FACULTAD DE CIENCIAS EXACTAS Y NATURALES

UNIVERSIDAD DE BUENOS AIRES

Desarrollo de un framework para la predicción probabilística de precipitación en mediana escala en regiones de Argentina

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

Tesista: Esp. Alfredo Luis Rolla

Directora: Dra. Marcela Hebe González

Fecha: 12-03-2024

Agradecimientos

A mis padres y hermano que siempre me apoyaron para desarrollarme académica y profesionalmente.

A mi esposa María Inés , mis hijas Mariana y Victoria por ser mi motor en la vida.

A Marcela González por permitirme trabajar a su lado en conjunto con su grupo de investigación, que han sido pilares fundamentales para la evolución de esta tesis.

A Mario Núñez por contagiarme su entusiasmo por la disciplina de la meteorología.

A mis compañeros de la Maestría, que hicieron que fuera una experiencia amena, compartiendo conmigo su tiempo y amistad.

Resumen

Se describe en este trabajo la implementación de un sistema para realizar un pronóstico climático (trimestral - bimensual - mensual) en cualquier región de la que se tengan observaciones provenientes de estaciones meteorológicas con registros largos, utilizando reanálisis y con el conocimiento de los forzantes climáticos que actúan en esas regiones. La implementación de pronósticos de precipitación a escalas mensuales es importante para los tomadores de decisiones en diferentes áreas como la agricultura, la industria y la generación de energía. Dentro de las metodologías de pronóstico de mediana escala se encuentran las técnicas estadísticas que brindan la posibilidad de aprender de situaciones pasadas para pronosticar futuras. Las técnicas de minería de datos son actualmente una herramienta poderosa para abordar estos problemas. En este caso se consideran las redes neuronales, la regresión de soporte vectorial y los modelos aditivos generalizados, además de la metodología de regresión lineal múltiple más utilizada en el pasado, para obtener modelos de predicción de precipitaciones. Los resultados indican que las técnicas de minería de datos mejoran los pronósticos derivados de otras metodologías, aunque la eficiencia de las diferentes metodologías depende en gran medida del mes y la región. Además, se tiene la posibilidad de generar ensambles de varios modelos y derivar pronósticos probabilísticos que es una alternativa muy recomendable para realizar pronósticos que este sistema además permite.

Abstract

This paper “Development of a framework for the probabilistic prediction of medium-scale precipitation in regions of Argentina” describes the implementation of a system to perform a climate forecast (quarterly - bimonthly - monthly) in any region in which there are observations from meteorological stations with long records using reanalysis and knowledge of the climatic forcing that act in those regions. The implementation of precipitation forecasts at monthly scales is important for decision makers in different areas such as agriculture, industry, and power generation. Among the medium scale forecast methodologies are statistical techniques that provide the possibility of learning from past situations to forecast future ones. Data mining techniques are currently a powerful tool to address these problems. In this case, neural networks, support vector regression and generalized additive models are considered, in addition to the multiple linear regression methodology most used in the past, to obtain rainfall prediction models. The results indicate that data mining techniques improve forecasts derived from other methodologies, although the efficiency of the different methodologies is highly dependent on month and region. Additionally, the possibility of generating ensembles of several models and deriving probabilistic forecasts is a highly recommended alternative to carry out the forecast that this system allows.

Índice de contenidos

Índice de Figuras.

Capitulo II

Capítulo III

Capítulo IV

Índice de Tablas.

Capítulo II

Capítulo III

Capítulo IV

Capítulo V

CAPÍTULO I : INTRODUCCIÓN

1.1. El desafío de pronosticar el clima.

Pronosticar el clima es un ejercicio cada vez más intensivo en datos. Los modelos de predicción meteorológica numérica (NWP) se están volviendo más complejos, con resoluciones más altas, y hay un número cada vez mayor de modelos diferentes en funcionamiento. Si bien la capacidad de predicción de los modelos continúa mejorando, el número y la complejidad de estos modelos plantean un nuevo desafío para el meteorólogo operacional: ¿cómo se debe combinar la información de todos los modelos disponibles, cada uno con sus propios sesgos y limitaciones, para proporcionar a las partes interesadas pronósticos probabilísticos bien calibrados para usar en la toma de decisiones?

Desde su modesto comienzo en la década de 1950, los pronósticos meteorológicos numéricos han evolucionado en escala y complejidad para convertirse en una parte integral de innumerables procesos de toma de decisiones en todo el mundo. Aunque no están tan difundidos como los pronósticos meteorológicos, muchos centros de investigación operativa producen pronósticos climáticos estacionales (Doblas-Reyes et al., 2013). Estos pronósticos estiman los valores medios mensuales y estacionales de las variables climáticas (por ejemplo, temperatura y precipitación) con 1 a 12 meses de anticipación para poder prever posibles eventos extremos como por ejemplo, sequías o inundaciones relacionadas con patrones de flujo atmosférico a gran escala, como El Niño Oscilación del Sur (ENSO; Rasmusson y Wallace, 1983, Kousky et al., 1984) o la Oscilación del Atlántico Norte (NAO; Hurrell (1996)). Con los aumentos en el poder de cómputo combinados con los avances en la ciencia del clima y en la calidad y cantidad de los datos de observación, existe un interés creciente en la viabilidad de los pronósticos climáticos decenales (comúnmente denominados pronósticos climáticos), es decir, pronósticos de mayor escala temporal, que brindan estimaciones de un año a varias décadas de antemano (Smith et al., 2007). La precisión de estos pronósticos ha progresado sustancialmente en los últimos 20 años (por ejemplo, consulte Doblas-Reyes et al. (2013) y Meehl et al. (2014), para más información sobre pronósticos estacionales y decenales).

1.2. Escalas de tiempo.

La escala de tiempo estacional se ocupa de los pronósticos para tiempos futuros que oscilan entre más de dos semanas y un poco más de 1 año. Los pronósticos meteorológicos y sub-estacionales tratan escalas de tiempo más cortas, mientras que las predicciones climáticas para tiempos futuros más allá del primer año del pronóstico y hasta 30 años están cubiertos por la predicción decenal. Para escalas de tiempo más largas, las proyecciones climáticas apuntan a estimar las posibles evoluciones del clima durante varias décadas en función de los escenarios de forzamiento pasados y futuros. Los límites entre estas escalas climáticas son imprecisos. Existe una superposición sustancial entre ellos porque, entre otras razones, muchos de los procesos involucrados son comunes. La viabilidad de la predicción estacional se basa en gran medida en la existencia de variaciones lentas y predecibles de ciertas variables como por ejemplo, la humedad del suelo, la capa de nieve, el hielo marino y la temperatura de la superficie del océano y en el conocimiento de la interacción de ellas con la atmósfera . A escalas temporales estacionales, el almacenamiento de calor y humedad por parte del océano y la tierra y la presencia o ausencia de nieve y hielo marino se convierten en factores importantes. El Niño-Oscilación del Sur (ENOS) es el principal proceso que contribuye a la calidad del pronóstico en escalas de tiempo estacionales. Una anomalía en la temperatura de la superficie del mar cálido (SSTA) en el océano Pacífico tropical conduce a un aumento del flujo de calor desde el océano a la atmósfera que, si es lo suficientemente grande, puede alterar la capa límite atmosférica y, en última instancia, cambiar la estructura de la lluvia y la liberación de calor latente en la troposfera. La liberación de calor latente adicional afectará la circulación atmosférica y produce un tren de ondas de Rossby que de desplaza desde los trópicos hacia extra trópicos y hacia el este (Mo, 2000) , de forma tal que provocará anomalías climáticas en regiones remotas del mundo, fenómeno conocido como “teleconexiones”. Sin embargo, el ENOS no es el único forzante que sirve para pronosticar el clima en escalas estacionales. Por ejemplo, el Dipolo del océano Indico (Saji et. al, 1999), el Monzón Sudamericano (Kousky, 1988.), la Oscilación Antártica (Thompson and Wallace,2000) entre otros,

también pueden influenciar el clima en dichas escalas. El acoplamiento de la atmósfera y los océanos o el suelo, generan mecanismos de feedback que son muy complejos y es por ello que los pronósticos tienen bastante incertidumbre. Los métodos de aprendizaje automático son entonces, una herramienta muy útil para lidiar con este tema.

1.3. Verificación de los pronósticos.

Independientemente de la escala de tiempo en consideración, es esencial evaluar la calidad del pronóstico a través de la verificación del mismo. La verificación de pronósticos se lleva a cabo comparando pronósticos de eventos pasados (también conocidos como pronósticos retrospectivos) con sus observaciones correspondientes (o productos de observación, denominados referencias en lo sucesivo). La verificación implica cuantificar la precisión (la correspondencia entre los pronósticos y las referencias) y la asociación (la fuerza de la relación entre los pronósticos y las referencias) (Potts, 2003) del sistema de pronóstico. Idealmente, se deben considerar varias métricas diferentes, ya que una sola medida no puede caracterizar completamente la calidad del pronóstico (Bennett et al., 2013).

1.4. Tipos de modelos de pronósticos estacionales.

Existen tres tipos de métodos que se pueden utilizar para realizar pronósticos estacionales: empírico (o estadístico), dinámico e híbrido (estadístico-dinámico). Los métodos empíricos utilizan relaciones estadísticas entre los predictores y el predictando, generalmente involucrando algún tipo de modelo estadístico de regresión utilizando datos observacionales. Los predictores se identifican mediante el análisis físico de los mecanismos que controlan la predicción.

1.5. Ventajas y desventajas.

Algunas ventajas de los métodos empíricos son que requieren bajos recursos informáticos y son fáciles de implementar operativamente, según la Guía de la Organización Meteorológica Mundial sobre prácticas operativas para la predicción estacional objetiva (WMO, 2020). Están diseñados para ser consistentes con las observaciones (ya están corregidos por sesgo, al menos con respecto a los valores medios), y ofrecen predicciones en términos tanto de valores determinísticos como de probabilidades. Dadas las continuas mejoras en los sistemas de predicción estacional basados en modelos dinámicos, se espera que en el futuro, estos sistemas sean más efectivos que los métodos de predicción empíricos para realizar pronósticos estacionales precisos (WMO, 2020).

Una desventaja de usar métodos dinámicos es que los modelos dinámicos tienen sesgos en la representación de la media y varianza de variables como precipitación, nubosidad nieve, y con respecto a la reproducción correcta de los patrones espaciales (ubicación, extensión, forma) de las variables atmosféricas y oceánicas, desde la Temperatura de la Superficie del Mar (TSM) hasta los campos de circulación y las teleconexiones asociadas. La necesidad de evaluar estos sesgos a través de pronósticos retrospectivos (hindcasts) de temporadas pasadas y compararlos con los resultados observados también aumenta sustancialmente el costo y la complejidad de los sistemas dinámicos.

1.6. Herramientas para la predicción del clima.

Alguna de las herramienta de predicción estacional de tipo estadístico muy usada en los centros de predicción estacional es el Climate Predictability Tool (CPT; Mason y otros, 2017) desarrollada y puesta a disposición por el International Research Institute for Climate and Society (IRI) y la técnica de Análisis de Correlación Canónica (ACC; Barnston, 1994) para realizar predicciones. El CPT es un software que facilita el proceso de calibración de pronósticos con ACC. El software fue desarrollado en el lenguaje de

programación Fortran 90 y adaptado para uso práctico en los sistemas operativos Windows y Linux. Es capaz de producir predicciones climáticas usando correcciones estadísticas de los resultados de modelos climáticos globales, o producir predicciones usando campos de temperatura de la superficie del mar o predictores similares (altura geopotencial, componentes zonales y meridionales del viento a diferentes niveles atmosféricos, etc.). El ACC es una de las técnicas disponibles en el software para generar predicciones empíricas o calibrar las salidas de modelos dinámicos.

El objetivo de este trabajo es desarrollar un framework para realizar un pronóstico climático, determinístico y probabilístico estacional (mensual o trimestral) en cualquier región de la que se tengan observaciones provenientes de estaciones meteorológicas con registros largos utilizando además información global de forzantes provenientes de reanálisis de NCEP/NCAR (National Centers for Environmental Prediction/ National Center for Atmospheric Research). La implementación de pronósticos de precipitación a escalas estacionales es importante para los tomadores de decisiones en diferentes áreas como la agricultura, la industria y la generación de energía. Dentro de las metodologías de pronóstico de mediana escala se encuentran las técnicas estadísticas que brindan la posibilidad de aprender de situaciones pasadas para pronosticar futuras. Las técnicas de minería de datos son actualmente una herramienta poderosa para abordar estos problemas. En este caso se consideraran las redes neuronales, la regresión de soporte vectorial y los modelos aditivos generalizados, además de la metodología de regresión lineal múltiple más utilizada en el pasado, para obtener modelos de predicción determinística y probabilística de precipitaciones. Se evaluarán utilizando algunos índices que miden la eficiencia de los mismos. Se podrán comparar los resultados de los modelos estadísticos y dinámicos usando , por ejemplo RMSSS (Root mean square error skill scores) (Murphy et. al 1988) , y diagramas de confiabilidad (Reliability Diagrams) (Hartman H, et al. 2002).

Se presenta un nuevo enfoque estadístico para el pronóstico probabilístico de precipitación estacional para aplicar en diferentes regiones de Argentina.

CAPÍTULO II: FRAMEWORK

2.1. Explicación general del framework

El objetivo es de desarrollar un framework para realizar un pronóstico climático determinístico y probabilístico en escalas estacionales en cualquier región de la que se tengan observaciones provenientes de estaciones meteorológicas con registros largos utilizando además información global proveniente de reanálisis de NCEP/NCAR.

Todos los módulos del framework fueron escritos en lenguaje R (R Core Team (2021)) y RStudio (Posit team (2023)).

Los paquetes específicos usados en cada módulo serán definidos en cada una de las descripciones de los mismos.

Los resultados/salidas en general se escribieron en formato Excel para facilidad de visualización de los resultados.

A continuación (Figura 2.1) mostramos el diagrama general de funcionamiento del framework.

Como se observa hay tres grupos principales: **el entrenamiento** donde se preparan los predictores con varios subprocesos , **verificación y testing** donde se va avanzando en el tiempo año por año generando planillas de resultados por cada año mes. Cabe aclarar que, como este proceso se realiza disponiendo de las observaciones, es posible verificar los pronósticos. Finalmente el **pronóstico**, es el proceso que se realiza para un período posterior, usando los predictores que han sido definidos, evaluados unos días antes de fin del período actual.

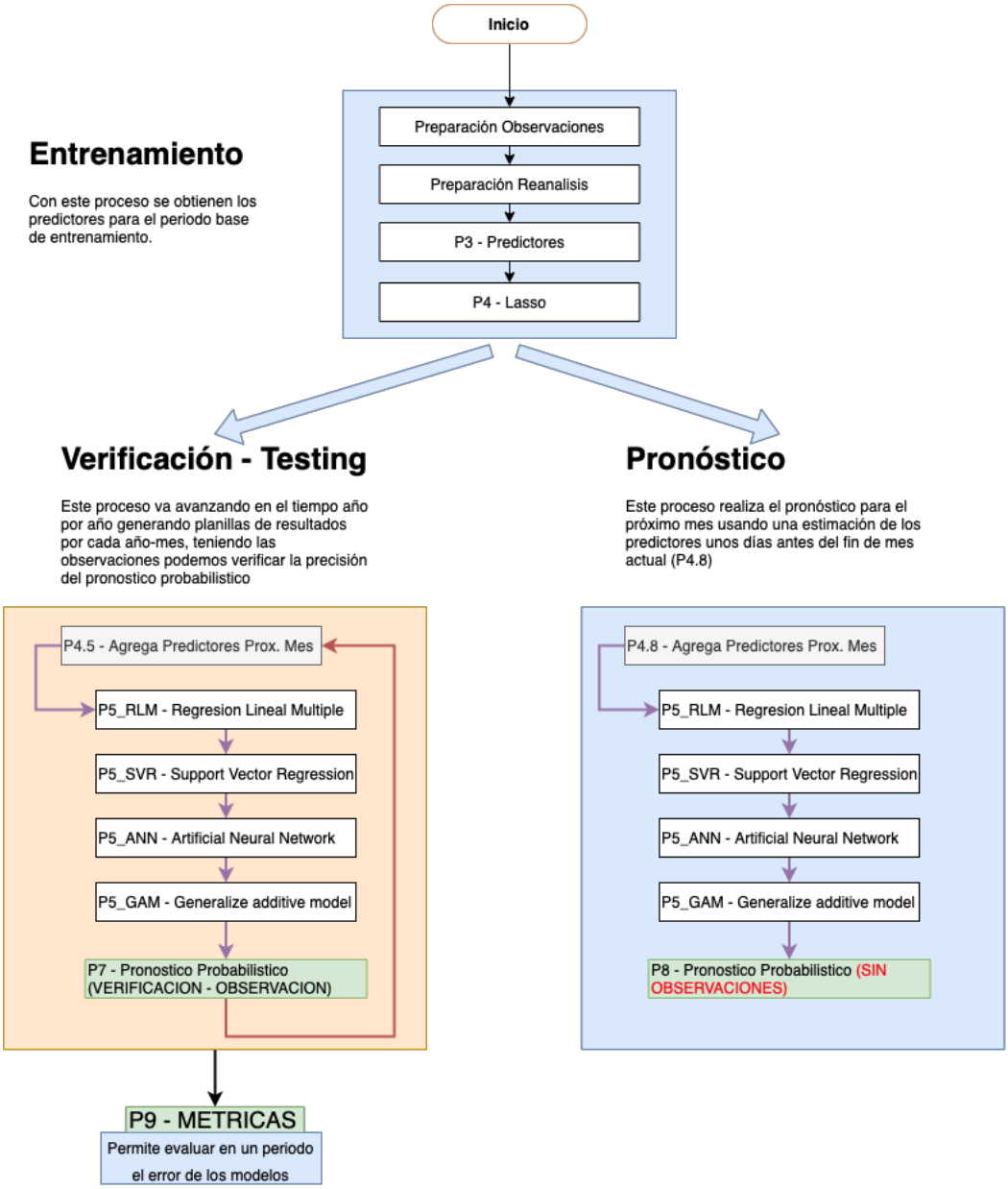


Figura 2.1. Diagrama General.

2.2. Evolución general del pronóstico en el tiempo

La Figura 2.2 muestra el esquema de evolución del pronóstico en el tiempo.

Primero se entrena entre 1981-201x y después avanza año por año usando los predictores de las regiones seleccionadas en el periodo 1981-201x hacia 2016, 2017, …, 201x... paso a paso.

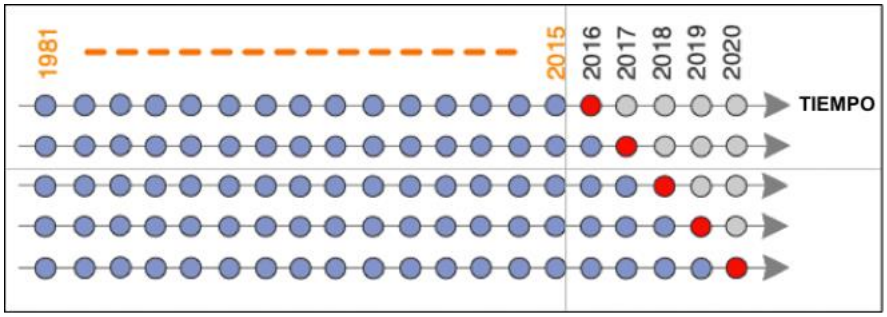


Figura 2.2. Evolución del pronóstico en el tiempo.

2.3. Entrenamiento de los modelos de pronóstico.

Figura 2.3. Diagrama de flujo de obtención de predictores.

Como se observa en el diagrama de la Figura 2.3 es necesario incorporar Observaciones y Reanálisis, para obtener los predictores a los modelos y su posterior reducción para no poner información redundante en los modelos generados.

Con este proceso se obtienen los predictores para el periodo base de entrenamiento de acuerdo a las observaciones disponibles de las regiones determinadas por alguna técnica de agrupamiento y los reanálisis espaciales para determinar el pronóstico.

A continuación se describen las 4 etapas principales para la generación de los predictores para la construcción de los modelos.

2.3.1. Preparación de Observaciones.

Las observaciones serán datos de precipitación mensuales provenientes de estaciones meteorológicas de la región de estudio que pueden ser previamente agrupadas en clusters con algún método estadístico de agrupamiento.

Los aspectos técnicos del modo de almacenamiento de las observaciones (Tabla 2.1) serán:

- Deben estar dentro del directorio "clusters" en el directorio de trabajo.
- Deben estar contenidas en un archivo Excel.
- Tiene que haber un archivo Excel por cada mes.
- Los archivos Excel deben llamarse "series.medias.pre.{mes} donde {mes} es 01, 02, ... hasta el 12.
- Como se observa en la figura habrá un columna para el año y una columna por cada clúster.

- Ejemplo de Excel con las observaciones de cada clúster.
- Cada columna se debe llamar "cluster1, cluster2, ... clusterN".

Tabla 2.1. Ejemplo de Excel con las observaciones de cada clúster.

	A	B	C	D	E
1		YEAR	cluster1	cluster2	cluster3
2	1	1981	164,2	24,1	137,1
3	2	1982	121,6	31,9	42,0
4	3	1983	80,4	67,4	121,8
5	4	1984	159,3	52,3	219,1
6	5	1985	138,4	15,6	181,8
7	6	1986	110,1	39,7	51,3
8	7	1987	107,9	53,6	108,3
9	8	1988	92,6	28,8	113,4
10	9	1989	50,3	16,1	52,3
11	10	1990	122,3	70,2	104,2
12	11	1991	59,2	8,5	67,5

2.3.2. Preparación de Reanálisis

Antes de continuar, describiremos a grandes rasgos que es un análisis y un reanálisis

2.3.2.1. Que es un Análisis climático?

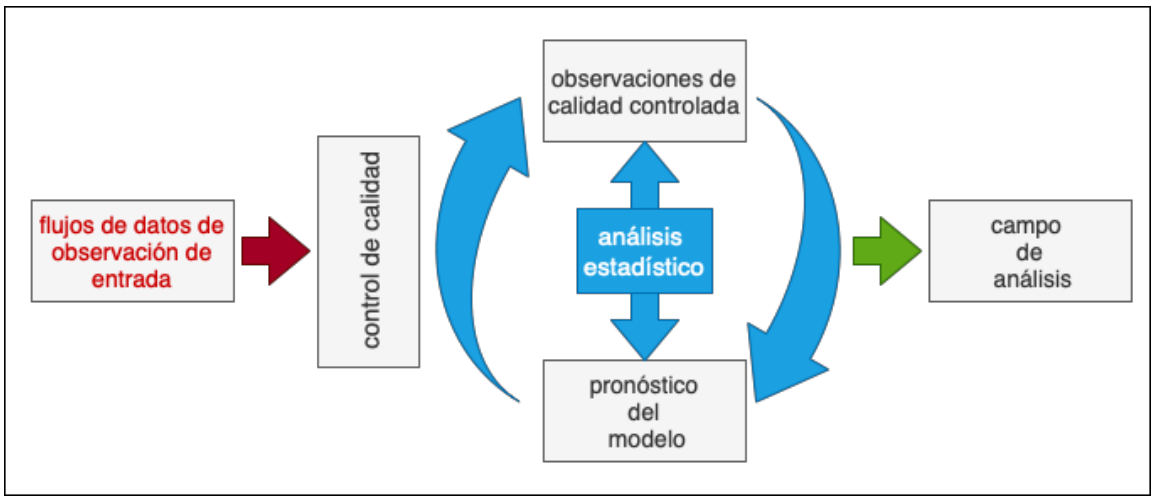


Figura 2.4. Esquema de generación de campos de análisis.

Los pronósticos de corto alcance y las observaciones se combinan estadísticamente para crear un "análisis" completo o un estado tridimensional reticulado de la atmósfera cada 6 horas/diario/mensual.

Este “ciclo de análisis” se desarrolló para la predicción numérica del tiempo, donde el modelo y el análisis estadístico se cambian con frecuencia (por ejemplo, mayor resolución) para mejorar los pronósticos (Figura 2.4).

2.3.2.2. Que es un RE-Análisis climático?

Todas las observaciones pasadas se reprocesan manteniendo el modelo y el análisis estadístico “congelados” (sin cambios).

Esto crea un análisis largo mucho más auto consistente. Pero, cambios en el sistema de observación aún pueden generar algunos “saltos” en el análisis climatológico.

2.3.2.3. Los reanálisis de NCEP-NCAR :

Son grillados globales de 144 x73 puntos, esto implica que tiene una resolución espacial de 250 x250 Km.

Son campos medios en nuestro caso de resolución temporal de un mes.

Los usamos a partir de Enero del año 1979 , porque en esa fecha empezaron a ingresar los datos de satélites que mejoran sustancialmente la calidad y precisión de los reanálisis.

Los nombre de las variables de los reanálisis considerados son:

- **hgt200**: geopotencial en 200 hPa.
- **hgt500**: geopotencial en 500 hPa.
- **hgt1000**: geopotencial en 1000 hPa.
- **sst**: temperatura superficie del mar.
- **pw**: agua precipitable.
- **lsm**: mascara de tierra y agua (es una variable estática).
- **u850:** componente de viento zonal en capas bajas.
- **v850:** componente de viento meridional en capas bajas.

La variable geopotencial se utiliza porque es representativa del movimiento del aire en diferentes niveles de la atmósfera. La SST es importante porque el océano interactúa con la atmósfera, generando mecanismos de teleconexión capaces de modificar la circulación atmosférica en zonas lejanas al lugar de origen donde se producen las anomalías de SST. La variable pw y u850, v850 son importantes porque influyen en la advección de aire húmedo en capas bajas, que guarda especial relación con la precipitación.

2.3.3. Preparación de predictores (P3-Predictores)

El objetivo de este módulo/programa es de obtener regiones cuya correlación este desfasada en 1 mes. La variable meteorológica observada en un mes determinado se correlaciona con las variables derivadas de los reanálisis en el mes previo y se identifican áreas de dichos campos de correlación donde se produzca un nivel umbral de correlación significativa. En este caso se utilizó un umbral del 95% de confianza, pero este valor puede modificarse en caso que sea necesario, en el programa de cálculo. Esas áreas permiten definir las series de predictores (promedio de la variable en dicha área) para que sirvan de entrada al programa de generación de modelos, previa selección de

aquellos predictores con sentido físico, que aporten información a los modelos generados.

Como se observa en el diagrama de funcionamiento las observaciones están en el directorio CLUSTERS que es un lugar fijo dentro del programa P3 y los reanálisis están en el directorio NNR que también es un lugar fijo y se generarán los mapas de correlación y se crearán archivos en formato Excel con las series de los predictores para el mes considerado.

Por ejemplo, si queremos pronosticar febrero, correlacionamos los reanálisis de enero con las observaciones de febrero y guardamos los predictores en las planillas Excel de febrero, con estos predictores se generarán los modelos para febrero.

Los reanálisis se descargan usando el script bash 'predictores.sh' descripto en el GIT HUB usando los comando Linux **wget**: un recuperador por red no interactivo, **cdo:** climate data operators y ncks: netcdf kitchen sink.

Link a los reanálisis: https://psl.noaa.gov/data/gridded/data.ncep.reanalysis.html

Las especificaciones de los reanálisis NCEP-NCAR 1 se muestran en la Figura 2.5:

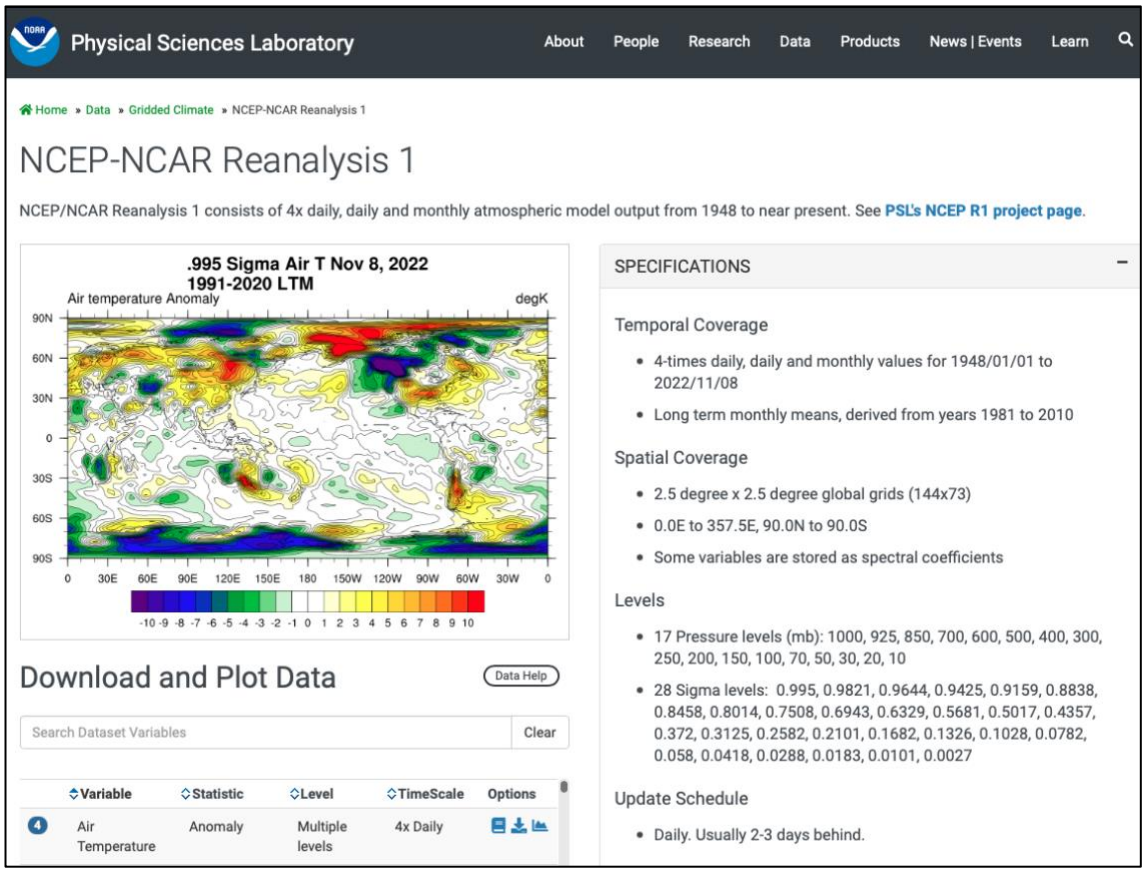


Figura 2.5. NCEP-NCAR Reanálisis 1.

2.3.3.1. Diagrama de funcionamiento.

El diagrama siguiente (Figura 2.6) muestra el esquema funcional del módulo P3.

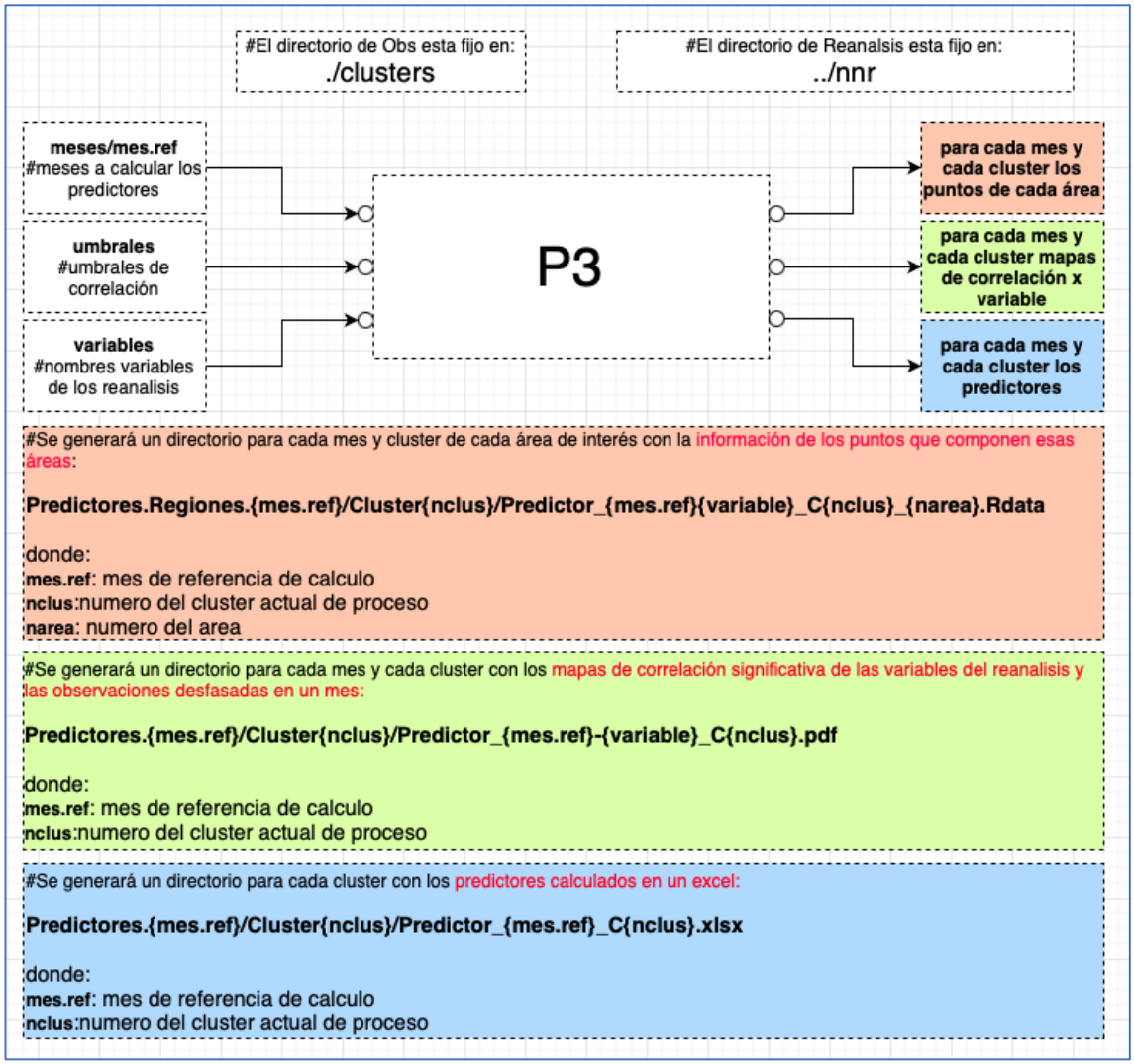


Figura 2.6. Esquema funcional del módulo P3.

