UNIVERSIDAD DE LA REPÚBLICA
FACULTAD DE INGENIERÍA

Tesis para optar al Título de Magíster en Ingeniería Matemática

BASES PARA UN SISTEMA DE PREDICCIÓN DE CAUDALES DE APORTE A RINCÓN DEL BONETE Y SALTO GRANDE

Autor: STEFANIE TALENTO

Director de Tesis: Dr. Ing. RAFAEL TERRA
Co-director de Tesis: Dr. MARCO SCAVINO

Montevideo, Uruguay
2011

AGRADECIMIENTOS

Primero que nada quiero agradecer a los tutores de esta tesis, Rafael Terra y Marco Scavino, por todo el trabajo y tiempo dedicados. Considero que fue un placer haber trabajado con ambos.

También quisiera agradecer a Gabriel Cazes-Boezio y Álvaro Díaz por ayudar en varias instancias del trabajo, incluida la redacción de la propuesta de tesis.

Por otro lado, quiero agradecer a los integrantes de la SCAPA de Ingeniería Matemática por el apoyo en todas las etapas de mis estudios de posgrado.

Finalmente agradezco a mi familia, en especial a Fabián, por el aliento y la paciencia continuos.

Tabla de Contenido

RESUMEN I
1. INTRODUCCIÓN 1
2. DATOS Y DESCRIPCIÓN DEL MODELO 8
2.1. Caudales 8
2.2. Datos atmosféricos 8
2.3. Datos oceánicos 9
2.4. Consideraciones respecto al período de estudio 10
2.5. Simulaciones con un modelo de circulación general de la atmósfera 11
3. ANÁLISIS PRELIMINAR DE DATOS 13
3.1. Caudales 13
3.2. Índice Niño 3.4 17
3.3. Climatologías observadas y simuladas 21
4. DETERMINACIÓN Y ANÁLISIS PRELIMINAR DE PREDICTORES 24
4.1. Circulación atmosférica regional 24
4.1.1. Aplicación del análisis de componentes principales a estudios geofísicos 25
4.1.2. Aplicación del análisis de Componentes Principales a la circulación atmosférica regional 25
4.2. Fenómeno El Niño Oscilación Sur 31
4.3. Caudales antecedentes 34
4.4. Resumen 35
5. MODELOS ESTADÍSTICOS DE PREDICCIÓN 38
5.1. Error de predicción 39
5.2. Regresión Lineal 41
5.3. Inestabilidad de estimaciones en regresión lineal múltiple 44
5.3.1. Selección de variables en el contexto de regresión lineal múltiple 45
5.3.2. Regresión por mínimos cuadrados parciales 46
5.4. Árboles de clasificación o regresión 47
5.5. Redes neuronales artificiales 50
5.6. Predicción mediante clustering 54
6. AJUSTE DE MODELOS ESTADÍSTICOS 56
6.1. Predicción con variables atmosféricas, oceánicas y caudales precedentes 58
6.1.1. Regresión lineal múltiple, selección de variables y PLSR 58
6.1.2. Árboles de regresión 60
6.1.3. Redes Neuronales 60
6.1.4. Clustering 60
6.1.5. Resultados 61
6.2. Predicción con variables atmosféricas y oceánicas 67
6.3. Predicción con variables oceánicas y caudales precedentes 71
7. RESULTADOS DE SIMULACIONES CON MODELO DE CIRCULACIÓN GENERAL DE LA ATMÓSFERA 74
7.1. Consideraciones generales sobre evaluación de pronósticos 74
7.2. Ajuste de modelos utilizando únicamente variables potencialmente predictibles 80
8. RESUMEN DE RESULTADOS Y CONCLUSIONES 87
Anexo A: Análisis de componentes principales (CPs) 93
Anexo B: Índice de abreviaciones 95
BIBLIOGRAFÍA 97

RESUMEN

En este trabajo se presentan las bases para el diseño de sistemas de predicción de caudales de aporte a las represas hidroeléctricas de Rincón del Bonete y Salto Grande. La predicción se realiza de forma independiente para cada embalse y para cada mes del año, siguiendo las metodologías de downscaling híbrido (cuando se utilizan predictores atmosféricos) o modelo orientado puramente por datos (cuando no se utilizan predictores atmosféricos).

A partir del análisis de la circulación atmosférica regional, índices asociados al fenómeno El Niño Oscilación Sur e información relacionada con caudales antecedentes se determina un conjunto inicial de variables predictoras. Luego, bajo la hipótesis de variables predictoras conocidas, se ajustan varios modelos estadísticos de regresión que relacionan a las mismas con el caudal, entre ellos: modelo lineal, modelo lineal acoplado con selección de variables, regresión por mínimos cuadrados parciales, árboles de regresión, redes neuronales y técnicas de clustering. En general, el modelo que presenta los mejores resultados de habilidad predictiva (estimada a través del error cross validation leave-one-out) es el modelo de regresión lineal acoplado con selección hacia atrás de variables.

Se encuentra que tanto en Rincón del Bonete como en Salto Grande los caudales de aporte medios mensuales son predictibles en todos los meses del año, exceptuando el caudal de aporte a Rincón del Bonete en agosto. Si bien para Rincón del Bonete no se distingue claramente un período de elevada predictibilidad, para Salto Grande las temporadas de marzo a mayo y de octubre a diciembre destacan como robustas en este sentido.

Los esquemas de predicción desarrollados presentan, en general, habilidad predictiva superior a la de pronosticar la media histórica, aún en situaciones de antecedencia del pronóstico que no permiten contar con los caudales precedentes seleccionados como predictores (antecedencias superiores a dos meses).

A pesar de que todo el trabajo fue realizado bajo la hipótesis de predictores conocidos, la utilización de un modelo de circulación general de la atmósfera (el de la Universidad de California, Los Ángeles) permitió evaluar el desempeño de los modelos desarrollados restringiendo las variables predictoras atmosféricas a aquellas que dicho modelo indica podrían ser predictibles, lo cual constituye una situación más cercana a la que debe ser enfrentada en modo operacional. Aún cuando el conjunto de predictores se restringe únicamente a aquellos potencialmente predictibles, los modelos desarrollados muestran habilidad predictiva superior a la de pronosticar la media histórica en ambos embalses en la mayoría de los meses, incluso bajo situaciones de antecedencia superiores a los dos meses. Aunque estos resultados deben considerarse cotas superiores de la habilidad predictiva en modo operacional (pues, de nuevo, se supusieron predictores conocidos) los resultados son alentadores.

1. INTRODUCCIÓN

El pronóstico de caudales es de vital importancia en varios aspectos y escalas temporales. Mientras que el pronóstico de caudales horarios o diarios es fundamental para, por ejemplo, alertas de inundaciones, el pronóstico de medias mensuales, estacionales o anuales es determinante para la operación de embalses y planificación de la disponibilidad hídrica y, en particular, hidro-energética futura.

En Uruguay, la mayor fuente de energía eléctrica es la de origen hidráulico, siendo las represas hidroeléctricas de Gabriel Terra (río Negro), en adelante llamada Rincón del Bonete, y Salto Grande (río Uruguay) las de mayor capacidad de embalse. El diseño de esquemas de predicción de caudales en escalas mensuales o estacionales y su eficiente incorporación al proceso de toma de decisiones en el sistema eléctrico pueden significar importantes beneficios económicos.

Para definir una estrategia de predicción de fenómenos físicos se deben considerar tres escalas de tiempo. Primero, la escala en la que se promedia la magnitud que se desea predecir. Por ejemplo, los enfoques de predicción suelen ser distintos según se deseen obtener predicciones de los promedios horarios, diarios, semanales, mensuales, estacionales o anuales. Segundo, la escala determinada por el tiempo de antelación con el que se va a realizar el pronóstico. Esta escala puede ir desde minutos hasta años de antelación. Tercero, la escala de los fenómenos físicos involucrados, que sustentan el grado de predictibilidad del sistema.

En el caso particular de caudales los modelos de predicción suelen ser categorizados en clases según sean orientados por procesos (también conocidos como modelos dinámicos) u orientados por datos (también conocidos como modelos estadísticos) (Wang, 2006). Los modelos orientados por procesos modelan los procesos físicos involucrados. Los modelos orientados por datos se basan en la identificación de una vinculación estadística directa entre el predictando (caudal) y ciertos predictores externos que se consideren adecuados; básicamente son modelos de caja negra que no consideran explícitamente los procesos físicos subyacentes. Otros modelos pueden mezclar conceptos de los anteriores y simular con base física algunos procesos y realizar estimación estadística de otros. La selección de la clase de esquema a utilizar en la predicción está fuertemente condicionada por las escalas temporales que plantee el problema particular a resolver y por el grado de entendimiento de los procesos físicos que fundamentan la predictibilidad.

Dependiendo de la cuenca, los caudales pueden estar mayormente influenciados por la ocurrencia de precipitaciones o por procesos tales como deshielo, aportes subterráneos u otros. Esto genera diferencias en las estrategias a seguir y, también, en el grado de predictibilidad potencial. En casos como el de Uruguay la predicción de caudales está notablemente atada a la predicción de precipitaciones.

La atmósfera es un sistema caótico: si dos realizaciones son iniciadas a partir de condiciones apenas diferentes, las dos soluciones evolucionarán hasta que, eventualmente, divergirán (Lorenz, 1963, 1969). Aunque se tuviera un modelo matemático perfecto de la física que gobierna la atmósfera, debido a que es imposible obtener una observación perfecta del estado atmosférico en cierto instante (que defina exactamente el estado inicial) es imposible determinar el estado del sistema en un momento futuro arbitrariamente lejano. En conclusión, la naturaleza caótica de la atmósfera hace que sea imposible realizar pronósticos precisos de precipitación con antelaciones mayores a su umbral de predictibilidad determinístico, el cual es del entorno de unos pocos días (Lorenz, 1982).

Surge, entonces, la pregunta ¿serán posibles pronósticos de caudal, precipitación u otras variables hidrometeorológicas, con antelaciones mayores a unos pocos días? La respuesta es que, bajo ciertas condiciones, sí. Aunque para predicciones atmosféricas a corto plazo las observaciones precisas del estado inicial del sistema son cruciales, para predicciones mensuales o estacionales su influencia, a pesar de ser detectable, se ve claramente debilitada. Si bien el estado atmosférico preciso no es predictible más allá de días, promedios mensuales o estacionales de algunas variables hidrometeorológicas sí son potencialmente predictibles, en muchas regiones del planeta. La principal fuente de esta predictibilidad reside en los componentes del sistema climático de variación más lenta.

Muchas de las condiciones de borde que fuerzan a la atmósfera tienen la característica de evolucionar lentamente y, por tanto, poder ejercer su influencia durante largos períodos. La principal condición de borde que afecta a la atmósfera es la temperatura de superficie de mar (TSM), en especial, la TSM en las regiones tropicales del planeta. La TSM es de evolución particularmente lenta debido a la gran capacidad calorífica de los océanos y su dinámica inherentemente más lenta, en comparación con la de la atmósfera. Otras condiciones de borde importantes, aunque usualmente menos influyentes, son: la humedad del suelo, el cubrimiento de nieve, el cubrimiento de hielo en los océanos y la vegetación. La lentitud de la variación de estas condiciones de borde implica que anomalías importantes puedan extenderse por varios meses ocasionando cambios sustanciales en la probabilidad de ocurrencia de ciertos fenómenos climáticos (Goddard et al., 2001).

Si las anomalías de las condiciones de borde que fuerzan la atmósfera son predictibles entonces ciertos aspectos del clima, dinámicamente acoplados con estas anomalías, también podrían serlo. Es así que pronósticos del comportamiento estadístico del caudal a mediano y largo plazo podrían llegar a ser posibles en escalas mensuales o estacionales en ciertas regiones del mundo.

En conclusión, el pronóstico mensual o estacional de caudales puede ser posible, en sentido probabilístico, y, como ya se indicó, existen tres estrategias para su realización: modelos orientados puramente por datos, orientados puramente por procesos o esquemas híbridos que combinan aspectos de los dos anteriores.

En el caso de modelos orientados puramente por datos se construye un modelo estadístico en el que se analiza información histórica y se identifican relaciones entre ciertas variables predictoras y la variable a predecir (predictando). Implícito en cualquier predicción estadística está que los valores anteriores, actuales o futuros de las variables predictoras puedan ser utilizados para predecir el estado futuro o evolución de la variable predictando, basándose únicamente en las relaciones matemáticas entre ellos (Goddard et al., 2001). Este tipo de enfoque de predicción podría presentar dificultades en situaciones en las que no pueda asumirse la estacionareidad del clima, o ante la ocurrencia de eventos no acontecidos previamente en los registros históricos, es decir situaciones en las que el clima pasado podría no ser representativo de la variabilidad futura. Por otro lado, este enfoque tiene la ventaja de que puede representar procesos físicos arbitrariamente complejos e incluso, en ciertas situaciones y a pesar de que no los modela directamente, el análisis de su estructura y parámetros puede llegar a proveer información sobre los citados procesos físicos. Dado que, día a día, la calidad y cantidad de datos disponibles se incrementa y que las técnicas y capacidad computacionales para procesarlos acompañan dicho crecimiento, este tipo de técnica se ha vuelto extremadamente popular. Dentro de esta categoría de esquemas de predicción se encuentran los modelos de series temporales, en particular los modelos periódicos auto-regresivos, los cuales no incluyen ningún tipo de información climática observada o predicha, por lo que son de

poca utilidad en situaciones en las que la persistencia temporal de los caudales es débil y los forzantes climáticos de gran escala son importantes. Otro ejemplo son las predicciones estadísticas en las que sí se incorpora información climática la cual, típicamente, incluye algún indicador del estado de la TSM (ver, por ejemplo, Lima y Lall (2010), Soukup et al. (2009), Westra y Sharma (2009)).

Para la predicción puramente dinámica de caudales se construyen esquemas que modelan físicamente todo el proceso de causalidades en el sistema climático. Actualmente estos esquemas se conciben en dos pasos. En el primer paso se generan predicciones del estado atmosférico futuro. En el segundo paso estas predicciones son ingresadas, como insumo, a algún modelo hidrológico de macro-escala el cual relaciona las variables atmosféricas con el caudal.

Para la predicción del estado atmosférico futuro se modela numéricamente y con base física la evolución del sistema acoplado atmósfera-océano. Actualmente los modelos numéricos utilizados para este tipo de predicciones trabajan en una o en dos etapas y son denominados tier-1 o tier-2, respectivamente. En los sistemas tier-1 la predicción se realiza con un modelo de circulación general acoplado atmósfera-océano (MCGAO). Estos modelos simulan la evolución tanto de la atmósfera como de los océanos teniendo en cuenta los fenómenos físicos de interacción entre ellos. Ambos, la atmósfera y los océanos, pueden evolucionar libremente e influenciarse mutuamente. Los MCGAO consisten en un modelo de circulación general de la atmósfera (MCGA) y en uno de circulación general del océano (MCGO) que se acoplan intercambiando información. Por su parte, en los sistemas tier-2 la predicción se realiza de manera desacoplada entre la atmósfera y el océano. Las condiciones de borde (TSM) son predichas primero y luego utilizadas para forzar a la atmósfera, a través de un MCGA. Los sistemas tier-2 se basan en la noción de que anomalías estacionales resultan, básicamente, de cambios en la TSM y que anomalías atmosféricas estacionales podrían ser predictibles prediciendo la TSM. Las predicciones del estado de la TSM pueden ser obtenidas mediante técnicas dinámicas, estadísticas o simplemente a través de la persistencia de las anomalías más recientemente observadas, superpuestas al ciclo anual climatológico.

En teoría, el enfoque tier-1 es superior al tier-2 debido a que el primero tiene el potencial de representar realísticamente la interacción atmósfera-océano. En el enfoque tier-2 al océano no se le permite responder a la influencia de la atmósfera, lo que lleva a flujos de calor no realistas en varias regiones del planeta. Sin embargo, en la generación actual de MCGAO las TSM suelen alejarse de valores realistas a medida que la simulación progresa forzando, a su vez, patrones no realistas en la componente atmosférica (Goddard et al., 2001). Consecuentemente, en la actualidad, el enfoque tier-2 es más utilizado que el tier-1.

Para la predicción dinámica de caudales, las predicciones atmosféricas obtenidas por sistemas tier-1 o tier-2 deben ser, finalmente, ingresadas a modelos hidrológicos. Tanto los sistemas tier-1 como tier-2 generan sus predicciones en escalas espaciales de cientos de kilómetros, pero para poder ser utilizadas por los modelos hidrológicos son necesarias mayores resoluciones. Estas mayores resoluciones se obtienen a través de una reducción de escala (downscaling), también, dinámico. El downscaling dinámico se materializa ya sea mejorando la resolución de los modelos de circulación general o anidando un modelo de mayor resolución únicamente en la región de interés. En ocasiones también son necesarias reducciones de las escalas temporales. Finalmente, a partir de este modelo hidrológico se obtienen las predicciones de caudal. Un ejemplo de esta metodología puede encontrarse en Wood et al. (2002).

La última estrategia para predicción de caudales mensuales o estacionales es la usualmente

denominada downscaling híbrido (Goddard et al., 2001). En este caso el proceso de pronóstico consiste de 2 etapas en las que se entremezclan el enfoque puramente estadístico con el puramente dinámico. En la primer etapa se construye un modelo orientado por datos en el que se relaciona, estadísticamente, la variable a predecir (caudal, en este caso) con ciertos índices predictores entre los cuales existen, al menos, algunos asociados al estado atmosférico futuro. Luego en la segunda etapa, al igual que en el caso puramente dinámico, se modela numéricamente y con base física la evolución del sistema acoplado atmósfera-océano y se generan predicciones del estado atmosférico futuro, también mediante sistemas tier-1 o tier-2, pero no se efectúa la reducción dinámica de escala. Finalmente, las predicciones del estado atmosférico futuro junto con predicciones de los demás índices predictores seleccionados en la primer etapa son utilizadas como insumo para el modelo orientado por datos para generar el pronóstico del caudal. Ejemplos de la utilización de esta técnica pueden ser encontrados en Landman et al. (2001).

De los 3 enfoques mencionados para la predicción mensual o estacional de caudales no hay uno que sea notablemente superior que otro para todas las regiones y épocas del año. Como fue expresado antes, todo aquello que involucre modelos estadísticos podría presentar dificultades en el caso de que el clima utilizado para la generación del modelo no represente la variabilidad futura. Por otro lado, los modelos físicos actualmente presentan problemas en la simulación del clima observado, por lo que su habilidad puede ser restringida según la región o época del año que se quiera estudiar. Adicionalmente, los procesos de downscaling dinámico son complejos, computacionalmente demandantes y presentan variedad de problemas originados por los cambios de resolución. Además, los modelos de circulación general (tanto en modo tier-1 como tier-2) no están totalmente exentos de los problemas de los modelos estadísticos, ya que todos los procesos de escala menor al tamaño de grilla utilizado para la resolución de las ecuaciones son estimados mediante ajustes empíricos. Es decir que, estrictamente, a la fecha no existen modelos puramente dinámicos.

Entre la predicción mensual o estacional de caudales no existen diferencias sustanciales en los enfoques a utilizar pero sí en los resultados que se puedan obtener. Ciertamente, para el diseño de estrategias de operación de embalses las predicciones a escala mensual son preferibles, y más útiles que las de escala estacional. Sin embargo, hay que tener en cuenta que predicciones mensuales serán, posiblemente, más impactadas por la componente no predictible de la variabilidad atmosférica y que, por ende, valores significativos de predictibilidad serán más difíciles de alcanzar.

A pesar de que los valores de predictibilidad puedan ser inferiores (peor relación señal/ruido), dada la importancia y utilidad que tienen los pronósticos más detallados el objetivo de este trabajo será el diseño de sistemas de predicción de caudales de aporte a Rincón del Bonete y Salto Grande, en escala mensual. Se diseñarán sistemas de predicción basados, únicamente, en los esquemas de modelo puramente orientado por datos y de downscaling híbrido. Aunque los caudales de aporte a Rincón del Bonete y Salto Grande no son independientes entre sí cada uno de los embalses será considerado por separado. Adicionalmente, debido a que partes importantes de los sistemas estarán basadas en la selección de predictores adecuados y el análisis de relaciones estadísticas y que estos pueden ser distintos según la época del año, cada mes del año se tratará, también, de forma separada. En resumen, la predicción de los caudales mensuales será realizada de forma independiente para cada embalse y para cada mes del año, siguiendo las metodologías de modelo orientado puramente por datos o downscaling híbrido.

Para cada embalse y mes existen infinidad de variables, muchas de ellas relacionadas entre sí, que podrían ser relevantes para la predicción de los caudales. En este trabajo se determinará un conjunto de variables predictoras a partir del análisis de: la circulación atmosférica regional, la evolución del campo global de TSM y la componente de persistencia representada a través de caudales en meses

anteriores. La posibilidad de utilizar a los caudales en meses anteriores como predictores depende de la antelación del pronóstico, para disponer de estos datos es necesario que, al momento de generación del pronóstico, las observaciones de los mismos ya estén disponibles.

Otras variables que, aunque importantes, no se considerarán como potenciales predictores son, por ejemplo: precipitación y estado de los suelos (humedad contenida en los mismos). A pesar de la estrecha relación que, en esta región, tiene la precipitación con los caudales se decidió no incorporarla, directamente, como potencial predictor debido a que se tiene como objetivo pronósticos con antelaciones mayores al umbral de predictibilidad del tiempo. Por su parte, la predictibilidad estacional del clima será capturada, directamente, por los otros predictores considerados. Es sabido que el estado del suelo puede tener consecuencias importantes en la evolución hidroclimática regional (Grimm et al., 2007), por lo que la inclusión de algún índice representativo de esta condición podría ser aconsejable. En este trabajo consideramos que este estado está, aunque sea parcialmente, representados por las variables relacionadas con los caudales antecedentes y es por ello que no se estimó necesaria su inclusión en otra forma.

Como se mencionó antes, la principal fuente de predictibilidad a escala mensual o estacional reside en la evolución del campo global de TSM. En particular, lo que acontezca en el Océano Pacífico tropical es de fundamental importancia para el hidroclima global, tanto por la extensión de la región como por la amplitud y duración de las anomalías que allí suelen ocurrir. El fenómeno El Niño Oscilación Sur (ENOS) es una interacción cuasi-periódica entre la atmósfera y el Océano Pacífico tropical. Los eventos de ENOS se definen cuando existen anomalías significativas de TSM en el Pacífico tropical ecuatorial: ante anomalías positivas se denominan eventos El Niño y ante anomalías negativas La Niña. Los eventos de ENOS pueden generar efectos significativos en el hidroclima en varias regiones del planeta, ya sean cercanas o distantes del Pacífico tropical. El grado con el cual un evento ENOS impacta el hidroclima de una cierta región depende de la época del año, amplitud y distribución espacial de las anomalías de TSM asociadas con el mismo.

Varios estudios han documentado, en base a observaciones, patrones hidroclimáticos anómalos asociados al fenómeno ENOS. En particular, se ha documentado el efecto de eventos ENOS sobre el clima de América del Sur extratropical (Aceituno, 1988, 1989; Ropelewski y Halpbert, 1987, 1989). La región sudeste de América del Sur (SESA), comprendida por el sur de Brasil, Uruguay y parte del noreste de Argentina, ha sido identificada como una de las que presenta respuestas estadísticamente significativas al fenómeno.

En SESA, Pisciottano et al. (1994) y Cazes-Boezio et al. (2003) encuentran que el fenómeno ENOS tiene efectos estadísticamente significativos durante la primavera austral de un año con presencia de un evento y, aunque de manera más débil, durante el otoño siguiente con tendencia a anomalías de precipitación sobre SESA positivas durante eventos El Niño y negativas durante eventos La Niña. También se han reconocido relaciones entre el fenómeno ENOS y los caudales de ríos en SESA. Mechoso y Pérez-Irribarren (1992) estudian la relación entre ENOS y los caudales de los ríos Uruguay y Negro; encuentran, en ambos ríos, una tendencia a anomalía negativa de caudal desde junio a diciembre de un año La Niña y una tendencia, un poco más débil, a anomalía positiva de caudal de noviembre de un año El Niño a febrero del año siguiente. Por lo recién mencionado consideramos importante la inclusión de, al menos, algún índice relacionado con este fenómeno en el conjunto de potenciales predictores de caudales.

Para el modelo orientado puramente por datos el conjunto de variables predictoras a considerar estará formado por algún índice representativo del fenómeno ENOS y caudales circulantes en meses anteriores. El índice representativo de ENOS puede estar basado en observaciones

disponibles antes de la generación del pronóstico o bien ser una predicción. Por su parte los caudales circulantes en meses anteriores, obviamente, sólo estarán disponibles si la antelación para la generación del pronóstico lo permite. De todas formas, en este trabajo para evaluar la predictibilidad de los caudales y la habilidad de los modelos generados consideraremos el caso en que los predictores son conocidos, es decir que en caso de requerir un pronóstico de alguna variable predictora asumiremos que éste es perfecto. Este procedimiento generará una cota superior a la habilidad del pronóstico, ya que en modo operativo no es posible obtener pronósticos perfectos de los predictores si los mismos no fueron aún observados.

Para el esquema de predicción por downscaling híbrido, al conjunto de variables predictoras considerado en el modelo orientado puramente por datos agregaremos variables relacionadas a la circulación atmosférica regional. En modo operativo, los pronósticos de éstas últimas variables serán generados a través de un MCGA forzado con pronósticos de TSM global. Nuevamente, para cuantificar la predictibilidad y habilidad del esquema de predicción, trabajaremos bajo la hipótesis de predictores conocidos. Hacia el final de este trabajo analizaremos la habilidad un MCGA particular (el de la Universidad de California, Los Ángeles (UCLA)) para predecir algunas variables asociadas a las circulación atmosférica regional.

En las Figuras 1.1 y 1.2, se plantean esquemas de los dos tipos de sistemas (orientado puramente por datos y downscaling híbrido) de predicción de caudales mensuales a Rincón del Bonete y Salto Grande. Los esquemas representan la predicción de caudales en un mes y embalse cualquiera. Las Figuras pretenden ejemplificar el proceso de predicción en modo operativo. En la Figura 1.1 se representa la metodología del sistema orientado puramente por datos mientras que en la Figura 1.2 se representa la metodología asociada al sistema de downscaling híbrido.

En la sección 2 se describen los datos y las principales características del MCGA a utilizar en el trabajo. En la sección 3 se realiza un estudio exploratorio de los datos. En la sección 4 se analizan posibles predictores de caudales y se conforma, para cada mes de año y cada embalse, un conjunto de 12 índices predictores. En la sección 5 se describen, brevemente, las técnicas estadísticas de predicción a utilizar para el desarrollo del modelo orientado por datos. En la sección 6 se presentan los resultados de aplicación de los modelos estadísticos de predicción a los casos en estudio. En la sección 7 se presentan los resultados que se obtienen al utilizar el MCGA de UCLA. Finalmente, en la sección 8 se presenta un resumen con los resultados más importantes y se extraen conclusiones.

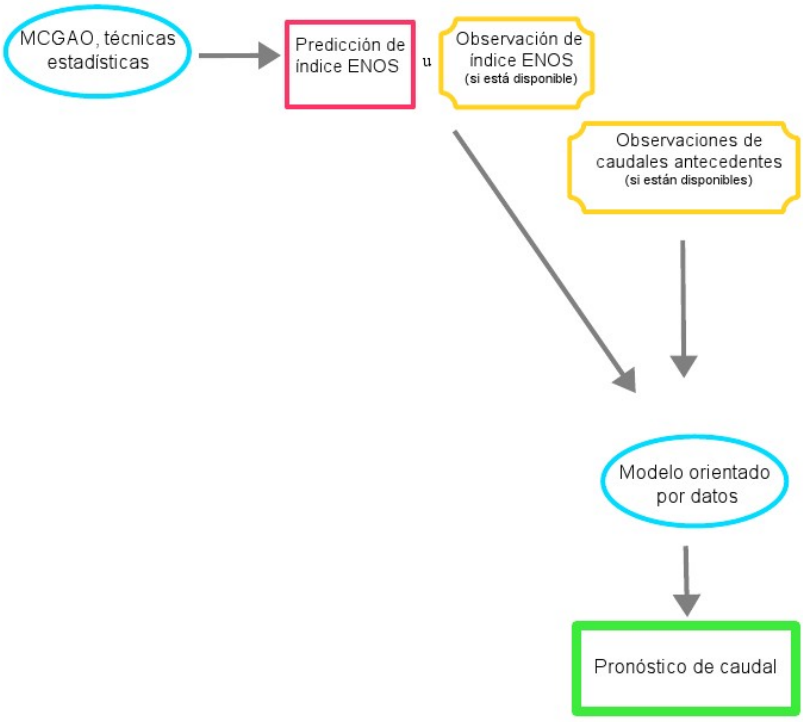

Figura 1.1: Sistema de predicción de caudales mediante modelo orientado puramente por datos.

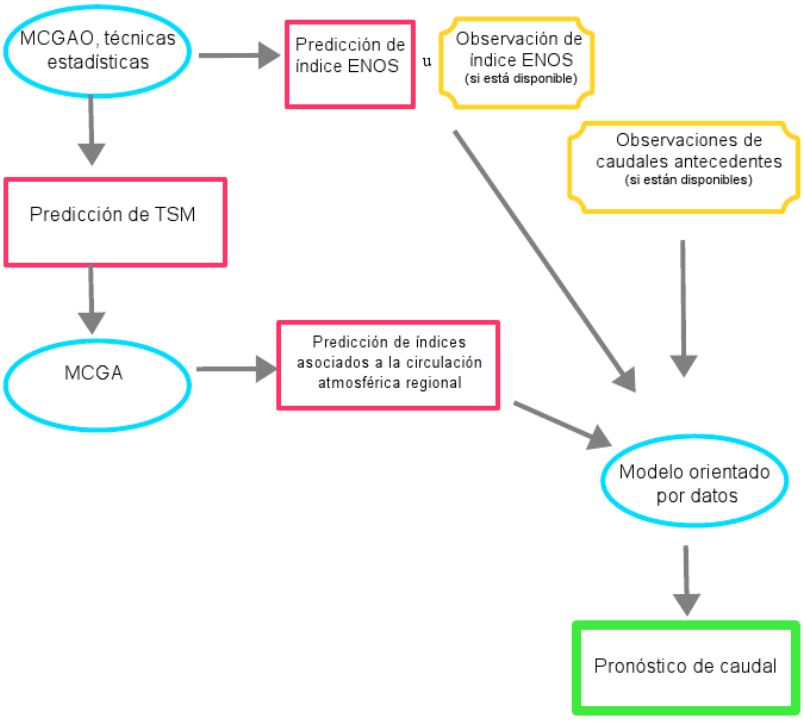

Figura 1.2: Sistema de predicción de caudales mediante downscaling híbrido.

2. DATOS Y DESCRIPCIÓN DEL MODELO

En esta sección se presentan los datos a utilizar: caudales, datos atmosféricos y datos oceánicos. Dada la importancia de la selección del período de estudio se discute la temática en una sub-sección independiente. Hacia el final del capítulo se brinda una breve descripción del MCGA UCLA y se detallan las simulaciones realizadas con el citado modelo.

2.1. Caudales

Se dispone de las series de caudales de aporte mensuales a las represas de Rincón del Bonete y Salto Grande. La serie de caudales de aporte a Rincón del Bonete corresponde a los caudales circulantes por el río Negro en la actual ubicación de la represa de Rincón del Bonete (Dr. Gabriel Terra) y está disponible desde enero de 1908 hasta diciembre de 2007. Por su parte, la serie de caudales de aporte a Salto Grande corresponde a los caudales circulantes por el río Uruguay en la actual ubicación de la represa binacional de Salto Grande y comprende el período enero de 1909 – diciembre de 2008. Todos los caudales están expresados en km^3/hora. Los datos se obtuvieron a través de U.T.E. y de la Comisión Técnica Mixta de Salto Grande.

Ambas series de caudal son no naturalizadas por lo que podrían estar impactadas por variaciones en los usos del agua y suelo en las cuencas de aporte. En particular, el caso de Salto Grande la serie de caudales podría estar influenciada por la operación y evaporación ocurrida en embalses que han sido construidos aguas arriba.

2.2. Datos atmosféricos

El viento se define como aire en movimiento relativo a la superficie de la Tierra. Para representar al viento, en cada instante de tiempo, se utiliza un vector tridimensional el cual queda determinado por tres componentes: zonal, meridional y vertical. La componente zonal se define como la componente del viento según la dirección del paralelo local de latitud. La componente meridional se define como la componente de viento según el meridiano local. Por último la componente vertical se define como la componente en la dirección vertical local. En un sistema coordenado, fijo localmente, con un eje dirigido hacia el este, otro hacia el norte y otro en dirección opuesta a la superficie de la Tierra el viento zonal se considera positivo si fluye del oeste hacia el este, el meridional se considera positivo si fluye de sur a norte y la componente vertical se considera positiva si fluye alejándose de la superficie de la Tierra.

La altura geopotencial es la altura de un punto dado en la atmósfera, referida al nivel medio de la superficie del mar. La relación entre la altura geométrica (h) y la altura geopotencial (hgt) es:

$$hgt = \frac{1}{g_0}\int_0^h g\,(lat\,,z)\,dz$$

en donde g_0 es el promedio global de la aceleración gravitatoria a nivel medio del mar,

$g_0 = 9.80\, m / s^2$ g es la aceleración gravitatoria y lat es la latitud.

Para producir una imagen, lo más precisa posible, del verdadero estado de la atmósfera, océano u otro sub-sistema climático en un cierto momento se realiza un procedimiento denominado asimilación de datos. Por asimilación de datos se entiende al proceso a través del cual toda la información disponible es utilizada para estimar el verdadero estado del sistema. La información disponible consiste, esencialmente, en observaciones y en las leyes físicas que gobiernan la evolución del sistema. Las primeras no están uniformemente distribuidas en el espacio y tiempo y contienen errores propios del proceso de medición. Las últimas, en la práctica, están sólo disponibles bajo la forma de modelos numéricos los cuales, a pesar de brindar información uniformemente distribuida en el espacio y tiempo, no son una representación matemática perfecta de estas leyes. El sistema de asimilación de datos permite propagar la información de regiones con abundancia de datos a otras con carencias de los mismos. Usualmente, la asimilación de datos resulta en la proyección de dicha combinación de información a una grilla regular.

Los reanálisis son un cierto tipo de sistema de asimilación de datos que se distingue de otros básicamente en dos aspectos. Primero, los reanálisis no son calculados en tiempo real. Segundo, el modelo de predicción numérico que se utilizan no cambia durante el período a analizar. Los datos de reanálisis se presentan en una grilla global regular e incluyen datos sobre varios campos derivados (por ejemplo: humedad del suelo) para los cuales observaciones directas son casi inexistentes. En particular, los reanálisis de NCEP/NCAR (Kalnay et al., 1996) tienen como objetivo el producir análisis atmosféricos utilizando datos históricos, desde 1948 en adelante. Estos reanálisis utilizan un sistema de asimilación de datos global, junto con observaciones provenientes de tierra, océanos (barcos y boyas marinas), aeronaves, satélites y otros para reproducir campos globales de varios parámetros meteorológicos.

En este trabajo se utilizan los reanálisis mensuales de NCEP/NCAR de las componentes zonal y meridional del viento y la altura geopotencial, todos en el nivel de 200 hPa. Estos datos están disponibles en una grilla regular global de espaciamiento 2.5º en la latitud (desde 90ºN a 90ºS) y 2.5º en la longitud (desde 0ºE a 357.5ºE) de forma mensual desde enero de 1948 hasta el presente.

Si bien los reanálisis están disponibles desde 1948, hay que hacer ciertas consideraciones en cuanto al grado de confianza que se puede tener en que los mismos representen el verdadero estado de la atmósfera desde su comienzo en todas las regiones del planeta. En particular, para variables atmosféricas de altura (200 hPa) en el Hemisferio Sur la calidad de los datos observacionales tiene un cambio muy importante debido a la incorporación de información satelital, ocurrida en 1979. Antes de la avenencia del satélite las observaciones de altura eran generadas, principalmente, por radiosondas meteorológicas. Estas radiosondas eran extremadamente escasas en todo el Hemisferio austral, por lo que la calidad de los reanálisis de variables de altura en esta región puede ser cuestionada en períodos anteriores a 1979.

2.3. Datos oceánicos

Se utiliza el índice Niño 3.4 de NOAA, el cual se obtiene promediando la TSM en la región comprendida por 5ºS-5ºN y 190ºE-240ºE. Para la TSM NOAA utiliza el análisis ERSST v3b. Este índice se encuentra disponible en forma mensual desde enero de 1950 hasta el presente a través de http://www.cpc.noaa.gov/data/indices/.

Además, para las simulaciones numéricas se utiliza el campo completo de TSM observado. Se utilizan los datos de Reynolds et al., 2002.

2.4. Consideraciones respecto al período de estudio

Es necesario considerar un período en el que todos los datos seleccionados estén disponibles, esto sería posible considerando cualquier período de tiempo posterior a enero de 1950. Sin embargo, como se mencionó antes la calidad de los reanálisis atmosféricos de variables de altura en esta región del mundo podría no ser adecuada antes de la incorporación de la información de satélites.

Por otro lado, como se mencionó en la introducción, siempre que se consideren relaciones estadísticas es necesario tener presente que la no estacionariedad del clima podría implicar cambios según el período en estudio. En particular en este caso, existen diferencias en la relación entre los campos atmosféricos y los caudales en Rincón del Bonete y Salto Grande si se consideran, por ejemplo, los períodos 1948-1978 o 1979-2007. A modo de ejemplo en la Figura 2.4.1 se muestra, en cada punto de grilla, la correlación entre la componente meridional del viento en 200 hPa en el mes de abril en el punto de grilla y el caudal simultáneo en Rincón del Bonete para ambos períodos. El umbral de significancia estadística se obtiene mediante un test de Student unidireccional con tantos grados de libertad como observaciones tengan las series a correlacionar. De acuerdo a este test para el primer período (31 grados de libertad) valores de correlación superiores a 0.30 son estadísticamente significativos a un nivel de 95%, mientras que para el segundo (29 grados de libertad) el umbral de correlaciones significativas es 0.31. En la Figura 2.4.1 sólo se indican los valores de correlación estadísticamente significativos, según este criterio. En ambos períodos se observa que existen correlaciones negativas significativas en SESA, indicando que un aumento en el viento en 200 hPa desde el norte se asocia con un aumento en los caudales circulantes, en el mes de abril. Sin embargo, mientras que las correlaciones en el primer período apenas alcanzan el valor de -0.4, en el segundo llegan hasta -0.9. Estas diferencias podrían deberse, por un lado, a la inclusión de la información de satélites en los reanálisis luego de 1979 pero también podrían estar asociadas al climate shift de finales de los años 70 (Trenberth, 1990; Miller et al., 1994), variabilidad natural interdecaderial, cambio climático antropogénico u otros.

Teniendo en cuenta estos aspectos se selecciona como período de trabajo al posterior a enero de 1979. Se considera que este período presenta datos de calidad adecuada y, al mismo tiempo, es representativo del clima actual.

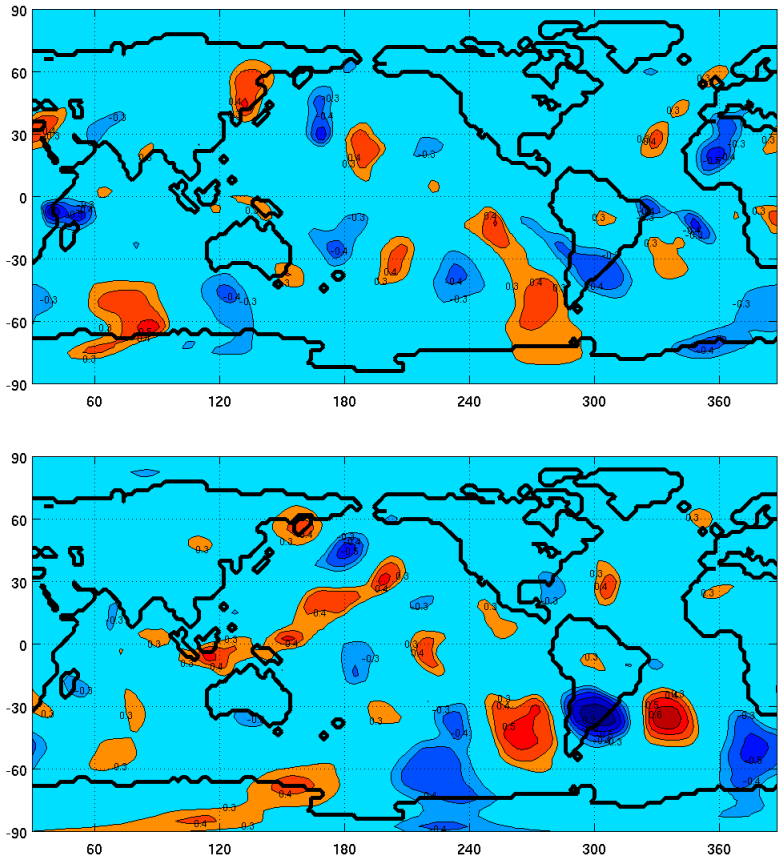

Figura 2.4.1: Correlación entre la componente meridional de viento en 200 hPa en el mes de abril y el caudal simultáneo en Rincón del Bonete para el período 1948-1978 (arriba) y para el período 1979-2007 (abajo). Los tonos azules indican correlación negativa y los rojos positiva. El intervalo de contorno es 0.1, y no se muestran los valores entre -0.2 y 0.2.

2.5. Simulaciones con un modelo de circulación general de la atmósfera

Las simulaciones son realizadas con el MCGA UCLA.

Un MCGA es un modelo numérico basado en las ecuaciones físicas del movimiento. Las ecuaciones atmosféricas en estos modelos pueden ser resueltas tanto por diferencias finitas en ciertos puntos de grilla o de manera espectral. En este tipo de modelos las condiciones de borde que fuerzan la atmósfera deben ser prescritas para poder simular la respuesta atmosférica. Las predicciones estacionales realizadas con estos modelos consisten, típicamente, en un conjunto de simulaciones (o ensemble) y no en una única realización. Debido a la sensibilidad atmosférica a las condiciones iniciales, si se realizan varias simulaciones con las mismas condiciones de borde pero partiendo de condiciones atmosféricas iniciales diferentes se obtendrán resultados diferentes. Las simulaciones ensemble consisten en, justamente, realizar varias simulaciones con los mismos forzantes pero distintas condiciones iniciales, a los efectos de poder separar el efecto de las condiciones de borde de la variabilidad interna. En general, la porción de la señal debida a las condiciones de borde se estima promediando los resultados de las distintas simulaciones, en la creencia de que este promedio contrarrestará la variabilidad interna.

El modelo de UCLA es un modelo de diferencias finitas. La versión utilizada en este trabajo tiene una resolución de 2º en la latitud, 2.5º en la longitud y 29 niveles en la dirección vertical, los cuales se extienden desde la superficie terrestre hasta, aproximadamente, 50km sobre el nivel medio del mar. El modelo predice viento horizontal, temperatura potencial, proporción de mezcla de vapor de agua, proporción de mezcla de agua líquida y hielo en las nubes, espesor de la capa límite planetaria, presión y temperatura en superficie y profundidad de la capa de nieve sobre el suelo. La coordenada vertical utilizada es la coordenada sigma (Suarez et al., 1983) en donde la capa más baja se corresponde con la capa límite planetaria. El grosor de la capa de hielo sobre océano se prescribe siguiendo a Alexander y Mobley (1976). El albedo de superficie y la rugosidad sobre tierra se prescriben de acuerdo a Dorman y Sellers (1989); la rugosidad sobre tierra varía acorde al tipo de vegetación. Los valores diarios son determinados a partir de los valores mensuales por interpolación lineal. Las principales características y parametrizaciones de los procesos físicos de escala menor al tamaño de grilla son descritas en Farrara et al. (2000) y en Konor et al. (2009).

Se realizan simulaciones para testear la posibilidad de utilizar este modelo para el pronóstico de los índices atmosféricos que se identifiquen como predictores de los caudales de circulación en Rincón del Bonete y Salto Grande. Una manera de estimar la confiabilidad de este esquema de pronóstico atmosférico es a través de experimentos en los que la TSM histórica observada es utilizada como condición de borde para correr el modelo. La circulación atmosférica así obtenida es, luego, contrastada con la circulación atmosférica observada. Este tipo de evaluación de la habilidad de un modelo, cuando se extiende por un período suficientemente largo, indica cuán realísticamente el modelo responde a la TSM cuando la TSM es conocida. Es por esto que usualmente se denomina a estas simulaciones como experimentos con "pronóstico perfecto" de TSM. En un esquema de pronóstico operacional, la TSM no será conocida sino que también deberá ser pronosticada, introduciendo un margen adicional de error. En conclusión, las estimaciones de la confiabilidad de un pronóstico climático realizadas a través de experimentos de "pronóstico perfecto" de la TSM deben ser consideradas como una cota superior. Estimaciones más realistas de la confiabilidad de un pronóstico pueden ser obtenidas mediante un ejercicio de "pronóstico retrospectivo", en el cual como condición de borde para el modelo se utilizan pronósticos de TSM, en lugar de TSM observadas.

