

Pronóstico de precipitación mensual en Argentina aplicando técnicas de machine learning

Eugenia M. Garbarini¹, Marcela H. González^{2,3}, Alfredo L. Rolla^{2,3}

¹ Servicio Meteorológico Nacional, Av. Dorrego 4019, Ciudad Autónoma de Buenos Aires, Argentina.

² Centro de Investigaciones del Mar y la Atmósfera (CIMA – CONICET), Intendente Güiraldes 2160 Pabellón II-2do. Piso, Buenos Aires, Argentina.

³ Institut Franco-Argentin d'etudes Sur Le Climat Et Ses Impacts, Unite Mixte Internationale, Intendente Güiraldes 2160 Pabellón II-2do. Piso, Buenos Aires, Argentina.
egarbarini@smn.gob.ar

Abstract. Las precipitaciones en Argentina presentan múltiples características debido a su extenso territorio y, al afectar a muchos sectores socio-económicos, la previsibilidad con un mes de antelación es de gran interés para una planificación y toma de decisiones eficaces. El objetivo de este trabajo es desarrollar un pronóstico de precipitación mensual mediante el uso de un conjunto de modelos estadísticos que combinan tanto técnicas convencionales como de aprendizaje automático. Utilizando datos de precipitación acumulada de 91 estaciones meteorológicas para el período 1981-2020 junto con datos de reanálisis de NCEP/NCAR se seleccionó un conjunto de predictores para cada mes y región. Los resultados indican que las técnicas de aprendizaje automático presentan errores menores especialmente en regiones con precipitaciones limitadas como la Patagonia, donde muchas actividades dependen de la disponibilidad de agua. Esto subraya la importancia de incorporar técnicas de aprendizaje automático para mejorar el pronóstico de precipitaciones mensuales en Argentina.

Keywords: aprendizaje automático, precipitación mensual, servicios climáticos.

1 Introducción

En el norte de Argentina, la cordillera de los Andes actúa bloqueando el aire húmedo proveniente del océano Pacífico y el flujo está gobernado por vientos predominantes del noreste que advectan vapor de agua en niveles bajos provenientes de regiones tropicales de Sudamérica y del océano Atlántico. En regiones como el Litoral, el noroeste, centro y centro-este del país, las mayores precipitaciones se registran en el semestre cálido (de octubre a marzo) mientras que en los meses de abril a septiembre se registran los mínimos de precipitación. Al sur de los 38°S, la altura de los Andes se reduce, y así el aire húmedo puede fluir desde el océano Pacífico, estableciendo un flujo desde el oeste durante todo el año. Esto genera un intenso contraste entre las

áreas de vegetación densa cercanas a la cordillera de los Andes y las llanuras secas que se extienden hacia la costa atlántica. De esta manera, sobre la cordillera las mayores precipitaciones suelen registrarse en los meses de invierno (de abril a agosto) mientras que, hacia la costa atlántica, las precipitaciones son escasas durante todo el año, con un leve máximo en otoño y primavera. La diversidad de climas y suelos en Argentina expone a la producción en todo el país a riesgos climáticos, con numerosos estudios que destacan esta vulnerabilidad ([1] [2] entre otros). Dado que la precipitación impacta sectores clave como la agricultura y la energía, la predictibilidad de su variabilidad interanual ha sido muy discutida últimamente.

La predictibilidad climática depende de la influencia de variables como la temperatura superficial del mar en la circulación atmosférica y, por consecuencia, en otras variables meteorológicas. Sin embargo, tanto los modelos dinámicos como los estadísticos que abordan este problema enfrentan desafíos para pronosticar el clima en Sudamérica ([3] [4] [5] [6] entre otros). Aunque se han desarrollado modelos estadísticos para predecir la precipitación en diversas regiones con buenos resultados, estos pueden presentar errores debido a la metodología empleada o a la variabilidad aleatoria inherente [7]. Si bien las variables meteorológicas tienen un componente aleatorio imposible de predecir, hay otras fuentes de error diferentes que pueden abordarse, como ser la metodología utilizada para construir el modelo de pronóstico. Investigaciones recientes [8] sugieren que un enfoque híbrido, combinando procesos climáticos y aprendizaje automático, podría mejorar la precisión de los pronósticos. En este trabajo, con el objetivo de mejorar los pronósticos de precipitación mensual en Argentina, se han propuesto integrar técnicas convencionales como la regresión lineal múltiple (RLM) y los Modelos Aditivos Generalizados (GAMs), junto con técnicas de aprendizaje automático como Redes Neuronales Artificiales (ANNs) y Máquinas de Soporte Vectorial (SVRs).

