, el diagnóstico de residuos es significativo (con 95% de confianza) en todos los alcances, bajo las mismas condiciones que en el resto de los modelos, como se puede observar en la Tabla 20. En este caso, como antes para C4, se hizo la prueba de quitar variables, vale decir no ajustar el modelo utilizando la totalidad de las 7. En este sentido se probó en primera instancia quitando una variable por vez y recalculando el modelo para los 5 alcances, sin resultados satisfactorios en términos de no significatividad del test de Ljung-Box. Seguidamente, se probó con combinaciones de 5 variables, y en este caso sí el test resultó no significativo en 7 combinaciones, y de entre ellas se eligió la que tenía RMSE más bajo, resaltada en azul en la Tabla 20.

Tabla 20 Resumen de modelos de regresión dinámica ajustados para el sitio DSA

Sitio	Alcance	ARIMA	AICc	Prueba de Ljung_Box	RMSE
DSA1	1	(2,1,1)(2,0,0)[12]	510.75	0.00	0.75
DSA1	3	(2,1,1)(2,0,0)[12]	521.21	0.00	0.69
DSA1	7	(2,1,1)(2,0,0)[12]	526.92	0.02	0.78
DSA1	14	(2,1,1)(2,0,0)[12]	529.91	0.01	0.92
DSA1	21	(2,1,1)(2,0,0)[12]	529.28	0.02	0.84
DSA1	7	(2,1,1)(2,0,0)[12]	522.49	0.05	0.77

Por último, en término del diagnóstico de errores, como se puede ver en la Fig. 72, se observa que la serie de errores es aproximadamente homosedástica, no se observa autocorrelación significativa en ningún retardo y la distribución de errores se presenta aproximadamente normal.

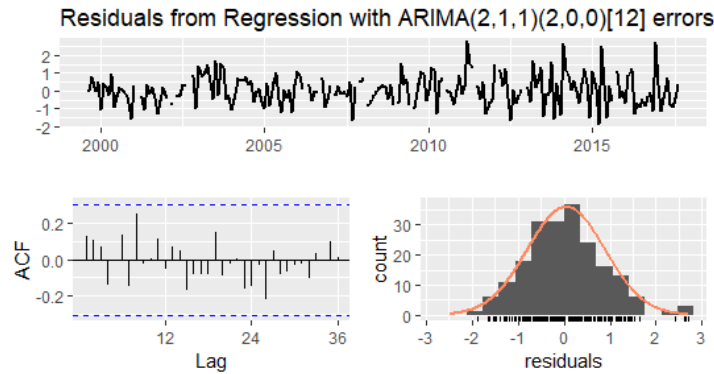


Fig. 72 Diagnóstico de residuos del modelo seleccionado del sitio DSA (DSA1)

Por otra parte, la Fig. 73 muestra la curva de predicción del modelo seleccionado para el sitio DSA junto con los intervalos de confianza correspondientes y la curva real. La curva de predicción parece ajustar bien la curva real, siempre contenida en los intervalos de confianza.

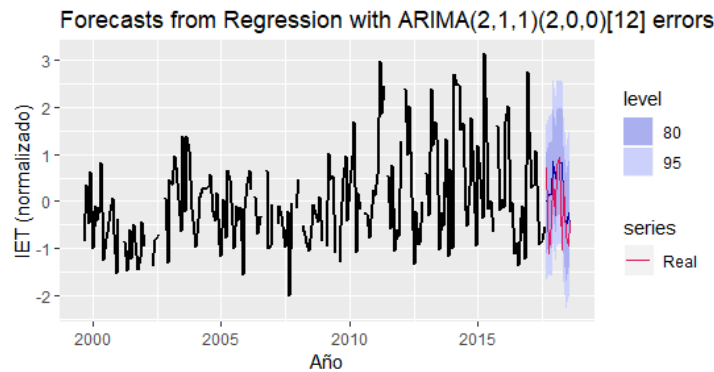


Fig. 73 Predicciones del modelo seleccionado del sitio DSA (DSA1)

A continuación, se presentan las Ecuaciones de los modelos resultantes (en pares: siguiendo las Ecuaciones generales (5) y (6)): C1 (21) y (22); C2 (23) y (24); C4 (25) y (26); C5 (27) y (28); TAC1 (29) y (30); TAC2 (31) y (32); TAC4 (33) y (34); TAC5 (35) y (36); DCQ1 (37) y (38); y DSA1 (39) y (40). Cabe recordar que cada una de las series intervinientes fue previamente estandarizada como se indicó en la metodología (ver sección 2.7.1.4). En color azul se encuentran resaltados aquellos coeficientes que resultan significativos a partir de la prueba t con el 95% de confianza (ver sección 2.7.1.5). Vale aclarar, además, que en los casos en que las series se han diferenciado (siempre una sola vez), los términos correspondientes en el modelo aparecen acompañados por el símbolo “prima” (') y el mismo se denomina “modelo en diferencias”.

$$IET_t(C1)' = -0,05Nv_t^{3'} + 0,51Tm_t^{3'} + 0,17PA_t^{3'} - 0,15HR_t^{3'} - 0,11RS_t^{3'} + 0,07Pr_t^{3'} - 0,08Vv_t^{3'} + \eta_t(C1)' \quad (21)$$

$$\eta_t(C1)' = 0,51\eta_{t-1}' - 0,25\varepsilon_{t-4} + 0,10\varepsilon_{t-3} + 0,65\varepsilon_{t-2} - 1,47\varepsilon_{t-1} + \varepsilon_t, \quad \varepsilon_t \sim N(0, 0,60) \quad (22)$$

$$IET_t(C2)' = -0,07Nv_t^{1'} + 0,42Tm_t^{1'} + 0,07PA_t^{1'} - 0,05HR_t^{1'} - 0,08RS_t^{1'} + 0,13Pr_t^{1'} - 0,11Vv_t^{1'} + \eta_t(C2)' \quad (23)$$

$$\eta_t(C2)' = 0,23\eta_{t-1}' + 0,17\eta_{t-2}' + 0,20\eta_{t-3}' - 0,97\varepsilon_{t-1} + \varepsilon_t, \quad \varepsilon_t \sim N(0, 0,55) \quad (24)$$

$$IET_t(C4) = 0,07RS_t^1 + 0,21Pr_t^1 + \eta_t(C4) \quad (25)$$

$$\eta_t(C4) = 1,09\eta_{t-1} - 0,71\eta_{t-2} + 0,44\varepsilon_{t-12} + 0,78\varepsilon_{t-2} - 0,85\varepsilon_{t-1} + \varepsilon_t, \quad \varepsilon_t \sim N(0, 0,56) \quad (26)$$

$$IET_t(C5)' = 0,25Nv_t^{21'} + 0,16Tm_t^{21'} - 0,08PA_t^{21'} + 0,07HR_t^{21'} + 0,03RS_t^{21'} + 0,16Pr_t^{21'} + 0,02Vv_t^{21'} + \eta_t(C5)' \quad (27)$$

$$\eta_t(C5)' = -0,88\varepsilon_{t-1} + \varepsilon_t, \quad \varepsilon_t \sim N(0, 0,67) \quad (28)$$

$$IET_t(TAC1)' = -0,05Nv_t^{1'} + 0,55Tm_t^{1'} + 0,06PA_t^{1'} - 0,07HR_t^{1'} - 0,10RS_t^{1'} + 0,15Pr_t^{1'} - 0,09Vv_t^{1'} + \eta_t(TAC1)' \quad (29)$$

$$\eta_t(TAC1)' = -0,89\varepsilon_{t-1} + \varepsilon_t, \quad \varepsilon_t \sim N(0, 0,70) \quad (30)$$

$$IET_t(TAC2)' = 0,02Nv_t^{1'} + 0,41Tm_t^{1'} - 0,00PA_t^{1'} + 0,01HR_t^{1'} - 0,02RS_t^{1'} + 0,19Pr_t^{1'} - 0,09Vv_t^{1'} + \eta_t(TAC2)' \quad (31)$$

$$\eta_t(TAC2)' = 0,82\eta_{t-1}' + 0,59\varepsilon_{t-2} - 1,59\varepsilon_{t-1} + \varepsilon_t, \quad \varepsilon_t \sim N(0, 0,66) \quad (32)$$

$$IET_t(TAC4) = 0,11Nv_t^7 + 0,17Tm_t^7 - 0,01PA_t^7 + 0,12HR_t^7 + 0,10RS_t^7 + 0,11Pr_t^7 - 0,07Vv_t^7 + \eta_t(TAC4) \quad (33)$$

$$\eta_t(TAC4) = 0,30\eta_{t-1} + 0,25\varepsilon_{t-12} + \varepsilon_t, \quad \varepsilon_t \sim N(0, 0,68) \quad (34)$$

$$IET_t(TAC5) = -0,23Nv_t^7 + 0,28Tm_t^7 - 0,14PA_t^7 + 0,19HR_t^7 + 0,06RS_t^7 - 0,08Pr_t^7 - 0,10Vv_t^7 + \eta_t(TAC5) \quad (35)$$

$$\eta_t(TAC5) = 0,17\eta_{t-12} + 0,19\varepsilon_{t-2} + 0,20\varepsilon_{t-1} + \varepsilon_t, \quad \varepsilon_t \sim N(0, 0,58) \quad (36)$$

$$IET_t(DCQ1)' = -0,30Nv_t^{1'} + 0,21Tm_t^{1'} - 0,04PA_t^{1'} + 0,11HR_t^{1'} - 0,13RS_t^{1'} - 0,00Pr_t^{1'} - 0,16Vv_t^{1'} + \eta_t(DCQ1)' \quad (37)$$

$$\eta_t(DCQ1)' = -0,74\eta_{t-1}' + 0,03\eta_{t-2}' + 0,23\eta_{t-3}' - 0,62\varepsilon_{t-2} - 0,28\varepsilon_{t-1} + \varepsilon_t, \quad \varepsilon_t \sim N(0, 0,61) \quad (38)$$

$$IET_t(DSA1)' = 0,30Tm_t^{3'} + 0,19PA_t^{3'} - 0,02RS_t^{3'} - 0,05Pr_t^{3'} - 0,05Vv_t^{3'} + \eta_t(DSA1)' \quad (39)$$

$$\eta_t(DSA1)' = 0,18\eta_{t-1}' + 0,21\eta_{t-2}' + 0,16\eta_{t-12}' + 0,18\eta_{t-24}' - 0,98\varepsilon_{t-1} + \varepsilon_t, \quad \varepsilon_t \sim N(0, 0,69) \quad (40)$$

3.2.2 Resultados del modelo de espacio de estados multivariado

Como alternativa a los modelos de regresión dinámica, se tienen los modelos de espacios de estados, que permiten combinar en un único modelo datos provenientes de distintos puntos de muestreo para estimar tendencias, como se aplica en Holmes et al. (2018a). En este sentido, el objetivo es modelar la dinámica del IET promedio en los perfiles superficiales y el basado en fósforo total en los perfiles más profundos, asumiendo distintas condiciones para los procesos estocásticos que gobiernan los sitios en que se desarrollan. En términos de la especificación en la sección 2.7.2, el proceso de observaciones estará determinado por la variable respuesta IET, tal como se introdujo al comienzo de este capítulo y variables hidrológicas y meteorológicas como covariables (las mismas empleadas en la sección anterior, 3.2.1). Concretamente, se ensayaron tres tipos de modelo, dos correspondientes al modelado en profundidad: un modelo dedicado al centro del lago y otro al área de presa TAC, y un tercero dedicado al modelado horizontal a nivel subsuperficial, que incluye los cuatro sitios de muestreo. En lo que respecta a las variables meteorológicas e hidrológicas se emplearon los alcances que produjeron los mejores modelos en la sección 3.2.1. Por último, se ajustaron 144 modelos en cada caso, adaptando los criterios que se mencionan a continuación para los parámetros de las Ecuaciones (10), (11) y (12):

1. Coeficientes todos distintos: n procesos distintos, siendo n el número de puntos incluidos en el modelo.
2. Coeficientes todos iguales: proceso único
3. Coeficientes compartidos:
 - a. Modelos de profundidad: un proceso modela la región próxima a la superficie mientras que otro proceso modela la región próxima al fondo.
 - b. Modelo subsuperficial: un proceso modela la región de desembocadura de los ríos tributarios mientras que otro proceso modela la región central del lago (Centro y TAC).
4. Coeficientes iguales para varianza y coeficientes iguales para covarianza

En lo que respecta a los coeficientes de los procesos, matrices B (proceso de estados) y Z (proceso observacional), se aplicaron los criterios del 1 al 3; mientras que, para los coeficientes de la varianza de los errores asociados, matrices Q (proceso de estados) y R (proceso observacional), se aplicaron los criterios del 1 al 4. En lo que respecta a los coeficientes de los errores propiamente dicho, se asumen 1. En lo relativo a los demás parámetros, a y u se asumen cero (puesto que los datos están normalizados no se tienen interceptos), mientras que los

puntos de comienzo de las trayectorias se asumen todos distintos comenzando en t_1 . Por último, en cuando a C y D (coeficientes de las covariables en los procesos de estado y observacional, respectivamente), se asumen que las covariables influncian al proceso observacional, por lo tanto, d contiene las covariables y D contiene los coeficientes no restringidos (cada par de valores: punto de muestreo-variable meteorológica puede adoptar cualquier valor), mientras que c y C son cero. De esta manera, los modelos ensayados adoptan la forma de las Ecuaciones (41), (42) y (43).

$$x_t = B_t x_{t-1} + w_t, \quad w_t \sim NM(0, Q_t) \quad (41)$$

$$y_t = Z_t x_t + D_t d_t + v_t, \quad v_t \sim NM(0, R_t) \quad (42)$$

$$x_1 = NM(\pi, \Lambda) \quad (43)$$

De esta manera, se tiene un proceso de estados autorregresivo de orden 1, que afecta a un proceso observacional a su vez influenciado por una serie de covariables.

Adicionalmente, la cantidad máxima de iteraciones se fijó en 2500. De entre los 144 resultados obtenidos se filtraron únicamente aquellos modelos que alcanzaron la convergencia.

La Tabla 21 muestra la parametrización de los primeros modelos ordenados por AICc para el caso del sitio Centro, que cumplen el requisito de convergencia y presentan el mejor diagnóstico de residuos.

Tabla 21 Parametrización de los primeros modelos de espacios de estados multivariados para el sitio Centro, ordenados por AICc

Param. Z	Param. B	Param. Q	Param. R	AICc
3	3	4	3	1771.33
3	2	4	4	1787.88
1	2	4	4	1791.19

Nota: los números para cada parámetro hacen referencia a los criterios mencionados anteriormente.

Como se puede observar, los mejores resultados se obtienen asumiendo errores del proceso de estado (Q) donde la varianza es la misma y la covarianza también. Por lo demás, la configuración varía para los demás parámetros siendo más parecidos los segundos dos modelos entre sí que con respecto al primero. La diferencia en términos de AICc es de más de 16 puntos entre el primero y el segundo, se opta por elegir el modelo con mejor AICc. Este modelo presenta coeficientes compartidos en todos los parámetros a acepción de Q. La Ecuación (44) describe el modelo resultante para el sitio Centro. En color azul se encuentran resaltados aquellos coeficientes que resultan significativamente distintos de 0 con el 95% de confianza a partir de prueba paramétrica por *bootstrapping* (ver sección 2.7.2.5).

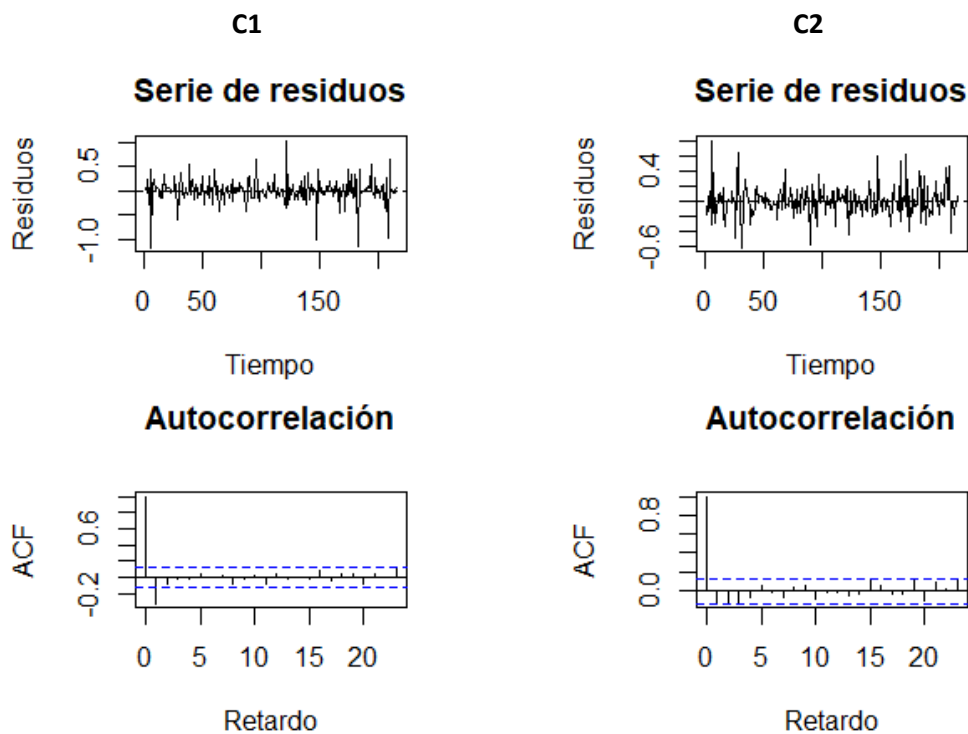
$$diag(B) = \begin{matrix} 0,99 & (C1) \\ 0,99 & (C2) \\ 0,98 & (C4) \\ 0,98 & (C5) \end{matrix} \quad diag(Z) = \begin{matrix} 0,97 & (C1) \\ 0,97 & (C2) \\ 0,45 & (C4) \\ 0,45 & (C5) \end{matrix} \quad (44)$$

$$Q = \begin{matrix} & (C1) & (C2) & (C4) & (C5) \\ \begin{matrix} (C1) \\ (C2) \\ (C4) \\ (C5) \end{matrix} & \begin{matrix} 0,73 \\ 0,71 \\ 0,71 \\ 0,71 \end{matrix} & \begin{matrix} 0,71 \\ 0,73 \\ 0,71 \\ 0,71 \end{matrix} & \begin{matrix} 0,71 \\ 0,71 \\ 0,73 \\ 0,71 \end{matrix} & \begin{matrix} 0,71 \\ 0,71 \\ 0,71 \\ 0,73 \end{matrix} \end{matrix} \begin{matrix} (C1) \\ (C2) \\ (C4) \\ (C5) \end{matrix}$$

$$diag(R) = \begin{matrix} & 0,11 & (C1) \\ & 0,11 & (C2) \\ & 0,62 & (C4) \\ & 0,62 & (C5) \end{matrix} \quad x_1 = \begin{matrix} -1,07 & (C1) \\ -0,85 & (C2) \\ -2,31 & (C4) \\ -0,00 & (C5) \end{matrix}$$

$$C = \begin{matrix} & (HR) & (Nv) & (PA) & (Pr) & (RS) & (Tm) & (VV) \\ \begin{matrix} (C1)(3) \\ (C2)(1) \\ (C4)(1) \\ (C5)(21) \end{matrix} & \begin{matrix} 0,004 \\ 0,057 \\ 0,120 \\ 0,180 \end{matrix} & \begin{matrix} -0,110 \\ -0,100 \\ 0,250 \\ 0,240 \end{matrix} & \begin{matrix} 0,021 \\ -0,004 \\ -0,038 \\ -0,130 \end{matrix} & \begin{matrix} 0,026 \\ 0,088 \\ 0,200 \\ 0,150 \end{matrix} & \begin{matrix} 0,062 \\ 0,100 \\ 0,047 \\ 0,150 \end{matrix} & \begin{matrix} 0,27 \\ 0,24 \\ 0,110 \\ -0,003 \end{matrix} & \begin{matrix} -0,089 \\ -0,071 \\ -0,039 \\ 0,008 \end{matrix} \end{matrix}$$

En la Fig. 74 se muestran los resultados del diagnóstico del modelo. Como se puede apreciar, los residuos muestran una varianza aproximadamente constante alrededor de la media cero. No se observa autocorrelación significativa con retardo en los casos de C2, mientras que para C1, C4 y C5 se observa alguna autocorrelación significativa, mayormente en C4. Por otra parte, la Fig. 75 muestra el diagnóstico de residuos del modelo del sitio Centro correspondiente a normalidad. Como se puede observar, en el caso de C4 y C5 se observa mayor normalidad, mientras que en C1 y C2 se observan desviaciones en los extremos.



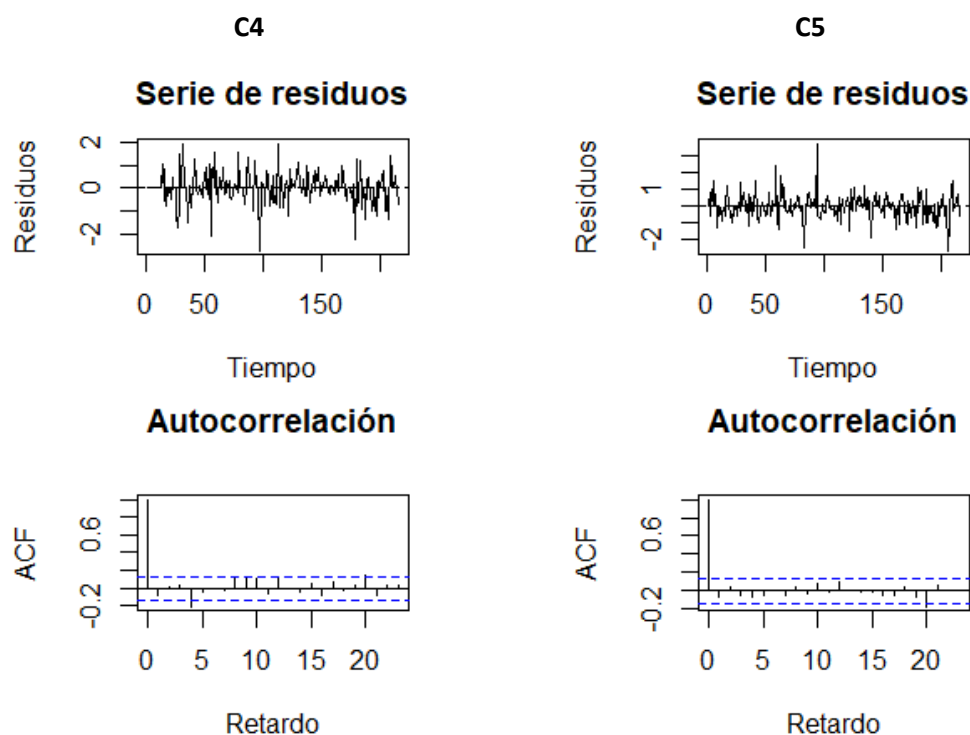


Fig. 74 Diagnóstico de residuos del modelo de espacios multivariado del sitio Centro: varianza y autocorrelación

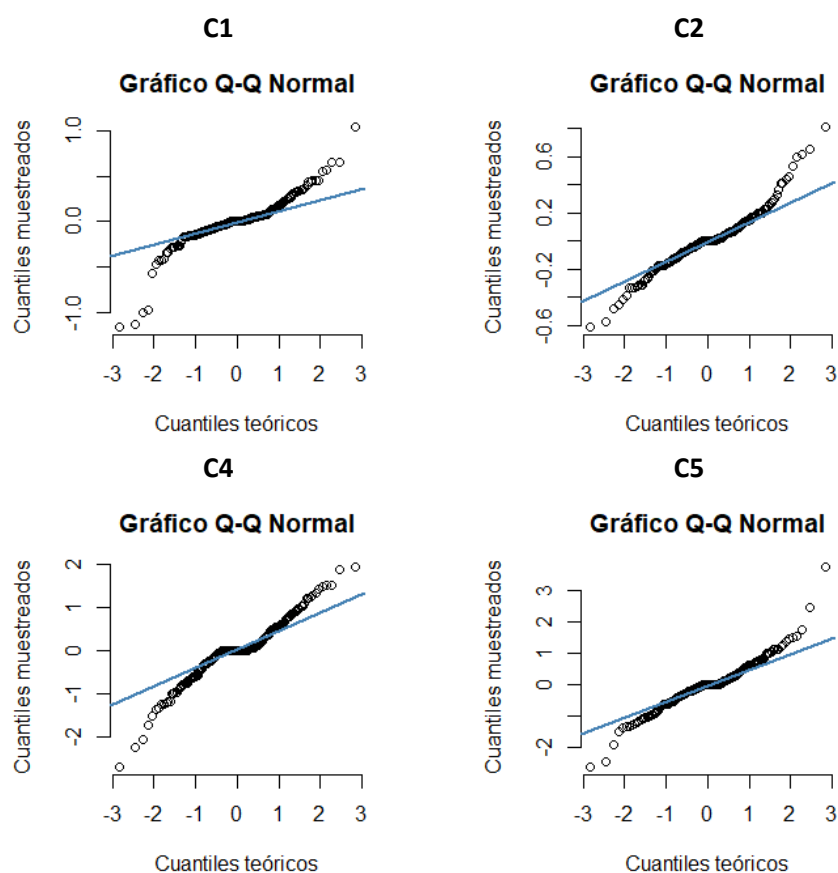


Fig. 75 Diagnóstico de residuos del modelo de espacios multivariados del sitio Centro: normalidad

Por último, la Fig. 76 muestra las curvas de ajuste resultantes del modelo junto con la serie de datos original en cada punto de muestreo del sitio Centro. Ambas curvas están representadas en dos tramos: para el ajuste, la parte negra corresponde al suavizado de Kalman, mientras que la parte más clara y en color corresponde a la predicción basada en Kalman; por otro lado, en cuanto a la serie original: en color intenso se tiene la serie utilizada para el modelado y en color más tenue la serie utilizada para prueba. Las curvas puenteadas representan los intervalos de confianza del 95%.

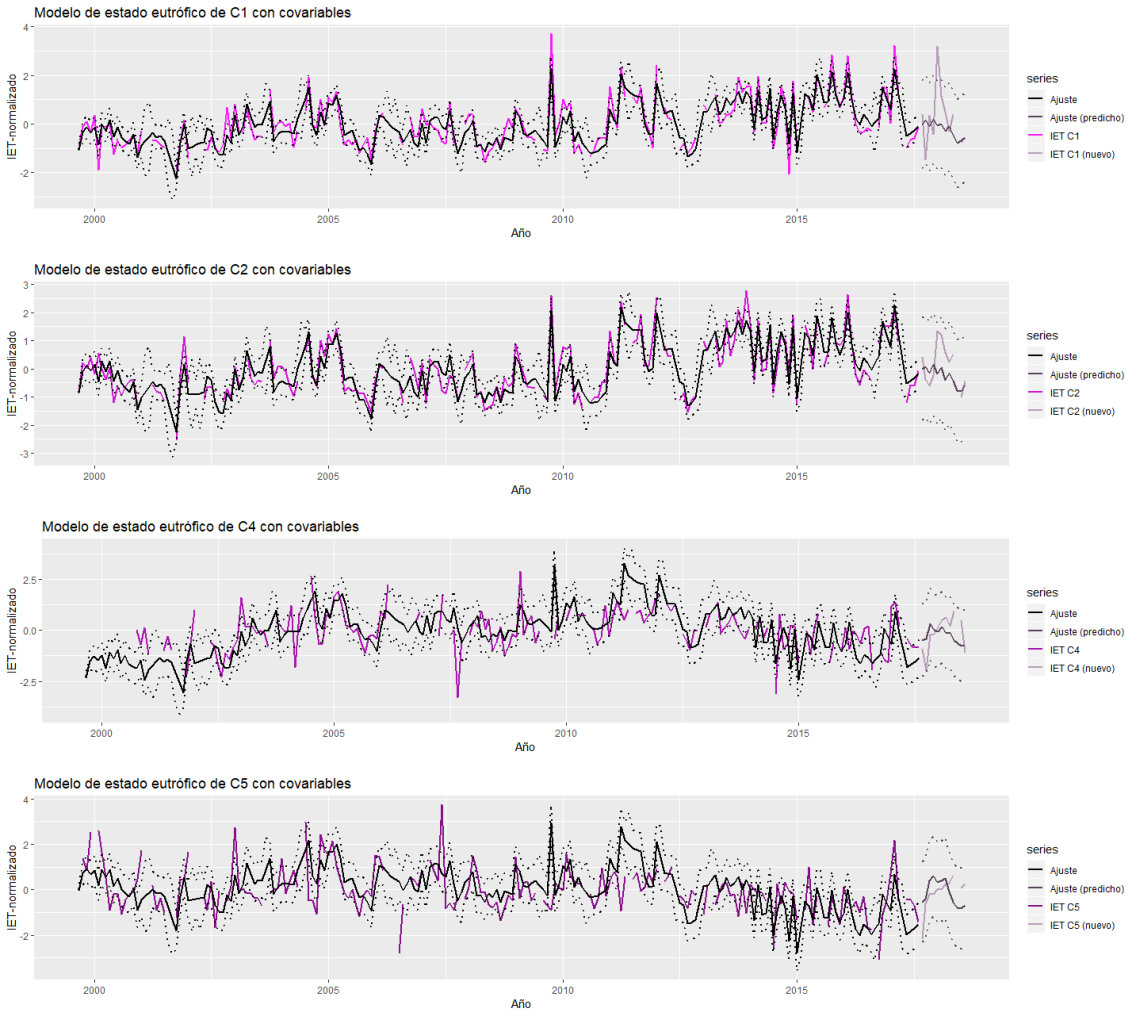


Fig. 76 Curvas de ajuste resultantes del modelo de espacio de estados multivariado junto con las series de datos originales del sitio Centro

La Tabla 22 muestra el error (RMSE) para las distintas respuestas por perfil para el modelo único de espacio de estados para el sitio Centro.

Tabla 22 Error por respuesta del modelo de espacio de estados para el sitio Centro

Perfil	RMSE
C1	1.23
C2	0.76
C4	0.83
C5	0.82

Con respecto al modelo para el sitio TAC, la Tabla 23 muestra los parámetros seleccionados para los diez primeros modelos ordenados por AICc, tal que convergen y presentan los mejores diagnósticos de residuos.