Este tipo de estimaciones, tanto en el modo de "pronóstico perfecto" de TSM o de "pronóstico retrospectivo", pueden tener la desventaja de que instancias de elevada habilidad ocurran únicamente bajo ciertas situaciones climáticas.

Se simula el período enero 1979 – diciembre de 2008 forzando al modelo con TSM mensual observada (Reynolds et al., 2002). El modelo deduce la variabilidad diaria de la TSM a partir de los valores mensuales, mediante el procedimiento descrito en Farrara et al. (2000). Se realizan 6 simulaciones variando las condiciones iniciales. Se genera, además, una simulación denominada ensemble mean la cual se obtiene promediando los resultados de las 6 simulaciones realizadas. Los resultados se presentan en la secciones 3.3 y 7.

3. ANÁLISIS PRELIMINAR DE DATOS

Comenzamos el análisis realizando un estudio exploratorio sobre los datos.

3.1. Caudales

En la Figura 3.1.1 se presenta la serie temporal de caudales mensuales en Rincón del Bonete desde enero de 1979 hasta diciembre de 2007. Se aprecia una gran variabilidad y, en principio, no se identifican valores extremadamente anómalos (outliers).

Los correlogramas son gráficos que indican las correlaciones de una serie temporal con sigo misma retrasada en el tiempo. Este tipo de gráfico suele presentarse de modo que las abscisas indiquen el retraso y las ordenadas el valor de la correlación. Evidentemente, cuando el retraso es 0 el correlograma mostrará el valor máximo de correlación: 1. Adicionalmente, suelen indicarse ciertos límites de significancia de las correlaciones. En este trabajo se seleccionan como límites de significancia los valores +/- 2/ sqrt(n), en donde n indica la longitud de la serie temporal. Estos valores se corresponden con el umbral de significancia estadística al nivel del 95% para un proceso gaussiano de ruido blanco.

En la Figura 3.1.2 se presenta el correlograma de la serie temporal de caudales de aporte a Rincón del Bonete. Se aprecian correlaciones significativas con hasta dos meses de retraso y picos, también significativos, a los 12 y 24 meses de retraso indicando alguna componente de comportamiento anual.

En la Figura 3.1.3 se presentan los valores medios mensuales (conocido como ciclo anual) junto con la desviación estándar de cada mes. Valores mínimos de los promedios mensuales se presentan en los meses de enero y febrero, y existen dos máximos: uno en el mes de mayo y otro en octubre. Las desviaciones estándar máximas se encuentran en los meses de otoño: desde abril a junio.

A continuación se repite el análisis para los caudales en Salto Grande. En las Figuras 3.1.4, 3.1.5 y 3.1.6 se presentan la serie temporal, correlograma y ciclo anual con desviación estándar. En la serie temporal se aprecia una gran variabilidad y un marcado período de elevados caudales que va de enero a octubre de 1998. El correlograma muestra correlaciones significativas con hasta 5 meses de retraso que, además, son superiores a las análogas para Rincón del Bonete. Al igual que antes, existen picos significativos en las correlaciones en 1 y 2 años, indicando la componente anual del comportamiento de la serie. El ciclo anual tiene un comportamiento similar al de Rincón del Bonete, con mínimos en la temporada de verano y dos máximos: en mayo y octubre, aunque en este caso el valor promedio de octubre es mayor al de mayo. Para Salto Grande las desviaciones estándar de las medias mensuales son mayores en los meses de abril, octubre y noviembre.

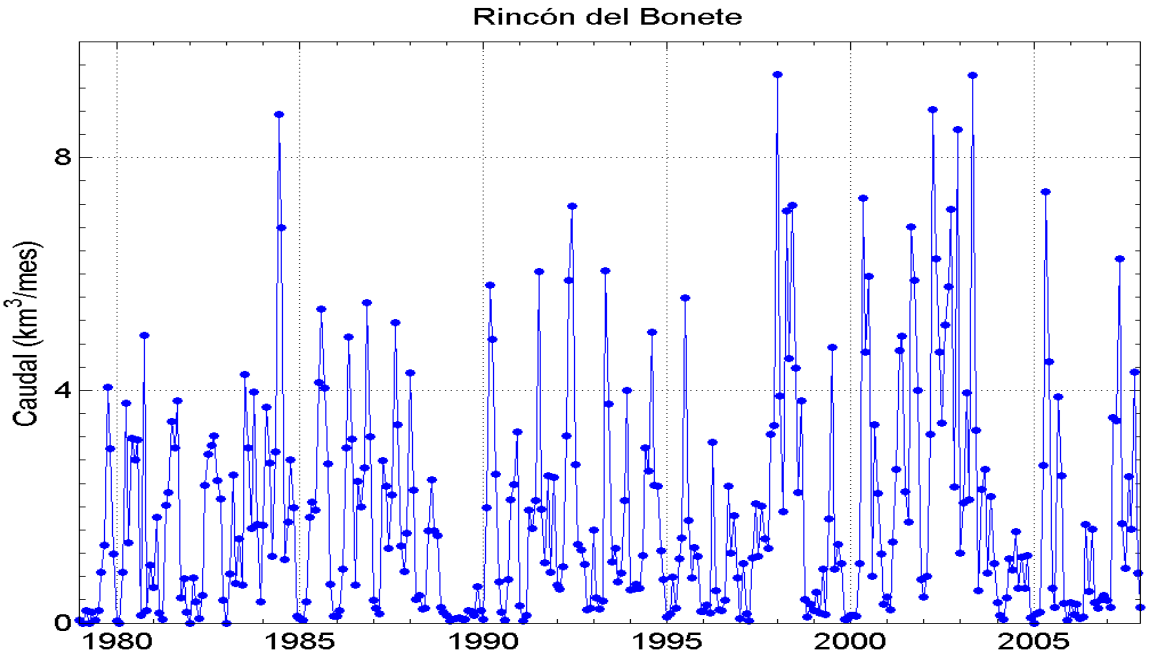

Figura 3.1.1: Serie temporal de caudales mensuales en Rincón del Bonete, desde enero de 1979 a diciembre de 2007.

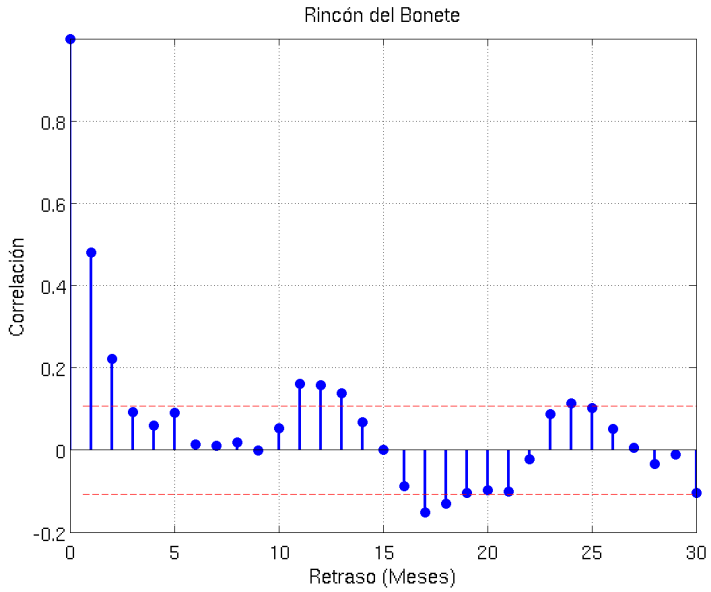

Figura 3.1.2: Correlograma para la serie temporal de caudales mensuales en Rincón del Bonete desde enero de 1979 a diciembre de 2007. Las líneas rojas se corresponden con el umbral de significancia estadística al nivel del 95% para un proceso gaussiano de ruido blanco.

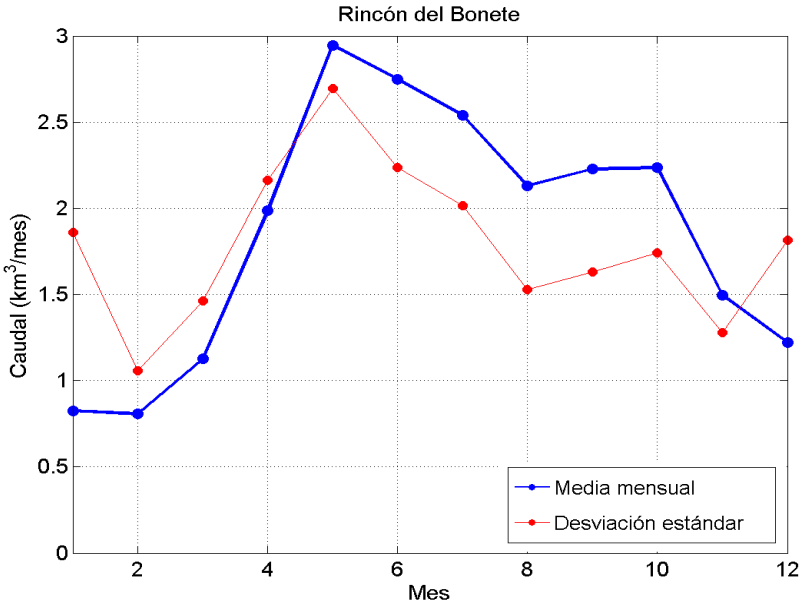

Figura 3.1.3: Caudales medios mensuales y desviación estándar de la serie de caudales de Rincón del Bonete de enero de 1979 a diciembre de 2007.

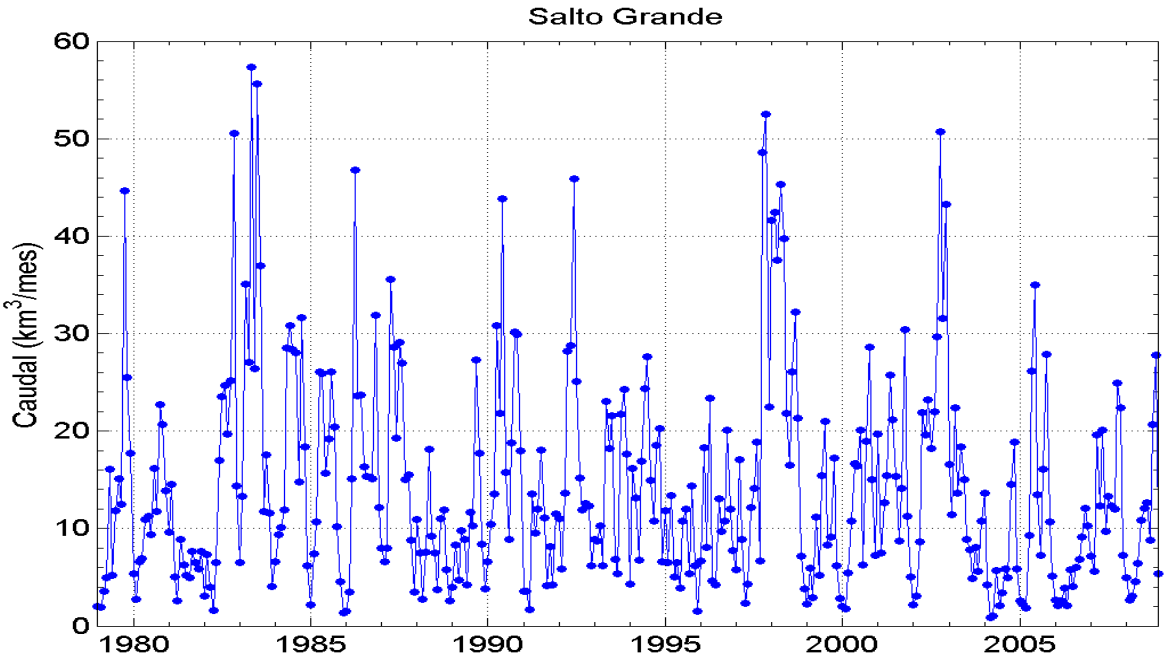

Figura 3.1.4: Serie temporal de caudales mensuales en Salto Grande, desde enero de 1979 a diciembre de 2008.

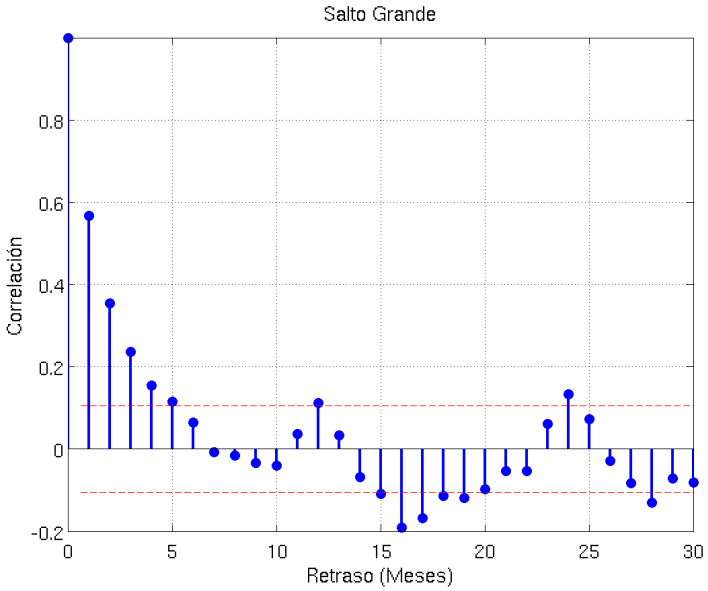

Figura 3.1.5: Correlograma para la serie temporal de caudales mensuales en Salto Grande desde enero de 1979 a diciembre de 2008. Las líneas rojas se corresponden con el umbral de significancia estadística al nivel del 95% para un proceso gaussiano de ruido blanco.

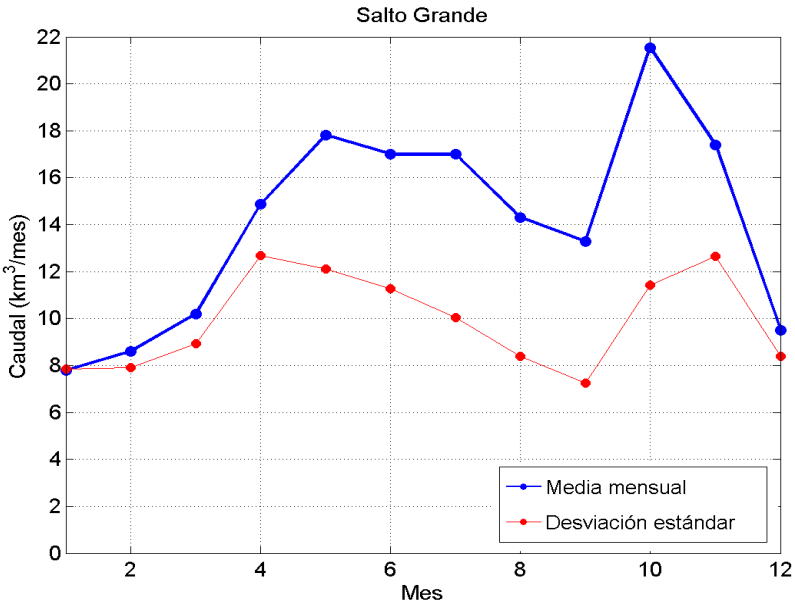

Figura 3.1.6: Caudales medios mensuales y desviación estándar de la serie de caudales de Salto Grande de enero de 1979 a diciembre de 2008.

En la Tabla 3.1.1 se presenta un resumen de las principales características de las series de caudal utilizadas en este trabajo.

	Rincón del Bonete	**Salto Grande**
Período	enero 1979 – diciembre 2007	enero 1979 – diciembre 2008
Mínimo	0 km^3/mes	0.85 km^3/mes
Primer cuartil	0.32 km^3/mes	6.17 km^3/mes
Mediana	1.17 km^3/mes	11.46 km^3/mes
Tercer cuartil	2.73 km^3/mes	19.07 km^3/mes
Máximo	9.44 km^3/mes	57.33 km^3/mes
Media	1.86 km^3/mes	14.11 km^3/mes

Tabla 3.1.1

3.2. Índice Niño 3.4

En la Figura 3.2.1 se presenta la serie temporal del índice Niño 3.4 entre enero de 1979 y diciembre de 2008. Se aprecia que, en general, hay una tendencia a una evolución lenta. Esto puede ser verificado en el correlograma de la serie temporal (Figura 3.2.2), en donde se encuentran correlaciones significativas hasta con 6 meses de retraso. Nuevamente con 1 año de retraso existen correlaciones significativas, lo que evidencia un comportamiento anual. Existe también un pico de correlaciones negativas y significativos al considerar entre 18 y 20 meses de retraso. En la Tabla 3.2.1 se resumen los principales estadísticos de esta serie temporal.

En la Figura 3.2.3 se presentan el ciclo anual (nuevamente, calculado como los valores medios mensuales) junto con la desviación estándar. El ciclo anual alcanza un máximo en el mes de junio y un mínimo en febrero. Por su parte, la desviación estándar es más elevada en los meses de noviembre a febrero.

Por otro lado, en la Figura 3.2.4 se presenta la serie temporal de las anomalías mensuales del índice Niño 3.4, la cual constituye la forma usual de presentar este índice. Las anomalías se obtienen como desviaciones respecto del ciclo anual. En la Figura 3.2.5 se presenta el correlograma de la serie temporal de anomalías mensuales. Se aprecia que existen correlaciones significativas hasta con 8 meses de retraso; un pico importante de correlación negativa significativa se da en el entorno de los 24 meses (2 años) de retraso. Por último, en la Tabla 3.2.2 se resumen los principales estadísticos de la misma.

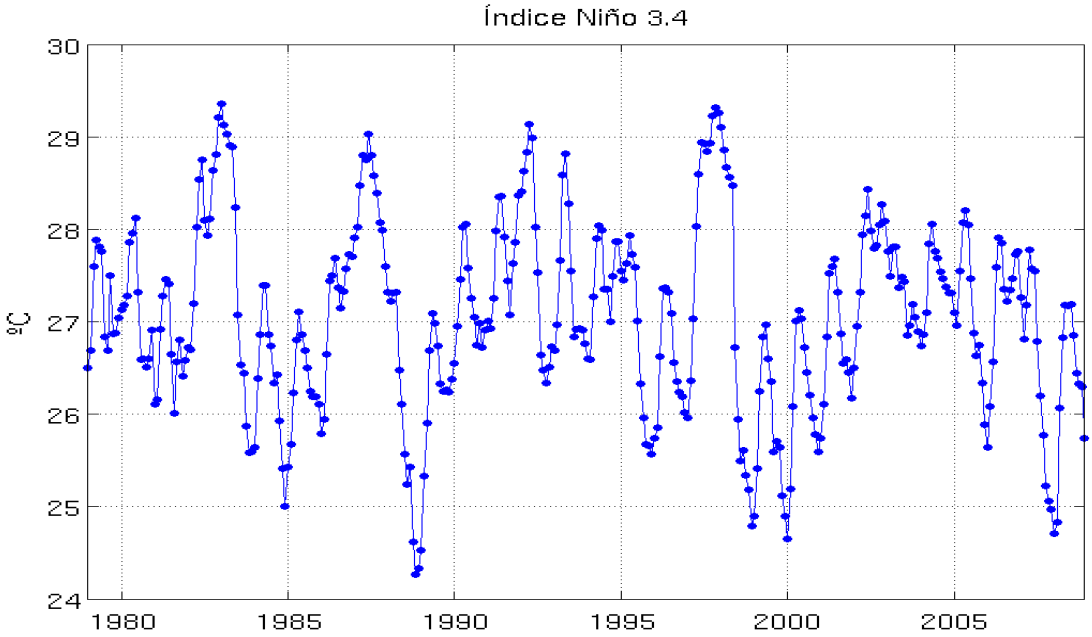

Figura 3.2.1: Serie temporal del índice Niño 3.4, desde enero de 1979 a diciembre de 2008.

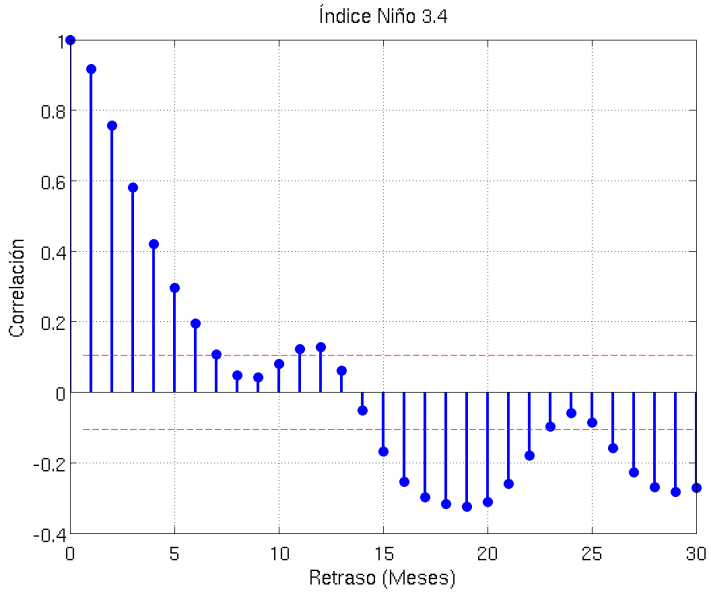

Figura 3.2.2: Correlograma para la serie temporal de índice Niño 3.4, desde enero de 1979 a diciembre de 2008. Las líneas rojas se corresponden con el umbral de significancia estadística al nivel del 95% para un proceso gaussiano de ruido blanco.

Indice Niño 3.4	
Período	enero 1979 – diciembre 2008
Mínimo	26.43 ºC
Primer cuartil	28.14 ºC
Mediana	28.67 ºC
Tercer cuartil	29.09 ºC
Máximo	29.83 ºC
Media	28.57 ºC

Tabla 3.2.1

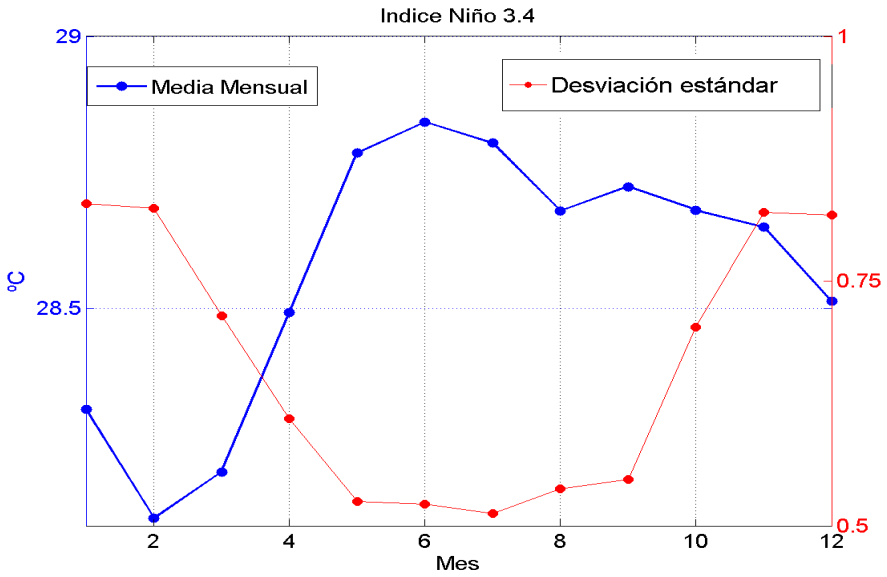

Figura 3.2.3: Valores medios mensuales y desviación estándar de la serie del índice Niño 3.4, de enero de 1979 a diciembre de 2008.

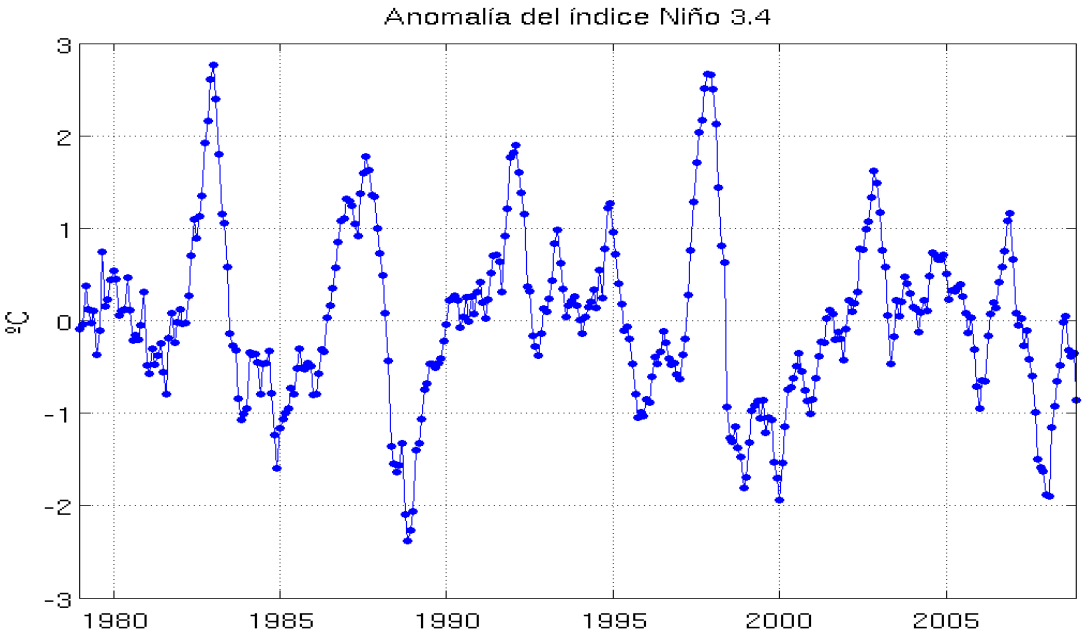

Figura 3.2.4: Serie temporal del anomalías mensuales del índice Niño 3.4, desde enero de 1979 a diciembre de 2008.

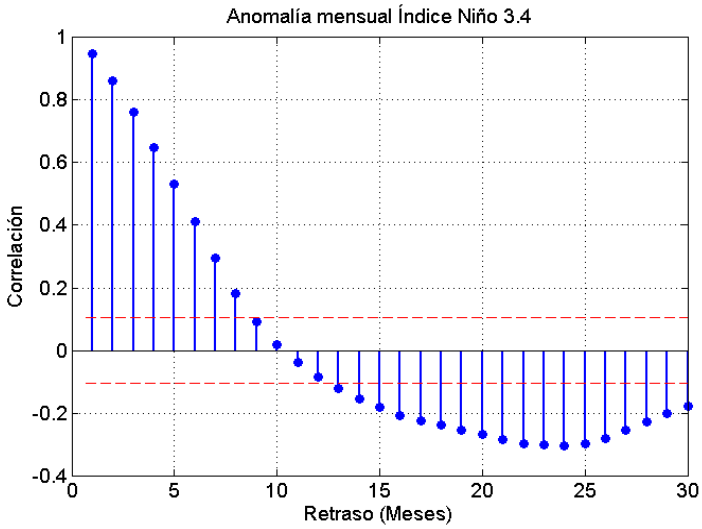

Figura 3.2.5: Correlograma para la serie temporal de anomalías mensuales del índice Niño 3.4, desde enero de 1979 a diciembre de 2008. Las líneas rojas se corresponden con el umbral de significancia estadística al nivel del 95% para un proceso gaussiano de ruido blanco.

Anomalía mensual Indice Niño 3.4	
Período	enero 1979 – diciembre 2008
Mínimo	-2.39 ºC
Primer cuartil	-0.55 ºC
Mediana	0 ºC
Tercer cuartil	0.47 ºC
Máximo	2.79 ºC
Media	0 ºC

Tabla 3.2.2

3.3. Climatologías observadas y simuladas

A modo de ejemplo en la Figura 3.3.1 se presentan las climatologías (promedio en el período 1979-2008) observadas de viento zonal en 200 hPa para los meses de enero y julio. Durante el invierno del hemisferio boreal, el viento zonal en 200 hPa tiene 2 máximos de velocidad (denominados chorros) localizados aguas abajo del plateau del Tíbet y las montañas Rocallosas sobre los océanos Pacífico y Atlántico, respectivamente; en ambos casos el viento proviene del oeste (viento zonal positivo). Durante el invierno del hemisferio austral, la banda de vientos a mayor velocidad se ubica sobre el Océano Pacífico sur, cercano a los -30º de latitud, y también fluyendo en la dirección oeste-este. Se observa que en 200 hPa, sobre casi todo el globo los vientos zonales son del oeste (viento zonal positivo), quedando confinados los vientos del este, prácticamente, sólo a la región comprendida entre los -30º y 30º de latitud.

En la Figura 3.3.2 se presentan las climatologías (promedio en el período 1979-2008) simuladas por la corrida ensemble mean (promedio entra las 6 corridas realizadas) para el viento zonal en 200 hPa en enero y julio. Se aprecia que el modelo es capaz de reproducir las principales características de la circulación zonal de altura como lo son los vientos de oeste a este en las regiones tropicales, las corrientes en chorro con su fortalecimiento a la salida de los continentes y la variación de la intensidad de las mismas según la estación del año. De todas formas, el modelo sobre-estima la magnitud de dichas corrientes, generando una climatología de decenas de metros por segundo más intensas en ciertos lugares.

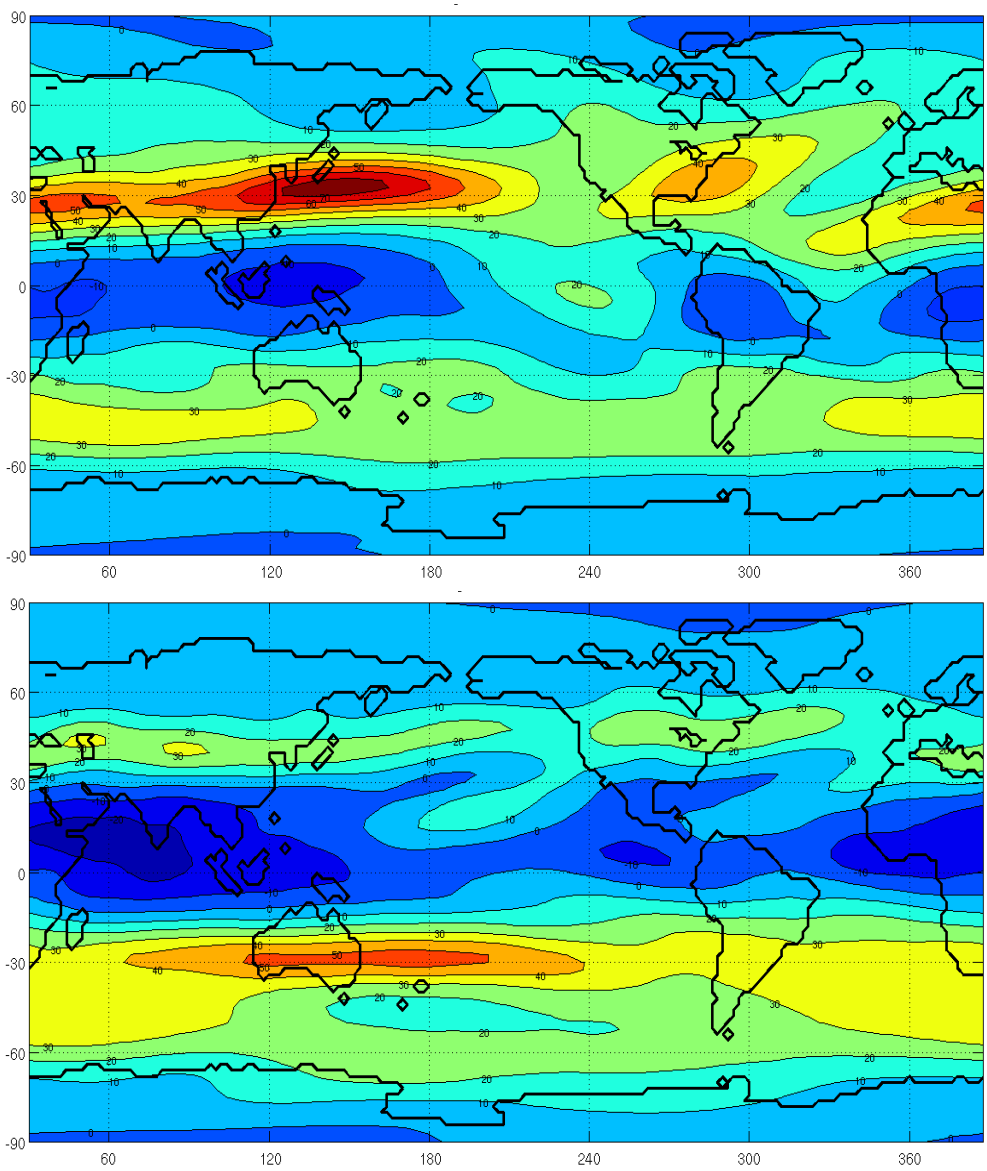

Figura 3.3.1: Climatologías observadas (promedio en el período 1979-2008) para viento zonal en enero (arriba) y julio (abajo). Intervalo de contorno: 10 m/s.

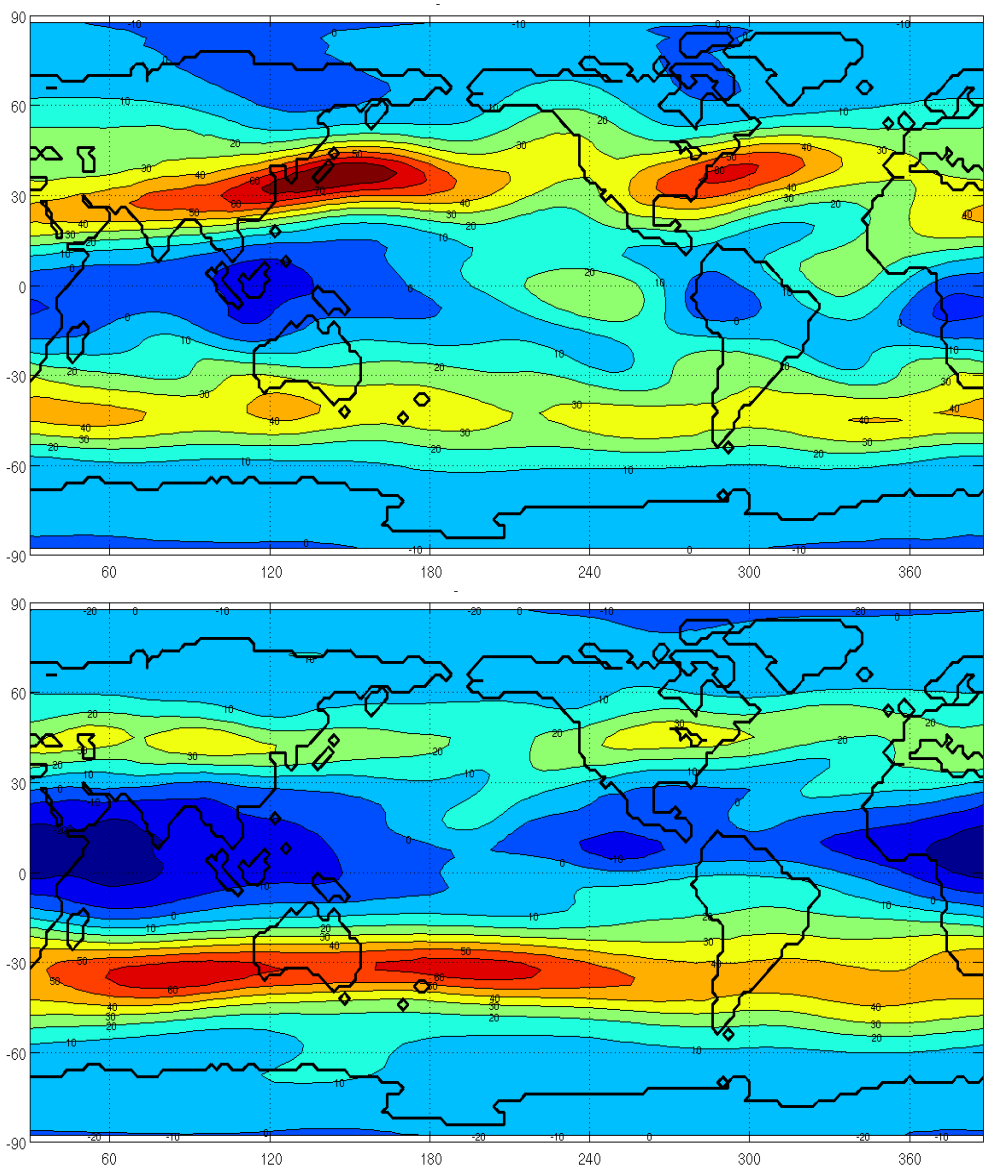

Figura 3.3.2: Climatologías simuladas por la corrida ensemble mean (promedio en el período 1979-2008) para viento zonal en enero (arriba) y julio (abajo). Intervalo de contorno: 10 m/s.

4. DETERMINACIÓN Y ANÁLISIS PRELIMINAR DE PREDICTORES

En esta sección se seleccionará un conjunto de variables predictoras de caudales para cada embalse y cada mes del año a partir de la circulación atmosférica regional, el fenómeno ENOS y los caudales precedentes.

4.1. Circulación atmosférica regional

Para el estudio de la relación entre caudal y circulación atmosférica regional debe tenerse en cuenta que existe un cierto retraso entre la ocurrencia de fenómenos a nivel atmosférico y la manifestación de su respuesta en términos de caudal efluente de una cuenca. Es por ello que se considera apropiado estudiar promedios bimestrales de los campos atmosféricos, con el objetivo de relacionarlos con el caudal observado durante el segundo mes del bimestre.

A los efectos de estudiar la circulación atmosférica regional se considera la región comprendida entre 50ºS-10ºS y 280ºE-330ºE. Dicha región comprende la porción del continente de América del Sur localizada al sur de 10ºS. Denominaremos a esta región AS.

Los reanálisis de NCEP/NCAR de los campos atmosféricos a estudiar tienen 357 puntos de grilla localizados dentro de la región AS. Dada la alta dimensionalidad del problema y el objetivo de representar la relación entre los caudales y los campos atmosféricos de forma simple, se opta por comenzar el estudio sometiendo a los campos atmosféricos a algún procedimiento de reducción de dimensionalidad.

Al lidiar con conjuntos de datos de alta dimensionalidad existe la posibilidad de proyectar los mismos en algún sub-espacio de menor dimensión, sin perder aquella información importante contenida en las variables originales. Una forma usual de lograr esto es mediante la generación de un conjunto restringido de nuevas variables formadas a partir de transformaciones (lineales o no) de las variables originales. Una de las técnicas actualmente más populares para la reducción de dimensionalidad es el análisis de componentes principales (CPs). El análisis de CPs fue introducido por Hotelling (1933) como una técnica que permite derivar un conjunto reducido de proyecciones lineales ortogonales a partir de variables correlacionadas y presentarlas ordenadas según la cantidad de información que cada una contiene.

En el presente estudio someteremos a los campos atmosféricos en consideración al análisis de componentes principales (CPs) y utilizaremos algunas de las nuevas variables generadas como índices predictores de caudales en Rincón del Bonete y Salto Grande. En el Anexo A puede encontrarse un desarrollo de la técnica de análisis de CPs.

4.1.1. Aplicación del análisis de componentes principales a estudios geofísicos

En las ciencias atmosféricas se suele utilizar una nomenclatura particular para el análisis de CPs. Siguiendo la notación que se introduce en el Anexo A, las nuevas variables $\xi_1, ..., \xi_r$ se denominan componentes principales y los vectores propios ($b_1, \ldots, b_r$) funciones empíricas ortogonales o, simplemente, EOF por su sigla en inglés.

La mayoría de las aplicaciones del análisis de CPs a campos geofísicos involucra múltiples observaciones de uno o varios campos. Un problema usual es aquel en el que se tienen múltiples observaciones, en forma de serie temporal, que han sido obtenidas en distintas locaciones geográficas (típicamente distintos puntos de una cierta grilla o ubicaciones de estaciones meteorológicas). Una manera de presentar estos datos al análisis de CPs es definiendo tantas variables como locaciones geográficas se tengan: la variable X_i toma el valor del campo geofísico en la ubicación i, por lo que la variable X_i resulta ser la serie temporal formada por el valor del campo geofísico en la ubicación i.

En el caso anterior, el resultado de un análisis de CPs puede ser desplegado en mapas: cada vector propio contiene exactamente tantos elementos como locaciones geográficas se están considerando, por lo que cada uno de esos elementos puede ser dibujado en la ubicación geográfica correspondiente e incluso luego pueden dibujarse contornos, como es usual para variables hidrometeorológicas. Estos tipos de mapas muestran qué locaciones geográficas contribuyen en mayor medida a cada uno de los vectores propios o, visto de otra forma, indican la distribución espacial de las nuevas variables.

Otra práctica usual para la aplicación del análisis de CPs a campos geofísicos es la utilización de anomalías (respecto a la media en algún período de tiempo que se considere adecuado, por ejemplo anomalías respecto del ciclo diario o ciclo anual), en lugar de la aplicación de la técnica directamente a las variables originales.

4.1.2. Aplicación del análisis de Componentes Principales a la circulación atmosférica regional

En este trabajo efectuaremos un análisis de CPs, en la región AS, para cada una de las medias bimestrales y para cada uno de los tres campos atmosféricos en consideración: viento zonal, viento meridional y altura geopotencial en el nivel de 200 hPa. Con esta técnica se espera reducir, sustancialmente, la cantidad de variables a considerar sin perder, en el proceso, la información más relevante. Realizaremos el análisis de CPs sobre anomalías respecto al ciclo anual. Para cada bimestre del año y cada campo atmosférico las anomalías se obtienen, en cada punto de grilla, primero calculando el promedio bimestral del campo y luego restando el promedio bimestral del mismo campo en el período 1979-2008.

En resumen, el análisis de CPs en la región AS se realiza considerando 357 variables (una variable por cada punto de grilla dentro de la región) que son observadas un total de 30 veces cada una: una vez cada año entre 1979 y 2008 y es, por lo tanto, un análisis de la variación interanual de la circulación atmosférica en la región.

Como se discute en el Anexo A, el análisis de CPs puede realizarse en base a la matriz de covarianza o a la matriz de correlaciones. Dado que, en este caso, el análisis es realizado para cada campo atmosférico de forma independiente, la inclinación hacia la utilización de la matriz de correlaciones no es obligatoria. Sin embargo, podría ocurrir que por la magnitud de la región seleccionada para el análisis ciertos puntos de grilla presentaran valores mucho mayores de variabilidad y por lo tanto, fueran solamente esos puntos los que dominaran en el análisis si sólo se utilizara la matriz de covarianzas. En consecuencia, se calcularon las CPs con ambos enfoques pero luego de apreciar que no existen diferencias significativas para los datos en estudio, se optó por presentar los resultados obtenidos únicamente con la matriz de covarianzas, lo cual representa el enfoque más tradicional.

El análisis de CPs en la región AS, para cada bimestre y cada campo atmosférico considerado, arroja entonces 357 variables (denominadas CPs) ordenadas según el porcentaje de la varianza interanual total explicada. La reducción de la dimensionalidad del problema puede lograrse considerando solamente algunas de estas 357 CPs. Existen varios criterios que intentan estimar, de una manera objetiva, qué cantidad de CPs retener en el análisis. Entre ellos uno de los más utilizados consiste en retener tantas CPs como sean necesarias para alcanzar un cierto porcentaje de varianza explicada, comenzando por las CPs que mayor porcentaje expliquen. Siguiendo este criterio, al requerir que un 50% de la varianza total sea explicada, en general, para todos los campos y todos los bimestres se deben retener las primeras 2 o las primeras 3 CPs. Para uniformizar se decide retener, en todos los casos, las primeras 3 CPs.

Las CPs obtenidas son series temporales, consistentes en una realización por año. Para los mapas de EOF (vectores propios) se despliegan cada una de sus 357 componentes, en cada uno de los 357 puntos de grilla contenidos en la región AS. Para lograr una idea más global de los patrones también es usual, en las ciencias atmosféricas, dibujar en cada punto de grilla fuera de la región considerada una extensión de las EOF. La extensión de la EOF asociada a una CP de cierto campo atmosférico en un punto de grilla (i,j) es el coeficiente de primer orden que se obtiene al hacer la regresión lineal del campo en el punto (i,j) contra la CP. Tanto el campo en el punto (i,j) como la CP son series temporales conformadas por una observación por año (dado que el análisis es interanual).

A modo de ejemplo, en la Figuras 4.1.2.1-4.1.2.4, presentamos la primer CP estandarizada del viento zonal en 200hPa así como la EOF asociada, en distintos bimestres: enero-febrero, abril-mayo, julio-agosto y octubre-noviembre. En todos los gráficos la región AS es indicada mediante una línea.