2 Datos y metodologías

Se utilizaron datos de precipitación acumulada de 91 estaciones meteorológicas para el período 1981-2020 para calcular las ondas anuales de precipitación. A ellas se les aplicó la metodología de Lund [9] con un umbral de 0,68 para el coeficiente de correlación, definiendo 6 clusters o regiones según su comportamiento anual (Figura 1). Por otro lado, se consideraron como predictores climáticos las variables globales de temperatura superficial del mar (TSM), altura geopotencial (HGT) en los niveles de 1000, 500 y 200 hPa, agua precipitable (TCW) en la columna atmosférica y viento zonal y meridional en el nivel de 850 hPa (U850 y V850) provenientes del reanálisis NCEP/NCAR [10] para el período 1981-2015. Para cada región definida y cada mes, se calculó la serie de precipitación media espacial característica para luego ser correlacionadas con las series temporales de las variables globales con un mes de antelación. Las áreas con correlación significativa, usando un test de distribución normal con un 95% de confianza, fueron definidas como predictores. Finalmente la técnica de Least Absolute Shrinkage and Selection Operator (LASSO) [11] se aplicó para determinar el grupo definitivo de predictores que fuesen independientes entre sí para eventualmente evitar el fenómeno de overfitting [12].

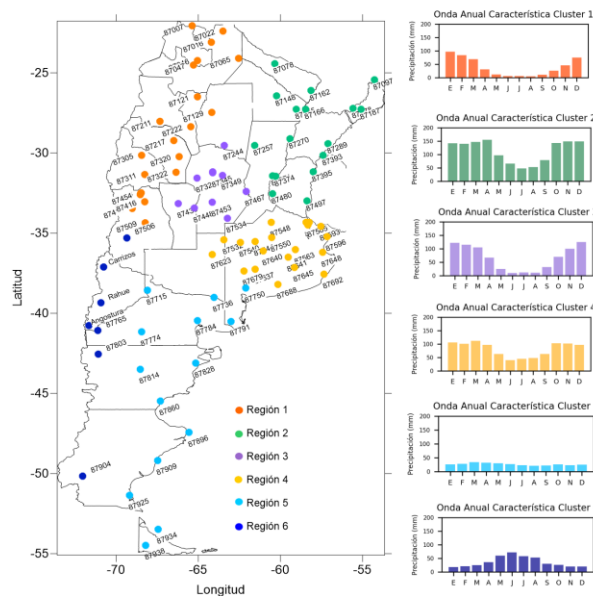


Fig. 1. Estaciones meteorológicas consideradas (*puntos*) para la clusterización junto con las ondas anuales características (*gráficos de barras*). Los colores identifican las distintas regiones o clusters: región NOA (*cluster 1, naranja*), región Litoral (*cluster 2, verde*), región central (*cluster 3, violeta*), región Buenos Aires (*cluster 4, amarillo*), región Patagonia este (*cluster 5, celeste*), región Patagonia oeste (*cluster 6, azul*).

Se consideró el periodo de entrenamiento 1981-2015 para generar modelos estadísticos con los predictores definidos previamente con el fin de pronosticar el año 2016. Luego se iteró este proceso avanzando un año en el periodo de entrenamiento para pronosticar el siguiente hasta finalmente pronosticar el año 2023 (Figura 2). Esto definió el periodo de verificación como 2016-2023.