Tabla 23 Parametrización de los diez primeros modelos de espacios de estados multivariados para el sitio TAC, ordenados por AICc

Param. Z	Param. B	Param. Q	Param. R	AICc
2	3	3	4	1876.61
3	3	3	4	1878.79
3	1	2	4	1879.45
2	1	3	4	1879.51
2	3	1	4	1879.87
3	1	3	4	1881.71
1	3	3	4	1882.07
3	3	1	4	1882.07
1	1	2	4	1882.65
2	1	1	4	1882.71

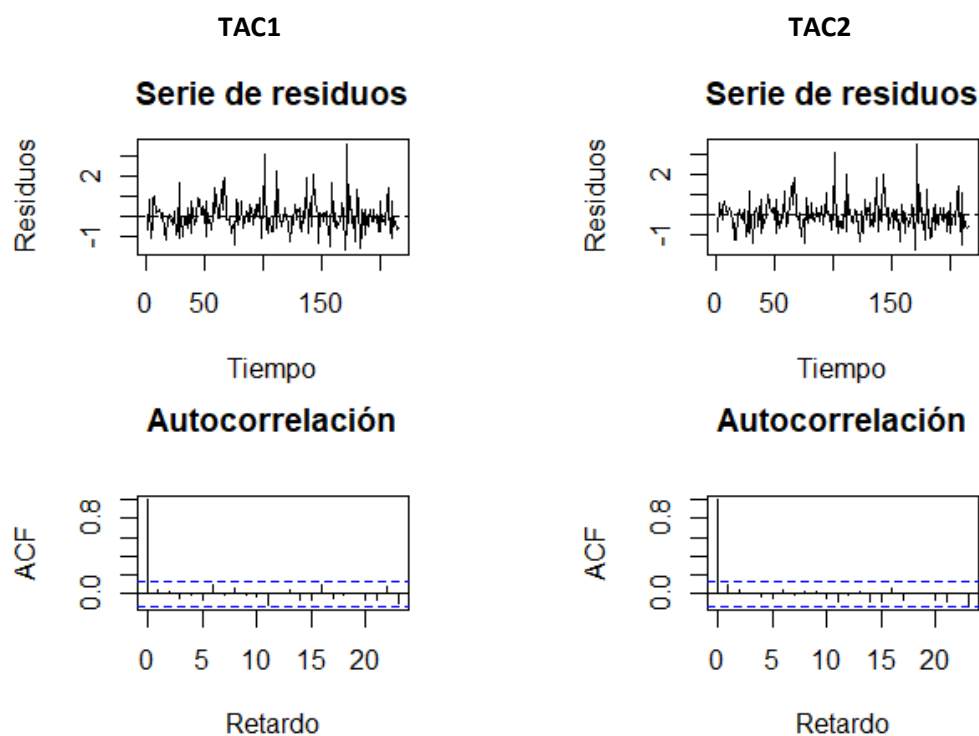
Nota: los números para cada parámetro hacen referencia a los criterios mencionados anteriormente.

En este caso, a diferencia del anterior, los mejores resultados de ajuste se obtienen considerando los errores del proceso observacional (R), a diferencia de los errores del proceso de estado (Q), con varianza igual y covarianza igual para los 10 primeros modelos ordenados por AICc. Los primeros dos lugares, corresponden asimismo a modelos que presentan coeficientes compartidos en el proceso de estados (B) y en los errores del proceso de estados. Por lo demás, el escenario es variable. En cuanto a AICc, se observa poca variación a lo largo de los primeros 10 modelos, habiendo un salto de casi 2 puntos entre el primero y el segundo, y menos de 1 punto entre los siguientes 4. El modelo seleccionado es el de AICc más alto, y asume además coeficientes iguales para el proceso observacional (Z). La Ecuación (45) muestra los coeficientes resultantes de este modelo, resaltando en color azul aquellos que son significativos con el 95% de confianza.

$$\begin{aligned}
 \text{diag}(B) &= \begin{matrix} 0,97 & (TAC1) \\ 0,97 & (TAC2) \\ 0,15 & (TAC4) \\ 0,15 & (TAC5) \end{matrix} & \text{diag}(Z) &= \begin{matrix} 0,98 & (TAC1) \\ 0,98 & (TAC2) \\ 0,98 & (TAC4) \\ 0,98 & (TAC5) \end{matrix} \\
 R &= \begin{matrix} (TAC1) & (TAC2) & (TAC4) & (TAC5) \\ 0,58 & 0,51 & 0,51 & 0,51 & (TAC1) \\ 0,51 & 0,58 & 0,51 & 0,51 & (TAC2) \\ 0,51 & 0,51 & 0,58 & 0,51 & (TAC4) \\ 0,51 & 0,51 & 0,51 & 0,58 & (TAC5) \end{matrix} & & (45) \\
 \text{diag}(Q) &= \begin{matrix} 0,01 & (TAC1) \\ 0,01 & (TAC2) \\ 0,96 & (TAC4) \\ 0,96 & (TAC5) \end{matrix} & x_1 &= \begin{matrix} -0,32 & (TAC1) \\ -0,09 & (TAC2) \\ 0,44 & (TAC4) \\ 1,29 & (TAC5) \end{matrix}
 \end{aligned}$$

	(HR1)	(Nv1)	(PA1)	(Pr1)	(RS1)	(Tm1)	(VV1)	
	0,014	-0,071	0,079	0,076	-0,029	0,510	-0,052	(TAC1)(1)
$C =$	0,058	-0,039	0,034	0,099	0,017	0,390	-0,041	(TAC2)(1)
	0,270	0,140	-0,019	-0,076	0,150	0,110	-0,003	(TAC4)(7)
	0,300	-0,240	-0,110	-0,170	0,078	0,280	0,009	(TAC5)(7)

En la Fig. 77 se muestran los resultados del diagnóstico del modelo. Como se puede observar, los residuos presentan una varianza aproximadamente constante alrededor de la media y no se observa autocorrelación significativa con retardo en ningún caso. Por otro lado, la Fig. 78 muestra el diagnóstico de residuos del modelo del sitio TAC correspondiente a normalidad. Como se puede apreciar, para los distintos puntos de muestreo se observa una distribución aproximadamente normal.



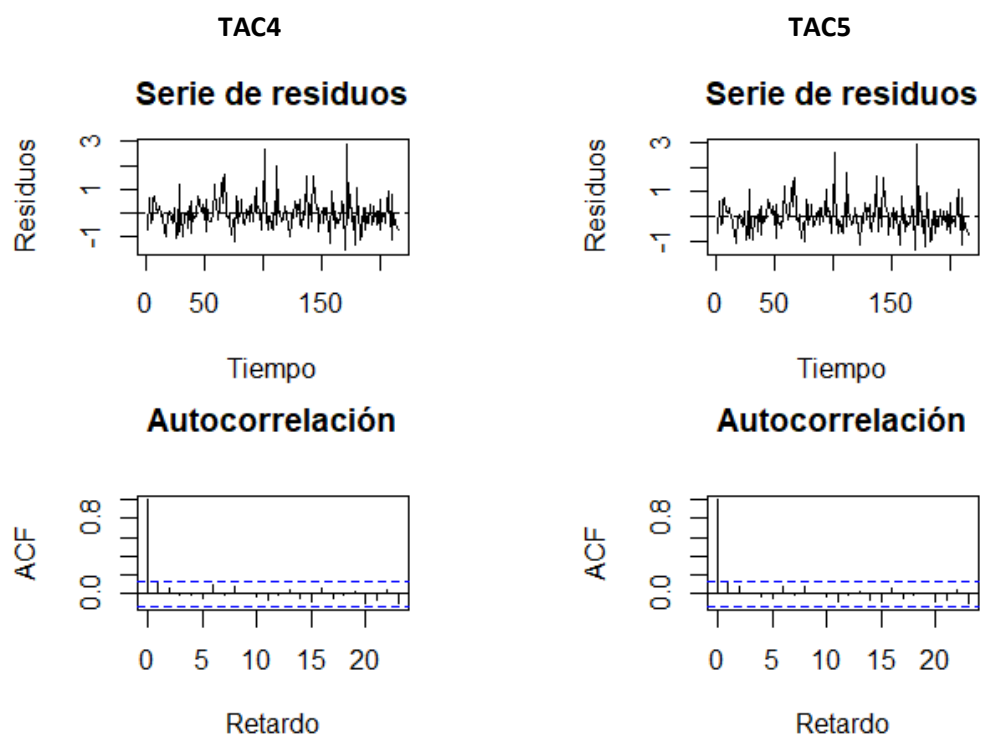


Fig. 77 Diagnóstico de residuos del modelo de espacios multivariado del sitio TAC: varianza y autocorrelación

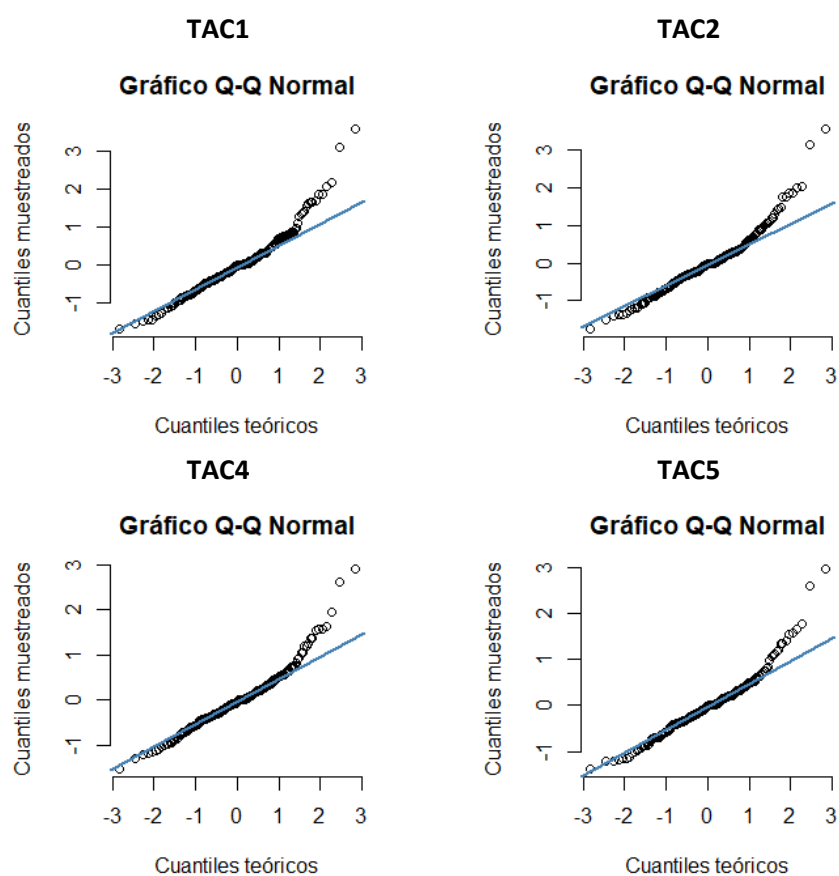


Fig. 78 Diagnóstico de residuos del modelo de espacios multivariados del sitio TAC: normalidad

Finalmente, la Fig. 79 muestra las curvas de ajuste resultantes del modelo junto con la serie de datos original y las predicciones en cada punto de muestreo del sitio TAC. Se siguió el mismo criterio que para el sitio Centro para representar los ajustes de suavizado y predicción y los datos originales de entrenamiento y prueba. Los intervalos de confianza del 95% se encuentran representados con líneas punteadas, y los mismos se muestran más bien estrechos en TAC4 y TAC5 y mucho más suaves en TAC1 y TAC2.

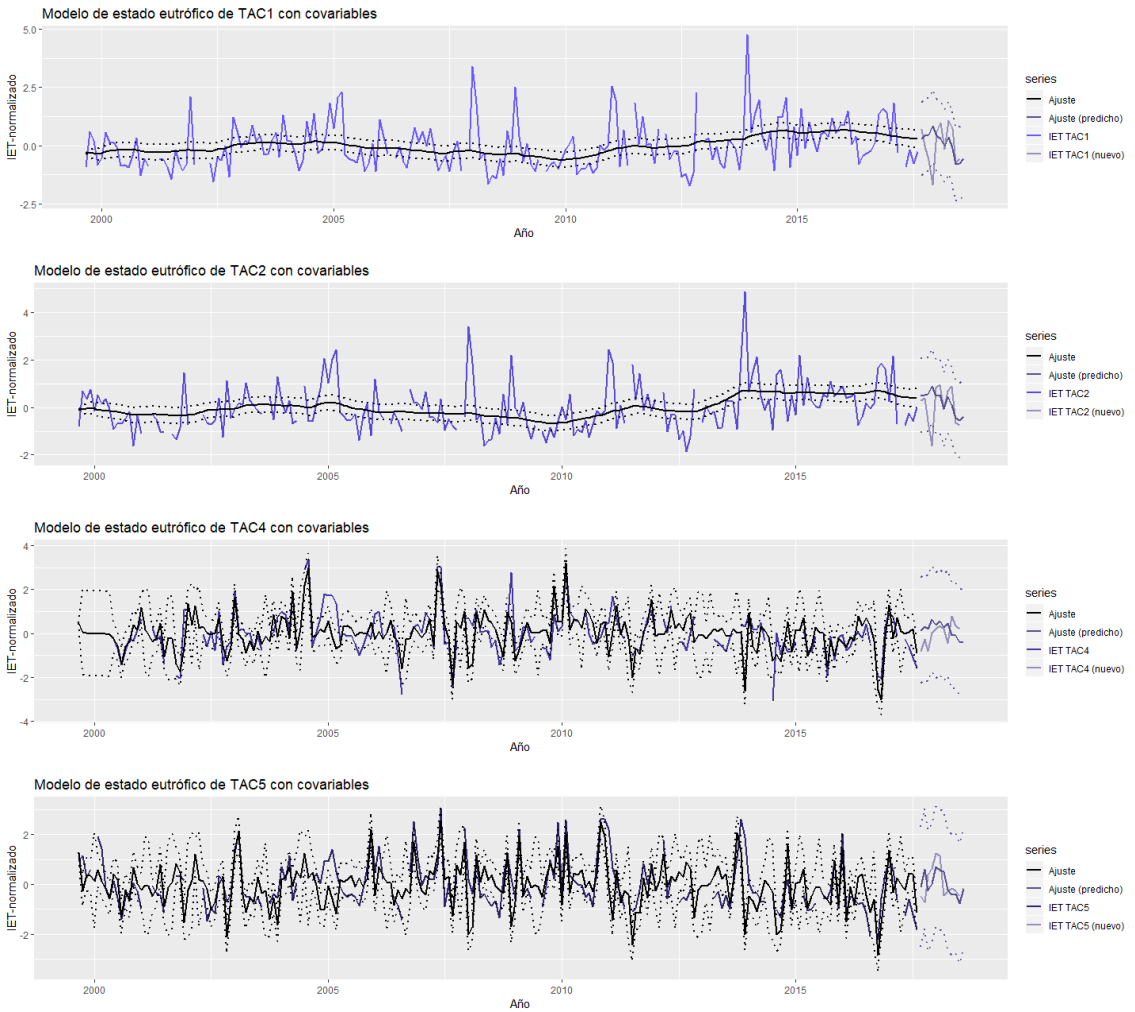


Fig. 79 Curvas de ajuste resultantes del modelo de espacio de estados multivariado junto con las series de datos originales del sitio TAC

La Tabla 24 muestra el error (RMSE) para las distintas respuestas por perfil para el modelo único de espacio de estados para el sitio TAC.

Tabla 24 Error por respuesta del modelo de espacio de estados para el sitio TAC

Perfil	RMSE
TAC1	0,94
TAC2	0,89
TAC4	0,66
TAC5	0,63

Por último, en lo que respecta al modelo “horizontal”, que contempla todos los puntos de muestreo a nivel subsuperficial, los resultados de los modelos en términos de AICc se muestran en la Tabla 25, todos los cuales cumplen el requisito de convergencia y presentan el más aceptable diagnóstico de residuos.

Tabla 25 Parametrización de los primeros modelos de espacios de estados multivariados para todos los sitios a nivel subsuperficial, ordenados por AICc

Param. Z	Param. B	Param. Q	Param. R	AICc
1	1	4	2	1990.87
1	1	4	4	1992.27
1	1	4	2	1992.88
1	1	4	1	1995.02
1	2	4	2	1995.67
1	2	4	3	1996.73
1	2	4	4	1996.90
1	3	4	4	1997.79
1	3	4	3	1998.14
1	2	4	1	1999.73
1	3	4	1	2001.28
3	3	4	1	2005.88
1	1	1	4	2053.36

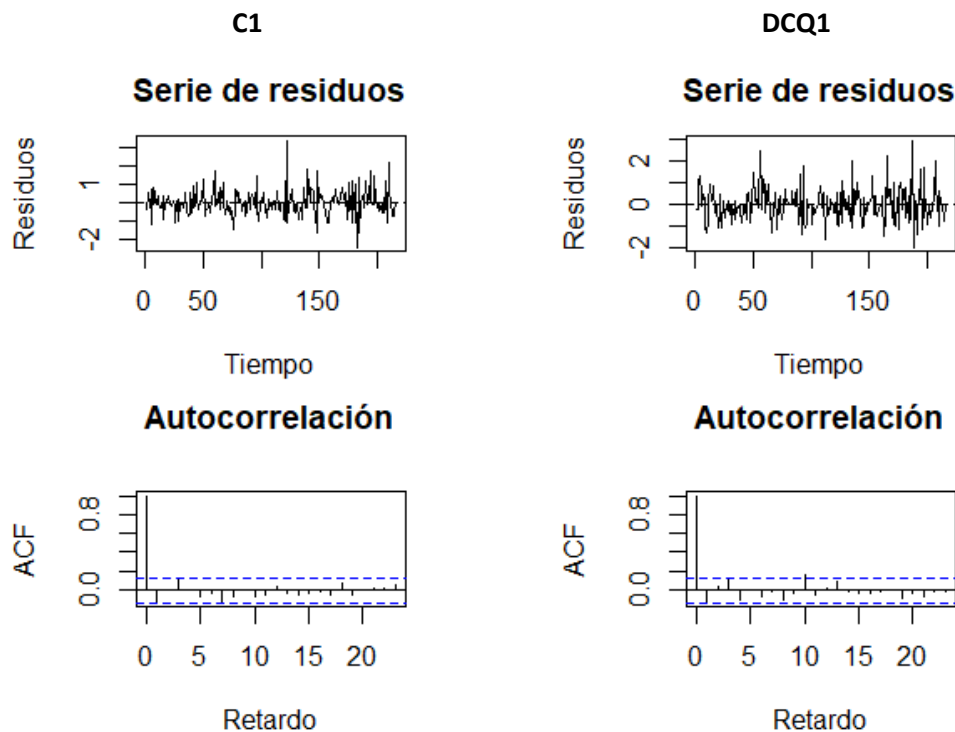
Nota: los números para cada parámetro hacen referencia a los criterios mencionados anteriormente.

En el caso de este modelo que combina varios sitios, los mejores resultados en términos de AICc se obtienen para Q con varianza igual y covarianza igual, a excepción del último modelo. En lo que respecta a los demás parámetros tanto el primer como el segundo modelo asumen Z y B con parámetros todos distintos. Incluso en todos los modelos a excepción de uno, Z siempre tiene parámetros distintos para cada punto. El último modelo, que asume valores distintos de Q, converge en menos cantidad de iteraciones y presenta mejor suavizado que el primero, a pesar de que la diferencia en términos de AICc es de más de 62 puntos, el RMSE es mejor. El modelo elegido es este último, y la Ecuación (46) muestra los coeficientes resultantes de este modelo.

$$\begin{aligned}
 \text{diag}(B) &= \begin{matrix} 0,95 & (C1) \\ 0,98 & (DCQ1) \\ 0,84 & (DSA1) \\ 0,90 & (TAC1) \end{matrix} & \text{diag}(Z) &= \begin{matrix} 0,49 & (C1) \\ 0,22 & (DCQ1) \\ 0,76 & (DSA1) \\ 0,44 & (TAC1) \end{matrix} \\
 R &= \begin{matrix} (TAC1) & (TAC2) & (TAC4) & (TAC5) \\ 0,61 & 0,05 & 0,05 & 0,05 & (C1) \\ 0,05 & 0,61 & 0,05 & 0,05 & (DCQ1) \\ 0,05 & 0,05 & 0,61 & 0,05 & (DSA1) \\ 0,05 & 0,05 & 0,05 & 0,61 & (TAC1) \end{matrix} & & (46) \\
 \text{diag}(Q) &= \begin{matrix} 0,10 & (C1) \\ 0,10 & (DCQ1) \\ 0,14 & (DSA1) \\ 0,13 & (TAC1) \end{matrix} & x_1 &= \begin{matrix} -0,83 & (C1) \\ -2,74 & (DCQ1) \\ -0,29 & (DSA1) \\ -1,02 & (TAC1) \end{matrix}
 \end{aligned}$$

	(HR)	(Nv)	(PA)	(Pr)	(RS)	(Tm)	(VV)	
	-0,120	-0,084	0,120	0,032	-0,100	0,460	-0,093	(C1)(3)
$C =$	0,170	-0,300	-0,048	-0,013	-0,110	0,170	-0,130	(DCQ1)(1)
	-0,010	0,086	0,230	-0,044	-0,130	0,380	-0,068	(DSA1)(7)
	-0,059	-0,024	0,045	0,160	-0,075	0,530	-0,081	(TAC1)(1)

En la Fig. 80 se muestran los resultados del diagnóstico del modelo. Como se puede observar, los residuos muestran una varianza aproximadamente constante alrededor de la media y no se observa autocorrelación significativa con retardo en DSA1 ni en TAC1, en el caso de C1 y DCQ1 se observa cierta autocorrelación menor. Por otro lado, la Fig. 81 muestra el diagnóstico de residuos del modelo de todos los sitios a nivel subsuperficial correspondiente a normalidad. Como se puede apreciar, se observa una distribución aproximadamente normal, siendo mejor en los casos de las desembocaduras.



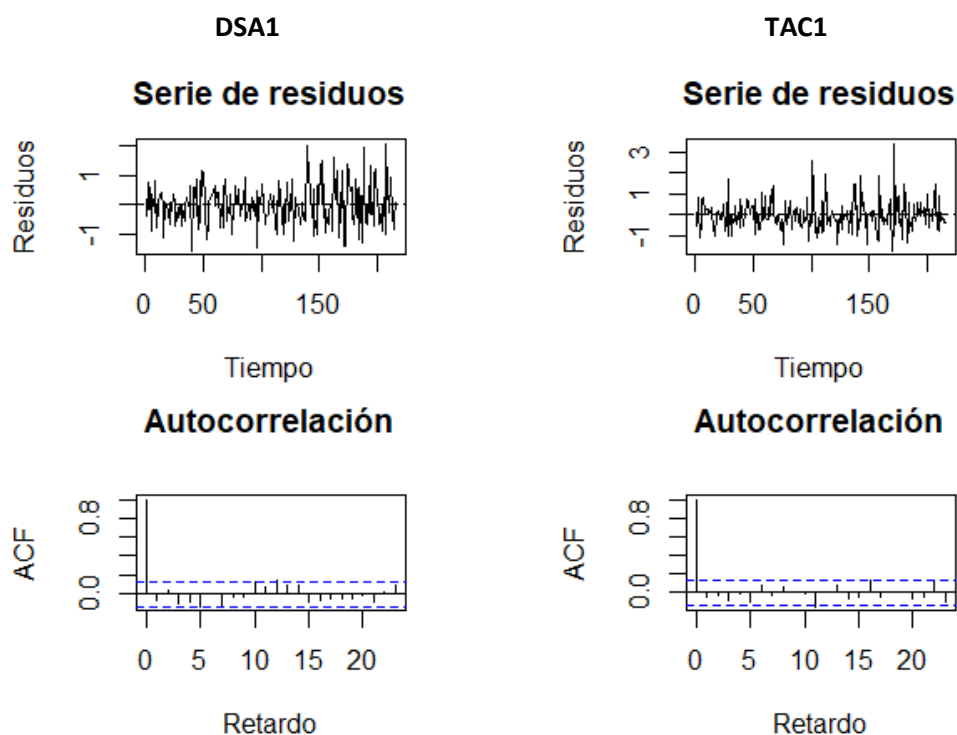


Fig. 80 Diagnóstico de residuos del modelo de espacios multivariado de todos los sitios a nivel subsuperficial: varianza y autocorrelación

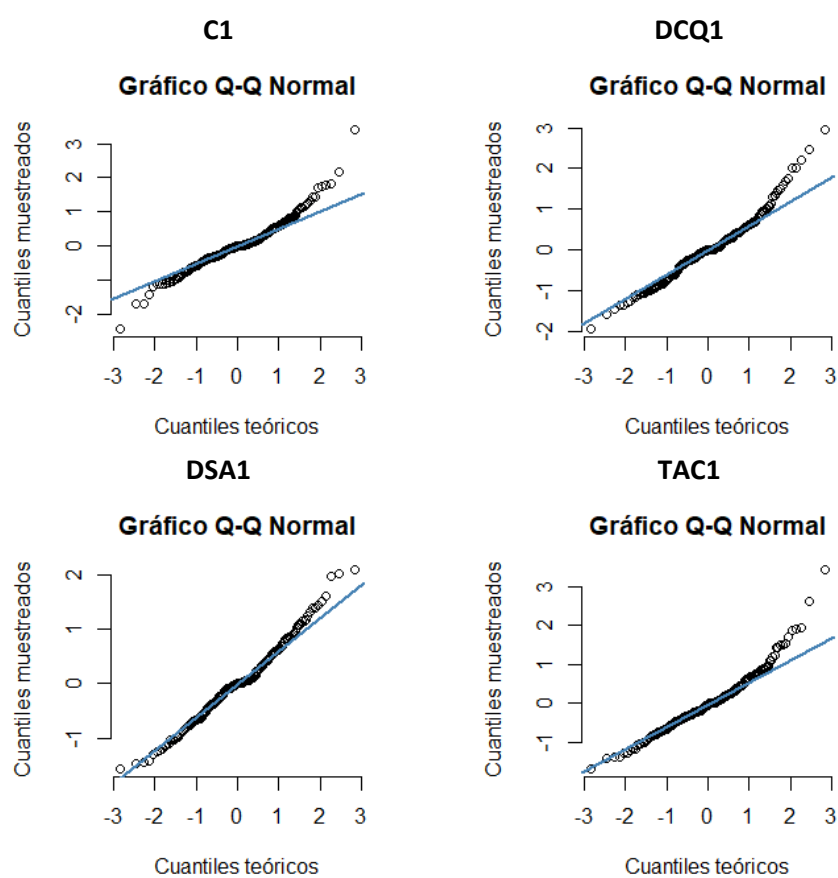


Fig. 81 Diagnóstico de residuos del modelo de espacios multivariados del todos los sitios a nivel subsuperficial: normalidad

Por último, la Fig. 82 muestra las curvas de ajuste resultantes del modelo junto con la serie de datos original en cada punto de muestreo a nivel subsuperficial. Los intervalos de confianza del 95% se encuentran representados con líneas punteadas, y los mismos se muestran más bien suaves.

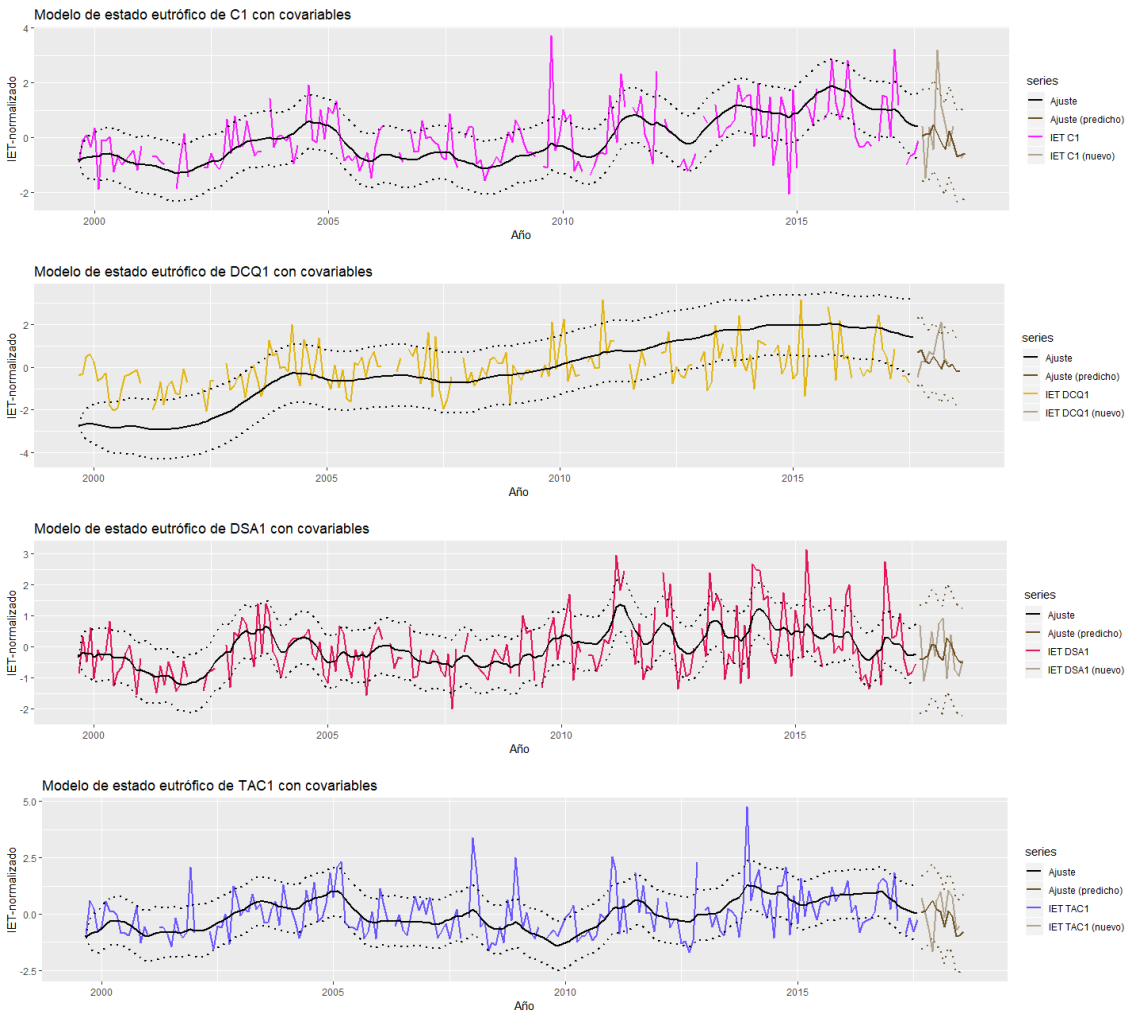


Fig. 82 Curvas de ajuste resultantes del modelo de espacio de estados multivariado junto con las series de datos originales de todos los sitios a nivel subsuperficial

La Tabla 26 muestra el error (RMSE) para las distintas respuestas por sitio para el modelo único de espacio de estados subsuperficial.

Tabla 26 Error por respuesta del modelo de espacio de estados para el modelo subsuperficial

Sitio	RMSE
C1	1,20
DCQ1	0,92
DSA1	0,78
TAC1	0,93

Cabe mencionar por último que se realizaron pruebas adicionales para comparar los resultados obtenidos de la especificación del modelo tal cual se planteó al comienzo de esta sección versus los correspondientes a una segunda configuración del modelo donde las covariables influyen el proceso de estados en lugar de influenciar el proceso observacional. De esta manera, se pudo

comprobar que, bajo la segunda configuración, en general la convergencia es escasa y el ajuste del modelo no es tan bueno en términos de las métricas evaluadas. Esto confirma la elección de la parametrización propuesta del modelo de espacio de estados multivariado.