En las Figuras 4.1.2.1- 4.1.2.4 se aprecia que el patrón de la primer EOF del viento zonal en 200 hPa en los distintos bimestres seleccionados consiste, principalmente, de una estructura de vórtice centrado en (30ºS, 310ºE) el cual queda comprendido, casi en su totalidad, dentro de la región AS. Como se aprecia, la nitidez de la estructura depende del bimestre, siendo julio-agosto el bimestre en el que ésta aparece menos clara (de entre los bimestres desplegados aquí). Es importante apreciar, también, que en octubre-noviembre el citado vórtice se asocia con anomalías de viento zonal importantes en la cuenca del Océano Pacífico, mientras que en los otros bimestres esto no sucede. Este hecho fue notado y analizado en detalle por Cazes-Boezio et al. (2003).

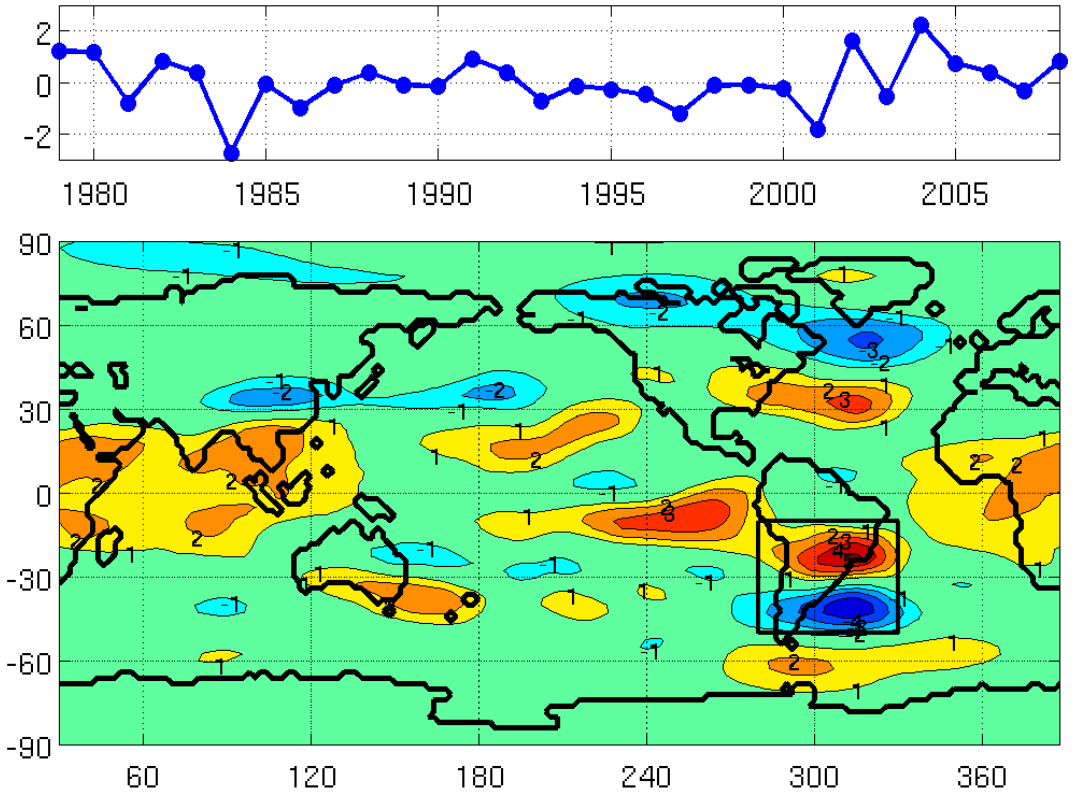

Figura 4.1.2.1: Primer CP del viento zonal en 200hPa en el bimestre enero-febrero (arriba) y EOF asociada, extendida a todo el globo, intervalo de contorno:1 m/s (abajo), se omite el contorno 0.

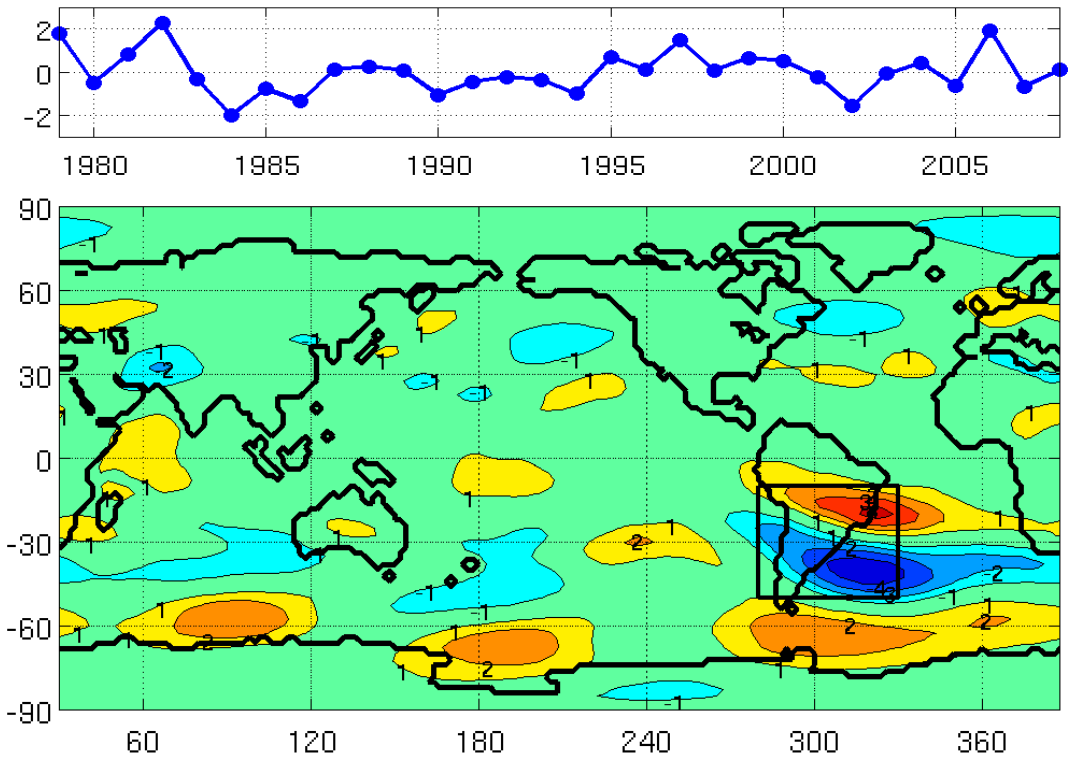

Figura 4.1.2.2: Idem Figura 4.1.2.1 para el bimestre abril-mayo.

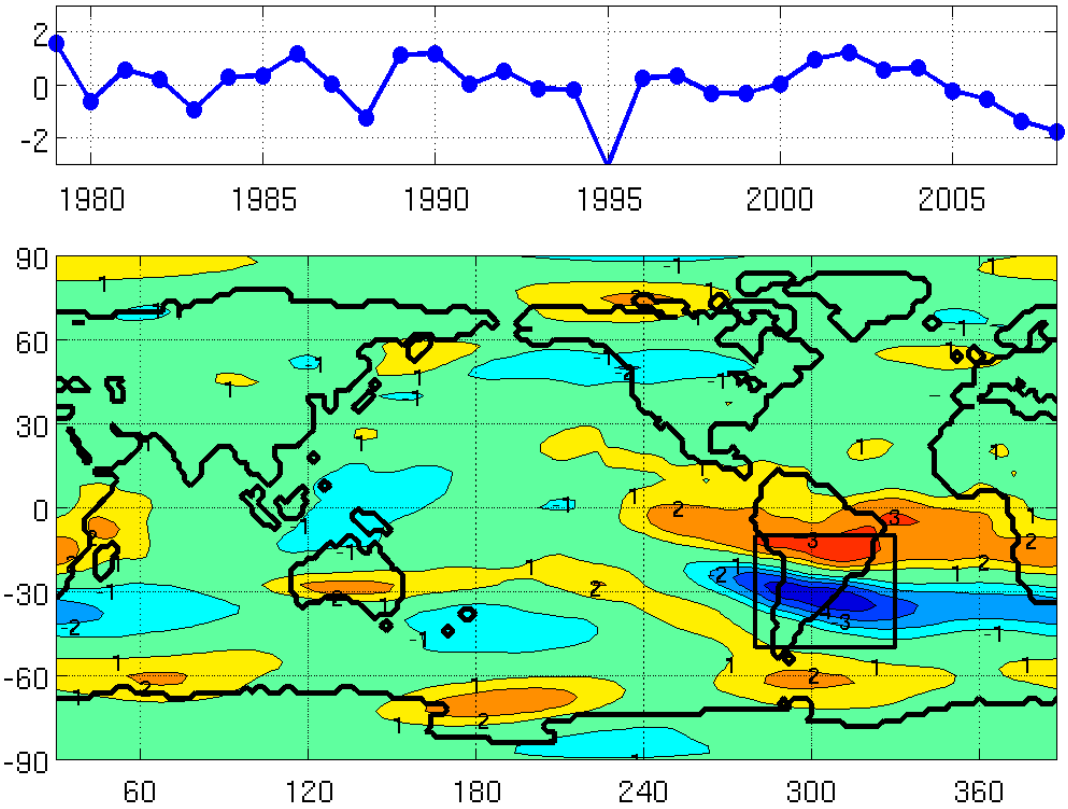

Figura 4.1.2.3: Idem Figura 4.1.2.1 para el bimestre julio-agosto.

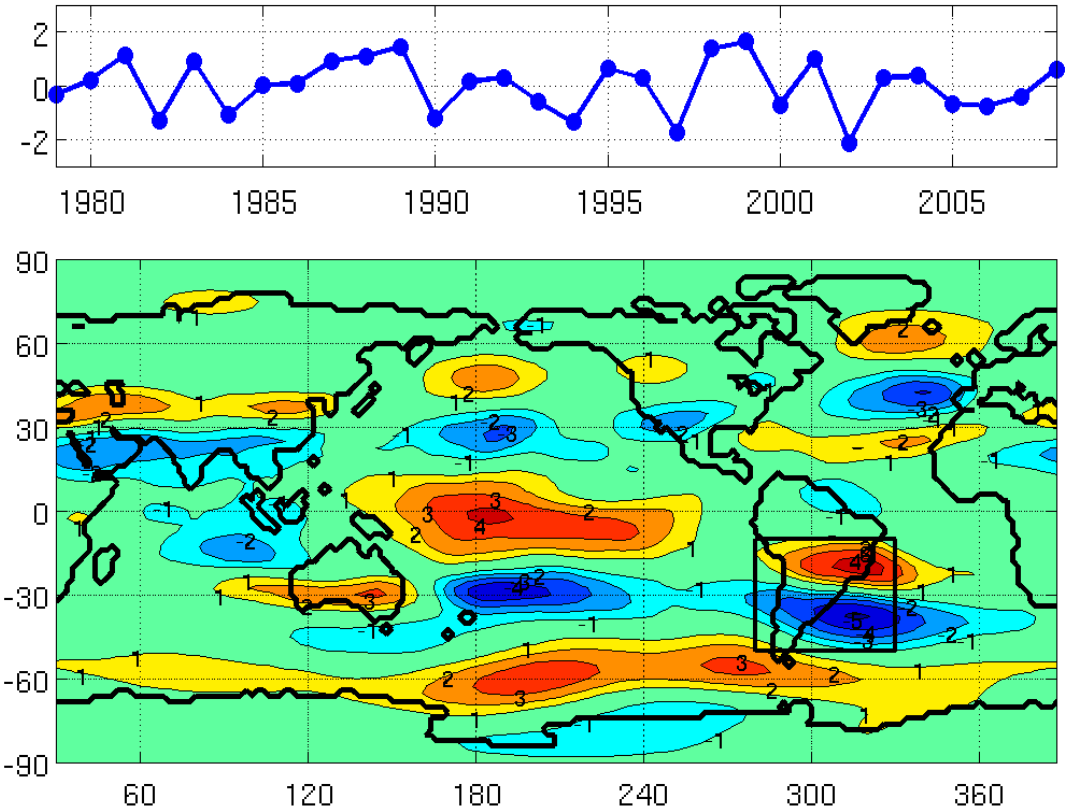

Figura 4.1.2.4: Idem Figura 4.1.2.1 para el bimestre octubre-noviembre.

La selección de la región para el análisis de CPs es, en principio, arbitraria. Cualquier dominio que, al menos, contenga la ubicación de las represas de Rincón del Bonete y Salto Grande podría significar una elección adecuada. La decisión final de utilizar la región AS fue tomada luego de realizar varias pruebas considerando regiones entre las cuales se incluyen algunas con mayor y otras con menor cantidad de puntos de grilla o en ubicaciones un poco distintas. Regiones con poca cantidad de puntos de grilla podrían ocasionar que los patrones identificados representaran únicamente modos de variabilidad particulares de la pequeña región en consideración y fueran, por tanto, extremadamente difíciles de predecir. Por el contrario, regiones muy amplias podrían dar lugar a modos de variabilidad que poca relación tuvieran con la variabilidad que afecta los caudales de circulación en Uruguay. Creemos que la región AS logra un equilibrio entre patrones lo suficientemente importantes como para poder ser predichos y los suficientemente particulares en cuanto a ser representantes de la variabilidad de la región que afecta a los caudales. Otra característica que cumple la región AS es que, en su gran mayoría, los patrones de las primeras EOF de los campos atmosféricos considerados tienen sus estructuras principales dentro de los límites de la propia región. Nuevamente, a modo de ejemplo, en la Figura 4.1.2.5 se presentan los patrones geográficos de las EOF (y sus extensiones globales) asociadas a la primer CP del viento zonal en 200hPa en el bimestre enero-febrero al considerar distintos dominios para el análisis de CPs.

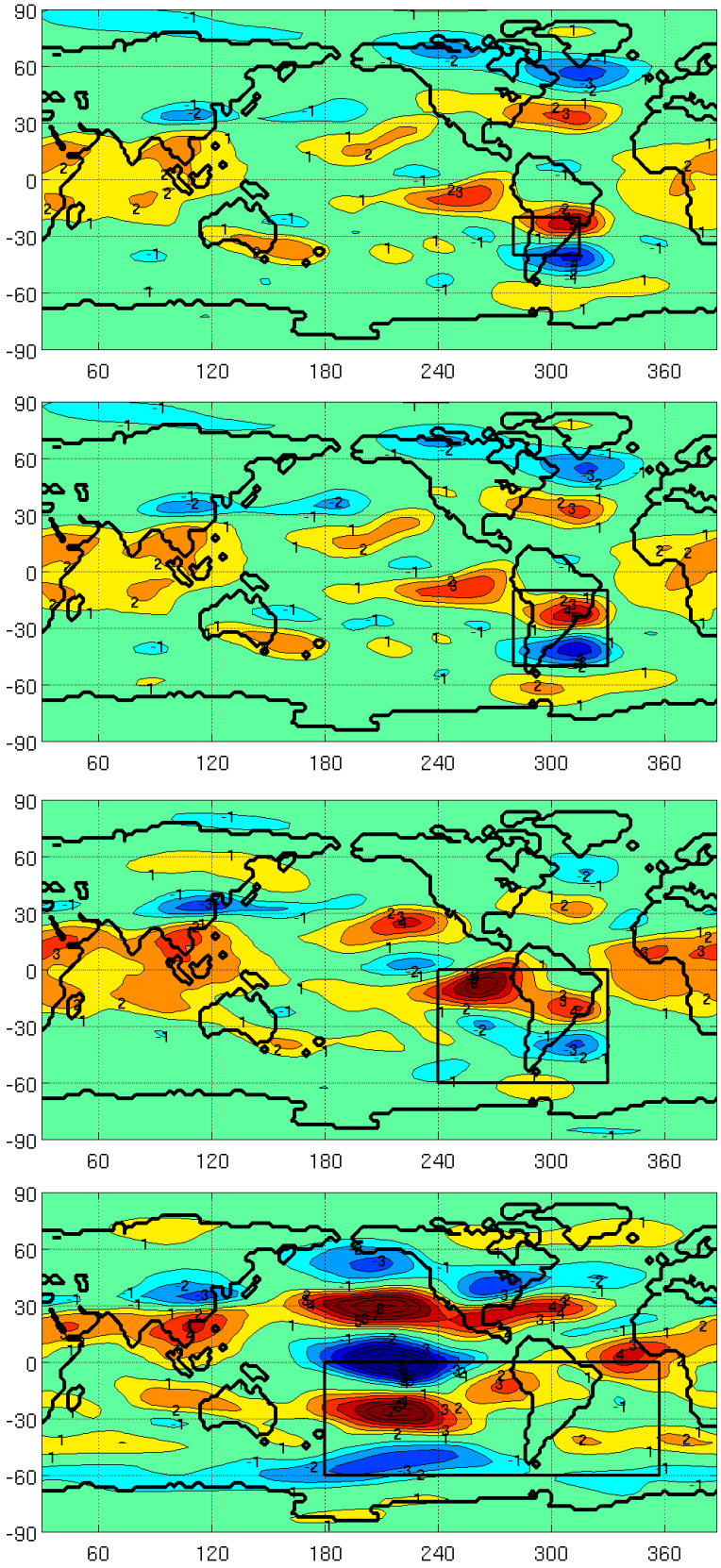

Figura 4.1.2.5: EOF (y extensión global) asociada a la primer CP del viento zonal en 200hPa en el bimestre enero-febrero, al utilizar distintos dominios para el análisis de CPs. Los dominios correspondientes son indicados mediante rectángulos. Intervalo de contorno: 1 m/s, se omite el contorno 0.

En síntesis, para cada uno de los 12 bimestres del año, se generaron 9 índices (las CPs retenidas) que reflejan la variabilidad interanual en la región AS: 3 de ellos asociados al viento zonal en 200hPa, 3 asociados al viento meridional en 200hPa y otros 3 a la altura geopotencial en 200hPa. Estos 9 índices serán utilizados como variables predictoras de los caudales en Rincón del Bonete y Salto Grande. El procedimiento seguido para la obtención de los índices no garantiza de forma alguna que éstos puedan resultar de utilidad en el proceso de predicción, aunque dada la relación entre caudal y circulación atmosférica una tal utilidad es esperable. Una manera simple de evaluar el potencial de estos índices como variables predictoras de caudal es calcular sus correlaciones con las series de caudal que esperamos sean capaces de predecir. Más adelante presentaremos estos resultados, complementados con los restantes índices predictores que serán utilizados en el trabajo y que desarrollaremos a continuación, en las siguientes dos sub-secciones.

4.2. Fenómeno El Niño Oscilación Sur

Como fue mencionado previamente, la influencia del fenómeno ENOS en el hidroclima de SESA es ampliamente reconocida. En esta sección buscamos determinar alguna variable relacionada con ENOS que posea potencial para oficiar de variable predictora de caudales en Rincón del Bonete y/o Salto Grande.

A modo de ejemplo, en la Figura 4.2.1 presentamos la correlación entre el caudal circulante en Salto Grande en el mes de noviembre y el campo global de TSM simultáneo, considerando el período 1979-2008. En dicha figura se observan correlaciones muy elevadas del caudal con la TSM en la región centro-este del Océano Pacífico ecuatorial, región en la que se desarrolla el fenómeno ENOS. En particular, en la Figura 4.2.1 puede apreciarse la ubicación de la región Niño 3.4.

A pesar de que en el caso particular mostrado en la Figura 4.2.1 la TSM en la región Niño 3.4 no es la que mejor se correlaciona (en el sentido de valor absoluto de la correlación simultánea) con el caudal las correlaciones que muestra son, de todas formas, muy elevadas. Al considerar caudales mensuales en otros meses y correlacionarlos con el campo global de TSM (simultáneo o con cierta antecedencia) esta situación o bien se repite o bien se encuentra que sí es en esta región donde se alcanzan las magnitudes de correlación máximas. Dada la facilidad con la que predicciones del índice Niño 3.4 pueden obtenerse, y con ánimos de simplificar la selección de variables predictoras, en esta etapa del trabajo se intenta identificar algún potencial predictor de los caudales solamente a partir de los registros históricos del índice Niño 3.4.

Variaciones interanuales de ciertas variables atmosféricas (como la temperatura media troposférica en los trópicos y otras) tienden a seguir los cambios de la TSM en el Océano Pacífico ecuatorial este, obteniéndose respuestas máximas luego de una o dos estaciones (Kumar y Hoerling, 2003; Su et al., 2005 y referencias en ellos). Por otro lado, también son conocidas relaciones entre índices asociados al fenómeno ENOS y la TSM en regiones oceánicas diferentes al Pacífico tropical, de modo que el fenómeno ENOS es el factor antecedente (Su et al., 2005 y referencias allí). Estos hechos motivan que para identificar un predictor de caudales relacionado con ENOS estudiemos el estado del fenómeno no sólo en simultaneidad con los caudales a predecir, sino también con varios meses de antelación. Para cada mes del año, se calcularon las correlaciones entre las series de caudales mensuales y los promedios bimestrales del índice Niño 3.4, en bimestres con hasta 1 año de antelación al mes en estudio. El índice bimestral con p meses de antecedencia al caudal en el mes (m), es el promedio bimestral del índice en los meses (m-p) y (m-p+1); por tanto, el índice bimestral

con 1 mes de antelación es el correspondiente a los meses (m-1) y (m). La consideración de promedios bimestrales es sólo a fines de suavizado de la serie.

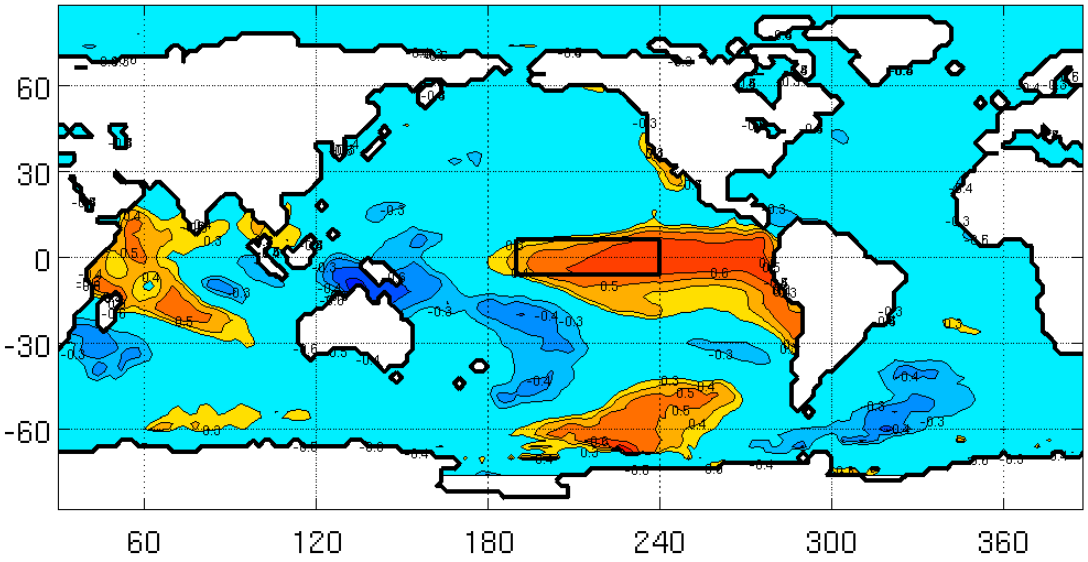

Figura 4.2.1: Correlación entre caudal circulante en Salto Grande en el mes de noviembre y campo global de TSM simultáneo, considerando el período 1979-2008. Sólo se muestran las correlaciones estadísticamente significativas al 95%. Se indica, además, la región Niño 3.4.

En las Figuras 4.2.2 (4.2.3) se muestran diagramas que indican los valores de correlación entre los caudales mensuales en Rincón del Bonete (Salto Grande) y el índice Niño 3.4 con distintas antelaciones. La significancia estadística se vuelve a calcular en base a un test de Student de 29 o 30 grados de libertad, según sea Rincón del Bonete o Salto Grande, por lo que valores de correlación superiores a 0.32 o 0.31 son estadísticamente significativos a un nivel del 95% de confianza, respectivamente. Para ambos embalses todas las correlaciones estadísticamente significativas son positivas.

Para Rincón del Bonete la relación caudal – índice Niño 3.4 parece seguir un comportamiento estacional. Sólo se obtienen correlaciones significativas con los caudales circulantes en los meses de noviembre a febrero (verano). Aún dentro de la temporada de verano, parecen co-existir 2 tipos diferentes de estructuras en la relación caudal – índice Niño 3.4: una para los caudales en noviembre y diciembre y otra para los de enero y febrero. Para los caudales en noviembre y diciembre la correlación máxima se obtiene contra el índice Niño 3.4 bimestral con 2 o 3 meses de antelación, mientras que para los caudales de enero y febrero las correlaciones con el índice con pocos meses de anticipación son relativamente bajas, o no significativas, pero aumentan hasta niveles muy altos y se tornan máximas a los 6 y 8 meses de anticipación, respectivamente.

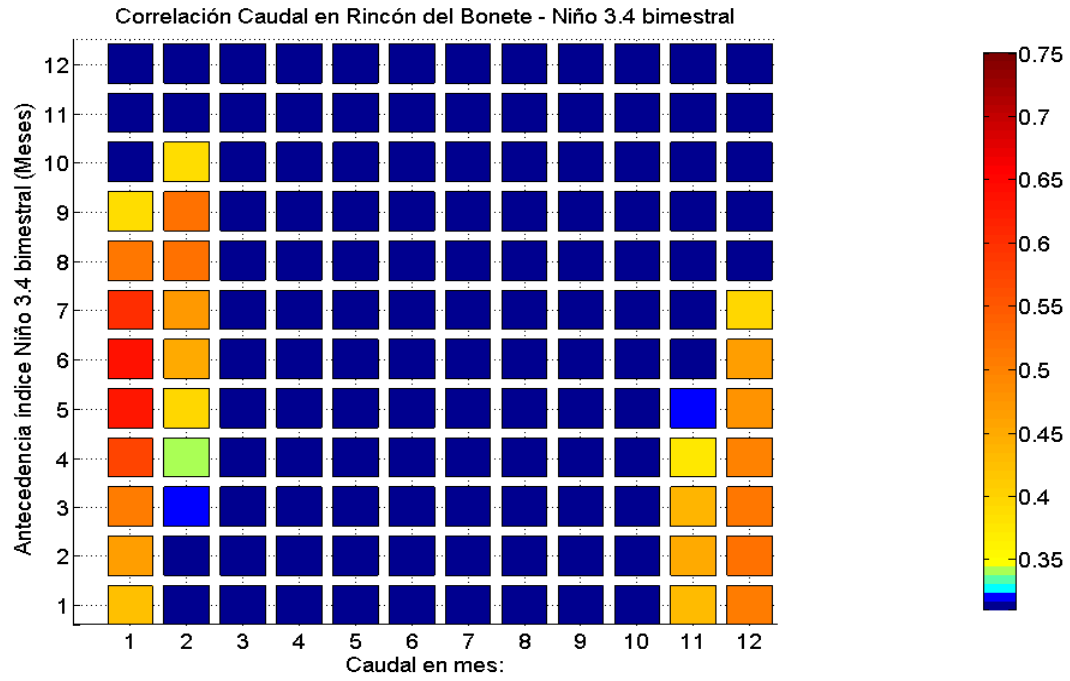

Figura 4.2.2: Correlación entre caudal mensual en Rincón del Bonete e índice Niño 3.4 bimestral con distintas antecedencias. El mes correspondiente al caudal se indica en el eje de las abscisas mientras que la antecedencia del índice Niño 3.4 se indica, en meses, en el eje de las ordenadas. Aquellas correlaciones que no alcanzan el límite de significancia estadística el 95% se indican en azul oscuro. El período considerado para el cálculo de la correlación es 1979-2007.

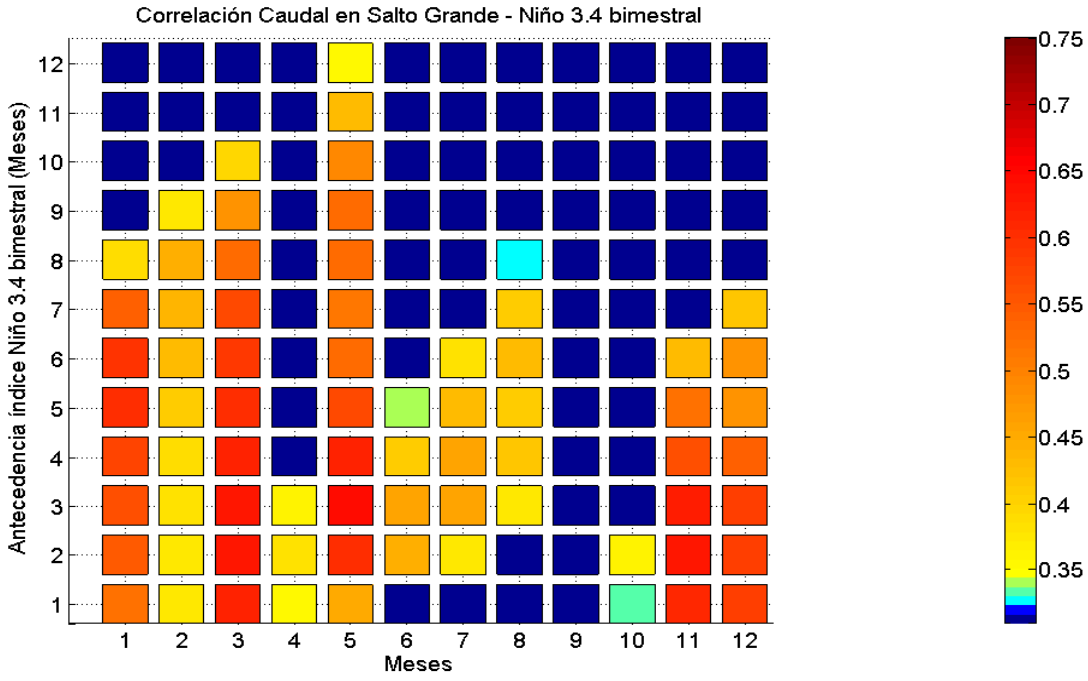

Figura 4.2.3: Idem Figura 4.2.2 para Salto Grande. El período considerado para el cálculo de la correlación es 1979-2008.

Finalmente, para cada embalse y cada mes del año, se selecciona como predictor asociado al fenómeno ENOS al índice Niño 3.4 antecedente en el bimestre en que la correlación con el caudal del embalse sea máxima (en el sentido de valor absoluto máximo). Denominaremos a este predictor índice Niño 3.4 óptimo. En casos en que con ninguna antecedencia se alcancen valores de correlación significativos, se utilizará como índice Niño 3.4 óptimo aquel con 1 mes de antelación.

4.3. Caudales antecedentes

El estudio de las series temporales de caudal realizado en la sección de estudio preliminar de datos indicaba una cierta componente de persistencia en las mismas, es decir una relación entre los caudales de meses sucesivos. Para profundizar en esa dirección haremos un análisis similar pero considerando a cada mes por separado, dado que este tipo de relaciones podría ser distinta según la época del año.

En las Figuras 4.3.1 y 4.3.2 presentamos los valores de correlación entre las series de caudales mensuales y los caudales antecedentes, con hasta 2 meses de antelación, para Rincón del Bonete y Salto Grande, respectivamente.

Para Rincón del Bonete (Figura 4.3.1) se observa que la correlación con caudales de 1 mes de antecedencia es positiva y significativa en casi todos los meses del año (se exceptúan enero y agosto); se destaca el elevado valor para el mes de febrero. Para el mismo embalse las correlaciones con los caudales de 2 meses previos caen marcadamente y sólo son significativas (y también positivas) para 4 meses en el año. Por su parte para Salto Grande, en general, los valores de correlación son mayores que los análogos en Rincón del Bonete siendo las correlaciones con los caudales del mes precedente positivas y significativas durante todo el año y las correlaciones con caudales con 2 meses de antecedencia también positivas y significativas en verano, otoño e invierno. Para Salto Grande (Figura 4.3.2), incluso, en ciertos meses las correlaciones con caudales con 2 meses de antelación son superiores a las obtenidas con 1 mes de antecedencia. También se destaca que, a pesar de ser significativas, las correlaciones con el caudal del mes precedente tienen una fuerte caída en los meses de setiembre y octubre.

Dados los resultados mostrados en las Figuras 4.3.1 y 4.3.2, se incorpora al conjunto de variables predictoras los caudales con 1 y 2 meses de antecedencia.

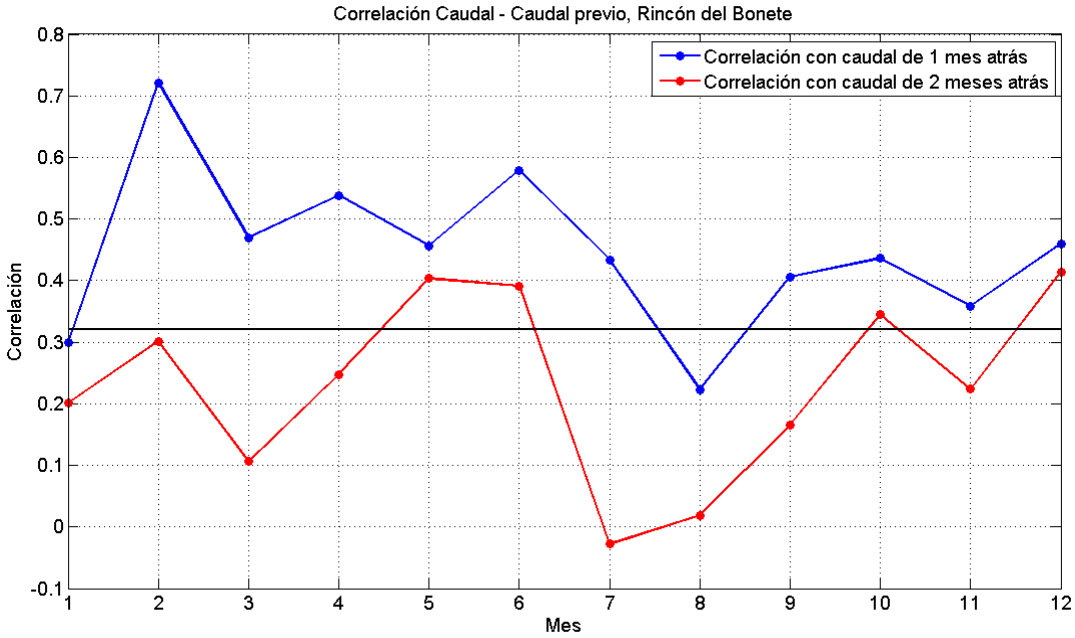

Figura 4.3.1: Correlación entre caudal mensual y caudal con 1 y 2 meses de antelación en Rincón del Bonete. Se indica en nivel de 95% de significancia estadística.

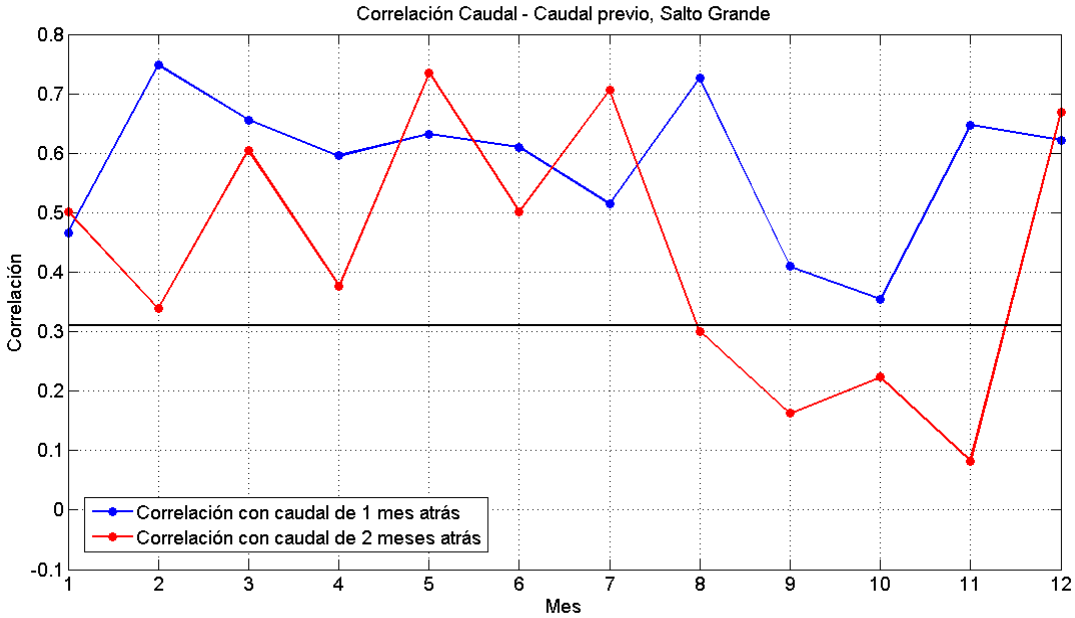

Figura 4.3.2: Idem Figura 4.3.1 para Salto Grande.

4.4. Resumen

Para cada embalse y cada mes del año se han seleccionado un total de 12 variables predictoras: 9 asociadas a la circulación atmosférica regional, 1 relacionada con el fenómeno ENOS y 2 que representan la componente de persistencia de caudales.

Para denotar a cada una de las 9 variables predictoras asociadas a la circulación atmosférica regional utilizaremos la siguiente nomenclatura: Pc (por ser una componente principal) + número (1, 2, 3 según sea la primera, segunda o tercera) + variable (u, v, hgt según corresponda a viento zonal, meridional o altura geopotencial). El índice Niño 3.4 óptimo elegido como predictor será notado N3.4 y los caudales con 1 y 2 meses de antecedencia Q1 y Q2, respectivamente.

A modo de resumen, en las Figuras 4.4.1 y 4.4.2 se presentan diagramas que indican el valor absoluto de las correlaciones entre las series de caudal mensuales y las 12 variables predictoras para Rincón del Bonete y Salto Grande, respectivamente. Al igual que antes, sólo se indican las correlaciones estadísticamente significativas al nivel del 95%. Como observación general se destaca que ninguno de los 12 predictores seleccionados para el caudal en Rincón del Bonete en agosto alcanza valores significativos de correlación.

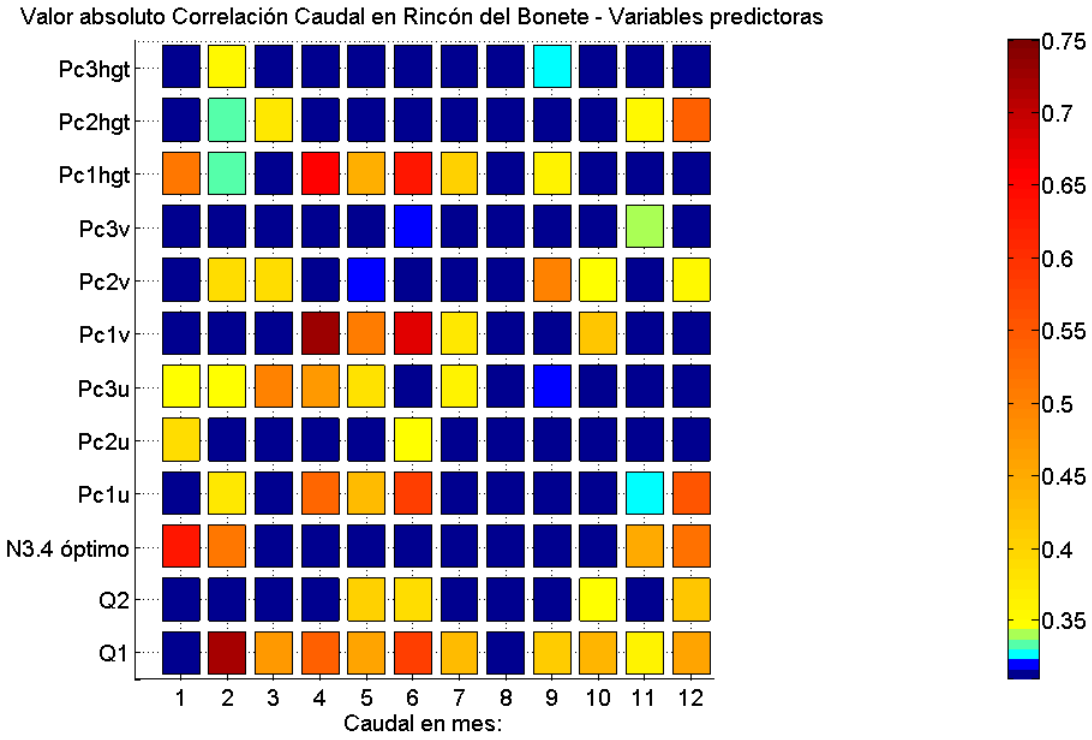

Figura 4.4.1: Valor absoluto de la correlación entre los 12 predictores seleccionados y los caudales mensuales en Rincón del Bonete. En las abscisas se indica el mes del caudal y en las ordenadas la variable con la que se calcula la correlación. Correlaciones no significativas se indican en azul oscuro.

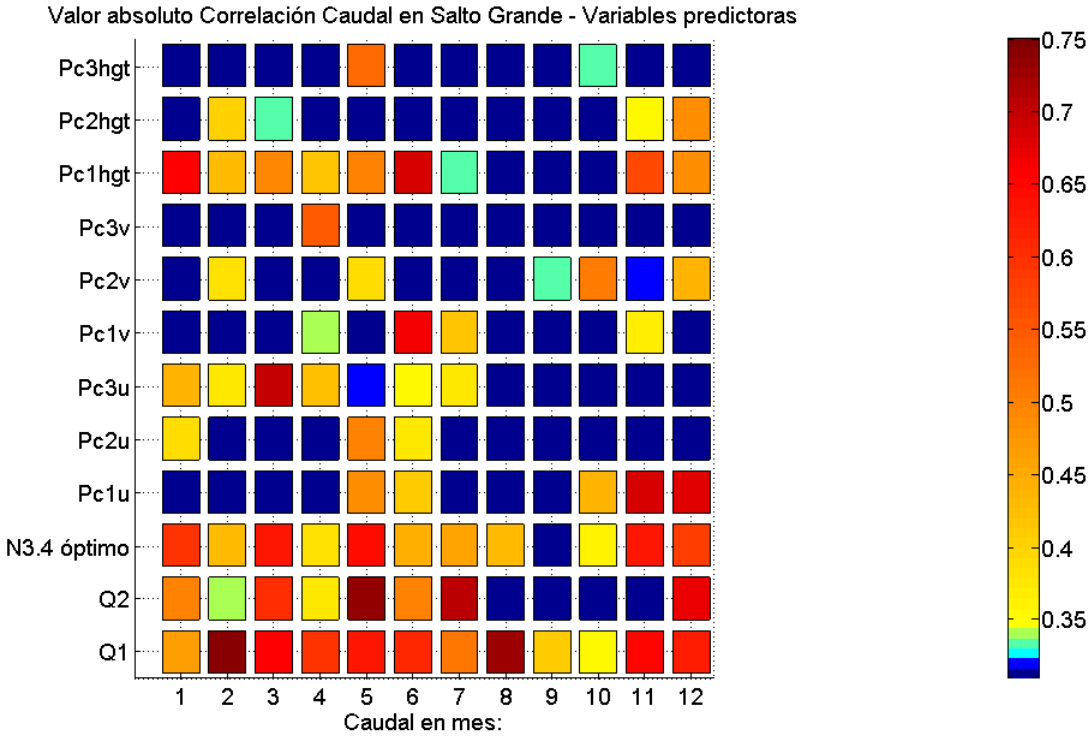

Figura 4.4.2: Idem Figura 4.4.1 para Salto Grande.

5. MODELOS ESTADÍSTICOS DE PREDICCIÓN

El análisis estadístico suele ser utilizado tanto con fines descriptivos como predictivos. Dentro de las actividades descriptivas se encuentran la búsqueda de relaciones, estructuras, tendencias, agrupamientos y patrones no evidentes en los datos en bruto, así como también la identificación de observaciones particularmente anómalas (outliers). Por su parte, cuando el objetivo es predictivo se intenta construir, y evaluar, modelos que permitan predecir el comportamiento de nuevas observaciones. Ambos tipos de actividades pueden ser enmarcadas en los denominados problemas de aprendizaje, ya que se espera generar un cierto aprendizaje a partir de los datos disponibles.

Dentro de los problemas de aprendizaje, a su vez, suele distinguirse entre aprendizaje supervisado o no supervisado.

En la categoría de aprendizaje supervisado se encuentran problemas en los cuales el algoritmo de aprendizaje recibe un conjunto de variables de entrada y una o varias variables de salida conocidas, las cuales pueden ser observadas o provistas por un "maestro" explícito. En estos problemas el algoritmo intenta encontrar una función de las variables de entrada para aproximar las salidas conocidas. Variables de salida continuas dan lugar a los denominados problemas de regresión y variables de salida categóricas a los problemas de clasificación. Dentro de aprendizaje supervisado las distintas técnicas pueden ser categorizadas según su enfoque sea lineal o no lineal, paramétrico o no paramétrico. En las técnicas (no) lineales se asume que la relación entre las variables predictoras y las de salida es (no) lineal. Por su parte, la división en técnicas paramétricas o no paramétricas se genera según la relación entre las variables predictoras y las de respuesta sea conocida a menos de una cantidad finita de parámetros o no.