Existen dos directorios de entrada uno con la información de datos observacionales de entrada (‘clusters’) y otro con los reanálisis de NCEP a utilizar (‘nnr’). Los reanálisis son archivos de tipo mensuales.

La Tabla 2.2 muestra las librerías R que se utilizan en este módulo:

Tabla 2.2. Librerías R utilizadas en módulo P3.

Nombre	Descripción
ncdf4	Pierce D (2023). Interfaz para datos de formato Unidata netCDF (versión 4 o anterior).
fields	Douglas Nychka et. al (2021). Herramientas para datos espaciales.
sp	Roger S. Bivand et. al (2013). Análisis de datos espaciales aplicado con R .

maptools	Bivand R, Lewin-Koh N (2023). Herramientas para el manejo de objetos espaciales.
openxlsx	Schauberger P, Walker A (2023). Leer, escribir y editar archivos xlsx.
raster	Hijmans R (2023). Análisis y modelado de datos geográficos.

Los parámetros de entrada al módulo P3 son los meses para calcular los predictores, los umbrales de correlación significativa y los nombres de las variables de los reanálisis.

Las salidas del módulo P3 son: los puntos que definen los convexos de las regiones de probabilidad significativa, para cada mes y cada clúster, los mapas de correlación por variable y además para cada mes y cada clúster los predictores asociados.

El módulo P3 (naranja) generará un directorio para cada mes y clúster de cada área de interés con la información de los puntos que componen esas áreas llamado:

Predictores.Regiones.{mes.ref}/Cluster{nclus}/Predictor_{mes.ref}{variable}_C{nclus}_{narea}.Rdata

donde:

mes.ref: mes de referencia de cálculo

nclus: número del clúster actual de proceso

narea: número del área

El módulo P3 (verde) generará un directorio para cada mes y cada clúster con los mapas de correlación significativa (Figura 2.7) de las variables del reanálisis y las observaciones desfasadas en un mes o trimestre:

Predictores.{mes.ref}/Cluster{nclus}/Predictor_{mes.ref}-{variable}_C{nclus}.pdf

donde:

mes.ref: mes de referencia de cálculo

nclus: número del clúster actual de proceso

El módulo P3 (celeste) generará un directorio para cada clúster con los predictores calculados en un Excel (Tabla 2.3):

Predictores.{mes.ref}/Cluster{nclus}/Predictor_{mes.ref}_C{nclus}.xlsx

donde:

mes.ref: mes de referencia de cálculo

nclus: número del clúster actual de proceso

Tabla 2.3. Ejemplo de tabla con predictores.

	A	B	C	D	E	F	G	H
1	sst_C2_7	hgt500_C2_4	hgt1000_C2_6	hgt200_C2_3	u850_C2_17	u850_C2_21	u850_C2_27	u850_C2_28
2	22,5823633	5521,28251	91,52620977	11753,8009	4,37437439	-6,8187504	-8,9554568	8,80925864
3	22,5101591	5556,90568	107,1383062	11802,2742	4,01228873	-4,1781273	-8,7209098	9,86351635
4	24,3714331	5525,11631	98,73870905	11755,6769	3,53603935	-3,2262497	-11,00091	9,73185052
5	23,7263281	5479,02444	67,98749981	11700,9978	5,08416557	-1,8543758	-7,7800016	8,03925804
6	22,8931892	5498,14418	79,8645166	11727,6869	5,7847894	-5,0025043	-7,4263652	6,92129517
7	23,3932488	5543,05426	109,5588711	11779,2336	1,97458204	-5,5856276	-9,8827307	9,11425725
8	23,6652727	5578,76392	104,4661294	11854,6475	3,46604156	-2,3050003	-10,224548	8,87722072
9	23,7634832	5545,64906	104,1149197	11789,6084	4,02687263	-4,3162508	-8,8690907	9,64703454
10	23,5090016	5536,90959	101,6487905	11768,4828	2,59353987	-4,203126	-9,7863672	9,97351668
11	22,7164824	5590,06504	127,7814515	11867,361	1,05853939	-3,5512505	-10,241819	10,5744423
12	22,8066433	5537,89542	102,3729834	11782,4571	3,37937323	-6,0343761	-9,5645488	10,9337014

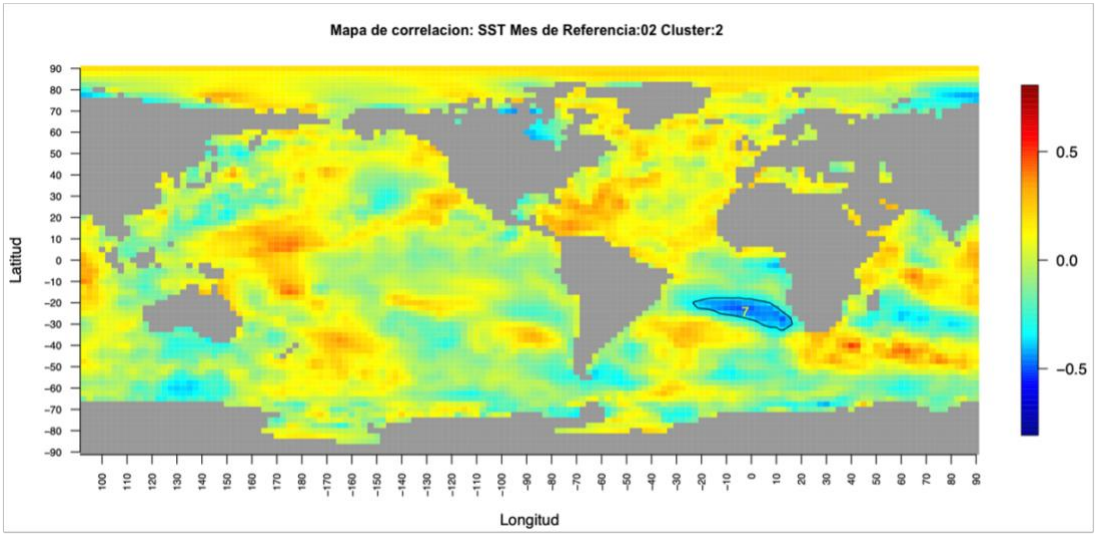


Figura 2.7. Ejemplo de mapa de correlación de precipitación clúster 2 vs. temp. de sup. del mar.

Se puede observar en la figura la región marcada como 7 (en azul) que tiene un nivel de significancia alto y es posible de ser usado como predictor en un modelo resultante.

2.3.4. Reducción de predictores (P4-LASSO).

El objetivo de este módulo/programa es de probar la independencia de los predictores generados en el módulo anterior P3-Predictores y por lo tanto reducir el número de los mismos. En el caso particular que todos los predictores sean independientes, todos ellos serán considerados como entrada a los modelos. Este paso es muy importante para evitar el problema de multicolinealidad que puede ser causa de inestabilidad en los modelos.

Previo a utilizar este módulo se realiza un proceso de selección manual de predictores de acuerdo a factibilidad meteorológica y física, es decir que todos los predictores que ingresan a un modelo deben guardar con la variable meteorológica una relación que sea explicable conceptualmente a través de un proceso físico.

En Estadística y Aprendizaje Automático, la metodología LASSO (least absolute shrinkage and selection operator, por sus siglas en inglés), es un método de análisis de regresión que realiza selección de variables y regularización para mejorar la exactitud e interpretabilidad del modelo estadístico producido por este. Fue introducido por Robert Tibshirani en 1996 basado en el trabajo de Leo Breiman sobre el Garrote No-Negativo.

Lasso fue formulado originalmente para el método de mínimos cuadrados y este caso simple revela una cantidad substancial acerca del comportamiento del estimador, incluyendo su relación con ridge regresión y selección de subconjuntos (de variables) y la conexión entre los coeficientes estimados con Lasso y el llamado 'soft thresholding'. También revela que (al igual que la Regresión Lineal estándar) los coeficientes estimados no necesariamente son únicos si las variables independientes son colineales.

Robert Tibshirani introdujo LASSO para mejorar la exactitud de la predicciones e interpretabilidad de los modelo estadísticos de regresión al alterar el proceso de construcción del modelo seleccionando solamente un subconjunto de (y no todas) las variables provistas para usar en el modelo final. Está basado en el Garrote No-Negativo de Breiman, que tiene propósitos similares, pero funciona de manera un poco diferente. Antes de LASSO, el método más usado para decidir qué variables incluir en un modelo era stepwise selection, que sólo mejora la exactitud de las predicciones en ciertos casos, como cuando sólo unas pocas variables tienen una relación fuerte con la variable

independiente. Sin embargo, en otros casos, puede agravar los errores de predicción. Además, en ese momento, ridge regression era la técnica más popular para mejorar la exactitud de las predicciones. Ridge Regression es un método para estimar los coeficientes de modelos de regresión múltiple en escenarios donde las variables independientes están altamente correlacionadas. Ridge regression mejora los errores de predicción al reducir en tamaño los coeficientes de regresión que sean demasiado grandes para reducir el 'sobreajuste'(overfitting) , pero no realiza selección de variables y por tanto no produce un modelo más interpretable.

Siendo x_i las variables predictoras, y_i los predictandos y β_i la estimación del LASSO, el problema de optimización se puede expresar con la siguiente expresión:

$$\sum_{i=1}^{n}(y_i - \beta_0 - \sum_{j=1}^{p}\beta_j x_{ij})^2 + \lambda \sum_{j=1}^{p}|\beta_j| = sum\ of\ square\ residuals + \lambda \sum_{j=1}^{p}|\beta_j|$$

2.3.4.1. Diagrama de funcionamiento

El diagrama siguiente (Figura 2.8) muestra el esquema funcional del módulo P4.

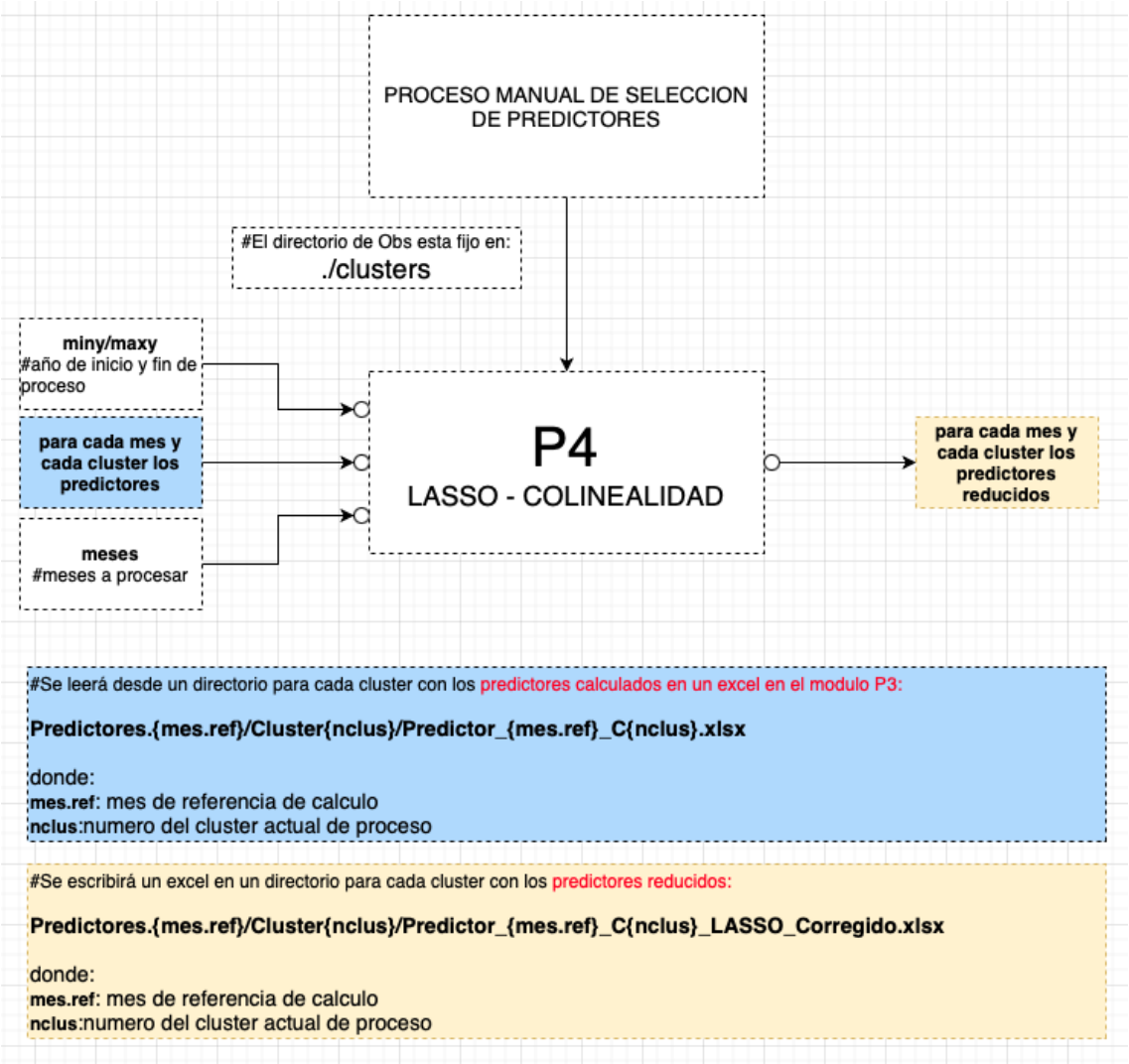


Figura 2.8. Esquema funcional del módulo P4.

[cálculo se escribe: cálculo cuando es sustantivo -esdrújula-, si lo pones sin acento es verbo -1ra persona singular-]

La Tabla 2.4 muestra las librerías R que se utilizan en este módulo:

Tabla 2.4. Librerías R utilizadas en el módulo P4.

Nombre	Descripción
glmnet	Friedman J, Tibshirani R, Hastie T (2010). Rutas de Regularización para Modelos Lineales Generalizados vía Descenso de Coordenadas.
plotmo	Milborrow S (2022). Trazar gráficos de residuos, respuesta y dependencia parcial de un modelo
openxlsx	Schauberger P, Walker A (2023). Leer, escribir y editar archivos xlsx.

Existe un directorio de entrada con la información de datos observacionales de entrada (‘clusters’).

Los parámetros de entrada al módulo P4 son los años de inicio y fin de proceso para calcular los predictores, para cada mes y cada clúster los predictores y por último los meses a procesar.

Las salidas del módulo P4 son para cada mes y cada clúster, predictores reducidos.

El módulo P4 (azul) leerá desde un directorio para cada clúster los predictores calculados en un Excel provenientes del módulo P3:

Predictores.{mes.ref}/Cluster{nclus}/Predictor_{mes.ref}_C{nclus}.xlsx

donde:

mes.ref: mes de referencia de cálculo

nclus: número del clúster actual de proceso

El módulo P4 (amarillo) escribirá un Excel en un directorio para cada clúster con los predictores reducidos:

Predictores.{mes.ref}/Cluster{nclus}/Predictor_{mes.ref}_C{nclus}_LASSO_Corregido.xlsx

donde:

mes.ref: mes de referencia de cálculo

nclus: número del clúster actual de proceso

2.3.5. Cálculo de predictores para el año a pronosticar (hindcasts) (P4.5).

El objetivo de este módulo/programa es, para cada mes y cada clúster, leer los predictores del paso anterior P4-LASSO-COLINEALIDAD y, para cada mes y cada clúster, agregar los predictores para el año a pronosticar.

Nota: Cada vez que se usa este módulo se sobrescribe el resultado.

2.3.5.1. Diagrama de funcionamiento.

El diagrama siguiente (Figura 2.9) muestra el esquema funcional del módulo P4.5.

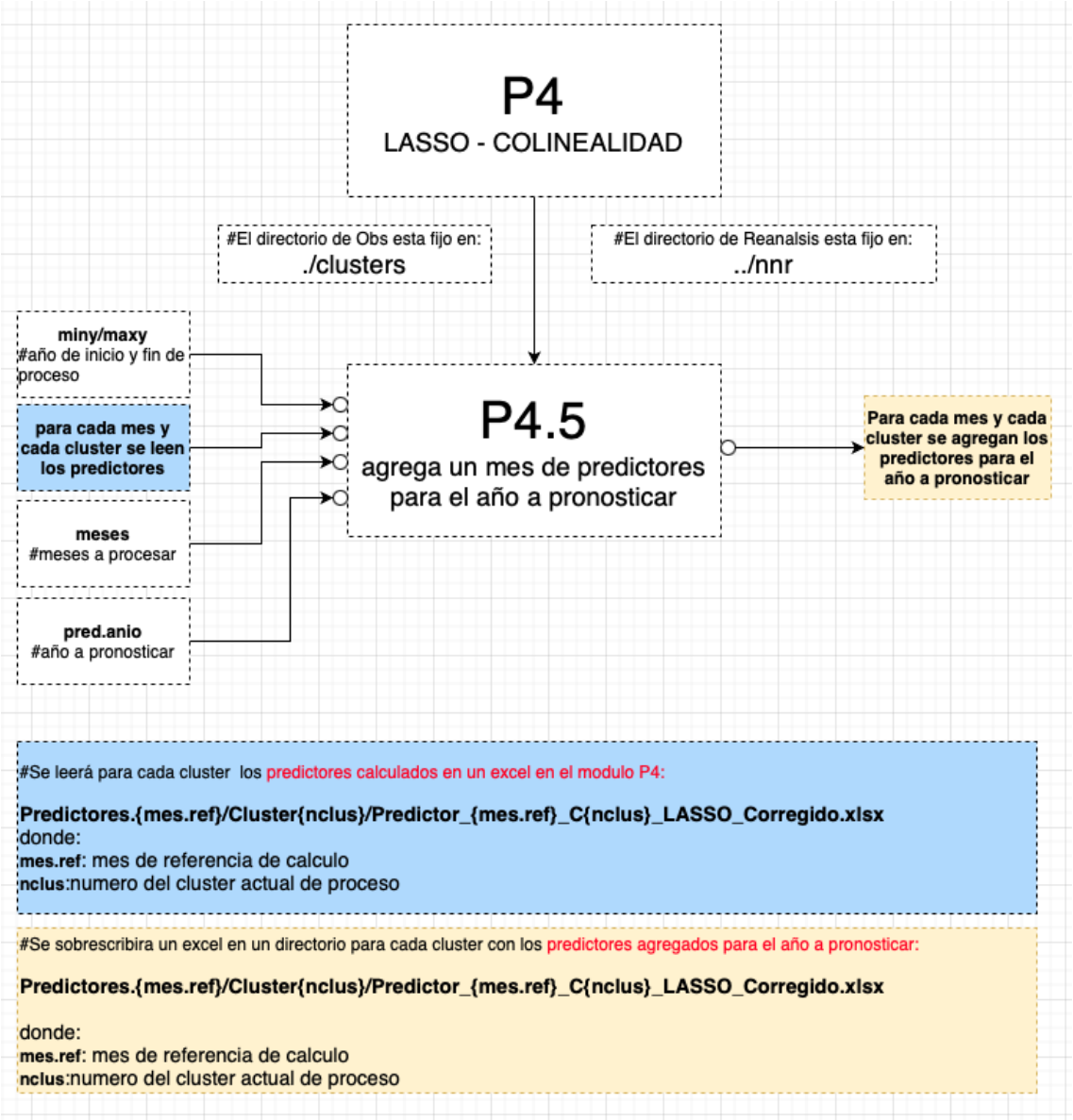


Figura 2.9. Esquema funcional del módulo P4.5.

La Tabla 2.5 muestra las librerías R que se utilizan en este módulo:

Tabla 2.5.Librerías R utilizadas en módulo P4.5.

Nombre	Descripción
ncdf4	Pierce D (2023). Interfaz para datos de formato Unidata netCDF (versión 4 o anterior).
fields	Douglas Nychka et. al (2021). Herramientas para datos espaciales.
sp	Roger S. Bivand et. al (2013). Análisis de datos espaciales aplicado con R .
maptools	Bivand R, Lewin-Koh N (2023). Herramientas para el manejo de objetos espaciales.

openxlsx	Schauberger P, Walker A (2023). Leer, escribir y editar archivos xlsx.
raster	Hijmans R (2023). Análisis y modelado de datos geográficos.

Existen dos directorios de entrada uno con la información de datos observacionales de entrada (‘clusters’) y otro con los reanálisis de NCEP a utilizar (‘nnr’).

Los parámetros de entrada al módulo P-4.5 son los años de inicio y fin de proceso para calcular los predictores, para cada mes y cada clúster los predictores y por último los meses a procesar y por último el año a pronosticar.

Las salidas del módulo P-4.5 son para cada mes y cada clúster, los predictores reducidos se agregan a los predictores para el año a pronosticar.

El módulo P-4.5 (azul) leerá para cada clúster los predictores calculados en un Excel en el módulo P4:

Predictores.{mes.ref}/Cluster{nclus}/Predictor_{mes.ref}_C{nclus}_LASSO_Corregido.xlsx

donde:

mes.ref: mes de referencia de cálculo

nclus: número del clúster actual de proceso

El módulo P-4.5 (amarillo) sobrescribirá un Excel en un directorio para cada clúster con los predictores agregados para el año a pronosticar:

Predictores.{mes.ref}/Cluster{nclus}/Predictor_{mes.ref}_C{nclus}_LASSO_Corregido.xlsx

donde:

mes.ref: mes de referencia de cálculo

nclus: número del clúster actual de proceso

2.3.6. Cálculo de predictores para el próximo periodo a pronosticar (P4.8)

El objetivo de este módulo/programa es, para cada mes y cada clúster, leer los predictores del paso anterior P4-LASSO-COLINEALIDAD y, para cada mes y cada clúster, los predictores para el año a pronosticar.

La diferencia con el P-4.5 es que este programa se puede utilizar antes de finalizar el mes o estación y calcular con los días disponibles la media mensual del último mes/estación.
Para lograr esto, es necesario descargar archivos diarios y transformarlos en mensuales antes de la finalización del mes/estación utilizando el utilitario escrito en bash 'predictProno.sh' al que se le pasa el parámetro del año de referencia.
(se pasa como parámetro, el año de referencia)

2.3.6.1. Diagrama de funcionamiento.

El diagrama siguiente (Figura 2.10) muestra el esquema funcional del módulo P4.8.

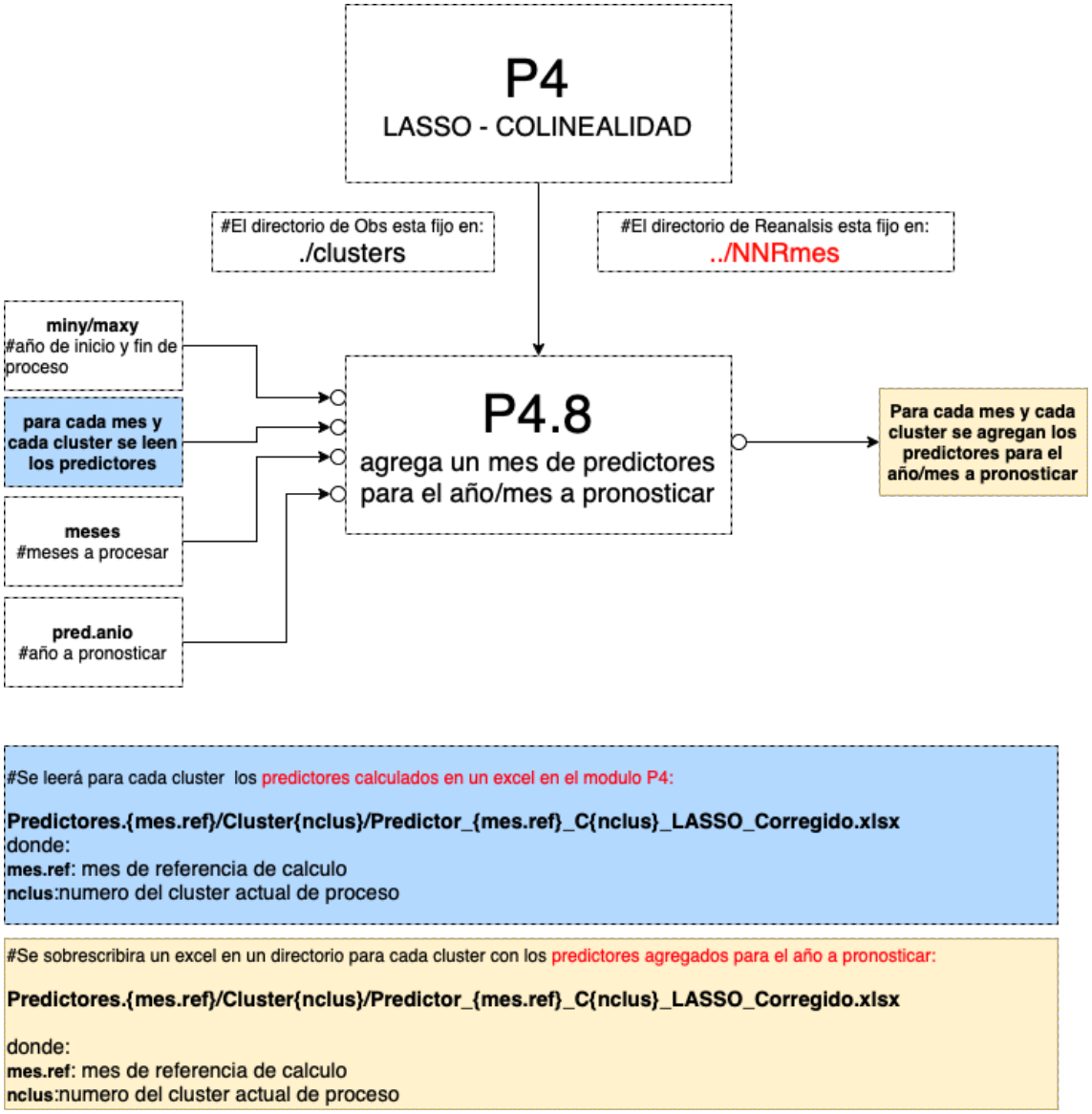


Figura 2.10. Esquema funcional del módulo P4.8.

La Tabla 2.6 muestra las librerías R que se utilizan en este módulo:

Tabla 2.6. Librerías R utilizadas en módulo P4.8.

Nombre	Descripción
ncdf4	Pierce D (2023). Interfaz para datos de formato Unidata netCDF (versión 4 o anterior).
fields	Douglas Nychka et. al (2021). Herramientas para datos espaciales.
sp	Roger S. Bivand et. al (2013). Análisis de datos espaciales aplicado con R .
maptools	Bivand R, Lewin-Koh N (2023). Herramientas para el manejo de objetos espaciales.
openxlsx	Schauberger P, Walker A (2023). Leer, escribir y editar archivos xlsx.

raster	Hijmans R (2023). Análisis y modelado de datos geográficos.

Existen dos directorios de entrada, uno con la información de datos observacionales de entrada (‘clusters’) y otro con los reanálisis de NCEP transformados a utilizar (‘NNRmes’).

Los parámetros de entrada al módulo P-4.8 son los años de inicio y fin de proceso para calcular los predictores, los meses a procesar y el año a pronosticar.

Las salidas del módulo P-4.8 son para cada mes y cada clúster, los predictores reducidos se agregan a los predictores para el año a pronosticar.

El módulo P-4.8 (azul) leerá para cada clúster los predictores calculados en un Excel en el módulo P4:

Predictores. {mes.ref}/Cluster{nclus}/Predictor_{mes.ref}_C{nclus}_LASSO_Corregido.xlsx

donde:

mes.ref: mes de referencia de cálculo

nclus: número del clúster actual de proceso

El módulo P-4.5 (amarillo) sobrescribirá un Excel en un directorio para cada clúster con los predictores agregados para el año a pronosticar:

Predictores.{mes.ref}/Cluster{nclus}/Predictor_{mes.ref}_C{nclus}_LASSO_Corregido.xlsx

donde:

mes.ref: mes de referencia de cálculo

nclus: número del clúster actual de proceso

2.3.7. Modelos de regresión lineal múltiple (P5-RLM).

El objetivo de este módulo/programa es, para cada mes y cada clúster, crear todos los modelos de regresión lineal múltiple que cumplen con los umbrales propuestos de R^2 ajustado (coeficiente de determinación ajustado).

En estadística, la regresión lineal o ajuste lineal es un modelo matemático usado para aproximar la relación de dependencia entre una variable dependiente Y, m variables independientes X_i con $m \in \mathbb{Z}^+$ y un término aleatorio ε. Este método es aplicable en muchas situaciones en las que se estudia la relación entre dos o más variables o para predecir un comportamiento, algunas incluso sin relación con la tecnología. En caso de que no se pueda aplicar un modelo de regresión a un estudio, se dice que no hay correlación lineal entre las variables estudiadas.

Este modelo puede ser expresado como:

$$Y = \beta_0 + \beta_1 \boldsymbol{X}_1 + \cdots + \beta_m \boldsymbol{X}_m + e$$

$$e \sim N(0; \sigma^2)$$

donde:

Y: es la variable dependiente o variable de respuesta.

$X_1, X_2, \ldots, X_m$: son las variables explicativas, independientes o regresoras.

$\beta_0, \beta_1, \beta_2, \ldots, \beta_m$: son los parámetros del modelo, miden la influencia que las variables explicativas tienen sobre el regresor.

El término β_0 es la intersección o término "constante", las β_i ($i \geq 1$) son los parámetros respectivos a cada variable independiente, y 'm' es el número de parámetros independientes a tener en cuenta en la regresión. La regresión lineal puede ser contrastada con la regresión no lineal.

El uso del coeficiente de determinación ajustado se justifica debido que a medida que añadimos variables a una regresión, el 'coeficiente de determinación sin ajustar' tiende a aumentar, incluso cuando la contribución marginal de cada una de las nuevas variables añadidas no tenga relevancia estadística.

Por lo tanto, al añadir variables al modelo, el coeficiente de determinación podría aumentar y podríamos pensar, de manera errónea, que el conjunto de variables elegido es capaz de explicar una mayor parte de la variación de la variable independiente. A este problema se le conoce comúnmente como "sobreestimación del modelo".

Para solucionar este problema, muchos investigadores (**Hyndman** et al, 2022) sugieren ajustar el coeficiente determinación mediante la siguiente fórmula:

$$R^2{}_{adj} = 1 - \frac{(1 - R^2)(n - 1)}{(n - k - 1)}$$

Donde:

n: número de datos.

k: número de predictores.

Teniendo en cuenta que $1\text{-}R^2$ es un número constante y que n es mayor que k, a medida que añadimos variables al modelo, el cociente entre paréntesis se hace más grande. Consecuentemente también el resultado de multiplicar este por $1\text{-}R^2$. Con lo cual vemos que la fórmula está construida para ajustar y penalizar la inclusión de coeficientes en el modelo.

2.3.7.1. Diagrama de funcionamiento.

El diagrama siguiente (Figura 2.11) muestra el esquema funcional del módulo P5-RLM.

Existen un directorio de entrada con la información de datos observacionales de entrada ('clusters').

Los parámetros de entrada al módulo P5-RLM son los años de inicio y fin de proceso, para cada mes y cada clúster, se leen los predictores y por último los meses a procesar.

Las salidas del módulo P5-RLM son para cada mes y cada clúster: se crean todos los modelos combinando los predictores de entrada y que superen del umbral de R^2 ajustado definido.

El módulo P5-RLM (azul en la Figura 2.11) leerá para cada clúster los predictores calculados en un Excel en el módulo P4.5/P4.8:

Predictores.{mes.ref}/Cluster{nclus}/Predictor_{mes.ref}_C{nclus}_LASSO_Corregido.xlsx

donde:

mes.ref: mes de referencia de cálculo

nclus: número del clúster actual de proceso

El módulo P5-RLM (azul en la Figura 2.11) escribirá un Excel en un directorio para cada clúster todos los modelos generados:

Modelos.{mes.ref}_{minx}_{may}/Cluster{nclus}/Modelos_{mes.ref}_C{nclus}_LASSO_RLM.xlsx

donde:

mes.ref: mes de referencia de cálculo

minx: año inicio

maxy: año fin

nclus: número del clúster actual de proceso

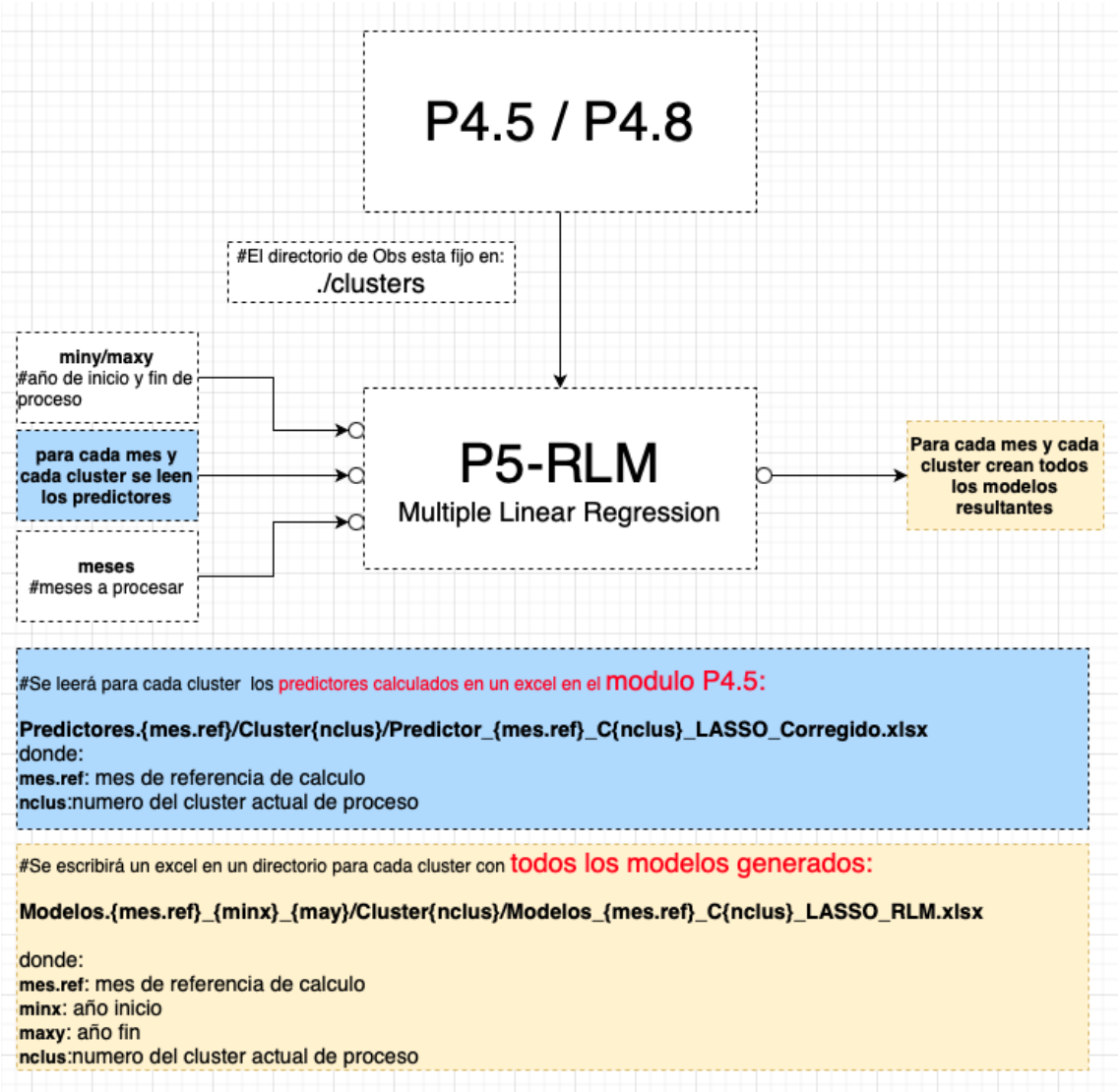


Figura 2.11. Esquema funcional del módulo P5-RLM.