3.2.3 Resultados del modelo de regresión bayesiano jerárquico

Por medio de la aplicación del paradigma bayesiano a la estimación del IET, se busca obtener un modelo probabilístico de distribuciones de los parámetros como se indicó en la sección 2.7.3. En este sentido, se plantearon distintas configuraciones de modelos, todas las cuales parten de distribuciones a priori no informativas para los coeficientes beta de la componente de regresión del modelo ARIMAX, con la misma media y varianza, así como también para los coeficientes del modelo autorregresivo y de media móvil. Por otra parte, se asume que la varianza de la verosimilitud es común a todas las respuestas. Así se tiene un modelo jerárquico con las características que se detallan en la Ecuación (47).

$$\begin{aligned}\theta_j &\sim N(0; 10) \\ \varphi_j &\sim N(0; 10) \\ \beta_{ij} &\sim N(0; 10) \\ \sigma &\sim U(0,10) \\ y_j &\sim N(\mu_j; \sigma^2)\end{aligned}\tag{47}$$

Donde:

j es uno de los diez puntos de muestreo

i es una de las siete variables meteorológicas

θ_j es un coeficiente del término de media móvil de la respuesta j -ésima y se distribuye normal con media 0 y varianza 10

φ_j es un coeficiente del término autorregresivo de media móvil de la respuesta j -ésima y se distribuye normal con media 0 y varianza 10

β_{ij} es un coeficiente de la variable meteorológica i -ésima correspondiente a la respuesta j -ésima y se distribuye normal con media 0 y varianza 10

σ es el desvío estándar de cada una de las diez respuestas y se distribuye uniforme entre 0 y 10.

y_j es la respuesta j -ésima y se distribuye normal con media μ_j y varianza σ^2

μ_j es una función de los parámetros θ_j , φ_j y β_{ij}

Con respecto a la forma que adopta μ_j , depende de la combinación de presunciones que se consideraron, respecto de cada una de las tres componentes (autorregresiva, media móvil y regresión de covariables):

- El proceso autoregresivo depende de la profundidad (cuatro distintos) o es independiente de la profundidad (diez distintos) y a su vez puede ser de orden 2, 1 o no estar presente. En el primer caso de procesos combinados se denomina AC y en el otro caso se denomina AI.

- El proceso de media móvil puede ser de orden 2, 1 o no estar presente.

Cabe destacar que todas las distribuciones parametrizadas (en la Ecuación (47)) se eligieron *no-informativas* en el sentido que abarcan un gran espectro de posibilidades (porque no se dispone de conocimiento a priori que justifique la elección de distribuciones o parámetros más específicos o estrechos). Asimismo, se eligió trabajar sobre las series originales sin aplicar diferenciación, ya que el modelo no impone una pauta al respecto cuando se tienen series con distinto tipo de estacionalidad y tendencia (ver 3.1).

La combinación de cada uno de estos criterios, da lugar a 14 posibilidades todas las cuales se evaluaron por medio de DICc como se describe en la sección 2.7.3.3. Es decir que, en este caso se ensayaron modelos que integran todos los datos disponibles. Se utilizaron 1000 iteraciones y 3 cadenas para cada modelo.

La Tabla 27 muestra los resultados obtenidos ordenados por DICc en base a las combinaciones mencionadas (todas las cuales se modelaron para los parámetros beta, phi, theta y sigma). Como se puede observar el mejor modelo corresponde al uso de coeficientes combinados del proceso autorregresivo de orden 2 más 1 término de media móvil. Los siguientes tres modelos, en orden de bondad de ajuste, presentan una diferencia de 6 puntos respecto del primero y distintas configuraciones de términos autorregresivos y de media móvil, con al menos uno en cada caso.

Tabla 27 Resultados de los modelos bayesianos jerárquicos

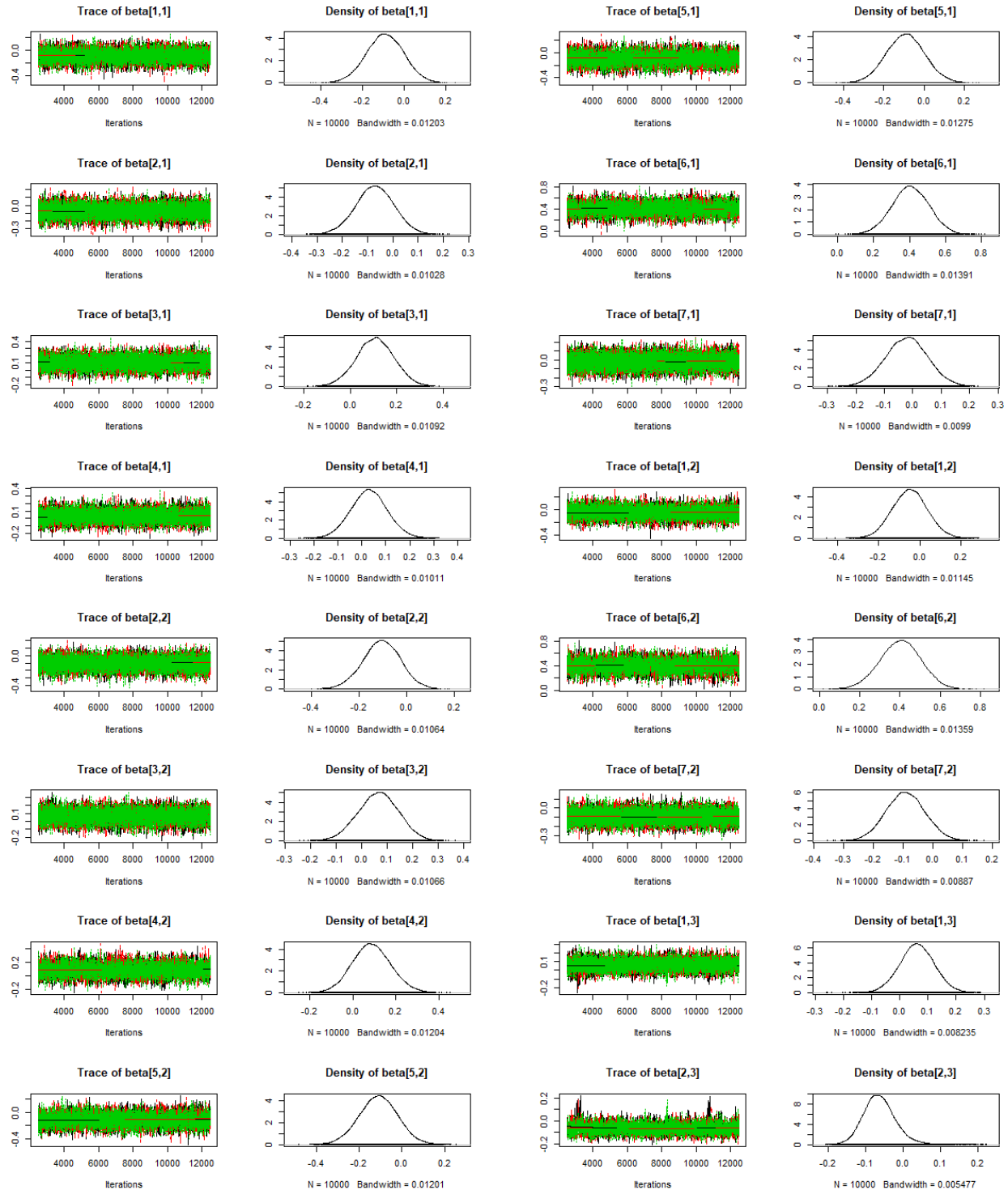
Modelo AR	Orden(AR)	Orden(MA)	DICc
AR Combinado	2	1	4216
AR Independiente	1	2	4222
AR Independiente	2	1	4222
AR Combinado	1	2	4222
AR Combinado	2	0	4223
AR Combinado	2	2	4223
AR Independiente	2	0	4232
AR Independiente	2	2	4232
AR Combinado	1	1	4249
AR Independiente	1	1	4251
Sin AR	0	2	4293
AR Independiente	1	0	4301
AR Combinado	1	0	4301
Sin AR	0	1	4351

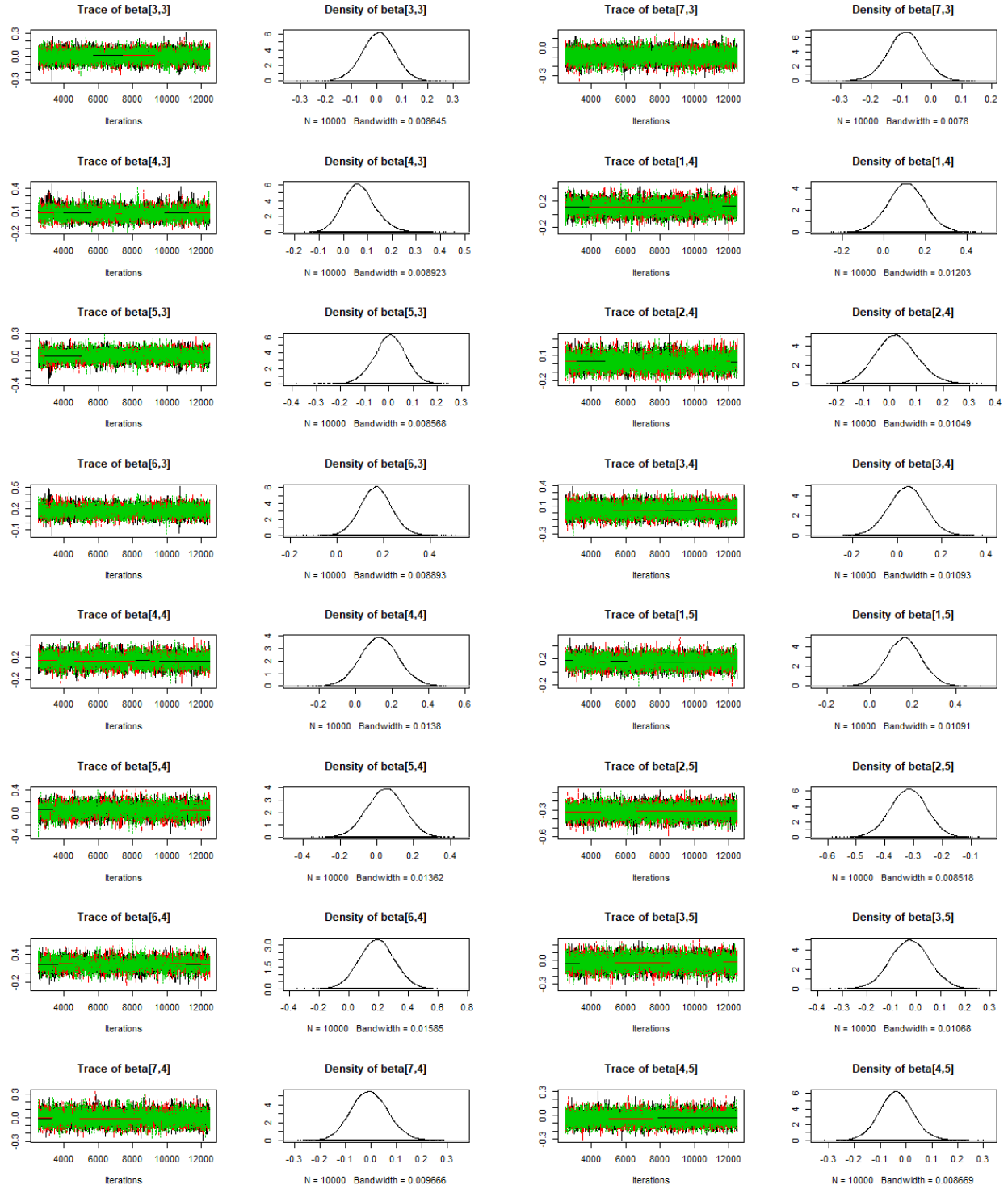
Como última prueba, se tuvo en cuenta un modelo donde la distribución de las respuestas es *t* de student en vez de normal, ya que al tener colas más pesadas es adecuada para contemplar valores extremos o *outliers*, y el estadístico DICc resultante es mejor: 4098. La Ecuación (48) muestra la distribución de la verosimilitud empleando la distribución *t* y sus parámetros.

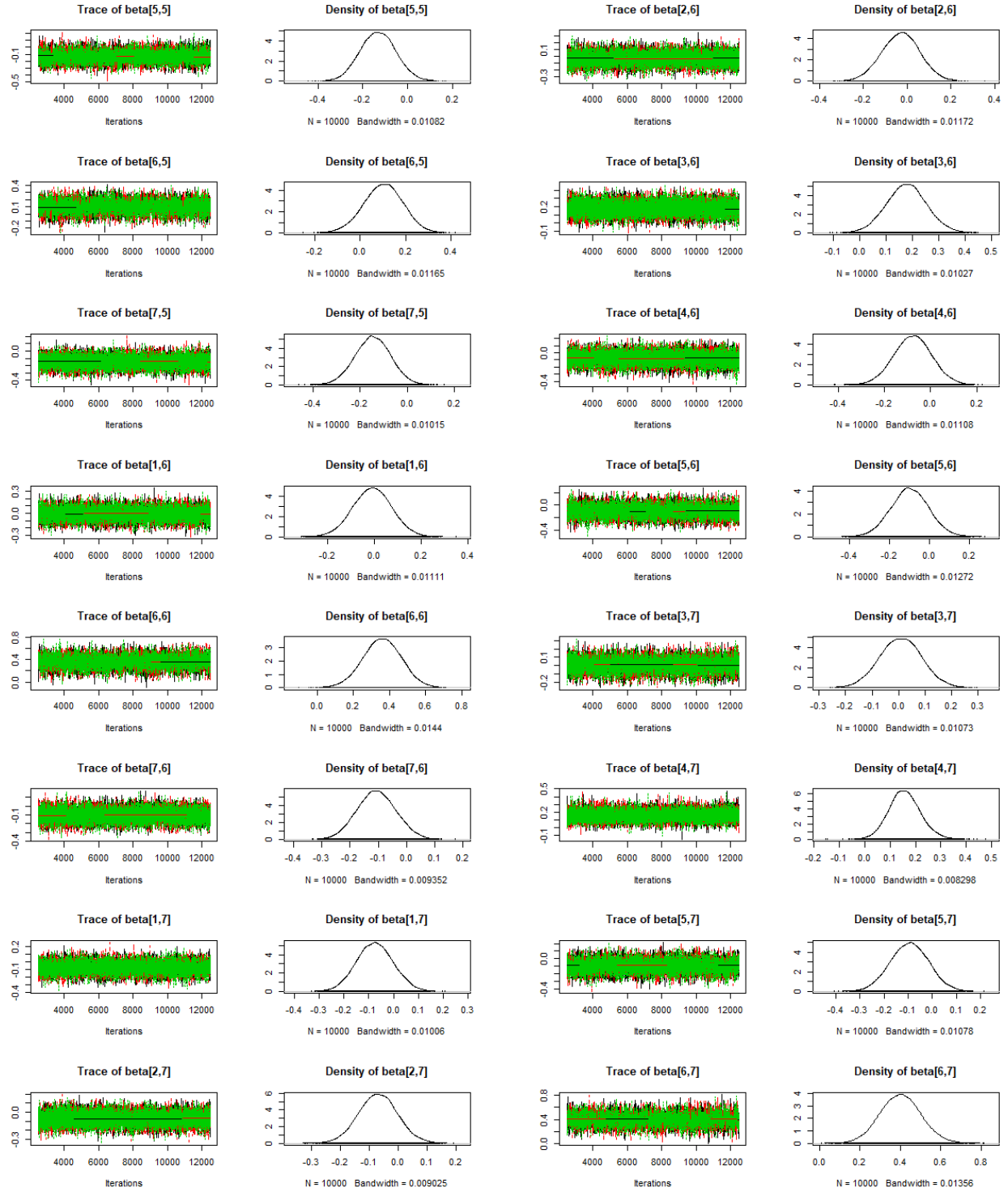
$$\begin{aligned}
 \tau &\sim G(2,5; 25) \\
 \eta &\sim Exp(1) \\
 y_j &\sim t(\mu_j; \tau; \eta + 2)
 \end{aligned}
 \tag{48}$$

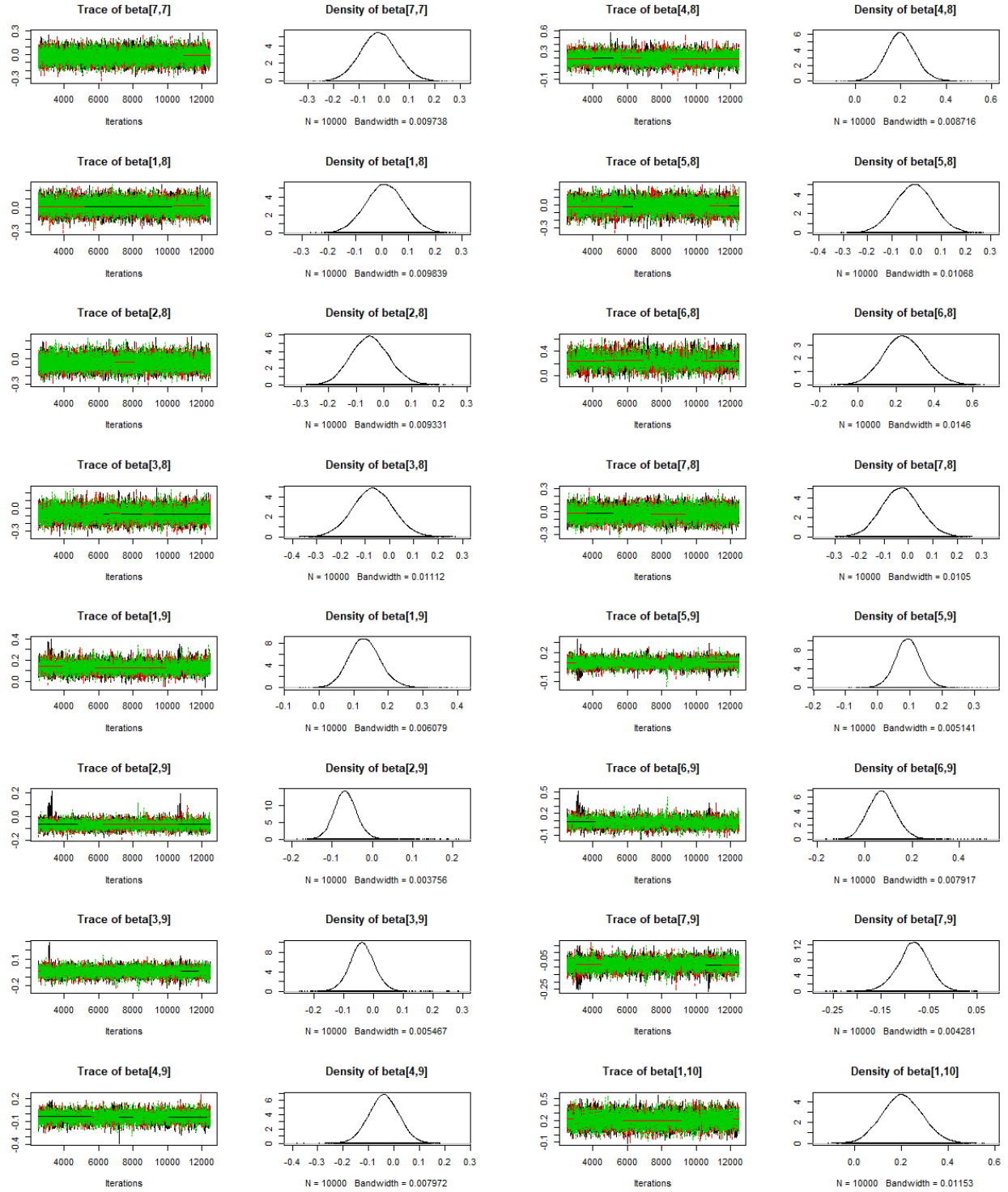
A continuación, en la Fig. 83 se pueden observar las trazas de las 3 cadenas (*traces*) y las distribuciones a posteriori (*density*) de los parámetros del modelo: sigma refiere a la varianza compartida por todas las respuestas, beta[i,j] al coeficiente de la covariable i (1 a 7: HR, Nv, PA, Pr, RS, Tm y VV) para la respuesta j (1 a 10: C1, C2, C4, C5, DCQ1, DSA1, TAC1, TAC2, TAC4 y TAC5), phi[k,m] al coeficiente k (1, 2) del proceso autorregresivo de orden 2 para el perfil de la

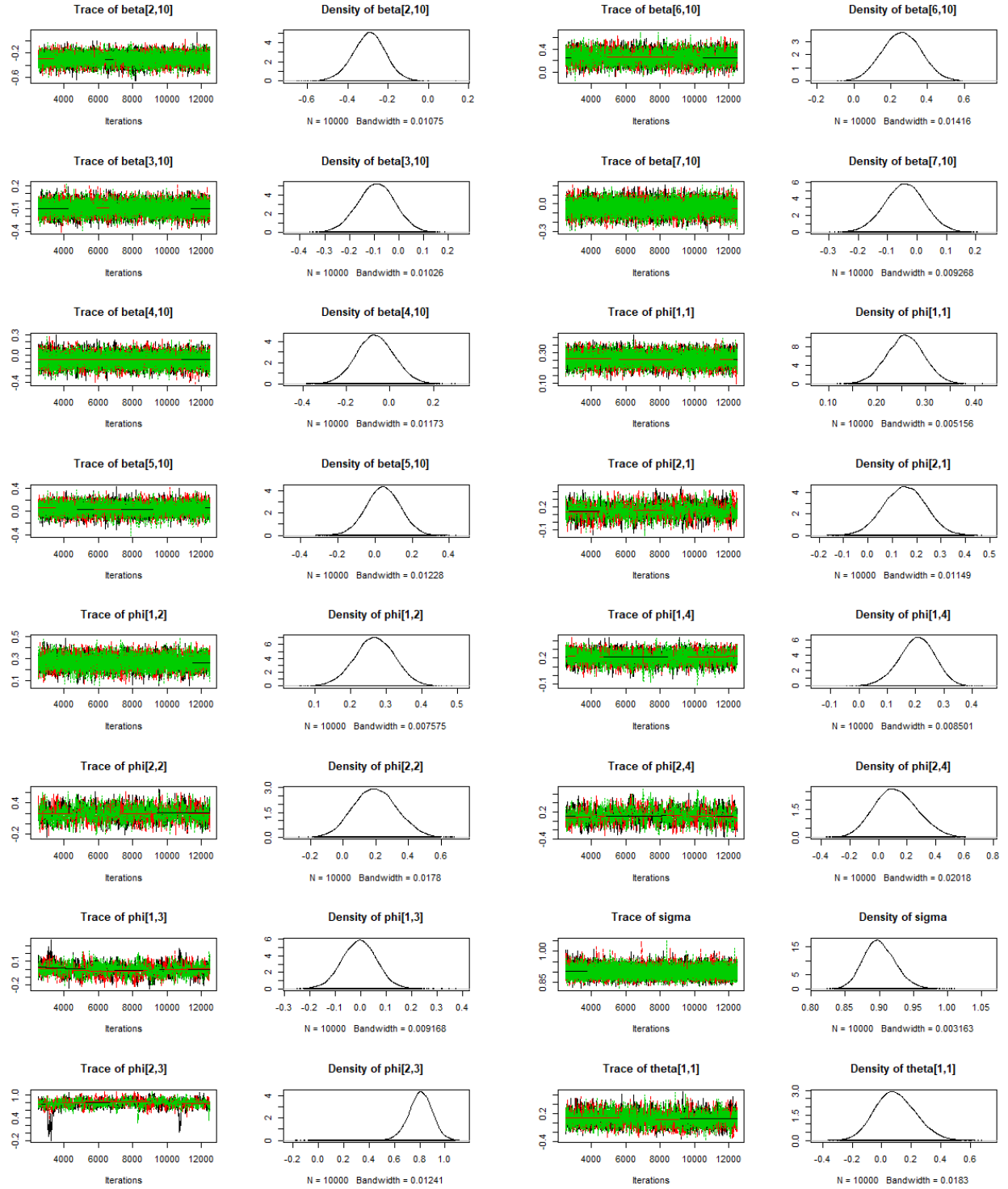
respuesta m (1 a 4: subsuperficial, zona fónica, zona intermedia y fondo) y $\theta_{n,j}$ al coeficiente n (1) del proceso de media móvil de orden 1 para la respuesta j. Las trazas no presentan patrones lo cual es un indicio de convergencia de las cadenas en 10.000 iteraciones, con 500 pasos de calentamiento y *thining* de 2 (es el intervalo entre valores sucesivos de una serie de tiempo, es un parámetro que evita la autocorrelación de las muestras, ver a Plummer et al. (2006) para mayor detalle).











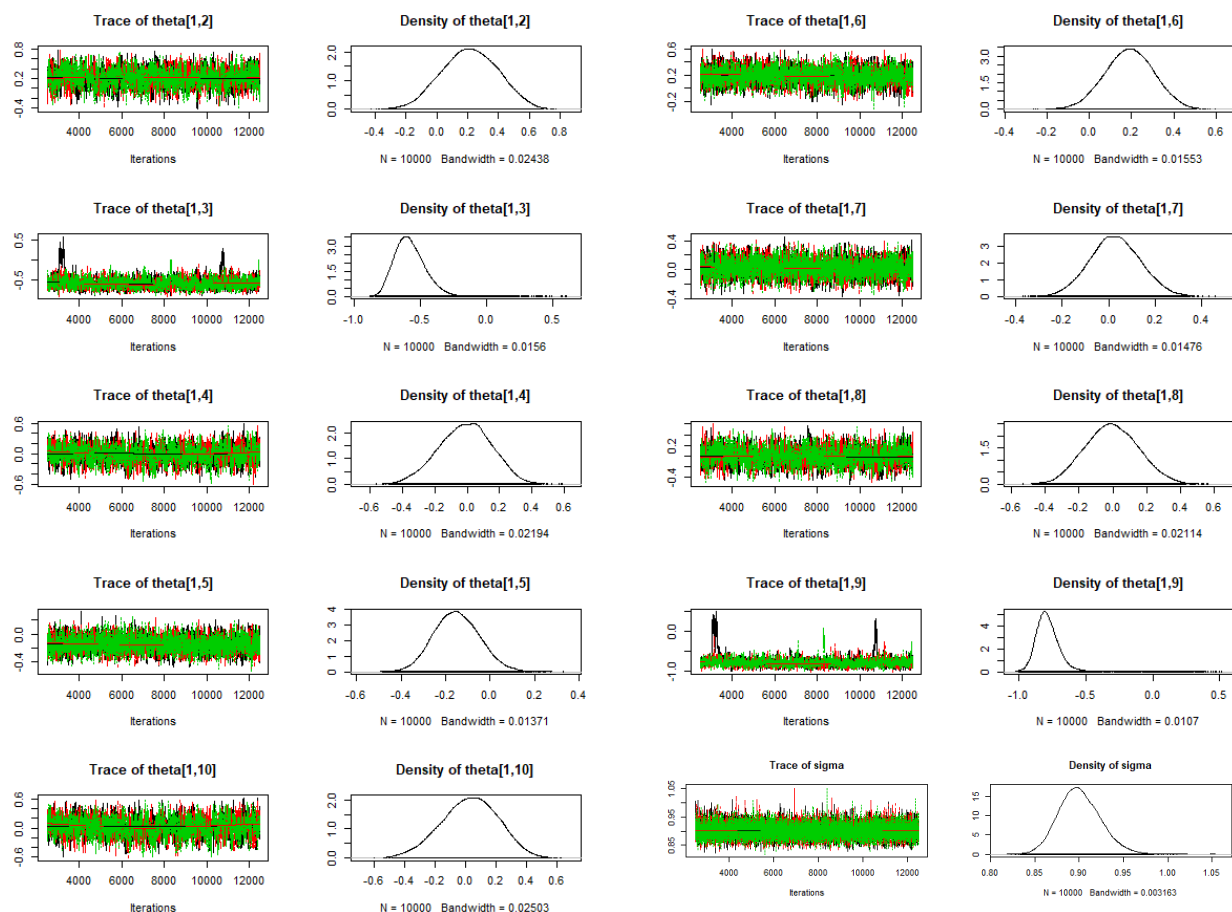


Fig. 83 Trazas de las cadenas de Markov y distribución a posteriori de los parámetros del modelo

La Tabla 28, en tanto, representa media e intervalos de credibilidad del 95% de los coeficientes. Aquellos resaltados en color azul son significativamente distintos de 0.