Dentro de aprendizaje no supervisado se engloban todos aquellos problemas en los cuales la o las variables de salida no son conocidas, es decir que no hay ningún "maestro" explícito. El enfoque en los problemas de aprendizaje no supervisado difiere del análogo de aprendizaje supervisado en que en aprendizaje supervisado se estudian relaciones entre las variables de entrada y salida mientras que en aprendizaje no supervisado las únicas variables a explorar son las de entrada.

En general, las variables de entrada se conocen también como variables explicativas, independientes o predictoras, y las variables de salida se denominan variables de respuesta, dependientes o predictandos.

En este trabajo enfocaremos el problema de predicción de caudales mensuales como un ejercicio de regresión. Para cada mes del año, las variables a predecir son los caudales mensuales circulantes en Rincón del Bonete y Salto Grande. Dado que los factores que influyen en los caudales mensuales podrían ser diferentes para cada uno de los embalses se decide desarrollar, para cada mes, dos problemas de regresión independientes: uno en el que la variable de respuesta es el caudal en Rincón del Bonete y otro en el que la variable de respuesta es el caudal en Salto Grande.

Las técnicas de predicción permiten obtener, en base a la información de las observaciones, modelos que relacionan las variables predictoras con la de respuesta. Estos modelos pueden, luego, ser utilizados para predecir el comportamiento de observaciones futuras. Tan importante como las técnicas que permiten construir los modelos son los mecanismos que permiten evaluar la habilidad predictiva de los mismos.

En este capítulo se desarrollarán, en forma breve, los fundamentos de algunas técnicas estadísticas de regresión que, más tarde, serán aplicadas al problema concreto de la predicción de caudales mensuales utilizando como variables predictoras aquellas indicadas en la sección 4.4. Primero, por ser de fundamental importancia, se desarrollará la noción de error de predicción así como también formas de estimarlo en la práctica. Segundo, se estudiará la técnica de regresión lineal (técnica lineal y paramétrica), junto con procedimientos de selección de variables o generación de nuevas variables que apuntan a mejorar los resultados. Tercero, se presentará la técnica de árboles de regresión (no lineal y no paramétrica). Cuarto, se analizará regresión mediante la utilización de redes neuronales artificiales (no lineal y paramétrica). Por último, se describirá una técnica de predicción basada en conceptos de aprendizaje no supervisado: predicción mediante clustering.

Por desarrollos y discusiones más profundas sobre las técnicas analizadas en este capítulo, así como también muchas otras, y nociones generales de estadística multivariada dirigirse a Izenman (2008), libro en el cual se basa este capítulo.

En todo este capítulo $X_1, \ldots, X_r$ denotan a r variables predictoras, mientras que Y denota a la variable de respuesta.

Un conjunto de m observaciones está dado por las (r+1)- uplas:

$$D = \left\{ (X_1^i, \ldots, X_r^i, Y^i),\ i = 1, \ldots, m \right\}$$

donde: X_j^i denota a la observación i-ésima de la variable X_j
Y^i denota a la observación i-ésima de la variable Y

En la sección de redes neuronales artificiales, dado que no presenta una mayor complejidad, el desarrollo se realiza considerando que pudieran existir varias variables de respuesta, notadas como $Y_1, \ldots, Y_s$.

5.1. Error de predicción

Para medir la precisión de una predicción se utiliza el concepto de error de predicción. En un problema de regresión, para definir el error de predicción de un modelo es necesario considerar observaciones independientes de aquellas utilizadas para desarrollar el modelo.

Para la deducción del modelo se utilizan los datos contenidos en un conjunto de aprendizaje L:

$$L = \left\{ (X_1^i, \ldots, X_r^i, Y^i),\ i = 1, \ldots, n \right\}$$

Mediante alguna técnica de regresión se obtiene un modelo para Y, a partir de las variables $X_1, \ldots, X_r$. Este modelo es: $modelo_L(X_1, \ldots, X_r)$

Supongamos que $(X_1^{nuevo}, \ldots, X_r^{nuevo}, Y^{nuevo})$ denota una observación de las variables $X_1, \ldots, X_r$ e Y que no fue utilizada para la deducción del modelo que se quiere evaluar.

Dados los nuevos valores $(X_1^{nuevo}, \ldots, X_r^{nuevo})$ se utiliza el modelo y se genera una predicción sobre el valor de Y^{nuevo} :

$$Predicción\, de\, Y^{nuevo} = modelo_L(X_1^{nuevo}, \ldots, X_r^{nuevo})$$

La predicción resultante, $Predicción\ de\ Y^{nuevo}$ se compara con el valor efectivamente observado Y^{nuevo}

La habilidad predictiva del modelo de regresión, con X_1, ..., X_r aleatorias, se cuantifica mediante su error de predicción para lo cual, usualmente, se selecciona el error cuadrático:

$$E(Y^{nuevo} - Predicción\ de\ Y^{nuevo})^2$$

Generalmente, si existe suficiente cantidad de datos, el procedimiento más utilizado para estimar el error de predicción consiste en formar, a partir de los mismos, dos conjuntos independientes y disjuntos: un conjunto de aprendizaje y otro de testeo. Los datos contenidos en el conjunto de aprendizaje se utilizan para la construcción del modelo de predicción mientras que aquellos contenidos en el conjunto de testeo son utilizados únicamente para evaluar la habilidad predictiva del modelo desarrollado con los datos del conjunto de aprendizaje. Sin embargo, en casos en los que esta división no sea practicable debido a cantidades restringidas de observaciones suelen utilizarse técnicas alternativas.

La estimación más simple del error de predicción puede obtenerse a través del denominado error de re-sustitución o error aparente. En ésta técnica el conjunto de aprendizaje está formado por la totalidad de las observaciones disponibles, es decir que, todas las observaciones (n) son utilizadas para estimar el modelo. Sea $\tilde{\mu}(X_1, \ldots, X_r)$ el modelo de regresión obtenido.
Luego, el error cuadrático aparente se define como:

$$Error\ cuadrático\ aparente = \frac{1}{n}\sum_{i=1}^{n}(Y^i - \tilde{\mu}(X^i_1, \ldots, X^i_r))^2 = \frac{RSS}{n}$$

RSS es por Residual Sum of Squares.

Esta estimación del error de predicción suele ser demasiado optimista, ya que los mismos datos utilizados para la estimación del modelo son, luego, utilizados para evaluar qué tan bien el modelo encontrado ajusta a los mismos datos, para los que fue optimizado. En otras palabras, el error aparente estima la habilidad predictiva de un modelo re-utilizando los datos con los que éste se entrenó.

La otra gran clase de técnicas de estimación del error de predicción de un modelo con X aleatorio la constituyen las técnicas de re-muestreo, siendo cross-validation (CV) (Stone, 1974) la más popular. Una reseña de métodos de re-muestreo puede encontrarse en Molinaro et al. (2005).

Supóngase que n = Vm, donde V ≥ 2 es un número natural. Se divide, aleatoriamente, el conjunto de datos D en V conjuntos disjuntos Tv, v = 1,...,V de igual tamaño. Luego, se generan V versiones del conjunto de datos: cada versión tiene un conjunto de aprendizaje Lv formado por (V-1) de los V conjuntos y un conjunto de test formado por el conjunto sobrante (Tv = D – Lv). Utilizando únicamente los datos dentro del conjunto de aprendizaje Lv se obtiene la función de regresión $\tilde{\mu_{-V}}(X)$. Luego, se evalúa esta función de regresión en las observaciones pertenecientes al conjunto de test Tv. El error de predicción se estima repitiendo este procedimiento a través de cada una de las parejas de conjunto de aprendizaje (Lv) y test (Tv), para v = 1,...,V. Esta técnica es denominada V-fold CV.

El error cuadrático CV es, luego: $\frac{1}{V}\sum_{v=1}^{V}\sum_{(X^i, Y^i)\in T_v}(Y^i - \tilde{\mu_{-v}}(X^i))^2$

La versión computacionalmente más intensiva de éste método ocurre cuando m = 1 y, por tanto, V = n y se denomina CV leave-one-out. En dicha versión se divide la muestra, de n observaciones, en n conjuntos con 1 observación cada uno. En cada paso se considera como conjunto de aprendizaje al conjunto formado por todas las observaciones menos una. Esta observación, que es dejada fuera, es utilizada para testear el modelo de regresión que se obtiene con las restantes (n-1) observaciones. Este procedimiento se repite alternando la observación dejada fuera del conjunto de aprendizaje.

En este trabajo utilizaremos el procedimiento de CV leave-one-out para la estimación del error de predicción. A efectos de que el error de predicción sea medido en las mismas unidades que la variable que se intenta predecir (caudales, en este caso) se evaluará la raíz cuadrada del error cuadrático CV leave-one-out.

5.2. Regresión Lineal

La técnica de regresión más popular es la de regresión lineal. Dentro de los problemas de regresión lineal las técnicas a utilizar varían según la cantidad de variables predictoras y la cantidad de variables de respuesta. El caso básico trata aquellos problemas en los que existe una variable predictora y una variable respuesta y se denomina regresión simple. En casos en que existen varias variables predictoras pero sólo una de respuesta el problema se denomina regresión múltiple. Por último, cuando el problema involucra varias variables predictoras y, también, varias variables respuesta se denomina regresión multivariada. La diferencia entre las técnicas para tratar problemas de regresión múltiple o multivariada está en que en el caso multivariado no sólo las variables de entrada podrían estar relacionadas entre sí sino que, además, las variables de salida también podrían estarlo.

En este trabajo sólo analizaremos casos de regresión lineal múltiple, por lo que el desarrollo siguiente sólo abarcará este caso.

En el ámbito de la regresión lineal múltiple se asume que la variable de respuesta Y está linealmente relacionada con las variables predictoras $X_1, ..., X_r$ de la forma:

$$Y = b_o+b_1X_1+ ... +b_rX_r + e$$

donde e, término de error en el modelo, es una variable aleatoria no observable (con media 0 y varianza s^2) y $b_o, ..., b_r$ son los (r+1) parámetros desconocidos a determinar. La linealidad del modelo es consecuencia de la linealidad en los parámetros $b_o, b_1, ..., b_r$. Por lo tanto, transformaciones de las variables predictoras (tales como potencias X^m y productos X_iX_j) pueden ser introducidas.

La aplicación de la técnica de regresión lineal múltiple requiere determinar los parámetros $b_o, ..., b_r$ y s^2 así como también evaluar el impacto de cada una de las variables predictoras sobre el comportamiento de Y.

Sea X el vector formado por las r variables predictoras: $X = (X_1, ..., X_r)^t$.
Se supone que X e Y son variables aleatorias con distribución conjunta P(X,Y), con medias $E(X) = \mu_X$, $E(Y) = \mu_Y$ y matrices de covarianzas Σ_{XX}, $\Sigma_{YY} = s_Y^2$ y Σ_{XY}.

El problema consiste en predecir Y a través de una función de X: f(X). La precisión del modelo de predicción se mide mediante la función real de pérdida: L(Y,f(X)), la cual estima la pérdida incurrida al predecir Y por f(X). La pérdida esperada es la denominada función de riesgo:

$$R(f)=E(L(Y,f(X)))$$

Si como medida de pérdida se utiliza el error cuadrático R(f) se transforma en:

$$R(f)=E(Y-f(X))^2=E_X(E_{Y|X}(Y-f(X))^2|X)$$

Luego, R(f) puede ser minimizada punto a punto (en cada x).

Si se define $\mu(x)=E_{Y|X}(Y|X=x)$ puede escribirse: $Y-f(x)=(Y-\mu(x))+(\mu(x)-f(x))$

Elevando al cuadrado ambos miembros de la ecuación anterior, y tomando esperanza condicional se tiene:

$$E_{Y|X}((Y-f(x))^2|X=x)=E_{Y|X}((Y-\mu(x))^2|X=x)+(\mu(x)-f(x))^2 \quad (1)$$

donde el producto cruzado se anula porque $E_{Y|X}(Y-\mu(x)|X=x)=0$.

Por lo tanto, (1) es minimizada, respecto de f, tomando: $\tilde{f}(x)=\mu(x)=E_{Y|X}(Y|X=x)$ de donde:

$$E_{Y|X}((Y-\tilde{f}(x))^2|X=x)=E_{Y|X}((Y-\mu(x))^2|X=x)$$

En consecuencia, el mínimo de la función de riesgo respecto de f es: $min\,R(f)=E((Y-\mu(X))^2)$ y el mejor predictor de Y en X=x es μ(x).

Más específicamente, si se asume que el término del error e no está correlacionado con las variables predictoras, entonces μ(X) está dado por:

$$\mu(X)=b_0+\sum_{i=1}^{r}b_iX_i=b_0+X^tb=Z^t\alpha$$

donde $b=(b_1,\ldots,b_r)^t$, $\alpha=(b_0,b_1,\ldots,b_r)^t$ y $Z=(1,X_1,\ldots,X_r)^t$.

Se deben elegir b_0 y b para minimizar la función objetivo: $S(\alpha)=E((Y-Z^t\alpha)^2)$

Por tanto, derivando respecto de α e igualando a cero se obtienen los coeficientes buscados (los cuales se identifican mediante un tilde):

$$\frac{\partial S(\alpha)}{\partial\alpha}=-2E(ZY-ZZ^t\alpha)=0 \qquad \tilde{\alpha}=[E(ZZ^t)]^{-1}E(ZY)$$

Re- escribiendo, según las definiciones de α, Z e Y, el vector de coeficientes buscado puede expresarse como producto de dos términos (1 y 2):

$$\tilde{\alpha}=\begin{bmatrix}\tilde{b_0}\\ \tilde{b_1}\\ \vdots\\ \tilde{b_r}\end{bmatrix}=E\left(\begin{bmatrix}1\\ X_1\\ \vdots\\ X_r\end{bmatrix}\begin{bmatrix}1 & X_1 & \ldots & X_r\end{bmatrix}\right)^{-1}E\left(\begin{bmatrix}1\\ X_1\\ \vdots\\ X_r\end{bmatrix}Y\right)=\underbrace{E\begin{bmatrix}1 & X_1 & \ldots & X_r\\ X_1 & X_1^2 & \ldots & X_1X_r\\ \vdots & \vdots & \vdots & \vdots\\ X_r & X_1X_r & \ldots & X_r^2\end{bmatrix}^{-1}}_{2}\underbrace{E\begin{bmatrix}Y\\ X_1Y\\ \vdots\\ X_rY\end{bmatrix}}_{1}$$

Desarrollando el término 1:

$$\begin{bmatrix} E(Y) \\ E(X_1 Y) \\ \vdots \\ E(X_r Y) \end{bmatrix} = \begin{bmatrix} \mu_Y \\ cov(X_1, Y) + E(X_1)E(Y) \\ \vdots \\ cov(X_r, Y) + E(X_r)E(Y) \end{bmatrix} = \begin{bmatrix} \mu_Y \\ cov(X_1, Y) + \mu_{X_1}\mu_Y \\ \vdots \\ cov(X_r, Y) + \mu_{X_r}\mu_Y \end{bmatrix}$$

Para simplificar la expresión del término 2, se utiliza la siguiente fórmula de inversión de matrices en bloque en donde las matrices A y D deben ser simétricas y, adicionalmente, la matriz A debe ser invertible.

$$\begin{bmatrix} A & B \\ B^t & D \end{bmatrix}^{-1} = \begin{bmatrix} A^{-1} + FE^{-1}F^t & -FE^{-1} \\ -EF^t & E^{-1} \end{bmatrix} \text{ donde } E = D - B^t A^{-1} B, \quad F = A^{-1} B$$

Para el caso del término 2, A=1, por lo que operando se obtiene:

$$F = B = (E(X_1, \ldots, X_r)) = \mu_X^t$$

$$E = \begin{bmatrix} E(X_1^2) & \ldots & E(X_1 X_r) \\ \vdots & \vdots & \vdots \\ E(X_1 X_r) & \ldots & E(X_r^2) \end{bmatrix} - \begin{bmatrix} E(X_1) \\ \vdots \\ E(X_r) \end{bmatrix} [E(X_1), \ldots, E(X_r)] = \Sigma_{XX}$$

Finalmente, los coeficientes buscados se obtienen multiplicando las expresiones de los términos 1 y 2:

$$\begin{bmatrix} \tilde{b}_0 \\ \tilde{b}_1 \\ \vdots \\ \tilde{b}_r \end{bmatrix} = \begin{bmatrix} 1 + \mu_X^t \Sigma_{XX}^{-1} \mu_X & -\mu_X^t \Sigma_{XX}^{-1} \\ -\Sigma_{XX} \mu_X & \Sigma_{XX}^{-1} \end{bmatrix} \begin{bmatrix} \mu_Y \\ \Sigma_{XY} + \mu_X \mu_Y \end{bmatrix}$$

$$\tilde{b}_0 = \mu_Y + \mu_X^t \Sigma_{XX}^{-1} \mu_X \mu_Y - \mu_X^t \Sigma_{XX}^{-1} (\Sigma_{XY} + \mu_x \mu_Y) = \mu_Y - \mu_X^t \Sigma_{XX}^{-1} \Sigma_{XY}$$

$$\begin{bmatrix} \tilde{b}_1 \\ \vdots \\ \tilde{b}_r \end{bmatrix} = -\Sigma_{XX} \mu_X \mu_Y + \Sigma_{XX}^{-1} (\Sigma_{XY} + \mu_X \mu_Y) = \Sigma_{XX}^{-1} \Sigma_{XY}$$

En la práctica, μ_X, μ_Y, Σ_{XX} y Σ_{XY} no son conocidos, por lo cual es necesario estimarlos a partir de los datos disponibles.

Sea $D = \{(X_1^i, \ldots, X_r^i, Y^i),\ i = 1, \ldots, n\}$ el conjunto de todas las observaciones disponibles en donde X_j^i es la observación i-ésima de la variable X_j e Y^i es la i-ésima observación de la variable Y.

Se define $X^i = (X_1^i, \ldots, X_r^i)^t$ como el vector formado por las observaciones i-ésimas de las variables $X_1, \ldots, X_r$.

Se define el vector $Y = (Y^1, \ldots, Y^n)^t$ como el vector formado por las n observaciones de la variable Y.

Luego, μ_X se estima por $\bar{X} = \frac{1}{n}\sum_{j=1}^{n} X^j$ y μ_y por $\bar{Y} = \frac{1}{n}\sum_{j=1}^{n} Y^j$

Sean $\bar{X} = (\bar{X}, \ldots, \bar{X})^t$ una matriz (n x r) y $\bar{Y} = (\bar{Y}, \ldots, \bar{Y})^t$.

Se define la matriz (n x r), $X = \begin{bmatrix} X_1^1 & \ldots & X_r^1 \\ \vdots & \vdots & \vdots \\ X_1^n & \ldots & X_r^n \end{bmatrix}$, en donde el elemento en la posición (i,j) es la observación i-ésima de la variable X_j.

Se definen las variables centradas: $X_c = X - \bar{X}$ y $Y_c = Y - \bar{Y}$

Y, finalmente, se estiman las matrices de covarianzas mediante:

$$\Sigma_{XX} \sim \frac{1}{n} X_c^t X_c \qquad y \qquad \Sigma_{XY} \sim \frac{1}{n} X_c^t Y_c$$

5.3. Inestabilidad de estimaciones en regresión lineal múltiple

Dado que las estimaciones de los coeficientes {β} dependen de $(X_c^t X_c)^{-1}$, en casos en los que $X_c^t X_c$ sea no invertible, o casi no invertible, se experimentarán complicaciones numéricas. Tales situaciones pueden ocurrir cuando la matriz X_c esté mal condicionada, sus columnas sean co-lineales (o casi) o ante situaciones en las que existan más variables que observaciones (r > n).

Para un cierto problema, los datos están mal condicionados cuando las cantidades a ser calculadas son sensibles a pequeños cambios en los datos. Cuando ese es el caso los resultados computacionales, especialmente aquellos que involucren inversión de matrices, pueden resultar numéricamente inestables.

En modelos de regresión lineal los problemas ocasionados por datos mal condicionados o por variables casi co-lineales coinciden. Si X_c está mal condicionada pequeños cambios en sus elementos llevarán a grandes cambios en $(X_c^t X_c)^{-1}$. Luego, las estimaciones de los coeficientes serán numéricamente inestables y las mismas podrían tener magnitudes muy grandes o hasta con signo equivocado. Si las variables son casi co-lineales se genera el mismo problema ante la inversión de matrices. Como consecuencia, aunque el modelo de regresión pudiera ser un buen ajuste a los datos en el conjunto de aprendizaje, no se tendrá un buen desempeño al ser enfrentado con nuevos datos.

Existen varias maneras de tratar con este tipo de situaciones: 1. Eliminación de variables predictoras mediante algún procedimiento se selección de variables. 2. Construcción de nuevas variables no correlacionadas, a partir de las variables originales. 3. Técnicas de estimación de coeficientes sesgadas.

A continuación brindaremos una breve descripción de algunos procedimientos de selección de variables y como ejemplo de técnica de regresión basada en la construcción de nuevas variables la denominada Regresión por mínimos cuadrados parciales.

5.3.1. Selección de variables en el contexto de regresión lineal múltiple

En la ecuación de regresión se pueden incluir tantas variables predictoras como se desee. Si el número de predictores es muy elevado, la cantidad de parámetros a determinar también lo será. Por otro lado, si muy pocos predictores se incluyen en el modelo la función de regresión corre el riesgo de generar una pobre explicación de los datos. La noción de qué características hacen que una variable sea importante no es aún muy claro, pero una interpretación es que una variable se vuelve importante si su exclusión del modelo afecta seriamente la precisión de las predicciones (Izenman, 2008 y referencias allí).

La idea de la selección de variables responde a la necesidad de la obtención de un modelo de regresión simple que garantice una buena habilidad predictiva. Existen varios procedimientos de selección de variables en problemas de regresión entre los que se destacan los métodos por pasos (stepwise), el de todos los conjuntos posibles y LASSO. En este trabajo sólo utilizaremos métodos por pasos.

De entre los métodos por pasos los dos tipos principales de técnicas son: eliminación hacia atrás (Backward Elimination) o selección hacia adelante (Forward Selection).

La técnica de Backward Elimination comienza con el conjunto entero de variables. En cada paso se elimina del modelo la variable cuyo índice F sea más pequeño. El índice F se define como:

$$índice\ F = \frac{(RSS_0 - RSS_1)/(df_1 - df_0)}{RSS_1/df_1}$$

donde RSS_0 es RSS para el modelo reducido, RSS_1 es RSS para el modelo con mayor cantidad de variables. df_1-df_0 = 1 y df_1 = n-k-1 siendo k la cantidad de variables en el modelo con mayor cantidad. Luego, se re-ajusta el modelo eliminando alguna variable y se repite el procedimiento. Una variante de éste método es la denominada Backward stepwise que consiste en cada paso poder volver a incluir una variable antes eliminada.

La técnica de Forward Selection comienza con un conjunto vacío de variables. En cada paso se selecciona la variable con el mayor índice F (df_1-df_0 = 1 y df_1 = n-k-2 donde k es el número de variables en el conjunto más pequeño). Se agrega esa variable al modelo de regresión, se re-ajusta el mismo y se repite el procedimiento.

Los procedimientos de selección de variables por pasos son usualmente criticados porque no existen garantías de que los modelos seleccionados con uno u otro tipo de esquema (backward o forward) conduzcan al mismo conjunto de variables, ni al "mejor conjunto de variables" en algún sentido. Más en general, los métodos de selección de variables son criticados porque utilizan los propios datos para agregar o eliminar variables y, por lo tanto, cambian el modelo que se asumía tenía variables predictoras determinadas a priori. Esto podría implicar que si los datos cambian levemente, las variables seleccionadas podrían también cambiar, haciendo a estos procedimientos

muy inestables (Izenman, 2008 y referencias allí).

5.3.2. Regresión por mínimos cuadrados parciales

La técnica de regresión por mínimos cuadrados parciales es conocida como PLSR (Partial Least-Squares Regression). En PLSR nuevas variables (conocidas como variables latentes) se construyen especialmente para retener la mayor cantidad de información de las variables X_1, …, X_r que ayuda a predecir la variable Y reduciendo, al mismo tiempo, la dimensionalidad del problema de regresión. PLSR utiliza información proveniente tanto de las variables X_1, …, X_r como de la variable respuesta Y. En general, PLSR se obtiene como el resultado de la utilización de un algoritmo en lugar de un procedimiento de optimización.

El algoritmo más popular (Wold et al., 1983) comienza con un conjunto vacío de variables y agrega una variable latente en cada uno de los pasos sucesivos. En el paso k-ésimo la variable latente Z_k es un promedio ponderado de los residuos en las variables X_1, …, X_r del paso anterior. Los pesos son proporcionales a la covarianza entre los residuos de las variables X_1, …, X_r y los residuos de la variable Y, del paso anterior.

El algoritmo está formado por los tres pasos que se describen a continuación:

Paso 1

Sea $vect\ X_j = (X_j^1, ..., X_j^n)^t$ el vector formado por las n observaciones de la variable X_j .

Considero que el vector está estandarizado y tiene media 0.

Sea $vect\ Y = (Y^1, ..., Y^n)^t$ el vector formado por las n observaciones de la variable de respuesta Y.

Defino $vect\ X_j^{(0)} = vect\ X_j\ ,\ j = 1, ..., r$

$vect\ Y^{(0)} = vect\ Y \qquad predicción\ Y^{(0)} = \bar{Y}\ Id_{nx1}$ donde la barra denota el promedio.

Paso 2

Para k=1,..., t

2.1 Realizo la regresión de $vect\ Y^{(k-1)}$ en función de $vect\ X_j^{(k-1)}$, de donde se obtienen los coeficientes $\widetilde{b_j^{(k-1)}} = \dfrac{cov(X_j^{(k-1)}, Y^{(k-1)})}{var(X_j^{(k-1)})}$

2.2 Defino $Z_k = \sum_{j=1}^{r} \widetilde{b_j^{(k-1)}}\ vect\ X_j^{(k-1)}\ w^{(k-1)}$

2.3 Calculo la regresión de $vect\ Y^{(k-1)}$ en función de Z_k , de donde se obtiene el coeficiente $\widetilde{\theta_k} = \dfrac{cov(Z_k, vect\ Y^{(k-1)})}{var(Z_k)}$

2.4 Calculo el residuo $vect\ Y^{(k)} = vect\ Y^{(k-1)} - \widetilde{\theta_k} Z_k$

2.5 Para j=1,...,r

Calculo la regresión de $vect\ X_j^{(k-1)}$ en función de Z_k , de donde se obtiene el coeficiente $\widetilde{\phi_{kj}} = \dfrac{cov(Z_k, vect\ X_j^{(k-1)})}{var(Z_k)}$

2.6 Calculo el residuo $vect\ X_j^{(k)} = vect\ X_j^{(k-1)} - \widetilde{\phi_{jk}} Z_k$

2.7 Parar cuando $\sum_{j=1}^{r} var(vect\ X_j^{(k)}) = 0$

Paso 3
Predicción de $vect\ Y$ truncando el análisis con t componentes:

$$predicción\ Y^{(t)} = \bar{Y}\ Id_{nx1} - \sum_{k=1}^{t} \widetilde{\theta_k} Z_k$$

5.4. Árboles de clasificación o regresión

El método de árboles de clasificación o regresión (CART por su sigla en inglés) se basa en el algoritmo de particionamiento recursivo. El particionamiento recursivo es un proceso secuencial en el que un árbol de decisiones es construido separando o no cada nodo del árbol en dos nodos hijos. La metodología CART tiene su principal atractivo en que, debido a que el algoritmo realiza una serie de preguntas booleanas (por ejemplo, ¿es cierta variable menor a cierto valor?), la interpretación y análisis de algunos resultados puede resultar relativamente simple.

En la metodología CART el espacio $\mathbb{R}^r$ se particiona en regiones rectangulares (si r = 2) o cuboides (si r > 2) disjuntas, cada una de las cuales se considera como homogénea a los efectos de predecir una única variable de salida Y: a cada región se le asigna un valor constante para Y.

Un árbol se obtiene como resultado de realizar una sucesión ordenada de preguntas, el tipo de pregunta a realizar en cada paso depende de las respuestas a las preguntas previas de la sucesión. La sucesión finaliza con la predicción de una clase, en caso de clasificación, o con la predicción de un valor, en caso de regresión.

El único punto de entrada a un árbol se denomina raíz, y consiste en el conjunto completo de los datos de aprendizaje L. Un nodo está formado por un subconjunto de los datos de L y puede ser terminal o no. Un nodo no terminal se denomina nodo padre y es un nodo que se divide en 2 nodos hijos (es, por tanto, una división binaria). Una tal división binaria es determinada por una condición en el valor de una única variable, en donde la condición es o bien satisfecha o bien no satisfecha por el valor observado de dicha variable. Todas las observaciones en L que han alcanzado un nodo padre en particular pasan, luego, a un nodo hijo si satisfacen la condición o al otro si no lo hacen.

Un nodo que no se divide se denomina nodo terminal y al mismo se la asigna una etiqueta de clase (clasificación) o un valor (regresión) para la variable respuesta. Cada una de las observaciones en L

finaliza en alguno de los nodos terminales. Cuando una observación recorre el árbol y finaliza en un cierto nodo terminal, se le asigna la clase o valor de la variable de respuesta que dicho nodo terminal indique.

Para construir un árbol, en función de las observaciones disponibles en un conjunto de aprendizaje L, es necesario considerar 3 aspectos: 1. Selección de los criterios de división de nodos; 2. Determinación de la condición de parada del algoritmo (es decir, se debe definir cuándo un nodo se vuelve terminal); 3. Selección del criterio de asignación de clases (clasificación) o valores (regresión) para la variable de respuesta en los nodos terminales.

En lo que sigue sólo discutiremos el caso de árboles de regresión.

Supongamos que las observaciones están dadas por $D=\{(X^i,Y^i), i=1,..,n\}$ donde Y^i es la observación i-esima de la variable continua de salida Y y X^i el vector formado por las observaciones i-esimas de las r variables predictoras $X^i=(X_1^i, \dots, X_r^i)$. Se asume que Y está relacionado con $X=(X_1,...,X_r)$ y se desea utilizar la técnica de árboles de regresión para predecir dicha relación.

El procedimiento para la construcción de un árbol de regresión se conoce como regresión por particionamiento recursivo. En un árbol de regresión, la variable de salida en un nodo terminal τ es definida para tener un valor constante $\bar{Y}(\tau)$.

Este valor constante $\bar{Y}(\tau)$ se calcula como el promedio de la variable de salida Y en las observaciones pertenecientes al nodo τ, es decir:

$$\bar{Y}(\tau) = \frac{1}{n_\tau} \sum_{X^i \in \tau} Y^i$$

en donde n_τ es la cantidad de observaciones en el nodo τ.

El error aparente de un árbol T se define como la suma de los errores de re-sustitución en los distintos nodos terminales del árbol, es decir:

$$error\ aparente(T) = \sum_{i=1}^{t} e_{\Re e}(\tau_{terminal\,i}) = \sum_{i=1}^{t} \frac{1}{n} \sum_{j\,/\,X^j \in \tau_{terminal\,i}} (Y^i - \bar{Y}(\tau_{terminal\,i}))^2$$

en donde $e_{\Re e}$ denota error de re-sustitución y $\tau_{terminal\,i}$ indica al nodo terminal i-ésimo.

Como estrategia de división en un nodo dado se utiliza el criterio de seleccionar aquella división (de entre todas las posibles, al considerar todas las variables predictoras) que provoque una mayor reducción en el error aparente de árbol. Esto es, si τ_i y τ_d son los dos nodos hijos del nodo τ, se debe maximizar:

$$\Delta e_{\Re e}(\tau) = e_{\Re e}(\tau) - e_{\Re e}(\tau_i) - e_{\Re e}(\tau_d)$$

Cabe notar que si las variables predictoras son continuas entonces, para cada variable, el número de posibles divisiones en un nodo es uno menos que el número de valores observados distintos. Es decir, si una variable toma 3 valores distintos hay dos separaciones posibles a hacer entre los 3 valores. Para construir el árbol es necesario evaluar todas las posibles divisiones en todas las variables predictoras e identificar la mejor en el sentido mencionado antes.

El procedimiento para construir el árbol comienza en el nodo raíz, el cual consta de todas las observaciones en el conjunto L. Utilizando el criterio de minimización del error aparente, el algoritmo encuentra la mejor división en el nodo raíz. Luego de dividido el nodo raíz, se utiliza el mismo procedimiento para dividir los 2 nodos hijos con la diferencia que las únicas observaciones a utilizar en cada nodo son aquellas que verifican la condición para pertenecer al mismo. Este procedimiento de división secuencial se denomina particionamiento recursivo.

Si el procedimiento se continúa hasta que ninguno de los nodos puede ser dividido (es decir, existe sólo una observación en cada nodo a dividir) se dice que el árbol está saturado. En consecuencia, en árboles saturados el error aparente del modelo será 0 y el ajuste, en este sentido, será perfecto. Modelos de regresión generados con árboles saturados tendrán, posiblemente, grandes errores de predicción debido a un sobre-ajuste a los datos contenidos en el conjunto de aprendizaje. Una manera de evitar este tipo de situaciones es poner condiciones para restringir el crecimiento del árbol. Una de estas condiciones puede ser declarar a un nodo como terminal en el caso que la cantidad de observaciones en el mismo sea menor que un cierto umbral predeterminado; cuanto mayor sea este valor umbral más severa será la condición. Otra condición utilizada es la declaración de que un nodo es terminal si la reducción en el error aparente del árbol es menor que un límite también predeterminado. Otra técnica, que no analizaremos en este trabajo, con un punto de vista diferente, es dejar crecer el árbol hasta saturación y luego podarlo.

En la Figura 5.4.1 se presenta un esquema de un árbol de regresión en el que se tienen 5 nodos terminales. En cada nodo terminal el valor de la variable respuesta Y es constante.

Breiman et al. (1984) mencionan que la precisión de los modelos de regresión obtenidos mediante la técnica de árboles de regresión ha demostrado ser comparable con la de los modelos de regresión lineal, aunque observan que en problemas con fuerte estructura lineal el desempeño de los modelos lineales tiende a ser superior.

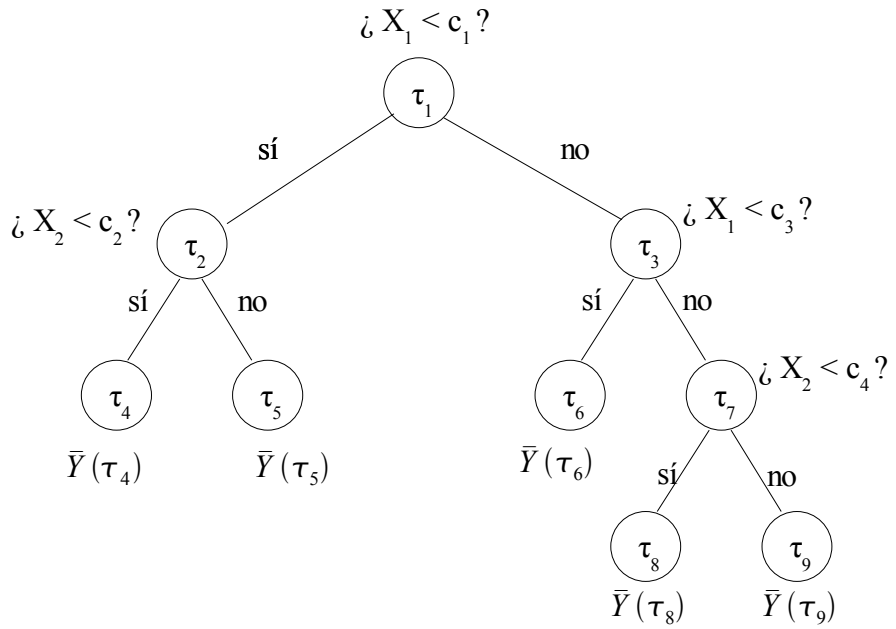

Figura 5.4.1: Esquema de árbol de regresión. X_1 y X_2 son las variables predictoras, Y la variable de respuesta y c_1, c_2, c_3 y c_4 números reales.

5.5. Redes neuronales artificiales

Las redes neuronales artificiales, o simplemente redes neuronales, surgen como un intento para modelar la actividad cerebral. Hoy en día las mismas son tratadas de forma más abstracta: redes que interconectan, de forma no lineal, elementos de cálculo. Son modelos paramétricos que pueden ser utilizados para aproximar cualquier tipo de relación funcional entre variables predictoras y de respuesta. A diferencia de lo que sucede con, por ejemplo, los modelos lineales no es necesario pre-definir el tipo de relación entre las variables a relacionar. Los datos observados se utilizan para entrenar la red, la cual aprende a aproximar la relación entre predictores y respuestas adaptando, de forma iterativa, sus parámetros. De entre los problemas en los que suelen aplicarse redes neuronales destacan el reconocimiento de patrones y problemas de predicción, especialmente casos con alta dimensionalidad y grandes cantidades de datos.

Una red neuronal multi-capa, conocida como perceptrón, es una técnica estadística multivariada que relaciona, de manera no lineal, una serie de variables de entrada $X_1, \ldots, X_r$ con una serie de variables de salida $Y_1, \ldots, Y_s$. Entre las variables de entrada y salida se generan unas variables ocultas, ordenadas en estructuras de forma de capa. Las variables de entrada, las ocultas y las de salida se denominan nodos, neuronas o unidades de cálculo.

Las redes neuronales pueden ser utilizadas para modelar problemas de regresión o clasificación.

Los nodos de los perceptrones multi-capa se organizan en tres grupos: nodos de entrada, nodos ocultos (ubicados en una o más capas) y nodos de salida. Los nodos de entrada están formados por las r variables de entrada: $X_1, \ldots, X_r$. Los nodos ocultos son nodos que se ubican entre los de entrada y salida, en distintas capas denominadas capas ocultas. Los nodos de salida están formados por las s variables de salida: $Y_1, \ldots, Y_s$. Los nodos de entrada y salida pueden tomar valores reales o discretos. Los nodos de salida que tomen valores reales son usualmente re-escalados para proveer salidas en el intervalo [0,1].

Los distintos nodos de la red están conectados. Una red completamente conectada tiene todos sus r nodos de entrada conectados a todos los nodos en la primera capa oculta, todos los nodos en la primer capa oculta conectados a todos los de la segunda capa oculta y así sucesivamente hasta tener todos los nodos de la última capa oculta conectados con todos los nodos de salida. Si algunas de las conexiones no existen, entonces, se tiene una red parcialmente conectada. Cada conexión entre nodos tiene asociado un peso β, el cual identifica la intensidad de la conexión. Los pesos pueden ser positivos, cero o negativos según representen señales excitatorias, inexistentes o inhibitorias. La arquitectura de una red involucra a los nodos, las conexiones entre nodos y los pesos de las conexiones.

Los nodos de entrada no realizan cálculos; simplemente toman valores introducidos por algún agente externo a la red. Dados los valores que se le ingresen a cada nodo oculto, éste computa un valor de activación calculando un promedio ponderado de los valores ingresados y sumando una constante. De manera similar, cada nodo de salida computa un valor de activación a partir del promedio ponderado de los valores que se le ingresen, los cuales provienen de los nodos de la última capa oculta, y suma una constante. Los valores de activación son, luego, filtrados a través de una función de activación para formar el valor salida de un nodo.

A modo de ejemplo consideremos una red neuronal con r nodos de entrada (X_m, m = 1,...,r), una capa oculta con t nodos ocultos (Z_j, j = 1,...,t) y s nodos de salida (Y_k, k = 1,...,s). Sea β_{mj} el peso de

la conexión entre los nodos X_m y Z_j y α_{jk} el peso de la conexión entre los nodos Z_j e Y_k. Se consideran conexiones β_{0j} y α_{0k} como sesgos (o sea, las constantes que los nodos agregan al promedio ponderado de los datos de entrada que se les ingresa). Sean $\{f_j, j = 1,...,t\}$ y $\{g_k, k = 1,...,s\}$ las familias de funciones de activación para los nodos ocultos y de salida, respectivamente. En la Figura 5.5.1 se aprecia un esquema de esta red, para el caso particular en que r = 3, t = 2 y s = 2.

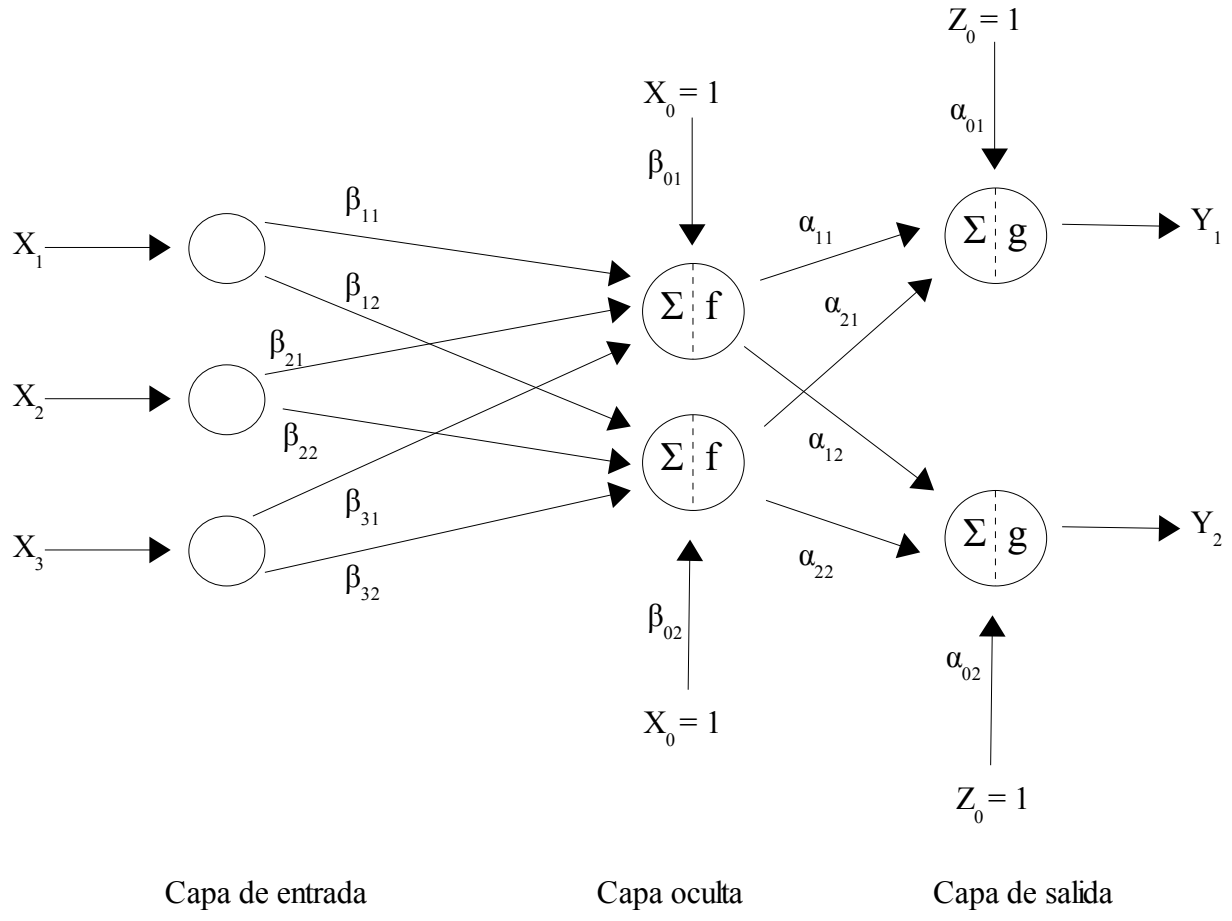

Figura 5.5.1: Esquema de un perceptrón con 3 nodos de entrada, una capa oculta con 2 nodos ocultos y 2 nodos de salida. Los coeficientes α y β indican los pesos de las diferentes conexiones y las funciones f y g representan las funciones de activación.

Sean $X = (X_1, ..., X_r)^t \quad Z = (Z_1, ..., Z_t)^t$

Sea $U_j = \beta_{0j} + X^t \beta_j, \quad V_k = \alpha_{ok} + Z^t \alpha_k$ donde $\beta_j = (\beta_{1j}, ..., \beta_{rj})^t, \quad \alpha_k = (\alpha_{1k}, ..., \alpha_{tk})^t$

Luego, $Z_j = f_j(U_j),\ j=1,...,t \quad \mu_k(X) = g_k(V_k),\ k=1,...,s$

Finalmente, la forma explícita del valor del k-ésimo nodo de salida Y_k es: $Y_k = \mu_k(X) + \epsilon_k$ donde para k=1,...,s

$$\mu_k(X) = g_k(\alpha_{0k} + \sum_{j=1}^{t} \alpha_{jk} f_j(\beta_{oj} + \sum_{m=1}^{t} \beta_{mj} X_m))$$

Como funciones $\{f_j\}$, usualmente, se consideran funciones sigmoidales (tienden a 0 en $-\infty$ y a 1 en $+\infty$). Por su parte, en problemas de regresión, como funciones $\{g_k\}$ suelen considerarse funciones lineales.