La Tabla 2.7 muestra las librerías R que se utilizan en este módulo:

Tabla 2.7. Librerías R utilizadas en módulo P5-RLM.

Nombre	Descripción
openxlsx	Schauberger P, Walker A (2023). Leer, escribir y editar archivos xlsx.
fpp	Hyndman RJ (2013). Data for "Forecasting: principles and practice
forecast	Hyndman R, Athanasopoulos G, Bergmeir C, Caceres G, Chhay L, O'Hara-Wild M, Petropoulos F, Razbash S, Wang E, Yasmeen F (2023). Funciones de pronóstico para series de tiempo y modelos lineales

A continuación se muestra (Tabla 2.8) un ejemplo de salida de este módulo P5-RLM en formato Excel de los modelos de regresión lineal múltiple, se observa la columna de R^2 y el ajustado AdjR^2 para este tipo de modelo.

Tabla 2.8. Ejemplo de salida del módulo P5-RLM.

A	B	C	D	E	F
Formula	CV	AIC	AICc	BIC	AdjR2
y.dato ~ u850_C1_8	1458,02748	256,240013	257,014207	260,906058	0,19482856
y.dato ~ hgt200_C1_2	1415,066	256,293516	257,06771	260,95956	0,19359679
y.dato ~ u850_C1_14	1513,81995	258,611899	259,386092	263,277943	0,13837221
y.dato ~ u850_C1_8+hgt200_C1_2	1342,51675	252,907579	254,240912	259,128971	0,28700699
y.dato ~ u850_C1_8+u850_C1_14	1441,79547	254,773135	256,106468	260,994527	0,24797225
y.dato ~ hgt200_C1_2+u850_C1_14	1262,74703	252,964468	254,297801	259,18586	0,28584715
y.dato ~ u850_C1_8+hgt200_C1_2+u850_C1_14	1320,27177	251,514627	253,583593	259,291367	0,33200624

Donde:

CV: Cross-Validation.
Es un procedimiento en el que :
1) Se remueve la observación t del dataset, se ajusta el modelo usando los datos restantes. Luego se calcula el error $(e_t^* = y_t - \hat{y}_t)$ para la observación faltante
2) Se repite el paso 1) para t=1,….,T.
3) Se calcula el MSE desde $e_1^*, \dots, e_T^*$. Esto se llama CV

AIC: Criterio de información de Akaike.
El modelo con el mínimo valor de AIC es el mejor modelo para pronostico.

AICc: Criterio de información de Akaike corregido.
Debido a que AIC tiende a seleccionar demasiados predictores, esta versión de bias corregido fue desarrollado. AIC debe ser minimizado.

BIC: Criterio de información Bayesiano de Shwarz's.
Como con AIC, se intenta minimizar el BIC para obtener el mejor modelo. El modelo seleccionado por BIC penaliza el número de parámetros más fuertemente que AIC.

AdjR^2: Es una versión modificada de R^2 que agrega precisión y confiabilidad al considerar el impacto de las variables independientes adicionales que tienden a sesgar los resultados de las mediciones de R^2 .

2.3.8. Modelos de Regresión de Soporte Vectorial (Support Vector Regression) (P5-SVR).

El objetivo de este módulo/programa es, para cada mes y cada clúster, crear todos los modelos de support vector regression (SVR) que cumplen con los umbrales propuestos de R^2 ajustado.

Support Vector Regression es una variante del modelo de análisis Support Vector Machine (SVM) utilizado para clasificar. Sin embargo, con esta variante el modelo de soporte vectorial se utiliza como un esquema de regresión para predecir valores.

SVR es similar a la regresión lineal en que la ecuación de la línea es $y= wx + b$. Esta línea recta se conoce como hiperplano. A diferencia de otros modelos de regresión que intentan minimizar el error entre el valor real y el predicho, el SVR intenta ajustar la mejor línea dentro de un valor de umbral (distancia entre el hiperplano y la línea límite). Para una regresión no lineal, la función kernel transforma los datos a una dimensión mayor y realiza la separación lineal utilizando una función kernel de base radial.

En SVR, la regresión se realiza en una dimensión superior. La función de kernel es con la que es posible realizar esta transformación, asignar los puntos de un conjunto de datos de menor dimensión a otro de mayor, facilitando la búsqueda de un hiperplano en un espacio de mayor dimensión al mismo tiempo que reduce el costo de computación.

En los modelos de clasificación SVM (Figura 2.12) los hiperplanos son las líneas empleadas para separar los conjuntos de datos en clases. Aunque, en el caso de SVR, los hiperplanos son las líneas que ayudan a predecir el valor objetivo, el vector de soporte es el objeto que se usa para definir el hiperplano.

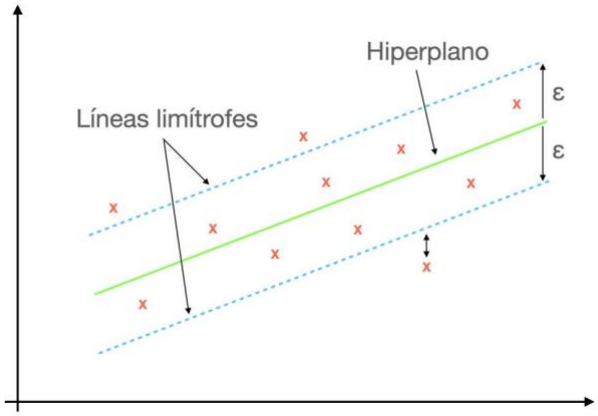


Figura 2.12. Descripción support vector regression.

2.3.8.1. Diagrama de funcionamiento.

El diagrama siguiente (Figura 2.13) muestra el esquema funcional del módulo P5-SVR.

Existen un directorio de entrada con la información de datos observacionales ('clusters').
Los parámetros de entrada al módulo P5-SVR son los años de inicio y fin de proceso. Para cada mes y cada clúster se leen los predictores y por último los meses a procesar.
Las salidas del módulo P5-SVR son, para cada mes y cada clúster, todos los modelos combinando los predictores de entrada y que superen del umbral de R^2 ajustado definido. Se calculan, además, los valores de RMS, Epsilon y cost.

El módulo P5-SVR (azul en la Figura 2.13) leerá para cada clúster los predictores calculados en un Excel en el módulo P4.5/P4.8:

Predictores.{mes.ref}/Cluster{nclus}/Predictor_{mes.ref}_C{nclus}_LASSO_Corregido.xlsx

donde:

mes.ref: mes de referencia de cálculo

nclus: número del clúster actual de proceso

El módulo P5-SVR (azul en la Figura 2.13) escribirá un Excel en un directorio para cada clúster conteniendo todos los modelos generados:

Modelos.{mes.ref}_{minx}_{may}/Cluster{nclus}/Modelos_{mes.ref}_C{nclus}_LASSO_RLM.xlsx

donde:

mes.ref: mes de referencia de cálculo

minx: año inicio

maxy: año fin

nclus: número del clúster actual de proceso

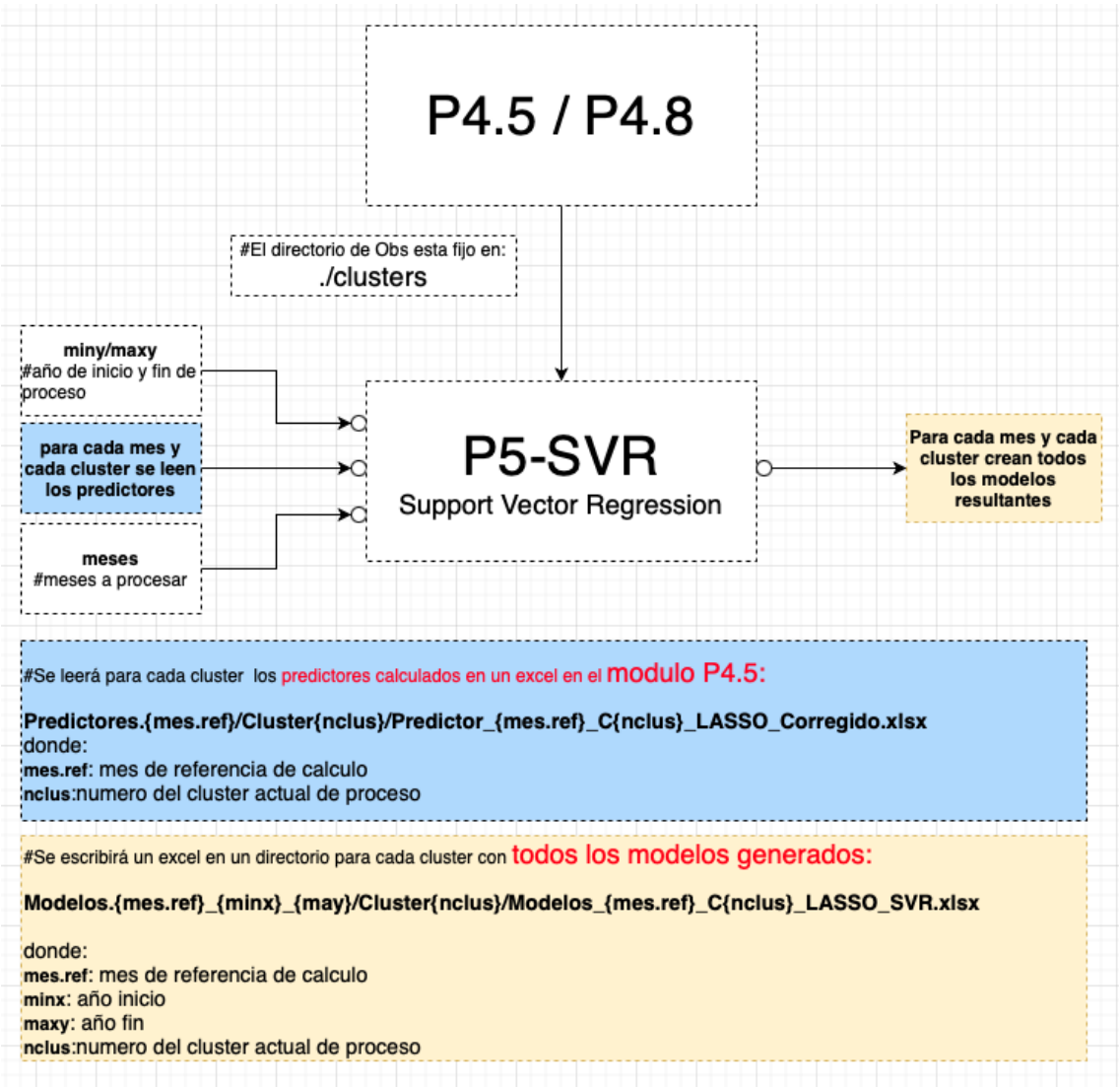


Figura 2.13. Esquema funcional del módulo P5-SVR.

La Tabla 2.9 muestra las librerías R que se utilizan en este módulo:

Tabla 2.9. Librerías R utilizadas en módulo P5-SVR.

Nombre	Descripción
openxlsx	Schauberger P, Walker A (2023). Leer, escribir y editar archivos xlsx.
fpp	Hyndman RJ (2013). Data for "Forecasting: principles and practice
forecast	Hyndman R, Athanasopoulos G, Bergmeir C, Caceres G, Chhay L, O'Hara-Wild M, Petropoulos F, Razbash S, Wang E, Yasmeen F (2023). Funciones de pronóstico para series de tiempo y modelos lineales
e1071	Meyer D, Dimitriadou E, Hornik K, Weingessel A, Leisch F (2023). Funciones Misceláneas del Departamento de Estadística, Grupo de Teoría de la Probabilidad

A continuación se muestra (Tabla 2.10) un ejemplo de salida del módulo P5-SVR en formato Excel de los modelos de support vector regresión, se observa la columna de R^2 y AdjR^2 para este tipo de modelo.

Tabla 2.10. Ejemplo de salida del módulo P5-SVR.

B	C	D	E	F	G
Formula	Rsquared	Radj	RMS	Epsilon	Cost
y.dato ~ u850_C1_8+hgt200_C1_2+u850_C1_14	0,624	0,588	24,7420409	0,72	4
y.dato ~ u850_C1_8+u850_C1_14	0,536	0,507	27,5132709	0,24	8
y.dato ~ u850_C1_8+hgt200_C1_2	0,512	0,482	28,2027087	0,76	8
y.dato ~ hgt200_C1_2+u850_C1_14	0,48	0,448	29,1024355	0,42	4
y.dato ~ u850_C1_8	0,346	0,326	32,651327	0,42	4
y.dato ~ hgt200_C1_2	0,289	0,267	34,0379747	0	4
y.dato ~ u850_C1_14	0,215	0,191	35,7808198	0	4

RMS: Error cuadrático medio.
Epsilon , Cost: Son hiperparámetros para usar en el espacio de búsqueda de minimización.

2.3.9. Modelos Aditivos generalizados (Generalize Additive Models) (P5-GAM).

El objetivo de este módulo/programa es, para cada mes y cada clúster, crear todos los modelos de Generalize Additive Models que cumplen con los umbrales propuestos de R^2 ajustado.

En estadística, un modelo aditivo generalizado (GAM) es un modelo no lineal generalizado en el que la variable de respuesta lineal depende linealmente de funciones suaves desconocidas de algunas variables predictoras, y el interés se centra en la inferencia sobre estas funciones suaves.

Los GAM fueron desarrollados originalmente por Trevor Hastie y Robert Tibshirani (Hastie, T. J.; Tibshirani, R. J. ,1990) para combinar propiedades de modelos lineales generalizados con modelos aditivos. Pueden interpretarse como la generalización discriminativa del modelo generativo de Bayes.

El modelo relaciona una variable de respuesta univariada Y, con algunas variables predictoras x_i. Se especifica una distribución de familia exponencial para Y (por ejemplo, distribuciones normales, binomiales o de Poisson) junto con una función de enlace g (por ejemplo, las funciones de identidad o logarítmica) que relacionan el valor esperado de Y con las variables predictoras a través de una estructura como:

$$g\big(E(Y)\big) = \beta_0 + f_1(x_1) + f_2(x_2) + \cdots + f_m(x_m)$$

Las funciones f_i pueden ser funciones con una forma paramétrica específica (por ejemplo, un polinomio o una spline de regresión no penalizada de una variable) o pueden especificarse de forma no paramétrica o semiparamétrica, simplemente como 'funciones suaves', para ser estimada por medios no paramétricos. Entonces, un GAM típico podría usar una función de suavizado de diagrama de dispersión, como una media ponderada localmente, para f_1 (x_1), y luego usar un modelo factorial para f_2 (x_2). Esta flexibilidad para permitir ajustes no paramétricos con suposiciones relajadas sobre la relación real entre la respuesta y el predictor, brinda el potencial para mejores ajustes

a los datos que los modelos puramente paramétricos, pero posiblemente con cierta pérdida de interpretabilidad.

2.3.9.1. Diagrama de funcionamiento.

El diagrama siguiente (Figura 2.14) muestra el esquema funcional del módulo P5-GAM.

Existen un directorio de entrada con la información de datos observacionales de entrada (‘clusters’) .

Los parámetros de entrada al módulo P5-GAM son los años de inicio y fin de proceso. Para cada mes y cada clúster se leen los predictores y por último los meses a procesar.

Las salidas del módulo P5-GAM son para cada mes y cada clúster, se crean todos los modelos combinando los predictores de entrada y que superen el umbral de R^2 ajustado definido.

El módulo P5-GAM (azul en la Figura 2.14) leerá para cada clúster los predictores calculados en un Excel en el módulo P4.5/P4.8:

Predictores.{mes.ref}/Cluster{nclus}/Predictor_{mes.ref}_C{nclus}_LASSO_Corregido.xlsx

donde:

mes.ref: mes de referencia de cálculo

nclus: número del clúster actual de proceso

El módulo P5-GAM (azul en la Figura 2.14) escribirá un Excel en un directorio para cada clúster conteniendo todos los modelos generados:

Modelos.{mes.ref}_{minx}_{may}/Cluster{nclus}/Modelos_{mes.ref}_C{nclus}_LASSO_RLM.xlsx

donde:

mes.ref: mes de referencia de cálculo

minx: año inicio

maxy: año fin

nclus: número del clúster actual de proceso

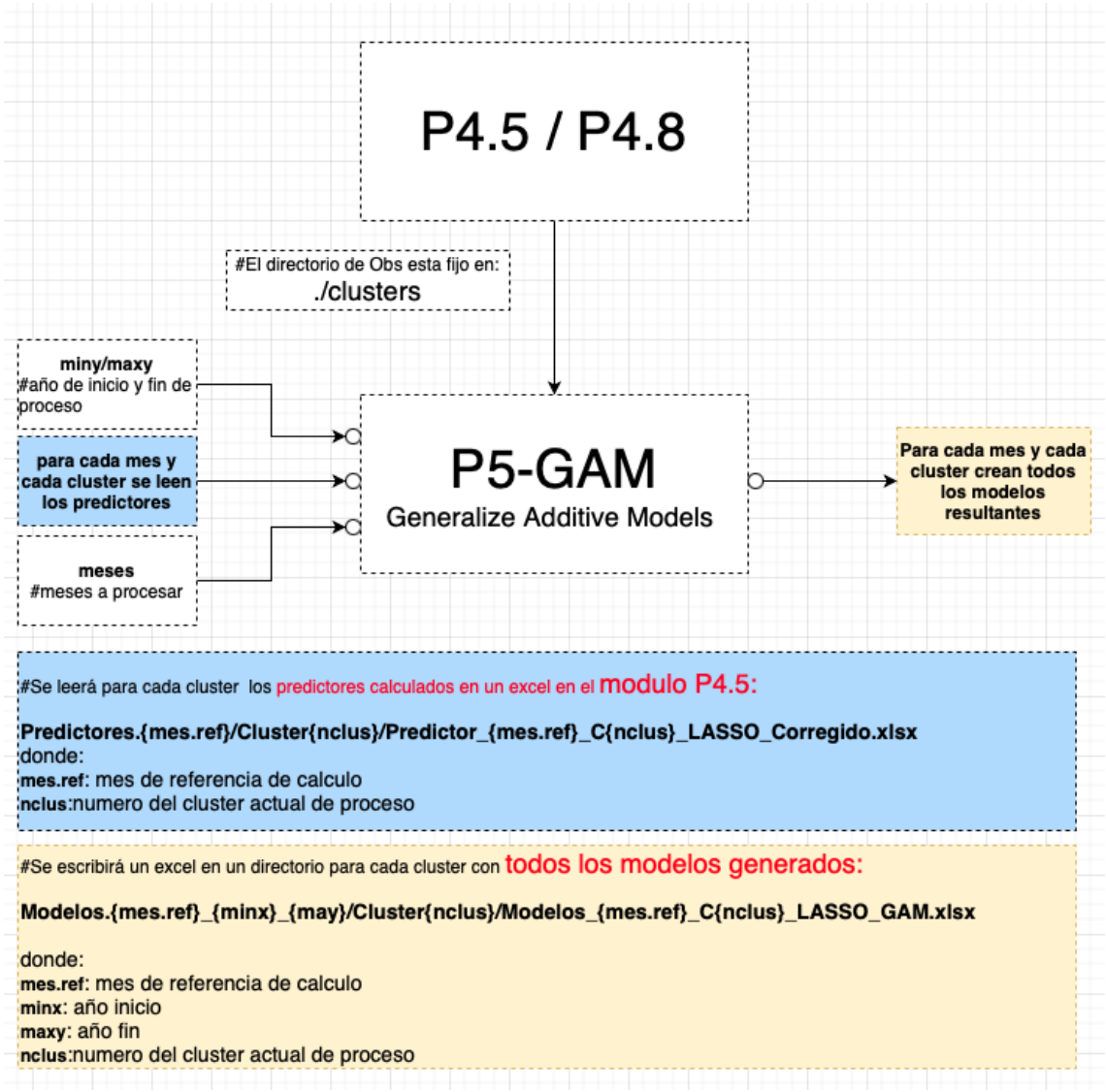


Figura 2.14. Esquema funcional del módulo P5-GAM.

La Tabla 2.11 muestra las librerías R que se utilizan en este módulo:

Tabla 2.11. Librerías R utilizadas en módulo P5-GAM.

Nombre	Descripción
openxlsx	Schauberger P, Walker A (2023). Leer, escribir y editar archivos xlsx.
fpp	Hyndman RJ (2013). Data for "Forecasting: principles and practice
forecast	Hyndman R, Athanasopoulos G, Bergmeir C, Caceres G, Chhay L, O'Hara-Wild M, Petropoulos F, Razbash S, Wang E, Yasmeen F (2023). Funciones de pronóstico para series de tiempo y modelos lineales
mgcv	Wood, S.N. (2011). Estimación rápida estable de máxima verosimilitud restringida y verosimilitud marginal de modelos lineales generalizados semiparamétricos.

A continuación, se muestra (Tabla 2.12) un ejemplo de tabla en Excel correspondiente a la salida de GAM. En ella también se muestra la columna de R^2 (Rsquared) y R^2 ajustado (Radj) para este tipo de modelo.

Tabla 2.12.ejemplo de salida del módulo GAM.

B	C	D
Formula	Rsquared	Radj
y.dato ~ s(u850_C1_8,k=3)+s(hgt200_C1_2,k=3)	0,472	0,402
y.dato ~ s(u850_C1_8,k=3)+s(hgt200_C1_2,k=3)+s(u850_C1_14,k=3)	0,503	0,396
y.dato ~ s(hgt200_C1_2,k=3)+s(u850_C1_14,k=3)	0,361	0,276
y.dato ~ s(u850_C1_8,k=3)	0,297	0,253
y.dato ~ s(u850_C1_8,k=3)+s(u850_C1_14,k=3)	0,339	0,251
y.dato ~ s(hgt200_C1_2,k=3)	0,245	0,198
y.dato ~ s(u850_C1_14,k=3)	0,164	0,112

2.3.10. Modelos de Redes Neuronales Artificiales (Artificial Neural Networks) (P5-ANN)

El objetivo de este módulo/programa es, para cada mes y cada clúster, crear todos los modelos de Artificial Neural Netwoks (ANN) que cumplen con los umbrales propuestos de R^2 ajustado. Para poder usar este módulo es necesario instalar Keras.

Las ANN se construyen mediante algoritmos de aprendizaje automático, en nuestro caso utilizando el marco Keras (Cholet et al, 2015), que ajusta las conexiones entre las neuronas para modelar un conjunto de datos determinado. De manera análoga a la estructura del cerebro, las ANN están compuestas por una gran cantidad de unidades de procesamiento simples (llamadas neuronas) conectadas entre sí en base a una arquitectura diseñada. Las neuronas se agrupan en capas, comenzando con una capa de entrada, para adquirir los datos, y terminando con una capa de salida para devolver los resultados. Una vez definida la arquitectura, se definen los pesos de la conexión entre neuronas, que son parámetros para ajustar los datos. Esto permite que el sistema "aprenda" y generalice ese aprendizaje. La salida de una neurona Yi se obtiene transformando la suma ponderada de las entradas que recibe mediante una función de activación:

$$Y_i = f\left(\sum_{J=1}^{n} w_{ij}x_j - \theta_i\right) = f(\sum_{j=0}^{n} w_{ij}x_j)$$

donde w_{ij} son los pesos, f es la función de activación y θi es el umbral de activación.

En este trabajo se utilizó la función de activación conocida como unidad lineal rectificada (ReLU). Es una función lineal por partes que generará la entrada directamente si es positiva; de lo contrario, generará cero. Se ha convertido en la función de activación predeterminada para muchos tipos de redes neuronales porque un modelo que la usa es más fácil de entrenar y, a menudo, logra un mejor rendimiento.

Se definieron cuatro arquitecturas de red como se muestra en la figura 2.15:

- 2 capas de 16 y 32 neuronas con dropout de 0,1 y 0,2 con ReLU como función de activación.
- 2 capas de 32 neuronas con dropout de 0,1 y 0,2 con ReLU como función de activación.

El término "dropout" (abandono) se refiere al dropout de los nodos (capa de entrada y oculta) en una red neuronal (como se ve en la Figura 2.15). Todas las conexiones hacia adelante y hacia atrás con un nodo descartado se eliminan temporalmente, creando así una nueva arquitectura de red a partir de la red principal. Los nodos se descartan con una probabilidad de abandono (dropout) de 'p'.

En el problema de overfitting (sobreajuste), el modelo aprende el ruido estadístico. Para ser precisos, el motivo principal del entrenamiento es disminuir la función de pérdida, dadas todas las unidades (neuronas). Entonces, en el sobreajuste, una unidad puede cambiar de una manera que corrige los errores de las otras unidades. Esto conduce a coadaptaciones complejas, lo que a su vez conduce al problema de sobreajuste porque esta coadaptación compleja no logra generalizar en el conjunto de datos invisible.
Ahora, si usamos dropout, evita que estas unidades corrijan el error de otras unidades, evitando así la coadaptación, ya que en cada iteración la presencia de una unidad es altamente poco confiable. Entonces, al soltar aleatoriamente algunas unidades (nodos), obliga a las capas a asumir más o menos responsabilidad por la entrada al adoptar un enfoque probabilístico.
Esto asegura que el modelo se generalice y, por lo tanto, reduzca el problema de sobreajuste.

2.3.10.1. Diagrama de funcionamiento.

La Figura 2.15 muestra la arquitectura definida para las redes neuronales de tipo BPN (Back-Propagation-Network).

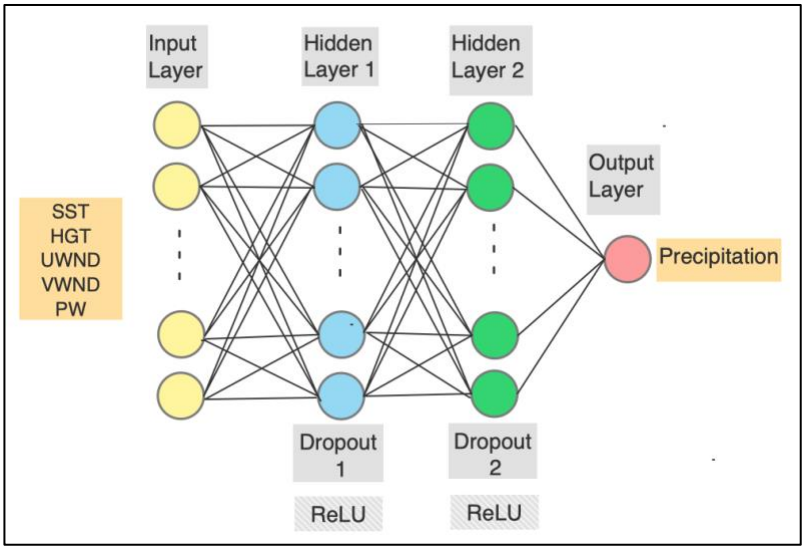


Figura 2.15. Arquitectura de las redes neuronales.

El diagrama siguiente (Figura 2.16) muestra el esquema funcional del módulo P5-ANN.

Existen un directorio de entrada con la información de datos observacionales de entrada ('clusters') .

Los parámetros de entrada al módulo P5-ANN son los años de inicio y fin de proceso. Para cada mes y cada clúster, se leen los predictores y por último los meses a procesar.

Las salidas del módulo P5-ANN son, para cada mes y cada clúster, todos los modelos que superen del umbral de R^2 ajustado definido.

El módulo P5-ANN (azul en la Figura 2.16) leerá para cada clúster los predictores calculados en un Excel en el módulo P4.5/P4.8:

Predictores.{mes.ref}/Cluster{nclus}/Predictor_{mes.ref}_C{nclus}_LASSO_Corregido.xlsx

donde:

mes.ref: mes de referencia de cálculo

nclus: número del clúster actual de proceso

El módulo P5-ANN (amarillo) escribirá un Excel en un directorio para cada clúster con todos los modelos generados:

Modelos.{mes.ref}_{minx}_{may}/Cluster{nclus}/Modelos_{mes.ref}_C{nclus}_LASSO_ANN.xlsx

Modelos.{mes.ref}_{minx}_{may}/Cluster{nclus}/Modelos_{mes.ref}_C{nclus}_LASSO_ANN{red}.xlsx

donde:

mes.ref: mes de referencia de cálculo

minx: año inicio

maxy: año fin

nclus: número del clúster actual de proceso

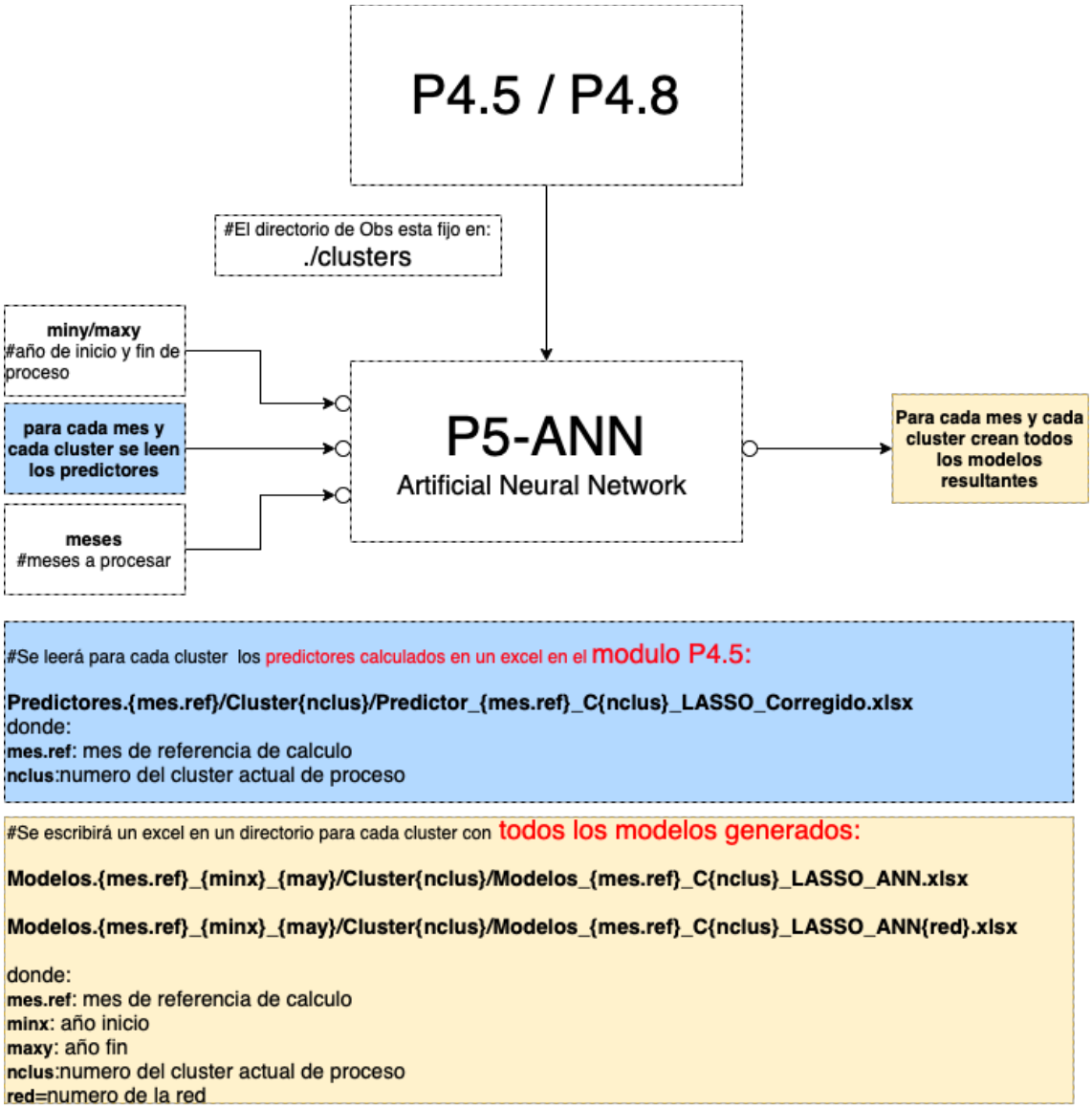


Figura 2.16. Esquema funcional del módulo P5-ANN.

La Tabla 2.13 muestra las librerías R que se utilizan en este módulo:

Tabla 2.13. Librerías R utilizadas en módulo P5-ANN.

Nombre	Descripción
openxlsx	Schauberger P, Walker A (2023). Leer, escribir y editar archivos xlsx.
fpp	Hyndman RJ (2013). Data for "Forecasting: principles and practice
forecast	Hyndman R, Athanasopoulos G, Bergmeir C, Caceres G, Chhay L, O'Hara-Wild M, Petropoulos F, Razbash S, Wang E, Yasmeen F (2023). Funciones de pronóstico para series de tiempo y modelos lineales
keras	Allaire J, Chollet F (2023). R Interface to 'Keras'.
tensorflow	Allaire J, Tang Y (2022). R Interface to 'TensorFlow'.