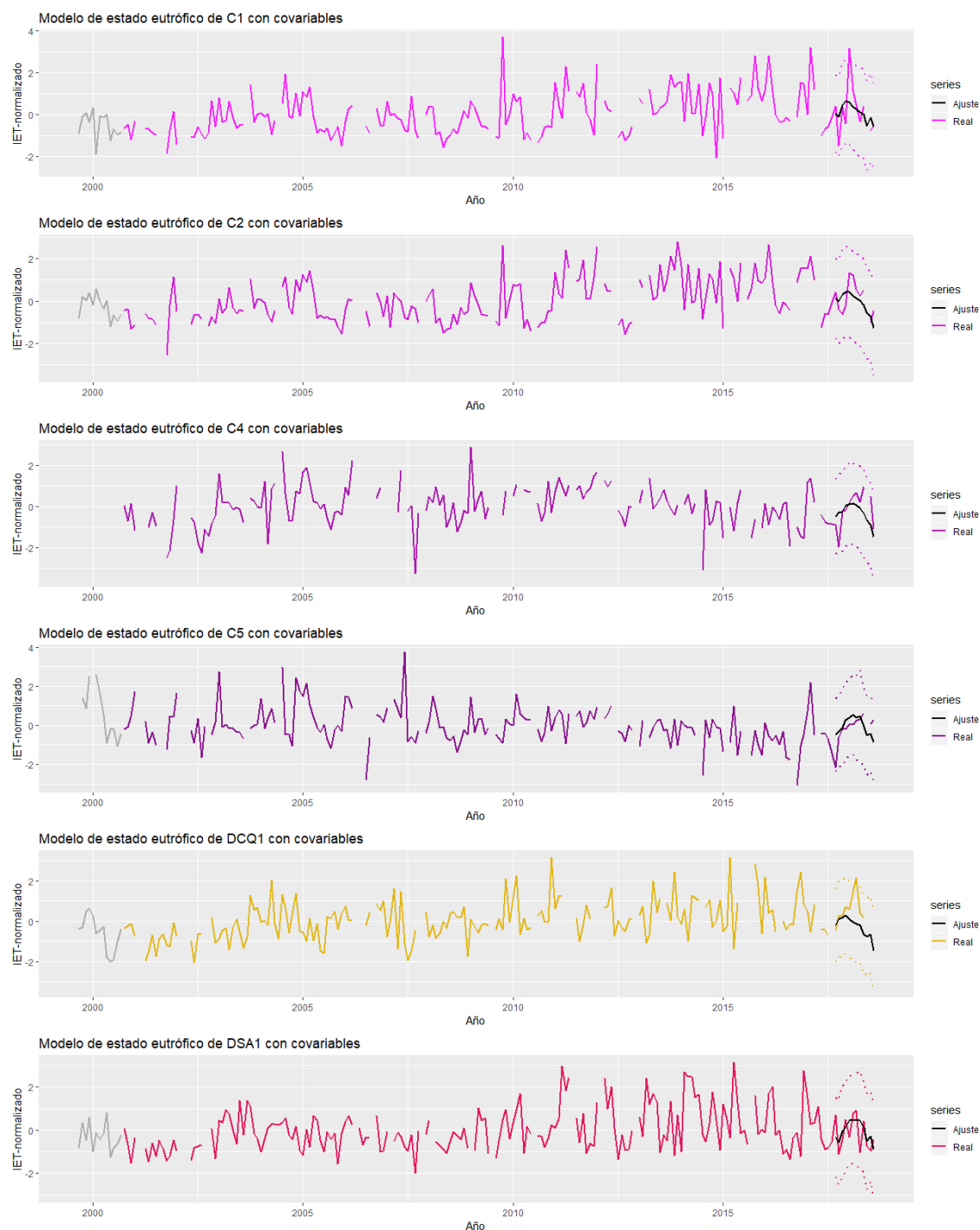
Tabla 28 Media de los coeficientes e intervalo de credibilidad (95%) del modelo bayesiano jerárquico

Coefficiente	Sitio	Variable(Alcance)	Media	2,50%	97,50%
beta[1,1]	C1	HR(3)	-0,09104	-0,26814	0,082675
beta[2,1]	C1	Nv(3)	-0,07137	-0,2252	0,078764
beta[3,1]	C1	PA(3)	0,107962	-0,05245	0,267283
beta[4,1]	C1	Pr(3)	0,031123	-0,1167	0,183659
beta[5,1]	C1	RS(3)	-0,09119	-0,27426	0,096487
beta[6,1]	C1	Tm(3)	0,404713	0,200547	0,60617
beta[7,1]	C1	VV(3)	-0,02116	-0,16587	0,123111
beta[1,2]	C2	HR(1)	-0,04843	-0,21257	0,120358
beta[2,2]	C2	Nv(1)	-0,10499	-0,26376	0,049862
beta[3,2]	C2	PA(1)	0,064256	-0,09215	0,216662
beta[4,2]	C2	Pr(1)	0,08709	-0,08509	0,26788
beta[5,2]	C2	RS(1)	-0,10936	-0,28546	0,065507
beta[6,2]	C2	Tm(1)	0,402009	0,202264	0,595743
beta[7,2]	C2	VV(1)	-0,0914	-0,21928	0,038723
beta[1,3]	C4	HR(1)	0,061046	-0,06028	0,181985
beta[2,3]	C4	Nv(1)	-0,06319	-0,14078	0,030196
beta[3,3]	C4	PA(1)	0,007325	-0,12284	0,133143
beta[4,3]	C4	Pr(1)	0,066688	-0,05963	0,215487
beta[5,3]	C4	RS(1)	0,003151	-0,12758	0,127574
beta[6,3]	C4	Tm(1)	0,167687	0,039662	0,302312
beta[7,3]	C4	VV(1)	-0,08629	-0,20379	0,030217
beta[1,4]	C5	HR(21)	0,108684	-0,0691	0,287757

beta[2,4]	C5	Nv(21)	0,024872	-0,12221	0,185946
beta[3,4]	C5	PA(21)	0,043729	-0,11561	0,203284
beta[4,4]	C5	Pr(21)	0,131252	-0,06628	0,335404
beta[5,4]	C5	RS(21)	0,047134	-0,15148	0,24243
beta[6,4]	C5	Tm(21)	0,18863	-0,04065	0,419084
beta[7,4]	C5	VV(21)	-0,00428	-0,13905	0,140294
beta[1,5]	DCQ1	HR(1)	0,156946	-0,00016	0,316269
beta[2,5]	DCQ1	Nv(1)	-0,3185	-0,44508	-0,19655
beta[3,5]	DCQ1	PA(1)	-0,02577	-0,18198	0,129177
beta[4,5]	DCQ1	Pr(1)	-0,03779	-0,16584	0,097044
beta[5,5]	DCQ1	RS(1)	-0,13165	-0,28509	0,027144
beta[6,5]	DCQ1	Tm(1)	0,101473	-0,07218	0,267861
beta[7,5]	DCQ1	VV(1)	-0,14704	-0,29543	-0,0007
beta[1,6]	DSA1	HR(7)	-0,00768	-0,16979	0,15506
beta[2,6]	DSA1	Nv(7)	-0,02827	-0,1999	0,14114
beta[3,6]	DSA1	PA(7)	0,178061	0,027166	0,33203
beta[4,6]	DSA1	Pr(7)	-0,07762	-0,24127	0,085132
beta[5,6]	DSA1	RS(7)	-0,09603	-0,28137	0,090406
beta[6,6]	DSA1	Tm(7)	0,358381	0,145455	0,566053
beta[7,6]	DSA1	VV(7)	-0,10571	-0,2393	0,032348
beta[1,7]	TAC1	HR(1)	-0,07521	-0,21925	0,071345
beta[2,7]	TAC1	Nv(1)	-0,06934	-0,20066	0,064117
beta[3,7]	TAC1	PA(1)	0,008313	-0,14701	0,166085
beta[4,7]	TAC1	Pr(1)	0,15776	0,037111	0,290308
beta[5,7]	TAC1	RS(1)	-0,09485	-0,25141	0,062229
beta[6,7]	TAC1	Tm(1)	0,401723	0,206394	0,602193
beta[7,7]	TAC1	VV(1)	-0,0207	-0,1618	0,123777
beta[1,8]	TAC2	HR(1)	0,011796	-0,13048	0,156549
beta[2,8]	TAC2	Nv(1)	-0,04938	-0,18025	0,091062
beta[3,8]	TAC2	PA(1)	-0,06797	-0,22981	0,095753
beta[4,8]	TAC2	Pr(1)	0,199174	0,06653	0,3366
beta[5,8]	TAC2	RS(1)	-0,01215	-0,16813	0,14374
beta[6,8]	TAC2	Tm(1)	0,242078	0,038052	0,461369
beta[7,8]	TAC4	VV(1)	-0,03079	-0,18138	0,124356
beta[1,9]	TAC4	HR(7)	0,132001	0,045506	0,228648
beta[2,9]	TAC4	Nv(7)	-0,06625	-0,12286	-0,00467
beta[3,9]	TAC4	PA(7)	-0,03832	-0,12266	0,045893
beta[4,9]	TAC4	Pr(7)	-0,03981	-0,15594	0,076679
beta[5,9]	TAC4	RS(7)	0,093864	0,015618	0,1717
beta[6,9]	TAC4	Tm(7)	0,072983	-0,03979	0,197653
beta[7,9]	TAC4	VV(7)	-0,08275	-0,15184	-0,01874
beta[1,10]	TAC5	HR(7)	0,205278	0,042759	0,376778
beta[2,10]	TAC5	Nv(7)	-0,29486	-0,46013	-0,13686
beta[3,10]	TAC5	PA(7)	-0,09132	-0,24308	0,056339
beta[4,10]	TAC5	Pr(7)	-0,06722	-0,23641	0,105706
beta[5,10]	TAC5	RS(7)	0,044944	-0,13731	0,222159
beta[6,10]	TAC5	Tm(7)	0,260251	0,057854	0,465766
beta[7,10]	TAC5	VV(7)	-0,04176	-0,178	0,094478
phi[1,1]	X1	NA	0,257383	0,181311	0,331988
phi[2,1]	X1	NA	0,150006	-0,01627	0,315704
phi[1,2]	X2	NA	0,267456	0,154972	0,376234
phi[2,2]	X2	NA	0,197276	-0,05393	0,460686
phi[1,3]	X4	NA	-0,00608	-0,14109	0,129066
phi[2,3]	X4	NA	0,797026	0,577979	0,97992
phi[1,4]	X5	NA	0,203697	0,072376	0,321122
phi[2,4]	X5	NA	0,123256	-0,15205	0,429871
sigma	X	NA	0,900538	0,857479	0,950398
theta[1,1]	C1	NA	0,093352	-0,15608	0,373074
theta[1,2]	C2	NA	0,214262	-0,13662	0,559227
theta[1,3]	C4	NA	-0,56944	-0,77226	-0,26457

theta[1,4]	C5	NA	-0,01171	-0,33742	0,291913
theta[1,5]	DCQ1	NA	-0,14933	-0,33929	0,055518
theta[1,6]	DSA1	NA	0,189625	-0,03686	0,411278
theta[1,7]	TAC1	NA	0,028531	-0,18307	0,246816
theta[1,8]	TAC2	NA	-0,00518	-0,30243	0,308208
theta[1,9]	TAC4	NA	-0,77665	-0,93132	-0,53686
theta[1,10]	TAC5	NA	0,027283	-0,35028	0,372074

La Fig. 84 muestra las series de tiempo originales de IET normalizado junto con el ajuste para cada sitio y perfil.



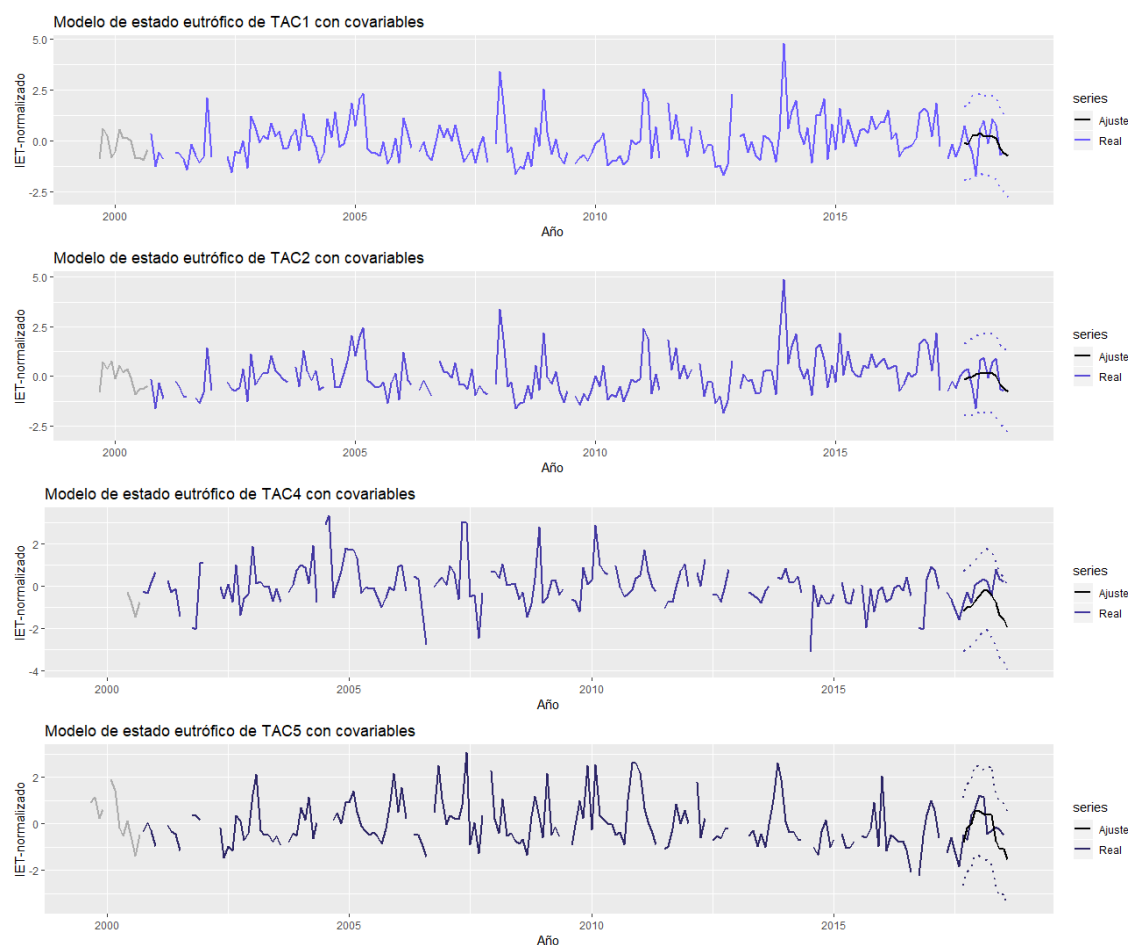


Fig. 84 Ajuste resultante del modelo bayesiano jerárquico para cada serie de tiempo

Asimismo, la Tabla 29 muestra el RMSE por sitio y perfil.

Tabla 29 RMSE correspondiente al modelo bayesiano jerárquico por sitio y perfil

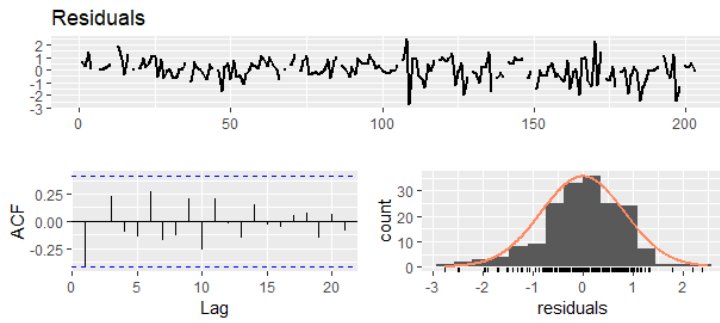
Sitio y perfil	RMSE
C1	1.01
C2	0.67
C4	0.81
C5	0.71
DCQ1	1.07
DSA1	0.67
TAC1	0.80
TAC2	0.73
TAC4	0.98
TAC5	0.61

En lo que se refiere al diagnóstico de correlación de las cadenas, se observa el decaimiento esperado conforme se incrementan los retardos. Por otro lado, el diagnóstico de Gelman-Rubin confirma que no se tienen problemas de convergencia de los parámetros (1,03 es el valor más alto). Por último, con respecto al tamaño de muestra efectivo, no se observan anomalías.

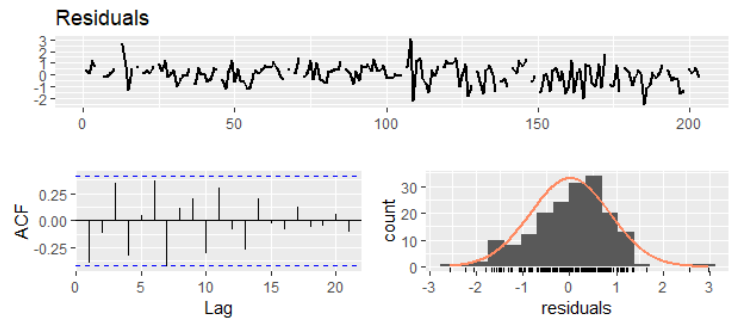
En lo que respecta al diagnóstico de residuos, como se puede ver en la Fig. 85, se observan residuos de varianza más o menos constante, de media cero y sin autocorrelación significativa, excepto en DCQ1 y TAC5 donde es escasa. En lo que respecta a normalidad, los mismos siguen

una distribución que se aproxima al ruido blanco, con ciertas desviaciones más notorias en C2, C5 y DSA1.

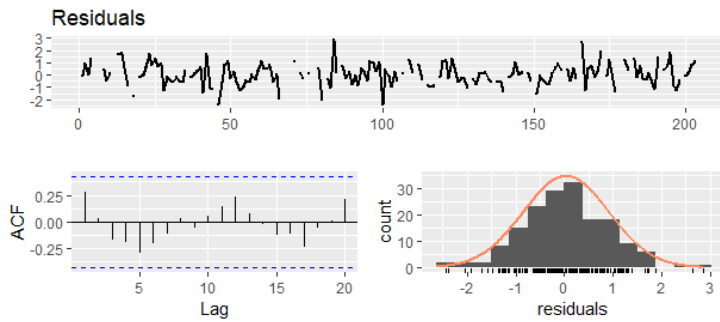
C1



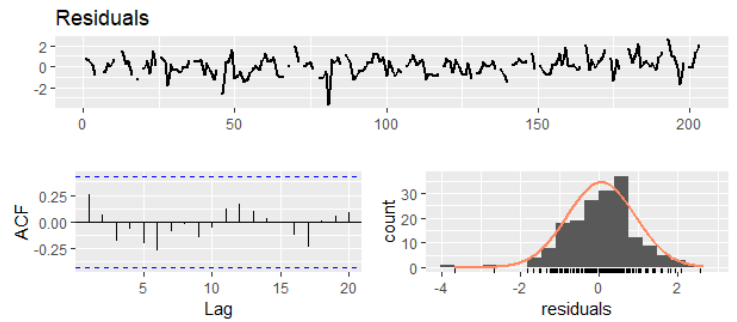
C2



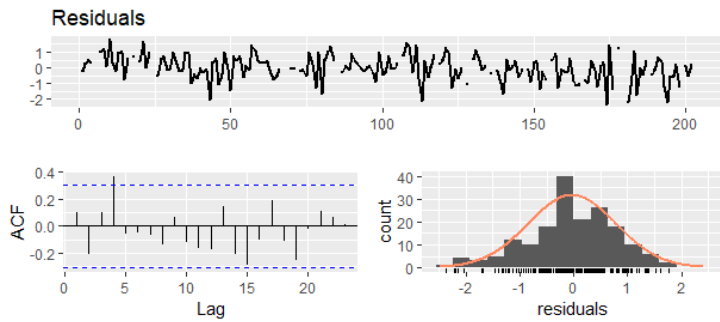
C4



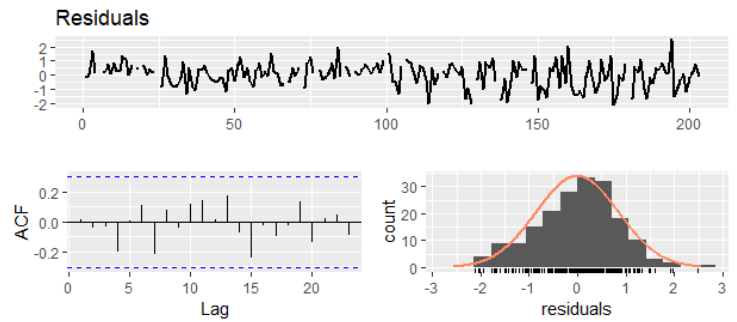
C5



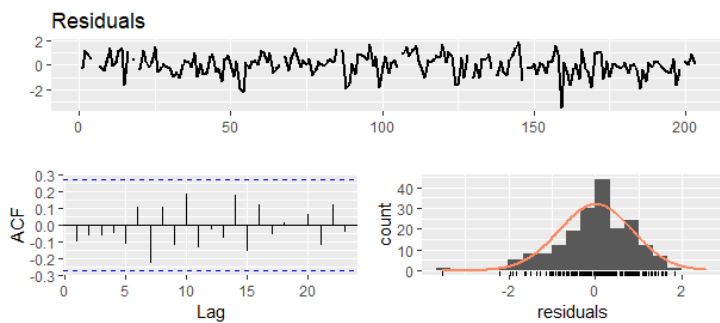
DCQ1



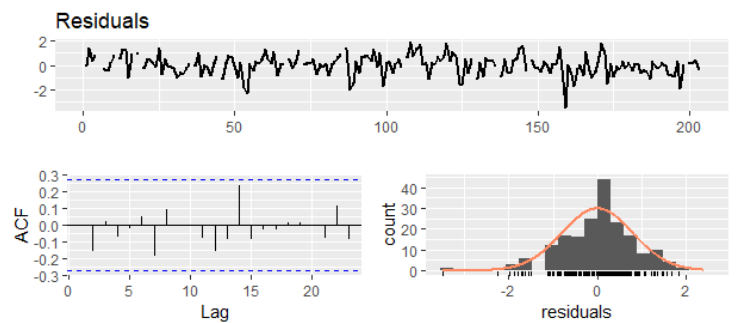
DSA1

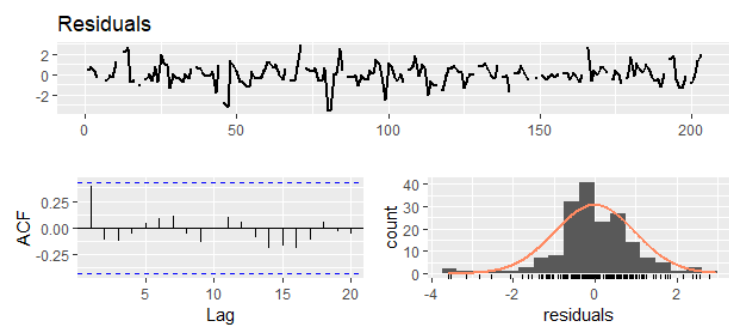
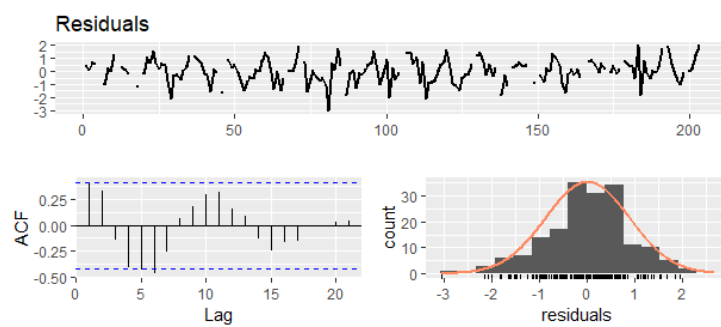


TAC1



TAC2



TAC4**TAC5***Fig. 85* Diagnóstico de residuos del modelo bayesiano jerárquico

4 Discusión y conclusiones

En el presente trabajo se estudió la aplicación de técnicas de Minería de Datos, para la resolución de interrogantes en el campo de la limnología y los recursos hídricos, tomando como caso de estudio al ESR (Embalse San Roque). Por un lado, se llevó a cabo el análisis exploratorio de los datos, para comprender más a fondo la naturaleza de las variables y las relaciones entre ellas, y por otro lado se propusieron modelos predictivos, respondiendo a la pregunta inicial: *¿cuál es la incidencia de los factores hidro-meteorológicos sobre la calidad del agua?*

En relación a la pregunta, vale la pena mencionar antes de proseguir que, cuando en este estudio se menciona la incidencia o el efecto de unas variables regresoras sobre la respuesta, el mismo debe ser entendido en términos de la relación predictiva que unas guardan con respecto a la otra (en el tiempo), mas no en términos de estricta causalidad. Según Kendall y Stuart (1961), “una relación estadística, no importa que tan fuerte o sugerente fuera, nunca puede establecer una conexión causal; las ideas de causalidad deben provenir de estadísticas externas, en última instancia de una teoría u otra”. Es decir, los modelos propuestos proporcionan una *estimación* de la medida en que los cambios en unas variables designadas independientes (en un tiempo dado) están estadísticamente relacionados (asociados, correspondidos) con los valores de una variable asumida dependiente. Podría suceder que, efectivamente, las variables hidrológicas y meteorológicas (y el estado trófico en sí mismo) tengan un efecto *causal* sobre el estado trófico en el tiempo, a la inversa, o bien que exista una variable de confusión (no observada) que tenga efecto sobre la variable respuesta y al menos una de las variables predictoras (Hyndman & Athanasopoulos, 2018). La segunda posibilidad, es decir que el estado trófico tenga efecto causal sobre las variables hidrológicas y meteorológicas, parece a todas luces contradecir el sentido común (es difícil pensar que el estado trófico del embalse esté afectando la radiación solar, por caso); mientras que la última alternativa parece poco probable al menos en el caso de las variables meteorológicas por ser esencialmente primarias. De cualquier manera, los resultados obtenidos debieran ser objeto de análisis teóricos más profundos para poder hablar de causalidad.

En lo que respecta al análisis exploratorio, se hizo una caracterización general de los ciclos y tendencias tanto de las variables seleccionadas de calidad de agua como de las variables hidrológicas y meteorológicas, concluyendo lo siguiente:

- En términos generales, existe una tendencia a largo plazo hacia el aumento de la radiación solar y el consiguiente aumento de la temperatura, que coincide con la tendencia hacia la disminución progresiva de la humedad relativa del aire.
- Se observa una tendencia al aumento de los niveles de clorofila-a en la mayoría de los puntos de muestreo. El fósforo total también presenta tendencia al aumento en los perfiles superficiales del sitio Centro y en las desembocaduras. Por otro lado, la transparencia (Secchi) tiende a la disminución en las desembocaduras de los ríos. Todo ello configura un cuadro que confirma la clara tendencia del embalse hacia la acentuación de la eutrofización.
- El ciclo de baja de la cota del lago en el período 2007-2012 da indicio de un período de escasez de agua, que coincide con el período de “sequía suave” mencionado por Elcano y Vicario (2018).
- Se observa un ciclo de empeoramiento de la condición trófica del ESR (aumento de fósforo total y clorofila-a, y disminución de Secchi) que se produce entre 2009 y 2011, y

posteriormente continúa. En este intervalo de tiempo ocurre simultáneamente un aumento de la temperatura y baja de la humedad relativa que se presenta concomitante con un descenso del nivel del lago (mínimo para todo el periodo de estudio), y un aumento de precipitación al inicio (que podría haber justificado la liberación de agua) seguido de descenso sobre el final del intervalo, para determinados alcances. Esta combinación de factores configuraría un escenario favorable a la eutrofización del ESR que debería evitarse. El bajo nivel del embalse acompañado de picos en precipitaciones podría explicarse por una regulación antrópica del volumen del mismo a través de la apertura de compuertas. Si bien este manejo podría evitar la inundación de la zona del peligro, por otro lado, parecería ser la peor combinación para el cuidado de la calidad del agua.

En lo que respecta a estacionalidad, existe una gran variabilidad en cuanto al momento en que se producen picos del fósforo total, fluctuando desde el inicio de la primavera al fin del verano, en el período 2008-2010, probablemente en consonancia con lo mencionado en el último punto. Este comportamiento resulta de interés para su estudio en futuros trabajos. En la misma línea, pero en este caso referido a la variable clorofila-a, el cambio en la estacionalidad de los picos podría estar asociado a la predominancia de uno u otro género de algas, que se desarrollan más intensamente en distintas épocas del año (ver Anexo III). Por ejemplo, en el caso de TAC se producen picos en épocas bien distintas (de primavera a otoño). En el caso de los picos y depresiones de clorofila-a, se debe tener en cuenta que, a diferencia del fósforo total, que es un parámetro químico, la clorofila-a responde a mayor o menor presencia de algas. Entonces un cambio en un pico puede que represente un cambio en un género que se desarrolla antes y otro después en función a las condiciones que se presentan en un momento u otro. Nuevamente es una hipótesis que se deja planteada para futuros trabajos.

En referencia a la cuestión predictiva, a través del desarrollo de tres modelos basados en series de tiempo, a saber: modelo de regresión dinámica, modelo de espacio de estados multivariado y modelo bayesiano jerárquico fue posible cuantificar el efecto de las variables hidrológicas y meteorológicas sobre el estado trófico del ESR por medio de coeficientes y establecer relaciones entre los mismos. Asimismo, se evaluó la capacidad predictiva de estos modelos a los fines de poder ser aplicados como herramientas de gestión de los recursos hídricos. Bajo distintas configuraciones y supuestos, los tres dan una visión distinta en términos de integración:

- El modelo de tipo regresión dinámica se focaliza en el efecto individual del proceso meteorológico sobre la calidad del agua en un determinado sitio y perfil, por ende, los coeficientes obtenidos son *dedicados*, es decir, permiten una interpretación local del fenómeno.
- El modelo de espacio de estados multivariados, permite una visión más integral, de mediano alcance, al compartirse los coeficientes entre los distintos perfiles de un mismo sitio o bien entre los distintos sitios de un mismo perfil subsuperficial. En este sentido, la interpretación ya no es en términos de un entorno reducido, sino que puede explicarse en un ámbito de más largo alcance espacial, como ventaja más destacable. Otra virtud del modelo, es explicar el fenómeno en términos de dos procesos: un proceso observacional, que como el nombre lo indica es directamente observable, y un proceso de estados que es desconocido a priori y afecta a los puntos modelados conjuntamente. Como desventaja este modelo no necesariamente produce un ajuste tan exacto.

- Por último, el modelo bayesiano jerárquico podría interpretarse como una síntesis de los anteriores, en términos de propuesta superadora. Tiene la ventaja de ofrecer la visión más general, porque se asume en este caso una varianza compartida entre los 10 puntos de muestreo, y también procesos autorregresivos compartidos por perfil, sin perder de vista que los coeficientes de regresión a partir de las covariables son dedicados a cada sitio y perfil, al igual que el proceso de media móvil.

Desde el punto de vista predictivo, no existe un tipo de modelo que se presente como el mejor en términos absolutos. Los tres modelos presentan RSME (raíz cuadrada del error cuadrático medio, ver sección 2.7.4) similares con un desempeño levemente mejor de uno u otro según el sitio o perfil. En lo que respecta a las curvas de ajuste, las correspondientes al modelo de regresión dinámica son más variables, mientras que en los otros dos casos se trata de curvas mucho más suaves. La Tabla 30 muestra la comparación de RMSE entre los tres tipos de modelo por sitio y perfil, donde se puede observar, que los mejores resultados se obtienen para el modelo bayesiano jerárquico y para el conjunto de los modelos de regresión dinámica, mientras que en varios casos, el modelo de espacio de estados multivariado tiene una muy leve diferencia con el segundo (caso de C2, DCQ1 y TAC5). Otro aspecto interesante es que este último tipo de modelo presenta mejor ajuste cuando se asume que los procesos de estado y observacional se desarrollan conjuntamente por perfil y no por sitio, así el tercer modelo de espacios de estado multivariado a nivel subsuperficial, presenta un mejor desempeño en términos de error frente a los correspondientes individuales por sitio para C1 y TAC1, como se encuentra resaltado en la tabla con un asterisco (*). Este último dato está en consonancia con el hecho de que el mejor modelo bayesiano jerárquico asume que los procesos autorregresivos se desarrollan por perfil. En este caso, se requieren dos términos del proceso autorregresivo mientras que el modelo de espacios multivariado es *markoviano* en el proceso de estados ya que el estado actual sólo depende del anterior, y quizás sea ésta una diferencia fundamental en términos del mejor ajuste de uno en relación al otro.

Tabla 30 Comparación de los errores entre los tres tipos de modelo por sitio y perfil

Sitio y perfil	Modelo de Regresión Dinámica (1)	Modelo de Espacio de Estados Multivariado (2)	Modelo Bayesiano Jerárquico (3)	Var % (1) respecto del mejor	Var % (2) respecto del mejor	Var % (3) respecto del mejor
C1	1.03	1.20 (*)	1.01	2%	19%	0%
C2	0.64	0.76	0.67	0%	19%	5%
C4	0.82	0.83	0.81	1%	2%	0%
C5	0.81	0.82	0.71	14%	15%	0%
DCQ1	0.90	0.92	1.06	0%	2%	18%
DSA1	0.77	0.78	0.67	15%	16%	0%
TAC1	0.88	0.93 (*)	0.80	10%	16%	0%
TAC2	0.76	0.89	0.73	4%	22%	0%
TAC4	0.58	0.66	0.98	0%	14%	69%
TAC5	0.60	0.63	0.61	0%	5%	2%

(*) Corresponde al modelo a nivel "horizontal" o subsuperficial

Por otro lado, comparando ahora el modelo bayesiano jerárquico con el modelo de regresión dinámica, cabe notar que, en los casos en que se obtuvieron mejores resultados para el segundo que para el primero, el modelo de regresión dinámica contempló más términos en el proceso autorregresivo (en el caso de C2 y DCQ1 se consideran 3 términos), se contempló autorregresión en la componente estacional, con retardo de 12 meses (caso de TAC5) o bien, un único término en el proceso autorregresivo (caso de TAC4). Por lo demás también es relevante la diferenciación, que se aplica en el modelo de regresión dinámica y no en los demás. De lo dicho

anteriormente, se puede notar que el proceso autorregresivo es clave para los modelos. Este aspecto ya podía vislumbrarse al observar autocorrelación y autocorrelación parcial de las variables meteorológicas (ver sección 3.1.2).