Si en el ejemplo anterior se toma s = 1, de modo de tener una única variable de respuesta, y se supone que los nodos ocultos tienen asociada la misma función de activación sigmoidal σ y que el nodo de salida tiene asociada una función g lineal, entonces la variable respuesta Y adquiere la forma:

$$Y = \mu(X) + \epsilon$$

$$\mu(X) = \alpha_0 + \sum_{j=1}^{t} \alpha_j \sigma(\beta_{0j} + \sum_{m=1}^{r} \beta_{mj} X_m) \quad (*)$$

Es importante notar que las funciones sigmoidales juegan un rol importante en el diseño de una red, ya que son flexibles como funciones de activación y pueden aproximar muchos otros tipos de funciones. Un resultado utilizado para motivar el uso de redes neuronales es el Teorema de Aproximación Universal de Kolmogorov, el cual expresa que: Cualquier función continua definida en un compacto de $\mathbb{R}^r$ puede ser uniformemente aproximada por una función de la forma (*). El teorema no especifica cómo encontrar la aproximación, es decir, cómo determinar los pesos de las conexiones ni la cantidad de nodos ocultos t. También se está en el contexto en el que se conoce la función continua a aproximar y que existe una cantidad arbitraria de nodos ocultos disponible.

Sea w el vector de tamaño st+rt+t+s, formado por todos los parámetros desconocidos de una red neuronal completamente conectada con r nodos de entrada, t nodos ocultos (en una capa oculta) y s nodos de salida. El vector w está formado, entonces, por los pesos de conexión y los sesgos. Para estimar los citados parámetros, en regresión, es usual minimizar la suma de los errores cuadráticos:

$$Suma\ Errores\ Cuadráticos = \sum_{i=1}^{n} \|Y_i - \tilde{Y}_i\|^2 = \sum_{i=1}^{n} \sum_{k \in Salida} (Y_k^i - \tilde{Y}_k^i)^2$$

donde Salida denota al conjunto de los nodos de salida, Y_k^i es la observación i-esima de la variable Y_k, $\tilde{Y}_k^i$ es el valor de salida predicho por la red al tomar como valores de entrada las observaciones i-ésimas de las variables $X_1, \ldots, X_r$:

$$\tilde{Y}_k^i = \mu_k(X_1^i, \ldots, X_r^i) = \mu_k(X_1^i, \ldots, X_r^i, w)$$

Debido a que $\tilde{Y}_k^i$ es una función no lineal de w, el criterio de minimización de la Suma de Errores Cuadráticos es una función no lineal de w. El vector w que minimiza la Suma de los Errores Cuadráticos no puede obtenerse de forma explícita y, por lo tanto, debe ser estimado utilizando algún tipo de algoritmo de optimización no lineal. El método numérico más utilizado para la estimación de los parámetros de una red neuronal es el algoritmo de propagación de errores hacia atrás (backpropagation of errors, en inglés).

El algoritmo de propagación de errores hacia atrás (Werbos, 1974) calcula las derivadas de la función de error, respecto de los parámetros de la red: $\{\alpha_{kj}\}$, $\{\beta_{jm}\}$. Luego, estas derivadas son

utilizadas para estimar los pesos de las conexiones minimizando la función de error, a través de un método iterativo de descenso por la dirección contraria a la del gradiente que se continúa hasta la identificación de un mínimo local.

Utilizando valores elegidos al azar para inicializar los pesos de conexión, se busca la dirección en la cual el error se disminuye. Las fórmulas de actualización en el algoritmo identifican dos etapas en los cálculos: una etapa hacia adelante y otra hacia atrás. Luego del paso inicial en el que se les asignan valores arbitrarios a los pesos de conexión se siguen las etapas:

- Etapa hacia adelante: Los valores de las variables predictoras para las observaciones en el conjunto de aprendizaje entran en la red, se siguen las conexiones, los nodos de cálculo calculan y se obtienen ciertos valores para las variables de salida, con los pesos de conexión actuales. Se calculan las sumas de errores cuadráticos, que miden la diferencia entre los valores de salida obtenidos por la red y los valores de salida verdaderos (es decir, los observados).
- Etapa hacia atrás: Se actualizan los valores de los pesos de conexión de la capa oculta a la capa de salida (α). Luego, se actualizan los pesos de conexión de la capa de entrada a la de salida (β).

En conclusión, en cada iteración los pesos de conexión se modifican de modo de acercar los valores respuesta de la red a los verdaderos valores respuesta. El proceso iterativo se continúa hasta que algún criterio pre-especificado se cumple. Una descripción más detallada de este algoritmo puede ser encontrada en Izenman (2008).

No existe prueba alguna que demuestre que el algoritmo de propagación hacia atrás de los errores siempre converge. De hecho hay evidencia que muestra que, en la práctica, la convergencia no está asegurada; esto se debe a que el algoritmo aprende lentamente, las estimaciones pueden resultar inestables y se puede caer en mínimos locales (Izenman, 2008).

Existen dos maneras en las que puede implementarse el algoritmo de propagación hacia atrás de errores: el modo incremental (u on-line) y el modo batch. En modo incremental las observaciones entran a la red individualmente, de forma secuencial, y los ajustes en los pesos de conexión y sesgos se realizan cada vez. En cambio, en modo batch, todas las observaciones en el conjunto de aprendizaje son presentadas a la red antes de actualizar ningún peso o sesgo. Luego de que todas las observaciones son introducidas se suman los gradientes de cada ejemplo y se actualizan los parámetros. En este trabajo utilizaremos el modo batch.

Al igual que puede ocurrir, en ciertas ocasiones, con los modelos lineales uno de los problemas comunes al entrenar redes neuronales es la escasa habilidad predictiva que se obtiene si ocurre un sobre-ajuste a los datos del conjunto de entrenamiento. Este tipo de problemas es especialmente usual ante situaciones con una reducida cantidad de datos. De entre las técnicas disponibles para evitar sobre-ajustes se encuentra la denominada detención temprana.

Para la detención temprana, en lugar de dividir los datos disponibles en aprendizaje y test, se introduce un tercer conjunto: validación. Los datos del conjunto de aprendizaje son los únicos que se utilizan para determinar los parámetros de red. Los datos del conjunto de test se utilizan como conjunto independiente para evaluar la habilidad predictiva del modelo. Por su parte, el error en los datos del conjunto de validación es monitoreado durante el proceso de entrenamiento de la red. Usualmente, el error sobre el conjunto de validación decrece durante la fase inicial del entrenamiento y aumenta cuando la red comienza a sobre-ajustarse a los datos de aprendizaje.

Cuando el error sobre el conjunto de validación crece durante una cantidad pre-especificada de iteraciones consecutivas del entrenamiento, éste se detiene. Ante esa situación los parámetros de red que devuelve el algoritmo son aquellos que se obtuvieron en la iteración en que se logró el menor error sobre el conjunto de validación.

Por último, cuando se diseña una red neuronal deben tomarse una serie de decisiones. Las variables de entrada podrían estar medidas utilizando distintas escalas, lo que podría afectar la contribución relativa de cada una de ellas en el análisis. Por ello, dichas variables suelen re-escalarse al intervalo [0,1], [-1,1], o estandarizarse para tener media 0 y desviación estándar 1. Otro problema de diseño consiste en tomar la decisión de cuántos nodos y capas ocultas incluir en la red. Esto, a su vez, determinará la cantidad de parámetros necesarios para el modelo. En la mayoría de las aplicaciones de las redes neuronales estos números se determinan ya sea por el contexto del problema o por prueba y error (Izenman, 2008).

5.6. Predicción mediante clustering

La técnica de clustering es la herramienta más conocida de aprendizaje no supervisado. La metodología consiste en varios algoritmos que buscan organizar un conjunto de datos en sub-grupos homogéneos, conocidos como clusters. Un cluster puede ser pensado como un agrupamiento de elementos en el que cada elemento de un cluster es "cercano" al elemento central de ese cluster y que los miembros de clusters distintos son "lejanos" entre sí.

Las técnicas de clustering pueden separarse en jerárquicas o no jerárquicas. En las jerárquicas existe una relación de jerarquía entre la solución que conforma K clusters y la que conforma K+1, en el sentido que la solución de K clusters es el punto de partida para la obtención de la solución con K+1 clusters. En cambio, en las técnicas no jerárquicas tal relación no existe.

Dentro de los procedimientos jerárquicos se distinguen los métodos aglomerativos y los divisivos. Los aglomerativos comienzan con cada elemento formando un cluster de 1 elemento, que luego se van uniendo hasta formar un único cluster (con todos los elementos). Las técnicas divisivas hacen lo opuesto, comienzan con un sólo cluster y lo van dividiendo hasta que cada elemento tiene el suyo propio.

Los métodos de clustering no jerárquicos simplemente dividen los datos en una cantidad K pre-determinada de clusters. Dado K, se busca una partición de los datos de modo que los elementos en cada cluster sean similares entre sí y, al mismo tiempo, diferentes de los pertenecientes a otros clusters. El método de K-means (MacQueen, 1967), catalogado en esta categoría, es uno de los más populares; el método tiene dos derivaciones muy utilizadas: K-medoides y Partitioning Around Medoids (PAM) (Vinod, 1969; Kaufman and Rousseeuw, 1990; Bock, 2007). En este trabajo utilizaremos el algoritmo PAM. A continuación se describen, brevemente, estos algoritmos.

El algoritmo K-means comienza asignando elementos a K clusters pre-determinados. Luego, calcula la posición de los centroides de los K clusters (por centroide se refiere a calcular los valores promedio de las variables en los elementos del cluster). A continuación, de manera iterativa, el algoritmo busca minimizar la suma cuadrática de las distancias Euclídeas al centroide (ESS, pos su sigla en inglés) re-asignando los elementos en los diferentes clusters. El procedimiento termina cuando ninguna re-asignación de elementos disminuye el valor de las distancias.

$$ESS = \sum_{k=1}^{K} \sum_{c(i)=k} (X^i - \overline{X^k})^t (X^i - \overline{X^k})$$

donde $X^i = (X_1^i, \ldots, X_r^i)$ es la observación i-ésima del vector de variables predictoras
$\overline{X_k}$ es el centroide del K-ésimo cluster
$c(i)$ es el cluster que contiene a X^i

En contraposición a K-means que busca K centroides, K-medoides busca K objetos representativos. Dada una configuración inicial de K clusters, dentro de cada cluster se identifica un elemento representativo (medoide) como aquel que minimiza la suma de los errores cuadráticos dentro del cluster. Luego, los centroides de K-means son reemplazados por estos medoides.

PAM consiste, a su vez, en una modificación a K-medoides: introduce una estrategia de intercambio en la cual el medoide de cada cluster puede ser reemplazado por otro elemento del cluster siempre y cuando el intercambio reduzca el valor de la función objetivo.

Aquí utilizaremos las técnicas de clustering como herramienta para desarrollar un modelo de predicción simple.

Supongamos que tenemos un conjunto $L = \left\{ (X_1^i, \ldots, X_r^i, Y^i),\ i = 1, \ldots, n \right\}$ de observaciones conocidas y queremos predecir el valor desconocido de Y^{nuevo} asociado a la nueva observación $(X_1^{nuevo}, \ldots, X_r^{nuevo})$.

La técnica que proponemos plantea: primero, tomar el conjunto de observaciones de las variables predictoras de L y agregarle la nueva observación $(X_1^{nuevo}, \ldots, X_r^{nuevo})$. Denominaremos a este conjunto (L+nuevo)$_X$: $(L+nuevo)_X = \left\{ (X_1^i, \ldots, X_r^i),\ i = 1, \ldots, n,\ (X_1^{nuevo}, \ldots, X_r^{nuevo}) \right\}$. Segundo, aplicar alguna técnica de clustering a (L+nuevo)$_X$ para construir una cantidad pre-determinada de clusters. Se observa que, hasta este paso del procedimiento, sólo se consideran las variables predictoras. Tercero, para predecir Y^{nuevo} identificar a qué cluster pertenece la observación $(X_1^{nuevo}, \ldots, X_r^{nuevo})$ y predecir Y^{nuevo} como el promedio de los valores que toma la variable Y asociada a las observaciones $(X_1^i, \ldots, X_r^i)$ pertenecientes a ese cluster.

En conclusión, la técnica plantea que para realizar una predicción de la variable de respuesta Y hay que identificar casos que son similares (en el sentido de las variables predictoras consideradas) y predecir el valor promedio de Y en esa sub-población de comportamiento similar.

6. AJUSTE DE MODELOS ESTADÍSTICOS

En esta sección presentamos los resultados obtenidos al ajustar distintos modelos estadísticos para la predicción de los caudales mensuales en Rincón del Bonete y Salto Grande. El procedimiento se realiza para cada mes y cada embalse de forma independiente. En cada caso se cuenta con un total de 12 variables predictoras y la única variable a predecir es el caudal mensual de un embalse. En cada situación, el conjunto de datos está formado por las observaciones de las variables ocurridas entre 1979 y 2007 para Rincón del Bonete (29 observaciones) y 1979 y 2008 para Salto Grande (30 observaciones).

Las 12 variables que podrían utilizarse como variables predictoras son: las primeras 3 CPs asociadas a la variación interanual del viento zonal bimestral en la región, las primeras 3 CPs asociadas a la variación interanual del viento meridional bimestral en la región, las primeras 3 CPs asociadas a la variación interanual de la altura geopotencial bimestral en la región, el índice N3.4 y los caudales antecedentes Q1 y Q2.

En todos los casos presentaremos los resultados de error cross-validation leave-one-out (error cv, de aquí en adelante). El error cv representa una estimación del error de predicción, dado que evalúa la habilidad predictiva del modelo al ser enfrentado a nuevos casos.

La variabilidad interanual de los caudales mensuales puede derivar en que el problema de predicción sea relativamente más complejo en algún mes y embalse que en otros. En consecuencia, para poder comparar distintos meses y embalses teniendo en cuenta este factor, se introduce lo que daremos en llamar modelo ymedio. Ante una observación no presente en el conjunto de aprendizaje, el modelo ymedio predice el caudal promedio en los casos pertenecientes al conjunto de aprendizaje. En modo de predicciones a futuro, el modelo ymedio representa una de las predicciones más simples que se pueden realizar: predecir el promedio histórico. Se espera que los modelos a desarrollar tengan errores cv menores a el error cv del modelo ymedio, es decir, que su habilidad predictiva sea superior a la habilidad predictiva de pronosticar el promedio histórico.

El desarrollo de los modelos y la estimación de su error de predicción se realizará suponiendo conocidas las variables predictoras, por lo que los resultados a obtener en esta sección deben considerarse como cotas superiores de la habilidad predictiva. En modo operativo, el sistema de predicción de caudales deberá ser acoplado con predicciones de aquellas variables predictoras no conocidas al momento de la realización del pronóstico lo que, inevitablemente, introducirá más errores al proceso.

Las 12 variables predictoras seleccionadas son de distinta naturaleza y ocurren (en el tiempo) en momentos que pueden ser anteriores o simultáneos con el caudal que se desea predecir. Las variables que anteceden al caudal a predecir tienen la ventaja de que, si la antecedencia del pronóstico lo permite, podrían ser observadas antes de realizar el pronóstico y, por tanto, serían efectivamente conocidas. Esta situación permite considerar el error cv como una estimación de la habilidad predictiva del modelo y no sólo como una cota superior de ésta. Por el contrario, aquellas variables predictoras que ocurren en simultaneidad con el caudal a predecir no podrán ser observadas antes de la ejecución del pronóstico y, por tanto, deberán ser pronosticadas, acarreando la incertidumbre asociada.

Para pronósticos del caudal del mes (i) a realizarse una vez que el mes (i-1) haya culminado, las

variables Q1 y Q2 serán conocidas (ya habrán sido observadas). Por el contrario, si la antelación del pronóstico es mayor a 2 meses, ninguna de las dos variables estará disponible (ni siquiera bajo la forma de pronóstico imperfecto ya que el problema de predicción de caudales es justamente el que estamos abordando). Por ende, para generar predicciones con más de 2 meses de antelación debe desarrollarse un sistema de pronóstico que no incluya ni a Q1 ni a Q2.

Por su parte, la antecedencia del índice N3.4 depende del mes y embalse. Los casos más comprometidos, en este sentido, lo representan aquellas situaciones en las que éste índice óptimo acontece en el bimestre mes (i-1)-mes (i), cuando se desea predecir el caudal del mes (i). Aunque en esta situación el índice no es conocido, dada la lenta evolución de la variable, estimar el promedio bimestral por el valor en el mes (i-1) puede resultar adecuado. En particular, la variable Niño 3.4 tiene gran predictibilidad en escala de pocos meses y existen varios centros internacionales de pronóstico que ofrecen predicciones habilidosas de la misma con más de 6 meses de antelación (ver, por ejemplo, http://portal.iri.columbia.edu/portal/server.pt?open=512&objID=945&mode=2 donde se encuentra una recopilación de pronósticos que utilizan tanto técnicas estadísticas como dinámicas). En resumen, dependiendo del mes, embalse y antelación del pronóstico es posible que la variable predictora índice N3.4 sea ya conocida; aún en los casos en que esto no suceda es factible obtener buenas predicciones de la misma.

Teniendo en cuenta todo lo anterior desarrollaremos esquemas de predicción de caudales en las siguientes tres situaciones:

1. Todas las variables predictoras atmosféricas, oceánicas y de caudal antecedentes están disponibles.

2. Q1 y Q2 no disponibles. Luego, las variables predictoras se reducen a aquellas relacionadas a la circulación atmosférica regional y el índice N3.4.

3. Q1 y Q2 disponibles, índice N3.4 disponible (ya observado o estimando el valor bimestral por el del primer mes del bimestre) y sin considerar las variables atmosféricas. En particular, y dados los pronósticos de calidad, también consideramos importante el desarrollo de algún esquema de pronóstico de caudales que se base solamente en el índice N3.4.

La sección está dividida en tres sub-secciones, según cada una de las situaciones anteriores. A modo de facilitar comparaciones posteriores cada uno de los modelos a desarrollar tendrá indicado (mediante un sub-índice) el grupo de variables predictoras que utiliza. La notación a utilizar para los mencionados sub-indices será la siguiente:

A: Grupo de variables atmosféricas, es decir, Pc 1, 2, 3 de u, v, hgt.
O: Grupo de variables oceánicas el cual, en este caso, se reduce al índice N3.4
Q: Grupo de variables de caudales precedentes, o sea, Q1 y Q2.

Vale la pena recordar que, según se desarrolló en el capítulo introducción, aquellos pronósticos de caudal que utilicen como variables predictoras a las variables relacionadas con la circulación atmosférica (casos 1 y 2 de la división presentada más arriba) pertenecen a la categoría de esquemas de predicción mediante downscaling híbrido (ver Figura 1.2). Por el contrario, aquellos pronósticos en los que únicamente se utilicen el índice Niño 3.4, Q1y Q2 pertenecen a la categoría de esquemas de predicción orientados puramente por datos (ver Figura 1.1).

6.1. Predicción con variables atmosféricas, oceánicas y caudales precedentes

Como se explicó, todos los modelos a desarrollar en esta sección contarán con el sub-índice AOQ que indica los grupos de variables predictoras que utilizan: atmosféricas, oceánicas y caudales precedentes.

Las generalidades sobre los modelos estadísticos a utilizar fueron descritas en la sección de Modelos estadísticos para predicción. A continuación se describen algunas decisiones particulares que fueron tomadas para el desarrollo de tales modelos en los casos en estudio. La implementación de los modelos fue realizada en R o Matlab. En particular se utilizaron los paquetes leaps (Lumeley, 2009) y clusters (Maechler et al., 2005).

6.1.1. Regresión lineal múltiple, selección de variables y PLSR

El primer modelo a desarrollar será el de regresión lineal múltiple, utilizando todas las variables predictoras de los grupos A, O y Q, los cuales totalizan 12 elementos: X_1, ..., X_{12}. No introducimos en el modelo ninguna transformación de estas variables por lo que la ecuación de regresión, para cada mes y embalse, puede expresarse como:

$$y = b_0 + b_1 X_1 + ... + b_{12} X_{12} + \epsilon$$

donde $b_0, ..., b_{12}$ son los coeficientes a determinar.

Dado el elevado número de variables predictoras y, también, las elevadas correlaciones entre algunas de ellas (las Figuras 6.1.1.1 y 6.1.1.2 muestran las correlaciones entre las variables para Rincón del Bonete en abril y Salto Grande en diciembre, respectivamente) el modelo anterior podría tener problemas al intentar predecir la respuesta en casos no presentes en el conjunto de aprendizaje. Como fue desarrollado anteriormente, existen varias técnicas que intentan lidiar con ésta problemática. De entre ellas trabajaremos sólo con dos: el método de selección de variables eliminación hacia atrás (modo stepwise) y PLSR.

Al combinar la regresión lineal múltiple con la técnica de selección de variables (eliminación hacia atrás stepwise, en este caso; de aquí en adelante llamada simplemente eliminación hacia atrás) se obtienen una serie de modelos con distinta cantidad de variables predictoras. Determinaremos la cantidad óptima de variables a retener en el modelo a través de los errores cv que los distintos modelos generan. El procedimiento consiste en, primero, utilizar todos los datos disponibles (es decir, todos aquellos que luego serán divididos en aprendizaje y test) para implementar la técnica de eliminación hacia atrás. Esta técnica genera, como resultado, una lista con las variables predictoras seleccionadas para conformar modelos con entre 1 y 12 variables. Segundo, para seleccionar cuál de los 12 modelos es más adecuado se calcula el error cv de cada uno de ellos. El modelo que presente el menor error cv será designado como óptimo, entre los modelos que se obtienen al combinar regresión lineal múltiple con eliminación hacia atrás.

Análogamente, para la determinación de la cantidad óptima de componentes a retener en el análisis PLSR se evalúan los errores cv de los distintos modelos generados. En resumen, para cada mes y embalse, el modelo PLSR tendrá la cantidad de componentes que generen el menor error cv.

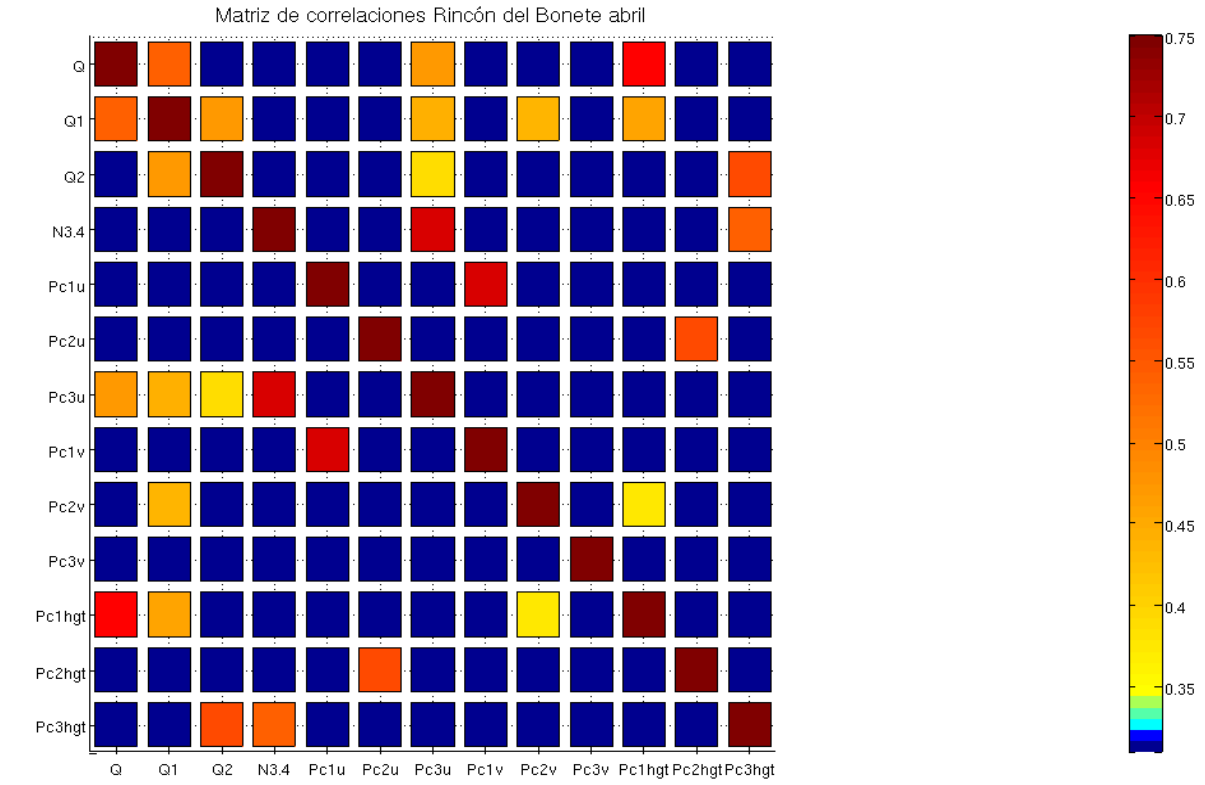

Figura 6.1.1.1: Matriz de correlaciones cruzadas entre variables predictoras y de respuesta para Rincón del Bonete en abril. Correlaciones no significativas al 95% son indicadas en azul oscuro.

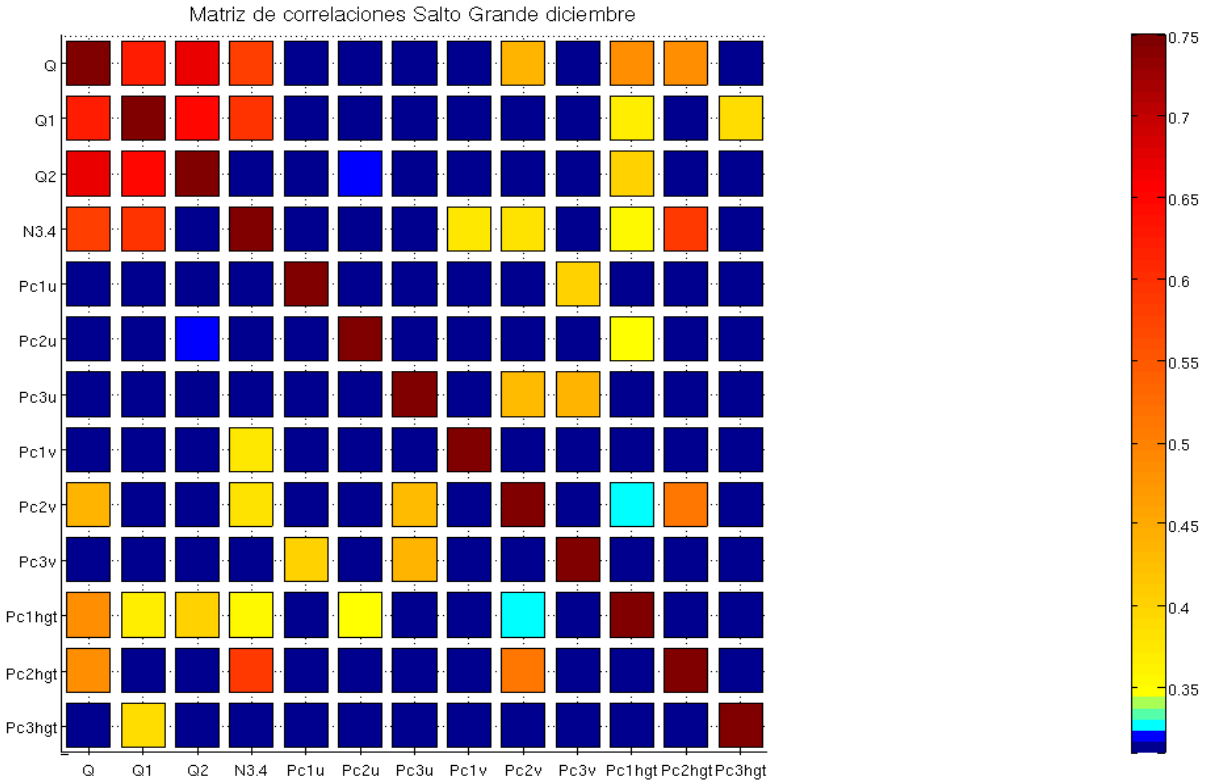

Figura 6.1.1.2: Idem que Figura 6.1.1.1 para Salto Grande en diciembre.

6.1.2. Árboles de regresión

Para la predicción mediante árboles de regresión, definimos un criterio para evitar la saturación de los árboles (y los consecuentes sobre-ajustes y escasa habilidad predictiva). El criterio consiste en prescribir la cantidad mínima de observaciones que deben pertenecer a un cierto nodo para que éste pueda ser dividido luego en 2 nodos hijos; llamaremos nmin a éste número. El número nmin óptimo será determinado, también, a través de la evaluación del error cv: el nmin que genere un menor error cv será seleccionado como óptimo. Se varía nmin en el rango 4-10.

6.1.3. Redes Neuronales

En redes neuronales es necesario pre-definir la arquitectura de red a utilizar. En este caso consideraremos, únicamente, redes neuronales con 1 capa oculta con 2 nodos ocultos dentro de ella. Dado que existen 12 variables predictoras, 3 sesgos y 1 variable de salida, en total deberían estimarse 29 parámetros. Visto que este número casi iguala la cantidad de observaciones disponibles, se decide realizar una modificación. En lugar de utilizar las 12 variables predictoras seleccionadas se utilizan solamente 4 de ellas: tomamos las 4 variables determinadas por el método de eliminación hacia atrás. Aunque la selección de variables hacia atrás se basa en el ajuste de modelos lineales consideramos, de todas formas, que es un procedimiento adecuado para identificar variables importantes para el proceso de predicción. Luego, con 4 variables predictoras, 3 sesgos y 1 variable de salida el número de parámetros a determinar se reduce a 13.

Las funciones de activación de los nodos ocultos y de salida son tangentes hiperbólicas (ejemplo de función sigmoidal). Para el entrenamiento de la red las variables se re-escalan al intervalo [-1,1].

Para intentar evitar el sobre-ajuste a los datos se utiliza la técnica de detención temprana del algoritmo de propagación hacia atrás de los errores. Para cada caso se seleccionan, al azar, el 20% de los datos del conjunto de aprendizaje para conformar el conjunto de validación. Si el error en el conjunto de validación crece por 6 iteraciones consecutivas del algoritmo, éste se detiene y se devuelven como parámetros de red aquellos correspondientes a la iteración en la que se obtuvo el mínimo error sobre el conjunto de validación.

Dada la complejidad del problema y la inicialización a través de valores aleatorios, es altamente probable que cada vez que se ejecute el algoritmo para determinación de los parámetros de red el resultado sea diferente. Es usual, entonces, implementarlo varias veces y, luego, seleccionar aquella colección de parámetros que generan el menor error aparente (es decir, el menor error sobre el conjunto de aprendizaje). En este caso ejecutamos el algoritmo 100 veces. De todas formas, no existe certeza de que este mecanismo finalmente resulte en la identificación del mínimo absoluto de la función de error.

6.1.4. Clustering

Como algoritmo de clustering se utiliza PAM. Para asignar una predicción de caudal a una observación se divide el conjunto en 7 clusters, según las variables predictoras. La cantidad de 7 clusters es arbitraria, aunque consideramos que dado el número de datos disponibles es una elección

adecuada. En caso de que la observación a predecir quede separada en un cluster individual, se repite el algoritmo solicitando división en 6 clusters. Si se repite la situación, sucesivamente se ejecuta el algoritmo reduciendo la cantidad de clusters a formar hasta que el cluster al que pertenece la observación a predecir tenga, al menos, 2 elementos.

Para la predicción mediante clustering la noción de error aparente, es decir, el promedio de los errores sobre el conjunto de aprendizaje coincide plenamente con la noción de error cv leave-one-out. Esto se debe a que, en ésta técnica, predecir la respuesta de una observación en el conjunto de aprendizaje es idéntico a predecir la respuesta de una observación cualquiera fuera de este conjunto.

6.1.5. Resultados

En las Figuras 6.1.5.1 y 6.1.5.2 se resumen, para Rincón del Bonete y Salto Grande, los resultados de error aparente para los distintos modelos ajustados: regresión lineal múltiple con todos los predictores de los grupos A, O y Q (lm_{AOQ}), regresión lineal múltiple acoplado con eliminación hacia atrás óptimo (lm_{AOQ} -óptimo), PLSR con la cantidad de variables óptima ($PLSR_{AOQ}$ -óptimo), árboles de regresión con el criterio nmin óptimo ($Árbol_{AOQ}$-óptimo), redes neuronales (Red_{AOQ}) y predicción vía algoritmo de clustering ($Clusters_{AOQ}$). En todos los casos los errores son presentados como cocientes respecto del error aparente del modelo ymedio, por lo que valores superiores (inferiores) a 1 indican un ajuste peor (mejor) que el del modelo ymedio.

Para Rincón del Bonete (Figura 6.1.5.1) se observa que todos los modelos, salvo redes neuronales y clustering, tienen en todos los casos errores aparentes menores que el modelo ymedio, o sea que ajustan mejor a los datos de aprendizaje. A excepción de las redes neuronales, que muestran valores muy variables a lo largo de los meses, el resto de los modelos mantienen, en general, un cierto orden según el grado de ajuste a los datos. Este orden es, de peor a mejor ajuste: $Clusters_{AOQ}$, $PLSR_{AOQ}$-óptimo, lm_{AOQ}-óptimo, lm_{AOQ}, $Árbol_{AOQ}$-óptimo.

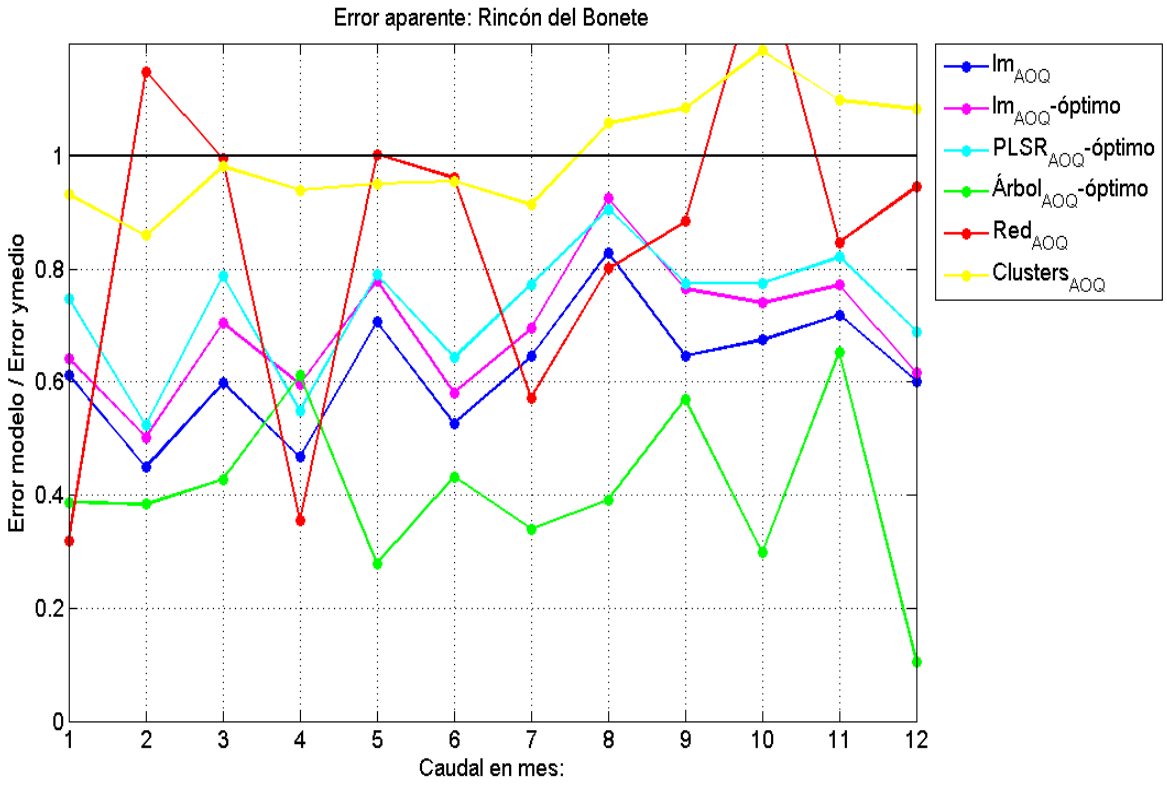

Figura 6.1.5.1: Errores aparentes de modelos para predicción de caudales en Rincón del Bonete. Los errores se expresan como el cociente por el error aparente del modelo ymedio. La línea negra indica errores iguales a los del modelo ymedio.

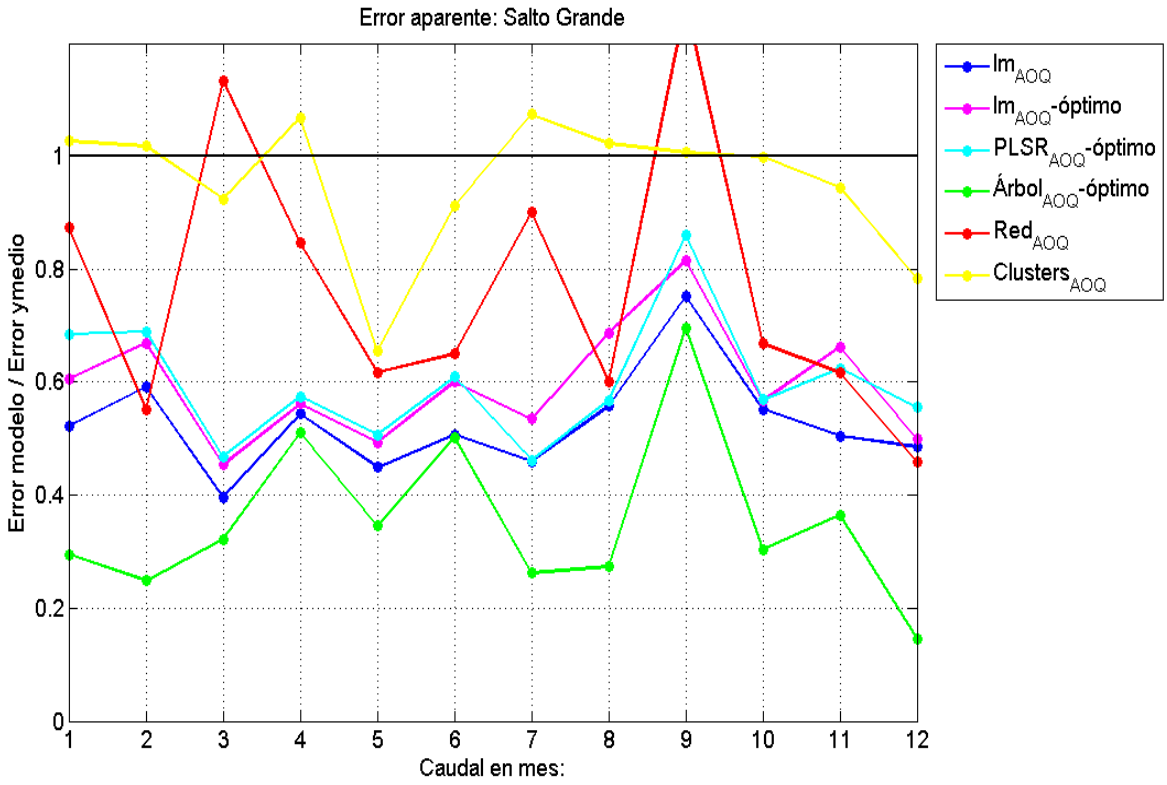

Figura 6.1.5.2: Idem que Figura 6.1.5.1 para Salto Grande.

Para Salto Grande (Figura 6.1.5.2) vuelve a repetirse la situación de que el modelo de redes neuronales tiene un comportamiento muy errático. De los restantes modelos Clusters es el único que presenta errores aparentes mayores que el ymedio en algunos meses. En general los modelos mantienen el mismo orden según grado de ajuste que se notó en el gráfico análogo de Rincón del Bonete.

Cabe recordar que considerar al error aparente como un estimador del error de predicción de un determinado modelo puede originar resultados poco realistas y quizás, también, demasiado optimistas. Como fue desarrollado antes, en este trabajo seleccionamos como estimación del error de predicción al error cv.

En las Figuras 6.1.5.3 y 6.1.5.4 se resumen, para Rincón del Bonete y Salto Grande, los resultados de error cv para los mismos modelos presentados en las Figuras 6.1.5.1 y 6.1.5.2. Nuevamente, los errores son presentados como cocientes respecto del error cv del modelo ymedio, por lo que valores superiores (inferiores) a 1 indican un desempeño peor (mejor) que el del modelo ymedio. Así mismo, el desempeño relativo del modelo lm_{AOQ}-óptimo respecto al modelo ymedio puede considerarse como un estimador del grado de predictibilidad de los caudales mensuales: en caso de que el desempeño de lm_{AOQ}-óptimo supere al del modelo ymedio diremos que existe predictibilidad y en caso contrario que no existe predictibilidad.

Para Rincón del Bonete (Figura 6.1.5.3) se observa que el modelo con el mejor desempeño general (mejor en 10 de 12 meses) es lm_{AOQ}-óptimo. En el mes de enero el lm_{AOQ}-óptimo es superado en desempeño por $Árbol_{AOQ}$-óptimo y en agosto, cuando iguala en desempeño al modelo simple ymedio, es superado por las redes neuronales. El error cv de lm_{AOQ}-óptimo es siempre menor al del modelo ymedio, salvo en el mes de agosto donde son iguales. El modelo lm_{AOQ}-óptimo es claramente el de mejor desempeño seguido por $PLSR_{AOQ}$-óptimo (el cual también supera en desempeño al modelo ymedio exceptuando el mes de agosto), en tercer lugar de desempeño se puede ubicar al $Clusters_{AOQ}$, aunque éste modelo es mejor que ymedio sólo entre enero y julio. Salvo pocas excepciones los modelos lm_{AOQ}, Red_{AOQ} y $Árbol_{AOQ}$-óptimo no presentan buen desempeño. En cuanto a la estacionalidad del desempeño de los distintos modelos no se aprecian comportamientos notables salvo la dificultad de predicción de caudales en el mes de agosto. Esta dificultad de predicción en agosto era, de cierta forma, esperable ya ninguno de los predictores seleccionados para este mes y embalse alcanzaban niveles de correlación significativos con la serie de caudal (Figura 4.4.1).

Para Salto Grande (Figura 6.1.5.4), nuevamente, el modelo con el mejor desempeño general es lm_{AOQ}-óptimo, seguido por $PLSR_{AOQ}$-óptimo. Ambos modelos tienen desempeños siempre superiores al del modelo ymedio. En este embalse el modelo $Clusters_{AOQ}$ supera, en desempeño, al ymedio en todos los meses salvo abril y julio. Los modelos lm_{AOQ}, Red_{AOQ} y $Árbol_{AOQ}$-óptimo son los de peores resultados. Si se selecciona como modelo de predicción de caudales el lm_{AOQ}-óptimo se concluye que, para Salto Grande, los caudales circulantes en setiembre, enero y febrero son los más difíciles de predecir, ocurriendo las temporadas de elevada predictibilidad entre el otoño - mediados de invierno y primavera.

La relación error cv / error cv modelo ymedio es, en general para todos los meses y modelos, menor para Salto Grande que para Rincón del Bonete, indicando una mayor predictibilidad potencial de caudales en Salto Grande.

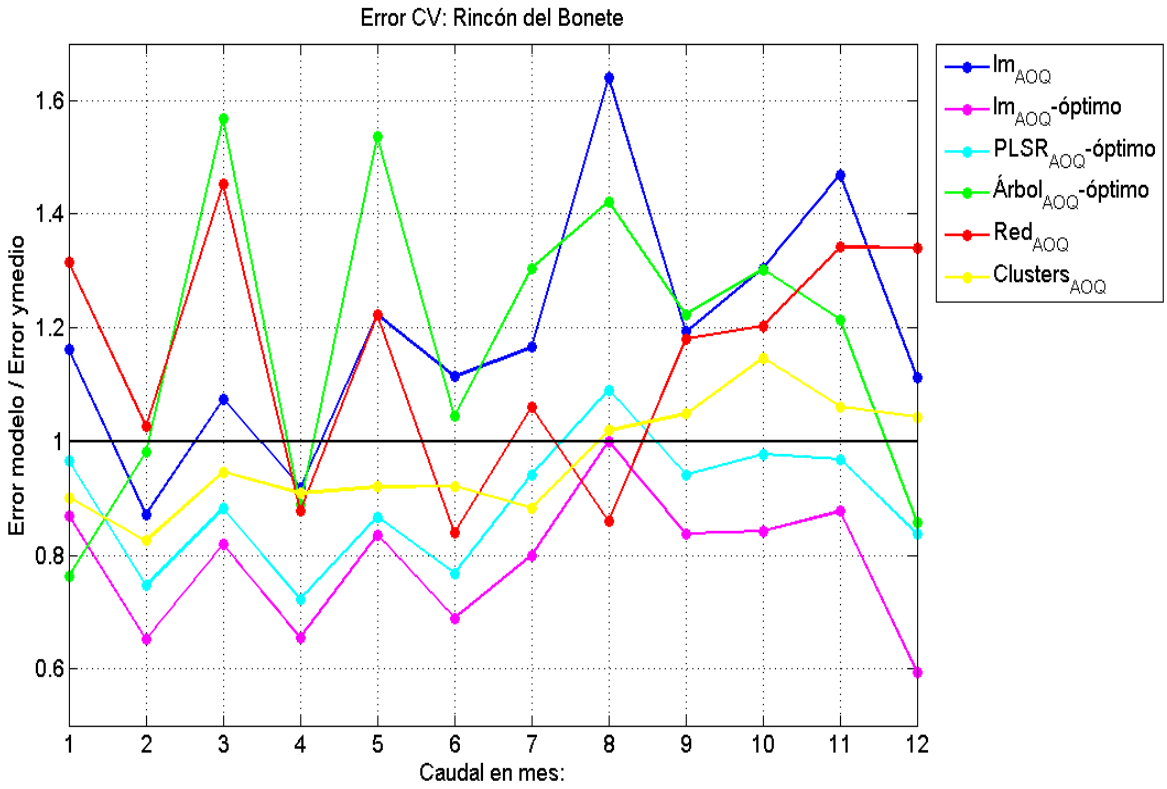

Figura 6.1.5.3: Errores cv de modelos para predicción de caudales en Rincón del Bonete. Los errores se expresan como el cociente por el error cv del modelo ymedio. La línea negra indica errores iguales a los del modelo ymedio.