A continuación se muestra (Tabla 2.14) un ejemplo del Excel de salida de ANN, se muestra la columna de R^2 (Rsquared) y R^2 ajustado (Radj) para este tipo de modelo.

Tabla 2.14. Ejemplo de salida de módulo P5-ANN.

A	B	C	D	E	F	G	H	I	J	K	L	M	N
	#Modelo	Capas	Entradas	Salidas	Error	Epochs	Capa	Tipo	Neuronas	Activacion	Rate	Rsquared	Radj
		4	3	1	0,02189175	1000						0,634	0,599
							0	dense	16	relu			
							1	drop	0,2				
							2	dense	32	relu			
							3	dense	1	relu			
	#Modelo	Capas	Entradas	Salidas	Error	Epochs	Capa	Tipo	Neuronas	Activacion	Rate	Rsquared	Radj
		4	3	1	0,01169063	1000						0,867	0,854
							0	dense	32	relu			
							1	drop	0,2				
							2	dense	32	relu			
							3	dense	1	relu			
	#Modelo	Capas	Entradas	Salidas	Error	Epochs	Capa	Tipo	Neuronas	Activacion	Rate	Rsquared	Radj
		4	3	1	0,01589964	1000						0,796	0,776
							0	dense	16	relu			
							1	drop	0,1				
							2	dense	32	relu			
							3	dense	1	relu			
	#Modelo	Capas	Entradas	Salidas	Error	Epochs	Capa	Tipo	Neuronas	Activacion	Rate	Rsquared	Radj
		4	3	1	0,013471	1000						0,869	0,856
							0	dense	32	relu			
							1	drop	0,1				
							2	dense	32	relu			
							3	dense	1	relu			

Una "epoch" en una red neuronal es el entrenamiento de la red neuronal con todos los datos de entrenamiento durante un ciclo.

El tipo de neurona “dense” es la cantidad de neuronas en la capa.

2.4. Pronóstico Probabilístico (P7)

El objetivo de este módulo/programa es generar un pronóstico probabilístico, evaluando todos los modelos generados con los módulos P5 y verificando con las observaciones.

El diagrama siguiente (Figura 2.17) muestra el esquema funcional del módulo P7.

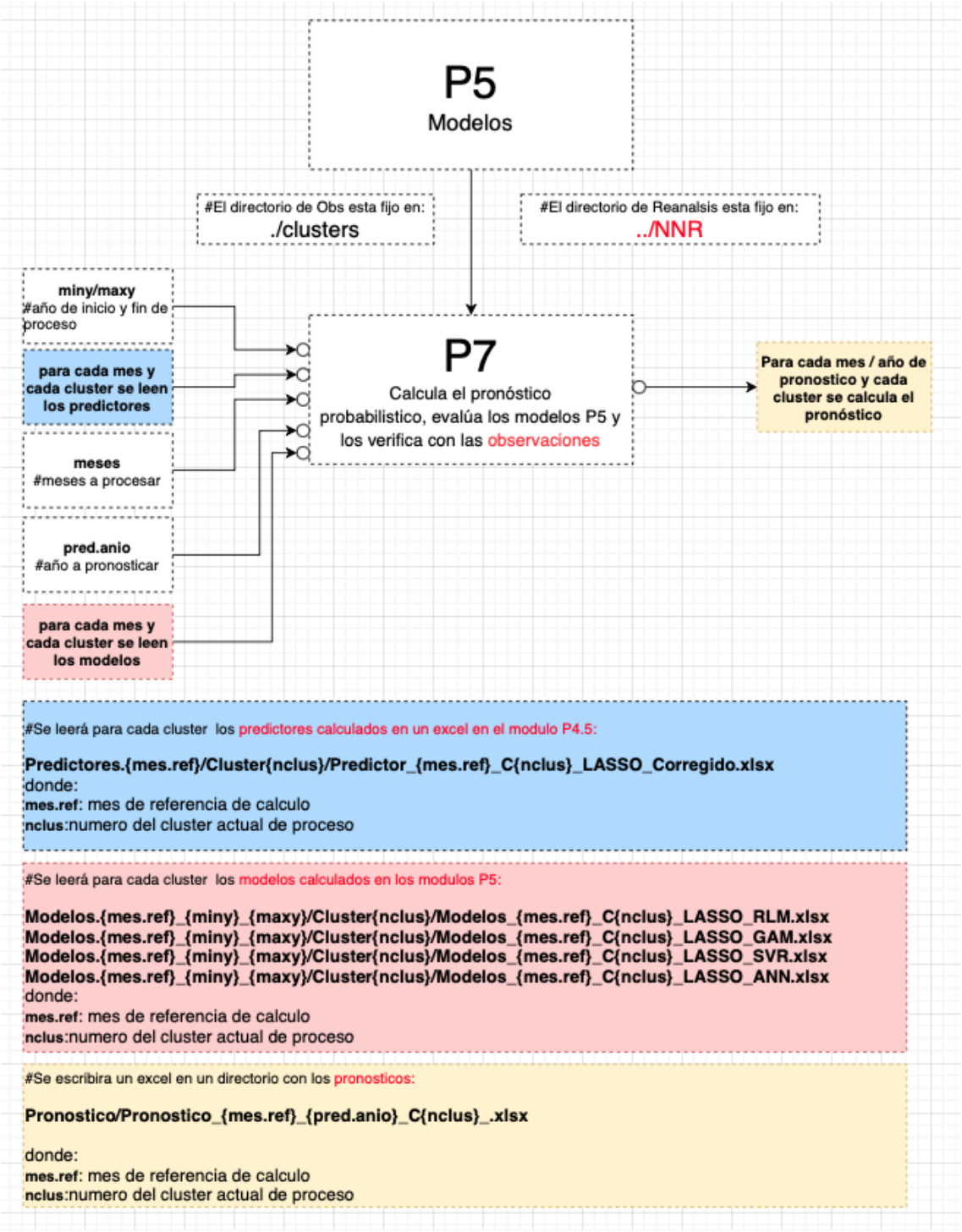


Figura 2.17. Esquema funcional del módulo P7.

La Tabla 2.15 muestra las librerías R que se utilizan en este módulo:

Tabla 2.15. Librerías R utilizadas en módulo P7.

Nombre	Descripción
dplyr	Wickham H, François R, Henry L, Müller K, Vaughan D (2023). Una gramática de la manipulación de datos.
openxlsx	Schauberger P, Walker A (2023). Leer, escribir y editar archivos xlsx.

A continuación se muestra (Tabla 2.16 a 2.19) ejemplos de las distintas partes de la tabla en Excel de salida del módulo P7 y se explicita su contenido:

Tabla 2.16. Ejemplo de la Parte 1 del Excel.

A	B	C	D	E	F	G	H
N.Modelo	adjR	Umbral	Pred.Anio	Formula	Tipo	Prono	obs
1	0,519	0,5	2016	y.dato ~ s(v850_C2_38,k=3)+s(u850_C2_21,k=3)+s(hgt200_C2_11,k=3)	GAM	86,07	32,825
2	0,657	0,5	2016	y.dato ~ v850_C2_38+hgt200_C2_11	SVR	50,99	32,825
3	0,565	0,5	2016	y.dato ~ u850_C2_21+hgt200_C2_11	SVR	41,38	32,825
4	0,503	0,5	2016	y.dato ~ v850_C2_38+u850_C2_21+hgt200_C2_11	SVR	44,62	32,825
5	0,616	0,5	2016	ANN_1	ANN	37,64	32,825
6	0,856	0,5	2016	ANN_2	ANN	19,92	32,825
7	0,845	0,5	2016	ANN_3	ANN	80,58	32,825
8	0,917	0,5	2016	ANN_4	ANN	59,02	32,825

- Se observa que 8 modelos superaron el umbral de 0.5 de R^2 ajustado.
- Se ven los predictores que quedaron en las fórmulas de pronóstico.
- Sólo superaron el umbral (0.5), 3 de los cuatro tipos de modelos: GAM, SVR, ANN
- El tipo RLM no generó modelos que superen el umbral de 0.5
- Se muestra el pronóstico y la observación para verificar

Tabla 2.17.Ejemplo de la Parte 2 del Excel.

J	K	L	M
Quintiles	P.mayor.que	lim.inf	Proba
7.70 - 16.70	P[pre> 7.70]	7,7	1
16.70 - 30.68	P[pre> 16.70]	16,695	1
30.68 - 41.41	P[pre> 30.68]	30,6807692	0,88
41.41 - 54.69	P[pre> 41.41]	41,415	0,62
54.69 - 108.75	P[pre> 54.69]	54,689	0,38

- La primer columna son los quintiles de las observaciones
- La segunda columna indica la probabilidad "mayor que ..." el límite inferior del intervalo
- La tercera columna, el límite inferior del intervalo
- La cuarta columna, la probabilidad acumulada asociada a cada intervalo

Tabla 2.18.Ejemplo de la Parte 3 del Excel.

O	P	Q	R	S	T	U
Quintiles	P.entre	Proba	obs	i.obs	i.prono	IDX
7.70 - 16.70	P[pre > 7.70 Y pre < 16.70]	0	32,825	3	5	-2
16.70 - 30.68	P[pre > 16.70 Y pre < 30.68]	0,12				
30.68 - 41.41	P[pre > 30.68 Y pre < 41.41]	0,25				
41.41 - 54.69	P[pre > 41.41 Y pre < 54.69]	0,25				
54.69 - 108.75	P[pre > 54.69 Y pre < 108.75]	0,38				

- La primer columna son los quintiles de las observaciones
- La segunda columna indica la probabilidad del intervalo "entre quintiles ..."
- La tercer columna, la probabilidad asociada
- La cuarta columna, la OBSERVACION
- La quinta columna, el intervalo en que cae la observación
- La sexta columna, el intervalo en el que cae la mayor cantidad de modelos
- La séptima columna, IDX: el intervalo diferencia o error definido como (i.obs-i.prono)

Tabla 2.19.Ejemplo de la Parte 4 del Excel.

Tipo	N	MEDIA	DESVIO
ANN	4	49,29	26,28
GAM	1	86,07	
SVR	3	45,66	4,89
TOTAL	8	52,53	22,13

- La primer columna es el tipo de modelo
- La segunda columna es la cantidad de modelos usados
- La tercer columna es la media por tipo de modelo
- La cuarta columna es el desvío estándar
- La última fila es el ensamble medio de los modelos

2.5. Pronóstico probabilístico operativo para el próximo período (P8)

El objetivo de este módulo/programa es similar al Modulo P7 pero tiene como objetivo realizar el pronóstico en forma operativa. Entonces, como estamos pronosticando con anticipación y por lo tanto, NO se dispone de la observación, TODO lo relacionado a verificación es omitido en el Excel resultante.

El diagrama siguiente (Figura 2.18) muestra el esquema funcional del módulo P7.

Antes de usar este módulo, se deben completar los archivos de predictores con el predictor del último mes, pero calculado con los días hasta el momento de realizar el pronóstico. Esto se debe a que operativamente el pronóstico para un mes determinado, se realiza los últimos días del mes anterior y por lo tanto no se dispone de la información completa del mes.

Para ello, se tiene que usar el script "predicProno.sh {año}" con el parámetro del año correspondiente (ej.: "predicProno.sh 2021" en la máquina virtual Linux de VirtualBox.

SCRIPT predicProno.sh:

El script "predictprono.sh {año}" creará el directorio NNRMes con los reanálisis de {hgt200.nc, hgt500.nc, hgt1000.nc, st.nc, tcw.nc, u850.nc, v850.nc} de todos los meses de los dos últimos años y el mes actual con los días que tiene hasta el momento de realizar el pronóstico.

El contenido de ese directorio NNRMes hay que copiarlo en nuestra máquina Windows usando el directorio compartido Windows <-> Linux.

Una vez que lo copiamos correr el P4.8 para agregar los predictores del último mes en el Excel correspondiente.

SCRIPT predictores.sh:

Cada vez que termina un año, hay que usar el script "predictores.sh" en la máquina virtual, para tener los reanálisis completos en el periodo desde 1979 hasta el año actual.

Cambiar en el script predictores.sh cambiar la línea 69 con el año que hay que descargar "for ((a=1979; a<= 2021; a++)); do" acá cambiar 2021 por el año que termino ej.; 2022

Este script crea el directorio NNR conteniendo los archivos de las variables de los reanálisis de todos los años.

El contenido del directorio hay que copiarlo al directorio Windows "NNR"

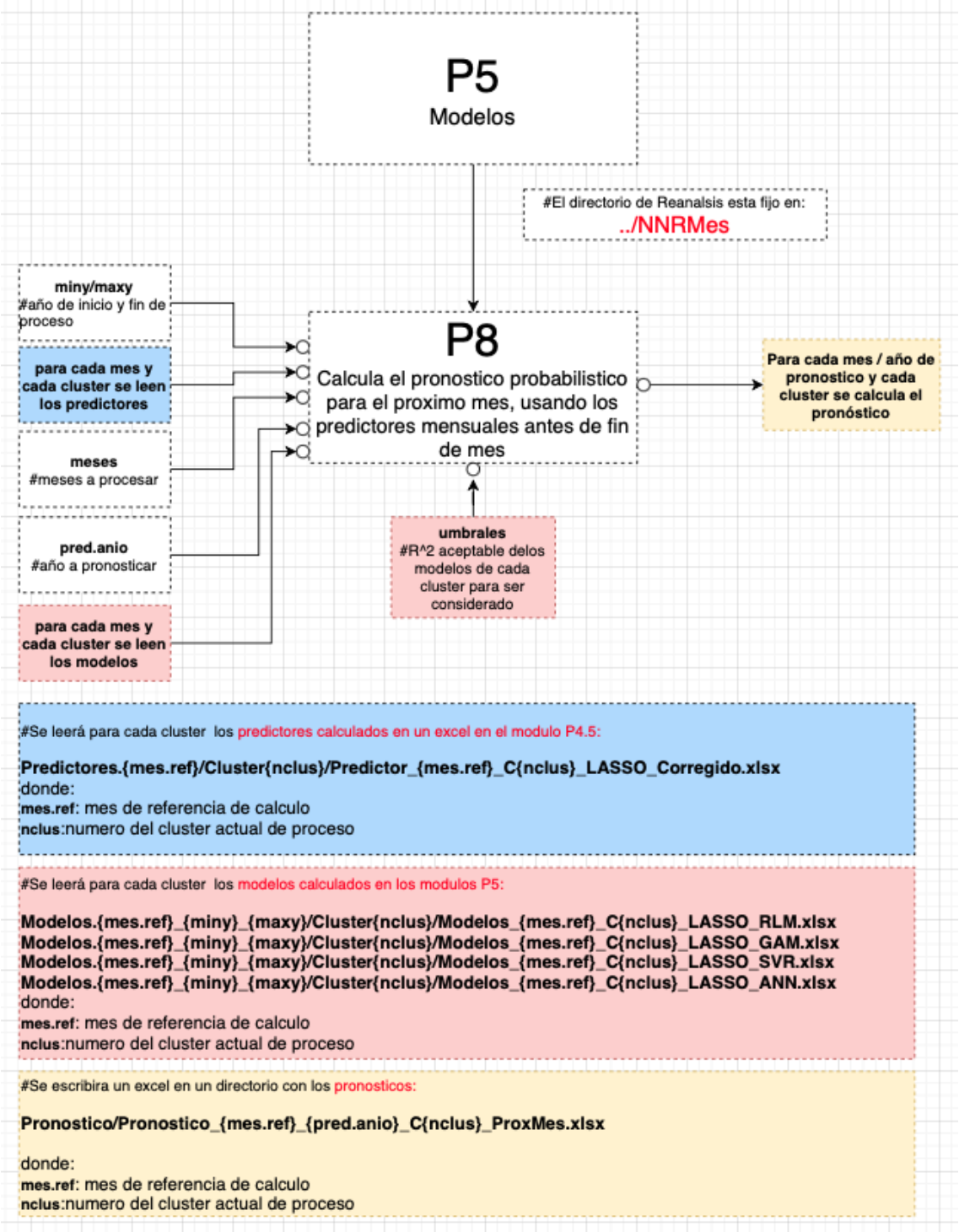


Figura 2.18. Esquema funcional del módulo P8.

La Tabla 2.20 muestra las librerías R que se utilizan en este módulo:

Tabla 2.20.Librerías R utilizadas en módulo P8.

Nombre	Descripción
openxlsx	Schauberger P, Walker A (2023). Leer, escribir y editar archivos xlsx.
fpp	Hyndman RJ (2013). Forecasting: principles and practice

e1071	Meyer D, Dimitriadou E, Hornik K, Weingessel A, Leisch F (2023). Funciones Misceláneas del Departamento de Estadística, Grupo de Teoría de la Probabilidad
mgcv	Wood, S.N. (2011). Estimación rápida estable de máxima verosimilitud restringida y verosimilitud marginal de modelos lineales generalizados semiparamétricos.
dplyr	Wickham H, François R, Henry L, Müller K, Vaughan D (2023). Una gramática de la manipulación de datos.

La Tabla 2.21 muestra los modelos con los cuales se calcularan las probabilidades.

Tabla 2.21. Ejemplo de los modelos de Sep-Oct-Nov, año 2021 Clúster 1 para calcular probabilidades.

N.Modelo	adjR	Umbral	Pred.Anio	Formula	Tipo	Prono
1	0,523	0,5	2021	y.dato ~ s(sst_	GAM	20,6
2	0,522	0,5	2021	y.dato ~ s(sst_	GAM	1,46
3	0,894	0,5	2021	y.dato ~ sst_C	SVR	207,69
4	0,887	0,5	2021	y.dato ~ sst_C	SVR	214,53
5	0,848	0,5	2021	y.dato ~ sst_C	SVR	170,26
6	0,794	0,5	2021	y.dato ~ sst_C	SVR	150,63
7	0,752	0,5	2021	y.dato ~ hgt2(	SVR	190,1
8	0,74	0,5	2021	y.dato ~ hgt2(	SVR	180,43
9	0,694	0,5	2021	y.dato ~ sst_C	SVR	197,23
10	0,685	0,5	2021	y.dato ~ hgt2(	SVR	149
11	0,679	0,5	2021	y.dato ~ sst_C	SVR	210,96
12	0,678	0,5	2021	y.dato ~ sst_C	SVR	157,15
13	0,666	0,5	2021	y.dato ~ hgt2(	SVR	182,99
14	0,658	0,5	2021	y.dato ~ u85(	SVR	226,43
15	0,645	0,5	2021	y.dato ~ u85(	SVR	194,8
16	0,637	0,5	2021	y.dato ~ sst_C	SVR	216
17	0,611	0,5	2021	y.dato ~ hgt2(	SVR	206,84
18	0,597	0,5	2021	y.dato ~ sst_C	SVR	134,47
19	0,573	0,5	2021	y.dato ~ sst_C	SVR	146,35
20	0,523	0,5	2021	y.dato ~ hgt2(	SVR	92,04
21	0,925	0,5	2021	ANN_1	ANN	115,31
22	0,946	0,5	2021	ANN_2	ANN	75,17
23	0,976	0,5	2021	ANN_3	ANN	75,17
24	0,985	0,5	2021	ANN_4	ANN	75,17

A continuación se muestran (Tabla 2.22) los quintiles definidos y las probabilidades calculadas:

Tabla 2.22. Quintiles definidos y las probabilidades calculadas

Quintiles	P.entre	Proba
75.17 - 119.60	P[pre > 75.17 Y pre < 119.60]	0,21
119.60 - 172.21	P[pre > 119.60 Y pre < 172.21]	0,25
172.21 - 199.89	P[pre > 172.21 Y pre < 199.89]	0,21
199.89 - 248.09	P[pre > 199.89 Y pre < 248.09]	0,25
248.09 - 434.33	P[pre > 248.09 Y pre < 434.33]	0

La visualización de este cuadro es un diagrama de la probabilidad de que la precipitación supere cierto umbral (Figura 2.19). A continuación se muestra este diagrama:

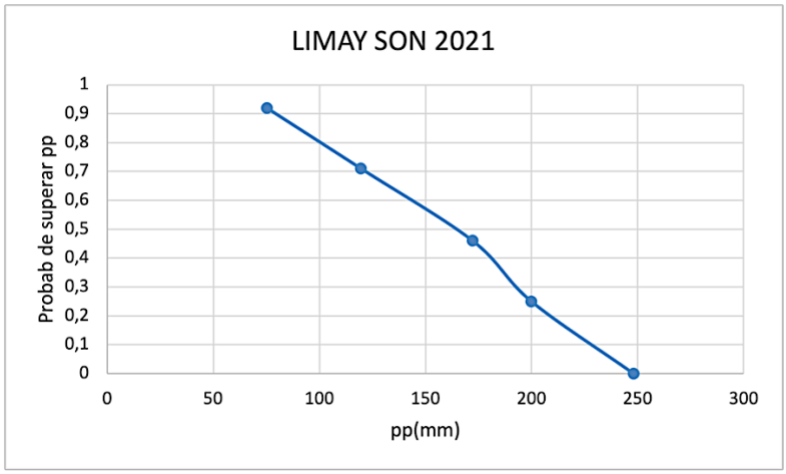


Figura 2.19. Ejemplo diagrama de probabilidad de pronóstico.

2.6. Verificación del pronóstico y errores (P9)

Este módulo lee todos los pronósticos realizados para la verificación y arma un archivo en formato Excel con 4 pestañas cada una conteniendo resúmenes de los errores de todos los pronósticos.

El diagrama siguiente (Figura 2.20) muestra el esquema funcional del módulo P7.

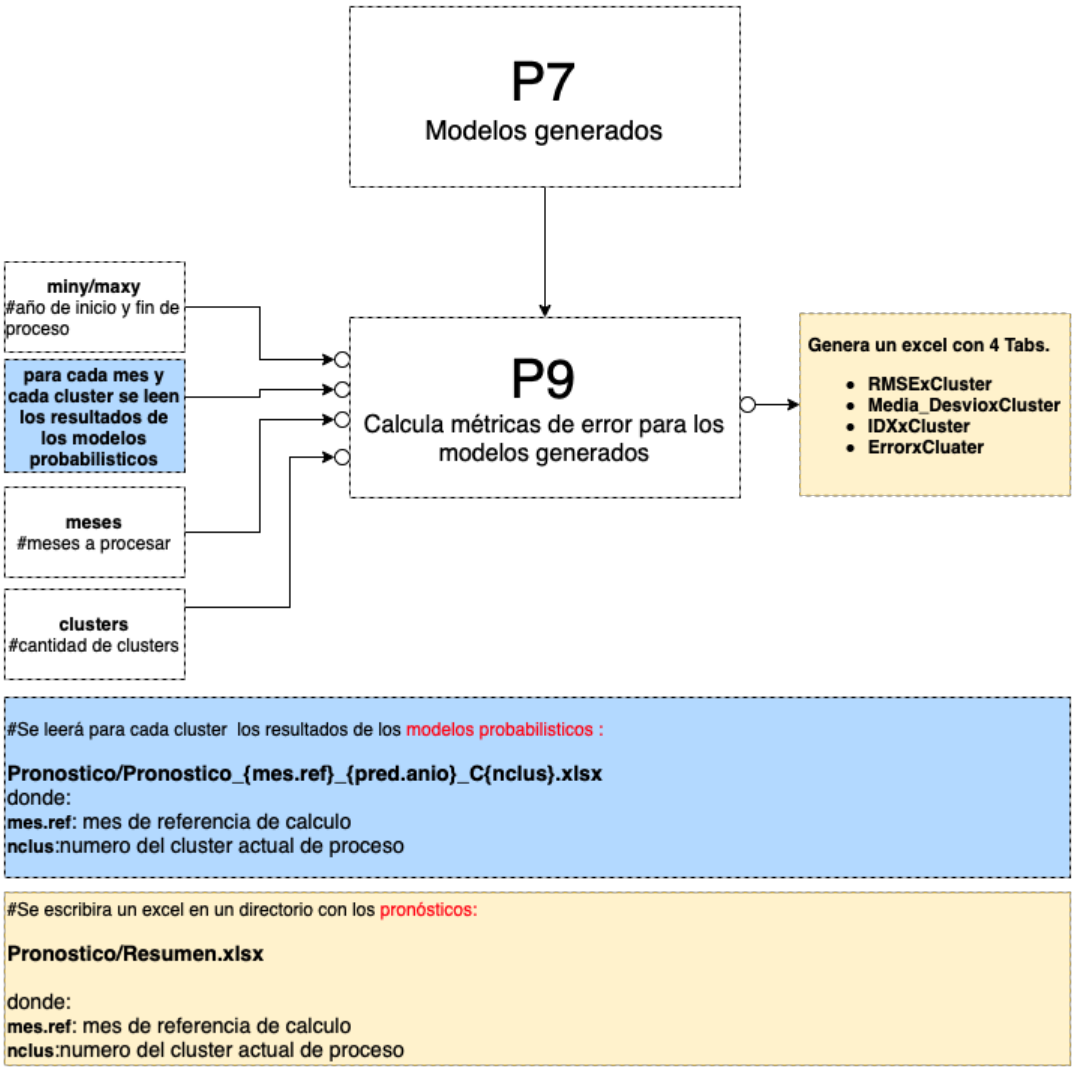


Figura 2.20. Esquema funcional del módulo P9.

Pestaña 1 (**RMSExCluster**):

Contiene la Raíz cuadrada del Error Cuadrático medio por cada clúster y cada mes (Tabla 2.23), además del RMSE del ensamble medio. A continuación se muestra un ejemplo de esta salida:

Tabla 2.23.Pestaña 1 (RMSExCluster).

CLUSTER: 1					
mes	RMSE_RLM	RMSE_GAM	RMSE_SVR	RMSE_ANN	RMSE_ENS
01			77,7	92,7	89,5
02	74,7	74,5	56,2	55,2	61,8
03	90,1	70,7	121,1	76,8	103
04	168,7	157,6	154,9	163,6	160,8
05	296,7	277,2	229,2	232,9	254,4
06	130	139,2	137,3	189,6	161,1
07	165,4	169,9	109,4	210,9	142,8
08	74,2	98,7	88,8	91	88,3
09	116,6	96,5	68,8	82,7	79
10			78,4	75,7	76,1
11	65,2	31,3	46,8	80,9	53,7
12	44,6		41,9	84,5	66,2

Pestaña 2 (**Media_DesvioxCluster**):

Contiene las medias y desvíos de los pronósticos por clúster y mes. A continuación se muestra (Tabla 2.24) un ejemplo de esta salida:

Tabla 2.24.Pestaña 2 (Media_DesvioxCluster).

CLUSTER: 1										
mes	MEAN_RLM	DESV_RLM	MEAN_GAM	DESV_GAM	MEAN_SVR	DESV_SVR	MEAN_ANN	DESV_ANN	MEAN_ENS	DESV_ENS
01					76,8	37	18,8	25,4	32,2	37,2
02	174,6	59,1	179	72,4	152,4	39,2	154,4	53,5	160,1	52
03	161	86,1	189,7	47,5	251	89,3	180,4	57,8	212,5	88,5
04	257,9	84,7	245,3	68,7	356,9	101,8	215	66,6	289,5	104
05	603,2	120,3	582,7	141,3	512,6	113,3	426,1	166,3	540,4	131,3
06	482,2	72,2	487,6	76,9	519,7	68,4	265,4	60,9	402,6	136,4
07	479,3	94,5	487,6	92,3	419,7	66,2	525,6	102,1	451,8	87,8
08	279,7	5,5	268,2		281,3	36,9	287,3	22,5	283,1	29,7
09	143,9	100,9	151,5	89,9	177,4	54	160,8	91,2	169,7	68,9
10					160,7	118,4	90,5	38,9	99,6	56,2
11	159,7	34,9	125,7	35,8	140,8	28,5	154,1	52,5	141,6	37,5
12	149,6				106,6	22,9	137,6	58,7	122,8	46,3

Pestaña 3 (**IDXxCluster**):

Contiene el cálculo de los IDXmedios y desvíos de los pronósticos por clúster y mes. A continuación se muestra (Tabla 2.25) un ejemplo de esta salida:

Tabla 2.25.Pestaña 3 (IDXxCluster).

CLUSTER: 1							
mes	IDX_MEAN	IDX_DESV	X2016	X2017	X2018	X2019	X2020
01	1,2	1,8	0	2	4	0	0
02	-0,6	1,1	-1	-2	1	-1	0
03	0,2	1,5	-2	0	2	1	0
04	0,8	1,3	-1	0	1	2	2
05	-1,2	2	-3	-2	-3	1	1
06	1,2	1,1	0	2	0	2	2
07	-1,8	1,1	-2	-2	-3	-2	0
08	0	2	2	2	0	-2	-2
09	-0,2	1,9	-3	1	2	0	-1
10	1,8	1,1	3	3	1	1	1
11	-1,4	1,3	0	-2	-2	-3	0
12	0,2	1,9	1	1	-3	0	2

Pestaña 4 (**ErrorxCluster**):

Contiene el cálculo de los errores y desvíos de los pronósticos por clúster y mes. A continuación se muestra un ejemplo de esta salida:

Tabla 2.26.Pestaña 4 (ErrorxCluster y por mes).

CLUSTER: 1 MES: 01						
Mes	Metodo	2016	2017	2018	2019	2020
01	ANN	-0,7	-116	-152,4	-42,9	-45
01	GAM					
01	RLM					
01	SVR	58	-61,5	-101,2		5,9
01	ENS	18,9	-105,1	-135,3	-42,9	-34,8

.....

CLUSTER: 1 MES: 12						
Mes	Metodo	2016	2017	2018	2019	2020
12	ANN	81,6	-4,5	93,5	129	-48,2
12	GAM					
12	RLM		44,6			
12	SVR	-1,4	2	74,9	35,1	-18,9
12	ENS	35,5	3,8	84,2	88,7	-33,5

CAPÍTULO III:
PRONOSTICO DETERMINÍSTICO DE PRECIPITACIÓN MENSUAL EN LA REGIÓN GRAN CHACO ARGENTINO

En este capítulo se aplicará el FRAMEWORK que se ha detallado en el capítulo 2 para el pronóstico determinístico de precipitación en la región del Gran Chaco Argentino para la época estival. El pronóstico determinístico implica pronosticar la precipitación mensual para cada mes de verano con diferentes técnicas estadísticas. Esto permite obtener varios pronósticos de precipitación y por ende se puede construir con ellos el ensamble medio y verificar la bondad del método.

3.1. Importancia del pronóstico en esta región

El desarrollo de las predicciones estacionales ha tenido una evolución muy importante en las últimas décadas. Los grandes centros mundiales han desarrollado con éxito modelos dinámicos y estadísticos. Todos ellos se utilizan para generar conjuntos que mejoran los pronósticos individuales. En general, los modelos estadísticos se diseñan cuando se van a realizar pronósticos en áreas limitadas. Esto es lo que se muestra en este trabajo para la región del Gran Chaco argentino. El "Gran Chaco Americano" es una ecorregión forestal de excepcional diversidad ambiental y social. Con una superficie de 1.100.000 km2, ocupa el segundo lugar como ecorregión forestal en América del Sur, después de la Amazonía, e incluye territorios de Argentina, Paraguay, Bolivia y Brasil. Sus ambientes -bosques, matorrales, pastizales, sabanas, pantanos y humedales- la convierten en una región única en el mundo y es un ecosistema forestal que funciona como transición entre el trópico y el templado. Durante las últimas décadas, el Chaco se ha convertido en una de las tres regiones con mayores tasas de transformación a nivel mundial, superando a la cuenca amazónica. La porción del Gran Chaco argentino mostró las mayores tasas de deforestación (0,89% anual), en comparación con el promedio a escala nacional, continental y mundial (0,82%, 0,51% y 0,13%, anual respectivamente) (FAO, 2011) . La precipitación en esta región es en parte producida por el proceso de evapotranspiración y reciclado del vapor in situ. El cambio de la cobertura vegetal puede modificar este proceso. Por estas razones, el seguimiento y la previsión de las precipitaciones en escalas mensuales son importantes para los responsables de la toma de decisiones.

Un aspecto importante en la región es la alta variabilidad espacial de la precipitación, por lo que es un desafío poder pronosticarla con certeza utilizando métodos dinámicos o estadísticos. El primero se basa en el uso de un sistema de ecuaciones que gobiernan la atmósfera. El segundo modela la precipitación mediante técnicas estadísticas que aprenden de situaciones pasadas relacionando las condiciones atmosféricas previas con los valores de precipitación que se registran posteriormente. Ambos métodos todavía tienen una gran incertidumbre en los resultados. Muchos autores han señalado las dificultades aún detectadas a la hora de pronosticar la precipitación estacional y subestacional (Barnston et al. 2005; Leetmaa 2003; Coelho et al. 2005; Kumar 2006). En

América del Sur se ha realizado una evaluación del pronóstico estacional (Goddard 2003, Nobre et al. 2005; Barreiro 2009) que muestra la limitada eficiencia. En el caso de los pronósticos estadísticos, los errores pueden deberse a múltiples causas: los datos utilizados (González y Rolla 2019), la definición de los predictores y las metodologías de modelado. Las técnicas de minería de datos son herramientas muy importantes , aunque no suficientemente utilizada en meteorología, que permite asociar patrones atmosféricos previos para determinar el comportamiento de una variable en el futuro (Ebert-Uphoff y Hilburn 2020). Se aplicaron la metodología de regresión lineal múltiple (RLM) que asume solo asociaciones lineales entre los predictores y la variable a predecir. Los modelos aditivos generalizados (GAM) (Wood 2006) permiten relaciones no lineales entre las variables predictoras y el resultado. La regresión de vector de soporte (SVR) es un algoritmo de aprendizaje automático supervisado que se puede emplear tanto para fines de clasificación como de regresión (Cortés y Vapnik 1995). Finalmente, las redes neuronales brindan una herramienta que incorpora la posibilidad de reconocer relaciones subyacentes en un conjunto de datos a través de un proceso que imita la forma en que opera el cerebro humano (ANN) (Boukabara et al. 2019; Reichstein et al. 2019; Lee et al. 2018). Este capítulo de la tesis intenta comparar la eficiencia de las metodologías antes mencionadas y cuya programación fue detallada en el capítulo anterior, generando conjuntos de modelos estadísticos de pronóstico utilizando diferentes predictores/metodologías para el caso particular de precipitación en escalas mensuales en la región del Gran Chaco Argentino.