Comparando los tipos de modelos, puede notarse como el modelo de regresión dinámica y el modelo bayesiano funcionan mejor al asumir mayor independencia de los procesos que se desarrollan en cada sitio y perfil: en un caso porque el modelo es individual, y en el otro porque lo que se asume compartido es la varianza global y los procesos autorregresivos, pero por perfil. En el caso del modelo de espacios multivariados, los mejores modelos asumen varianza y covarianza compartida en los errores del proceso observacional o del proceso de estados, pero también coeficientes del proceso observacional y el proceso de estados compartidos entre perfiles superficiales y perfiles de fondo, con excepción del último modelo “horizontal” que asume coeficientes todos distintos por sitio y perfil, consiguiéndose resultados levemente mejores.

Posiblemente, y en lo relativo a la gestión, podría recomendarse el uso del modelo bayesiano jerárquico en cuanto representa una solución integral, en un escenario normal de dinámica del cuerpo de agua. El modelo de regresión dinámica podría ser conveniente en casos en que por particularidades excepcionales de alguno de los sitios convenga analizar la evolución del estado trófico de manera localizada (por ejemplo, los sitios de desembocadura, que de vez en cuando se ven afectados por la sequía de los ríos tributarios).

Por último, en términos de los ajustes, debe tenerse en cuenta que, existen otras variables de tipo físicas, químicas y biológicas que afectan la calidad del agua, pero no se han incluido como parte de los modelos para poder apreciar aisladamente la incidencia de las variables meteorológicas, y porque además su efecto se encuentra más ampliamente estudiado. Por otra parte, existen modelos de aprendizaje automático enmarcados en la categoría de “caja-negra”, tal es el caso de la familia de modelos de redes neuronales o los modelos no paramétricos como los procesos gaussianos en el ámbito específico de las series temporales que, aun utilizando únicamente las variables hidrológicas y meteorológicas, podrían haber producido mejores resultados en cuanto al ajuste. Sin embargo, estos modelos privan al analista del conocimiento de las condiciones, en tanto relaciones entre los parámetros, que posibilitaron la obtención de los resultados. Vale decir, la falta de transparencia habría dificultado cuando no impedido posiblemente responder a la pregunta inicial. Se sugiere para trabajos futuros poner a prueba este tipo de modelos.

Otro aspecto de interés, ya relativo al impacto de las variables meteorológicas es el alcance de las mismas en términos de la cantidad de días que se contemplan para analizar su efecto. En este sentido, pudo comprobarse en el caso del modelo de regresión dinámica que la variabilidad del estado trófico se explica mejor considerando distintos alcances según la profundidad del perfil, teniendo en cuenta que los ensayos se hicieron considerando que todas las variables hidrológicas y meteorológicas actúan con igual alcance. Así, en los niveles subsuperficial y zona fótica el alcance que mejor explica la variable respuesta es menor que en los niveles de zona intermedia y fondo para los sitios Centro y TAC: entre 1 y 3 para los perfiles superficiales y entre 7 y 21 para los perfiles de fondo, con excepción de C4 donde 1 es el mejor alcance. En lo que respecta a las desembocaduras, en el caso del río Cosquín, el alcance que mejor explica la variabilidad es 1, siguiendo la regla anterior para niveles superficiales, mientras que en el caso del río San Antonio, se tendría una excepción con alcance 7. Vale notar además que las excepciones se producen en las dos series donde fue necesario seleccionar variables específicas para que el modelo de regresión dinámica cumpla con los supuestos, por lo que no se hallan en

igualdad de condiciones a los demás en términos de esta comparación. Asimismo, los alcances elegidos para este modelo decidieron emplearse para los demás, de manera tal que los resultados fuesen comparables. En este sentido, cabe notar dos puntos: primero, podría haberse realizado el ensayo considerando todos los alcances para los modelos de espacio de estados multivariado y bayesiano jerárquico, tal cual se hizo para el modelo de regresión dinámica; y segundo, se podría haber incluso probado en todos los modelos todas las combinaciones posibles de alcances por variables. En un caso y en el otro, la combinatoria arroja una cantidad de casos que resulta prohibitiva en términos de tiempo de procesamiento para los plazos de este trabajo, por lo que se sugiere como punto interesante a retomar para trabajos futuros.

Por otro lado, y yendo un paso más allá, si bien es cierto que, por medio del uso de alcances, las regresoras en el momento t están desplazadas en el tiempo respecto de la respuesta (1, 3, 7, 14 o 21 días), esta estrategia constituye una agregación del pasado. También hubiera sido enriquecedor para el análisis individualizar el efecto diferido, es decir en el momento $t-1$, $t-2$, ... $t-k$, de las distintas variables. Esta variante se conoce en el contexto del modelado de regresión dinámica como “modelos de retardos distribuidos” (Pankratz, 1994). La misma idea hubiera podido ser extendida posiblemente al marco bayesiano, más no así al planteo de espacio de estados multivariados que se siguió en este trabajo, por su misma definición (ver Ecuaciones (10) y (11), donde se observa que las regresoras actúan en el momento t , ya sea en el proceso de estados o en el proceso observacional, respectivamente), es decir que su aplicación fue posible a partir del uso de alcances. Asimismo, este tipo de modelado habría requerido una previa determinación de cuál es, para cada variable y sitio-perfil, la cantidad de períodos a considerar previo planteo del modelo.

Siguiendo en la línea del análisis de las variables regresoras, hidrológicas y meteorológicas, y ya en vistas a responder a la pregunta inicial: *¿cuál es la incidencia de las condiciones hidro-meteorológicas sobre la calidad del agua?*, corresponde analizar los coeficientes de los modelos obtenidos por sitio y perfil.

La Tabla 31 muestra las variables con coeficientes significativos para cada tipo de modelo.

Tabla 31 Variables con coeficientes significativos por sitio y perfil para cada tipo de modelo

Sitio y perfil	Alcance	Modelo de Regresión Dinámica (1)	Modelo de Espacio de Estados Multivariado (2)	Modelo Bayesiano Jerárquico (3)
C1	3	Tm (+), PA (+)	Nv (-), RS (+), Tm (+), PA (+)	Tm (+)
C2	3	Tm (+)	HR (+), Nv (-), Pr (+), RS (+), Tm (+), VV (-)	Tm (+)
C4	1	Pr (+)	Nv (+), Pr (+)	Tm (+)
C5	21	Nv (+)	HR (+), Nv (+), PA (-), RS (+)	-
DCQ1	1	Nv (-), Tm (+)	HR (+), Nv (-), RS (-), Tm (+), VV (-)	Nv (-), VV (-)
DSA1	7	Tm (+), PA (+)	PA (+), Tm (+)	PA (+), Tm (+)
TAC1	1	Tm (+), Pr (+)	Pr (+), RS (-), Tm (+), VV (-)	Tm (+), Pr (+)
TAC2	1	Tm (+), Pr (+)	Pr (+), Tm (+), VV (-)	Tm (+), Pr (+)
TAC4	7	-	HR (+)	HR (+), Nv (-), RS (+), VV (-)
TAC5	7	Nv (-), Tm (+), PA (-), HR (+)	HR (+), Nv (-), Pr (-), Tm (+)	HR (+), Nv (-), Tm (+)

Nota: en color gris las variables del modelo (2) vertical (C o TAC) y en letra normal las variables del modelo (2) horizontal, en cursiva variables con coeficiente significativo en ambos. En negrita los modelos con mejor ajuste.

Como se puede observar, la **temperatura** tiene coeficiente significativo en nivel subsuperficial y zona fótica (a excepción de DCQ1 en modelo (3)), y siempre con signo positivo, vale decir que, a mayor temperatura, mayor Índice de Estado Trófico (IET). Por otro lado, en el caso de TAC, para

los mismos perfiles mencionados, se suma la variable **precipitación** en todos los modelos, asimismo ejerciendo un efecto positivo sobre la respuesta. En los perfiles más profundos, en zona fótica y fondo, mayor cantidad de variables presentan coeficientes significativos. Así la **humedad relativa** ejerce un efecto positivo, especialmente en TAC, mientras que el **nivel del embalse** ejerce efecto negativo en TAC y positivo en C. Esta última variable también ejerce efecto negativo en DCQ1. Continuando con las comparaciones por perfil, la **presión atmosférica** ejerce influencia positiva en los perfiles superficiales y negativa en los perfiles más profundos. Por último, cabe destacar el efecto significativo negativo de la **velocidad del viento** principalmente en los perfiles superficiales. En lo que respecta a la variable **radiación solar**, no pareciera haber un patrón general, dado que es significativa mayormente con signo positivo en todos los perfiles, pero en DCQ1 y TAC1 presenta efecto negativo. Teniendo en cuenta lo indicado en la introducción sobre el efecto positivo de la radiación sobre el aumento de clorofila-a, podría decirse que en este caso no se tiene evidencia suficiente para respaldar ese punto, ya sea porque la clorofila-a está promediada en un índice, por el mayor efecto de otras variables en el contexto del problema, o bien por la necesidad de mayor cantidad de mediciones (más estaciones). Cabe mencionar además, como puede observarse en las Ecuaciones (44) y (45), que conforme aumenta la profundidad, crece el efecto positivo de la humedad relativa, mientras que decrece el efecto negativo del viento. En lo que se refiere a la magnitud de los coeficientes significativos descriptos con anterioridad, y tomando como base los mejores modelos hallados puede observarse lo siguiente por sitio y perfil, asumiendo que todas las demás variables del modelo se mantienen constantes:

- C1: por cada desvío estándar adicional de temperatura se espera un incremento de 0,40 del desvío estándar del IET.
- C2: por cada desvío estándar diferenciado (1 vez) adicional de temperatura se espera un incremento de 0,42 del desvío estándar diferenciado (1 vez) del IET.
- C4: por cada desvío estándar adicional de temperatura se espera un incremento de 0,17 del desvío estándar del IET.
- DCQ1: por cada desvío estándar diferenciado (1 vez) adicional de temperatura se espera un incremento de 0,21 en el desvío estándar diferenciado (1 vez) del estado trófico; y por cada desvío estándar adicional del nivel del embalse se espera una disminución de 0,30 en el IET. En este caso, tratándose de la desembocadura de un río se observa como la influencia que ejerce el nivel supera a la correspondiente a la temperatura.
- DSA1: por cada desvío estándar adicional de presión atmosférica se espera un incremento de 0,18 del desvío estándar del IET; y por cada desvío estándar adicional de temperatura, se espera un incremento de 0,36 en el desvío estándar de la respuesta.
- TAC1: por cada desvío estándar adicional de temperatura, se espera un incremento de 0,40 del desvío estándar del estado trófico; y por cada desvío estándar adicional de precipitación, se espera un incremento de 0,16 del desvío estándar de la respuesta.
- TAC2: por cada desvío estándar adicional de temperatura, se espera un incremento de 0,24 del desvío estándar del IET; y por cada desvío estándar adicional de precipitación se espera un incremento de 0,20 del desvío estándar de la respuesta.
- TAC5: por cada desvío estándar adicional de temperatura, se espera un incremento de 0,28 del desvío estándar del estado trófico; por cada desvío estándar adicional de nivel se espera una disminución de 0,23 del desvío estándar de la respuesta; por cada desvío estándar adicional de presión atmosférica se espera una disminución de 0,14 del desvío estándar de la respuesta; y por cada desvío estándar adicional de humedad relativa se espera un incremento de 0,19 del desvío estándar de la respuesta.

Cabe notar que en el caso de los sitios C5 y TAC4 los mejores modelos no presentan coeficientes significativos para las variables hidrológicas y meteorológicas. Por otro lado, puede observarse como la temperatura es la variable cuyo efecto relativo es más intenso, mucho más intenso más cerca de la superficie que en el fondo. En el caso de la precipitación, como segunda variable relevante, el efecto relativo es más intenso a nivel fótico que a nivel subsuperficial. En cuanto al nivel del embalse, pareciera haber un efecto relativo más intenso cerca de la superficie que en el fondo.

En lo que se refiere a la significatividad de los coeficientes de términos autorregresivos y de media móvil, en el caso del modelo bayesiano jerárquico todos los términos autorregresivos de orden 1 son significativos excepto para C4 y TAC4 donde no hay ningún término autorregresivo significativo, mientras que sí lo es el término de media móvil en ambos casos, mientras que este mismo término no es significativo para ningún otro sitio y perfil. Por tal motivo, podría suponerse que, en la zona intermedia, caracterizada por la transición, el IET se explica menos por un fenómeno autorregresivo y responde más a eventos actuales siendo más fluctuante. Cabe notar además que, en los modelos de regresión dinámica, tanto C4 como TAC4 no fueron objeto de diferenciación y tienen un término de media móvil con retardo estacional, lo cual refuerza la idea anterior. En el modelo de regresión dinámica los términos autorregresivos y de media móvil son mayormente significativos en todos los casos, mientras que en el modelo de espacio de estados multivariado los coeficientes del proceso de estados autorregresivo son siempre significativos. Esto último sustenta con evidencia estadística la importancia de los procesos autorregresivos (y en menor medida de media móvil) como se mencionó con anterioridad en este capítulo.

Por último y relativo a la significatividad de los coeficientes en general, vale la pena mencionar que no existe un consenso al respecto de la prueba de hipótesis de significatividad de parámetros. Así, Hyndman (2019) cuestiona las conclusiones que se pueden extraer en relación a la significatividad de los coeficientes en los modelos de regresión dinámica. Por otra parte, en lo que respecta al paradigma bayesiano, no existe la noción de p-valor en la determinación de significatividad, sino que lo que se evalúa es el intervalo de credibilidad asociado al coeficiente (la probabilidad de que el coeficiente esté contenido en el intervalo es del 95%). Bajo este enfoque también sería discutible hablar de coeficientes significativos. A efectos de poder hacer múltiples comparaciones y en aras de la simplicidad y la máxima rigurosidad, se optó por incluir la significatividad de parámetros en este estudio, pero al mismo tiempo se reconoce que también sería válida una evaluación integral considerando todos los parámetros independientemente de su significatividad, en tanto todos contribuyen a la variabilidad de la serie de tiempo respuesta.

En lo que respecta a las investigaciones realizadas en torno a la pregunta que motiva este trabajo, existen varios estudios. Trabajos similares, muestran que la temperatura alta afecta fuertemente las tasas de crecimiento y el metabolismo del fitoplancton y es un factor importante que produce la floración de las algas (Weiyang et al., 2013), siendo la biomasa algal el principal indicador biológico que se busca capturar por medio del IET de Carlson, por la relevancia pública de los eventos de floración (Carlson, 1976). Por otro lado, en lo relativo a la precipitación y el nivel del embalse, se reporta que el estado trófico está relacionado a las fluctuaciones estacionales en la hidrología del sistema, controlado por las lluvias, lo cual es una característica común de los embalses de regiones semiáridas (Barbosa Chaves et al., 2014). En el trabajo citado los autores analizan el comportamiento hidrológico y la calidad del agua del embalse en un período de dos años. Así, se observaron valores bajos del IET durante la primera

mitad de 2010 (uno de los años menos lluviosos en varios años), durante el escaso período de lluvias, donde el nivel del embalse se encontraba alto producto de la inusualmente abundante precipitación registrada el año previo a causa del fenómeno del Niño. De esta manera, se observa el efecto negativo del nivel sobre el IET (a mayor nivel del embalse, menor IET). En la segunda mitad del año, en la estación seca, el volumen del embalse se redujo y se observó un incremento del IET, nuevamente evidenciando el efecto negativo del nivel. Así en 2011, siendo esas las condiciones iniciales de la temporada húmeda, donde se combinó bajo nivel del embalse y caída normal de lluvias, el estado trófico se mantuvo alto. Hacia la segunda mitad del año, el IET fue decreciendo gradualmente y la tendencia se mantuvo durante la temporada seca, se restableció el volumen del embalse a su nivel normal producto del agua acumulada, y el IET disminuyó, coincidiendo también con la menor cantidad de precipitación. Aquí se observa el efecto combinado de las dos variables: el negativo del nivel y el positivo de la lluvia.

Barbosa Chaves et al. (2014) indican además que, en embalses de regiones semiáridas se producen condiciones propicias para la floración de la división *Cyanobacteria*, en un contexto de nivel de agua reducido y alta temperatura, lo cual es consistente exactamente con lo que expresa el mejor modelo en DCQ1. En este caso, como en DSA1, ya se había observado además tendencia creciente de las variables que componen el estado trófico (ver 3.1.1). Se cita asimismo el enriquecimiento de nutrientes en el agua, tal es el caso del fósforo, como producto de las lluvias intensas que arrastran el material hacia el embalse, lo cual sería coincidente con la importancia de la lluvia para explicar el aumento del estado trófico que se observa en varios de los modelos obtenidos en este estudio. Cabe notar que, en ambos estudios citados, se empleó un promedio del IET (ponderado en el segundo estudio) tal cual se aplicó en el presente trabajo a los niveles superficiales, siendo una novedad en este caso el uso del índice en profundidad a partir del fósforo total únicamente.

Otro estudio de calidad de agua basado en diferentes indicadores a los empleados en este estudio, menciona que la presión atmosférica y la temperatura, junto con la radiación solar, son las principales variables que determinan la calidad del agua; mientras que precipitación, humedad relativa y velocidad del viento tienen un rol secundario (Samu et al., 2013). Las altas temperaturas, radiación y presión atmosférica, intensifican la producción de materia orgánica y falta de oxígeno en lagos eutróficos asociados también con mayor evaporación y reducción del nivel. Esto último es consistente con los modelos en perfil subsuperficial y zona fótica donde temperatura y presión atmosférica tienen un efecto positivo sobre la eutrofización y también, probablemente de manera menos intensa (sólo en modelo de espacio de estados multivariado), con el efecto negativo del nivel ya observado.

Por otro lado, un estudio realizado sobre el efecto del cambio climático a largo plazo sobre la calidad del agua, aporta pruebas para sustentar la teoría de que el calentamiento global empeora la condición trófica de los lagos (Komatsu et al., 2007). Esta conclusión es consistente con el efecto positivo de la temperatura en los distintos modelos como variable relevante, y además con la tendencia creciente en el aumento de la temperatura a lo largo de los 18 años observados, como se pudo apreciar en la sección 3.1.2 (ver análisis de la variable temperatura).

En lo que respecta a las principales limitaciones de este trabajo, podría señalarse, en primer término, que el estudio de los efectos temporales de las variables regresoras podría haber sido más exhaustivo en términos de combinaciones de alcances y opciones de modelado. Por otro lado, el estudio de las variables regresoras podría haber sido más extenso, por ejemplo, en

cuanto a selección, uso de variables categóricas, análisis de la correlación cruzada (temporal) entre pares de variables, o exploración del efecto de los alcances en la intensidad de la correlación. Finalmente, el hallazgo de patrones temporales (fundamentalmente ciclos y cambios en la estacionalidad) observados en el análisis exploratorio de las variables de calidad del agua requeriría indagar más en las razones y efecto aislado de los mismos en relación a las demás variables. Estas limitaciones bien podrían funcionar como disparadores para proyectos futuros que puedan mejorar y enriquecer las contribuciones del presente trabajo.

A modo de conclusión puede decirse que a partir de este estudio fue posible hacer un aporte a nivel metodológico y a nivel de resultados para permitir un conocimiento más profundo del efecto de las condiciones meteorológicas sobre la calidad del agua. En cuanto al primer punto, se propuso:

- Una metodología consistente de trabajo para el preprocesamiento de grandes volúmenes de datos provenientes de estaciones meteorológicas.
- Una metodología de prueba de tres tipos distintos de modelos (basados en los dos grandes paradigmas de la estadística: frecuentista y bayesiano, a partir de distintos supuestos sobre el alcance y la relación entre sitios y perfiles de agua.
- Un encuadre multidimensional de las series: tiempo, ubicación horizontal (sitio) y vertical (perfil) para el análisis del estado trófico, lo cual constituye un aporte innovador en la medida en que los estudios realizados al momento sólo implican un único nivel de profundidad (subsuperficial) acorde a la propuesta original de Carlson.
- Un procedimiento de predicción para el modelo de espacio de estados multivariados no disponible en el paquete de Holmes et al. (2018b).

En lo que se refiere a los resultados, se puede concluir lo siguiente:

- Existe una influencia del alcance de las variables meteorológicas sobre la variación de la calidad el agua, siendo este alcance mayor conforme se analiza a mayor nivel de profundidad.
- Los procesos autorregresivos y de media móvil son necesarios para explicar la dinámica del estado trófico afectado por variables hidrológicas y meteorológicas.
- La temperatura y la presión atmosférica, la precipitación y la humedad tienen una incidencia específica cuantificable sobre el empeoramiento de la calidad del agua; mientras que el nivel del embalse y la velocidad del viento inducen el efecto inverso.
- Existe una tendencia hacia el empeoramiento de la calidad del agua en las desembocaduras en las tres variables que constituyen el IET.

En lo que respecta a la gestión del cuerpo de agua, como primer aporte se realizó la revisión, control, depuración, estructuración y sistematización de datos hidrológicos y meteorológicos históricos producidos por INA-CIRSA que se encontraban dispersos y difícilmente accesibles. El segundo aporte, fue la consolidación y generación de una base de datos histórica, con datos consistentes y fidedignos, incorporada como parte del sistema de datos institucional del INA-CIRSA. En particular, en lo que respecta a los datos de precipitación, se logró completar las series de tiempo correspondientes a todas las estaciones que reportan on-line (desde 2017), y que se encuentran disponibles mediante acceso web para usuarios miembros. Este avance sienta precedente para hacer lo propio con las demás variables hidrológicas y meteorológicas.

Es indudable que para poder gestionar y avanzar en la investigación es imprescindible contar con una fuente de información única, completa, accesible y confiable, y en este sentido se ha hecho un aporte fundamental a partir del trabajo realizado.

Finalmente, los modelos propuestos (en particular los modelos bayesiano jerárquico y de regresión dinámica) permitirán hacer seguimiento de la evolución de los procesos de eutrofización, en especial en un contexto de clima cambiante. La aplicación de estos modelos, que a su vez se podría sumar a otras iniciativas en curso en INA-CIRSA tales como la estimación del estado trófico por medio de imágenes satelitales, podría contribuir de manera decisiva al entendimiento integral del fenómeno, la estimación de eventos relevantes y la toma de decisiones vinculadas principalmente al uso del agua para potabilización, riego y recreación.

En base a lo dicho anteriormente es de vital importancia concientizar, con base en resultados científicos, acerca del estado actual de eutrofización del ESR y las perspectivas a futuro. Si bien se considera esencial la asignación de recursos para el saneamiento ecológico, uno de los desafíos actuales es encontrar el modo en que los avances en el conocimiento del problema ambiental, al que este estudio pretende contribuir, puedan articularse cada vez más con la gestión de esta fuente de agua de vital importancia para el desarrollo de la región.

I Anexo: Monitoreo de variables de calidad del agua extendido

Datos de calidad de agua del lago

Los sensores para medir las variables físico-químicas *in-situ* por medio de la sonda multiparamétrica, mencionada en la sección 2.2.1, se detallan a continuación¹⁴ a partir de la hoja técnica de especificación del equipo (YSI, 2009):

- pH: el sensor mide el potencial de hidrógeno por medio del método del electrodo de vidrio.
- Oxígeno Disuelto (OD): en este caso el instrumento utiliza el principio polarográfico de estado estable, para medir la cantidad de oxígeno disuelto en el agua. Esta variable se mide en miligramos por litro (mg/L).
- Saturación de Oxígeno Disuelto (SOD): el sensor utiliza el mismo principio indicado para OD, para medir el porcentaje (%) de oxígeno disuelto en el agua.
- Conductividad: el dispositivo mide la conductividad eléctrica por medio de un sensor que consta de una celda de cuatro electrodos con auto rango. Este parámetro se mide en micro-siemens por centímetro ($\mu\text{S}/\text{cm}$) y está expresada en relación a una temperatura estándar de 25°C.
- Sólidos Disueltos Totales (SDT): este parámetro representa todas las sustancias solubles contenidas en el líquido, se calcula a partir de una conversión de la medición de conductividad, utilizando una constante: 0,65. La unidad de medición es gramos por litro (g/L).
- Temperatura: el sensor mide la temperatura del agua en grados centígrados (°C), por medio de un termistor.

En lo que respecta a las condiciones del entorno y a las condiciones generales del agua, se observan:

- Condiciones climáticas: el observador registra el estado del cielo, si llueve, si hay viento, entre otras.
- Turbidez algal: en este caso el observador analiza el aspecto del agua, para indicar el o los colores que se ven mayormente (verde, marrón, rojizo, entre otros).
- Aspecto de algas: en caso de observarse, se toma nota de las características de las algas observadas a simple vista en el agua, tales como disposición, si presentan floración o espuma, entre otras.

En lo que respecta a las variables que se determinan en laboratorio, se tienen además de las mencionadas en la sección 2.2.1, las siguientes:

- Nitrógeno de Amonio (N-NH_4^+): esta variable se mide en micro-gramos por litro ($\mu\text{g}/\text{l}$) y se determina por dos métodos: A – colorimétrico del indofenol por espectrofotometría SM 4500 N y B – flujo segmentado SM 4120.
- Nitrógeno de Nitrato (N-NO_3^-): esta variable se mide en mili gramos por litro (mg/l) y se determina por el método A – cromatografía iónica SM 4110.

¹⁴ A lo largo de los años, las sondas han cambiado y con el avance de la tecnología el principio de medición no ha sido siempre el mismo. En este apartado se describen los correspondientes a la sonda actualmente en uso.

- Nitrógeno de Nitrito (N-NO₂⁻): esta variable se mide en µg/l y se determina por medio de dos métodos: A – colorimétrico sulfanilamida espectrofotometría SM 4500 N y B: flujo segmentado SM 4120.
- Fósforo Reactivo Soluble (PRS): esta variable se mide en µg/l por medio de dos métodos: A – espectrofotometría SM 4500 P y B – Flujo segmentado SM 4120.
- Carbono Orgánico Total (COT): esta variable se mide en mg/l por el método de análisis infrarrojo.
- Fitoplancton: esta variable representa 78 géneros de algas que se miden en células por litro (Cel/L) y se determinan por el método de sedimentación y recuento con retícula de Whipple. Los géneros y su *phylum* o división se muestran en la Tabla 32 basado en la taxonomía propuesta en Scagel et al. (1991) y Bourrelly (1972).

Tabla 32 Géneros de alga y su división

Géneros	División
Aulacoseira, Aulacoseira 1, Cocconeis, Cyclotella, Cyclotella 1, Cymatopleura, Cymbella, Diatoma, Gomphonema, Gyrosigma, Hantzschia, Melosira, Melosira 1, Navicula, Nitzschia, Pinnularia, Stephanodiscus, Surirella, Mallomonas, Amphora, Amphipleura, Biddulphia, Epithemia, Fragillaria, Tabellaria	Chrysophyta
Closterium, Staurastrum, Staurodesmus, Ankitrodesmus, Botryococcus Coelastrum, Chlamydomonas, Carteria, Chlorella, Kirchneriella, Monoraphidium, Oocystis, Pediastrum, Planctonema, Planktosphaeria, Scenedesmus, Schroederia, Sphaerocystis, Volvocales (*), Actinastrum, Crucigenia, Dictyosphaerium, Fortiella, Lobocystis, Micractinium, Pandorina, Selenastrum, Spirogyra, Tetraedron, Ulothrix, Volvox	Chlorophyta
Chroomonas, Cryptomonas	Cryptophyta
Trachelomonas, Euglena	Euglenophyta
Ceratium, Peridinium	Pyrrhophyta
Dolichospermum, Dolichospermum 1, Aphanocapsa, Microcystis, Microcystis 1, Pseudoanabaena, Anabaenopsis, Aphanizomenon, Aphanothece, Merismopedia, Oscillatoria, Planktothrix, Raphydiopsis, Raphydiopsis 1, Romeria, Spirulina	Cyanobacteria

Notas: (*) excepcionalmente es orden y no género.

Los géneros acompañados de 1 indican un subgénero no identificado.

Como continuación de la Tabla 4, la Tabla 33 muestra el detalle de las restantes variables medidas en el lago, su alcance y determinación (por simplicidad, las variables referidas a concentraciones de sustancias químicas no incluyen el uso de subíndices ni separadores). Una particularidad de las variables del lago es que algunas resultan de un cálculo, ya sea directo como es el caso de PRS/PT (proporción de fósforo disuelto total sobre fósforo total); o indirecto en función de otras variables, como es el caso de condición de estratificación (si el lago se encuentra en condición de mezcla o estratificación, para mayor detalle referirse al Anexo II).

Tabla 33 Detalle de las variables medidas en el lago, su alcance y determinación (continuación). Las variables correspondientes a los géneros de algas se han agrupado bajo el título *Géneros de Algas*.

Variable	Alcance	Determinación
pH	Por profundidad	Medición
OD	Por profundidad	Medición
SOD	Por profundidad	Medición
Conductividad	Por profundidad	Medición
SDT	Por profundidad	Medición
Temperatura	Por profundidad	Medición
Condición Térmica	Por sitio	Cálculo
NNH4	Por perfil	Medición
< NNH4	Por perfil	Indicación de censura
NNO3	Por perfil	Medición
< NNO3	Por perfil	Indicación de censura
NNO2	Por perfil	Medición
< NNO2	Por perfil	Indicación de censura
PRS	Por perfil	Medición
< PRS	Por perfil	Indicación de censura
PRS/PT	Por perfil	Cálculo
COT	Por perfil	Medición
Algas Totales	Por perfil	Cálculo
<i>Géneros de alga</i>	Por perfil	Medición
Condiciones climáticas	Por sitio	Percepción del observador
Turbidez algal	Por sitio	Percepción del observador
Aspecto de algas	Por sitio	Percepción del observador

Datos de calidad de agua de los ríos

De manera análoga al registro de los datos de calidad de agua del lago, se produce el de datos de ríos, obtenidos de los distintos puntos de muestreo.

Los datos de ríos se estructuran a partir de las siguientes variables de referencia: muestra (identificador que asigna el laboratorio a cada muestra que se analiza), fecha, código (se corresponde a uno de los códigos de los puntos de muestreo que se muestran en la Tabla 2). Por cada combinación de estos valores, se tienen una serie de variables medidas. En este caso, a diferencia de los datos del lago, las mediciones se realizan a un solo nivel de profundidad.

La figura Fig. 86 muestra un ejemplo de la estructura de las variables de referencia de los datos de ríos, mientras que la Tabla 34 muestra el detalle de las variables que se miden y su determinación. En este caso, a diferencia de los datos del lago, cada medición corresponde a una combinación de las variables de referencia, es decir, por cada muestra, fecha y código se tiene un único conjunto de valores medidos.

MUESTRA	Fechas	CODIGO
215263	25-Oct-2011	RCQ
215265	25-Oct-2011	RLM
215264	25-Oct-2011	RLC
215262	25-Oct-2011	RSA

Fig. 86 Ejemplo de las variables de referencia de los datos de ríos

Tabla 34 Detalle de las variables medidas en ríos y su determinación

Variable	Determinación
Temperatura	Medición
Conductividad	Medición
NNH4	Medición
<NNH4	Indicación de censura
NNO3	Medición
<NNO3	Indicación de censura
NNO2	Medición
<NNO2	Indicación de censura
PRS	Medición
PT	Medición

De manera análoga a como ocurre en el caso de los datos del lago, se descartan variables adicionales de mediciones en ríos debido a la alta proporción de faltantes.