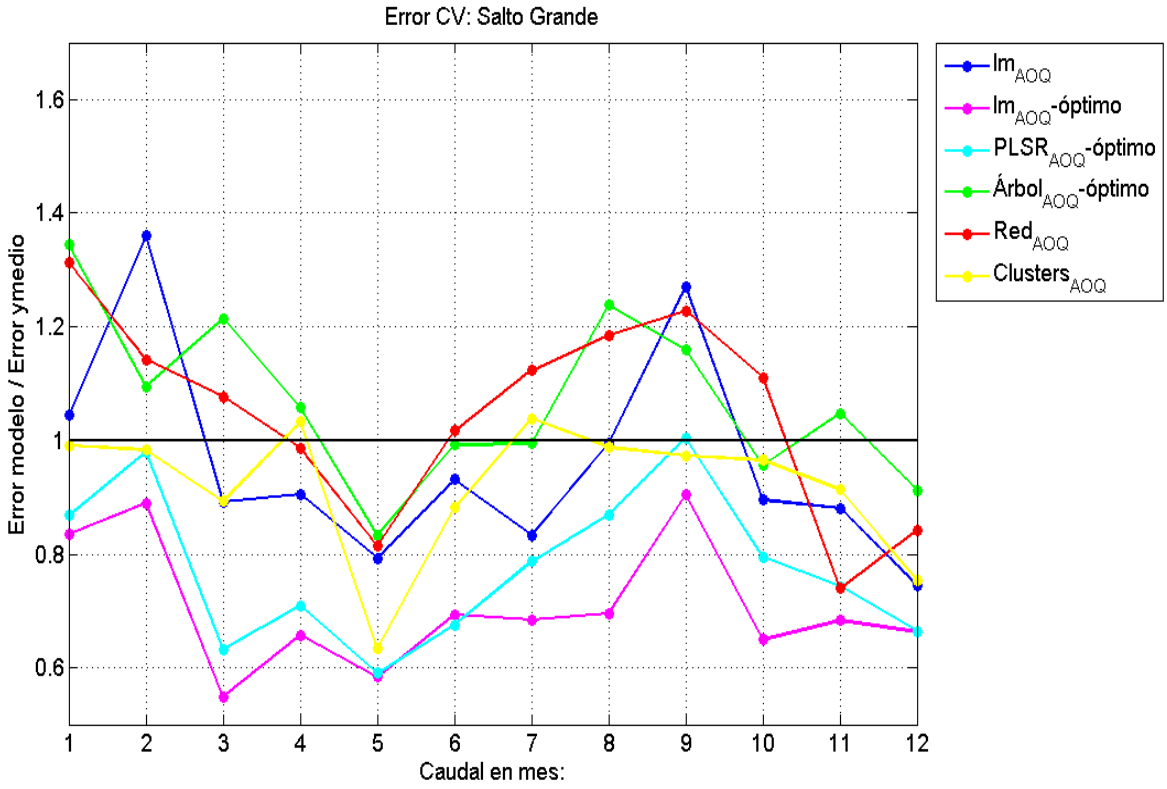

Figura 6.1.5.4: Idem que Figura 6.1.5.3 para Salto Grande.

Se recuerda que todos los ajustes de redes neuronales se realizan utilizando el concepto de conjunto de validación, para evitar el sobre-ajuste. El conjunto de validación es seleccionado al azar a partir del conjunto de aprendizaje y totaliza el 20% de los datos en él. Por tanto, la comparación entre redes neuronales y los restantes modelos debe ser cautelosa, ya que el conjunto de aprendizaje es usado para aprendizaje y validación en el caso de las redes . Asimismo, hay que tener en cuenta que, dada la poca cantidad de datos, los resultados podrían ser muy sensibles a la integración del conjunto de validación.

Tanto para Rincón del Bonete como para Salto Grande en la enorme mayoría de los meses el modelo de mayor habilidad predictiva es el lm_{AOQ}-óptimo. En las Figuras 6.1.5.5 y 6.1.5.6 se presentan cuadros que indican las variables que son seleccionadas por el procedimiento de eliminación hacia atrás para conformar los modelos lm_{AOQ}-óptimo en los distintos meses para cada uno de los embalses. En ambas figuras se indica en el eje de las abscisas el mes del caudal que se desea predecir y en el eje de las ordenadas la variable predictora, las variables que son seleccionadas para el modelo lm_{AOQ}-óptimo en cada caso son indicadas en color naranja.

Para Rincón del Bonete (Figura 6.1.5.5), en general, de entre las variables asociadas a la circulación atmosférica regional las CPs del viento zonal (u) son poco seleccionadas y no se aprecia ninguna estacionalidad en dicha selección; las CPs asociadas al viento meridional (v) son bastante seleccionadas apreciándose una continuidad en la selección en el período abril-julio; por su parte las CPs asociadas a la altura geopotencial (hgt) presentan una temporada muy marcada en la que son seleccionadas: el verano (desde diciembre a febrero). El índice N3.4 es seleccionado en 5 meses, pero no se aprecia una estacionalidad en esta selección. Por último, los caudales antecedentes son seleccionados con gran frecuencia, siendo Q1 más seleccionada que Q2. Se aprecia, también, que en un par de meses ambas Q1 y Q2 fueron seleccionadas por el modelo óptimo. En cuanto a la cantidad de variables seleccionadas por el modelo lm12-óptimo en cada caso, se aprecia que la misma varía entre 2 y 6 variables.

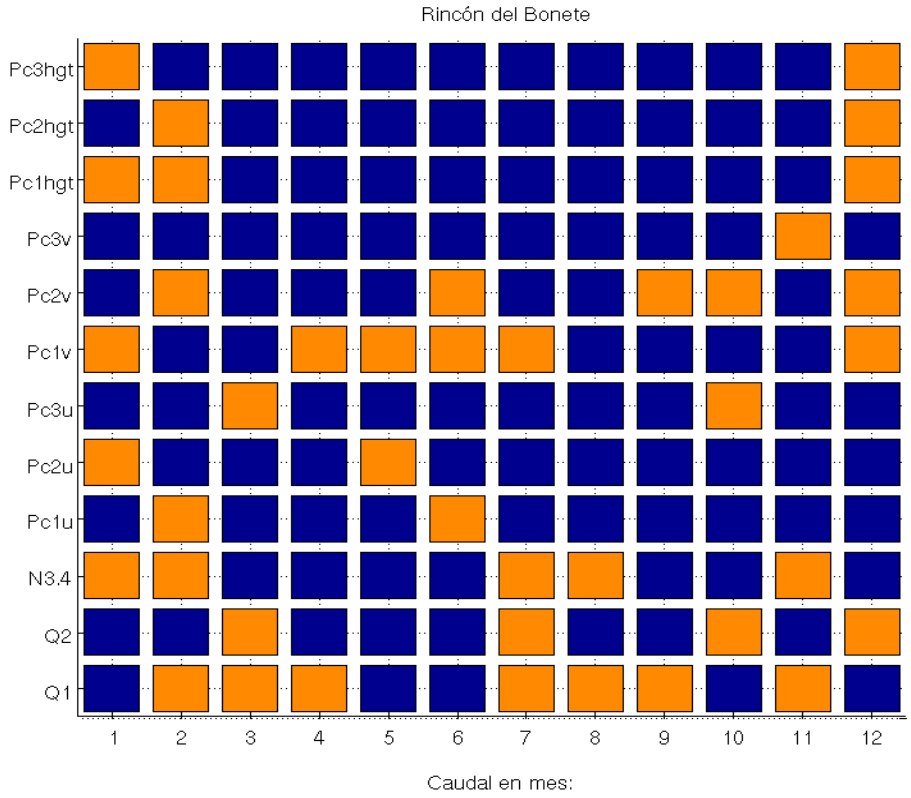

Figura 6.1.5.5: Variables seleccionadas por el proceso de eliminación hacia atrás para conformar el modelo lm_{AOQ}-óptimo para Rincón del Bonete. Las variables seleccionadas se indican en color naranja.

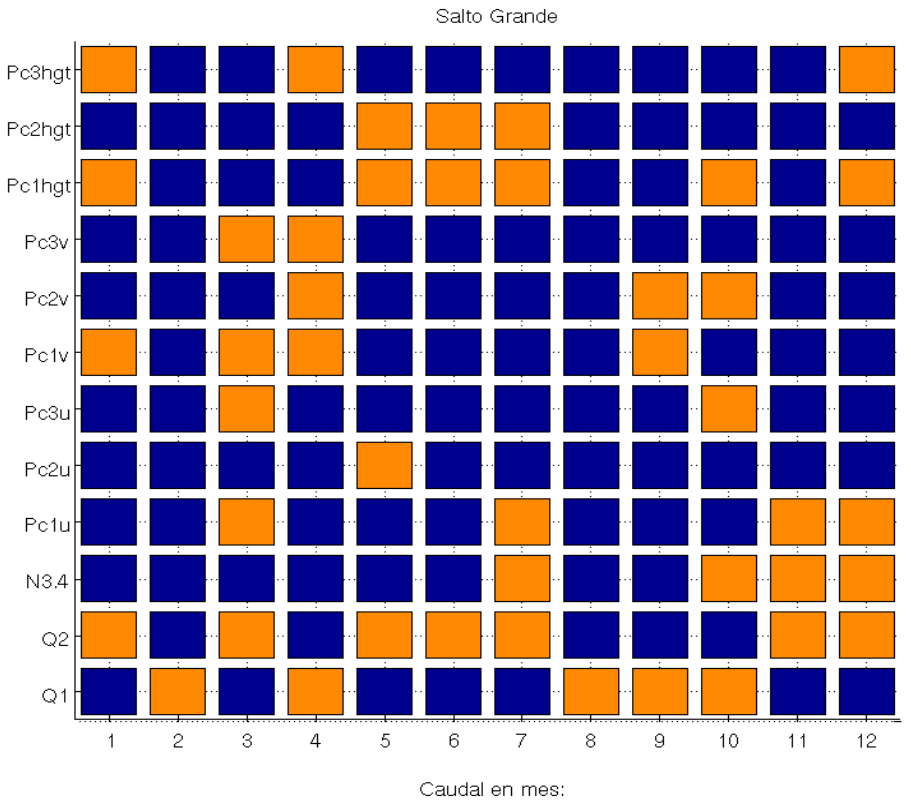

Figura 6.1.5.6: Idem que Figura 6.1.5.5 para Salto Grande.

Para Salto Grande (Figura 6.1.5.6) de las variables atmosféricas, nuevamente, aquellas asociadas al viento zonal (u) son las menos seleccionadas; las asociadas al viento meridional (v) parecen tener un breve pero robusto período de selección hacia comienzos del otoño (marzo y abril) y otro hacia comienzos de la primavera (setiembre y octubre); las asociadas a la altura geopotencial (hgt) son las más seleccionadas y se destaca un período (entre mayo y julio) en el que su selección es sistemática. El índice N3.4, a diferencia de lo que ocurre en Rincón del Bonete, es seleccionado, principalmente, en el período de primavera (octubre a diciembre). En todos los meses alguno de los dos caudales antecedentes es seleccionado, no ocurriendo la selección simultánea de ambos.

6.2. Predicción con variables atmosféricas y oceánicas

En esta sub-sección se analizan resultados frente a situaciones en las que no se cuenta con las variables de caudal antecedente (Q1 y Q2) dentro del conjunto inicial de variables predictoras. Por lo tanto, el conjunto inicial de predictores está formado por 10 elementos: el grupo de variables atmosféricas y la variable oceánica.

Visto que en la sub-sección anterior el modelo de mejor desempeño, en términos del error de predicción, fue el lineal acoplado con selección de variables en la presente situación sólo analizaremos los resultados del mismo. A partir del conjunto de 10 variables predictoras a utilizar en este sección (las 9 Cps atmosféricas y N3.4), implementaremos le técnica de eliminación hacia atrás de variables y determinaremos el modelo lineal con la cantidad óptima de variables, según el criterio de menor error cv. Es decir que, el procedimiento para la obtención del modelo lineal óptimo es análogo al descrito en la sub-sección anterior, pero sin utilizar ni Q1 ni Q2. Este modelo óptimo se denominará lm_{AO}-óptimo, para indicar que es óptimo a partir de los grupos de predictores atmosféricos y oceánicos. Sólo analizaremos los resultados de error cv.

En las Figuras 6.2.1 y 6.2.2 se presentan los errores cv del modelo lineal óptimo que se obtiene por selección de variables a partir del conjunto de predictores atmosféricos y oceánicos (lm_{AO}-óptimo) en Rincón del Bonete y Salto Grande, respectivamente. Para facilitar la comparación con los resultados de la sección anterior, en ambas figuras, se despliegan los errores cv de los modelos lm_{AOQ}-óptimo correspondientes a cada embalse.

Para Rincón del Bonete (Figura 6.2.1) el modelo lm_{AO}-óptimo presenta un desempeño superior al del modelo ymedio en todos los meses del año, distinguiéndose agosto como el mes de peor desempeño. Es necesario notar que en el mes de agosto aún cuando el modelo lm_{AOQ}-óptimo arroja resultados iguales a los del modelo ymedio, en esta ocasión (modelo lm_{AO}-óptimo) el desempeño es levemente superior; ésto simplemente indica que las variables seleccionadas para el modelo óptimo a partir de los predictores atmosféricos y oceánico fueron distintas a las seleccionadas cuando también se incluyen los predictores de caudal previo. Más allá de destacarse agosto como el mes más comprometido no existen períodos de claro peor o mejor desempeño. En general, para este embalse, las mayores pérdidas de predictibilidad al no utilizar los caudales antecedentes se aprecian en primavera y verano.

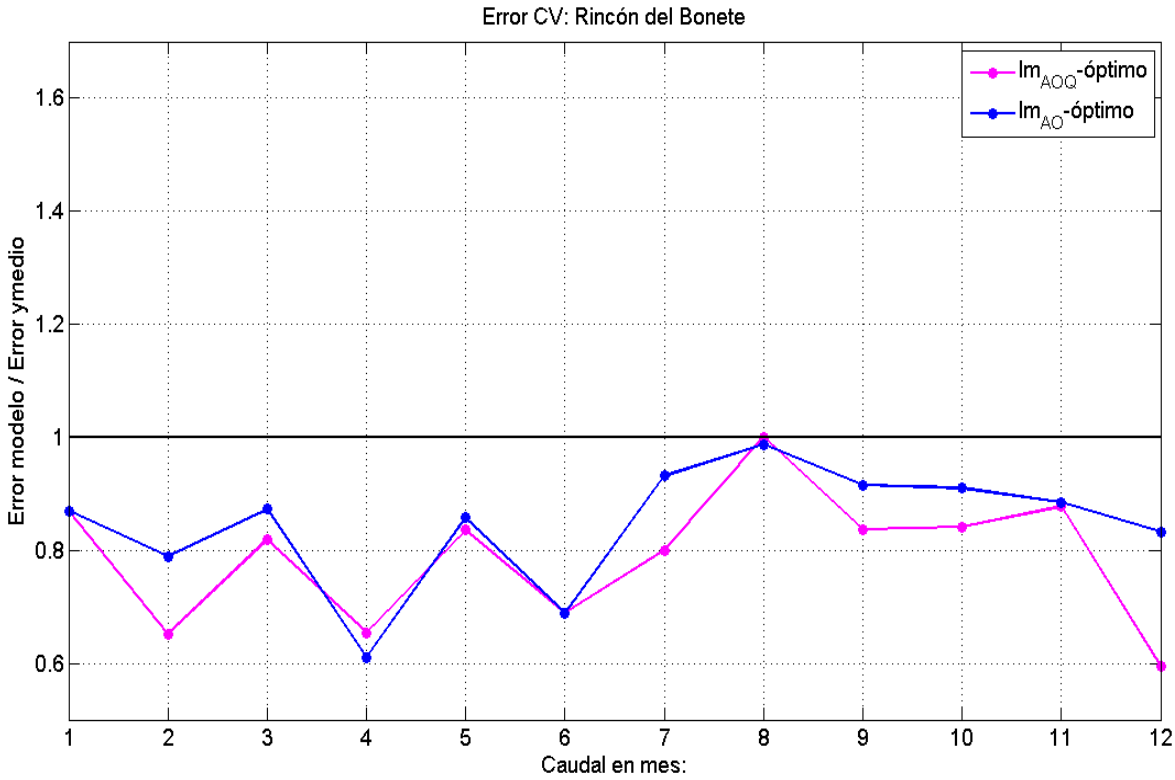

Figura 6.2.1: Errores cv de los modelos lm_{AO}-óptimo y lm_{AOQ}-óptimo para Rincón del Bonete. Los errores se expresan como el cociente por el error cv del modelo ymedio. La línea negra indica errores iguales a los del modelo ymedio.

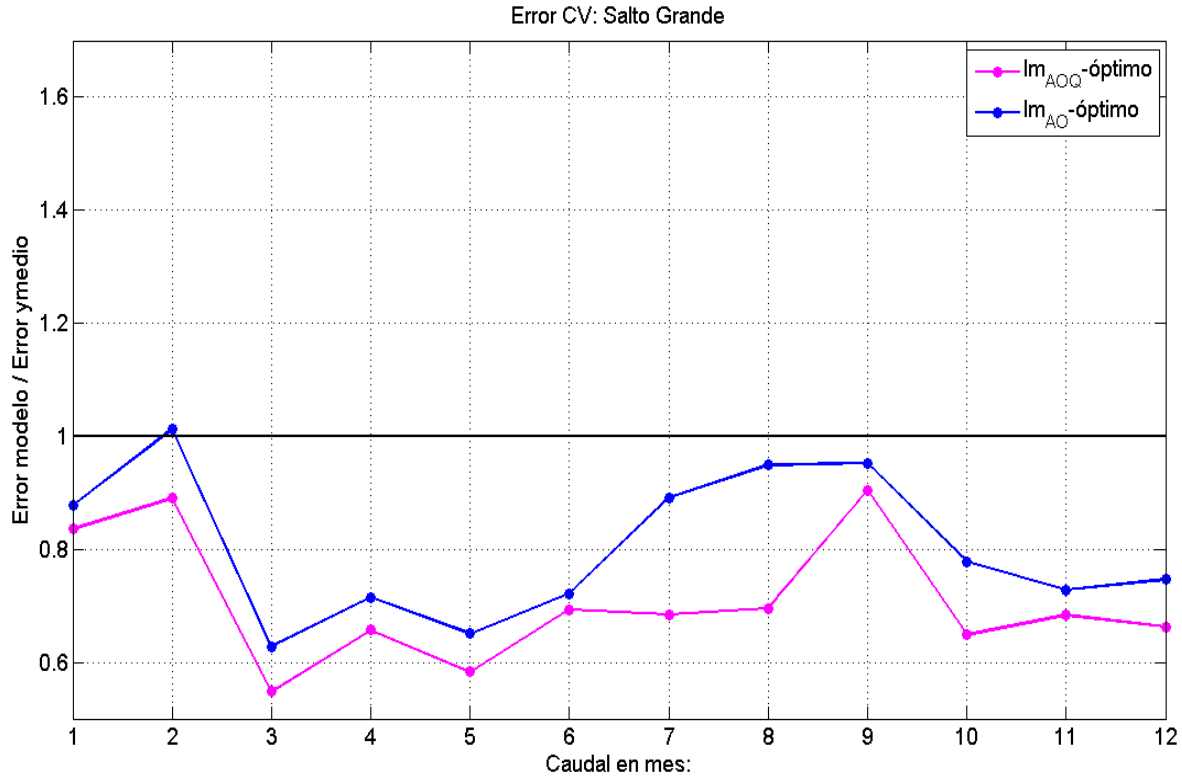

Figura 6.2.2: Idem Figura 6.2.1 para Salto Grande.

Para Salto Grande (Figura 6.2.2) se destacan dos períodos de marcado buen desempeño del modelo lm_{AO}-óptimo : desde marzo a junio (otoño) y de octubre a diciembre (primavera). El único mes en el que la habilidad de lm_{AO}-óptimo es inferior a la de ymedio es febrero. Para Salto Grande se aprecia que la disminución en la habilidad predictiva al dejar de considerar los caudales precedentes es importante en todos los meses del año; julio y agosto destacan como los meses en los que los errores cv de los modelos lm_{AO}-óptimo y lm_{AOQ}-óptimo se encuentran más alejados.

Por otro lado, ante la situación de no disponer de Q1 ni Q2 los valores del cociente error cv modelo lm_{AO}-óptimo / error cv modelo ymedio de un cierto mes no son muy diferentes para los dos embalses en consideración.

Por último, en las Figuras 6.2.3 y 6.2.4 se presentan cuadros que indican las variables seleccionadas por eliminación hacia atrás para formar los modelos lm_{AO}-óptimo en Rincón del Bonete y Salto Grande, respectivamente.

Para Rincón del Bonete (Figura 6.2.3), de entre las variables atmosféricas las CPs asociadas a la altura geopotencial (hgt) son las menos seleccionadas y las asociadas al viento meridional (v) las más seleccionadas. De forma similar a lo que ocurría cuando se utilizaban además de los predictores atmosféricos y oceánico los de caudal previo, las variables asociadas a la altura geopotencial parecen ser seleccionadas continuamente en la temporada de verano. En este caso también las variables asociadas al viento meridional presentan cierta estacionalidad siendo seleccionadas de forma persistente entre abril y junio. El índice N3.4 es seleccionado de forma alternada, no observándose ninguna estacionalidad en la selección.

Para Salto Grande (Figura 6.2.4), al eliminar las variables Q1 y Q2 del conjunto, las CPs asociadas con la altura geopotencial pasan de ser las más seleccionadas (de entre las variables atmosféricas) a ser las menos seleccionadas. De todas formas, la variable más seleccionada para este embalse es el índice N3.4, el cual es seleccionado para el modelo óptimo en 9 de 12 meses: únicamente no es seleccionado en marzo, setiembre y octubre.

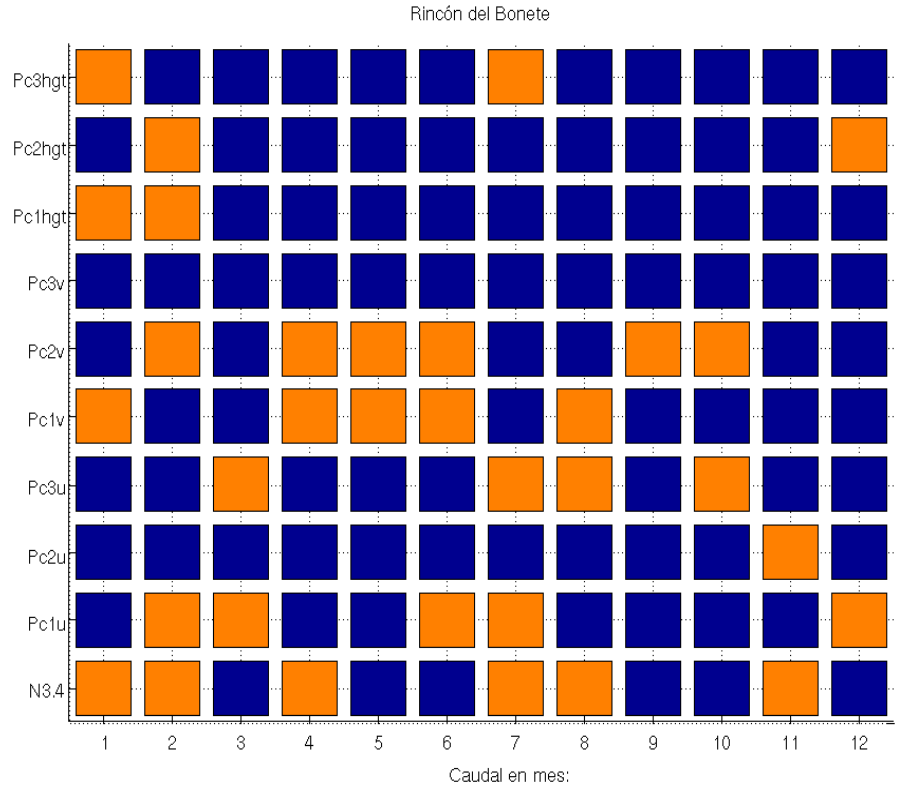

Figura 6.2.3: Variables seleccionadas por el proceso de eliminación hacia atrás para conformar el modelo lm_{AO}-óptimo para Rincón del Bonete. Las variables seleccionadas se indican en color naranja.

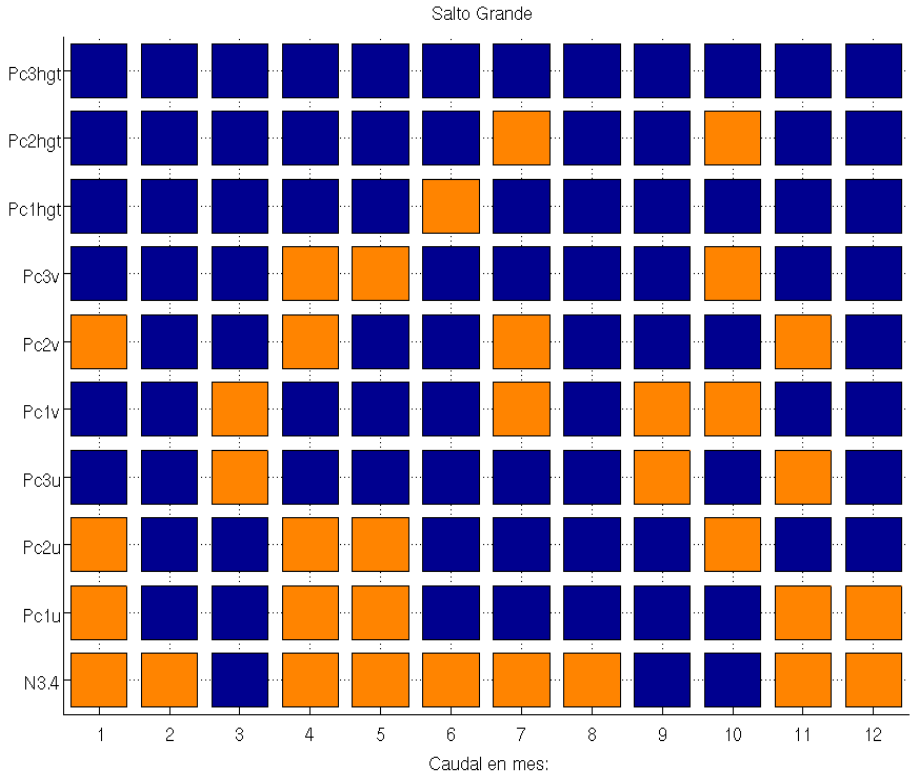

Figura 6.2.4: Idem Figura 6.2.3 para Salto Grande.

6.3. Predicción con variables oceánicas y caudales precedentes

Esta sub-sección es análoga a la anterior, pero tomando como conjunto inicial de predictores a N3.4, Q1 y Q2 solamente.

En las Figuras 6.3.1 y 6.3.2 se presentan, para Rincón del Bonete y Salto Grande, los errores cv de los modelos lineal óptimo correspondiente (lm_{OQ}-óptimo) y de regresión lineal utilizando solamente el índice N3.4 (lm_O). Al igual que antes, para facilitar comparaciones, se agregan a las figuras los resultados obtenidos previamente para los modelos lm_{AOQ}-óptimo. En analogía con las secciones precedentes, por lm_{OQ}-óptimo referimos a aquel que se obtiene aplicando el método de selección de variables eliminación hacia atrás y calculando el menor error cv.

Para Rincón del Bonete (Figura 6.3.1) el modelo lm_{OQ}-óptimo tiene un desempeño superior al del modelo ymedio en todos los meses, salvo agosto, siendo el período de febrero a julio el de menor error de predicción relativo al del modelo ymedio. Por el contrario, el modelo lm_O presenta habilidad predictiva superior al modelo ymedio en la mitad del año: desde finales de primavera a comienzos del otoño; el desempeño de este modelo es siempre inferior al del lm_{OQ}-óptimo. La pérdida de habilidad predictiva al no utilizar los predictores atmosféricos es importante en la mayoría de los meses del año.

En Salto Grande (Figura 6.3.2) el modelo lm_{OQ}-óptimo tiene desempeño superior al del modelo ymedio en todos los meses del año, destacando los períodos enero-febrero y setiembre-octubre como los de mayor error de predicción relativo a ymedio. Para este embalse lm_O presenta habilidad considerable (es decir, superior a la del modelo ymedio) en todos los meses salvo febrero y setiembre, aunque esta habilidad es inferior a la del modelo lm_{OQ}-óptimo. La pérdida de habilidad por no considerar los predictores atmosféricos es de importancia considerable, destacando octubre como el mes en el que ésta pérdida es más extrema.

Por último, en las Figuras 6.3.3 y 6.3.4 se presentan las variables seleccionadas por eliminación hacia atrás para conformar el modelo lm_{OQ}-óptimo para Rincón del Bonete y Salto Grande, respectivamente. En ambos cuadros se aprecia que la variable más seleccionada para ambos embalses es Q1; N3.4 en Rincón del Bonete es seleccionada de forma continua en la temporada de noviembre a marzo, es decir desde fines de primavera a fines del verano.

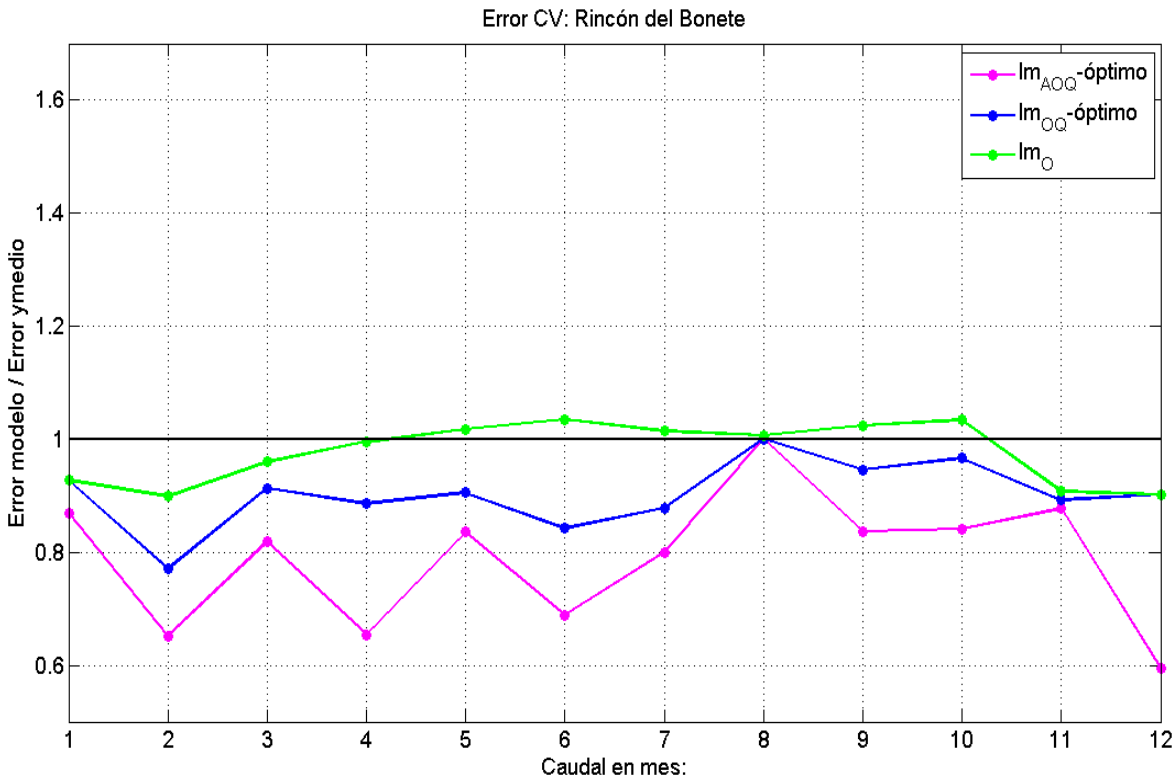

Figura 6.3.1: Errores cv de los modelos lm_{OQ}-óptimo lm_{O}-óptimo y lm_{AOQ}-óptimo para Rincón del Bonete. Los errores se expresan como el cociente por el error cv del modelo ymedio. La línea negra indica errores iguales a los del modelo ymedio.

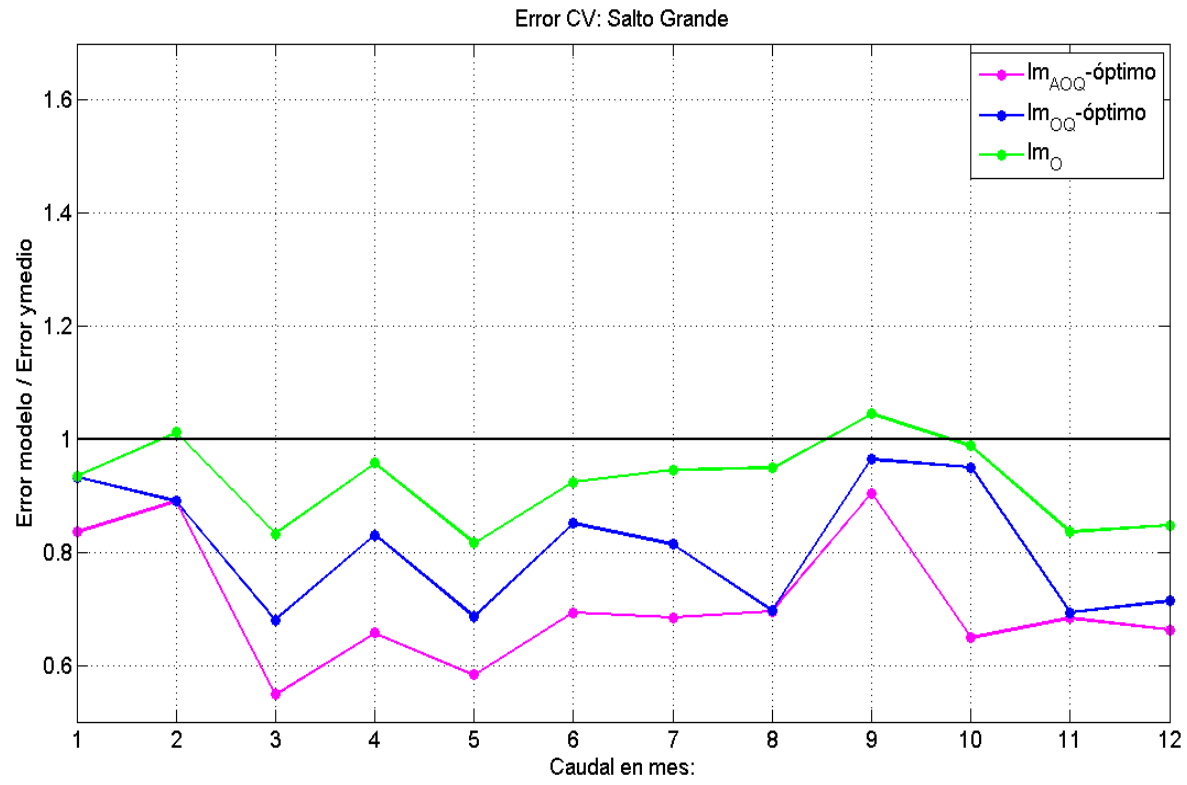

Figura 6.3.2: Idem Figura 6.3.1 para Salto Grande.

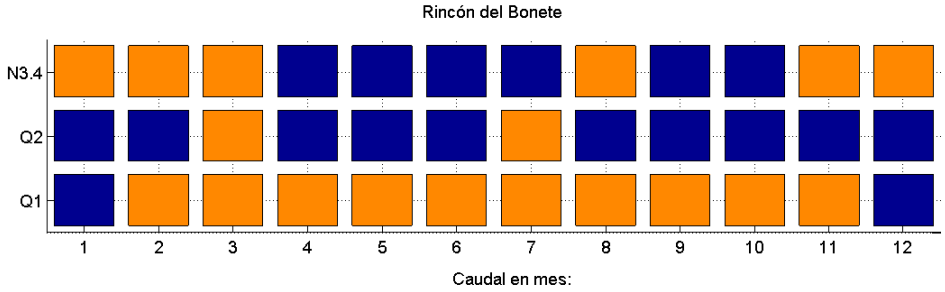

Figura 6.3.3: Variables seleccionadas por el proceso de eliminación hacia atrás para conformar el modelo lm_{OQ}-óptimo para Rincón del Bonete. Las variables seleccionadas se indican en color naranja.

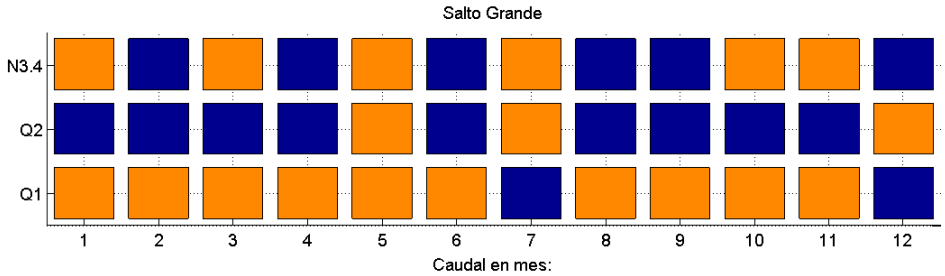

Figura 6.3.4: Idem Figura 6.3.3 para Salto Grande.

7. RESULTADOS DE SIMULACIONES CON MODELO DE CIRCULACIÓN GENERAL DE LA ATMÓSFERA

En la presente sección se evalúa la habilidad del MCGA UCLA para ser utilizado como herramienta de pronóstico de las condiciones atmosféricas en la región AS. En particular, es de interés evaluar la habilidad del modelo para predecir los índices atmosféricos identificados como predictores de caudales, es decir las primeras tres componentes principales del viento zonal, meridional y altura geopotencial en los distintos bimestres del año (Pc1, 2, 3 de u, v y hgt).

Una vez determinados sobre cuáles de los índices atmosféricos de interés el modelo tiene habilidad predictiva generaremos un subconjunto de variables predictoras de los caudales (las cuales serán en sí mismas potencialmente predictibles) conformado por: éstos índices, el índice N3.4 y los caudales antecedentes Q1 y Q2. Luego, con dicho conjunto reducido de variables predictoras y considerando que el modelo de regresión lineal acoplado con eliminación hacia atrás ha sido el que ha presentado los mejores resultados, ajustaremos dicho modelo para cada mes y embalse. Lo anterior será realizado ante dos situaciones de antecedencia del pronóstico: antecedencias que permiten contar con los caudales precedentes Q1 y Q2 y antecedencias que no lo permiten.

Es importante destacar que los resultados de esta sección, a diferencia de la anterior, dependen del modelo utilizado y que otros modelos podrían tener mayor o menor habilidad predictiva sobre las variables de interés.

La descripción del modelo y simulaciones fue realizada en la sección 2.5.

La presente sección está dividida en dos sub-secciones. En la primera de ellas se realizan consideraciones generales sobre la evaluación de pronósticos atmosféricos, se describe el procedimiento a utilizar para evaluar el potencial de predecir específicamente los índices en cuestión y, finalmente, se indica cuales de éstos son potencialmente predictibles con este modelo. En la segunda sub-sección se presentan los resultados del ajuste de los modelos lineales acoplados con selección de variables, utilizando únicamente aquellas variables contenidas en el subconjunto de variables potencialmente predictibles descrito antes.

7.1. Consideraciones generales sobre evaluación de pronósticos

Ante un pronóstico por ensemble es importante analizar, por un lado, la calidad de un pronóstico individual y, por otro, la dispersión de los miembros del ensemble, la cual ofrece una estimación de la incertidumbre asociada a la predictibilidad en sí. A continuación se describen algunas técnicas que permiten caracterizar, de forma sencilla, la calidad de un pronóstico individual.

Cuando el predictando es un campo atmosférico, para analizar la calidad de una predicción, suelen utilizarse métodos que analizan el pronóstico únicamente a través de los valores que éste toma en los puntos de una grilla espacial. La precisión del pronóstico es, usualmente, juzgada calculando medidas de la correspondencia entre los valores del campo pronosticado y el campo observado (o reanálisis, en este caso) en los diferentes puntos de la grilla. La correspondencia entre valores pronosticados y observados suele realizarse mediante medidas escalares. Aún para grillas de pequeño tamaño la evaluación de un pronóstico es un problema de alta dimensionalidad. Las tres

metodologías más utilizadas de esta clase son: error medio cuadrático, correlación de patrones (pattern correlation) y correlación de patrones anómalos (anomaly pattern correlation). En los próximos párrafos se brinda una breve descripción de cada una de estas técnicas. Por más detalles, discusiones y ejemplos dirigirse a Wilks (2006), texto en el que se basa este resumen.

En todo lo que sigue p_m y o_m denotan al pronóstico y la observación de un cierto campo atmosférico en el punto de grilla m, respectivamente. La grilla espacial cuenta con un total de M puntos.

El error cuadrático medio (MSE por su sigla en inglés) es el promedio espacial del cuadrado de las diferencias entre pronóstico y observación:

$$MSE = \frac{1}{M}\sum_{m=1}^{M}(p_m - o_m)^2$$

Evidentemente, ante el caso de un pronóstico perfecto el MSE es 0. En lugar de presentar el valor del MSE es usual presentar su raíz cuadrada (RMSE), para que las unidades sean las mismas de la variable en consideración.

En casos en los que se posean simulaciones de todo un período también es posible calcular el MSE (o RMSE) de las anomalías de un cierto campo atmosférico. En este caso se consideran como magnitudes a pronosticar las anomalías respecto de las climatologías simuladas y observadas.

La anomalía del campo simulado en el punto de grilla m se define como $ap_m = p_m - clima_m^{simulado}$ donde $clima_m^{simulado}$ denota al valor promedio que toma la variable p_m a lo largo de todo el período a evaluar.

Análogamente, la anomalía en el campo observado en el punto de grilla m se define como $ao_m = o_m - clima_m^{observado}$ donde $clima_m^{observado}$ denota al valor promedio que toma la variable o_m a lo largo de todo el período a evaluar.

Luego,

$$MSE_{anomalías} = \frac{1}{M}\sum_{m=1}^{M}(ap_m - ao_m)^2$$

Por pattern correlation se entiende al cálculo de la correlación entre el vector formado por los pronósticos en los M puntos de grilla y el vector formado por las observaciones en los M puntos de grilla.

Sean $p = (p_1, \ldots, p_M)$ y $o = (o_1, \ldots, o_M)$ los vectores de pronóstico y observación en los diferentes puntos de grilla. Luego,

$$Pattern\,Correlation = corr(p, o)$$

Esta medida puede variar en el intervalo [-1,1], donde 1 indica los resultados más positivos. El valor del pattern correlation está estrechamente relacionado con el parecido que los mapas de pronóstico y observación puedan tener. Sin embargo esta mediada tiene una importante desventaja: no penaliza sesgos. Es decir si, por ejemplo, en cada punto de grilla se pronostica un valor que resulta ser 10 veces el valor de la observación en ese punto, entonces, el pattern correlation valdrá 1 (indicando un

buen desempeño) a pesar de que desde otro punto de vista tal pronóstico podría ser considerado extremadamente insatisfactorio. Por lo tanto, es necesario considerar a la medida pattern correlation únicamente como un indicador parcial de habilidad.

En casos en los que se posean simulaciones de todo un período también es posible calcular lo que se denomina anomaly pattern correlation. Ésta es una medida similar a pattern correlation, pero que considera como variables a las anomalías respecto a las climatologías simulada y observada, respectivamente. Ante situaciones en las que no se posea información suficiente para generar la climatología simulada, es posible sustituir ésta por la observada.