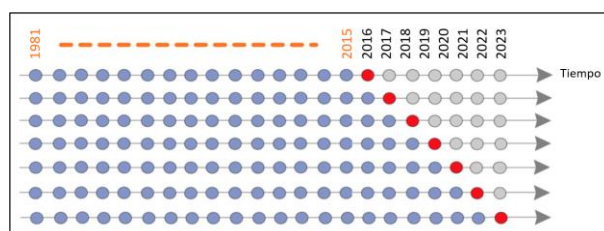


Fig. 2. Esquema de construcción de los modelos estadísticos de pronóstico.

Los modelos fueron generados teniendo en cuenta las metodologías LRM, GAM, SVR y ANN y utilizando los sets de predictores independientes que fueron definidos en los pasos previos. La metodología más simple considerada, LRM, aplica Forward Stepwise Regression [12], una técnica iterativa de construcción de modelos que suma en cada paso de la construcción una variable predictora y finalmente genera el modelo con aquel conjunto de predictores que explican la mayor varianza con el menor error posible. LRM sigue la expresión:

$$Y = \sum a_i X_i + e . \quad (1)$$

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

Donde Y es el predictando, X_i los predictores, a_i los coeficientes lineales y e el error. Para considerar la posibilidad de comportamientos no lineales, se agregó la metodología GAM [13], en la cual se parte de la base de la metodología anterior pero permite que los coeficientes del modelo sean funciones:

$$Y = \sum F(x_i) + e . \quad (3)$$

$$e \sim N(0; \sigma^2) . \quad (4)$$

Por otro lado, la metodología SVR [14] [15], es similar a la regresión lineal en la expresión:

$$y = wx + b . \quad (5)$$

Donde y es el predictando, x el predictor, w el coeficiente lineal y b la ordenada al origen. En este caso (5) se considera como un hiperplano y, a diferencia de otros modelos de regresión que pretenden minimizar el error entre el valor real y el predicho, la metodología de SVR intenta ajustar la mejor línea dentro de un umbral (distancia entre el hiperplano y la línea límite).

Finalmente, se consideraron redes neuronales (ANNs), las cuales se componen por una gran cantidad de unidades simples de procesamiento (neuronas) conectadas entre sí y agrupadas en capas, comenzando por una capa de entrada (input) que recibe los datos y finalizando por una de salida (output). El algoritmo de aprendizaje automático aprende a partir del ajuste de los pesos que conectan a las neuronas entre sí para modelar el conjunto de datos. La salida de una neurona Y_i se obtiene a partir de la transformación de la suma ponderada de las entradas que recibe mediante una función de activación conocida como Unidad Lineal Rectificada (ReLU por sus siglas en inglés) y dada por la expresión:

$$Y_i = f(\sum_{j=1}^n w_{ij} x_j - \theta_i) = f(\sum_{j=0}^n w_{ij} x_j) . \quad (6)$$

Donde w_{ij} son los pesos, f es la función de activación y θ_i es el umbral de activación. Esta expresión es una función lineal por partes que devuelve el input en caso de ser positiva y es nula en cualquier otro caso. Suele ser de las más utilizadas a la hora de construir redes neuronales ya que el modelo que utiliza suele ser fácil de

entrenar y generalmente alcanza mejor performance. Para este análisis se definieron 4 diferentes arquitecturas usando retro propagación (Figura 3):

- 2 capas, una de 16 y otra de 32 neuronas, con un dropout de 0,1 y 0,2 respectivamente, con función de activación ReLU.
- 2 capas de 32 neuronas cada una con un dropout de 0,1 y 0,2 respectivamente con función de activación ReLU.

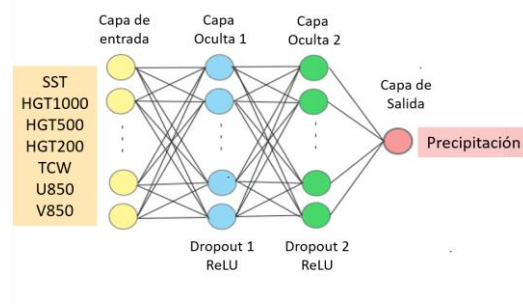


Fig. 3. Esquema de la arquitectura utilizada para la construcción de los modelos ANNs.