II Anexo: Preprocesamiento de datos extendido

Datos de calidad del agua

Datos del lago

Imputación por extensión

- **Condición térmica:** representa una condición del sitio del lago. Se observan un total de 58 casos en que, estando determinada en algunos cortes (al menos subsuperficial) de algunos sitios, no lo estaba en todos. Por tal motivo se completan estos casos. Por otro lado, cabe aclarar que en ningún caso de múltiples valores por sitio se observó discrepancia, es decir, en todos los casos se trataba del mismo estado.
- **Variables de conteo de algas:** en este caso se tiene un total de algas y un conteo por cada género. Si el conteo por género está faltante, pero el total tiene valor, es seguro imputar un valor cero al conteo de géneros, porque el faltante indica que no se observó en el microscopio. Se asume que en toda oportunidad el analista pudo haber detectado cualquiera de las especies cuyo conteo constituye variables de esta base, por lo que si el dato está faltante tuvo que ser porque no se observó la especie correspondiente. A la inversa, si se observa registro de cero para determinados conteos de género (y solo éstos tienen valor), pero el valor del total está faltante, se eliminó el valor cero por interpretarse que se trató de un error en la carga.
- **Variables categóricas:** las variables condiciones climáticas, turbidez algal, y aspecto de algas, se encuentran registradas de manera arbitraria mayormente a nivel subsuperficial, porque refieren todas ellas a atributos que se observan en el ambiente (condiciones climáticas) o en la superficie del lago (turbidez algal y aspecto de algas) en cada sitio. Por tal motivo, se hizo extensivo el registro de las apreciaciones en términos de estas variables a los demás cortes de profundidad.

Imputación teórica por cálculo

A efectos de completar determinadas variables con un alto porcentaje de valores faltantes a partir de otras, se procedió a su cálculo según determinados criterios teóricos. A continuación, se detallan los casos y la solución utilizada. Todos los procedimientos están basados en el conjunto de datos hasta agosto de 2017.

- **SOD.** Los datos faltantes se calcularon a partir de OD y temperatura (t), mediante las fórmulas anidadas en las Ecuaciones (49), (50) y (51), que se muestran a continuación (Water on the web, 2017):

$$\ln P = 5.25 * \ln \left(1 - \frac{h}{44.3} \right) \quad (49)$$

Donde:

P es presión (atm).

h es la altitud (km) relativa a la presión parcial estándar de 760 mm-Hg al nivel del mar.

$$C_p = C^* * P * \left[\frac{(1 - P_w/P) * (1 - \theta * P)}{(1 - P_{wv}) * (1 - \theta)} \right] \quad (50)$$

Donde:

C_p es la concentración de oxígeno en equilibrio a presión no estándar en mg/L.

C^* es la concentración de oxígeno en equilibrio a presión estándar de 1 atm, mg/L. $C^* = e^{7.7117 - 1.31403 * \ln(t + 45.93)}$.

P es la presión no estándar en atm.

P_{wv} es la presión parcial del vapor de agua, atm, calculada como: $\ln(P_{wv}) = 11,8871 - 3840,70/T - 216961/T^2$.

T es la temperatura en °K.

θ es $0,000975 - (1,426 * 10^{-5}t) + (6,436 * 10^{-8}t^2)$.

t es la temperatura en °C.

Finalmente,

$$SOD(\%) = \frac{100 * OD}{C_p} \quad (51)$$

Donde

OD en mg/L es el oxígeno disuelto

SOD el porcentaje de saturación de oxígeno disuelto

En primera instancia, a efectos de comprobación de la aptitud de la fórmula, se la aplicó a los 630 casos en que se contaba con un valor registrado de SOD (y las variables intervinientes en el cálculo tenían dato), de manera tal de comparar contra lo calculado. Se consideraron dos indicadores en este sentido. Por un lado, la raíz cuadrada del error cuadrático relativo, que establece la comparación del modelo de cálculo versus el modelo sencillo de promedio. Por otro lado, el coeficiente de determinación, que indica la proporción de la varianza de la variable SOD que se explica por medio de OD y Temperatura, establece la bondad de ajuste. Los resultados indican que el error es del 0,20 y el coeficiente de determinación de 0,96. Considerándose aceptables, se aplica al conjunto de datos de 1698 registros donde SOD está incompleto (y existen valores de OD y Temperatura), imputándose de esta manera el valor calculado.

- SDT. En este caso a efectos de completar los datos faltantes, se aplicó una fórmula de conversión en la Ecuación (52), basada en una constante (YSI, 2009).

$$SDT = 0.65 * CE/1000 \quad (52)$$

Donde

CE es la conductividad específica, es decir la conductividad eléctrica (C) medida a 25°C y escalada a nivel micro (ver Anexo I). Las sondas miden la conductividad eléctrica y estandarizan a la temperatura de 25°C de forma automática, aplicando la fórmula en la Ecuación (53) (reescrito de Horiba (1991)):

$$CE = C / 1 + (r * (t - 25)) \quad (53)$$

Donde

r es el coeficiente de corrección por temperatura (0,02), indica la variación de conductividad por grado de temperatura.

t es la temperatura de la solución.

Por lo tanto en este caso $CE = C$, y no es necesario aplicar un factor de corrección de temperatura para obtener CE , se aplica directamente la fórmula de SDT. Para probar la imputación, se procedió de manera análoga a lo hecho anteriormente para OD, utilizando 1027 registros con dato, se pudo observar que el error es de 0,13 y el coeficiente de determinación de 0,98. Se considera aceptable, aplicándose al conjunto de 1864 registros restantes.

- Estratificación térmica: siendo una variable secundaria calculada a partir de la profundidad y la temperatura, presenta valores faltantes en los sitios donde se realizan mediciones solo a nivel subsuperficial o fondo desde diciembre de 1998 (tal es el caso de los puntos de desembocadura de ríos). Cabe aclarar que se trata de una variable definida a nivel del sitio, y por ende todos los registros para una misma fecha y sitio presentan el mismo valor. Se genera una nueva variable calculada con el siguiente criterio (respetando el orden de precedencia en la evaluación de cada condición) (Andreoni et al., 2018):
 1. Si la variación en grados de temperatura es mayor o igual a 1 grado (con redondeo a un dígito decimal) entre dos registros con valores de Z consecutivos (con 1 metro o menos de diferencia, excepto para corte subsuperficial, donde se asume no determinante), entonces se considera que existe *estratificación*.
 2. Si la variación de temperatura es de 0,9 (el redondeo se hace a un dígito decimal) para una variación en profundidad que no supere el metro (excepto para corte subsuperficial, donde se considera no determinante), no puede determinarse, se requiere revisión y se indica como T, haciendo alusión al incumplimiento de la condición de temperatura. En estos casos, donde la variación de temperatura es próxima a 1 puede decidirse estratificación en función de otras variables.
 3. Si la variación de temperatura es superior al grado para una variación de profundidad que supera el metro, se requiere revisión y se indica como Z, haciendo alusión al incumplimiento de la condición de profundidad. En caso en que exista variación de temperatura superior al grado con paso de profundidad superior al metro, puede decidirse estratificación en función de otras variables.
 4. En cualquier otro caso en que la temperatura no supere el grado se considera *mezcla* (se tomaron todos los casos en que se tenían al menos dos puntos para comparar temperatura, aunque los demás valores de temperatura faltaran o la distancia en Z fuera superior a 1).

De esta manera, teniendo la variable original 1047 casos faltantes, pudieron calcularse 503 casos como mezcla y 189 casos como estratificación. De los restantes, en 23 casos es preciso analizar la condición de temperatura, en 13 la condición de profundidad y finalmente 319 permanecen sin completar.

Cabe mencionar, que, comparando los casos completos de la variable original respecto de la calculada, el 91% de los casos de mezcla fueron calculados como mezcla (teniendo en cuenta que, de los restantes un 7% se calculó como T o como Z y un 1% se calculó

como estratificación) y 81% de los casos de estratificación fueron calculados como estratificación (teniendo en cuenta que, de los restantes un 16% se calculó como T o como Z y un 2% se calculó como mezcla, verificándose posteriormente que era correcto en 3 casos que se corrigieron). Por último, excepcionalmente se habían identificado 4 casos como indeterminaciones, de los cuales 3 fueron calculados como Z y 1 como mezcla.

Se crearon adicionalmente, dos variables de referencia:

1. Z Termoclina: profundidad a la que se produce la estratificación (480 casos).
2. Delta T: variación de temperatura entre los códigos 1 (perfil subsuperficial) y 5 (fondo), para los sitios donde estén disponibles. Esta variable refleja la variación de temperatura entre las capas más distantes de agua.

Estructuración de variables categóricas

Las variables de tipo texto en la base de datos, presentan la característica de tener mucha variabilidad producto de que el registro no ha sido estructurado. Por tal motivo, con el objetivo de conseguir mayor homogeneidad se procede a, por un lado, discriminar los aspectos a los que se refiere cada variable, tratándolos como variables separadas; y por otro lado definir un cierto número de categorías específicas para cada una de estas nuevas variables, de manera de *mapear* o hacer corresponder los datos en la variable original.

- Condiciones climáticas: se trata de una variable de texto libre, donde se tienen 198 valores distintos presentes en 1089 registros. Se refiere en general a tres aspectos: tiempo, viento y presencia de neblina. Así se crearon tres variables nuevas, por categoría:
 - Nubosidad: DESPEJADO (presencia de palabras “despejado” y “soleado” y no de la palabra “nublado”) / PARCIALMENTE NUBLADO (presencia de palabra “parc”) / NUBLADO (presencia de palabra “nublado” y no de la palabra “parc”). Es importante la distinción entre “nublado” y “parcialmente nublado” como indicador de luminosidad baja o media, respectivamente.
 - Tipo lluvia: LLUVIA (presencia de palabra “lluvia”) / LLOVIZNA (presencia de la palabra “llovizna”) / SIN LLUVIA (se asume que no llovió en caso que no se especifique explícitamente en alguno de los registros no nulos, por tratarse de una situación excepcional).
 - Viento: CALMO (presencia de palabra “calmo”) / LEVE (presencia de palabras “leve”, “suave” en primer término frente a “calmo”) / INTENSO (presencia de palabras “fuerte”, “ventoso”, “moderado” en primer término frente a otras).
 - Presencia Neblina: S/N (presencia o ausencia de palabra “neblina”, se asume que no hubo neblina en caso que no se especifique explícitamente en alguno de los registros no nulos, por tratarse de una situación excepcional).

Vale aclarar que, dado que se trata de una apreciación subjetiva, se minimizó la cantidad de categorías para así evitar tener datos demasiado atados a la perspectiva del observador en cuestiones sensibles como el estado del tiempo.

- Turbidez Algal: en este caso se presentan 42 valores distintos, repartidos en 1051 registros. Se refiere a los colores que presenta el agua, los cuales se pueden estandarizar. En este caso se seleccionan unos pocos colores representativos, ya que en campo no se puede hacer una distinción mayor. Así se crea una nueva variable:

- Aspecto Turbidez Algal: MARRON (presencia de la palabra “MARRON” o “MARRÓN” en primer término frente a otras, excepto “VERDE”) / VERDE (presencia de la palabra “VERDE” en primer término frente a otras, excepto “MARRON”, “MARRÓN” y “AZUL”) / VERDE MARRON (presencia de las palabras “VERDE” y “MARRÓN” combinadas en primer término frente a otras) / VERDE AZUL (presencia de las palabras “VERDE” y “AZUL” combinadas en primer término frente a otras) / ROJIZO (presencia de la palabra “ROJIZO” en primer término frente a otras) / AMARILLO (presencia de la palabra “AMARILLENTO” en primer término frente a otras) / NEGRO (presencia de las palabras “OSCURO”, “NEGRUZCO”, “GRISASEO”, “TURBIO” en primer término frente a otras) / INCOLORO (presencia de las palabras “SIN TURBIDEZ”). La categoría VERDE MARRON se considera frecuente y representa, simultáneamente y en “balance”, la presencia de algas de un tipo de color y otro. En el caso de VERDE AZUL se trata de un fenómeno raro, pero de relevancia. Las demás combinaciones de colores no son importantes, y son absorbidas por la clase que se menciona en primera instancia.
- **Aspecto de Algas**: en este caso se tienen 39 categorías repartidas en 1000 registros. Se refiere principalmente a tres aspectos: visibilidad, tipo de alga y presencia de espuma. Por tal motivo se generan tres variables nuevas de manera tal de estructurar esta información de la forma más conveniente para su posterior análisis:
 - Visibilidad Algas: S (ausencia de las palabras “NO VISIBLE”, “ESPUMA” solamente, “FLORACION” solamente o “ESPUMA FLORACION” solamente) / N (presencia de las palabras “NO VISIBLE”).
 - Disposición Algas: NO VISIBLE (presencia de las palabras “NO VISIBLE”) / ACUMULOS (presencia de las palabras “ACUMULOS”, “ACUMULOS ESPUMA”, “DISPERSAS”, “CIANOS DISPERSAS”) / ACUMULOS Y SUSPENSION FINA (presencia de las palabras “ACUMULOS Y SUSPENSION FINA”, “DISPERSAS Y SUSPENSION FINA”, “CIANOS DISPERSAS”) / PINCELADAS (presencia de la palabra “PINCELADAS”) / MANCHA (presencia de la palabra “MANCHAS”) / PINCELADAS Y SUSPENSION FINA (presencia de las palabras “SUSPENSION FINA Y PINCELADAS”) / PINCELADAS Y ACÚMULOS (presencia de las palabras “PINCELADAS Y ACÚMULOS”) / SUSPENSIÓN FINA (presencia de las palabras “SUSPENSIÓN”, “CERATIUM”, en primer término o a continuación de otras palabras excepto alguna de las categorías previas). En todos los casos con y sin acento, y con y sin “y” en el caso de las conjunciones.
 - Espuma: S (presencia de la palabra “ESPUMA”) / N (ausencia de la palabra “ESPUMA”).
 - Floración: S (presencia de la palabra “FLORACION”) / N (ausencia de la palabra “FLORACION”).

En el caso de las variables binarias se asume que la no presencia de la indicación indica ausencia.

Datos de ríos

En el caso de los datos de ríos, debido a la presencia de faltantes que no pueden obtenerse por imputación teórica ni por extensión no fue necesario aplicar ningún procedimiento. Tampoco fue necesario estructurar variables.

III Anexo: Análisis exploratorio extendido

En este anexo se presentan las series correspondientes a variables adicionales de calidad del agua. En lo que respecta al lago, y como ya se mencionó, se tienen variables físico-químicas, biológicas y suplementarias. Dentro del primer grupo, se encuentran las siguientes series por sitio y perfil determinadas *in-situ*: pH (Fig. 87), oxígeno disuelto (Fig. 88), saturación de oxígeno disuelto (Fig. 89), conductividad eléctrica (Fig. 90), sólidos disueltos totales (Fig. 91) y temperatura del agua (Fig. 92); y las siguientes determinadas en laboratorio: nitrógeno de amonio (Fig. 93), nitrógeno de nitrato (Fig. 94), nitrógeno de nitrito (Fig. 95), fósforo reactivo soluble (Fig. 96), relación fósforo reactivo soluble sobre fósforo total¹⁵ (Fig. 97) y carbono orgánico total (Fig. 98). Dentro del grupo de las variables biológicas, se tienen las siguientes series¹⁶ determinadas en laboratorio: algas por perfil en sitio centro (Fig. 99), algas por perfil en sitio TAC (Fig. 100) y algas por sitio en perfil subsuperficial (Fig. 101); y estas otras determinadas *in-situ*, por sitio: aspecto de algas (Fig. 102) y disposición de algas (Fig. 103). Por último, dentro del grupo de variables suplementarias, se encuentran las siguientes series determinadas *in-situ* y por sitio: nubosidad (Fig. 104), viento (Fig. 105) y condición de estratificación térmica (Fig. 106).

Por otro lado, en lo relativo a los ríos, se tienen únicamente variables físico-químicas, parte determinadas *in-situ*, a saber, las siguientes series: nivel (Fig. 107), pH (Fig. 108), temperatura del agua (Fig. 109) y conductividad eléctrica (Fig. 110); y parte, en laboratorio: nitrógeno de nitrato (Fig. 111), nitrógeno de nitrito (Fig. 112), fósforo reactivo soluble (Fig. 113), fósforo total (Fig. 114) y relación fósforo reactivo soluble sobre fósforo total¹⁷ (Fig. 115).

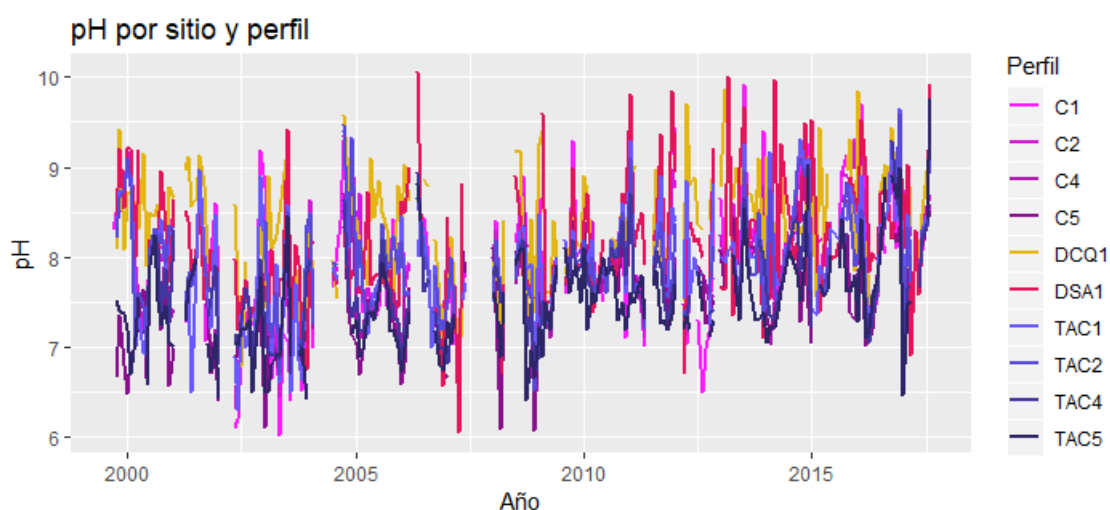


Fig. 87 Series de tiempo de pH por sitio y perfil

¹⁵ Se trata de un cálculo a partir de las variables determinadas en laboratorio.

¹⁶ Se aplica una transformación $\ln(x+1)$ para reducir el rango de valores (muy dispares), siendo x células por litro.

¹⁷ Se trata de un cálculo a partir de las variables determinadas en laboratorio.