3.2. Metodología y Datos

El área de estudio corresponde a la Argentina subtropical al norte de los 34°S. El centro de esta región corresponde al Gran Chaco Argentino. La previsión se hará para el semestre de verano que comienza en octubre y finaliza en marzo. Esta temporada es la de mayor precipitación en la región y es relevante para cultivos como la soja.

Se utilizaron datos de precipitación mensual en 34 estaciones pertenecientes a la red de medición del Servicio Meteorológico Nacional de Argentina (Figura 3.1) para el período 1980-2017. Los datos fueron cuidadosamente analizados para evaluar su calidad y de tal forma que los faltantes no superen el 5%. Se utilizaron datos globales del reanálisis del

NCEP (Kalnay et al. 1996) para definir los predictores para las regiones/ clusters, ya que estaban disponibles en tiempo real en el momento de este trabajo. Las variables utilizadas fueron los valores mensuales de temperatura superficial del mar (SST), altura geopotencial (HGT) en varios niveles de la atmósfera (1000, 500 y 300 hPa), agua precipitable (PW) en la columna atmosférica y viento (UWND, VWND) en capas bajas (850Hpa). Según fueron definidas en capitulo 1

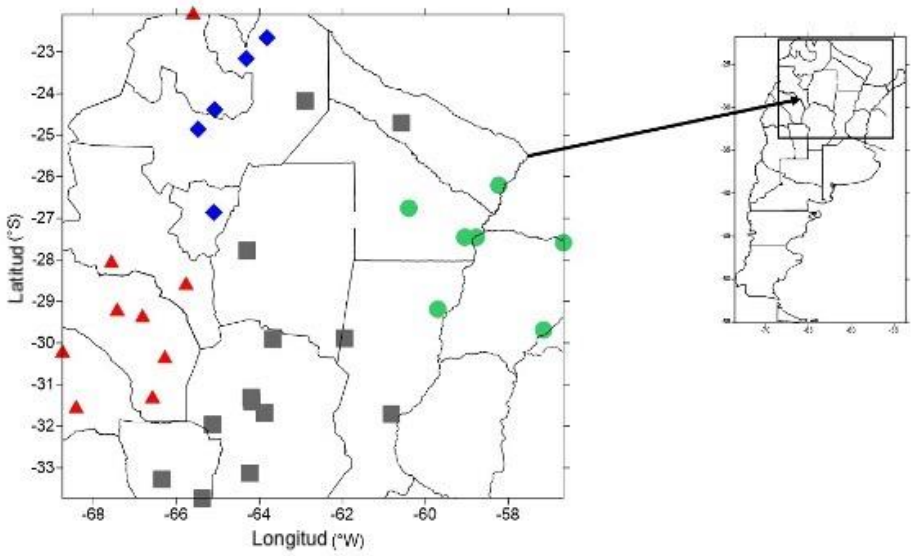


Figura 3.1.Área de estudio, análisis de clúster derivado de SOM (clúster 1 rojo, clúster 2 azul, clúster 3 gris y clúster 4 verde).

Se aplicó un análisis de agrupamientos para definir grupos utilizando todas las estaciones meteorológicas disponibles en la región. Se utilizaron diferentes métodos jerárquicos como enlace simple, enlace completo, promedio (Hartigan, J.A., 1985) y Ward (Murtagh at al., 2014) con matrices de distancia euclidiana y de correlación. También, métodos no jerárquicos como K-means (Hartigan, 1979) y métodos no supervisados como SOM (mapas autoorganizados) (Kaski S., 2011). Finalmente, se seleccionó el método SOM como el más apropiado debido a que definió muy bien el comportamiento climático regional utilizando un número reducido de cuatro regiones.
Para cada agrupamiento y para cada mes, la serie de precipitación media espacial se correlacionó con las variables meteorológicas observadas en el mes anterior para el período 1980-2008. Las variables globales utilizadas fueron la SST al sur de 10°N y la HGT en el hemisferio sur para determinar los forzantes climáticos y PW, UWND y VWND sobre la región de estudio para evaluar la circulación regional. Las áreas con correlación

estadísticamente significativa utilizando una prueba normal con un 95% de confianza (módulo de la correlación superior a 0,37), se utilizaron para definir los predictores. Se considera que del 2009 al 2017 no hubo cambios en la señal climática y por lo tanto la definición de los predictores se mantuvo durante todo el período de pronóstico. Además, configurar los predictores nos permite ver las diferencias reales que generan las diferentes metodologías. Sólo se consideraron aquellos predictores donde la relación entre ellos y la variable a pronosticar tuviera una explicación física razonable. Además, se definió un conjunto de predictores independientes para cada mes y cada grupo para evitar el problema de multicolinealidad, utilizando la técnica LASSO.

El proceso de construcción de los modelos se iteró avanzando un año en el período de entrenamiento para pronosticar el año siguiente como muestra la Figura 3.2.

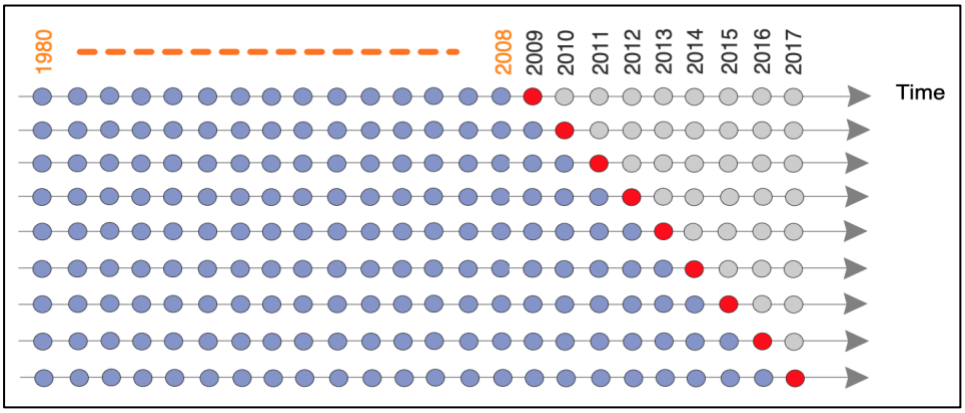


Figura 3.2. Esquema de construcción de modelos de pronóstico.

Para cada metodología se consideró el subconjunto de modelos que explican al menos el 50% de la varianza de la precipitación y su respectiva media del ensamble. Para cada clúster, mes y año pronosticado, se calcularon la media y la desviación estándar de la precipitación pronosticada derivada de cada metodología y la media del ensamble. El error cuadrático medio de la raíz también se calculó como:

$$RMSE = \sqrt{\sum(Yp_i - Y_i)^2}/_n$$

dónde Y_{pi} es el predictando y Y_i es la precipitación observada.

Se implementó un pronóstico probabilístico utilizando el conjunto de todos los modelos seleccionados. Se construyó una distribución probabilística utilizando los cinco intervalos de clase definidos por los quintiles de precipitación observada. En cada caso se podría predecir el intervalo con mayor probabilidad de ocurrencia. El índice IDX se definió como:

$$IDX = I_o - I_p$$

donde:

I_o: Número del intervalo de clase más probable de precipitación observada.

I_p: Número del intervalo de clase más probable para la precipitación pronosticada.

Con esta definición, un valor IDX negativo (positivo) indica que la precipitación fue sobreestimada (subestimada).

Para comparar los pronósticos de precipitación utilizando técnicas de aprendizaje automático con los pronósticos dinámicos derivados de los centros mundiales, se utilizaron datos de pronósticos estacionales del Servicio de Cambio Climático de Copernicus (C3S). Las instituciones que proporcionaron los modelos se detallan a continuación:

- ECMWF
- The Met Office
- Météo-France
- Deutscher Wetterdienst (DWD)
- Centro Euro-Mediterraneo sui Cambiamenti Climatici (CMCC)
- National Centers for Environmental Prediction (NCEP)
- Japan Meteorological Agency (JMA)
- Environment and Climate Change Canada (ECCC)

La siguiente tabla (Tabla 3.1) muestra información detallada de los modelos participantes

Tabla 3.1. información detallada de los modelos.

Institución	Cod. Modelo	Resolución (grados)	Mes inicio	Mes Fin	Num. Miembros
ECMWF	5	1x1	01-2009	12-2016	25
MeteoFrance	6	1x1	01-2009	12-2016	25
UKMetOffice	13	1x1	01-2009	12-2016	28
CMCC	35	1x1	01-2009	12-2016	40
JMA	2	2,5x2,5	01-2009	12-2016	10
DWD	21	1x1	01-2009	12-2016	30
NCEP	2	1x1	01-2009	12-2016	28
ECCC	2	1x1	01-2009	12-2016	10

La variable usada es **tprate** (tasa media de precipitación total mensual) en mm/mes

La región seleccionada es la que muestra la figura 3.3:

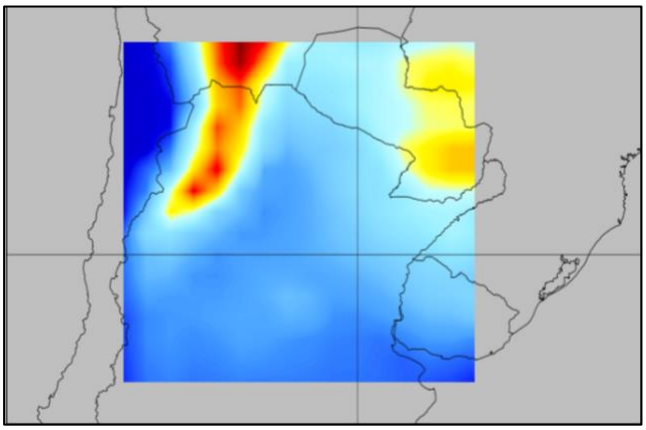


Figura 3.3. Región seleccionada. Las distintas tonalidad representan valores de tprate (mm/mes)

Los productos de predicción estacional de C3S se basan en datos de varios sistemas de predicción estacional de última generación. Están disponibles combinaciones multisistema, así como predicciones de los sistemas participantes individuales.

El RMSE se calculó a partir de las previsiones derivadas de cada centro mundial, promediadas sobre los puntos de la cuadrícula que componen cada uno de los clusters. Luego se calculó el RMSESS (root mean square error skill score) que se define como:

$$RMSESS = 1 - \ RMSEml/RMSEc3s$$

Donde $RMSE_{ml}$ es el RMSE derivado del conjunto de técnicas de aprendizaje automático y RMSEc3s es el del conjunto de modelos dinámicos C3S.

RMSESS mide si los valores de pronóstico de aprendizaje automático son mejores (positivos) o peores (negativos) que los derivados de los modelos C3

3.3. Resultados y discusión

3.3.1. El pronóstico determinístico de precipitación

El resultado del agrupamiento de las estaciones se muestra en la Figura 3.1 y la precipitación media anual para (1980-2017) en cada región en la Figura 3.4.

Se puede observar que la precipitación tiene un marcado ciclo anual, mayor hacia el oeste (Figura 3.4). El clúster 1 (rojo) tiene poca precipitación durante todo el período y es particularmente bajo en primavera. El clúster 2 (azul) muestra una gran diferencia en la precipitación acumulada a principios de la primavera en comparación con el verano. Los clusters 3 (gris) y 4 (verde) muestran valores altos de precipitación durante todo el período, aunque el clúster 3 presenta una mayor diferencia entre primavera y verano. Ambos tienen valores de precipitación acumulada anual más altos que los clusters 1 y 2.

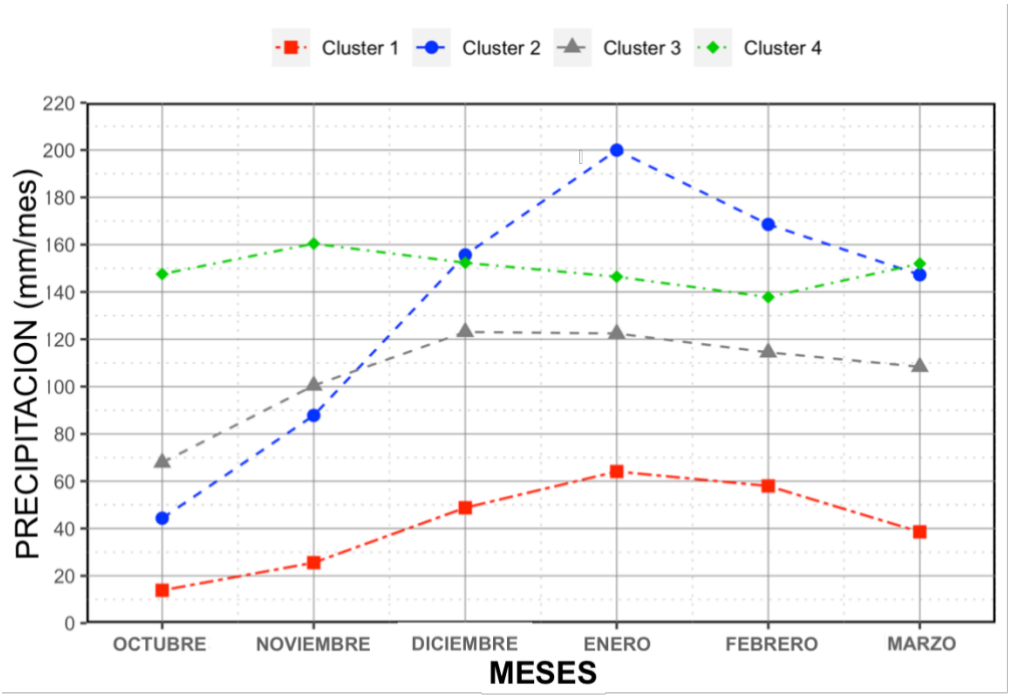


Figura 3.4.Evolución media anual (1980-2017) de los cuatro clústers.

Se aplicaron todas las metodologías de modelado para cada clúster y cada mes y se seleccionó el subconjunto de los mejores modelos. Este proceso se iteró para cada periodo de entrenamiento utilizado para pronosticar el año siguiente. Cabe señalar que en algunos casos ningún modelo cumplió la condición de explicar una varianza de la precipitación superior al 50%. La Tabla 3.2 muestra para cada estación (octubre a marzo), año pronosticado, clúster y el número de modelos considerados.

Tabla 3.2. Número de modelos considerados para cada estación, año pronosticado y clúster.

Estación	Año Pronosticado	Clúster	Numero de modelos				
			Total	ANN	GAM	RLM	SVR
(Oct-Mar)	2009	1	262	24	66	61	111
(Oct-Mar)	2009	2	136	24	22	27	63
(Oct-Mar)	2009	3	240	24	59	63	94
(Oct-Mar)	2009	4	135	24	27	36	48
(Oct-Mar)	2010	1	249	24	63	57	105
(Oct-Mar)	2010	2	128	24	20	28	56
(Oct-Mar)	2010	3	218	24	51	52	91
(Oct-Mar)	2010	4	133	24	29	34	46
(Oct-Mar)	2011	1	230	24	55	50	101
(Oct-Mar)	2011	2	121	23	17	22	59
(Oct-Mar)	2011	3	184	24	38	37	85
(Oct-Mar)	2011	4	106	24	19	24	39
(Oct-Mar)	2012	1	197	24	43	34	96
(Oct-Mar)	2012	2	122	23	19	20	60
(Oct-Mar)	2012	3	172	24	36	32	80
(Oct-Mar)	2012	4	114	23	23	23	45
(Oct-Mar)	2013	1	179	24	41	21	93
(Oct-Mar)	2013	2	112	23	16	17	56
(Oct-Mar)	2013	3	157	24	26	27	80
(Oct-Mar)	2013	4	96	23	12	20	41
(Oct-Mar)	2014	1	172	24	41	19	88
(Oct-Mar)	2014	2	107	23	14	15	55
(Oct-Mar)	2014	3	164	24	27	32	81
(Oct-Mar)	2014	4	88	24	7	16	41
(Oct-Mar)	2015	1	164	24	39	19	82
(Oct-Mar)	2015	2	106	24	14	15	53
(Oct-Mar)	2015	3	139	24	20	22	73
(Oct-Mar)	2015	4	70	23	3	10	34
(Oct-Mar)	2016	1	149	24	33	10	82
(Oct-Mar)	2016	2	101	23	12	15	51
(Oct-Mar)	2016	3	127	24	19	20	64
(Oct-Mar)	2016	4	60	23	0	5	32
(Oct-Mar)	2017	1	152	24	33	12	83
(Oct-Mar)	2017	2	74	23	1	0	50
(Oct-Mar)	2017	3	112	24	15	15	58
(Oct-Mar)	2017	4	53	21	0	2	30

En todos los casos se realizó el pronóstico categórico de precipitación. Para comprender mejor el proceso, se detalla el caso del pronóstico de precipitación de octubre de 2010 en el clúster 3. Los modelos se generaron utilizando el período de entrenamiento 1980-2009. La precipitación observada fue de 45,3 mm . La media y desviación estándar de los valores predichos utilizando las diferentes metodologías se muestran en la Tabla 3.3. La media pronosticada con la media del ensamble (50,9mm) fue la que más se acercó al valor observado aunque presenta una desviación estándar alta (15,29 mm). mm). Cada una de las metodologías predijo valores diferentes y la SVR fue la que presentó menor variabilidad dentro del grupo (9,85 mm).

Tabla 3.3. Media y desviación estándar de los valores predichos utilizando las diferentes metodologías para octubre de 2010, clúster 3.

Metodología	Número de modelos seleccionados	Precipitación media pronosticada (mm)	Desviación estándar (mm)
ANN	4	36,13	15,22
GAM	12	38,28	16,84
MLR	11	54,66	11,36
SVR	23	58,34	9,85
Ensamble medio	50	50,94	15,29

Los intervalos de clase se definieron utilizando los quintiles de la serie de precipitación observada en octubre en el período 1980-2009 (Tabla 3.4).

Tabla 3.4. Definición de intervalos de clase para octubre, clúster 3.

Intervalo de clase	Rango de precipitación (mm)
1	21,30 - 47,40
2	47,40 - 54,40
3	54,40 - 70,70
4	70,70 - 87,00
5	87,00 – 133,00

El valor de precipitación observado se ubica en el intervalo de primera clase que resulta en (I_o = 1). Con los 50 modelos que componen el ensamble se construyó la distribución de probabilidad (Figura 3.5) y se determinó el valor del intervalo pronosticado más probable I_p (I_p = 1) (Tabla 3.5). Como IDX = 0, el intervalo más probable fue bien pronosticado.

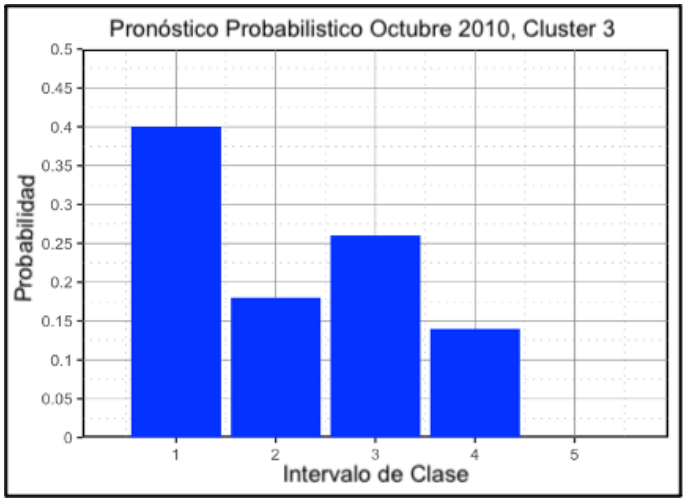


Figura 3. 5.Distribución de probabilidad de octubre de 2010, clúster 3.

Tabla 3.5. Valor del intervalo más probable predicho (Ip) y observado (Io) para octubre de 2010, clúster 3.

I_o	1
I_p	1
IDX	0

El período de verificación es 2009-2017. La previsión del año Aj se ha realizado utilizando como período de entrenamiento 1980-A_{j-1} con el que se construyeron los diferentes modelos. Los resultados obtenidos durante todo el período de verificación se resumen a continuación.

La Figura 3.6 muestra el RMSE calculado para el conjunto de modelos derivados de cada una de las metodologías y el ensamble medio (ENS) para cada uno de los meses de cada clúster. RMSE es mayor en los clusters 2 y 4 que en los clusters 1 y 3 porque la precipitación también es mayor. Es importante señalar que el RMSE es similar cuando se utilizan diferentes metodologías y, por lo tanto, la media del ensamble para cada mes. Ninguna metodología se destaca por su eficiente comportamiento. En el caso del clúster 2 en octubre, enero y febrero y en el clúster 4 en octubre, utilizando GAM no ha sido posible encontrar modelos que expliquen más del 50% de la varianza. Tampoco fue posible encontrar un modelo que supere el 50% de la varianza usando MLR en febrero en el clúster 2.

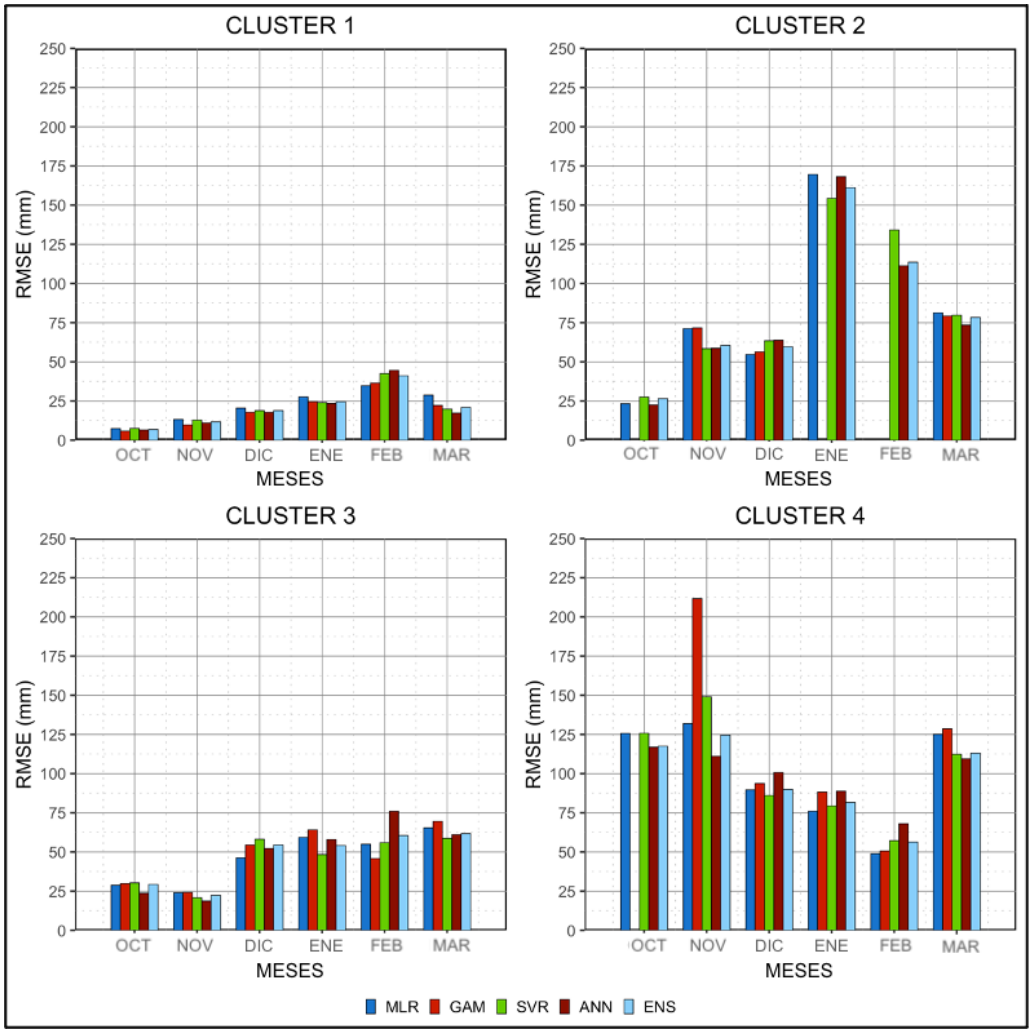


Figura 3. 6. RMSE para el conjunto de modelos derivados de cada una de las metodologías y de la media del ensamble para cada uno de los meses de cada clúster.

Tabla 3.6. Metodología con el RMSE más bajo para cada uno de los clusters y los meses (RMSE en mm), el mejor método para cada clúster en negrita cursiva.

Metodología (RMSE, mm)	Octubre	Noviembre	Diciembre	Enero	Febrero	Marzo
Clúster 1	***GAM*** (5,8)	***GAM*** (9,8)	***GAM*** (17,8)	ANN (23,5)	***GAM*** (36,5)	ANN (17,3)
Clúster 2	***ANN (22,5)***	SVR (58,5)	MLR (54,7)	SVR (154,3)	***ANN (111,3)***	***ANN (73,5)***
Clúster 3	***ANN (23,7)***	***ANN (18,7)***	***ANN (52,2)***	SVR (48,4)	GAM (45,8)	***ANN (61,1)***
Clúster 4	***ANN (116,9)***	***ANN (111,1)***	SVR (86,1)	MLR (76,0)	MLR (48,9)	***ANN (109,5)***

Para establecer la dispersión de los pronósticos derivados por cada una de las metodologías, se calculó el promedio y la desviación estándar de los valores de

precipitación pronosticada para cada metodología. Lo mismo se hizo para la media del ensamble.

La Tabla 3.6 muestra la metodología que produjo el menor RMSE en cada caso. Cabe señalar que las técnicas con menor RMSE son aquellas que internamente todos sus modelos producen valores menos dispersos (GAM y ANN). Aquellos con mayor RMSE estaban más dispersos (MLR y SVR). Para el clúster 4, en octubre, ANN tiene un CV alto solo debido a que las otras técnicas no han tenido modelos que expliquen más del 50% de la varianza.

La Tabla 3.7 muestra el coeficiente de variación (CV) definido como la relación entre la desviación estándar y el promedio, expresado en porcentaje, promediado para el período de verificación 2009-2017. Cuanto menor es el CV, el promedio dentro de cada técnica, mejor representa el conjunto de modelos, ya que tiene menor dispersión.

Tabla 3.7. Metodología con el mejor coeficiente de variación (CV en %) de los valores de precipitación pronosticados, el mejor método para cada clúster en cursiva negrita.

Metodología (CV, %)	Mes 10	Mes 11	Mes 12	Mes 01	Mes 02	Mes 03
Clúster 1	***SVR (27,1)***	RLM (17,5)	ANN (15,6)	***SVR (17,2)***	GAM (22,5)	***SVR (26,5)***
Clúster 2	ANN (63,9)	***RLM (11,4)***	***RLM (20,9)***	***RLM (5,2)***	SVR (29,6)	GAM (26,6)
Clúster 3	***SVR (24,9)***	***SVR (16,6)***	RLM (22,0)	RLM (9,2)	***SVR (18,0)***	***SVR (46,0)***
Clúster 4	ANN (95,4)	***RLM (16,7)***	GAM (33,2)	SVR (34,2)	***RLM (29,6)***	GAM (24,8)

La Figura 3.7 muestra los CV correspondientes al conjunto de todos los modelos sin discriminar por metodologías utilizadas (media del ensamble). Se puede concluir que el CV es bastante bajo (y por lo tanto el promedio representa todo el conjunto de modelos) en todos los meses en los clusters 1 y 3 y en el clúster 4, excepto en el mes de octubre. En el clúster 2, en cambio, el CV es bajo sólo en los meses de diciembre, febrero y marzo.

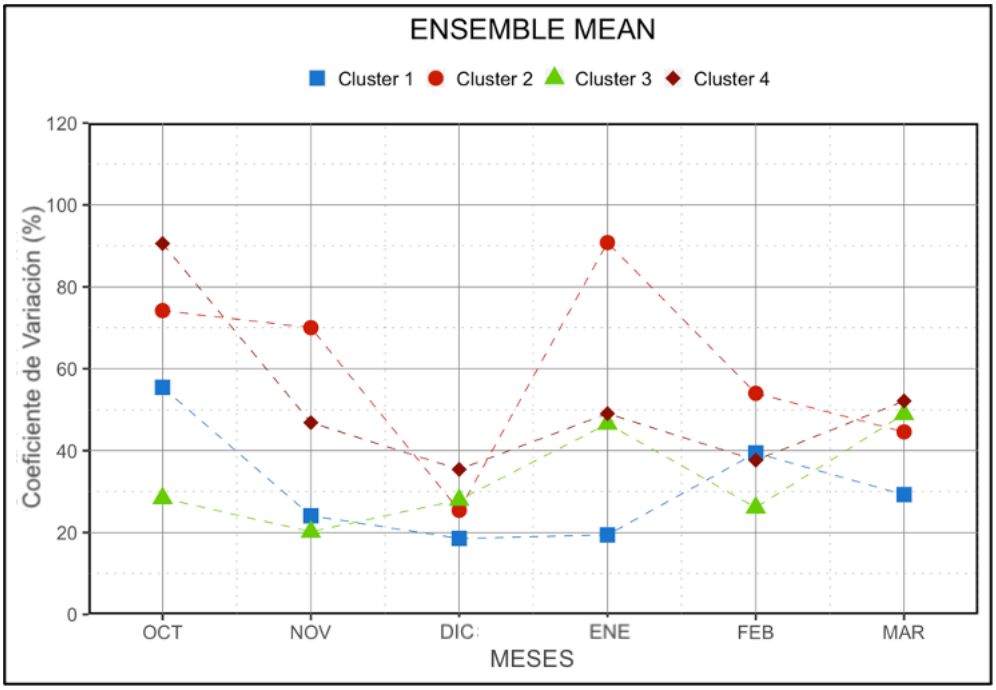


Figura 3.7. CVs correspondientes al conjunto de todos los modelos sin discriminar por metodologías utilizadas (ensamble medio).

El pronóstico categórico se realizó para cada temporada de pronóstico utilizando el conjunto de todos los modelos seleccionados (ensamble medio) en cada mes y cada clúster. La Figura. 3.8 muestra los valores IDX obtenidos. Recuerde que los valores negativos (positivos) indican una sobreestimación (subestimación) de la precipitación. Se puede observar una gran variabilidad respecto a los resultados en los diferentes años de pronóstico.

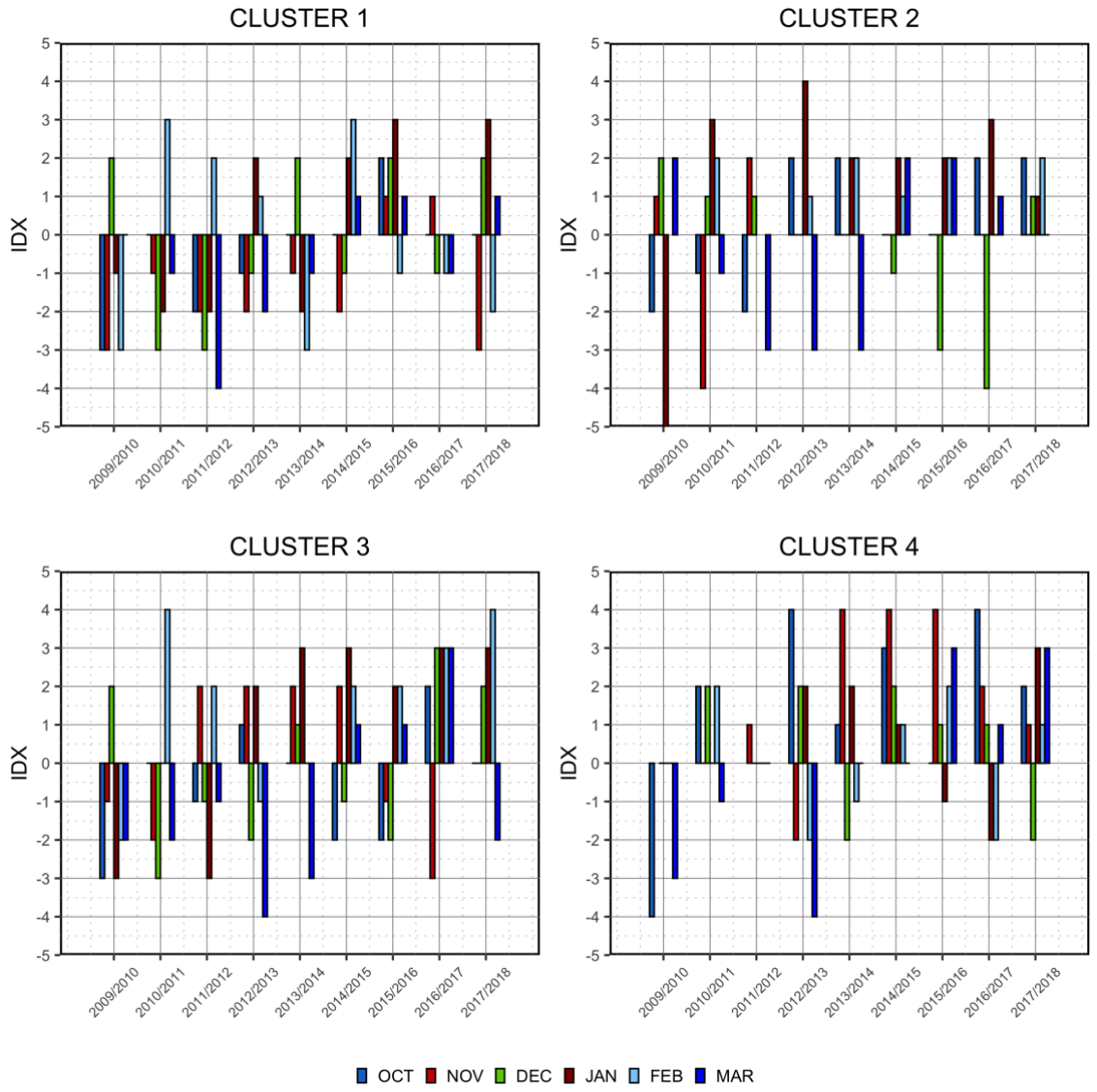


Figura 3.8. Índice IDX para cada temporada y mes de pronóstico.

Para resumir los resultados anteriores, se realizó un boxplot para la serie IDX de las estaciones de cada clúster (Figura 3.9). Estos gráficos permiten visualizar la distribución de los datos: el rango de los datos (línea vertical), el promedio (cruz), la mediana (línea horizontal dentro de la caja), el primer y tercer cuartil (límites de la caja) y el valores atípicos o fuera de rango (marcados con puntos fuera de los cuadros). Estos gráficos son especialmente útiles para realizar comparaciones de varias series y analizar su distribución.