Sean $ap = (ap_1, \ldots, ap_M)$ y $ao = (ao_1, \ldots, ao_M)$ los vectores de anomalía de pronóstico y anomalía de observación en los diferentes puntos de grilla. Luego,

$$Anomaly\ pattern\ correlation = corr(ap, ao)$$

Las tres medidas presentadas: error medio cuadrático, pattern correlation y anomaly pattern correlation, son medidas diseñadas para evaluar la calidad de un único pronóstico. En este caso es de interés analizar las simulaciones agrupadas según los bimestres del año para cada uno de los cuales se generaron 30 simulaciones (una por cada año entre 1979 y 2008) de cada uno de los 6 miembros del ensemble de simulaciones, totalizando 180 simulaciones por bimestre. Adicionalmente, una práctica usual en la metodología de pronóstico por ensemble consiste en promediar los miembros del ensemble para obtener un único pronóstico. Este pronóstico se denomina ensemble mean. Los beneficios de considerar el ensemble mean derivan del hecho que aspectos en los que hay desacuerdo entre los miembros del ensemble tienden a cancelarse y aspectos en los que hay acuerdo tienden a enfatizarse.

Por tanto, dividiremos el análisis de la calidad de las simulaciones según el bimestre del año para cada uno de los cuales contamos con 180 simulaciones más 30 resultantes del ensemble mean. En las Figuras 7.1.1, 7.1.2 y 7.1.3 se presentan, a modo de ejemplo, los valores de RMSE de anomalías, pattern correlation y anomaly pattern correlation para el viento zonal en la región AS. La grilla del modelo es interpolada a la grilla de los reanálisis, por lo que dentro de la región AS se consideran 357 puntos de grilla. Las Figuras se esquematizan utilizando el siguiente concepto de boxplot: el punto marcado con un círculo indica la mediana, los extremos superior e inferior de las barras gruesas indican los percentiles 75% y 25%, respectivamente y los extremos superior e inferior de las barras delgadas representan los percentiles 90% y 10%. Para cada bimestre los datos se organizan en las 180 realizaciones provenientes de los 6 miembros del ensamble y las 30 realizaciones asociadas al ensemble mean. Además para facilitar comparaciones, en el gráfico de RMSE (Figura 7.1.1) se agregan los resultados al considerar como pronóstico la climatología observada; es deseable que el modelo presente mayor habilidad predictiva que éste pronóstico simple. Por su parte, en el gráfico de pattern correlation (Figura 7.1.2) se indica la correlación entre los vectores formados por la climatología simulada y la observada.

Los RMSE de la anomalía de viento zonal (Figura 7.1.1) son de magnitudes y dispersiones similares en todos los bimestres del año. Se aprecia también que en 5 de los 12 bimestres del año la mediana de la simulación ensemble mean es menor a la análoga del pronóstico de climatología, siendo 4 de estos bimestres los que ocurren entre enero-febrero y abril-mayo.

Como era de esperar el gráfico de pattern correlation (Figura 7.1.2) presenta valores elevados de correlación que, a lo largo del año, siguen la forma de la curva de correlación entre las climatologías simulada y observada. En este sentido, se aprecia que la correlación entre las climatologías es

elevada, pero con un leve descenso en la temporada de invierno – comienzos de primavera.

El gráfico de anomaly pattern correlation (Figura 7.1.3) muestra gran dispersión en los resultados, indicando que existen años mejor simulados que otros.

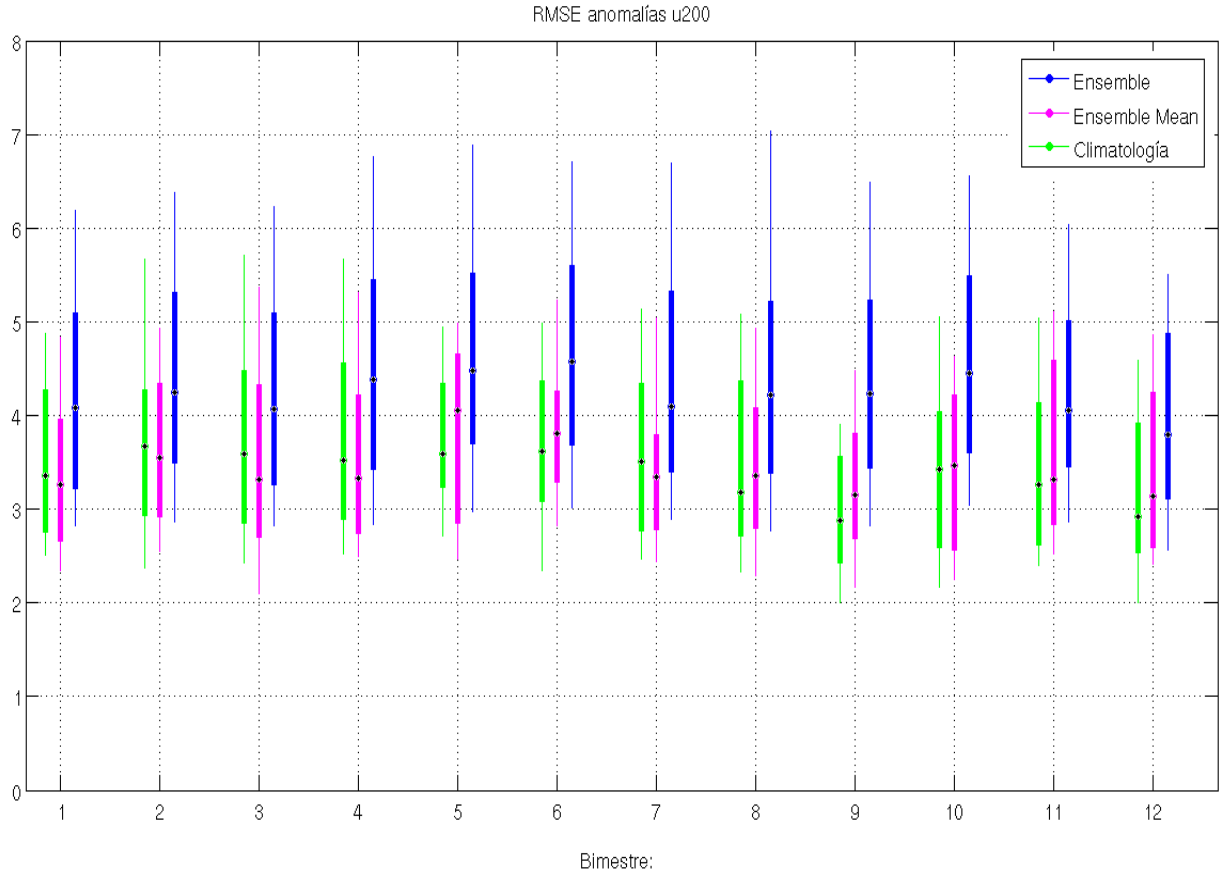

Figura 7.1.1: RMSE en el pronósticos de anomalías del viento zonal bimestral en la región AS. Las unidades son m/s. En el eje de las abscisas se indica el bimestre con la siguiente nomenclatura: bimestre (i) indica a los mese (i) e (i+1). Los datos se grafican en forma de boxplot donde el punto marcado con un círculo es la mediana, los extremos superior e inferior de las barras gruesas son los percentiles 75% y 25%, respectivamente y los extremos superior e inferior de las barras delgadas son los percentiles 90% y 10%, respectivamente.

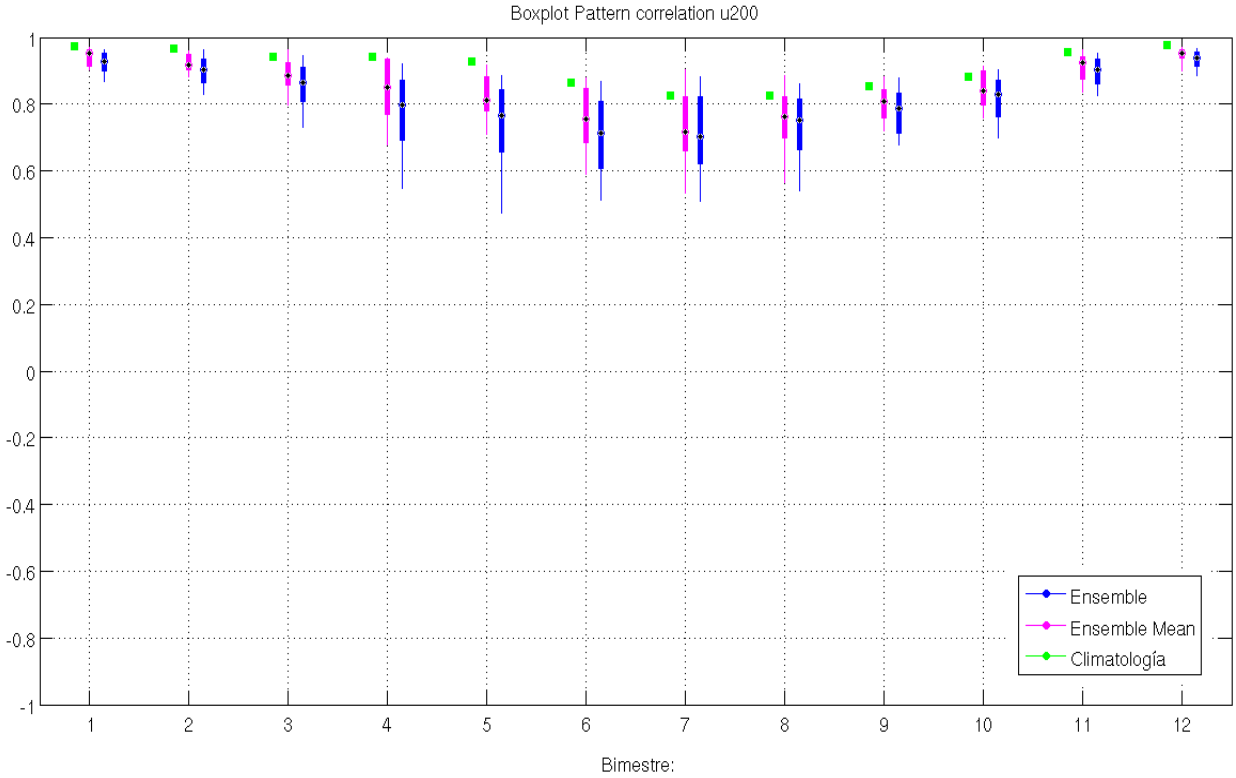

Figura 7.1.2: Idem Figura 7.1.1 pero para Pattern Correlation del viento zonal en AS.

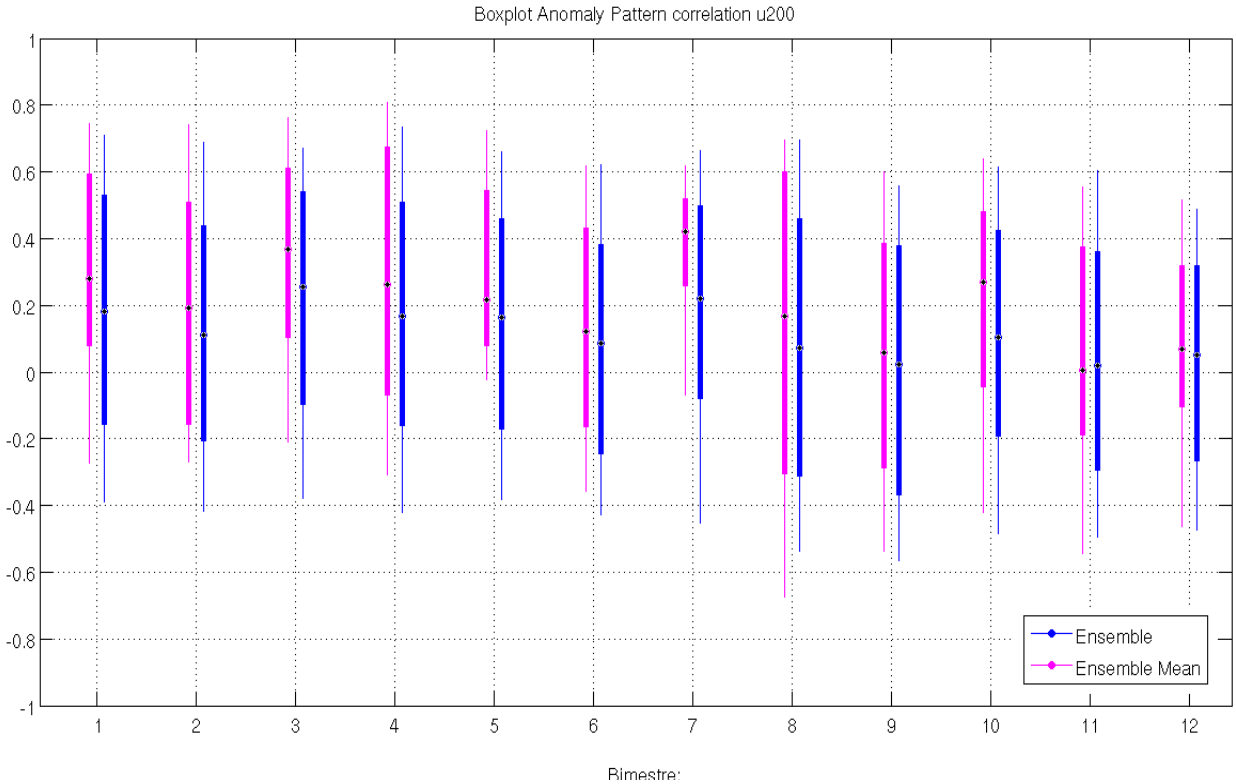

Figura 7.1.3: Idem Figura 7.1.1 para Anomaly Pattern Correlation del viento zonal en AS.

Cualquiera de las tres medidas RMSE, pattern correlation o anomaly pattern correlation pueden resultar extremadamente exigentes como indicadoras de la habilidad de un pronóstico. En nuestro

caso no es de interés tener habilidad predictiva sobre todos los aspectos de los campos atmosféricos u, v y hgt sino que sólo resulta necesario tener habilidad predictiva sobre las primeras tres componentes principales de los mismos. Considerando esto, se diseña un procedimiento para evaluar la habilidad del modelo en pronosticar las mencionadas componentes principales.

El método a emplear para evaluar la habilidad del modelo en simular una componente principal se desarrolla de forma independiente para cada bimestre del año y considerando, únicamente, la simulación ensemble mean. A continuación describimos el procedimiento a seguir para generar las que daremos en llamar Pc-ensemble mean. Dado un cierto bimestre, se cuenta con el resultado de la simulación ensemble mean para todos los años entre 1979 y 2008. Fijado el bimestre, para la corrida ensemble mean, se calculan la climatología simulada del campo u y las anomalías simuladas de u para cada uno de los años entre 1979 y 2008. Por otro lado, junto con Pc1u (observada) ya se calculó el campo de eof observado asociado: eof1u (ver, por ejemplo, Figura 4.1.2.4). Finalmente, Pc1u-ensemble mean se define como la proyección de la anomalía simulada por el ensemble mean de u sobre el patrón de eof1u observado. Luego, para cada bimestre del año se tiene una serie temporal de 30 elementos (uno por cada año entre 1979 y 2008) denominada Pc1u-ensemble mean.

En la Figura 7.1.4 se presentan, para cada bimestre del año, las correlaciones entre las diferentes Pc observadas y las Pc-ensemble mean. Por coherencia con la notación de secciones precedentes en el eje de las abscisas, en lugar de indicar el bimestre al cual corresponde las Pc, se indica el mes del caudal que se desea predecir con ella; en otras palabras, la Pc del bimestre mes (i)-mes (i+1) corresponde con la abscisa del caudal en el mes (i+1). Al igual que antes, en la Figura 7.1.4, las correlaciones no significativas (o significativas, pero negativas) se indican con color azul oscuro. Claramente la temporada con correlaciones más pobres entre las Pc observadas y simuladas es finales de primavera- comienzo de verano: en octubre y noviembre sólo 1 de las variables simuladas presenta correlación significativa y en diciembre ninguna. Por otro lado, la temporada con mejores correlaciones va desde febrero a mayo donde existen varias variables con valores elevados de correlaciones.

Finalmente, definimos como Pc potencialmente predecible por el modelo a aquella cuya correlación con la Pc-ensemble mean correspondiente sea significativa, tal como es indicado en la Figura 7.1.4. Se recuerda que estos resultados dependen del modelo utilizado.

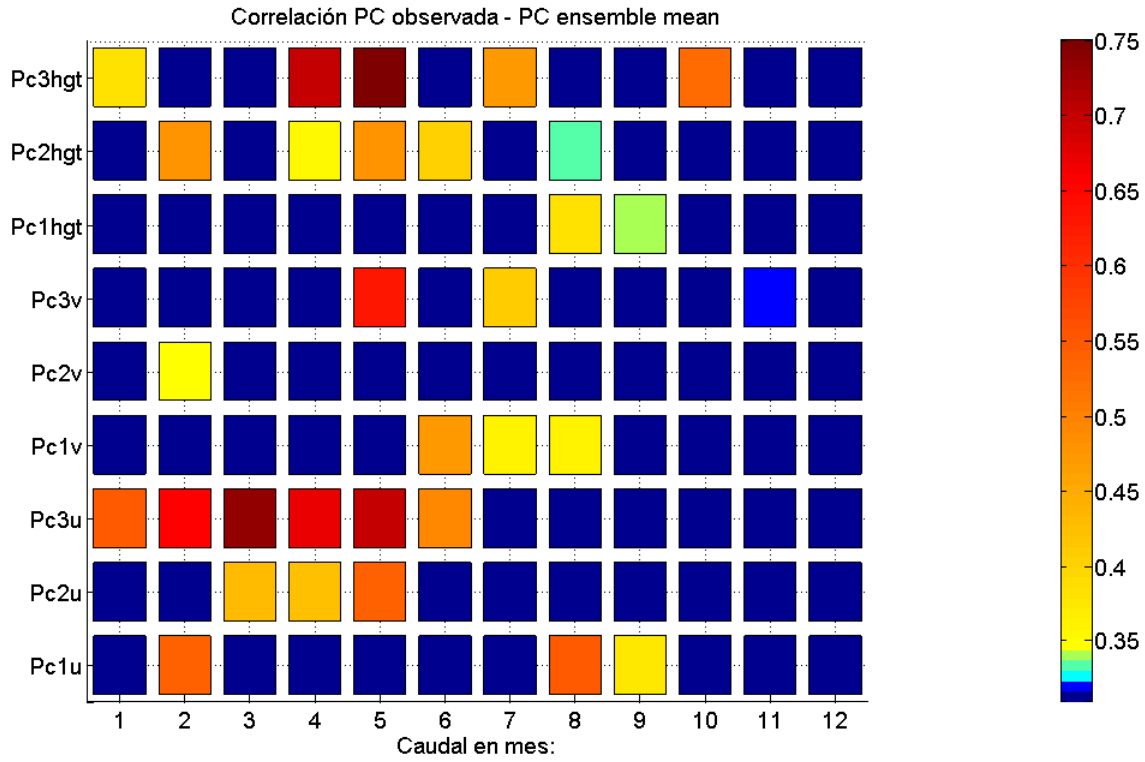

Figura 7.1.4: Correlaciones entre las Pc observadas y Pc-ensemble mean. Correlaciones no significativas al nivel de 95% o negativas se indican en color azul oscuro. El eje de las abscisas indica el mes del caudal que se desea predecir utilizando una Pc, es decir que la Pc del bimestre mes (i)-mes (i+1) se corresponde con la abscisa del caudal en el mes (i+1).

7.2. Ajuste de modelos utilizando únicamente variables potencialmente predictibles

Para cada mes del año seleccionamos a partir del conjunto inicial de variables predictoras de caudales (las 9 variables atmosféricas, la variable oceánica y las 2 variables de caudal precedente) únicamente aquellas que son potencialmente predictibles. Es decir que, para cada mes del año, se genera el grupo de variables potencialmente predictibles con: las variables atmosféricas que el MCGA-UCLA tiene potencial de predecir, la variable oceánica N3.4 y los caudales precedentes si es que la antecedencia del pronóstico a realizar permite disponer de Q1 y Q2.

Utilizando únicamente las variables potencialmente predictibles ajustaremos, nuevamente, el modelo lineal acoplado con el procedimiento de selección de variables hacia atrás y seleccionaremos el modelo óptimo según el criterio de menor error cv. Los resultados se presentan, por separado, para cada una de las dos situaciones de antecedencia del pronóstico: antecedencias que permiten disponer de los caudales previos y antecedencias que no lo permiten.

En las Figuras 7.2.1 y 7.2.2 se presentan los resultados para antecedencias que permiten disponer de Q1 y Q2 para Rincón del Bonete y Salto Grande, respectivamente. Para facilitar comparaciones, en los gráficos se indican los resultados para el modelo lineal óptimo ($lm_{A^{UCLA}_{OQ}}$-óptimo), obtenido realizando eliminación hacia atrás a partir de el conjunto de predictores potencialmente predictibles

y para el modelo lineal óptimo obtenido a partir de la utilización de todos los predictores atmosféricos, oceánicos y de caudal precedente (lm_{AOQ}-óptimo). Al igual que antes, todos los resultados de error cv son normalizados por el error cv del modelo ymedio.

Para Rincón del Bonete (Figura 7.2.1), el modelo $lm_{A}{}^{UCLA}{}_{OQ}$-óptimo presenta mayor habilidad predictiva que el modelo ymedio en todos los meses del año, salvo en agosto cuando estas habilidades coinciden. Los períodos de mayor habilidad predictiva respecto a la análoga del modelo ymedio son entre febrero y abril, junio y julio. Las mayores diferencias en desempeño entre los modelos $lm_{A}{}^{UCLA}{}_{OQ}$-óptimo y lm_{AOQ}-óptimo se presentan en el mes de diciembre.

Para Salto Grande (7.2.2) el modelo $lm_{A}{}^{UCLA}{}_{OQ}$-óptimo supera en habilidad predictiva a ymedio en todos los meses del año. Las temporadas de marzo a agosto y noviembre-diciembre destacan como aquellas en las que la habilidad predictiva es mayor, respecto de la análoga para ymedio. Para este embalse las mayores pérdidas de habilidad por considerar únicamente los predictores atmosféricos potencialmente predictibles por el modelo de UCLA se concentran en los meses de octubre y abril.

En las Figuras 7.2.3 y 7.2.4 se indican las variables seleccionadas por el modelo $lm_{A}{}^{UCLA}{}_{OQ}$-óptimo para cada mes del año para Rincón del Bonete y Salto Grande, respectivamente. Las variables que no se encuentran disponibles para seleccionar (por no pertenecer al conjunto de variables potencialmente predictibles) son indicadas en color blanco, las seleccionadas en naranja y las disponibles pero no seleccionadas en azul oscuro.

Para Rincón del Bonete (Figura 7.2.3) la variable más seleccionada es Q1; N3.4 es seleccionado de forma persistente en la temporada noviembre-febrero y la única temporada en la que variables atmosféricas son seleccionadas es de febrero a julio.

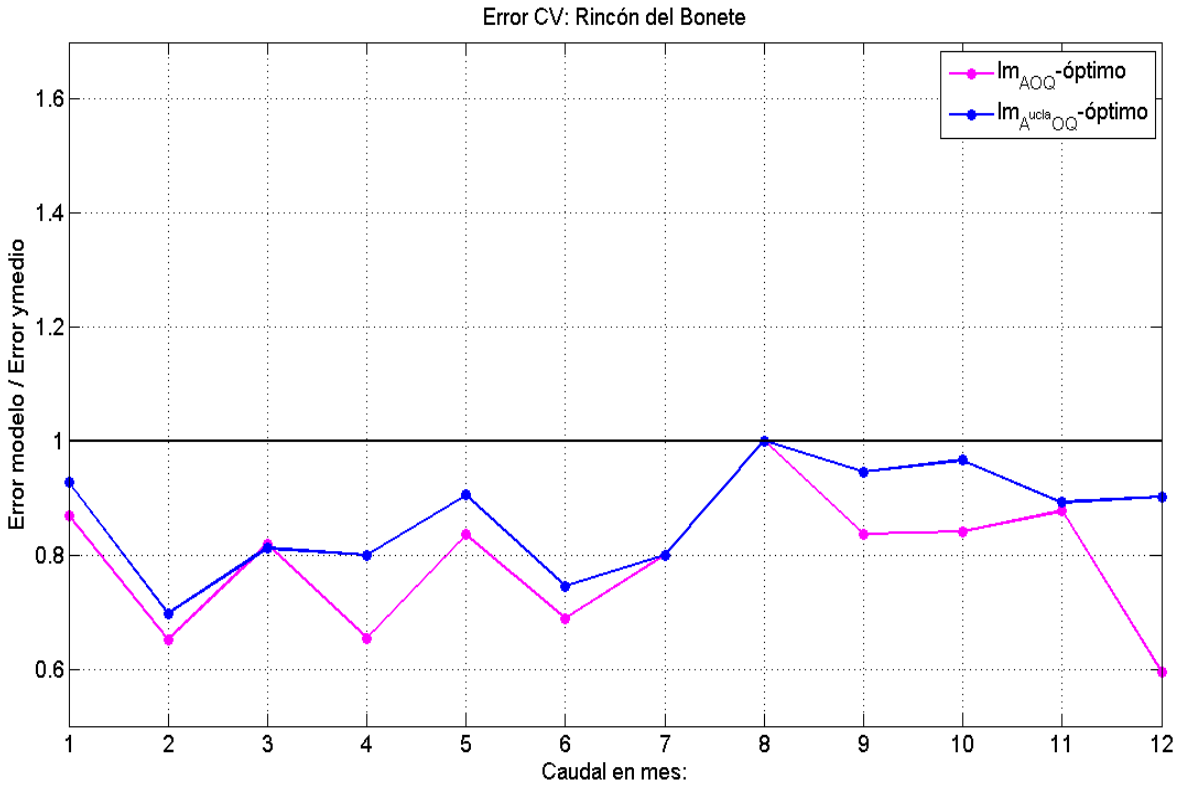

Figura 7.2.1: Errores cv de los modelos $lm_A{}^{UCLA}{}_{OQ}$-óptimo y lm_{AOQ}-óptimo para Rincón del Bonete. Los errores se expresan como el cociente por el error cv del modelo ymedio. La línea negra indica errores iguales a los del modelo ymedio.

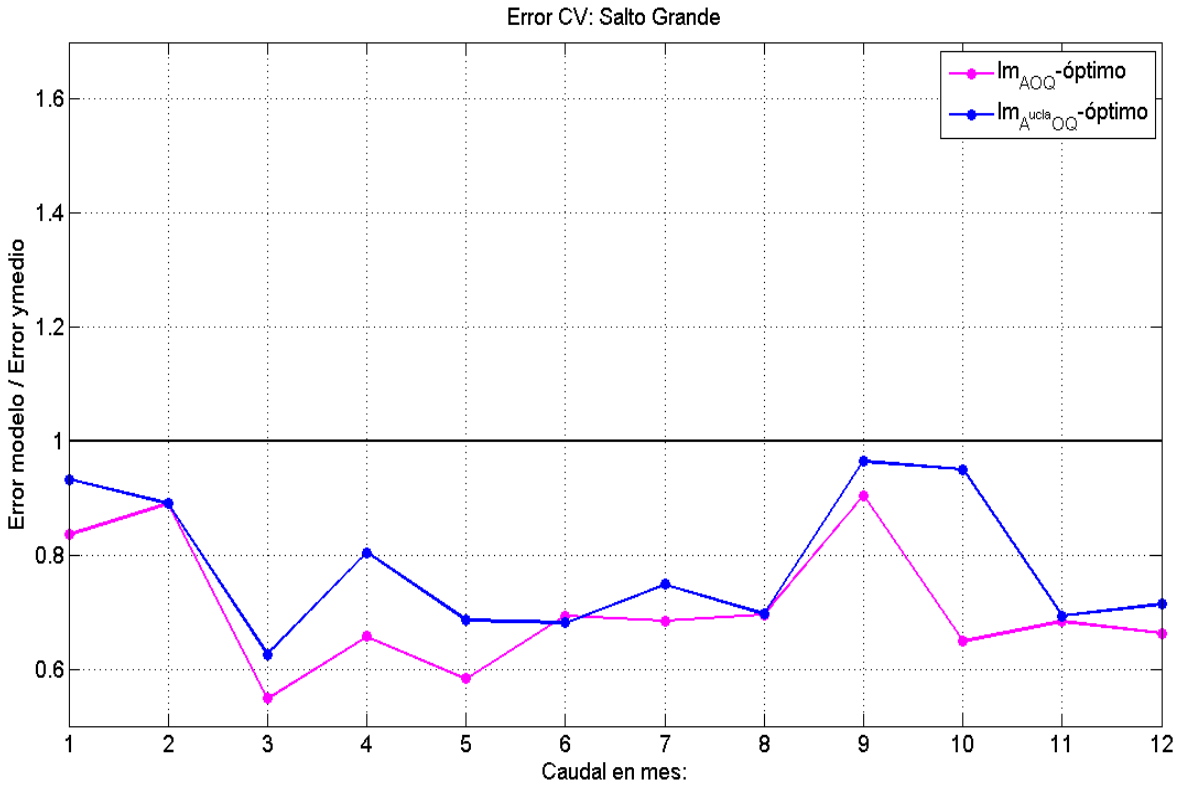

Figura 7.2.2: Idem Figura 7.2.1 para Salto Grande.

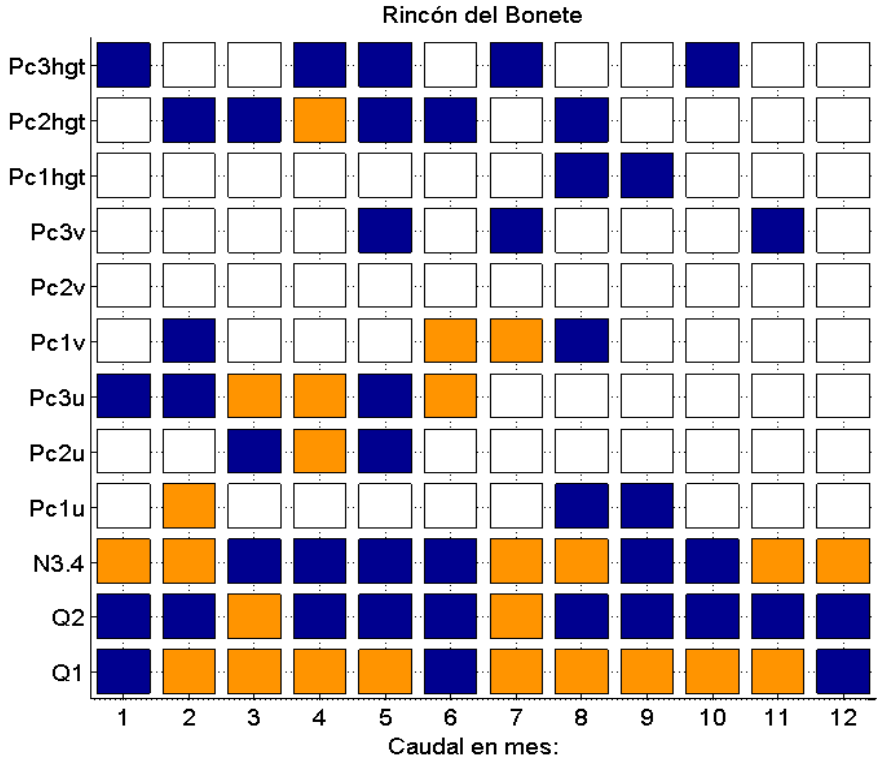

Figura 7.2.3: Variables seleccionadas por el proceso de eliminación hacia atrás para conformar el modelo $lm_{A^{UCLA}_{OQ}}$-óptimo para Rincón del Bonete. Las variables seleccionadas se indican en color naranja y las que no están presentes para seleccionar (por no pertenecer al conjunto de variables potencialmente predecibles) es color blanco.

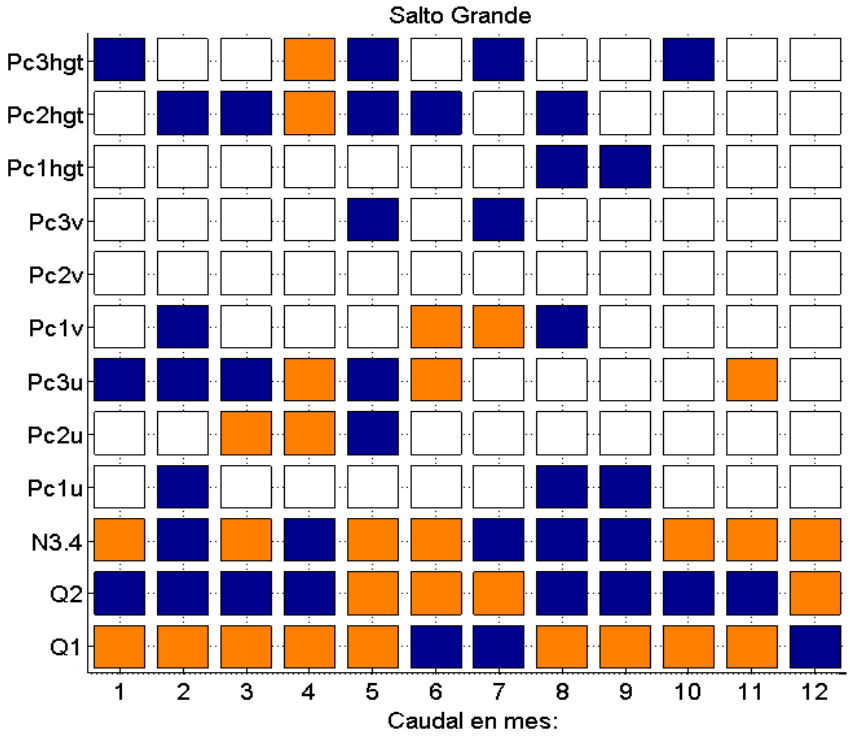

Figura 7.2.4: Idem Figura 7.2.3 para Salto Grande.

Por su parte para Salto Grande (Figura 7.2.4) en todos los meses del año al menos una de las dos variables de caudales precedentes es seleccionada. N3.4 es seleccionada de forma continua entre octubre y enero y, de forma similar a lo ocurrido en Rincón del Bonete, la selección de variables atmosféricas se restringe al período marzo-julio.

En las Figuras 7.2.5 y 7.2.6 se presentan los resultados de error cv para Rincón del Bonete y Salto Grande utilizando, únicamente, las variables potencialmente predictibles atmosféricas y oceánicas, es decir, sin utilizar ni Q1 ni Q2. En ambas figuras se presentan los resultados para el modelo lineal acoplado con selección de variables óptimo ($lm_{A^{UCLA}O}$-óptimo) y el modelo lineal óptimo generado a partir de todos los predictores atmosféricos y oceánicos (lm_{AO}-óptimo).

Para Rincón del Bonete (Figura 7.2.5) el modelo $lm_{A^{UCLA}O}$-óptimo presenta resultados superiores a los del modelo ymedio en todos los meses salvo la temporada agosto-octubre donde las habilidades de los mismos son muy similares. En los restantes meses la habilidad del modelo respecto a la habilidad de ymedio no presenta demasiadas diferencias entre los distintos meses, aunque se destaca el mes de junio como el mes en el que el cociente de los errores cv es menor. Las diferencias en habilidad entre los modelos $lm_{A^{UCLA}O}$-óptimo y lm_{AO}-óptimo son importantes en todos los meses, siendo abril el mes en el que ésta diferencia es mayor.

Para Salto Grande (Figura 7.2.6) el modelo $lm_{A^{UCLA}O}$-óptimo presenta una mayor habilidad predictiva que el modelo ymedio en todos los meses salvo febrero y setiembre. Se destacan períodos de elevada predictibilidad entre marzo y junio, noviembre y diciembre. La pérdida de habilidad del modelo $lm_{A^{UCLA}O}$-óptimo respecto del lm_{AO}-óptimo es más notoria entre marzo y mayo y entre octubre y diciembre.

En las Figuras 7.2.7 y 7.2.8 se presentan las variables seleccionadas por los modelo $lm_{A^{UCLA}O}$-óptimo para Rincón del Bonete y Salto Grande, respectivamente. Se sigue el mismo criterio que para las Figuras 7.2.3 y 7.2.4.

Para Rincón del Bonete (Figura 7.2.7) se destaca el período noviembre-febrero en el que la variable N3.4 se seleccionada continuamente. Para Salto Grande (Figura 7.2.8) vale la pena resaltar que N3.4 es seleccionada en 9 de los 12 meses.

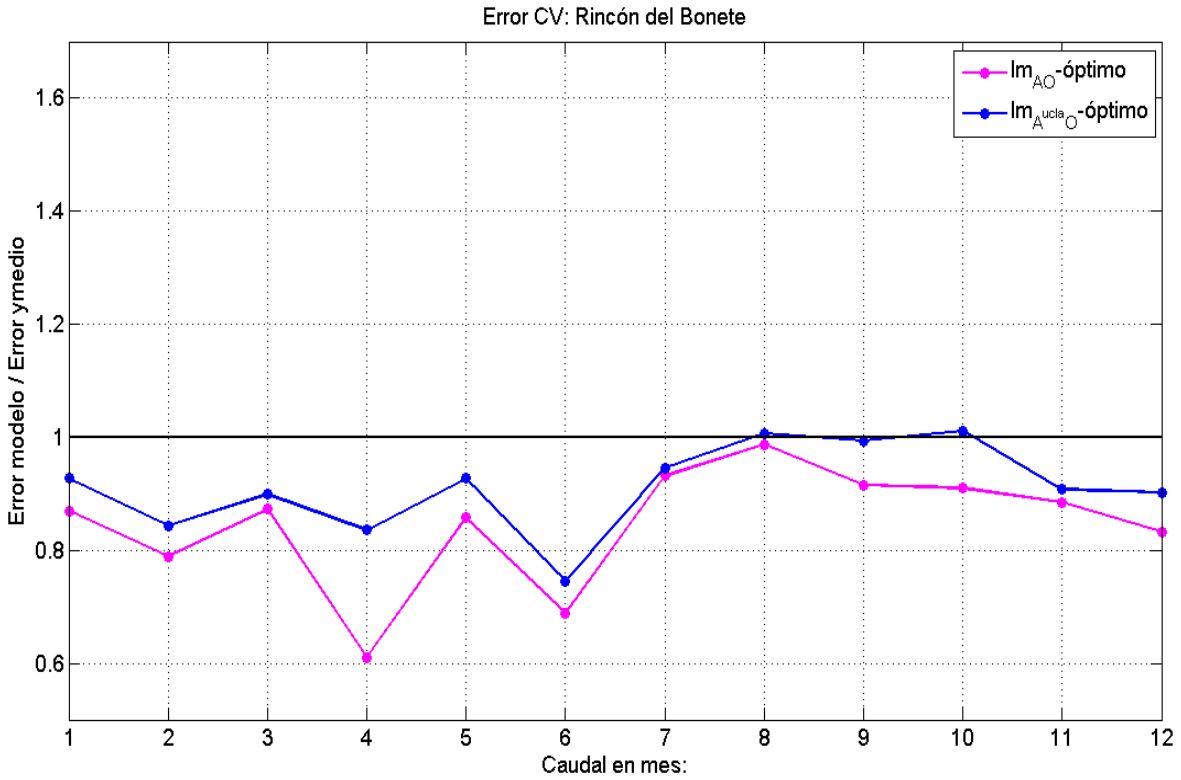

Figura 7.2.5: Errores cv de los modelos lm$_A{}^{UCLA}{}_O$-óptimo y lm$_{AO}$-óptimo para Rincón del Bonete. Los errores se expresan como el cociente por el error cv del modelo ymedio. La línea negra indica errores iguales a los del modelo ymedio.

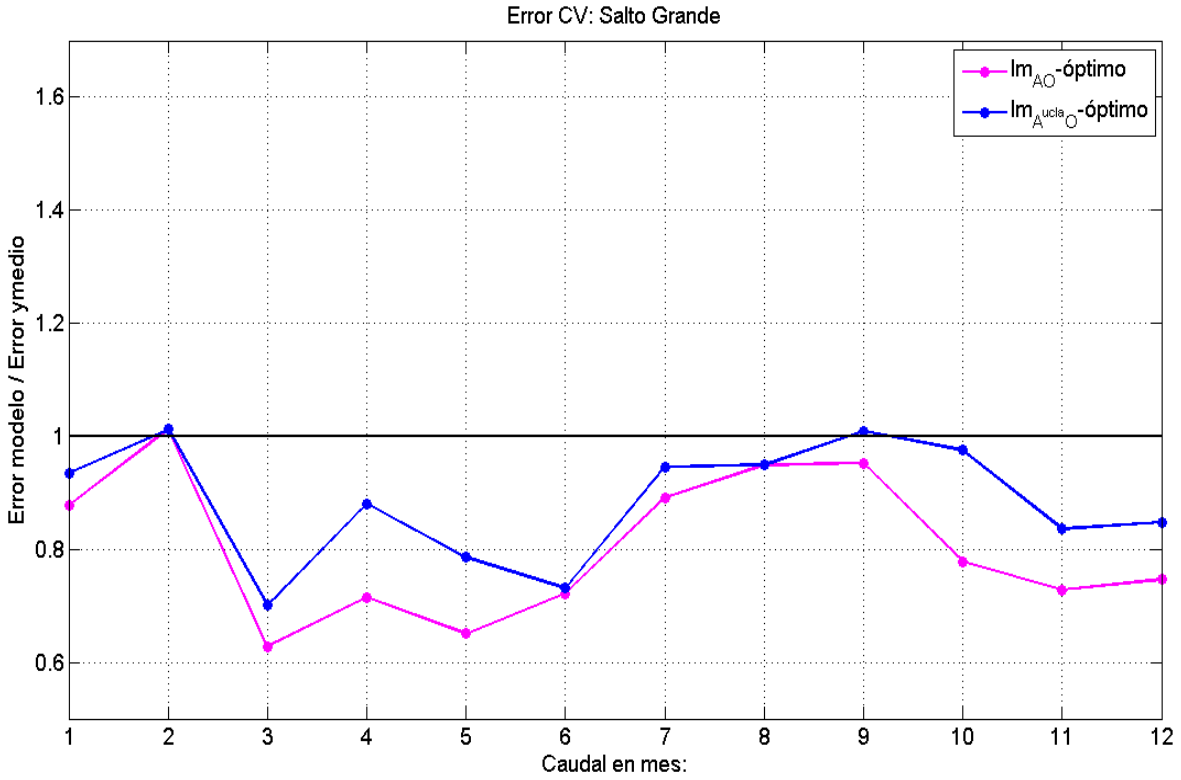

Figura 7.2.6: Idem Figura 7.2.5 para Salto Grande.

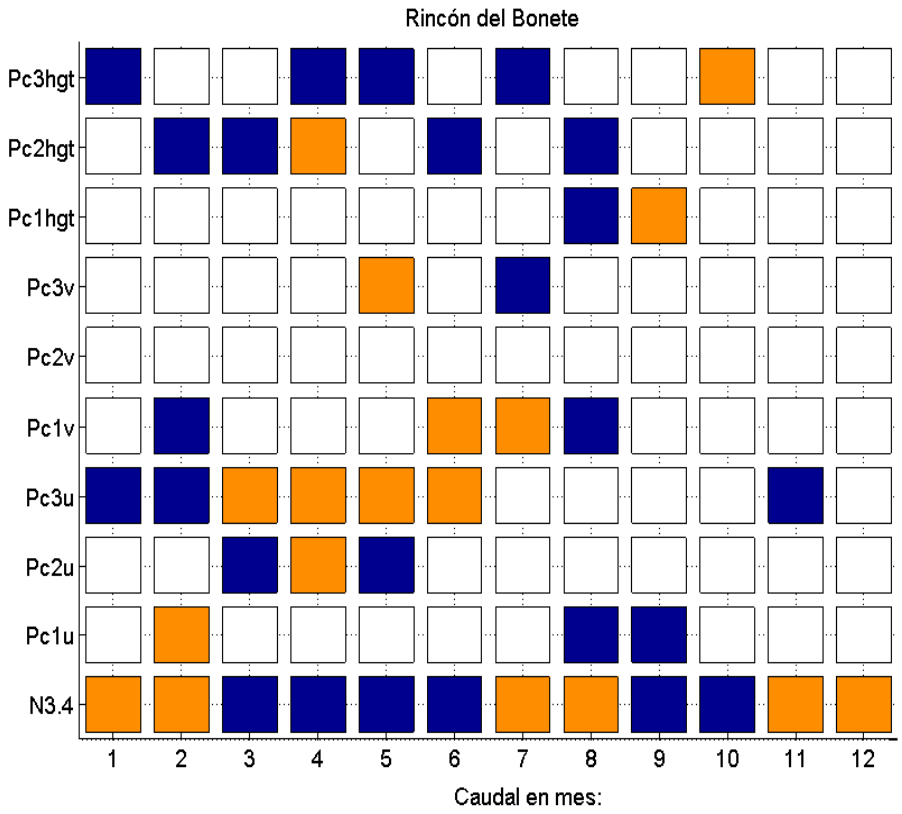

Figura 7.2.7: Variables seleccionadas por el proceso de eliminación hacia atrás para conformar el modelo $lm_A{}^{UCLA}{}_O$-óptimo para Rincón del Bonete. Las variables seleccionadas se indican en color naranja y las que no están presentes para seleccionar (por no pertenecer al conjunto de variables potencialmente predecibles) es color blanco.