Para cada una de las cuatro metodologías aplicadas se obtuvo un set de posibles modelos, de los cuales únicamente se seleccionaron aquellos que explicaran más de un 50% de la varianza de la precipitación a partir del coeficiente de determinación ajustado (R^2_{adj}) [16] dado por la expresión:

$$R^2_{adj} = 1 - \{[(1 - R^2)(n - 1)] / (n - k - 1)\} . \quad (7)$$

Donde n es la cantidad de datos, k el número de variables predictores y:

$$R^2 = 1 - (\sigma_r^2 / \sigma^2) . \quad (8)$$

Donde σ_r^2 es la varianza de la precipitación que no puede ser explicada por las variables del modelo (varianza residual) y σ^2 es la varianza de la variable dependiente.

Para cada cluster, mes y año de predicción, se calculó la media y el desvío estándar de la precipitación pronosticada derivada de cada metodología y del ensamble de metodologías, así como también el error cuadrático medio según:

$$RMSE = \{[\sum (Y_{p_i} - Y_i)^2 / n]\}^{1/2} . \quad (9)$$

Donde Y_{p_i} es la precipitación pronosticada e Y_i la observada, σ_p y μ_p son el desvío estándar y la media de la precipitación pronosticada. Así, cuanto más pequeño es un valor de RMSE, más cercanos son entre sí los valores observados y pronosticados.

3 Resultados y discusión

El resultado de la regionalización por clusters y el comportamiento de la precipitación en las distintas zonas pueden verse en la Figura 1. En los clusters 1 a 4 los máximos de precipitación se registran en los meses cálidos mientras que los mínimos ocurren desde aproximadamente abril a septiembre. El acumulado de precipitación varía según la región, siendo máximo en el Litoral (C2), seguido por el centro del país (C3) y la provincia de Buenos Aires (C4) y por último el noroeste argentino (NOA, C1). Por otro lado, en la Patagonia andina (C5) las mayores precipitaciones suelen registrarse en otoño e invierno y en la llanura patagónica (C6) la precipitación, si bien escasa, muestra un máximo muy leve en otoño y primavera.

Al realizar el pronóstico según el esquema de la Figura 2, se aplicaron todas las metodologías para cada cluster y cada mes, seleccionando los mejores modelos según los criterios mencionados en la sección anterior, y también se construyó el pronóstico por ensamble (ENS) teniendo en cuenta los modelos seleccionados de todas las metodologías en conjunto. Es importante mencionar que, en algunos casos y para algunas metodologías, ningún modelo logró cumplir el requisito de explicar al menos el 50% de la varianza.

La Figura 4 muestra el valor de precipitación observado en contraposición con el valor pronosticado por el ensamble de metodologías para cada región del país y cada mes de pronóstico comprendido en el periodo de verificación. Se puede ver que, en líneas generales, los valores se encuentran cerca de la línea de identidad, lo que indica una buena performance, particularmente en los meses de marzo, julio a septiembre y noviembre. Para el semestre cálido, donde las precipitaciones son superiores al norte del paralelo 38°S, la dispersión suele ser mayor en el NOA (C1) en enero y febrero, y en el Litoral (C2) de octubre a diciembre. Por otro lado, en la Patagonia Este (C5) la performance del pronóstico por ensamble es buena sin importar el mes.

La Figura 5 muestra el RMSE del ensamble para el periodo de verificación. Se puede ver que el RMSE es mayor en el Litoral, centro del país y Buenos Aires (C2, C3 y C4, respectivamente) ya que la precipitación es mayor en esas regiones, sin embargo, es importante notar que en los meses de abril a septiembre en C3 y de abril a julio en C4 el RMSE es menor. Esto demuestra que el ensamble presenta mejores resultados allí en esos meses, incluyendo los meses donde la precipitación es menor. Lo opuesto se puede observar en la Patagonia andina (C6) donde el RMSE del ensamble aumenta de mayo a octubre, abarcando los meses más lluviosos de esa región.