En el caso del clúster 1, los valores medios de IDX están bastante cerca del valor cero ideal. El más alejado es el mes de noviembre con un valor de -1,3. La mayor dispersión se presenta en verano (dic, ene y feb). No se observan valores atípicos. En el clúster 2, los valores medios más alejados de 0 son los de enero (1,5) y febrero (1,4). La mayor

dispersión se observa en octubre y marzo. Se observaron valores atípicos en la temporada 2010-2011 (IDX = -4) y 2011-2012 (IDX = 2) y en enero 2009-2010 (IDX = -5). En el clúster 3, los valores IDX promedio más alejados corresponden a los meses de enero (1,1) y febrero (1,6). La mayor dispersión se observa entre diciembre y marzo. No se observan valores atípicos.

En el clúster 4, los meses de octubre y noviembre presentan el IDX promedio más alto con valores de 1,5 y 1,8 respectivamente. La mayor dispersión se observa en octubre, noviembre y marzo. Tampoco se observan valores atípicos.

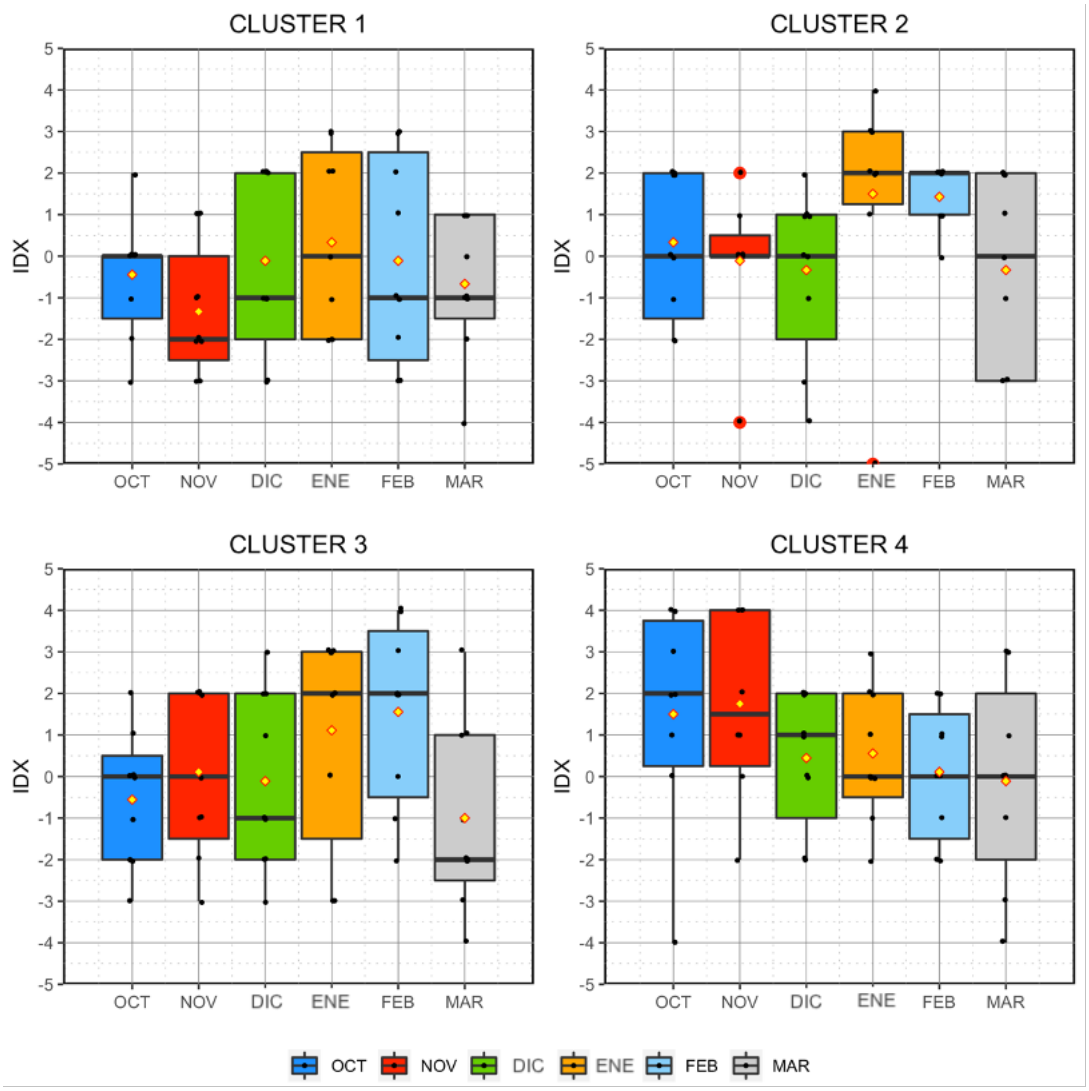


Figura 3.9.Boxplot para la serie IDX de las estaciones de cada clúster

La Tabla 3.8 muestra el IDX medio absoluto (sin considerar el signo) para cada clúster promediando todas las estaciones y todos los meses. Esto indica que al implementar un

pronóstico categórico y pronosticar el quintil más probable, el error está entre 1,5 y 1,9 quintil con un rango entre 1 y 1,3 quintil y la desviación estándar es igual o menor a 1,3.

Tabla 3.8. Media absoluta para cada clúster promediando todas las estaciones y todos los meses.

Clúster	Media IDX absoluta	Desviación estándar IDX absoluta
1	1.6	1.0
2	1.5	1.3
3	1.9	1.1
4	1.6	1.3

El porcentaje de error definido por la relación porcentual entre la precipitación prevista y la observada se calculó para cada clúster y mes de la temporada. El valor ideal es 100, valores mayores (menores) de 100 indican sobreestimaciones (subestimaciones). A partir de estos cálculos se consideró el conjunto de todos los modelos (media del ensamble) y se construyó un diagrama de caja para la serie del error porcentual para cada clúster (Figura 3.10).

En el clúster 1, las medias y medianas del porcentaje de error son similares en todos los meses, pero la dispersión es mayor en febrero y marzo. Hubo un gran pronóstico atípico en octubre de 2009 donde la media pronosticada fue de 12,9 mm cuando en realidad llovió 0,8 mm. Esto ocurre porque en el clúster 1 en octubre los registros de precipitación pueden ser muy bajos como fue el caso, lo que aumenta mucho el porcentaje de error. En el clúster 2 las medias son cercanas al 100%. La mayor dispersión se da en octubre y marzo. Los valores atípicos fueron en octubre de 2009 (se pronosticaron 42 mm y llovió 7,9 mm), en noviembre de 2010 (se pronosticaron 118,3 mm y llovió 31,8 mm) y en enero de 2009 (se pronosticaron 262,2 mm y llovió 86,1 mm). Cabe señalar que sólo 13 modelos cumplieron la condición para formar parte del ensamble. En el caso del clúster 3 se observa un buen desempeño. Los promedios se acercan al 100% y la mayor dispersión se registra en marzo y enero. El único caso atípico ocurrió en octubre de 2009 (se pronosticó 62,8 y llovió 21,3). En el clúster 4 las medias también están cerca del 100%. La mayor dispersión se registra en marzo. El único valor atípico fue en octubre de 2009 (se pronosticaron 213 mm y llovió 68,5 mm). En este caso, sólo 6 modelos componían el ensamble.

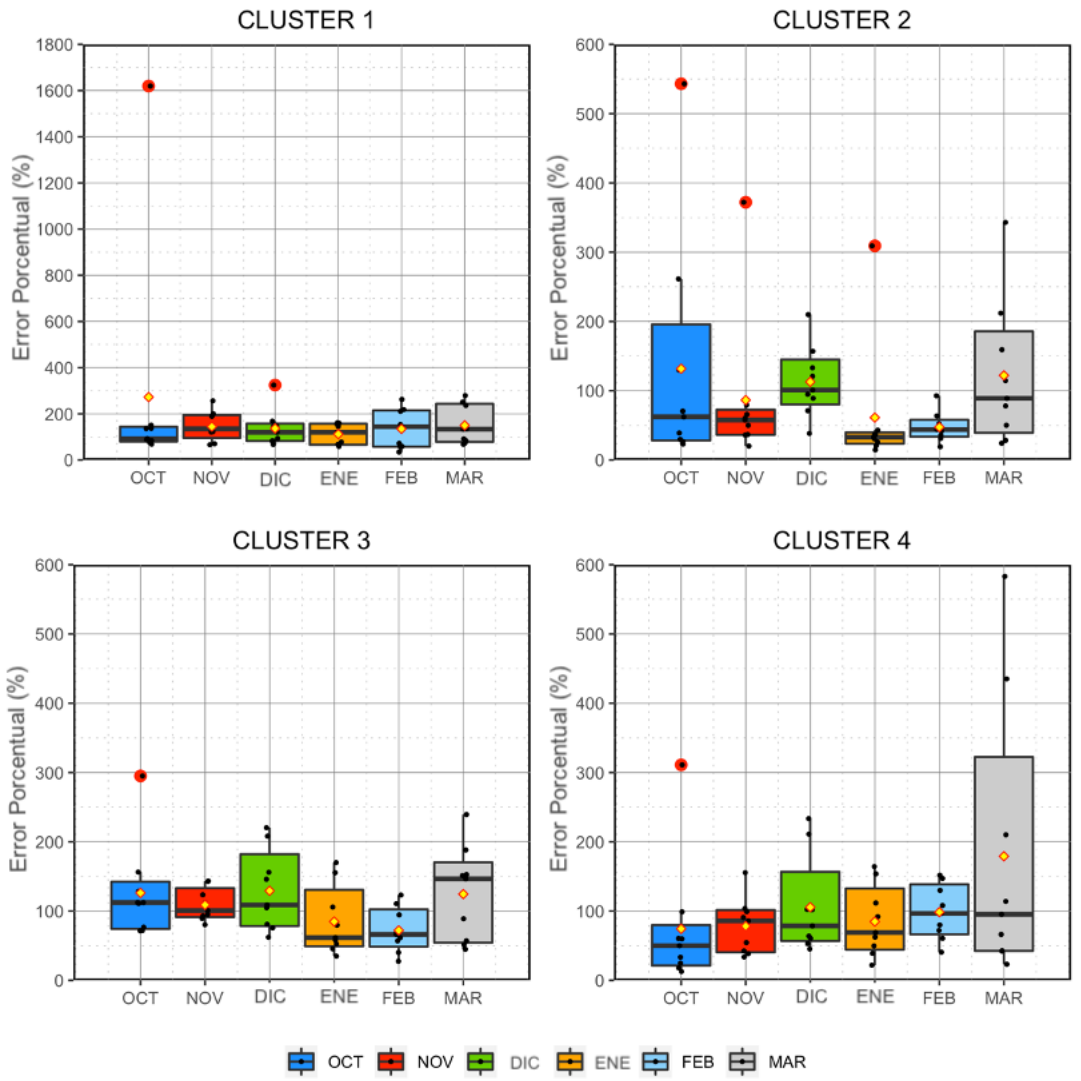


Figura 3.10.Error porcentual para cada clúster, mes.

3.1.2. Comparación con pronósticos derivados de centros mundiales

A continuación, se detalla el análisis de la comparación entre los resultados obtenidos con los pronósticos detallados en este trabajo y los modelos dinámicos de centros mundiales que se mencionaron en la sección 3.2. Para generar los valores de pronóstico mensual en cada clúster, se promediaron los pronósticos utilizando las máscaras de la

Figura 3.11. En el caso de JMA, las máscaras tienen una resolución de 2.5 x 2.5 grados y en el resto de los modelos es de 1x1 grado.

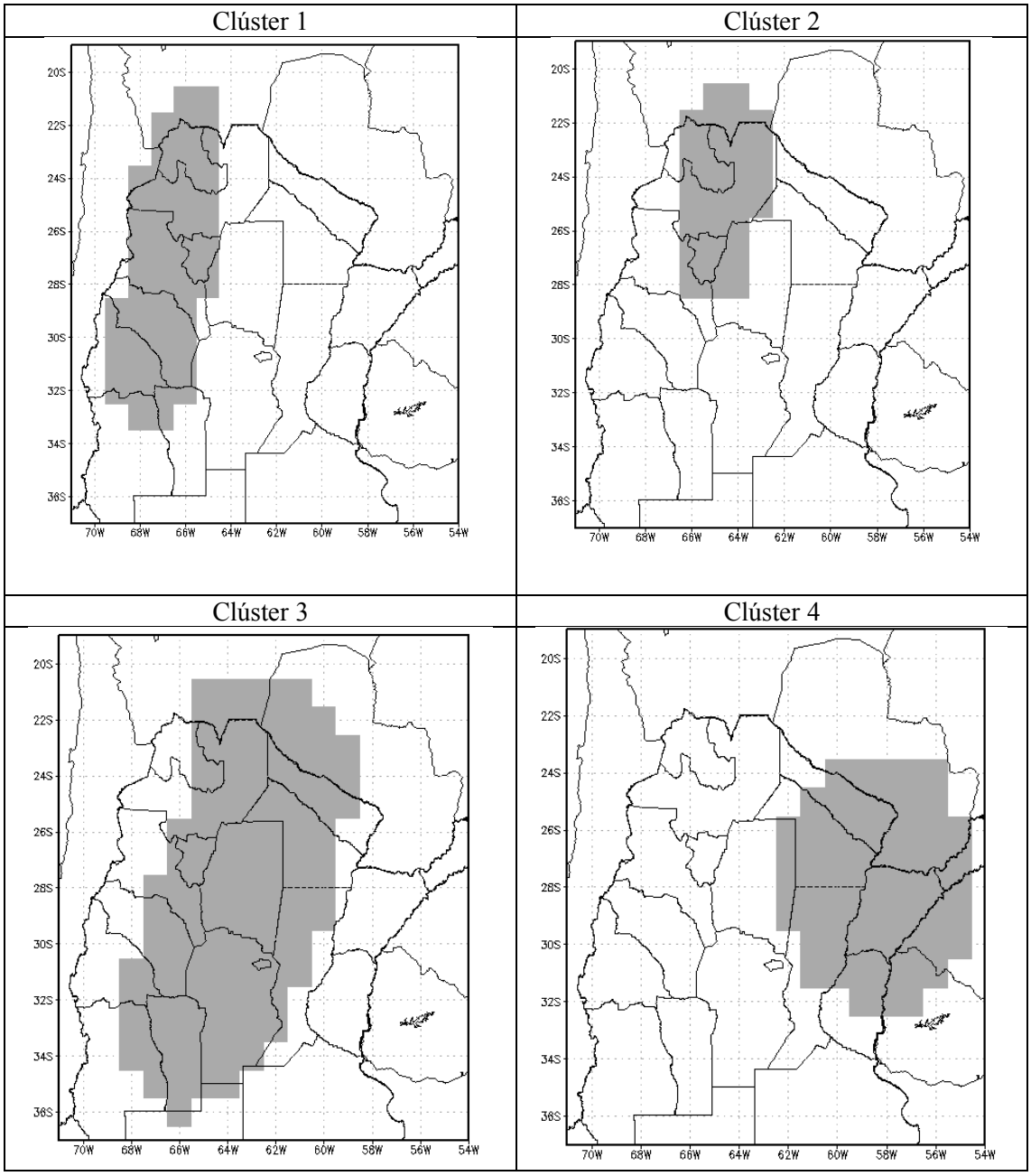


Figura 3.11.Máscaras utilizadas para derivar los pronósticos de los modelos dinámicos en cada clúster.

La Figura 3.12 muestra el RMSE derivado de los pronósticos del centro mundial detallados en la sección de metodología para los cuatro clusters y el valor correspondiente al ensamble de todos ellos. Se puede observar una gran variabilidad entre los diferentes modelos dinámicos con valores de RMSE más bajos en los clusters 3 y 4. Los modelos JMA y METEOFRANCE parecen funcionar peor en los clusters 1 y 2

que en los clusters 3 y 4, mientras que el modelo DWD es el mejor, especialmente en los clusters 1, 2 y 3.

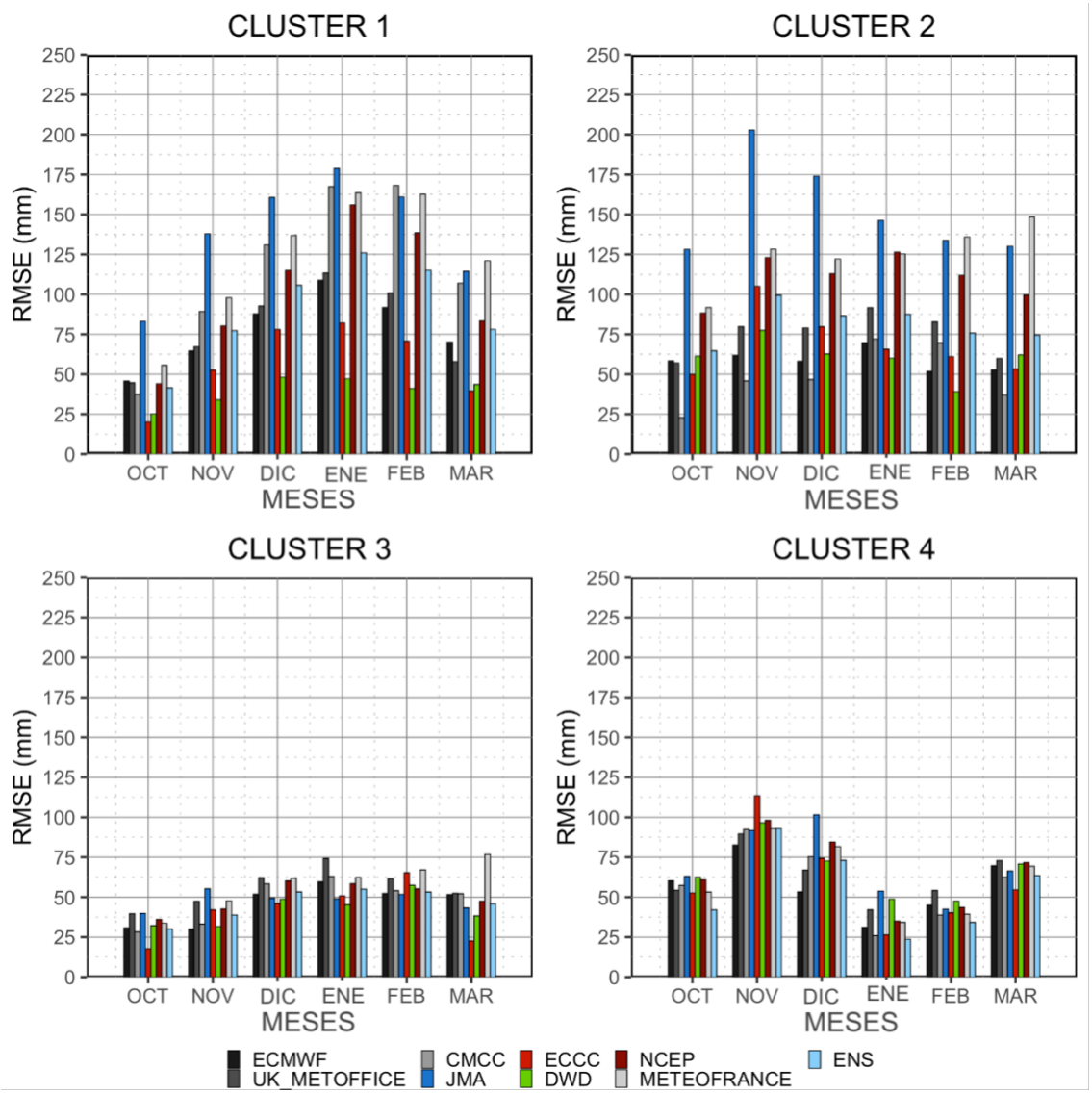


Figura 3.12. RMSE para los modelos dinámicos y la media del ensamble correspondiente.

La Figura 3.13 muestra el valor RMSESS del ensamble para cada clúster y para cada mes. Esto facilita la identificación de casos en los que la eficiencia del pronóstico derivado de las técnicas de aprendizaje automático supera a la de los modelos dinámicos de los centros globales. Se puede observar que las técnicas de aprendizaje automático han mejorado los pronósticos de los modelos dinámicos en el clúster 1 y en la primavera sobre el clúster 2. La habilidad es similar para esos métodos en el caso del clúster 3. Por

otro lado, los modelos dinámicos son más eficientes que la técnicas de aprendizaje automático en el caso del clúster 4.

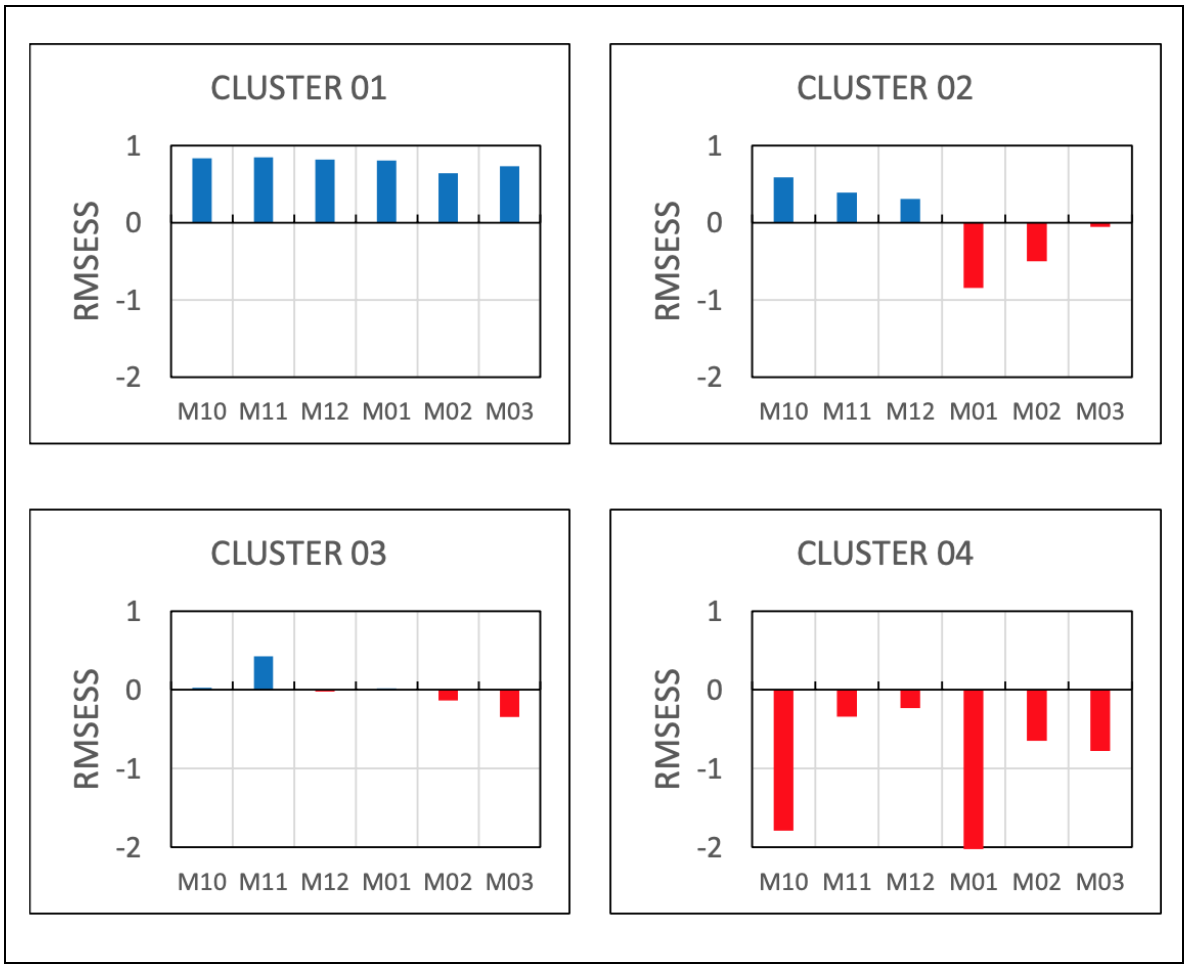


Figura 3.13. RMSESS se refirió al pronóstico con técnicas de aprendizaje automático para el ensamble en cada clúster para todos los meses (azul: modelos estadísticos, rojo: modelos dinámicos).

CAPÍTULO IV: PRONOSTICO PROBABILÍSTICO DE PRECIPITACIÓN ESTACIONAL EN LA REGIÓN DEL COMAHUE ARGENTINO

En esta sección se aplicará el FRAMEWORK que se ha detallado en el capítulo 2 para el pronóstico probabilístico de precipitación en la región del Comahue Argentino. El pronóstico probabilístico implica calcular la probabilidad de la precipitación estacional en alguna de las tres categorías definidas Subnormal , Normal, SobreNormal, basándose en los pronósticos derivados de las diferentes técnicas estadísticas aplicadas. Este pronóstico probabilístico se realiza para cada uno de los 12 trimestres del año y se ha verificado con estadísticos para tal fin.

4.1. Importancia del pronóstico en esta región.

La región de los Andes centrales en Argentina (Figura 4.1) comprende las cuencas de los ríos Neuquén (SRN) y Limay (SRL) al oeste y el río Negro (SRNe) al este. Toda la zona presenta un relieve que va cambiando desde las altas cumbres del oeste (alrededor de 1500 m de altura) hacia el océano Atlántico, atravesando el altiplano patagónico. Las precipitaciones se presentan principalmente en invierno en el oeste y disminuyen drásticamente hacia el este de las cuencas de los ríos Limay y Neuquén (Prohaska 1976, Paruelo et al. 1998; Castañeda y González 2008; Aravena y Luckman 2009; Garreaud et al. 2013; entre otros) . En cambio, la cuenca del Río Negro presenta precipitaciones escasas durante todo el año, aunque ligeramente superiores en verano. El área tiene una extensión aproximada de 140.000 km^2 y el 80% de la población vive en las inmediaciones de los ríos. La actividad frutícola-hortícola es la más importante de la cuenca del río Negro que se destaca por producir manzanas de muy alta calidad. La principal actividad en las otras dos cuencas (Limay y Neuquén) es la producción de energía hidroeléctrica, aportando alrededor del 25% de la energía eléctrica total de la Nación. Cinco represas están ubicadas en el río Limay (Alicurá, Piedra del Aguila, Pichi Picun Leufu, El Chocón y Arroyito) y cuatro en Neuquén (Complejo Cerros Colorados con las represas: Mari Menuco, El Chañar, Portezuelo Grande y Los Barreales). La precipitación invernal en la alta montaña del oeste es la principal responsable del aumento del caudal de los ríos (González et al. 2015). Además, se ha observado una disminución de la precipitación de 3 a 8 mm por año desde la década de 1950 y una reducción del caudal medio anual de hasta un 30% en los últimos 20 años (Saurral et al., 2017; UNCCF, 2015; IPCC, 2013; González y Vera 2010). González et al. (2021) mostraron que durante el período 2001-2016 se observaron tendencias negativas de precipitación en las subcuencas de los ríos Limay y Neuquén del orden de más de 35 mm/año, principalmente en invierno (más de 25 mm/año). Los autores demostraron que los cambios en las precipitaciones se correlacionan significativamente con la producción de energía hidroeléctrica en la región, lo que genera un riesgo socioeconómico importante para la región y para el país en general.

En este marco, es importante generar pronósticos probabilísticos estacionales de precipitación en la región. Muchos centros mundiales producen este tipo de pronóstico utilizando modelos deterministicos y estadísticos. En la región del Comahue se han desarrollado algunos modelos estadísticos con resultados alentadores. Por ejemplo, Romero et al. (2020) diseñaron modelos de pronóstico de precipitación anual utilizando técnicas de regresión lineal múltiple. Los resultados muestran que las temperaturas de la superficie del mar de los océanos Índico y Pacífico son buenos predictores para los modelos y explican el 42,8 % de la varianza de la precipitación. González (2015) derivó modelos estadísticos utilizando el método de pasos hacia adelante para predecir el índice de precipitación estándar (SPI) en la cuenca del río Neuquén que explicó el 42% de la varianza del SPI y retuvo dos predictores relacionados con la circulación sobre el océano Pacífico: uno de ellos muestra la relevancia por la intensidad del flujo zonal en latitudes medias, y el otro por la influencia de las bajas presiones cercanas a la cuenca del río Neuquén. González y Herrera (2014) derivaron diferentes modelos estadísticos: un modelo de promedio móvil integrado autorregresivo (ARIMA), Holt Winter (HW), Climate Prediction Tool (CPT) y un conjunto de todos, denominado multimodelo para predecir la lluvia invernal en la Patagonia. El principal resultado es que la consideración de ARIMA y HW en el conjunto multimodelo mejora la habilidad obtenida usando CPT.
Todos los modelos derivados tienen deficiencias que pueden deberse a varias razones: la selección de predictores, la metodología utilizada y la predictibilidad de la propia atmósfera en escalas estacionales (Coelho et al. 2005; Leetmaa 2003; Nobre et al. 2005).
En este capítulo de la tesis se presenta un método de pronóstico estadístico probabilístico para la precipitación estacional en las subcuencas de los ríos Limay, Neuquén y Negro, utilizando diversas técnicas estadísticas adicionadas para mejorar la destreza. La Sección 4.2 detalla los datos utilizados y los métodos aplicados. En la sección 4.3 se describen los resultados obtenidos y en la 5.3 las principales conclusiones.

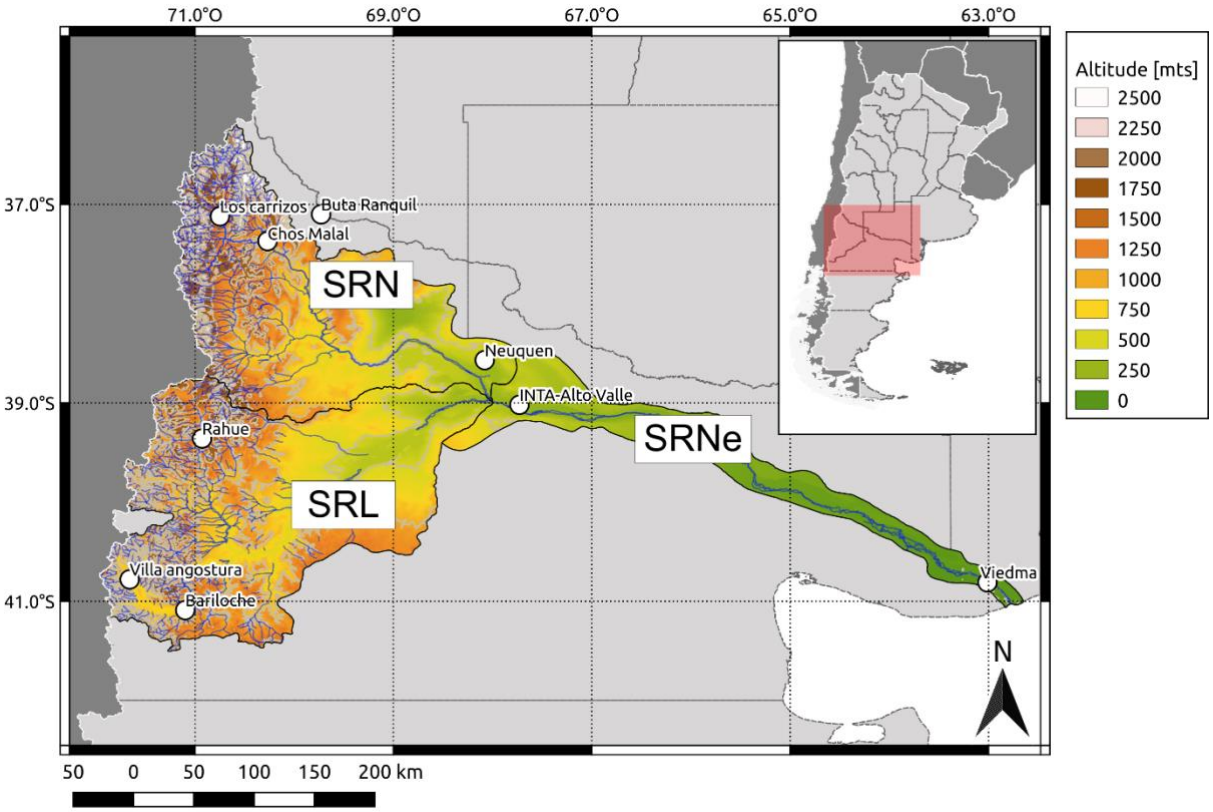


Figura 4.1.Región de estudio y estaciones utilizadas.

4.2. Metodología y Datos

Los límites de las subcuencas son los indicados en la Figura 4.1. Al noroeste se encuentra la subcuenca del río Neuquén (SRN), al suroeste la del Limay (SRL) y al este la del Negro (SRNe). Los círculos blancos son las posiciones de las estaciones de observación de precipitaciones.

Se consideraron series de precipitación acumulada trimestralmente para doce trimestres (de EFM: enero/febrero/marzo a DEF: diciembre/enero/febrero) del año en nueve estaciones de la región (Figura 4.1), tres de ellas ubicadas en SRL, cuatro en SRN y dos en SRNe. Los datos provienen de la red de medición del Servicio Meteorológico Nacional (SMN), el Instituto Nacional de Tecnología Agropecuaria (INTA) y la Autoridad Territorial de las cuencas de los ríos Limay, Neuquén y Negro (AIC).

La Figura 4.2 muestra la precipitación media y la variabilidad (1981-2020) en cada subcuenca. Es posible observar el predominio de la precipitación en invierno en las subcuencas de la región occidental, donde se ubican la mayoría de las represas, con valores de precipitación mayores en SRL que en SRN. La SRNe presenta valores muy inferiores a lo largo del año, aunque ligeramente superiores en verano. También existe una mayor variabilidad en los meses de mayor precipitación en cada subcuenca. La subcuenca del río Limay es la que presenta mayor variabilidad de todas.