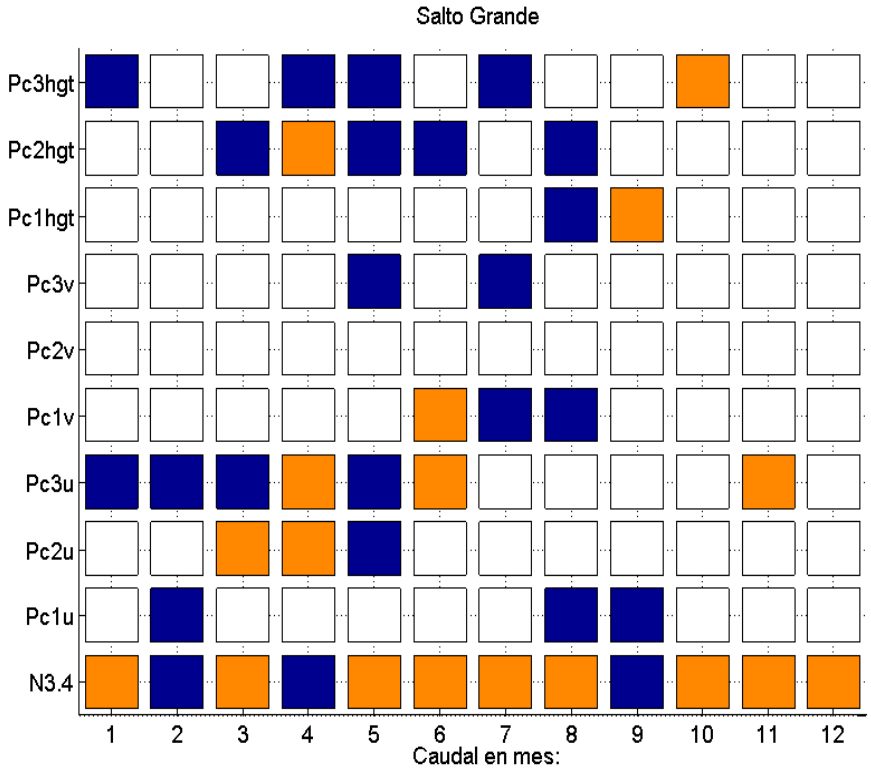

Figura 7.2.8: Idem que Figura 7.2.7 para Salto Grande.

8. RESUMEN DE RESULTADOS Y CONCLUSIONES

A partir del análisis de la circulación atmosférica regional, índices asociados al fenómeno ENOS e información relacionada con caudales antecedentes se determinó, para cada mes y embalse, un conjunto inicial de 12 variables predictoras de caudales: 9 asociadas a la circulación atmosférica regional (variables del grupo A, por atmosféricas), 1 asociada al fenómeno ENOS (variable del grupo O, por oceánica), Q1 y Q2 (variables del grupo Q, por caudales). La disponibilidad de las variables del grupo Q depende de la antelación con la que se desee realizar el pronóstico del caudal.

Según la clasificación presentada en la introducción, los sistemas de predicción que involucren variables de los grupos A, Q y O pertenecen a la categoría de esquemas de downscaling híbrido (Figura 1.2); por su parte, aquellos que no involucren a predictores del grupo A pertenecen a la categoría de predicción orientada puramente por datos (Figura 1.1).

Para cada mes y embalse se ajustaron varios modelos estadísticos evaluando, en cada caso, su desempeño predictivo. Utilizando como estimador del error de predicción al error cross validation leave-one-out (cv) se encontró que, en general, de entre los modelos ajustados el que presenta los mejores resultados es el lineal acoplado con la técnica de selección de variables hacia atrás, determinando la cantidad óptima de variables a incluir en el modelo mediante minimización del error cv (modelo lineal óptimo). En cuanto a los restantes modelos se destaca que PLSR-óptimo mostró un desempeño apenas inferior al modelo lineal óptimo, por lo que su utilización en problemas similares es también recomendable. Por el contrario, los modelos de árboles de regresión y redes neuronales no presentaron resultados satisfactorios en cuanto a habilidad predictiva; dado que ambas técnicas tienen, en principio, el potencial de detectar cualquier tipo de relación entre predictores y predictando (y no sólo relaciones lineales) su desempeño podría mejorar notoriamente si se contara con una mayor cantidad de observaciones. Por último, la metodología de predicción por clustering presentó resultados intermedios los cuales quizás también podrían mejorar ante la inclusión de una mayor cantidad de observaciones. Cabe recordar que en este tipo de aplicaciones las relaciones entre las variables pueden modificarse sustancialmente con el tiempo y, por ende, no siempre todas las observaciones disponibles podrán considerarse representativas del clima que se desea predecir.

A continuación presentamos un resumen de los resultados del error cv únicamente para los modelos lineales óptimos, segregados según si entre las variables predictoras iniciales están presentes las del grupo Q o no. Además, se presentan los resultados cuando en lugar de utilizar todos los predictores contenidos en el grupo A se utilizan sólo aquellos con mayor predictibilidad, según indican las simulaciones realizadas con el MCGA-UCLA. Se recuerda que en todos los resultados las variables predictoras se suponen conocidas (o perfectamente predictibles). En modo operativo la introducción de pronósticos imperfectos de los predictores generará, posiblemente, mayores errores en el sistema de predicción.

Las Figuras 8.1 y 8.2 muestran los resultados en situaciones de antecedencia de pronóstico que permiten utilizar a los caudales precedentes como variables predictoras para Rincón del Bonete y Salto Grande, respectivamente. En estas Figuras la diferencia entre los errores cv de los modelos lm_{AOQ}-óptimo y $lm_{A}{}^{UCLA}{}_{OQ}$-óptimo representa la pérdida de habilidad predictiva en la que se incurre al utilizar únicamente aquellos predictores atmosféricos que el MCGA-UCLA indica serían predictibles. Por su parte, la diferencia entre los errores cv de los modelos lm_{AOQ}-óptimo y lm_{OQ}-óptimo puede utilizarse como indicador de la importancia relativa de incluir predictores

atmosféricos en el esquema de predicción. Así mismo, el desempeño relativo del modelo lm_{AOQ}-óptimo respecto al modelo ymedio puede considerarse como un estimador del grado de predictibilidad de los caudales mensuales: en caso de que el desempeño de lm_{AOQ}-óptimo supere al del modelo ymedio diremos que existe predictibilidad y en caso contrario que no existe predictibilidad.

Para Rincón del Bonete (Figura 8.1) el modelo lm_{AOQ}-óptimo presenta habilidad predictiva superior al modelo ymedio en todos los meses del año, exceptuando agosto. Para este modelo y embalse no resulta clara la existencia de períodos de elevada predictibilidad (bajo error cv). La habilidad predictiva del modelo $lm_{A}{}^{UCLA}{}_{OQ}$-óptimo es claramente inferior a la del lm_{AOQ}-óptimo, aunque esta habilidad supera a la del modelo ymedio en todos los meses del año salvo agosto. La mayor diferencia en habilidad predictiva entre los modelos lm_{AOQ}-óptimo y $lm_{A}{}^{UCLA}{}_{OQ}$-óptimo se manifiesta en el mes de diciembre. Si no se considera ninguno de los predictores del grupo A (lm_{OQ}-óptimo) la habilidad predictiva desciende aún más, aunque también supera al modelo ymedio en todas las ocasiones a excepción de agosto. Se observa que las curvas de error cv para los modelos $lm_{A}{}^{UCLA}{}_{OQ}$-óptimo y lm_{OQ}-óptimo coinciden desde agosto a enero y en mayo; este comportamiento indica que, en los citados meses, o bien ninguna de las variables del grupo A tiene potencial de ser pronosticada por el modelo UCLA o bien aquellas que sí son potencialmente predictibles por el modelo no son seleccionadas por el procedimiento de selección de variables utilizado.

Para Salto Grande (Figura 8.2) el modelo lm_{AOQ}-óptimo presenta habilidad predictiva superior al modelo ymedio en todos los meses del año. Con éste modelo se destacan dos períodos de elevada predictibilidad: de marzo a mayo y de octubre a diciembre. Los meses donde la ganancia de predictibilidad respecto a la utilización del modelo ymedio es menor son: enero, febrero y setiembre. Para este embalse, el modelo $lm_{A}{}^{UCLA}{}_{OQ}$-óptimo también tiene habilidad predictiva superior a ymedio en todos los meses del año. La mayor pérdida de habilidad al pasar del conjunto A al A^{UCLA} se observa en el mes de octubre. Por su parte, el modelo lm_{OQ}-óptimo también presenta un desempeño superior al modelo ymedio en todos los meses del año. Para este embalse el modelo lm_{OQ}-óptimo sólo se diferencia del $lm_{A}{}^{UCLA}{}_{OQ}$-óptimo entre marzo y julio, período en el cual es clara la importancia de la inclusión de predictores atmosféricos.

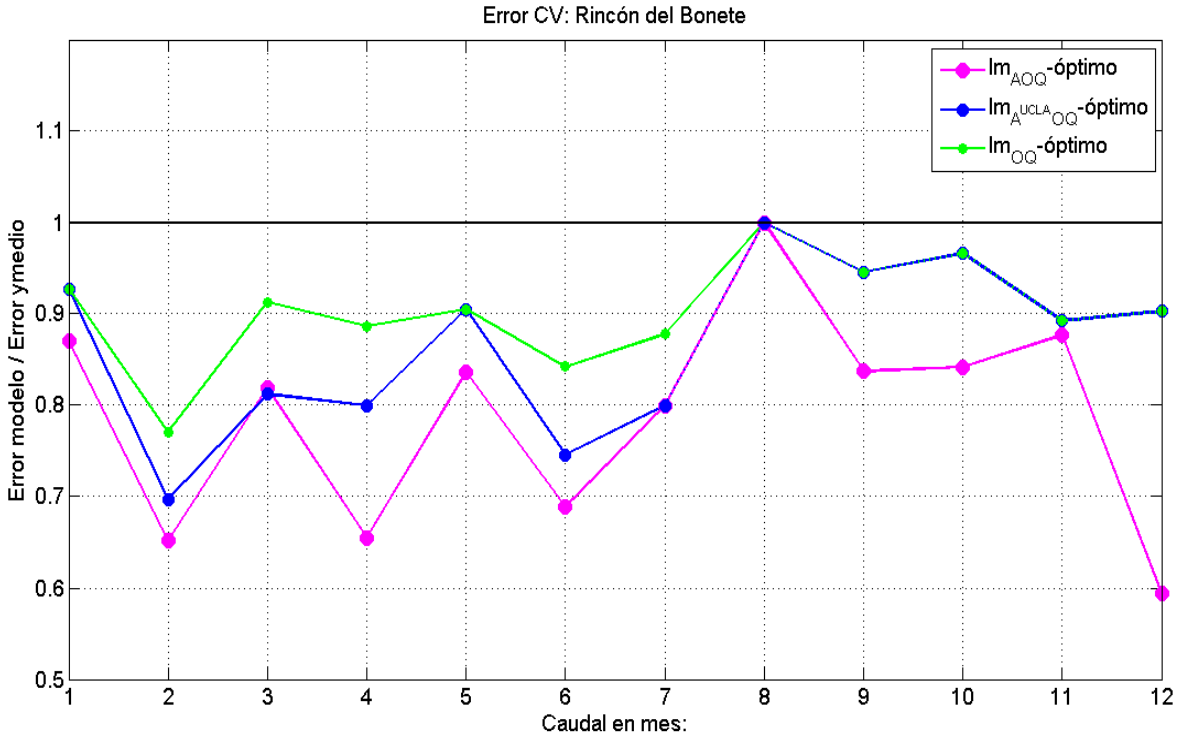

Figura 8.1: Errores cv de los modelos lm_{AOQ}-óptimo, $lm_A{}^{UCLA}{}_{OQ}$-óptimo y lm_{OQ}-óptimo para Rincón del Bonete. Los errores se expresan como el cociente por el error cv del modelo ymedio. La línea negra indica errores iguales a los del modelo ymedio.

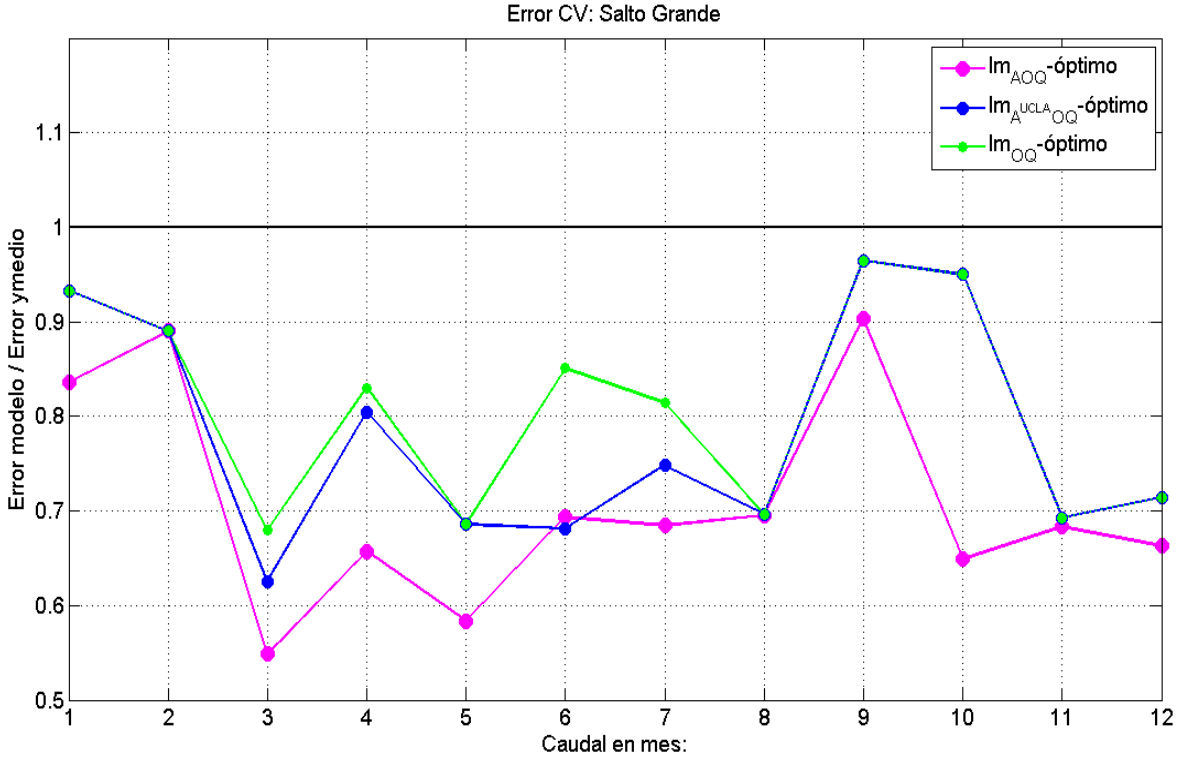

Figura 8.2: Idem Figura 8.1 para Salto Grande.

En las Figuras 8.3 y 8.4 se presentan los principales resultados para situaciones de antecedencia de pronóstico que no permiten disponer ni de Q1 ni de Q2 para Rincón del Bonete y Salto Grande, respectivamente.

Para Rincón del Bonete (Figura 8.3) el modelo lm_{AO}-óptimo presenta un desempeño superior al del modelo ymedio en todos los meses. Los meses donde la habilidad predictiva del modelo lm_{AO}-óptimo, relativa al modelo ymedio, es superior son abril y junio. El modelo $lm_A{}^{UCLA}{}_O$-óptimo presenta un desempeño superior al de ymedio en todos los meses a excepción de agosto y octubre. La mayor diferencia entre los errores cv de lm_{AO}-óptimo y $lm_A{}^{UCLA}{}_O$-óptimo se presenta en abril. Por su parte, el modelo lm_O tiene habilidad superior al modelo ymedio únicamente entre noviembre y abril. La importancia de la inclusión de los predictores atmosféricos es destacable, sobretodo, en la temporada de otoño-invierno: de marzo a julio.

Para Salto Grande (Figura 8.4) el modelo lm_{AO}-óptimo muestra habilidad predictiva superior al modelo ymedio durante todo el año salvo en el mes de febrero. Con este modelo se destacan dos períodos donde la mejora del desempeño respecto al modelo ymedio es muy importante: de marzo a junio (otoño) y de octubre a diciembre (primavera). Por su parte el modelo $lm_A{}^{UCLA}{}_O$-óptimo no supera en desempeño a ymedio en dos ocasiones: febrero y setiembre; las mayores diferencias en las habilidades predictivas de los modelos lm_{AO}-óptimo y $lm_A{}^{UCLA}{}_O$-óptimo se dan, justamente, en otoño y primavera. El modelo lm_O no supera en desempeño a ymedio en los mismos meses que $lm_A{}^{UCLA}{}_O$-óptimo y coincide con este modelo en 6 ocasiones: de noviembre a febrero, julio y agosto. Se destaca que las diferencias entre los modelos lm_{AO}-óptimo y lm_O son máximas de marzo a junio y de octubre a diciembre y que, por lo tanto, es en estos dos períodos que la inclusión de predictores atmosféricos se hace extremadamente beneficiosa.

En definitiva, si consideramos que el desempeño del modelo lm_{AOQ}-óptimo puede utilizarse como estimador de la predictibilidad se concluye que: tanto en Rincón del Bonete como en Salto Grande los caudales de aporte medios mensuales son predictibles en todos los meses del año, exceptuando el caudal de aporte a Rincón del Bonete en agosto. Si bien para Rincón del Bonete no se distingue claramente un período de elevada predictibilidad, para Salto Grande las temporadas de marzo a mayo y de octubre a diciembre destacan como robustas en este sentido.

En el contexto de predictores conocidos y a través de la utilización del modelo lineal acoplado con selección de variables fue posible construir esquemas de predicción que presentan, en general, habilidad predictiva superior a la de pronosticar la media histórica, aún en situaciones de antecedencia del pronóstico que no permiten contar con los caudales precedentes Q1 y Q2, es decir, antecedencias superiores a los dos meses.

Si bien todo el trabajo fue realizado bajo la hipótesis de predictores conocidos también se utilizó el MCGA-UCLA para evaluar desempeños restringiendo las variables predictoras atmosféricas a aquellas que dicho modelo indica podrían ser predictibles, lo cual constituye una situación más cercana a la que debe ser enfrentada en modo operacional. Aún cuando el conjunto de predictores se restringe a aquellos potencialmente predictibles (según indica el análisis con el MCGA-UCLA) y N3.4 (cuya predictibilidad a varios meses es muy alta), los modelos desarrollados muestran habilidad predictiva superior a la de pronosticar la media histórica en ambos embalses en la mayoría de los meses, incluso bajo situaciones de antecedencia superiores a los dos meses. Aunque estos resultados deben considerarse cotas superiores de la habilidad predictiva que los modelos puedan tener en modo operacional los mismos son alentadores.

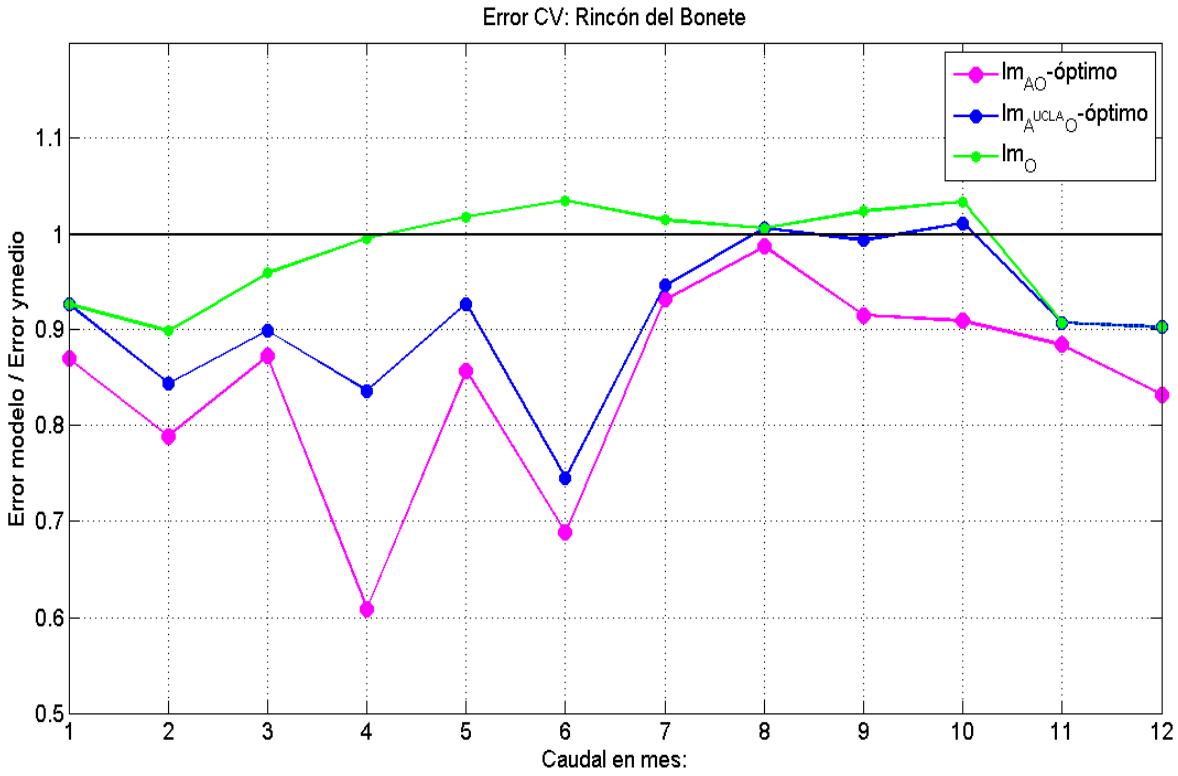

Figura 8.3: Errores cv de los modelos lm_{AO}-óptimo, $lm_A^{UCLA}{}_O$-óptimo y lm_O para Rincón del Bonete. Los errores se expresan como el cociente por el error cv del modelo ymedio. La línea negra indica errores iguales a los del modelo ymedio.

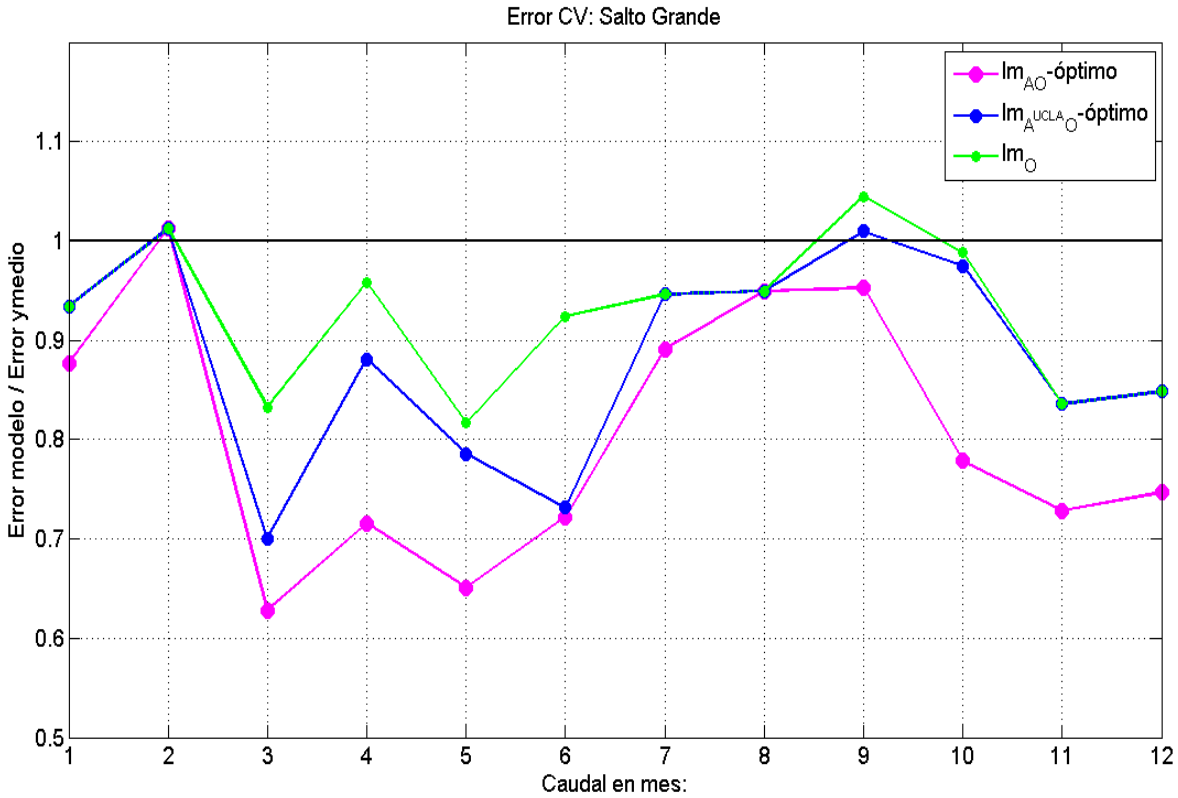

Figura 8.4: Idem Figura 8.3 para Salto Grande.

Para obtener una estimación real de la habilidad predictiva de los modelos generados la hipótesis de predictores conocidos debe suprimirse, realizando lo que se conoce como ejercicio de pronóstico retrospectivo. En particular, para sistemas de predicción que involucren variables del grupo A (esquemas de downscaling híbrido) dicho ejercicio, el cual va más allá del alcance del trabajo, consiste en utilizar como variables atmosféricas predictoras aquellas simuladas por un MCGA en modo pronóstico con la antecedencia requerida. Evidentemente los resultados dependerán del MCGA utilizado y de la antelación de la predicción. Dado que los pronósticos son siempre imperfectos, es esperable que los márgenes de ganancia respecto a la utilización de la media histórica se vean reducidos.

El objetivo de este trabajo se concentró en determinar la variación estacional y por embalse de la predictibilidad de los caudales mensuales y de las variables predictoras más importantes evaluando, a su vez, varios modelos de regresión. Por tanto, los resultados pueden ser utilizados como base para abordar una diversidad de aplicaciones muy relacionadas. La más directa es el diseño de un sistema de predicción mensual de caudales (para lo cual se requiere de un MCGA operativo), no obstante los resultados son fácilmente extensibles a otras escalas temporales. En particular, para escala estacional es esperable que los resultados de predictibilidad y habilidad predictiva sean superiores a los análogos encontrados para escala mensual (debido a una mayor relación señal/ruido). Por otro lado, a pesar de que para escalas menores a la mensual la relación señal/ruido es más débil, los resultados encontrados sobre la estacionalidad de la predictibilidad pueden orientar el diseño de sistemas de predicción adecuados a dichas escalas particulares. Si bien los modelos utilizados en este trabajo son determinísticos, la descripción de la predictibilidad que de ellos se obtiene es, también, valiosa para el diseño de sistemas de predicción probabilísticos como los que se utilizan en la actualidad en los modelos de apoyo a la gestión del sistema eléctrico.

Anexo A: Análisis de componentes principales (CPs)

Dado un conjunto de r variables correlacionadas X_1, ..., X_r, el análisis de CPs busca sustituir el conjunto de las r variables originales por un conjunto de t variables ξ_1, ..., ξ_t $t \leq r$, mutuamente no correlacionadas y ordenadas según una cierta medida de información. Las nuevas t variables se obtienen como combinaciones lineales de las variables originales, de modo de minimizar la pérdida de información debido a la sustitución.

Sea X el vector formado por las r variables X_1, ..., X_r: $X = (X_1, \ldots, X_r)^t$. Diremos que X tiene media μ_X y matriz de covarianzas V_{XX}. Entonces el análisis de CPs busca nuevas variables que se escriban como combinación lineal de las originales: $\xi_j = b_j^t X = b_{j1} X_1 + \ldots + b_{jr} X_r \qquad j = 1, \ldots, t$

En el análisis de CPs la información se interpreta como la "variación total" de las variables originales:

$$\sum_{j=1}^{r} var(X_j) = tr(V_{XX})$$

V_{XX} es una matriz simétrica, por lo que se puede escribir como $V_{XX} = U \Lambda U^t, \; U^t U = Id_r$ donde Λ es una matriz diagonal cuyos elementos son los valores propios $\{\lambda_j\}$ de V_{XX} y las columnas de U son los vectores propios de V_{XX} . Por lo tanto, la variación total es $tr(V_{XX}) = tr(\Lambda)$.

El j-ésimo coeficiente $b_j = (b_{j1}, \ldots, b_{jr})^t$ es elegido de modo que:

1. Las primeras t combinaciones lineales ξj, j=1,...,t estén ordenadas en orden decreciente según sus varianzas $\{var(\xi j)\}$.
2. ξ_j no esté correlacionada con ξ_k, k<j.

Las t combinaciones lineales ξ_1, ..., ξ_t se conocen como las primeras t componentes principales.

En la derivación original de Hotelling (1933) los coeficientes b_j eran obtenidos de manera secuencial de forma de que las varianzas de las variables derivadas $(var(\xi_j) = b_j^t \, V_{XX} \, b_j)$ estén ordenadas de forma descendiente, restrictas a la condición de normalización $b_j^t \; b_j = 1$ (j=1,...,t) y que las nuevas variables no estén correlaciondas con las derivadas previamente $(cov(\xi_k, \xi_j) = b_k^t \, V_{XX} \, b_j = 0 \; k<j)$.

La primer componente principal, ξ_1, se obtiene eligiendo los coeficientes $b_1 = (b_{11}, \ldots, b_{1r})^t$ de modo que la varianza $var(\xi_1)$ sea máxima, sujeto a la condición de normalidad $b_1^t \; b_1 = 1$.

A los efectos de resolver el problema de extremos condicionados se considera la función

$$f(b_1) = b_1^t V_{XX} b_1 - \lambda_1 (1 - b_1^t b_1)$$

donde λ_1 es un multiplicador de Lagrange. Se deriva respecto de b_1 y se iguala a cero, para la obtención del máximo.

$$\frac{\partial f(b_1)}{\partial b_1} = 2(V_{XX} - \lambda_1 Id_r) b_1 = 0$$

Si b_1 no es el vector nulo, entonces λ_1 debe satisfacer la ecuación: $det(V_{XX} - \lambda_1 Id_r) = 0$

Por lo tanto, λ_1 debe ser valor propio de V_{XX} y b_1 un vector propio asociado a λ_1. Es más, dado que $var(\xi_1) = b_1^t V_{XX} b_1 = b_1^t \lambda_1 b_1 = \lambda_1$, λ_1 debe ser el mayor valor propio de V_{XX}.

La segunda componente principal, ξ_2, se obtiene eligiendo los coeficientes $b_2 = (b_{21}, \ldots, b_{2r})^t$ de modo que $var(\xi_2)$ sea máxima entre todas las proyecciones lineales de X que no estén correlacionadas con ξ_1. Por lo tanto se debe maximizar $var(\xi_2) = b_2^t V_{XX} b_2$, sujeto a la condición de normalidad $b_2^t b_2 = 1$ y la condición de ortogonalidad $b_1^t b_2 = 0$.
Se genera la función

$$f(b_2) = b_2^t V_{XX} b_2 + \lambda_2 (1 - b_2^t b_2) + \mu b_1^t b_2$$

donde λ_2 y μ son multiplicadores de Lagrange. Derivando respecto de b_2 e igualando a cero:

$$\frac{\partial f(b_2)}{\partial b_2} = 2(V_{XX} - \lambda_2 Id_r) b_2 + \mu b_1 = 0$$

Pre-multiplicando la ecuación anterior por b_1^t y utilizando las condiciones se tiene que:

$$2 b_1^t V_{XX} b_2 + \mu = 0$$

Pre-multiplicando la ecuación $\frac{\partial f(b_1)}{\partial b_1} = 2(V_{XX} - \lambda_1 Id_r) b_1 = 0$ por b_2^t tenemos que:

$$b_2^t V_{XX} b_1 = 0$$

Luego, $\mu = 0$ y λ_2 debe cumplir: $(V_{XX} - \lambda_2 Id_r) b_2 = 0$ por lo que λ_2 debe ser el segundo mayor valor propio de V_{XX} y b_2 el vector propio asociado a λ_2.

Sucesivamente se obtienen los restantes vectores de coeficientes b_3, ..., b_r. Para determinar las componentes principales.

Una alternativa usual dentro del análisis de CPs, es la ejecución del mismo a partir de la matriz de correlaciones (en lugar de la matriz de covarianzas V_{XX}). La matriz de correlaciones se define como la matriz de covarianzas que se obtiene al considerar las variables previamente estandarizadas. Aunque las variables originales y las estandarizadas se relacionan fácilmente, los valores y vectores propios de las matrices de covarianzas asociados a estas no siguen una relación tan directa. Es por ello que las dos alternativas para el análisis de CPs no conducen, necesariamente, a la misma información y, por lo tanto, para cada aplicación es necesario seleccionar entre uno u otro enfoque. Si el objetivo del análisis es la identificación de los principales modos de variabilidad de una serie de datos, generalmente, la mejor alternativa es la utilización de la matriz de covarianzas. Sin embargo, si los datos a utilizar provienen de variables de distinta naturaleza, o si bien representen una misma magnitud están medidos con distintas unidades o se presenten órdenes de magnitud muy diferentes, siempre será preferible la utilización de la matriz de correlaciones. Un discusión al respecto de esta temática puede encontrarse en Jolliffe (2002).

Anexo B: Índice de abreviaciones

A: Grupo de variables predictoras atmosféricas.
A^{UCLA}: Grupo de variables atmosféricas potencialmente predictibles por el MCGA-UCLA.
AS: Región comprendida entre 50ºS-10ºS y 280ºE-330ºE.
$Árbol_{AOQ}$-óptimo: Modelo de árbol de regresión con el criterio nmin óptimo que utiliza las variables predictoras de los grupos A, O y Q.
CART: Árboles de clasificación y regresión.
$Clusters_{AOQ}$: Modelo de regresión vía algoritmo de clustering, que utiliza las variables predictoras de los grupos A, O y Q.
CPs: Componentes Principales.
CV: Cross-Validation.
ENOS: El Niño Oscilación Sur.
EOF: Función empírica ortogonal.
Error cv: Error cross-validation leave-one-out.
ESS: Suma cuadrática de distancias Euclídeas al centroide.
hgt: Altura geopotencial en 200hPa.
lm_{AO}-óptimo: Modelo de regresión lineal múltiple acoplado con eliminación hacia atrás óptimo partiendo del conjunto inicial de variables predictoras de los grupos A y O.
lm_{AOQ}: Modelo de regresión lineal múltiple utilizando todos los predictores de los grupos A, O y Q.
lm_{AOQ}-óptimo: Modelo de regresión lineal múltiple acoplado con eliminación hacia atrás óptimo partiendo del conjunto inicial de variables predictoras de los grupos A, O y Q.
$lm_{A^{UCLA}O}$-óptimo: Modelo de regresión lineal múltiple acoplado con eliminación hacia atrás óptimo partiendo del conjunto inicial de variables predictoras de los grupos A^{UCLA} y O.
$lm_{A^{UCLA}OQ}$-óptimo: Modelo de regresión lineal múltiple acoplado con eliminación hacia atrás óptimo partiendo del conjunto inicial de variables predictoras de los grupos A^{UCLA} ,O y Q.
lm_{O}: Modelo de regresión lineal múltiple utilizando todos los predictores del grupo O.
lm_{OQ}: Modelo de regresión lineal múltiple utilizando todos los predictores de los grupos O y Q.
lm_{OQ}-óptimo: Modelo de regresión lineal múltiple acoplado con eliminación hacia atrás óptimo partiendo del conjunto inicial de variables predictoras de los grupos O y Q.
MCGA: Modelo de circulación general de la atmósfera.
MCGAO: Modelo de circulación general acoplado atmósfera-océano.
MCGO: Modelo de circulación general del océano.
N3.4: Índice Niño 3.4 óptimo.
O: Grupo de variables predictoras oceánicas.
Pc1hgt: Primer componente principal de la altura geopotencial en 200hPa, en la región AS.
Pc2hgt: Segunda componente principal de la altura geopotencial en 200hPa, en la región AS.
Pc3hgt: Tercera componente principal de la altura geopotencial en 200hPa, en la región AS.
Pc1u: Primer componente principal del viento zonal en 200hPa, en la región AS.
Pc2u: Segunda componente principal del viento zonal en 200hPa, en la región AS.
Pc3u: Tercera componente principal del viento zonal en 200hPa, en la región AS.
Pc1v: Primer componente principal del viento meridional en 200hPa, en la región AS.
Pc2v: Segunda componente principal del viento meridional en 200hPa, en la región AS.
Pc3v: Tercera componente principal del viento meridional en 200hPa, en la región AS.
PLSR: Regresión por mínimos cuadrados parciales.
$PLSR_{AOQ}$-óptimo: Modelo de regresión por mínimos cuadrados parciales partiendo del conjunto inicial de variables predictoras de los grupos A, O y Q, con cantidad de variables óptima.
Q: Grupo de variables predictoras de caudales precedentes.
Q1: Caudal con 1 mes de antecedencia.
Q2: Caudal con 2 meses de antecedencia.

Red_{AOQ}: Modelo de redes neuronales (Red_{AOQ}) que utiliza las variables seleccionadas por selección hacia atrás partiendo del conjunto inicial de variables predictoras de los grupos A, O y Q.
RMSE: Raíz cuadrada del error medio cuadrático.
RSS: Residual Sum of Squares.
SESA: Sudeste de América del Sur.
TSM: Temperatura de superficie de mar.
u: Viento zonal en 200hPa.
UCLA: Universidad de California, Los Ángeles.
v: Viento meridional en 200hPa.
ymedio: Modelo que predice el valor de caudal promedio en los casos pertenecientes al conjunto de aprendizaje.

BIBLIOGRAFÍA

Aceituno P. 1988. On the Functioning of the Southern Oscillation in the South American Sector. Part I: Surface Climate. Monthly Weather Review, 116, 505-524.

Aceituno P. 1989. On the Functioning of the Southern Oscillation in the South American Sector. Part II. Upper-Air Circulation. Journal of Climate, 2, 341-355.

Alexander R.C. and Mobley R.L. 1976. Monthly Average Sea–Surface Temperatures and Ice–Pack Limits on a 1° Global Grid. Monthly Weather Review. 104, 143-148.

Bock H.H. 2007. Clustering Methods: a History of k-Means Algorithms. Selected Contributions in Data Analysis and Classification, 2, 161-172. Springer.

Breiman L., Friedman J.H., Olshen R.A. and Stone C.J. 1984. Classification and Regression Trees. Wadsworth.

Cazes-Boezio G.; Robertson A. and Mechoso R. 2003. Seasonal Dependence of ENSO Teleconnections over South America and Relationships with Precipitation in Uruguay. Journal of Climate, 16, 1159-1176.

Dorman, J.L. and Sellers P.J. 1989. A global climatology of albedo, roughness length and stomatal resistance for atmospheric general circulation models as represented by the Simple Biosphere model (SiB). Journal of Applied Meteorology, 28, 833-855.

Farrara, J. D.; Mechoso C. R. and Robertson A. W. 2000. Ensembles of AGCM two-tier predictions and simulations of the circulation anomalies during winter 1997-1998. Monthly Weather Review, 128, 3589-3604.

Goddard L.; Mason S.J.; Zebiak S.E.; Ropelewski C.F.; Basher R. and Cane M.A. 2001. Current approaches to seasonal-to-interannual climate predictions. International Journal of Climatology, 21, 1111-1152.

Grimm A.M.; Pal J.S. and Giorgi F. 2007. Connection between Spring Conditions and Peak Summer Monsoon Rainfall in South America: Role of Soil Moisture, Surface Temperature and Topography in eastern Brazil. Journal of Climate, 20, 5929-5945.

Hotelling H. 1933. Analysis of a complex of statistical variables into principal components. Journal of Educational Psychology, 24, 417-441 y 498-520.

Izenman A.J. 2008. Modern Multivariate Statistical Techniques. Springer Texts in Statistics.

Jolliffe, I.T. 2002. Principal Component Analysis, second edition, Springer.

Kalnay et al. 1996. The NCEP/NCAR 40-year reanalysis project, Bulletin of the American Meteorological Society, 77, 437-470.

Kaufman L. and Rousseeuw P.J. 1990. Introduction in Finding Groups in Data: An Introduction to Cluster Analysis. John Wiley & Sons, Inc.

Konor C.S.; Cazes-Boezio G.; Mechoso C.R. and Arakawa A. 2009. Parameterization of PBL Processes in an Atmospheric General Circulation Model: Description and Preliminary Assessment. Monthly Weather Review 137: 1061-1082.

Kumar A. and Hoerling M.P. 2003. The Nature and Causes for the Delayed Atmospheric Response to El Niño. Journal of Climate, 16, 1391-1403.

Landman W.A.; Mason S.J.; Tyson P.D. and Tennat W.J. 2001. Statistical downscaling of GCM Simulations to streamflow. Journal of Hydrology, 252, 221-236.

Lima C.H. and Lall U. 2010. Climate informed monthly streamflow forecasts for the Brazilian hydropower network using a periodic ridge regression model. Journal of Hydrology, 380, 438-449.

Lorenz E.N. 1963. Deterministic Nonperiodic Flow. Journal of the Atmospheric Sciences, 20, 130-141.

Lorenz E.N. 1969. Atmospheric Predictability as Revealed by Naturally Occurring Analogues. Journal of the Atmoepheric Sciences, 26, 636-646.

Lorenz E.N. 1982. Atmospheric predictability experiments with a large numerical model. Tellus, 34, 505-513.

Lumely T. 2009. Package Leaps. http://cran.r-project.org/

MacQueen J.B. 1967. Some methods for classification and analysis of multivariate observations. Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, 1, 281-297. University of California Press.

Maechler M.; Rousseeuw P.; Struyf A. and Hubert, M. 2005. Cluster Analysis Basics and Extensions. http://cran.r-project.org/

Mechoso C.R. and Pérez-Irribarren G. 1992. Streamflow in Southeastern South America and the Southern Oscillation. Journal of Climate, 5, 1535-1539.

Miller A.J.; Cayan D.R.; Barnett T.P.; Graham E.N. and Oberhuber J.M. 1994. The 1976-1977 climate shif of the Pacific Ocean. Journal of Oceanography, 7, 21-26.

Molinaro A.M.; Simon R. and Pfeiffer R.M. 2005. Prediction error estimation: a comparison of resampling methods. Bioinformatics, 21, 3301-3307.

Pisciottano G.; Díaz A.; Cazes G.. and Mechoso R. 1994. El Niño - Souther Oscillation impact on rainfall in Uruguay. Journal of Climate, 7, 1286-1302.

Reynolds R.W.; Rayner N.A.; Smith T.M.; Stokes D.C.; Wang W. 2002. An Improved In Situ and Satellite SST Analysis for Climate. Journal of Climate 15: 1609-1625.

Ropelewski C.F. and Halpert M.S. 1987. Global and Regional Scale Precipitation Patterns Associated with the El Niño/Southern Oscillation. Monthly Weather Review,115, 1606-1626.

Ropelewski C.F. and Halpert M.S. 1989. Precipitation Patterns Associated with the High Index

Phase of the Southern Oscillation. Journal of Climate, 2, 268-284.

Soukup T.L; Aziz O.A.; Tootle G.A.; Piechota T.C. and Wulff S.S. 2009. Long lead-time streamflow forecasting of the North Platte River incorporating oceanic-atmosheric climate variability. Journal of Hydrology, 368, 131-142.

Stone M. 1974. Cross-Validatory Choice and Assessment of Statistical Predictions. Journal of the Royal Statistical Society. Series B (Methodological), 36, 111-147.

Su H.; Neelin J.D. and Meyerson J.E. 2005. Mechanisms for Lagged Atmospheric Response to ENSO Forcing. Journal of Climate, 18, 4195-4215.

Suarez M.J.; Arakawa A. and Randall D.A. 1983: The parameterization of the planetary boundary layer in the UCLA general circulation model: Formulation and results. Monthly Weather Review, 111, 2224-2243.

Trenberth K.E. 1990. Recent observed interdecadal climate changes in the Northern Hemisphere. Bulletin of the American Meteorological Society, 71, 988-993.

Vinod H.D. 1969. Integer programming and the Theory of Grouping. Journal of the American Statistical Association, 64, 506-519.

Wang W. 2006. Stochasticity, nonlinearity and forecasting of streamflow processes. IOS Press.

Werbos P.J. 1974. Beyond regression: new tools for prediction and analysis in the behavioral sciences. PhD Dissertation, Harvard University.

Westra S. and Sharma A. 2009. Probabilistic Estimation of Multivariate Streamflow Using Independent Component Analysis and Climate Information. Journal of Hydrometeorology, 10, 1479-1492.

Wilks D.S. 2006. Statistical Methods in the Atmospheric Sciences. Second Edition. International Geophysics Series.

Wold S.; Martens H. and Wold H. 1983. The multivariate calibration problem in chemistry solved by the PLS method. Lecture Notes in Mathematics, 973, 286-293. DOI: 10.1007/BFb0062108

Wood A.W.; Maurer E.P.; Kumar A. and Lettenmaier D.P. 2002. Long-range experimental hydrologic forecasting for the eastern United States. Journal of Geophysical Research (Atmospheres), 107, D20, pp. ACL 6-1. DOI:10.1029/2001JD000659.