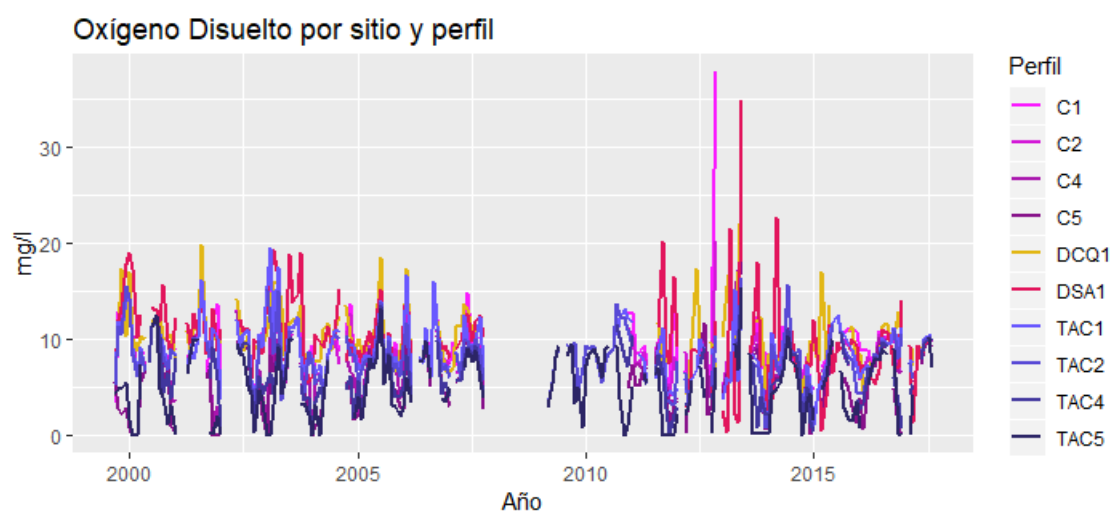


Fig. 88 Series de tiempo de oxígeno disuelto por sitio y perfil

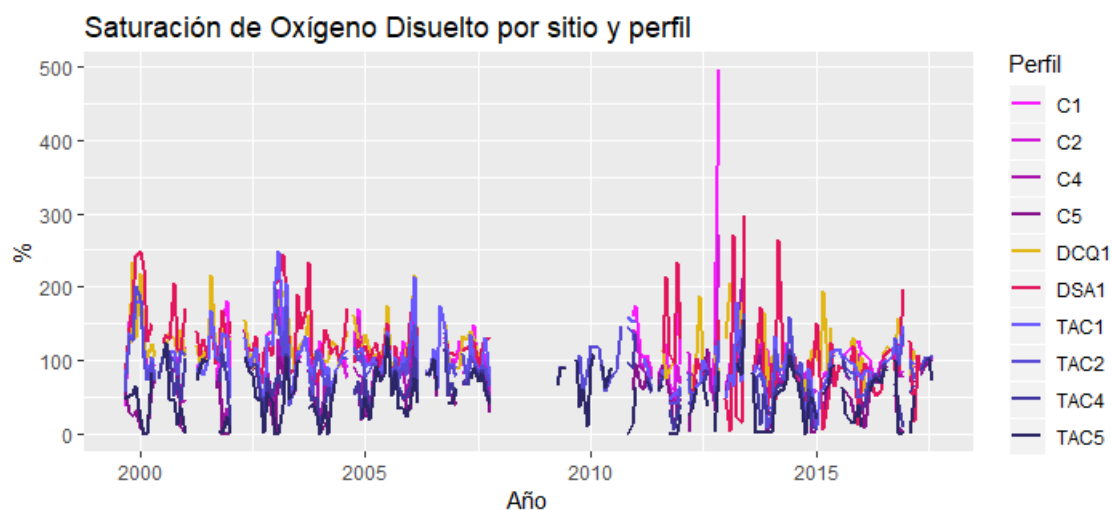


Fig. 89 Series de tiempo de saturación de oxígeno disuelto por sitio y perfil

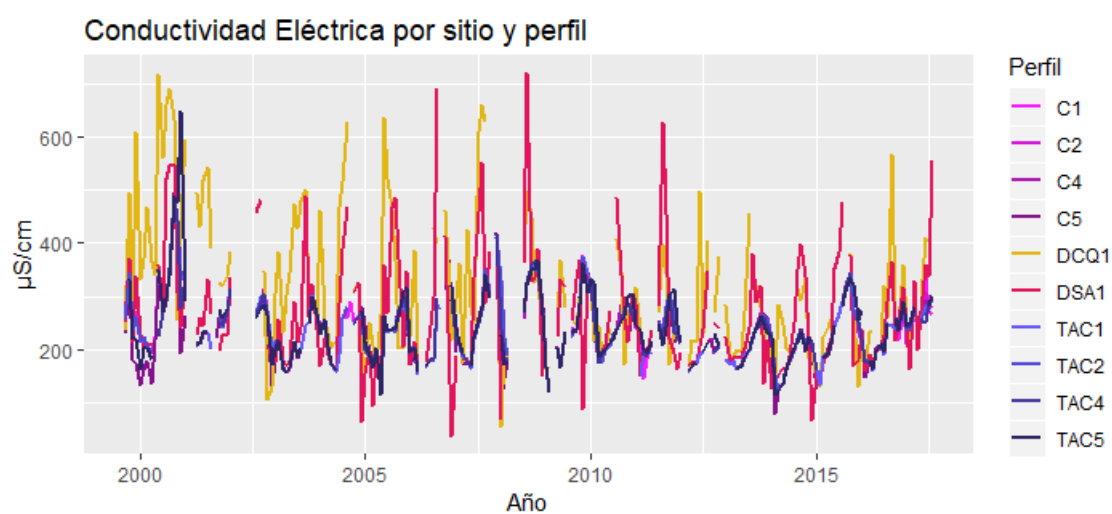


Fig. 90 Series de tiempo de conductividad eléctrica por sitio y perfil

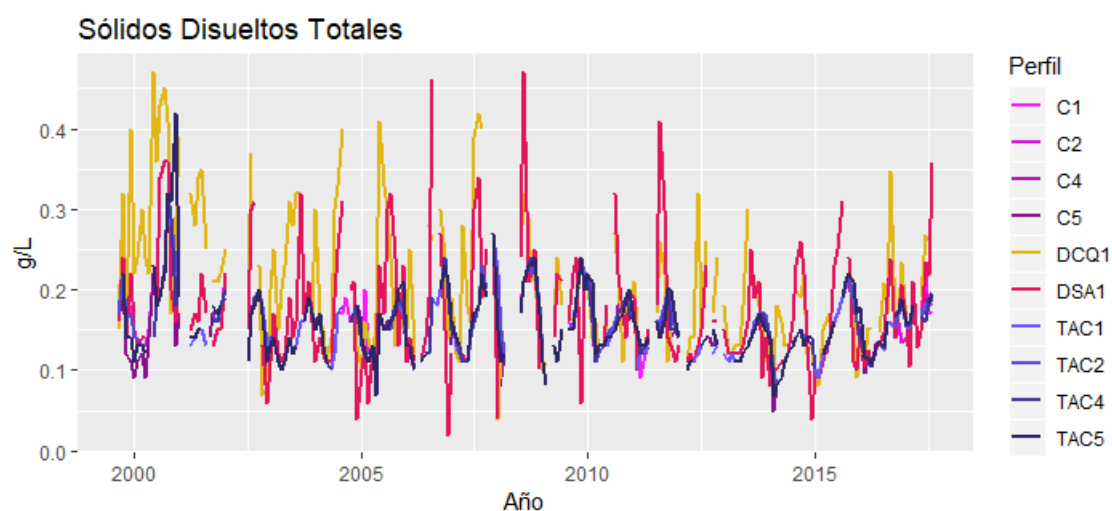


Fig. 91 Series de tiempo de sólidos disueltos totales por sitio y por perfil

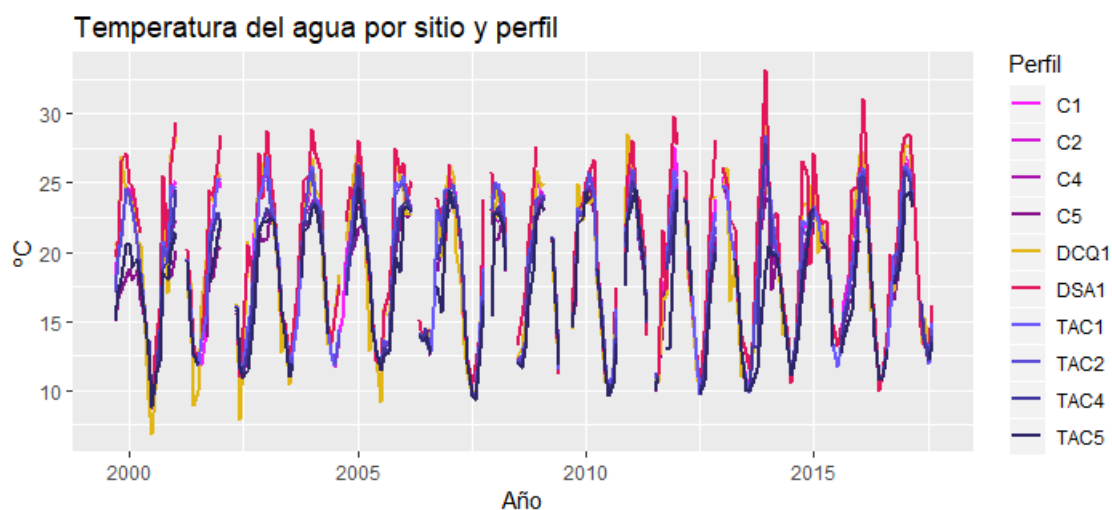


Fig. 92 Series de tiempo de la temperatura del agua por sitio y perfil

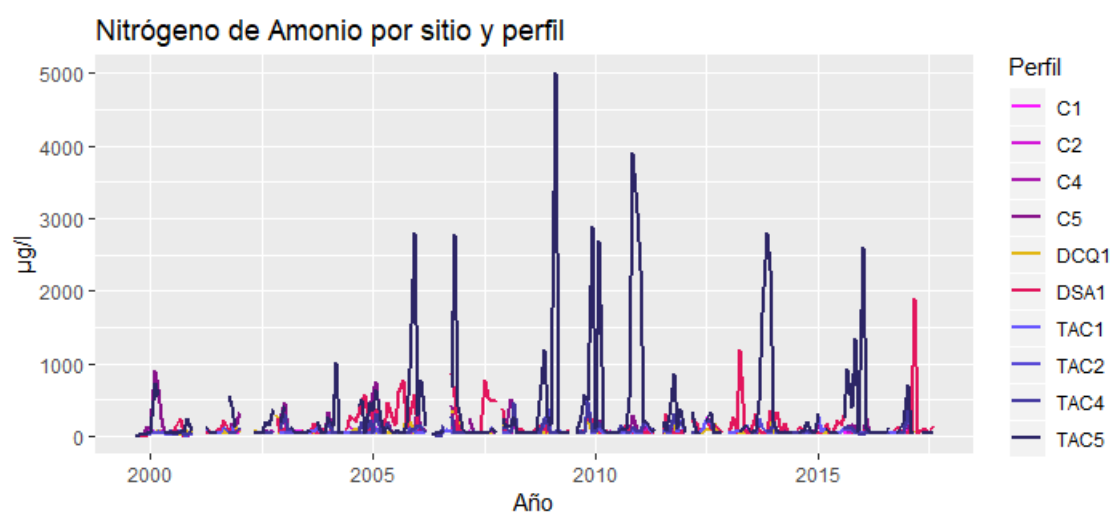


Fig. 93 Series de tiempo de nitrógeno de amonio por sitio y perfil

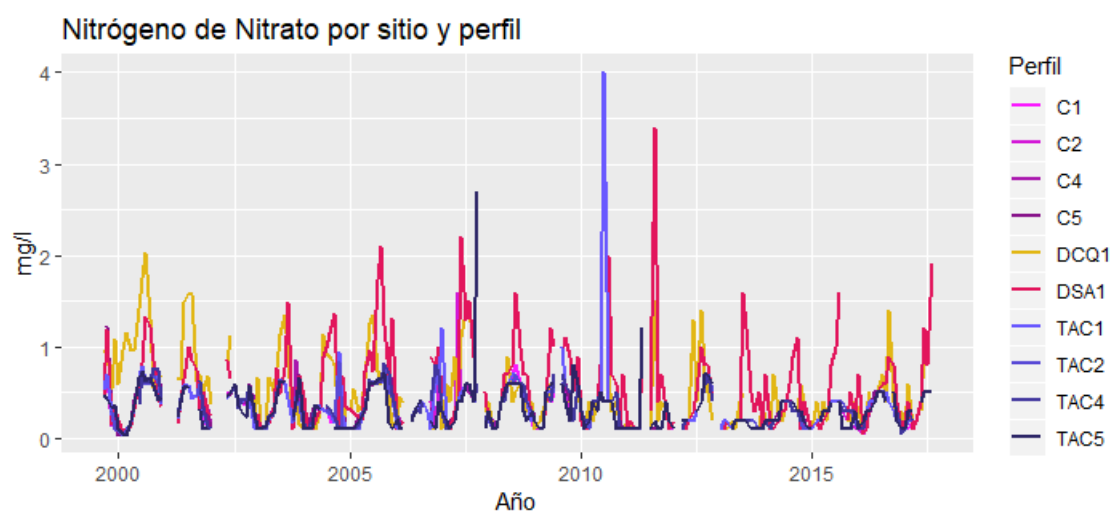


Fig. 94 Series de tiempo de nitrógeno de nitrato por sitio y perfil

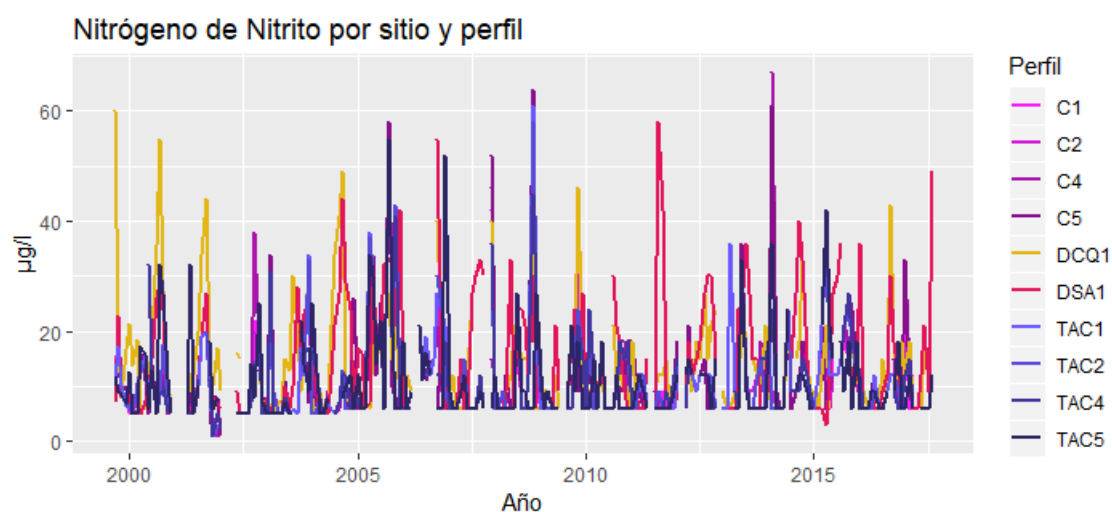


Fig. 95 Series de tiempo de nitrógeno de nitrito por sitio y perfil

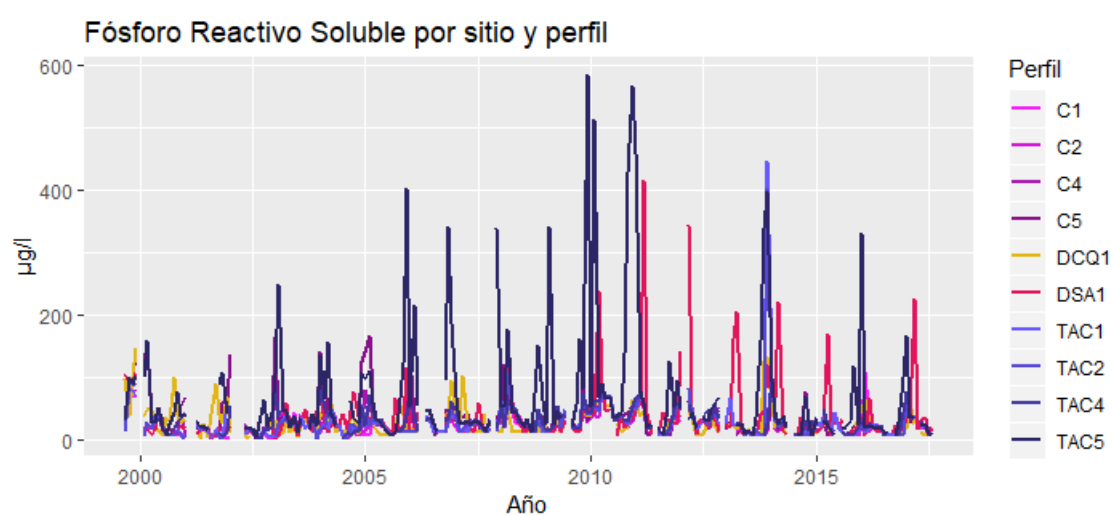


Fig. 96 Series de tiempo del fósforo reactivo soluble por sitio y por perfil

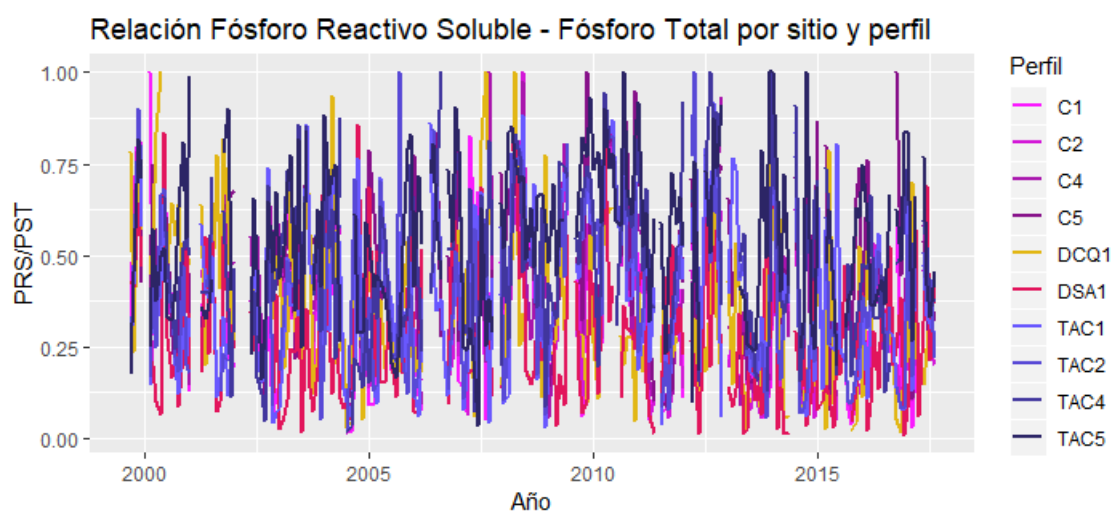


Fig. 97 Series de tiempo de la relación fósforo reactivo soluble sobre fósforo total por sitio y perfil

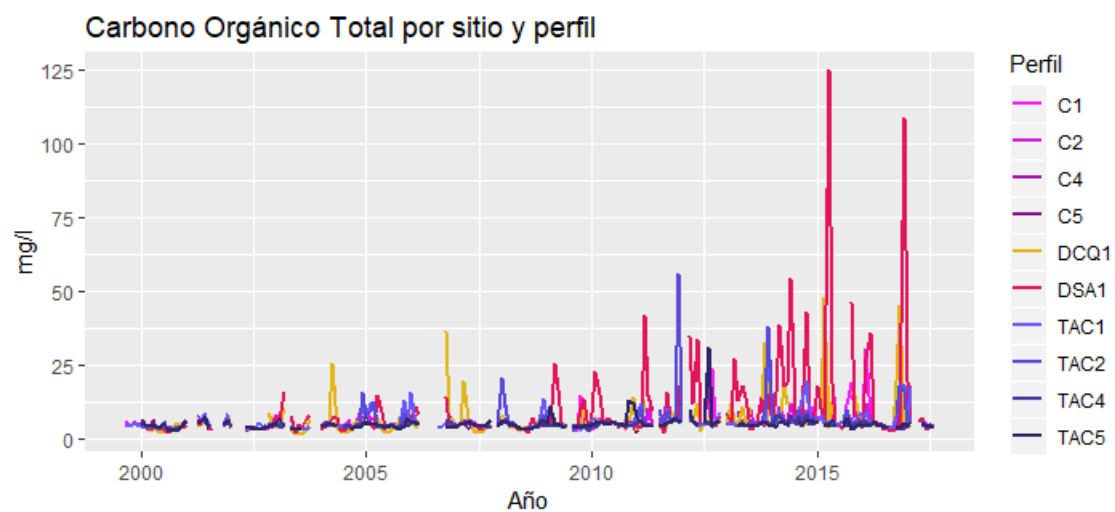


Fig. 98 Series de tiempo de carbono orgánico total por sitio y perfil

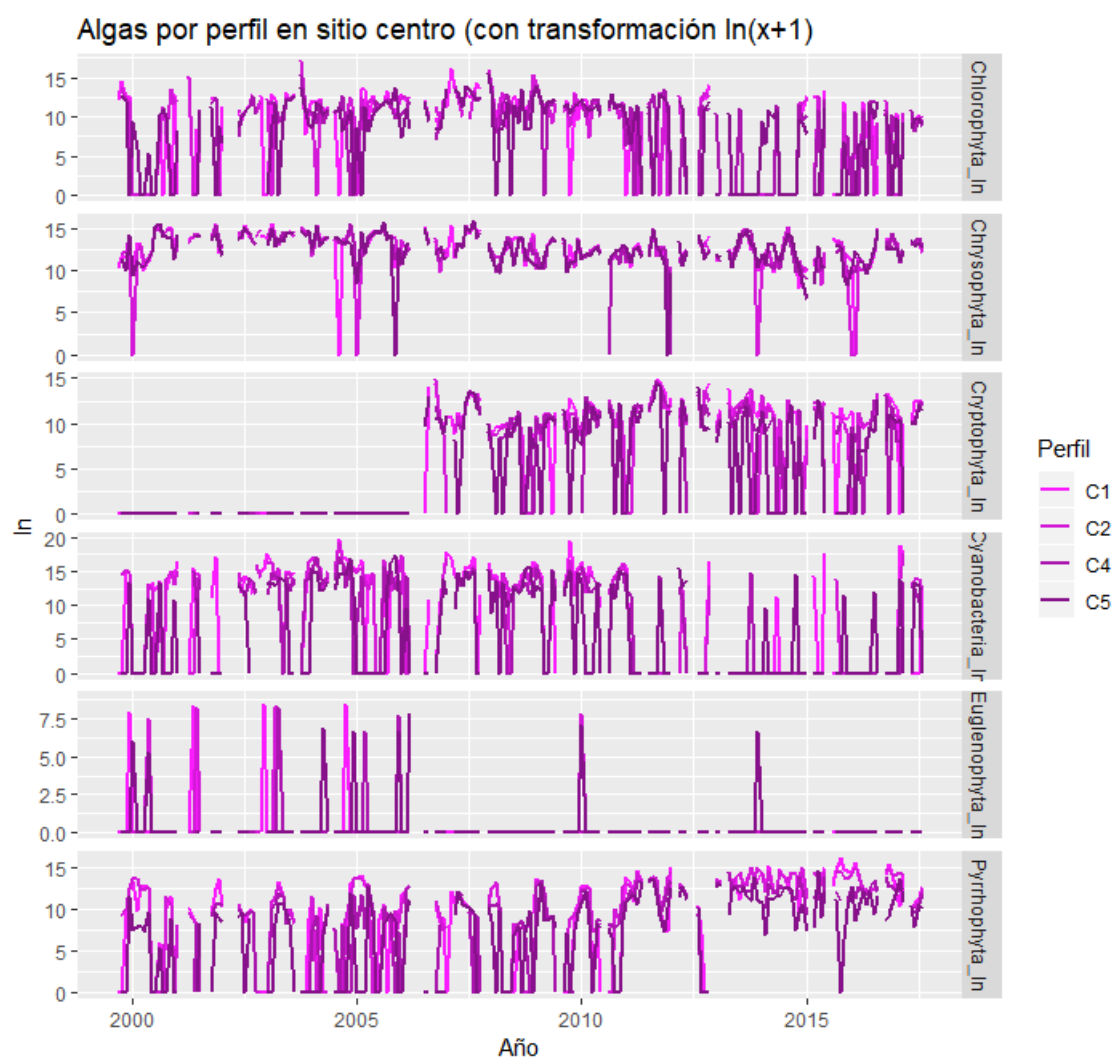


Fig. 99 Series de tiempo del logaritmo natural de las células por litro más uno para cada división de algas por perfil del sitio Centro

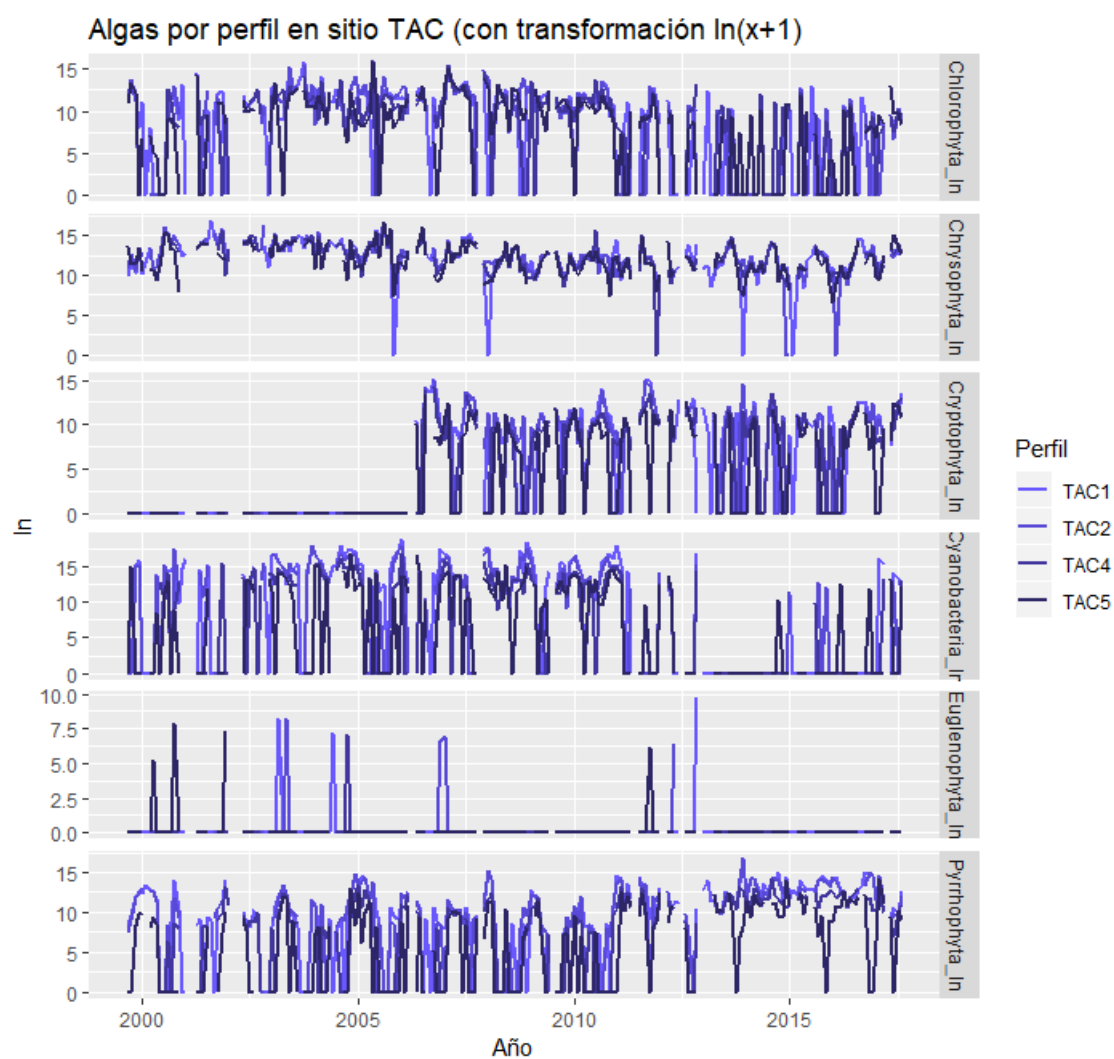


Fig. 100 Series de tiempo del logaritmo natural de las células por litro más uno para cada división de algas por perfil del sitio TAC

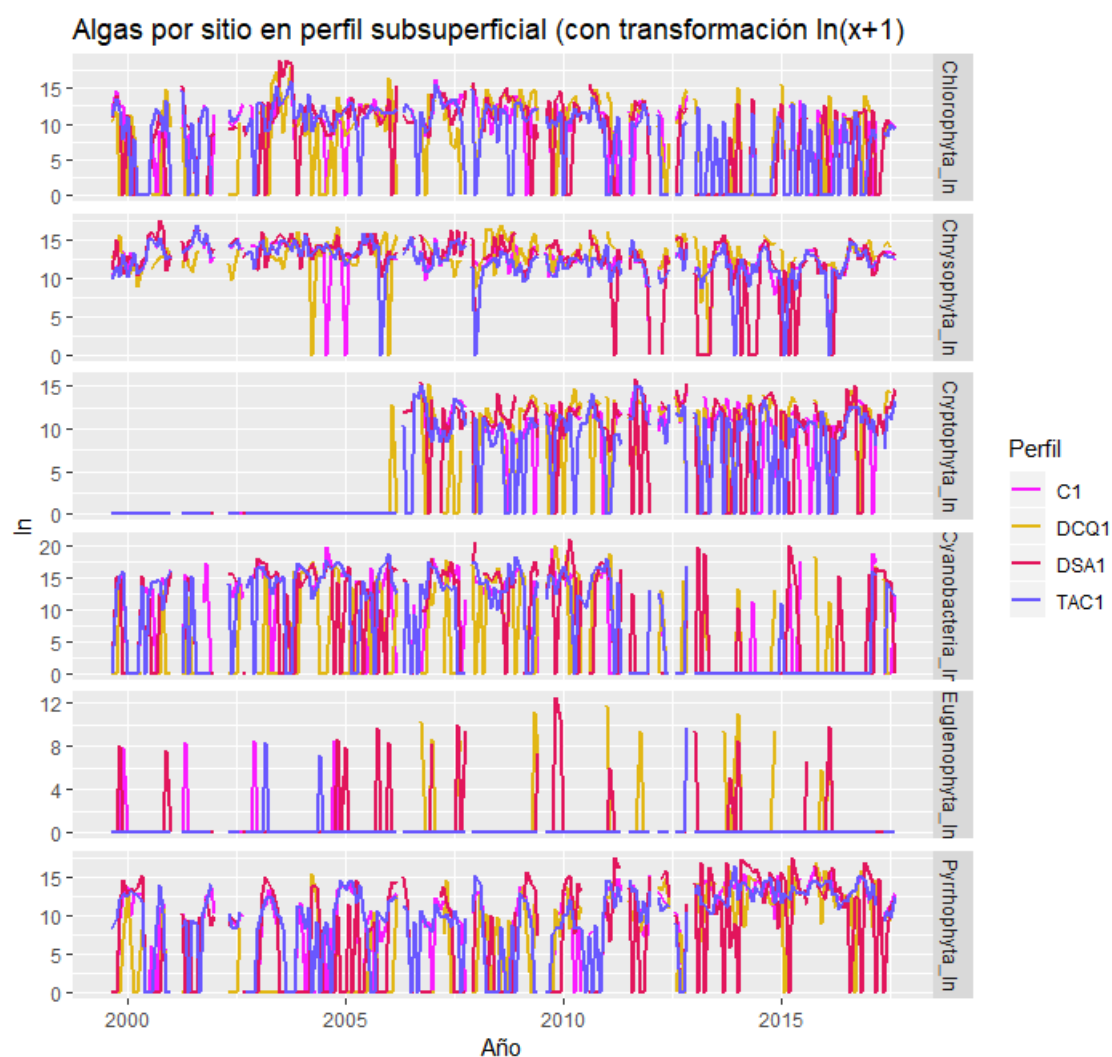


Fig. 101 Series de tiempo del logaritmo natural de las células por litro más uno para cada división de algas para cada sitio en perfil subsuperficial