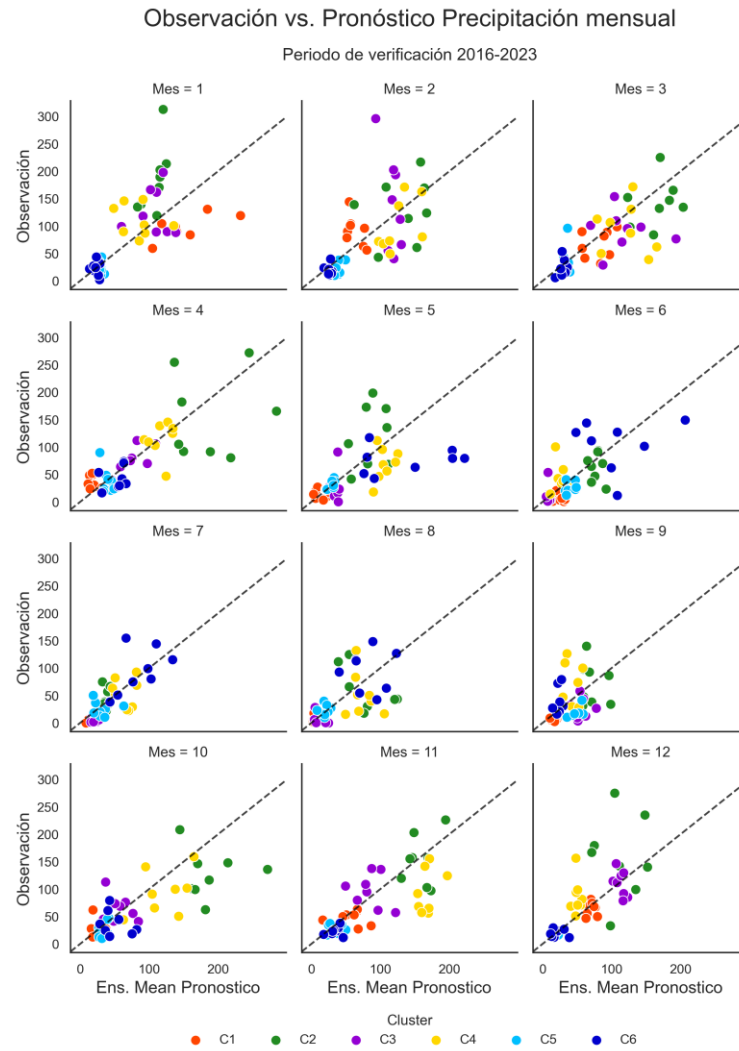


Fig. 4. Diagrama de dispersión del valor observado versus valor pronosticado junto con la línea de identidad (*línea punteada*) utilizando el ensemble para cada mes del periodo de verificación. Los colores representan los diferentes clusters.

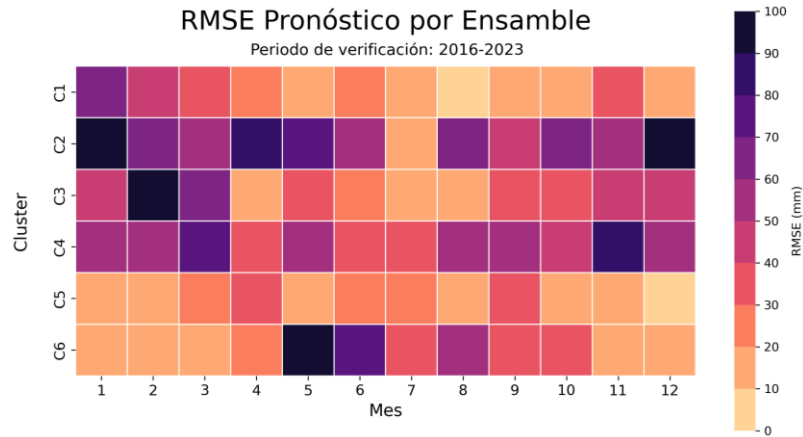


Fig. 5. Mapa de calor del RMSE del ensamble por clusters y mes.

Análogamente, la Tabla 1 muestra el menor valor de RMSE para cada mes y cada cluster teniendo en cuenta las cuatro metodologías por separado, de forma tal de identificar aquellas técnicas que funcionan mejor según la región y momento del año. Para esta comparación, se decidió hacer a un lado los meses de enero y febrero para C1 y julio para C4 ya que sólo una metodología fue capaz de construir modelos que expliquen el 50% de la varianza.