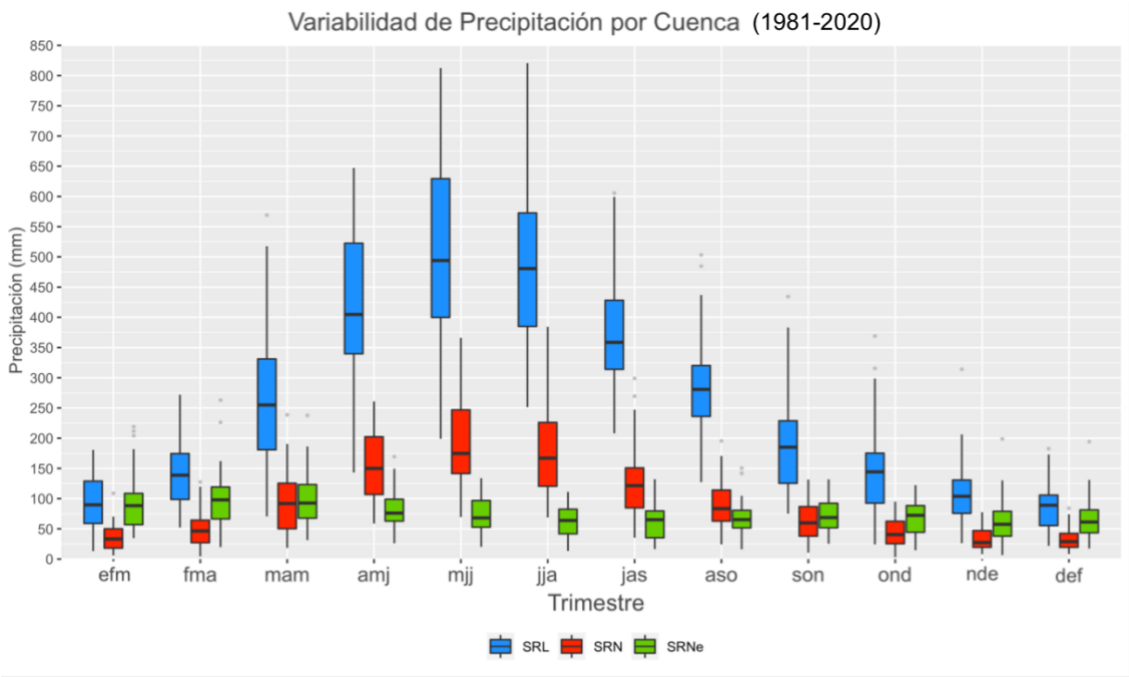


Figura 4.2. Diagrama de caja de la precipitación media promediada en SRL (azul), SRN (rojo) y SRNe (verde) (1981-2020).

La selección de las estaciones de observación de precipitación se realizó en base a la longitud de registro de datos completos en el período 1981-2020 y la posibilidad de que se distribuyan homogéneamente. Para cada subcuenca, se consideró la precipitación promedio de las estaciones ubicadas dentro de ella y estas series se utilizaron para entrenar los métodos de predicción.

Para justificar la representatividad de las estaciones en cada subcuenca, se calculó el valor absoluto del coeficiente de correlación de Pearson (r) entre la precipitación de las diferentes estaciones en cada subcuenca. Valores cercanos a 1 indican una asociación lineal, mientras que valores cercanos a 0 implican que no existe dependencia lineal entre

las estaciones. Posteriormente, se ajustó una curva de decaimiento exponencial de correlación con la distancia (d):

$$r(d) = ae^{bd}$$

La distancia crítica se consideró para r = 0.6, de esta forma para cada estación se define un radio crítico, y se espera que las estaciones dentro de la circunferencia sean linealmente dependientes de la estación centrada (Svoboda et al. 2014; Tokay et al. 2014 y Díaz et al. 2021). La Figura 4.3 muestra los resultados obtenidos, donde para el SRN se aprecia una cobertura total de la cuenca (círculos) indicando que estas estaciones son representativas de las lluvias registradas. En el caso de SRL se cubre casi toda la subcuenca, excepto en el este donde no existen estaciones meteorológicas con registros largos de precipitación. Para SRNe, el radio de las estaciones seleccionadas sólo cubre los tramos superior e inferior de la cuenca, quedando sin cobertura el tramo medio por la ausencia de estaciones de medición.

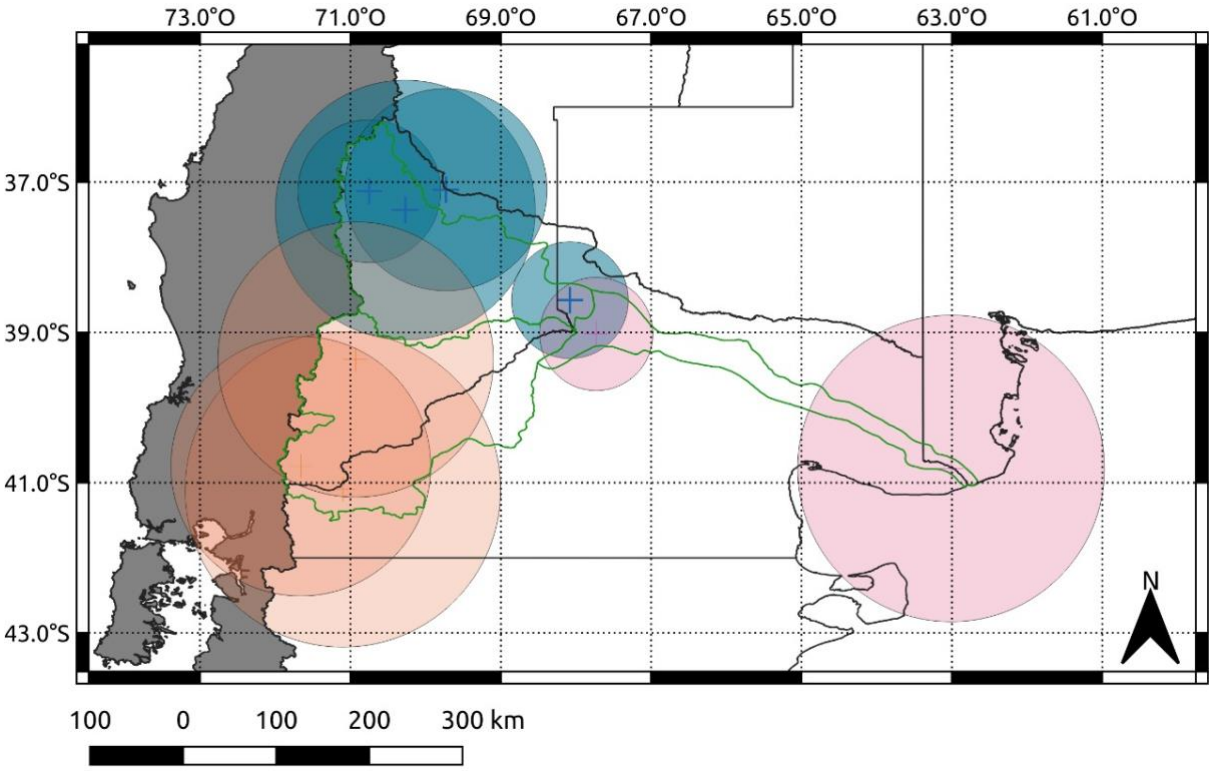


Figura 4.3.Representatividad de las estaciones seleccionadas para las subcuencas SRN (azul), SRL (rosa) y SRNe (morado claro).

Los forzantes de precipitación en la región del Comahue en diferentes temporadas han sido previamente estudiados (González y Vera 2010, González y Cariaga 2011, González 2015 y Romero et al. 2020). Se detectó una fuerte correlación entre las precipitaciones y la TSM en algunas áreas de los océanos Pacífico e Índico. Esto tiene relación con las teleconexiones del ENSO y las del Dipolo del océano Índico que han sido registradas por varios autores (González y Vera 2010, Garbarini et al. 2016, Garbarini et al. 2020). Las anomalías de la TSM en los océanos tropicales están relacionadas con el tren de ondas de Rossby que se extiende desde los océanos tropicales hacia las extratropicales, es decir hacia el sur y el este, como fue descrito por Mo (2000) y Kalnay et al (1986). Ingresan al sur de Argentina, donde la cordillera de los Andes es baja y los sistemas de precipitación asociados se desplazan luego hacia el noreste. La relación con la TSM en el océano Atlántico está asociada al ingreso de aire húmedo hacia Sudamérica a través de la alta subtropical semi-permanente del océano Atlántico (Garbarini et al. 2019). Por ello, se han definido predictores de precipitación como la TSM en algunas zonas de los océanos tropicales y subtropicales así como de los campos de circulación (altura geopotencial) que describen el desplazamiento de los sistemas.

El flujo regional en niveles bajos también tiene influencia sobre la precipitación, ya que la intensidad de los vientos del oeste en el océano Pacífico Central y en las latitudes medias y la componente del viento meridional, aumentan la advección de humedad desde el Pacífico y desde el norte, respectivamente. Obviamente, la cantidad de agua disponible sobre la región de estudio también colabora con la precipitación total, por lo que el TCW también se consideró como variable de entrada para el análisis.

Por esta razón, todas estas variables fueron consideradas para definir los predictores utilizando las áreas que tienen una correlación significativa con la precipitación observada que ocurre el siguiente trimestre.

Los predictores se definieron como el valor medio de la variable en todos los puntos de la grilla sobre las áreas con correlación estadísticamente significativa (superior a 0,37) (95% de confianza usando una prueba normal) y que cubren un área suficientemente grande (mínimo 1,25 10^6 km^2), por lo tanto, sólo se utilizaron áreas que tenían al menos

20 puntos de cuadrícula contiguos. Con el objetivo de que el pronóstico probabilístico sea robusto, se debe considerar un número significativo de miembros/modelos. Esto implica que el conjunto multimodelo que tiene el mayor número posible de miembros y la menor dispersión entre ellos, dará mayor certeza. Por eso es conveniente utilizar muchos modelos derivados de cada técnica estadística. Además, cada uno de estos miembros utiliza diferentes predictores y esto asegura la representatividad del mayor número de situaciones posibles en el pasado, lo que permite que los métodos "aprendan" a estimar las situaciones futuras.

El período de entrenamiento utilizado fue inicialmente 1981-2010 para construir los modelos estadísticos. La verificación se realizó para el período 2011-2020, con la particularidad que para el pronóstico del año n, los modelos fueron entrenados hasta el año (n-1) (Figura 4.4).

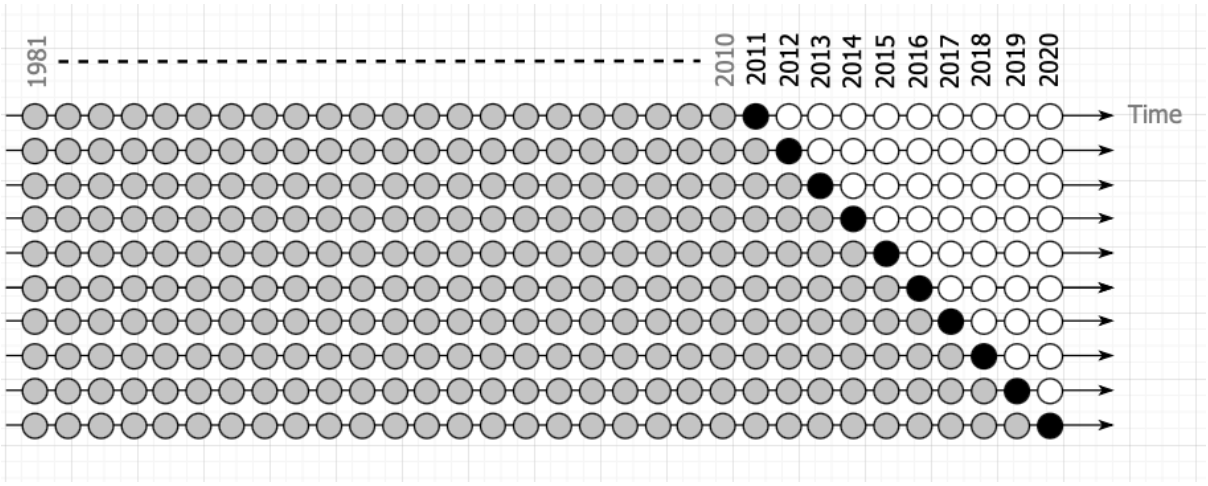


Figura 4.4. Esquema de los periodos de entrenamiento y verificación.

Se construyó una distribución probabilística de la precipitación acumulada estacional para cada región y trimestre utilizando los cinco intervalos de clase definidos por los quintiles de precipitación observada. Se pudo predecir el intervalo con mayor probabilidad de ocurrencia y compararlo con el observado. Se calculó el índice IDX definido como en el caso del Chaco argentino mostrado en la sección anterior.

Para evaluar el pronóstico probabilístico, se calcularon diagramas de confiabilidad y Brier score (Kumar et al., 2020), utilizando tres categorías: Inferior a la Normalidad (SubN) cuando el valor de precipitación es inferior al primer tercil, Superior a la Normalidad (SobN) cuando es superior al segundo tercil y Normal (NOR) cuando se encuentra entre el primer y segundo tercil.

Los diagramas de confiabilidad se basan en un diagnóstico de pronósticos probabilísticos para un conjunto predefinido de eventos, por lo que pueden construirse para cada una de las categorías (SubN, NOR y SobB).

Los diagramas de confiabilidad por categorías son útiles para indicar si la calidad de los pronósticos es buena.

La frecuencia relativa observada del evento se plotea versus la frecuencia relativa pronosticada.

La frecuencia relativa pronosticada es el número de pronósticos en cada intervalo dividido por el número total de pronósticos, la frecuencia relativa observada se calcula como el número de eventos dividido por el número de pronósticos.

Para cada valor discreto de la probabilidad de pronóstico, el diagrama de confiabilidad indica si el evento de pronóstico ocurrió con la frecuencia esperada.

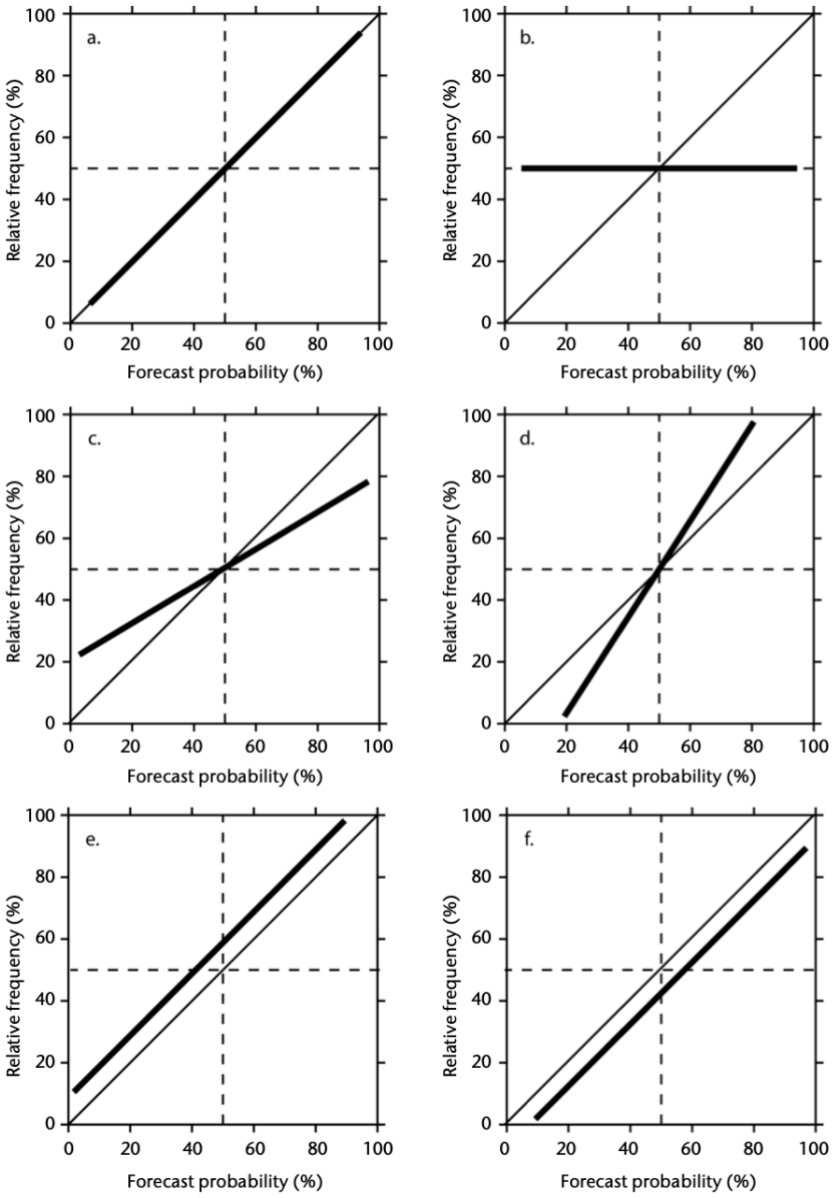


Figura 4.5. Ejemplos idealizados de curvas de confiabilidad que muestran sistemas de pronóstico con (a) confiabilidad perfecta, (b) sin resolución, (c) exceso de confianza, (d) confianza insuficiente, (e) pronóstico insuficiente, (f) pronóstico excesivo.

La interpretación de los diagramas de confiabilidad puede facilitarse al considerar algunos ejemplos idealizados, como se muestra en la Figura 4.5 (figura extraída del manual de WMO "Guidance on Verification of Operational Seasonal Climate Forecasts",2018). Si los pronósticos son perfectamente confiables, la frecuencia relativa observada será igual a la probabilidad del pronóstico para todos los valores de esa probabilidad y, por lo tanto, la curva de confiabilidad estará a lo largo de la diagonal de 45° (Figura 4.5a). La figura 4.5b ilustra el caso en que el evento ocurre con la misma frecuencia relativa independientemente de los pronósticos, por lo que los pronósticos

no tienen resolución y son inútiles. Más típicamente, los pronósticos tienen cierta resolución, pero no son perfectamente confiables y con frecuencia mostrarán un exceso de confianza (Figura 4.5c): el pronóstico sobreestima (subestima) las probabilidades observadas, para probabilidades mayores (menores) al 50%. La Figura 4.5d indica la situación inversa. Cuanto mayor sea el grado de exceso de confianza, menor será la pendiente de la curva. Si los pronósticos son poco confiables, la curva es más pronunciada que la diagonal. La Figura 4.5e ilustra el caso de las probabilidades de pronóstico que son consistentemente más bajas que las frecuencias relativas observadas. Lo que indica que el evento siempre ocurre con más frecuencia de lo anticipado y, por lo tanto, está por debajo del pronóstico. La Figura 4.5f ilustra el caso cuando lo contrario es cierto y el evento ocurre con menos frecuencia de lo anticipado y está sobre pronosticado.

En este capítulo de la tesis se construyeron los diagramas de confiabilidad para cada cuenca (Ríos Negro, Limay y Neuquén) y para dos periodos: Verano (de OND a MAM) e Invierno (de AMJ a SON).

El Brier Score (BS) mide la precisión de los pronósticos probabilísticos. Cuanto menor sea la puntuación de Brier para un conjunto de predicciones, mejor es la predicción. Se define como el cuadrado de la mayor diferencia posible entre la probabilidad predicha y el resultado real:

$$BS = \frac{1}{n}\sum_{t=1}^{N}(f_t - o_t)^2$$

donde f_t es la probabilidad pronosticada, o_t es el resultado real y n es el número de instancias de pronóstico.

El Brier Skill Score (BSS) mide si las predicciones son simplemente tan buenas como las que se toman como referencia (BSref) y se define como:

$$BSS = 1 - BS/BSref$$

en el caso de que BSref fueron los valores climatológicos:

$$BSref = \frac{1}{N}\sum(ym - yi)$$

donde y_m es el promedio de y_i y y_i es la precipitación observada. Los valores negativos (positivos) de BSS indican que el pronóstico es menos (más) preciso que el valor climatológico.

4.3. Resultados y discusión

Para generar un pronóstico probabilístico, es deseable tener una gran cantidad de modelos. Así, para cada período de entrenamiento y para cada trimestre del año, los modelos fueron entrenados con todas las metodologías: RLM, GAM, SVR y ANN. Sólo se retuvieron aquellos modelos que explicaban más del 50% de la varianza de la precipitación trimestral. Por lo tanto, el modelo múltiple incluye no solo el "mejor" modelo para cada método, sino también varios otros modelos, con un número variable de predictores, que también cumplen el criterio. En el caso particular de RLM, la técnica se aplicó al total de los predictores independientes que corresponden a cada subcuenca y trimestre, y a un número menor de ellos por lo que se generaron muchos modelos. En la Tabla 4.1 se detalla el número de modelos seleccionados en cada caso. Las cuatro arquitecturas ANN generaron siempre modelos que explican al menos el 50% de la varianza y la técnica SVR es la que proporciona un mayor número de modelos hábiles.

Tabla 4.1. Número de modelos que explican más del 50% de la varianza de la precipitación derivada del uso de todas las metodologías.

Trimestre	Periodo de entrenamiento	Año Pronosticado	subcuenca	Número de modelos				
				Total	ANN	GAM	RLM	SVR
EFM a DEF	1981-2010	2011	SRL	442	48	87	110	197
EFM a DEF	1981-2010	2011	SRN	250	48	48	52	102
EFM a DEF	1981-2010	2011	SRNe	339	48	72	80	139
EFM a DEF	1981-2011	2012	SRL	427	48	82	106	191
EFM a DEF	1981-2011	2012	SRN	263	48	51	55	109
EFM a DEF	1981-2011	2012	SRNe	343	48	78	75	142
EFM a DEF	1981-2012	2013	SRL	434	48	80	104	202
EFM a DEF	1981-2012	2013	SRN	239	48	44	44	103
EFM a DEF	1981-2012	2013	SRNe	355	48	79	78	150
EFM a DEF	1981-2013	2014	SRL	424	48	81	98	197
EFM a DEF	1981-2013	2014	SRN	231	48	42	40	101

EFM a DEF	1981-2013	2014	SRNe	349	48	81	74	146
EFM a DEF	1981-2014	2015	SRL	402	48	71	91	192
EFM a DEF	1981-2014	2015	SRN	227	48	41	40	98
EFM a DEF	1981-2014	2015	SRNe	354	48	82	76	148
EFM a DEF	1981-2015	2016	SRL	391	48	63	94	186
EFM a DEF	1981-2015	2016	SRN	221	48	33	37	103
EFM a DEF	1981-2015	2016	SRNe	346	48	77	75	146
EFM a DEF	1981-2016	2017	SRL	394	48	68	84	194
EFM a DEF	1981-2016	2017	SRN	209	48	29	38	94
EFM a DEF	1981-2016	2017	SRNe	313	48	66	55	144
EFM a DEF	1981-2017	2018	SRL	368	48	57	79	184
EFM a DEF	1981-2017	2018	SRN	203	48	28	30	97
EFM a DEF	1981-2017	2018	SRNe	301	48	61	57	135
EFM a DEF	1981-2018	2019	SRL	351	48	56	72	175
EFM a DEF	1981-2018	2019	SRN	198	48	21	28	101
EFM a DEF	1981-2018	2019	SRNe	277	48	51	49	129
EFM a DEF	1981-2019	2020	SRL	304	48	36	53	167
EFM a DEF	1981-2019	2020	SRN	160	47	7	15	91
EFM a DEF	1981-2019	2020	SRNe	243	48	37	34	124

La Figura 4.6 muestra los diagramas de dispersión entre la precipitación pronosticada y la observada cuando se usa la media del ensamble en cada trimestre para el período de verificación (2011-2020). Se puede notar que los valores están muy cerca de la línea de identidad indicando un buen desempeño, especialmente cuando los valores de precipitación son medios y altos. Para valores de precipitación bajos como los que se dan en verano en SRN y SRL, la dispersión es mayor. No hay mayor diferencia en la habilidad de los resultados en cada cuenca.

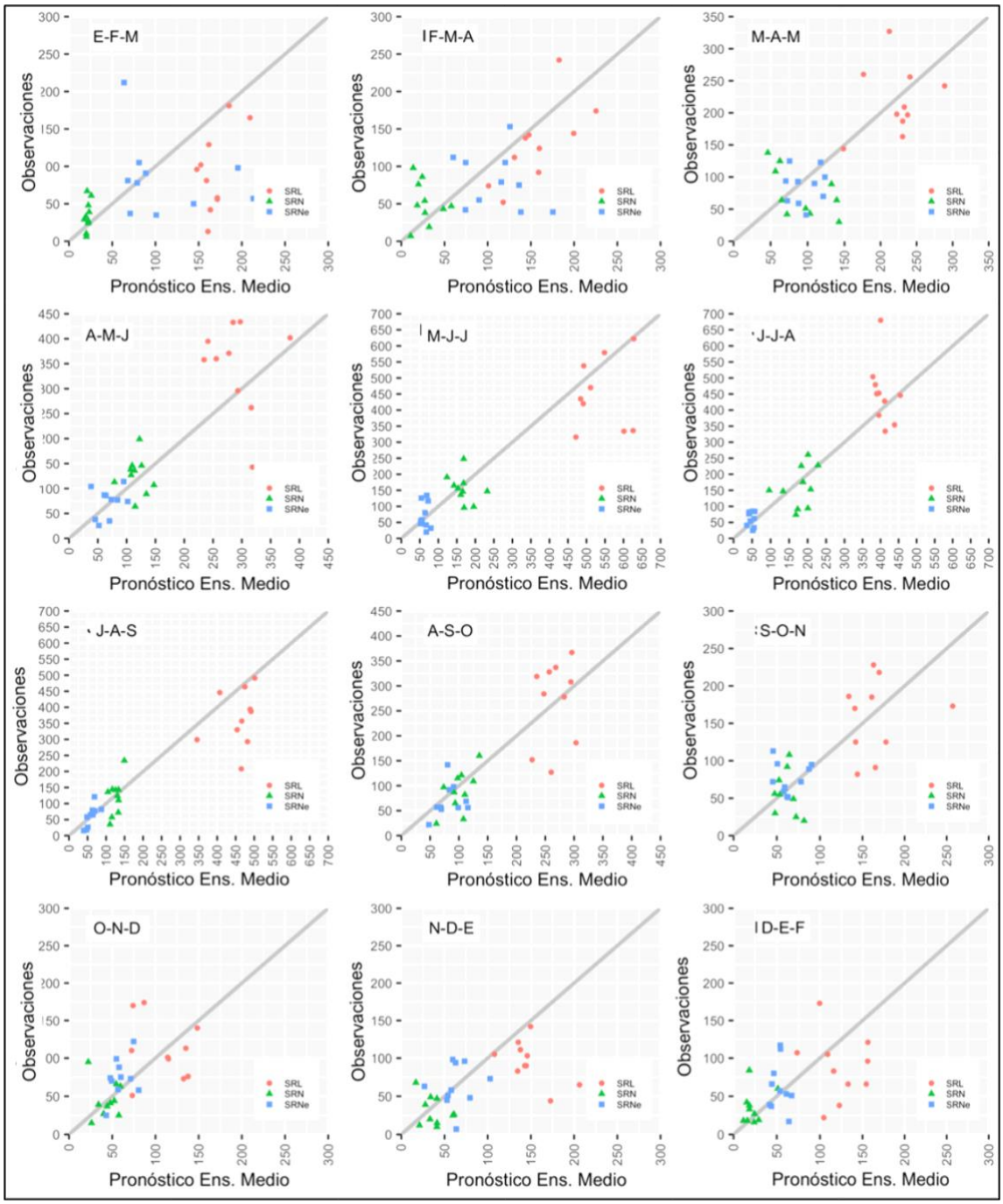


Figura 4.6. Diagramas de dispersión que muestran la precipitación observada frente a la pronosticada, utilizando la media del ensamble para el período de verificación.

La generación de varios modelos nos permite evaluar el pronóstico de forma probabilística. A modo de ejemplo, se muestran los resultados obtenidos para el pronóstico de precipitación para el trimestre JAS en 2017. Los modelos se generaron utilizando el período de entrenamiento 1981-2016. La Tabla 4.2 muestra el número total de modelos que se encontraron que explican más del 50% de la varianza de la

precipitación y que, por lo tanto, se incluyeron en la media del ensamble. También se muestran: el valor de precipitación pronosticado por el ensamble (promedio de todos los modelos) y la precipitación observada. Se registró una sobreestimación en la precipitación de la cuenca del Limay, una subestimación en las cuencas del Negro y del Neuquén.

A partir de los quintiles de lluvia observados en el período de entrenamiento se calcularon los valores de Io e Ip. La Figura 4.7 muestra las probabilidades asignadas a cada intervalo de quintiles en cada cuenca. Los colores indican la categoría predicha: muy subnormal (menos del primer quintil, marrón), subnormal (entre el primer y segundo quintil, marrón claro), normal (entre el segundo y tercer quintil, gris), supernormal (entre el tercero y el cuarto quintil, verde claro) y muy supernormal (mayor que el cuarto quintil, verde). La categoría más probable fue pronosticada correctamente en el río Neuquén (IDX nulo en la Tabla 4.2). IDc se define como el valor de IDX para valores climatológicos, cuando Ip es 3. Si se hubiera utilizado la climatología para pronosticar (Ip =3), resultaría un IDc como se muestra en la Tabla 4.2. En este caso, se puede concluir que sólo en el caso de la cuenca del río Neuquén, el pronóstico es mejor que la climatología. Las curvas de la Figura 4.8 muestran la probabilidad de que la precipitación pronosticada (eje Y) exceda el umbral (eje X) definido como los quintiles de precipitación.

Tabla 4.2. Número de modelos que explican más del 50% de la varianza de las precipitaciones en JAS 2017.

JAS 2017	Número total de modelos	Precip. observada	Precip. media pronosticada	Desv. Est. de la precipitación. pronosticada	Io	Ip	IDX	IDc
SRL	91	357	466	77	3	5	-2	0
SRN	81	142	126	39	4	4	0	1
SRNe	49	78	63	15	4	2	2	1

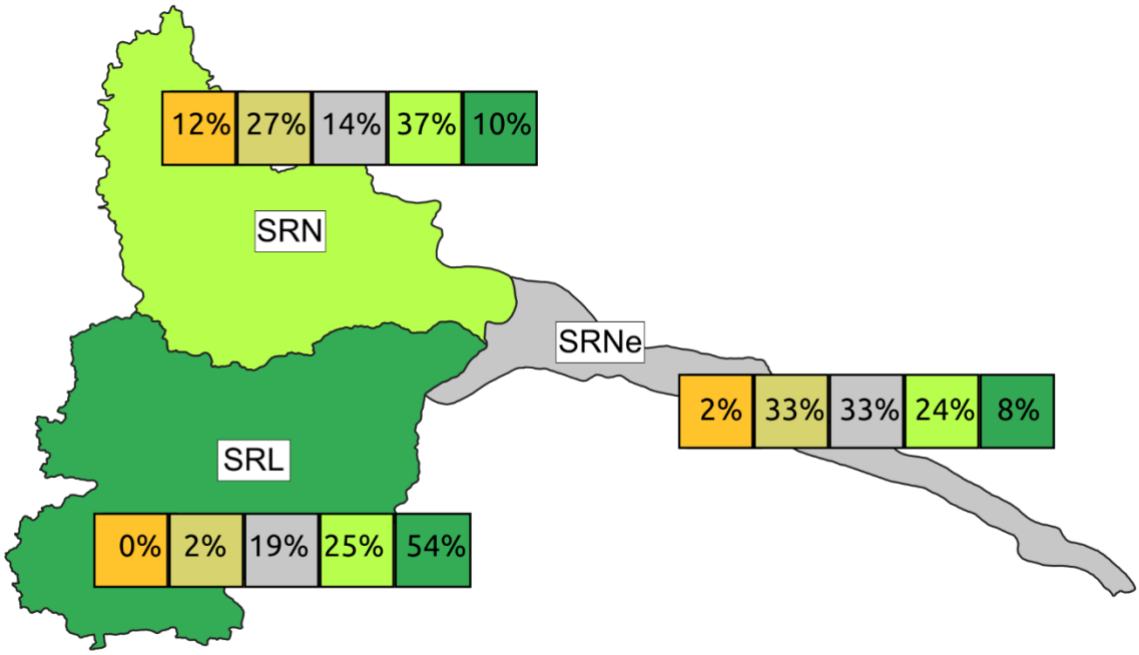


Figura 4.7. Pronóstico probabilístico de precipitación JAS 2017 utilizando las técnicas descritas.