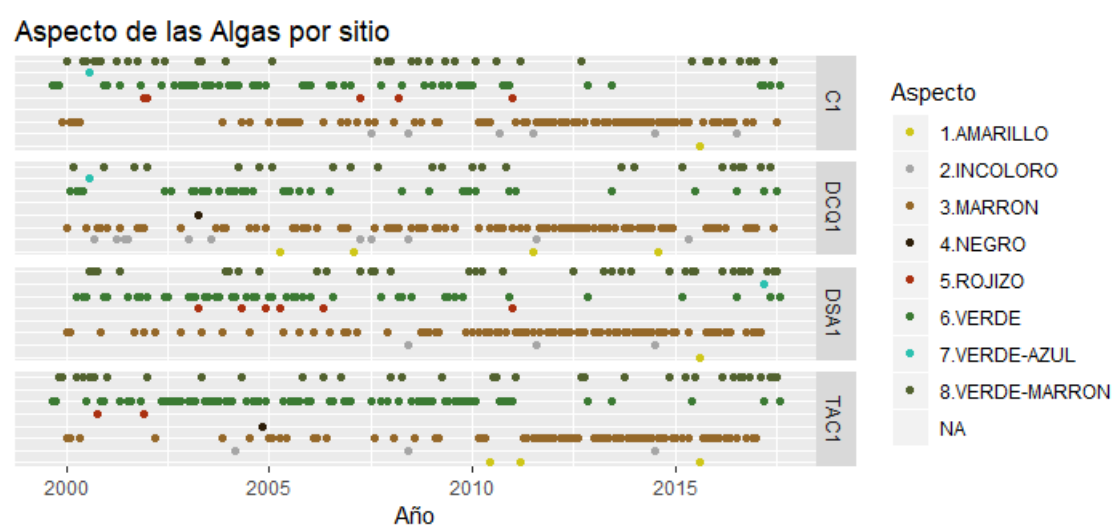


Fig. 102 Series de tiempo del aspecto de algas por sitio

Disposicion de las Algas por sitio

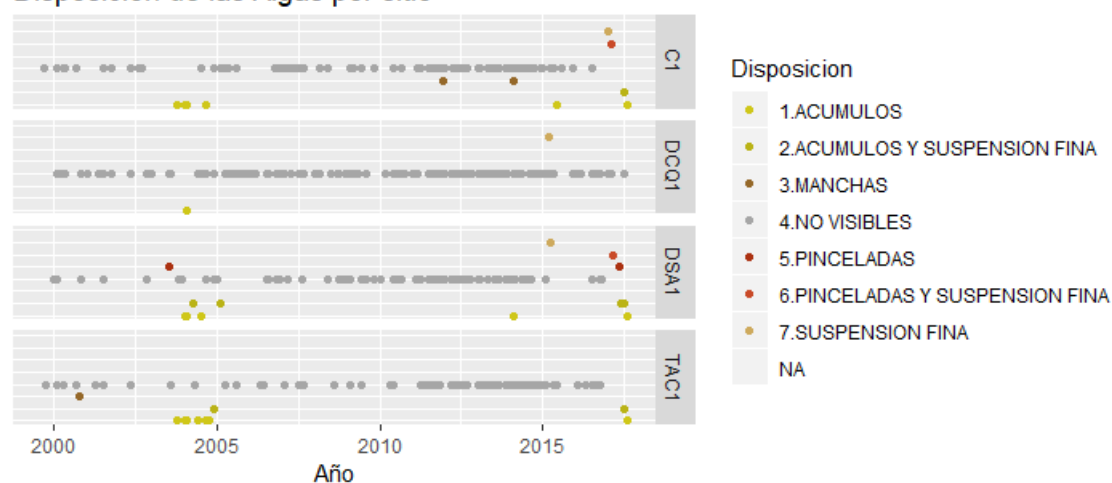


Fig. 103 Series de tiempo de la disposición de las algas por sitio

Nubosidad por sitio

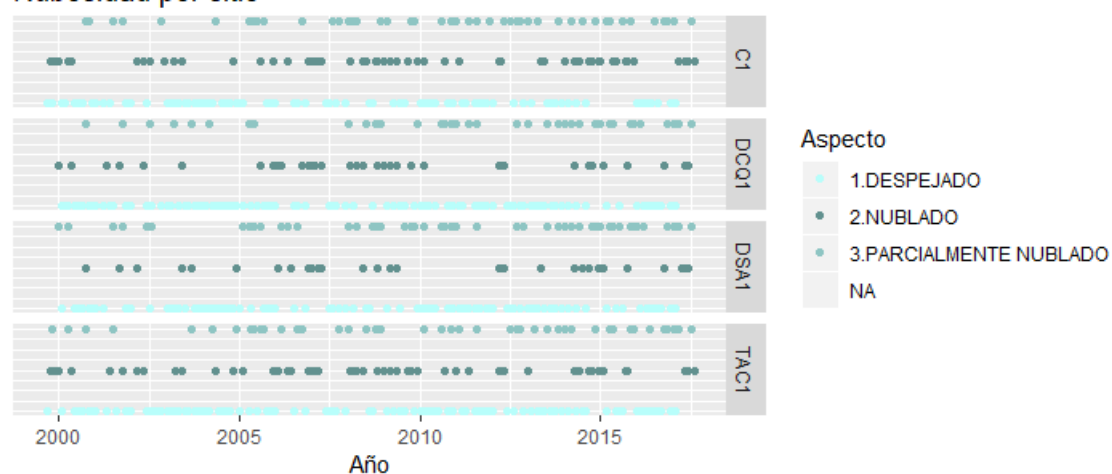


Fig. 104 Series de tiempo de nubosidad por sitio

Viento por sitio

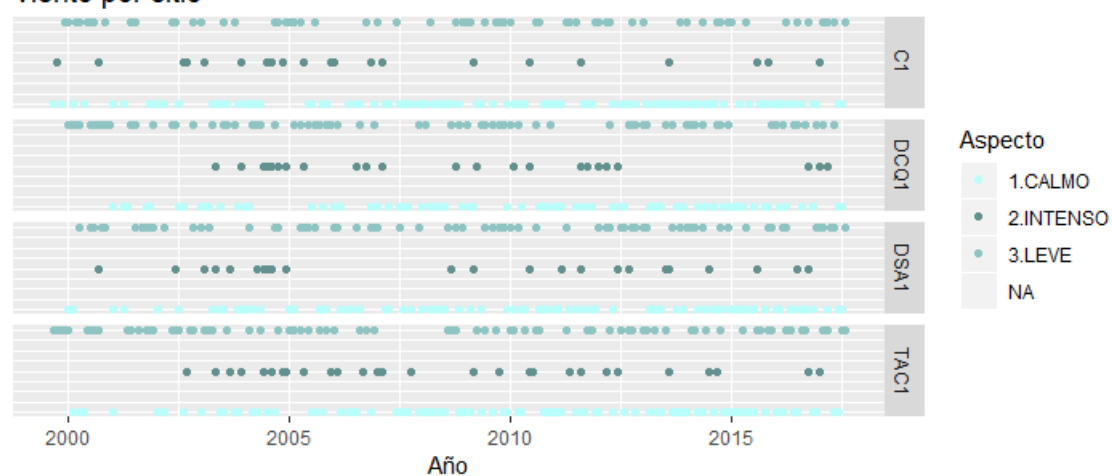


Fig. 105 Series de tiempo de viento por sitio

Condición de estratificación por sitio



Fig. 106 Series de tiempo de estratificación térmica por sitio

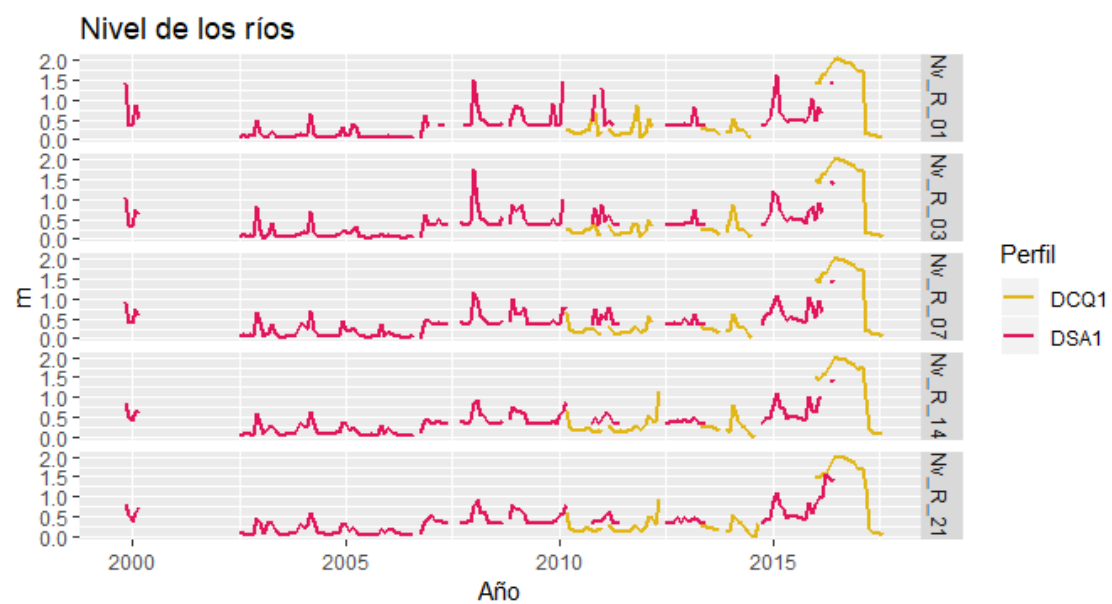


Fig. 107 Series de tiempo de los niveles, en distintos alcances, de los ríos

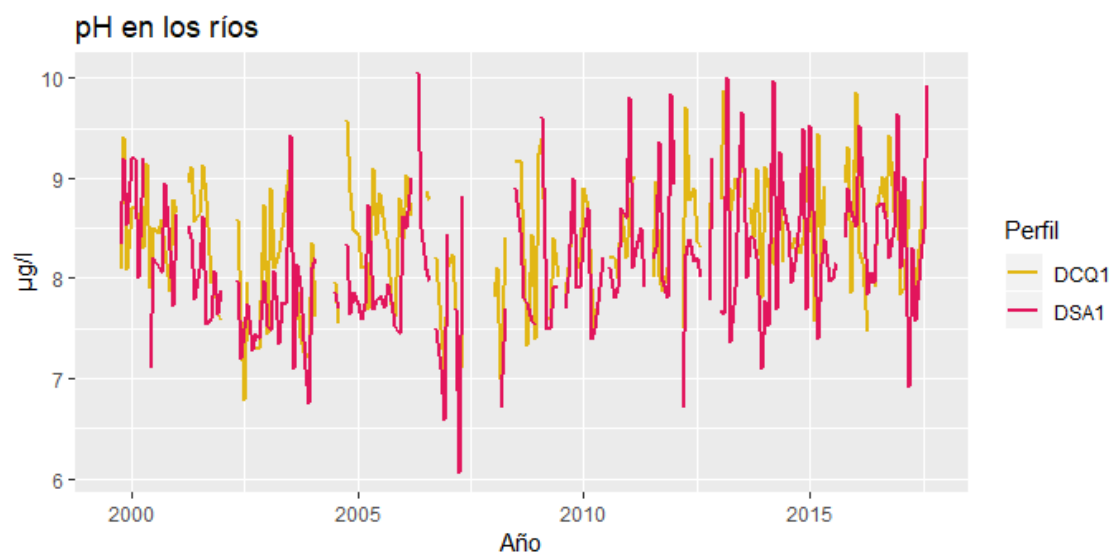


Fig. 108 Series de tiempo de pH de los ríos

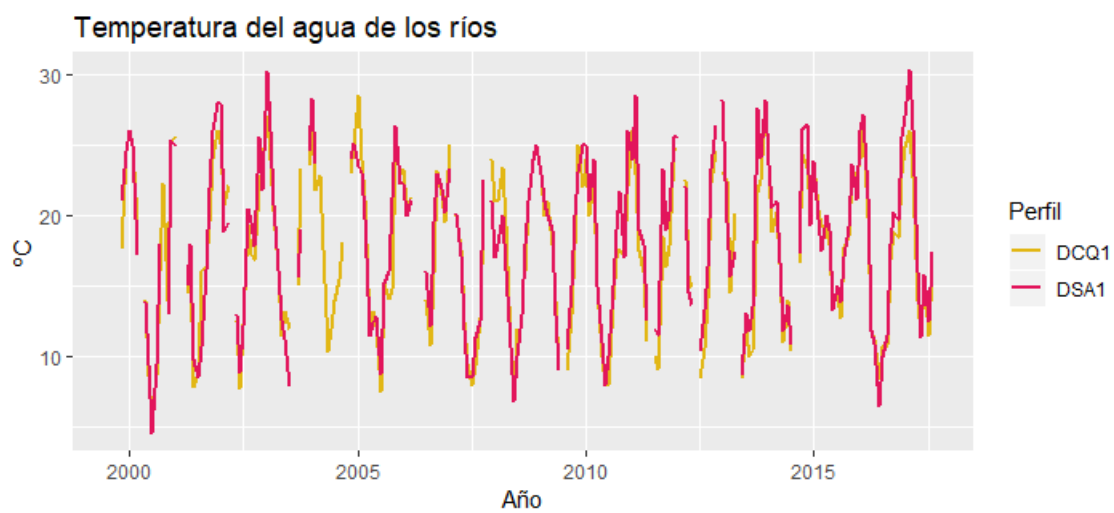


Fig. 109 Series de tiempo de la temperatura del agua de los ríos

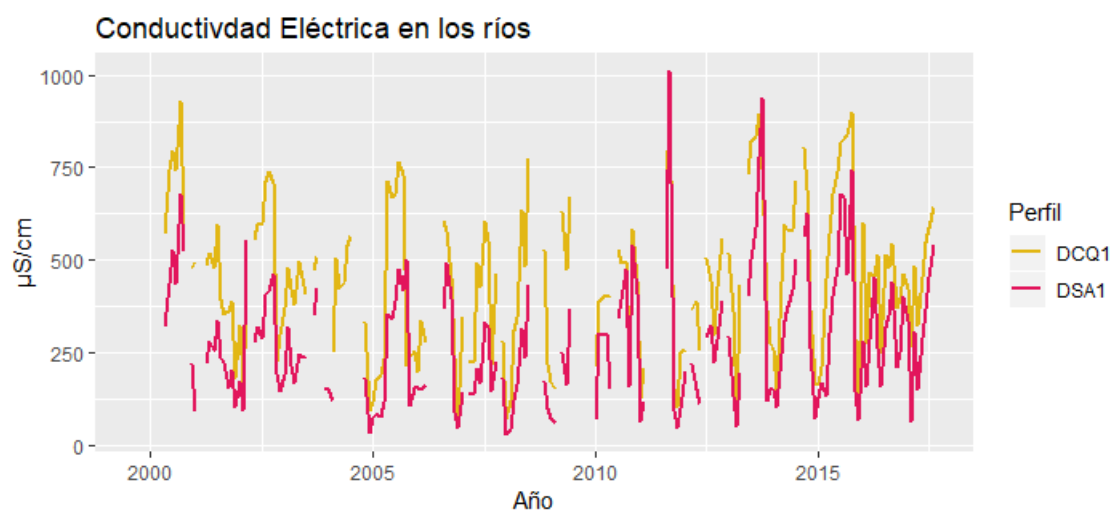


Fig. 110 Series de tiempo de conductividad eléctrica de los ríos

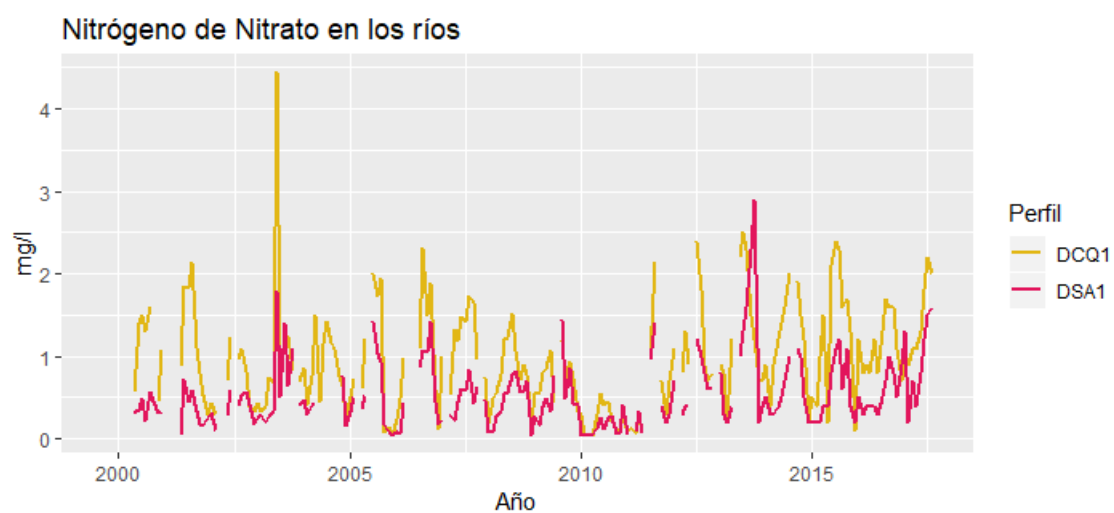


Fig. 111 Series de tiempo de nitrógeno de nitrato de los ríos

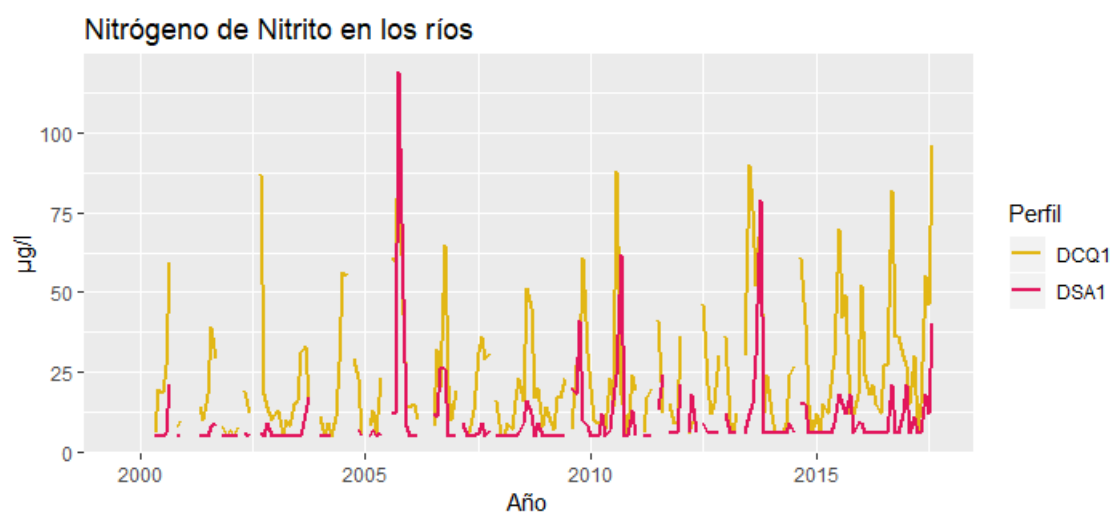


Fig. 112 Series de tiempo de nitrógeno de nitrito de los ríos

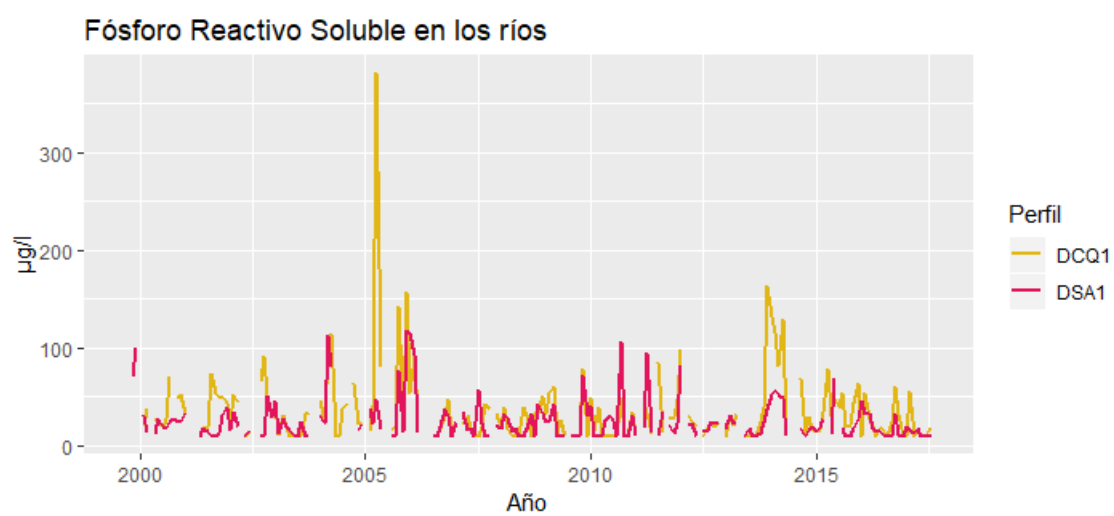


Fig. 113 Series de tiempo de fósforo reactivo soluble de los ríos

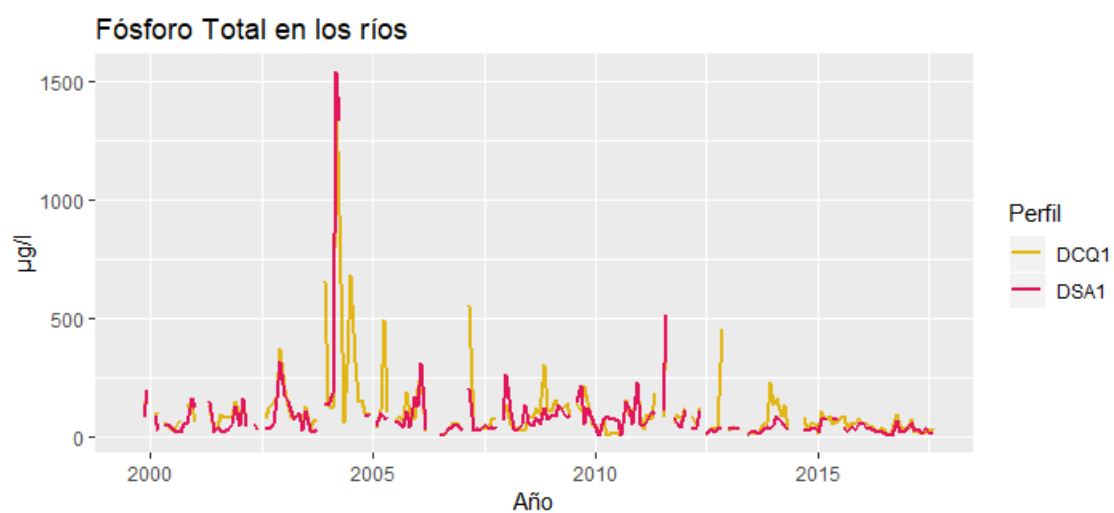


Fig. 114 Series de tiempo de fósforo total de los ríos

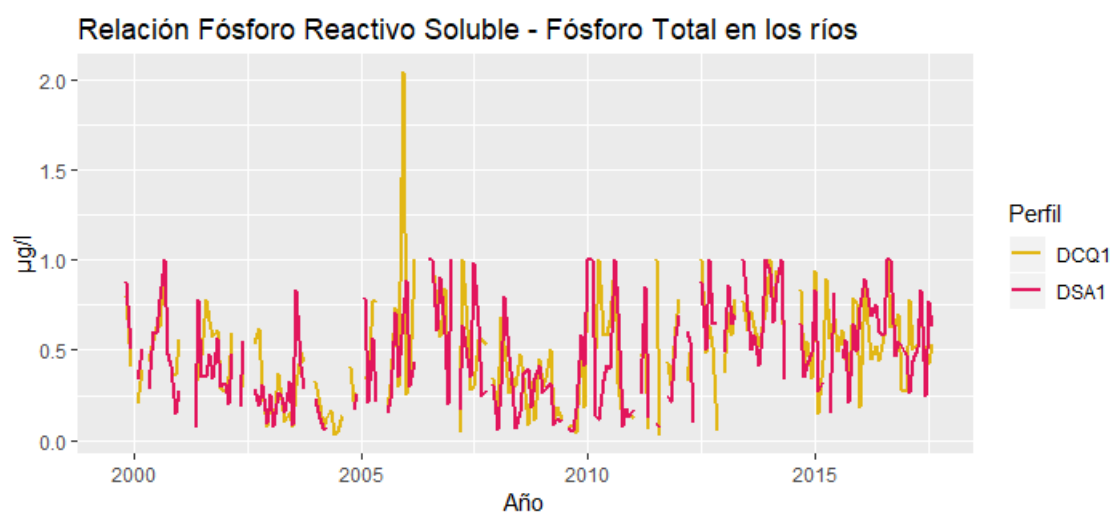


Fig. 115 Series de tiempo de la relación fósforo reactivo soluble sobre fósforo total de los ríos

IV Anexo: Software

Los siguientes programas y paquetes fueron empleados para el desarrollo del presente trabajo:

- **Base de datos y preprocesamiento de datos:** SQL Server 2014 Express.
- **Procesamiento general y cálculo:** R (R Core Team, 2018). Bibliotecas específicas:
 - Acceso a la base de datos: RODBC (Ripley & Lapsley, 2017).
 - Estructuras de datos: formattable (Ren & Russell, 2016), plyr (Wickham, 2011), reshape2 (Wickham, 2007), dplyr (Wickham et al., 2018), tidyverse (Wickham, 2017) y data.table (Dowle & Srinivasan, 2019).
 - Gráficos: ggplot2 (Wickham, 2016) y gridExtra (Aguie, 2017).
 - Cálculo de métricas: MLmetrics (Yan, 2016) y hydroGOF (Zambrano-Bigiarini, 2017).
 - Estadística: circular (Agostinelli & Lund, 2017).
 - Tratamiento espacial de datos: geosphere (Hijmans, 2017), sp (Pebesma & Bivand, 2005) y sphereplot (Robotham, 2013).
 - Análisis de series temporales: Kendall (McLeod, 2011), trend (Pohlert, 2018) y zoo (Zeileis & Grothendieck, 2005).
 - Regresión dinámica: fpp2 (Hyndman, 2018).
 - Modelo de espacio de estados multivariado: MARSS (Holmes et al., 2018c).
 - Modelado bayesiano: rjags (Plummer, 2018) y coda (Plummer et al., 2006).

5 Referencias

- Abunaka, K., Hosomi, M., & Murakami, A. (1997). Novel application of a back-propagation artificial neural network model formulated to predict algal bloom. *Water Science & Technology*, 36(5), 87-97.
- Agostinelli, C., & Lund, U. (Productor). (2017). *R package 'circular': Circular Statistics (version 0.4-93)*. Obtenido de <https://r-forge.r-project.org/projects/circular/>
- Ahrens, C. D., & Henson, R. (2017). *Meteorology Today*. Boston: Cengage.
- Amé, V., Ferral, A., & Solís, V. (2017). Eutrofización en el embalse San Roque y floraciones masivas de cianobacterias. Seguimiento por técnicas geoespaciales. *Unciencia*.
- American Public Health Association. (2017). *Standard method for examination of water and wastewater* (23 ed.). Washington: APHA, AWWA, WPCF.
- Andreoni, P., Rodriguez, M. I., Ruiz, M., Dasso, C., Vicario, L., Colladón, L., & Ruibal Conti, A. L. (2018). Gestión de Grandes Bases de Datos en el Área de los Recursos Hídricos: Aplicación de Técnicas de Exploración y Preprocesamiento de Datos (Minería de Datos). *4to Encuentro de Investigadores en Formación de Recursos Hídricos*.
- Auguie, B. (Productor). (2017). gridExtra: Miscellaneous Functions for "Grid" Graphics, R package version 2.3. Obtenido de <https://CRAN.R-project.org/package=gridExtra>
- Barbosa Chaves, F. Í., De Freitas Lima, P., Carrhá Leitão, R., Paulino, W. D., & Santaella, S. T. (2014). Influence of rainfall on the trophic status of a Brazilian semiarid reservoir. *Acta Scientiarum Biological Sciences*.
- Baroni, M. V., & Wunderlin, D. A. (2018). *The Suquia River Basin (Córdoba, Argentina) An Integrated Study on its Hydrology, Pollution, Effects on Native Biota and Models to Evaluate Changes in Water Quality; Capítulo: Integrating Data from Suquia River Basin: Chemometrics and Other Concepts*. Córdoba: Springer.
- Bechtel, Hansen, G., & Benito. (1992). Estudio de Factibilidad /Diseño Conceptual Para el Manejo de Descargas de Aguas Cloacales. *Inf. Int. N° 6, Inf. Prel. de est. de fact.*
- Bisgaard, S., & Kulahci, M. (2011). *Time Series Analysis and Forecasting by Example*. Singapur: Wiley.
- Bobbin, J., & Recknagel, F. (2001). Knowledge discovery for prediction and explanation of blue-green algal dynamics in lakes by evolutionary algorithms. *Ecological Modeling*, 146, 253-262.
- Bourrelly, P. (1972). *Les algues d'eau douce : initiation à la systématique*. Paris: Boubée.
- Brockwell, P. J., & Davis, R. A. (2002). *Introduction to Time Series and Forecasting*. Nueva York: Springer.
- Brockwell, P. J., & Davis, R. A. (2006). *Time Series: Theory and Methods*. Nueva York: Springer.

- Burriel Martí, F., Lucena Conde, F., Arribas Jimeno, S., & Hernández Méndez, J. (1985). *Química Analítica Cualitativa*. Madrid: Thomson Editores Spain.
- Carlson, R. E. (1976). A trophic state index for lakes. *American Society of Limnology and Oceanography*.
- Chen, Q., & Mynett, A. E. (2003). Integration of data mining techniques and heuristic knowledge in fuzzy logic modelling of eutrophication in Taihu Lake. *Ecological Modeling*, 16, 55-67.
- Cleveland, R. B., Cleveland, W. S., McRae, J. E., & Terpenning, I. (1990). STL: A Seasonal-Trend Decomposition Procedure Based on Loess. *Journal of Official Statistics. Statistics Sweden*, 3-73.
- Colladón, L. (2018). *Evaluación del desempeño de los sensores de nivel, a través del registro de picos de crecidas*. Retrieved from Villa Carlos Paz:
- Colladón, L., & Vélez, E. (2011). *Sistema de Monitoreo Automático de Ríos en las Sierras de Córdoba*. Artículo presentado en Memorias del Quinto Simposio Regional sobre Hidráulica de Ríos, Santiago del Estero.
- Dowle, M., & Srinivasan, A. (Productor). (2019). data.table: Extension of 'data.frame', R package version 1.12.0. Obtenido de <https://CRAN.R-project.org/package=data.table>
- Durbin, J., & Koopman, S. J. (2012). *Time Series Analysis by State Space Methods*. Oxford: Oxford University Press.
- Efron, B., & Tibshirani, R. (1994). *An Introduction to the Bootstrap*. New York: Chapman & Hall.
- Elcano, G., & Vicario, L. (2018). *Evaluación de sequías hidrológicas en los tributarios de la cuenca alta del río Suquia, provincia de Córdoba*. Retrieved from Córdoba:
- Francis, T. B., Wolkovich, E. M., Scheuerell, M. D., Katz, S. L., Holmes, E. E., & Hampton, S. E. (2014). Shifting Regimes and Changing Interactions in the Lake Washington, U.S.A., Plankton Community from 1962–1994.
- Gavilán, J. G. (1981). Study of Water Quality in the San Roque Reservoir. *Water Quality Bulletin, Environment*, 6(4), 136-158.
- Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Aki, V., & Rubin, D. B. (2014). *Bayesian Data Analysis*. Boca Raton: CRC Press.
- Gower, J. C. (1975). Generalized Procrustes Analysis. *Psychometrika*, 40(1).
- Han, J., & Kamber, M. (2006). *Data Mining: Concepts and Techniques*. San Francisco: Morgan Kaufmann.
- Hijmans, R. J. (Productor). (2017). geosphere: Spherical Trigonometry, R package version 1.5-7. Obtenido de <https://CRAN.R-project.org/package=geosphere>
- Hipel, K. W., & McLeod, A. I. (1994). *Times Series Modeling of Water Resources and Environmental Systems*. Amsterdam: Elsevier.

- Hirsch, R. M., Slack, J. R., & Smith, R. A. (1982). Techniques of Trend Analysis for Montly Water Quality Data. *Water Resources Research Vol 18 NO 1*, 107-121.
- Holmes, E. E. (2012). *Derivation of the EM algorithm for constrained and unconstrained MARSS models*. Retrieved from
- Holmes, E. E., Scheuerell, M. D., & Ward, E. J. (2019). *Applied Time Series Analysis for Fisheries and Environmental Sciences*. Seattle: NOAA Fisheries, Northwest Fisheries Science Center.
- Holmes, E. E., Ward, E. J., & Scheuerell, M. D. (2018a). Analysis of multivariate time-series using the MARSS package.
- Holmes, E. E., Ward, E. J., Scheuerell, M. D., & Wills, K. (2018b). Package 'MARSS'. CRAN. <https://cran.r-project.org/web/packages/MARSS/MARSS.pdf>
- Holmes, E. E., Ward, E. J., & Wills, K. (Productor). (2018c). MARSS: Multivariate Autoregressive State-Space Modeling, R package version 3.10.10.
- Horiba. (1991). *Water Quality Checker U-10, Instruction Manual*: Horiba Ltd.
- Hyndman, R. (Productor). (2019, Junio 15). RobJHyndman. *Statistical tests for variable selection*. Obtenido de <https://robjhyndman.com/hyndsight/tests2/>
- Hyndman, R. J. (Productor). (2018). fpp2: Data for "Forecasting: Principles and Practice" (2nd Edition), R package version 2.3. Obtenido de <https://CRAN.R-project.org/package=fpp2>
- Hyndman, R. J., & Athanasopoulos, G. (2018). *Forecasting: Principles and Practice*. Melbourne: OTexts.
- Hyndman, R. J., & Khandakar, Y. (2008). Automatic time series forecasting: The forecast package for R. *Journal of Statistical Software*, 27(21), 21–22.
- Instituto Nacional del Agua. (2014). *Informe periódico Nº 2: Actividad Permanente de Monitoreo del embalse San Roque y su cuenca*. Retrieved from Córdoba: https://www.ina.gov.ar/legacy/pdf/Cirsa_SanRoque_2014
- Instituto Nacional del Agua (Productor). (2019, Agosto 13). Actividad Permanente "Monitoreo del Embalse San Roque y Gestión de Información de Calidad de Aguas y Cianobacterias". *ina.gob.ar*. Obtenido de <http://ina.gob.ar/cirsa/index.php?seccion=1>
- Jorgensen, S. E., & Vollenweider, R. A. (1989). *Guidelines of Lake Management, Vol 1 Principles of Lake Management*. Otsu: UNEP.
- Kendall, M. G., & Stuart, A. (1961). *The advanced theory of statistics. Volume 2: Inference and relationship*. New York: Hafner publishing company.
- Komatsu, E., Fukushima, T., & Harasawa, H. (2007). A modeling approach to forecast the effect of long-term climate change on lake water quality. *Ecological Modeling - Elsevier (209)*, 351-366.

- Kratzer, C. R., & Brezonik, P., L. (1981). A Carlson-Type trophic state index for Nitrogen in Florida Lakes. *Water Resource Bulletin, American Water Resources Association*, 713-715.
- Liao, S.-H., Chu, P.-H., & Hsiao, P.-Y. (2012). Data mining techniques and applications – A decade review from 2000 to 2011. *Expert Systems with Applications*, 39.
- McLeod, A. I. (Productor). (2011). Kendall: Kendall rank correlation and Mann-Kendall trend test, R package version 2.2. Obtenido de <https://CRAN.R-project.org/package=Kendall>
- Osgood, R. A. (1982). Reply to discussion: "A Carlson-Type Trophic State Index for Nitrogen in Florida Lakes". *Water Resources Bulletin, American Water Resources Association*, 543-544.
- Pankratz, A. (1994). *Forecasting with Dynamic Regression Models*. Greencastle: Wiley.
- Pebesma, E. J., & Bivand, R. S. (Productor). (2005). Classes and methods for spatial data in R. R News 5 (2). Obtenido de <https://cran.r-project.org/doc/Rnews/>
- Plummer, M. (Productor). (2018). rjags: Bayesian Graphical Models using MCMC, R package version 4-8. Obtenido de <https://CRAN.R-project.org/package=rjags>
- Plummer, M., Best, N., Cowles, K., & Vines, K. (Productor). (2006). CODA: Convergence Diagnosis and Output Analysis for MCMC. Obtenido de <https://journal.r-project.org/archive/>
- Pohlert, T. (Productor). (2018). trend: Non-Parametric Trend Tests and Change-Point Detection, R package version 1.1.1. Obtenido de <https://CRAN.R-project.org/package=trend>
- R Core Team (Productor). (2018). R: A language and environment for statistical computing, R version 3.5.2. Obtenido de <https://www.R-project.org/>
- Recknagel, F., French, M., Harkonen, P., & Yabunaka, K. (1997). Artificial neural network approach for modelling and prediction of algal blooms. *Ecological Modeling*, 96, 11-28.
- Ren, K., & Russell, K. (Productor). (2016). formattable: Create 'Formattable' Data Structures, R package version 0.2.0.1. Obtenido de <https://CRAN.R-project.org/package=formattable>
- Ripley, B., & Lapsley, M. (Productor). (2017). RODBC: ODBC Database Access, R package version 1.3-15. Obtenido de <https://CRAN.R-project.org/package=RODBC>
- Robotham, A. (Productor). (2013). sphereplot: Spherical plotting, R package version 1.5. Obtenido de <https://CRAN.R-project.org/package=sphereplot>
- Rodríguez, M. I., Cossavella, A., Oroná, C., Larrossa, N., Avena, M., Rodríguez, A., Del Olmo, S., Bertucci, C., Muñoz, A., Castelló, E., Bazán, R., & Martínez, M. (2000). *Estudios Preliminares de la Calidad de Agua y Sedimentos del Embalse San Roque relacionados al proceso de Eutroficación*. Artículo presentado en XVIII Congreso Nacional del Agua, Santiago del Estero.

- Roldán Pérez, G., & Ramírez Restrepo, J. J. (2008). *Fundamentos de Limnología Neotropical*. Medellín: Editorial Universidad de Antioquía.
- Roth, G. D. (2003). *Meteorología. Formaciones nubosas y otros fenómenos meteorológicos. Situaciones meteorológicas generales. Pronósticos del tiempo*. Barcelona: Ediciones Omega.
- Samu, A., Zoltán, C., & Bárány-Kevey, I. (2013). Influence of meteorological variables to water quality in five lakes over the Aggtelek (Hungary) and Slovak Karst regions - a case study. *Acta Carsologica* 42/1, 121-133.
- Scagel, R. F., Bandoni, R. J., Maze, J. R., E., R., Schofield, W. B., & Stein, J. R. (1991). *Plantas No Vasculares*. Barcelona: Omega.
- Secretaría de Turismo de Villa Carlos Paz. (2015). *Informe de Demanda Turística, período comprendido: 1 al 15 de Enero de 2015*. Retrieved from Villa Carlos Paz: <https://www.villacarlosspaz.gov.ar>
- Shoba, G. (2014). Water Quality Prediction Using Data Mining techniques: A Survey. *International Journal Of Engineering And Computer Science*, 3.
- Shumway, R. H., & Stoffer, D. S. (2017). *Time Series Analysis and Its Applications With R Examples*. Nueva York: Springer.
- Sigee, D. C. (2005). *Freshwater Microbiology*. Chichester: John Wiley & Sons Ltd.
- Spiegelhalter, D. J., Best, N. G., Carlin, B. P., & Van der Linde, A. (2002). Bayesian measures of modelo complexity and fit. *Journal of the Royal Statistical Societey Series*, 583-639.
- Tizro, A. T., Ghashaghaie, M., Georgiou, P., & Voudouris, K. (2014). Time series analysis of water quality parameters. *Journal of Applied Research in Water and Wastewater*, 43-52.
- Vaisala. (2012). Vaisala HUMICAP® Sensor for Measuring Relative Humidity. *Vaisala*. <https://www.vaisala.com/sites/default/files/documents/HUMICAP-Technology-description-B210781EN-C.pdf>
- Water on the web (Productor). (2017, Marzo 28). Dissolved Oxygen. *Water on the web*. Obtenido de <http://www.waterontheweb.org/under/waterquality/oxygen.html>
- Weiyang, Z., In Chio, L., Yijun, K., Wai Kin, U., & Kai Meng, M. (2013). Eutrophication analyses and principle component regression for two subtropical storage reservoirs in Macau. *Desalination and Water Treatment*, 51:37-39.
- Wetzel, R. G., & Likens, G. E. (2000). *Limnological Analyses*. Springer: Nueva York.
- Wickham, H. (Productor). (2007). Reshaping Data with the reshape Package, *Journal of Statistical Software*, 21(12), 1-20. Obtenido de <http://www.jstatsoft.org/v21/i12/>
- Wickham, H. (Productor). (2011). The Split-Apply-Combine Strategy for Data Analysis, *Journal of Statistical Software* , 40(1), 1-29. Obtenido de <http://www.jstatsoft.org/v40/i01/>

- Wickham, H. (Productor). (2016). ggplot2: Elegant Graphics for Data Analysis. Obtenido de <http://ggplot2.org>
- Wickham, H. (Productor). (2017). tidyverse: Easily Install and Load the 'Tidyverse', R package version 1.2.1. Obtenido de <https://CRAN.R-project.org/package=tidyverse>
- Wickham, H., François, R., Henry, L., & Müller, K. (Productor). (2018). dplyr: A Grammar of Data Manipulation, R package version 0.7.8. Obtenido de <https://CRAN.R-project.org/package=dplyr>
- Yan, Y. (Productor). (2016). MLmetrics: Machine Learning Evaluation Metrics, R package version 1.1.1. Obtenido de <https://CRAN.R-project.org/package=MLmetrics>
- Yang, K., Yua, Z., Luoa, Y., Yang, Y., Zhao, L., & Zhoud, X. (2018). Spatial and temporal variations in the relationship between lake water surface temperatures and water quality - A case study of Dianchi Lake. *Science of the Total Environment*, 859-871.
- YSI. (2009). *YSI 556 Multiparameter System*: YSI Incorporated.
- Zambrano-Bigiarini, M. (Productor). (2017). hydroGOF: Goodness-of-fit functions for comparison of simulated and observed hydrological time series. R package version 0.3-10. Obtenido de <http://hzambran.github.io/hydroGOF/>. DOI:10.5281/zenodo.840087.
- Zeileis, A., & Grothendieck, G. (Productor). (2005). zoo: S3 Infrastructure for Regular and Irregular Time Series, *Journal of Statistical Software*, 14(6), 1-27. Obtenido de doi:10.18637/jss.v014.i06
- Zhang, L., Zhang, G. X., & Li, R. R. (2016). Water Quality Analysis and Prediction Using Hybrid Time Series and Neural Network Models. *J. Agr. Sci. Tech.*, 975-983.