Tabla 1. Metodología con el menor RMSE mensual (*entre paréntesis el valor en mm*) para el periodo de verificación 2016-2023. En negrita se muestran los casos en que solo dicha metodología logró cumplir el requisito de que los modelos expliquen al menos el 50% de la varianza.

Mes	C1	C2	C3	C4	C5	C6
1	ANN (62)	SVR (77)	RLM (25)	SVR (30)	SVR (13)	SVR (12)
2	ANN (49)	GAM (59)	GAM (71)	RLM (49)	SVR (15)	ANN (12)
3	GAM (27)	ANN (49)	SVR (50)	ANN (66)	RLM (15)	ANN (15)
4	RLM (5)	GAM (36)	ANN (15)	GAM (35)	SVR (30)	ANN (24)
5	SVR (10)	ANN (66)	RLM (14)	SVR (39)	SVR (9)	SVR (60)
6	SVR (7)	SVR (34)	ANN (18)	SVR (34)	SVR (15)	ANN (68)
7	SVR (8)	GAM (17)	ANN (16)	ANN (31)	SVR (8)	ANN (36)
8	SVR (7)	SVR (54)	SVR (14)	ANN (49)	SVR (11)	SVR (49)
9	SVR (9)	ANN (43)	SVR (35)	SVR (46)	SVR (18)	SVR (25)
10	GAM (15)	SVR (54)	SVR (27)	ANN (46)	GAM (13)	SVR (29)
11	SVR (27)	GAM (55)	ANN (42)	SVR (66)	SVR (8)	ANN (14)
12	ANN (15)	SVR (87)	SVR (33)	SVR (29)	SVR (5)	SVR (12)

Puede verse que para todo el país, la metodología que supone mejores resultados es SVR al ser la que menor RMSE presenta en la mayoría de los meses del año. Esto es especialmente notorio en la Patagonia Este (C5) donde ocurre dicho fenómeno en diez meses. Además, en la Patagonia andina (C6) tanto la técnica SVR como ANN presentan los menores RMSE, demostrando que, en líneas generales, la inclusión de metodologías de aprendizaje supervisado presupone una mejoría a los pronósticos de precipitación en el país, con especial énfasis en el sector patagónico donde la disponibilidad de agua es muy limitada. Un resultado similar se desprende a la hora de analizar los meses de invierno, donde predominan dichas técnicas como aquellas con mejores resultados. Para el mes de diciembre ocurre el mismo fenómeno, pero no se ve una predominancia tan clara hacia enero y febrero. Por otro lado, es importante mencionar que en la región del Litoral (C2), en donde se registran los mayores valores acumulados de precipitación, las técnicas con menor RMSE son SVR y GAM, lo que podría señalar la importancia de incluir procesos no lineales en dicha región.

4 Conclusiones

El pronóstico mensual de la precipitación es de vital importancia para las diversas actividades socio-económicas que se desarrollan en las distintas regiones del país. La inclusión de técnicas de aprendizaje supervisado en los modelos de pronóstico ha demostrado mejoras significativas, que varían según la región y el mes analizado. En términos generales, el ensamble de metodologías lineales, no lineales y de machine learning ha mostrado pronósticos cercanos a los valores observados, con resultados particularmente destacados en los meses de marzo, noviembre y de julio a septiembre. Sin embargo, en el Litoral, donde las precipitaciones son más abundantes, se observa una mayor dispersión en los pronósticos entre octubre y diciembre, aunque las técnicas de aprendizaje supervisado y la inclusión de procesos no lineales parecen mejorar los resultados en esta región. En el centro del país y la región de Buenos Aires, el ensamble muestra mejores resultados durante el semestre frío, cuando las precipitaciones son menores. El análisis del error por metodologías reveló que la técnica de máquinas de soporte vectorial ofrece los mejores resultados en la mayoría de los meses y regiones, con un impacto particularmente notable en la Patagonia. En el futuro se explorarán otras métricas que evalúen la precisión y confiabilidad de los pronósticos, con el objetivo de que puedan ser utilizados de manera operativa en el futuro cercano.