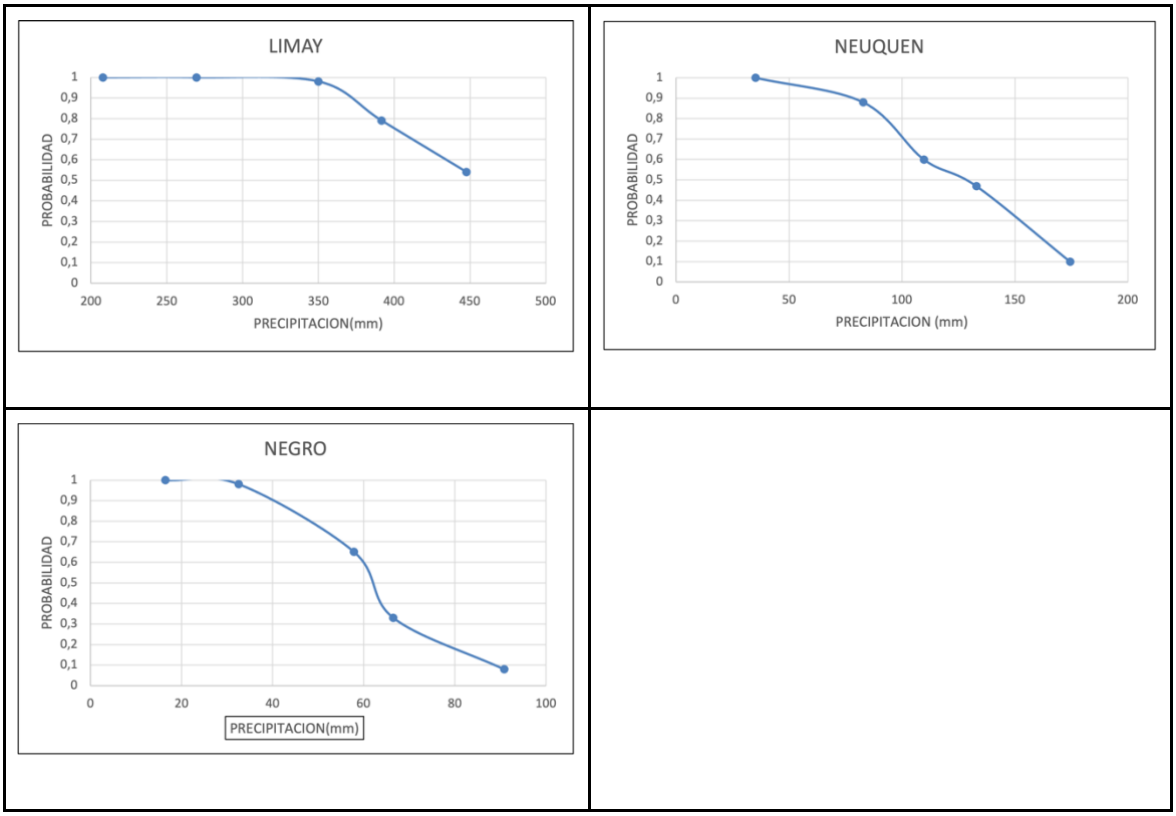
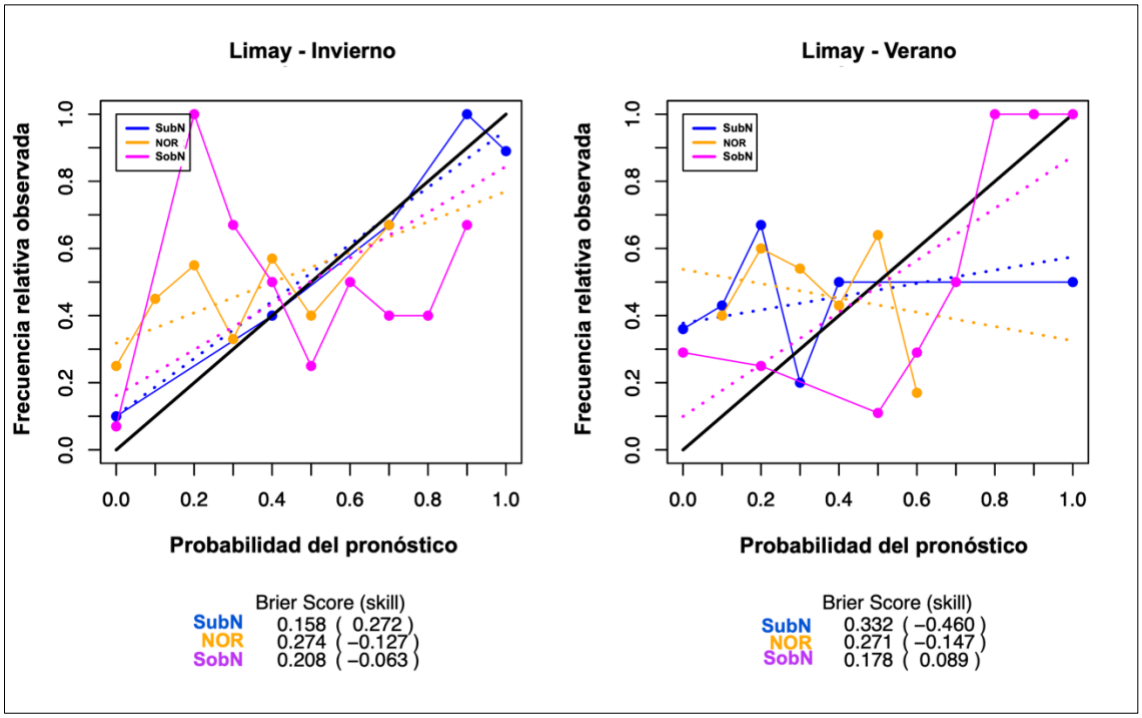


Figura 4.8. Probabilidad de que la precipitación pronosticada (eje Y) supere cierto umbral (eje X) en JAS 2017 en SRL (panel superior), SRN (panel central) y SRNe (panel inferior).

Este procedimiento se ha realizado para todos los trimestres del periodo de verificación 2011-2020. La confiabilidad del pronóstico de probabilidad se evaluó mediante la construcción de los diagramas de confiabilidad para las categorías SubN (SubNormal), NOR (Normal) y SobN (SobreNormal), como se detalla en la sección 4.2 de metodología.

La Figura 4.9 muestra los diagramas de confiabilidad correspondientes para río Negro (paneles superiores), río Neuquén (paneles medios) y río Limay (paneles inferiores) y para Invierno (paneles izquierdos) y Verano (paneles derechos). La línea de regresión ponderada para cada categoría se traza en el gráfico. Los valores de BS y BSS se detallan a continuación para cada gráfico. El pronóstico de verano es peor que el pronóstico de invierno en todas las cuencas, probablemente debido a la baja cantidad de precipitaciones de verano. Los pronósticos tienden a sobreestimar la ocurrencia de valores reales cuando la probabilidad es alta y los subestiman cuando

la probabilidad es baja. Los pronósticos son mejores en la cuenca del río Limay y Neuquén en invierno que en la cuenca del río Negro. Para el caso del invierno cuando la confiabilidad es mayor, BS presenta mejores valores para las cuencas Limay y Neuquén en las categorías SubN y SobN. BSS se calcula utilizando pronósticos de referencia con probabilidades climatológicas (1/3) y se detallan debajo de los gráficos en la Figura 4.9. Aunque el BSS tiene un máximo de 1, en la práctica, el BSS suele ser mucho menor que 1 para previsiones estacionales. Los valores de BSS indican que el pronóstico probabilístico es mejor que la climatología en invierno, en SRL y SRN para las categorías SubN y SobN. Este hecho es importante ya que las centrales hidroeléctricas son operadas en base a los caudales de los ríos en estas cuencas, que aumentan especialmente con las lluvias invernales (González et al. 2015). SRNe presenta valores de confiabilidad similares para todas las categorías y parece tener menos habilidad que las otras dos cuencas.



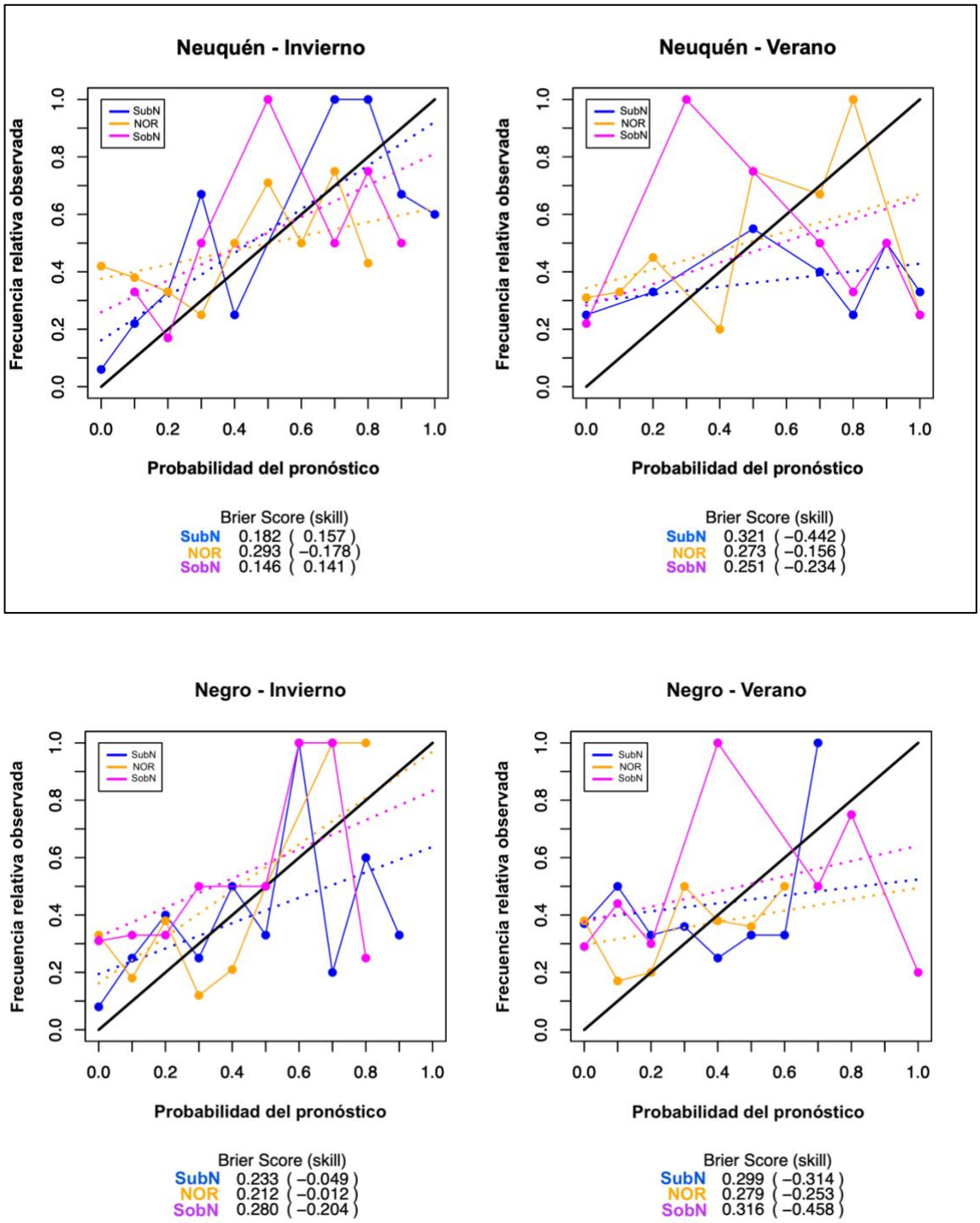


Figura 4.9. Diagramas de confiabilidad para río Negro (paneles superiores), río Neuquén (paneles centrales) y río Limay (paneles inferiores) y para Invierno (paneles izquierdos) y Verano (paneles derechos). La BS y sus tres componentes se detallan debajo de cada gráfico y se trazan las líneas de regresión.

CAPÍTULO V: CONCLUSIONES

5.1. Conclusiones del framework.

En este trabajo se desarrolló un framework que permite implementar pronósticos probabilísticos en cualquier región de Argentina, donde se disponga de registros largos de observaciones y conocimiento de posibles forzantes que influyan en la región. Este pronóstico es realizado estadísticamente utilizando un ensamble de modelos estadísticos con las técnicas de Regresión Lineal Múltiple, Modelos Aditivos Generalizados, Regresión de Soporte Vectorial y Redes Neuronales. Sólo forman parte del ensamble los modelos que explican más del 50% de la varianza de la precipitación. Los modelos derivados a partir de este framework se utilizaron para aplicarlos a 2 casos particulares: por un lado, derivar un pronóstico determinístico de precipitación mensual en la zona del Chaco argentino y por otro un pronóstico probabilístico de precipitación estacional en las cuencas del Comahue argentino.

Actualmente, este framework se usa operativamente para el pronóstico de precipitación experimental en el Comahue realizado por el grupo de "Perspectiva Climática" del Departamento de Ciencia de la Atmósfera y los Océanos (http://perspectiva.at.fcen.uba.ar/). Este grupo estudia la posibilidad de realizar pronósticos estadísticos en diferentes áreas de Argentina que puedan ser utilizados para distintos propósitos, como la generación de energía hidroeléctrica y la prevención de sequías e inundaciones, entre otros. Además, se enfoca en la interacción entre las dinámicas climática y social asociadas a eventos climáticos, relacionándose con instituciones nacionales tomadoras de decisiones.

Los resultados obtenidos son alentadores y animan a seguir trabajando para mejorar los resultados obtenidos. Dichas mejoras pueden probarse a partir de utilizar predictores con una mayor resolución, implementar técnicas de redes neuronales más complejas y aumentar el número de técnicas estadísticas utilizadas. Aún más, la comparación con modelos disponibles de centros mundiales mostró que en algunos casos los mejoran. Esta metodología es especialmente útil en áreas limitadas y donde se conocen los forzantes climáticos de precipitación, derivados de estudios previos.

Para finalizar, mostraremos un resumen de las conclusiones obtenidas para cada uno de los casos en que se aplicó el framework.

El repositorio del framework está disponible en:

https://github.com/alrolla/PronProbaEst

5.2. Conclusiones Región del Gran Chaco Argentino.

En términos generales, el pronóstico de precipitación mensual en los meses estivales en la región del Gran Chaco Argentino mostró que la consideración de procesos no lineales mejora los resultados, dado que GAM y ANN muestran un error cuadrático medio menor que RLM y SVR. Además, GAM parece funcionar mejor en el clúster 1 donde las precipitaciones son más escasas, especialmente en primavera. Sin embargo, estas técnicas muestran una gran dispersión interna, cuando se consideran todos los modelos derivados de cada una de ellas. Si se considera la media del ensamble de cada técnica, se puede ver que el coeficiente de variación es bajo para todos los clusters excepto el clúster 2, lo que indica que la media representa bien el conjunto de modelos. El pronóstico por ensambles de modelos múltiples mejorará los pronósticos en la región en comparación con un enfoque basado en un modelo único.

Se implementó un pronóstico categórico teniendo en cuenta el ensamble total de modelos. Esto permitió pronosticar el intervalo de precipitación, definido con los quintiles de precipitación, con mayor probabilidad de ocurrencia. Para ello se definió el índice IDX. En promedio, IDX mostró un valor de 1,5 , lo que indica que, en general, el intervalo más probable se predice con un intervalo y medio de error aproximadamente (30%). El pronóstico categórico trata de cuantificar la incertidumbre en una predicción, lo que puede ser un ingrediente esencial para una óptima toma de decisiones.

Se calculó el porcentaje de error para cada clúster y mes. Se observó que el error medio del ensamble aumenta cuando la precipitación es escasa o cuando no hay un gran número de modelos para generar la media del ensamble.

En resumen, la habilidad de las diferentes metodologías depende mucho del mes y la región. Por otro lado, se recomienda el uso de ensambles medios formados por modelos estadísticos de pronóstico derivados de diferentes metodologías. Los pronósticos derivados del aprendizaje automático mejoran los modelos dinámicos del centro mundial, especialmente en algunas regiones del área de estudio (Figura 3.13).

5.3. Conclusiones del pronóstico probabilístico de precipitación en la Región del Comahue

Se ha implementado un pronóstico probabilístico de precipitación estacional (trimestral para los 12 trimestres del año) en la región argentina del Comahue, utilizando un conjunto de modelos derivados de diversas metodologías estadísticas.

En este trabajo se ha trabajado en un área limitada y utilizando predictores que han sido previamente estudiados y se conoce su relación con la precipitación. Los resultados indicaron que este pronóstico produce resultados aceptables en las subcuencas SRN y SRL especialmente en invierno. Se ha observado que mejora la climatología simple en temporada de invierno, especialmente al pronosticar por encima y por debajo de las categorías normales. Esto es muy beneficioso para los tomadores de decisiones que operan las represas en la región, ya que son altamente dependientes de las lluvias invernales.

Como se detalló en la sección 5.1. esta metodología de pronóstico probabilístico está siendo probada en tiempo real de forma experimental y está publicada en el sitio web http://perspectiva.at.fcen.uba.ar/comahue-perspectiva.php

Además, en la página web se muestra un detalle de la eficiencia obtenida hasta ahora con la utilización de esta metodología de pronóstico.

El pronóstico se considera:

Óptimo: si hay coincidencia total entre la categoría pronosticada y la observada, o bien cuando se incluyen dos categorías (por ejemplo: normal a subnormal) si hay coincidencia entre la categoría extrema pronosticada y la observada.

Aceptable: si la categoría pronosticada difiere en una categoría con la observada.

Ineficiente: si la categoría pronosticada difiere en dos categorías con la observada, o bien cuando se incluyen dos categorías y la categoría extrema pronosticada es opuesta a la observada.

Para evaluar los resultados obtenidos se determina la eficiencia de los pronósticos, que va a estar dada por la suma de todos los casos calificados como “óptimos” y “aceptables”.

La Tabla 5.1 muestra la verificación de los pronósticos que se han realizado desde el año 2019 en adelante:

Tabla 5.1.Verificación de los pronósticos.

	subnormal
	sobrenormal
N	normal
S/PRON	SIN PRONÓSTICO

Las celdas coloreadas que además incluyen la letra "N" indican que existe la posibilidad de que se den ambas categorías.

		LIMAY		NEUQUÉN		NEGRO	
año	trimestre	pp estimada	pp observada	pp estimada	pp observada	pp estimada	pp observada
2018	NDE	N	N	N	N	N	N
	DEF	N		N		N	
2019	EFM	S/PRON		S/PRON		S/PRON	
	FMA	S/PRON		S/PRON		S/PRON	
	MAM	N		N		N	
	AMJ	N	N	N	N		N
	MJJ	N	N	N	N	N	N
	JJA	N	N	N	N	N	N
	JAS	N	N	N		N	
	ASO	N		N		N	
	SON	N	N	N	N	N	N
	OND	N	N	N	N	N	N
	NDE		N		N		N
	DEF	N	N	N	N	N	
2020	EFM	S/PRON	N	S/PRON	N	S/PRON	N
	FMA	N	N	N	N	N	N
	MAM		N		N	N	N
	AMJ	S/PRON	N	S/PRON	N	S/PRON	N
	MJJ	N		N		N	
	JJA	N	N	N	N	N	N
	JAS	N	N	N	N		N
	ASO	N	N	N	N	N	N
	SON	N	N	N	N	N	N
	OND	N	N	N	N	N	N
	NDE	N	N	N	N	N	N
	DEF	N	N	N	N	N	N
2021	EFM	N	N	N	N	N	N
	FMA	N	N	N	N	N	N
	MAM	N	N	N	N		N
	AMJ	N		N	N	N	N
	MJJ	N		N	N	N	N
	JJA				N	N	
	JAS	N		N	N	N	
	ASO	N	N	N	N		
	SON	N		N	N	N	N
	OND			N	N	N	N
	NDE	N	N	N	N	N	N
	DEF		N		N		
2022	EFM	N	N	N	N	N	
	FMA					N	N
	MAM	N	N	N	N	N	
	AMJ	N	N	N	N	N	N
	MJJ	N	N		N		
	JJA		N	N	N	N	N
	JAS	N	N	N	N	N	N
	ASO	N	N		N		N
	SON	N		N	N	N	
	OND	N	N	N	N	N	N
	NDE	N		N	N		N
	DEF	N		N	N	N	N
2023	EFM		N		N	N	
EFICIENCIA	Óptimo		30 %		32 %		28 %
	Aceptable		44 %		40 %		44 %
	TOTAL		74 %		72 %		72 %
	Ineficiente		26 %		28 %		28 %

REFERENCIAS

Allaire J, Chollet F (2023). *R Interface to 'Keras'.* https://CRAN.R-project.org/package=keras

Allaire J, Tang Y (2022). *R Interface to 'TensorFlow'.* https://CRAN.R-project.org/package=tensorflow

Aravena, J and Luckman B (2009) *Spatiotemporal rainfall patterns in Southern South America. Int J Climatol 29: 2106–2120.* https://doi.org/10.1002/joc.1761.

Barreiro M (2009) *Influence of ENSO and the South Atlantic Ocean on climate predictability over Southeastern South America. Climate Dynamics. DOI 10.1007/s00382-009-0666-9.*

Barnston A.G. (1994): *Linear Statistical Short-Term Climate Predictive Skill in the Northern Hemisphere. J. Climate, 7, 1513–1564,* https://doi.org/10.1175/1520-0442(1994)007<1513:LSSTCP>2.0.CO;2 .

Barnston A, Kumar A, Goddard L and Hoerling M (2005) *Improving seasonal prediction practices through attribution of climate variability. BAMS. 59-72.*

Bennett, N.D., Croke, B.F., Guariso, G., Guillaume, J.H., Hamilton, S.H., Jakeman, A.J., Marsili-Libelli, S., Newham, L.T., Norton, J.P., Perrin, C., et al., 2013. *Characterizing performance of environmental models. Environ. Model. Software 40, 1e20.*

Bivand R, Lewin-Koh N (2023). *maptools: Tools for Handling Spatial Objects_. R package version 1.1-7, <*https://CRAN.R-project.org/package=maptools*>.*

Boukabara S, Krasnopolsky V, Stewart JQ, Maddy ES, Shahroudi N and Hoffman RN (2019). *Leveraging modern artificial intelligence for remote sensing and NWP: Benefits and challenges. Bulletin of the American Meteorological Society. 100 (12). ES473–ES491.*

Castañeda E and González M H (2008) *Some aspects related to precipitation variability in the Patagonia region in Southern South America. Atmosfera 21 3: 303-317.*

Chollet et al. (2015) *Keras. https://keras.io. Accessed September 13th, 2022*

Coelho C, Stephenson S, Balmaseda M, Doblas Reyes F and Oldenborge G (2005) *Towards an integrated seasonal forecasting system for South America. J Climate 19: 3704-3721.*

Cortes C and Vapnik V (1995) *Support-vector networks. Mach Learn., 20: 273–297. https://doi.org/10.1007/BF00994018.*

Cortes C, Vapnik V (1995) *Support-vector networks. Mach Learn 20: 273–297. https://doi.org/10.1007/BF00994018*

Díaz G, Vita M, Hobouchian M P, Ferreira L and Giordano L (2021) *Expansión de la red de referencia empleando los datos de precipitación de las estaciones meteorológicas automáticas de terceros. Technical Note SMN, pp 2021-90.*

Doblas-Reyes Francisco J., Javier García-Serrano, Fabian Lienert, Aida Pintó Biescas, Luis R. L. Rodrigues (2013). *Seasonal climate predictability and forecasting: status and prospects. Wiley Interdisciplinary Reviews: Climate Change. Volume 4, Issue 4*

Douglas Nychka, Reinhard Furrer, John Paige, Stephan Sain (2021). *"fields: Tools for spatial data." R package version 14.1, https://github.com/dnychka/fieldsRPackage.*
Ebert-Uphoff I and Hilburn K (2020) *Evaluation, Tuning and Interpretation of Neural Networks for Working with Images in Meteorological Applications. Bulletin of the American Meteorological Society. https://doi.org/ 10.1175/BAMS-D-20-0097.1*

FAO (2011). *State of the World´s Forests. Food and Agriculture Organization of the United Nations, Rome, Italy.*

Friedman J, Tibshirani R, Hastie T (2010). *Regularization Paths for Generalized Linear Models via Coordinate Descent. Journal of Statistical Software. DOI: https://doi.org/10.18637/jss.v033.i01.*

Garbarini E M, Skansi M M, González M H and Rolla A L (2016) *ENSO Influence over Precipitation in Argentina. Advances in Environmental Research, Justin A. Daniels Ed., NOVA Publisher, NY, USA, pp 223-246*

Garbarini E M, González M H and Rolla A L (2019) *The influence of Atlantic High on seasonal rainfall in Argentina. International Journal of Climatology 39 12: 4688-4702. https://doi.org/10.1002/joc.6098*

Garbarini E M, González M H and Rolla A L (2020) *Connection between sea surface temperature patterns and low-level geopotential height in the South Atlantic Ocean. Atmosfera, 33: 175-185. https://doi.org/10.20937/ATM.52641.*

Garreaud R, Lopez P, Minvielle M, and Rojas M (2013) *Large-scale control on the Patagonian climate. J Climate 26: 215-230. https://doi.org/10.1175/JCLI-D-12-00001.1*

Goddard L, Barnston A and Mason S (2003). *Evaluation of the IRI´s "net assessment" seasonal climate forecasts. 1997-2001. BAMS. 1761-1781.*

González M H and Cariaga M L (2011) *Estimating winter and spring rainfall in the Comahue region (Argentine) using statistical techniques. Advances in Environmental Research, Justin A. Daniels Ed., NOVA Publisher, NY, USA, pp 103-118.*

González M H and Vera C S (2010) *On the interannual winter rainfall variability in Southern Andes. Int J Climatol 30: 643-657. https://doi.org/10.1002/joc.1910.*

González, M H and Herrera N (2014) *Statistical prediction of Winter rainfall in Patagonia (Argentina). Horizons in Earth Science Research, Benjamin Veress and Jozsi Szigethy Eds., NOVA Publisher, NY, USA, pp 221-238.*
González M H (2015) *Statistical seasonal rainfall forecast in Neuquén river basin (Comahue Region, Argentina). Climate 3: 349-364.*

González M H, Garbarini E M and Romero P E (2015) *Rainfall patterns and the relation to atmospheric circulation in northern Patagonia (Argentina). Advances in Environmental Research, Justin A. Daniels Ed., NOVA Publisher, NY, USA, pp 85-100.*

González M H, Rolla A L (2019) *Comparison between statistical precipitation prediction in northern Patagonia (Argentina) using ERA- INTERIM and NCEP reanalysis datasets. Agricultural Research Updates, Prathamesh Gorawala and Srushti Mandhari Eds., NOVA Science Publications, NY, USA, pp 117-128.*

González M H, Losano F and Eslamian S (2021) *Rainwater Harvesting Reduction Impact on Hydro-Electric Energy in Argentina. Handbook of Water Harvesting and conservation, S. Eslamian Ed., John Wiley & Sons, NY, USA, pp 251-260.*

Hartigan JA (1985). *Statistical theory in clustering. Journal of Classification 2: 63–76. https://doi.org/10.1007/BF01908064*

Hartigan JA and Wong MA (1979). *Algorithm AS 136: A K-means clustering algorithm. Applied Statistics 28: 100-108. DOI 10.2307/2346830.*

Hartmann, H.C., Pagano, T.C., Sorooshiam, S. and Bales, R. (2002). *Confidence builder: evaluating seasonal climate forecasts from user perspectives. Bull Amer. Met. Soc., 84, 683-698*.

Hastie, T. J.; Tibshirani, R. J. (1990). *Generalized Additive Models. Chapman & Hall/CRC. ISBN 978-0-412-34390-2.*

Hijmans R (2023). *raster: Geographic Data Analysis and Modeling_. R package version 3.6-20, <https://CRAN.R-project.org/package=raster>.*

Hyndman RJ (2013*). "Forecasting: principles and practice.* *https://CRAN.R-project.org/package=fpp*

Hyndman R, Athanasopoulos G, Bergmeir C, Caceres G, Chhay L, O'Hara-Wild M, Petropoulos F, Razbash S, Wang E, Yasmeen F (2023. *Forecasting functions for time series and linear models. R package version 8.21, https://pkg.robjhyndman.com/forecast/.*

Hyndman R J and Athanasopoulos G (2022) *Forecasting principles and practice. O Texts: Melbourne, Australia. http://otexts.org/fpp2/. Accessed September 13th, 2022.*

James G, Witten D, Hastie T and Tibshirani R (2013). *An Introduction to Statistical Learning, Springer, New York, 440 pp.*

Kalnay E, Mo K C and Paegle J (1986) *Large-Amplitude, Short Scale Stationary Rossby Waves in the Southern Hemisphere: Observations and Mechanistics Experiments on determine their origin. Journal of Atmospheric Science 3: 252-275.*

Kalnay E, Kanamitsu M, Kistler R, Collins W, Deaven D, Gandin L, Iredell M, Saha S, White G, Woollen J, Zhu I, Chelliah M, Ebisuzaki W, Higgings W, Janowiak J, Mo K C, Ropelewski C, Wang J, Leetmaa A, Reynolds R, Jenne R and Joseph D (1996) *The NCEP/NCAR Reanalysis 40 years-project. Bull American Meteorological Society 77: 437-471.*

Kaski S (2011). *Self-Organizing Maps. Sammut C., In: Webb GI (eds) Encyclopedia of Machine Learning. Springer, Boston, MA. https://doi.org/10.1007/978-0-387-30164-8_746*

Kousky, V.E., 1988. "*Pentad outgoing longwave radiation climatology for the South America sector", Revista Brasilera de Meteorología, 3, 217-231.*

Kousky, V.E., et al. (1984). *A Review of the Southern Oscillation: Oceanic-Atmospheric Circulation Changes and Related Rainfall Anomalies. Tellus A, 36, 490-504.* https://doi.org/10.1111/j.1600-0870.1984.tb00264.x

Kumar A (2006). *On the interpretation and utility of skill information for seasonal climate predictions. Mon.Wea. Rev 135: 1974 – 1984.*

Kumar A, Ceron J, Coelho C, Ferranti L, Graham R, Jones D, Merryfield W, Muñoz A, Pai S and Rodriguez E (2020) *Guidance on Operational Practices for Objective Seasonal Forecasting, World Meteorological Organization, WMO-No. 1246, Geneva 2, Switzerland, 106pp.*

Leetmaa A (2003) *Seasonal Forecasting. Innovation in practice and institutions. Bull American Meteorological Society 84: 1686 - 1691.*

Lee Y, Hall D, Stewart J and Govett M (2018). *Machine learning for targeted assimilation of satellite data. Joint European Conference on Machine Learning and Knowledge Discovery in Databases. Springer. 53–68.*

Leetmaa A (2003). *Seasonal Forecasting. Innovation in practice and institutions. BAMS 84: 1686 - 1691.*

Mason S.J., M.K. Tippett (2017): *Climate Predictability Tool version 15.6.1, Columbia University Academic Commons* https://doi.org/10.7916/D8SF37S0

Meehl G et. Al (2014). *Decadal Climate Prediction: An Update from the Trenches. BAMS 95:issue 2. DOI:* https://doi.org/10.1175/BAMS-D-12-00241.1

Meyer D, Dimitriadou E, Hornik K, Weingessel A, Leisch F (2023). *Misc Functions of the Department of Statistics, Probability Theory Group <*https://CRAN.R-project.org/package=e1071*>.*

Milborrow S (2022. *Plot a Model's Residuals, Response, and Partial Dependence Plots. R package version 3.6.2, DOI:* https://CRAN.R-project.org/package=plotmo

Mo K C (2000) *Relationships between low frequency variability in the Southern Hemisphere and sea surface temperature anomalies. J Climate 13: 3599-3610*

Mosavi A, Pinar Ozturk I and Kwok-wing C (2018) *Flood Prediction Using Machine Learning. Literature Review Water 10: 1-41.*

Murphy A, Epstein E (1988). *Skill Scores and Correlation Coefficients in Model Verification. Climate Analysis Center, National weather Service, Washington , D.C.*

Murtagh F and Legendre P (2014). *Ward's hierarchical agglomerative clustering method: which algorithms implement Ward's criterion? Journal of Classification 31: 274--295. DOI 10.1007/s00357-014-9161-z.*

Nobre C, Marengo J, Cavalcanti I, Obregon G, Barros V, Camilloni I, Campos N and Ferreira A (2005) *Seasonal to decadal predictability and prediction of South America Climate. J Climate 19, 23: 5988 - 6004.*

Paruelo J, Beltran A, Jobbagy E, Sala O and Golluscio R (1998) *The climate of Patagonia: General patterns and controls on biotic processes. Ecol Austral 8: 85–101.*

Pierce D (2023). *ncdf4: Interface to Unidata netCDF (Version 4 or Earlier) Format Data Files_. R package version 1.21,* https://CRAN.R-project.org/package=ncdf4 .

Posit team (2023). *RStudio: Integrated Development Environment for R. Posit Software, PBC, Boston, MA. URL* http://www.posit.co/.

Potts, J.M.,(2003). *Basic Concepts. Forecast Verification: a Practitioner's Guide in Atmospheric Science, Second Edition, pp. 11e29.*

Prohaska F (1976) *The climate of Argentina, Paraguay, and Uruguay. Climates of Central and South America. W. Schwerdtfeger Ed, World Survey of Climatology, Elsevier, pp 13-72.*

R Core Team (2021). R*: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria.URL* https://www.R-project.org/.

Rasmusson Eugene M. and Wallace John M. (1983). *Meteorological Aspects of the El Niño/Southern Oscillation. New Series, Vol. 222, No. 4629 (Dec. 16, 1983), pp. 1195-1202 (8 pages)*

Reichstein M, Camps-Valls G, Stevens V, Jung M, Denzler J, Carvalhais N and Coauthors (2019). *Deep learning and process understanding for data-driven earth system science. Nature 566 (7743): 195–204.*

Roger S. Bivand, Edzer Pebesma, Virgilio Gomez-Rubio (2013). *Applied spatial data analysis with R, Second edition. Springer, NY.* https://asdar-book.org/

Romero P E, González M H, Rolla A L and Losano F (2020) *Forecasting annual precipitation to improve the operation of dams in the Comahue region (Argentina). Hydrological Sciences Journal, 65 11:1974–1983.*

Saji, N.H., Goswami, B.N., Vinayachandran, P.N., and Yamagata, T., 1999. *A dipole mode in the tropical Indian Ocean, Nature 401, 360-363.*

Saurral R, Camilloni I and Barros V (2017) Low-frequency variability *and trends in centennial precipitation stations in southern South America. Int J Climatol 37 4: 1774-1793.*

Schauberger P, Walker A (2023). *openxlsx: Read, Write and Edit xlsx Files. R package version 4.2.5.2, <*https://CRAN.R-project.org/package=openxlsx*>.*

Smith DM, Cusack S, Colman AW, Folland CK, Harris GR, Murphy JM (2007). *Improved surface temperature prediction for the coming decade from a global climate model. Science 317:796–799. doi:10.1126/science.1139540*

Soares dos Santos T, Mendes D and Rodrigues Torres R (2016) *Artificial neural networks and multiple linear regression model using principal components to estimate rainfall over South America. Nonlin Processes Geophys 23: 13–20. https://doi.org/10.5194/npg-23-13-2016.*

Svoboda V, Máca P and Hanel M (2014) *Spatial correlation structure of monthly rainfall at a mesoscale region of north-eastern Bohemia. Theor Appl Climatol 121: 359–375.*

Tibshibari R (1996) *Regression Shrinkage and Selection via the Lasso. Journal of the Royal Statistical Society, Series B (Methodological) 58 1: 267–288.*

Thompson, D.W. and Wallace, J.M. 2000. *Annular modes in the extratropical circulation. Part I: Month-to-month variability. J. Climate 13: 1000-1016.*

Tokay, A, Roche R and Bashor P (2014) *An Experimental Study of Spatial Variability of Rainfall. Journal of Hydrometeorology 15 2: 801-812.*

Wickham H, François R, Henry L, Müller K, Vaughan D (2023). *dplyr: A Grammar of Data Manipulation_. R package version 1.1.2, https://CRAN.R-project.org/package=dplyr.*

Wilks D S (2011) *Statistical methods in the atmospheric sciences. 3rd. Edition, Academic Press, San Diego, California, USA, 704 pp.*

WMO, 2020: *Guidance on Operational Practices for Objective Seasonal Forecasting. WMO – No. 1246. ISBN 978-92-63-11246-9.*

WMO, 2018: *Guidance on Verification of Operational Seasonal Climate Forecasts. WMO – No. 1220. ISBN 978-92-63-11220-0.*

Wood S (2006) *Generalized Additive Models: An Introduction with R. 2nd. Edition, CRC Press, Taylor & Francis, 474 pp.*