Agradecimientos. Los datos de precipitación fueron proporcionados por el Servicio Meteorológico Nacional (SMN) y la Autoridad Territorial de la Cuenca del Comahue (AIC).

Referencias

1. Müller, G.V., Fernández Long M.E., Bosch E. 2015. Relación entre la temperatura de la superficie del mar de diferentes océanos y los rendimientos de maíz en la pampa húmeda. *Meteorológica*, 40,1, 5-16. ISSN: 0325-187X
2. Pántano V, Penalba O, Spescha L, Murphy G. 2017. Assessing how accumulated precipitation and long dry sequences impact the soil water storage. *International Journal of Climatology*. DOI: 10.1002/joc.5087.
3. Domínguez, D. A. y González, M. H. 2013. Variabilidad de la precipitación en el centro oeste de Argentina y un modelo de predicción estadística. *Meteorologica* 38.2, 105-120. versión On-line ISSN 1850-468X
4. Gulizia, C., Camilloni, I. y Doyle, M. 2013. Identification of the principal patterns of summer moisture transport in South America and their representation by WCRP/CMIP3 global climate models. *Theor Appl Climatol* 112, 227–241. DOI: 10.1007/s00704-012-0729-4
5. Zilli, M. T., Carvalho, L.M.V. 2021. Detection and attribution of precipitation trends associated with the poleward shift of the South Atlantic Convergence Zone using CMIP5 simulations. *Int J Climatol*. 2021; 1– 22. DOI: 10.1002/joc.7007
6. Silvestri, G y Vera C. 2003. Antarctic Oscillation signal on precipitation anomalies over southeastern South America. *Geophysical Research Letters* 30 (21)
7. Gonzalez, 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*, Vol. 27, Chapter 4, 117-128. Ed. Prathamesh Gorawala y Srushti Mandhari, NOVA Science Publications, Nueva York, USA. ISBN: 978-1-53615-917-2. 260p.
8. Reichstein, M., Camps-Valls, G., Stevens, B., Jung, M., Denzler, J., Carvalhais, N., y Prabhat. 2019. Deep learning and process understanding for data-driven Earth system science. *Nature*, 566, 195-204.
9. Lund, I. A. 1963. Map pattern classification by statistical methods. *Journal of Applied Met.* 2: 56-65. DOI: 10.1175/1520-0450(1963)002<0056:MPCBSM>2.0.CO;2.
10. Kalnay, E., Kanamitsu, M., Kistler, R., Collins, W., Deaven, D., Gandin, L., Iredell, M., Saha, S., White, G., Woollen, J. y Zhu, Y., 1996. The NCEP/NCAR 40-year reanalysis project. *Bulletin of the American meteorological Society*, 77,3, 437-472. DOI: 10.1175/1520-0477(1996)077<0437:TNYRP>2.0.CO;2
11. Tibshirani, R. 1996. Regression Shrinkage and Selection via the lasso. *Journal of the Royal Statistical Society. Series B (methodological)* 58 (1). Wiley: 267–88. <http://www.jstor.org/stable/2346178>
12. Wilks, D. S., 2011. *Statistical Methods in the Atmospheric Sciences*. 3rd. ed., Academic Press, 676p. ISBN: 9780123850225.
13. Wood, S. N. 2006. *Generalized additive models – An introduction with R*. Chapman and Hall, Lon-don. 410 pp. ISBN 9781498728331
14. James G, Witten D, Hastie T, Tibshirani R (2013). *An introduction to statistical learning*. Springer, New York, p 440
15. Cortes C, Vapnik V (1995) Support-vector networks. *Mach Learn* 20:273–297. [https:// doi.org/ 10. 1007/ BF009 94018](https://doi.org/10.1007/BF00994018)
16. Hyndman, R. J. y Athanasopoulos, G. 2018. *Forecasting: principles and practice*. 2nd edition. ISBN 978-0987507105. Consultado en <https://otexts.com/fpp2/>