PLAN DE TESIS

MAESTRÍA EN EXPLOTACIÓN DE DATOS Y DESCUBRIMIENTO DEL CONOCIMIENTO

Pronóstico estadístico de precipitación mensual para la región de Gran Chaco, Argentina.

Juan Ignacio Mazza

Directora: Dra. Marcela Hebe González

Codirector: Dr. Julio Cesar Rodríguez Martino

Facultad de Ciencias Exactas y Naturales

Universidad de Buenos Aires

Tabla de contenidos

Tabla de figuras

1. Resumen

Disponer con antelación del pronóstico de lluvias para una región resulta importante a la hora de planificar y ejecutar actividades que podrían verse afectadas por estas, con el fin de minimizar o mitigar los riesgos asociados a posibles temporadas de sequía o inundación.

En la actualidad el pronóstico de lluvias para la región de Gran Chaco, Argentina, es muy eficiente a corto plazo (hasta 10 días), pero su performance disminuye para escalas superiores al mes. En regiones localizadas, una forma de abordar el pronóstico en estas escalas más grandes es la utilización de métodos estadísticos que logran aprender del pasado y generan modelos de predicción a futuro. Para esto, han sido utilizados modelos de regresión lineal múltiple que combinan varias variables atmosféricas y oceánicas. Sin embargo, dada la naturaleza no lineal y caótica de la lluvia [1], otras metodologías podrían mejorar dichos pronósticos.

El presente trabajo propone explorar diversos modelos no lineales de machine learning para pronosticar las lluvias mensuales en la región de Gran Chaco, generando una solución automatizada aplicable para cualquier región de la Argentina simplemente cambiando los archivos de predictores y de lluvias.

2. Tema

Pronostico estadístico de la precipitación mensual para la región de Gran Chaco, Argentina

3. Antecedentes sobre el tema

Dada la importancia de tener un pronóstico extendido de precipitaciones que pueda ser utilizado para diferentes actividades, como la generación de energía hidroeléctrica, la agricultura, la ganadería, la prevención de incendios por sequias, etc., resulta de interés aplicar modelos predictivos no lineales que usen como predictores las variaciones lentas de las condiciones de los océanos (temperatura de la superficie marítima) y de la atmósfera (vientos, presión atmosférica medida a diferentes alturas, cantidad de agua precipitable en el aire, etc.). Las anomalías de estas variables pueden influenciar y generar fenómenos climáticos en lugares distantes, mediante la propagación de ondas, fenómeno conocido como “teleconexiones” [2][3][4][5][6][7]. En este

sentido resulta importante el estudio de los diferentes factores que influencian la variabilidad interanual de la precipitación estacional en el área [8] y que pueden ser usados para la definición de predictores de modelos estadísticos. Aunque el fenómeno El Niño-Oscilación Sur (ENOS) sea el principal forzante remoto [9][10][11][12], no se debe descartar la influencia de otros forzantes, que adquieren relevancia sobre todo en situaciones de ENOS neutrales. Existe una clara influencia de las anomalías de temperatura de la superficie del mar en el océano Indico sobre la precipitación, ya sea a través del patrón del dipolo del Indico [13][14] como de su calentamiento o enfriamiento generalizado [15]. La acción del monzón sudamericano [16][17][18] y la variabilidad interanual de la convección sobre la selva brasilera, sobre todo en las épocas de inicio y fin del verano, también afectan a la precipitación estacional [19][20]. Las ondas de Rossby desplazándose en el Pacífico y el efecto del modo anular del Sur (SAM, por sus siglas en inglés) también son condicionantes en la precipitación de Argentina [21][22][23][24][25]. Por otro lado, la posición e intensidad del anticiclón del Atlántico también regula los procesos advectivos de humedad hacia el continente que influencian la precipitación en el norte argentino [26][27].

Varios autores han abordado este tema. [28] Zheng y Frederiksen, han logrado demostrar que existe una fuerte correlación entre la temperatura de la superficie océano Indico y las lluvias de invierno y verano en Nueva Zelanda. [2] C. J. C. Reason, demostró como las lluvias de Sudafrica se relacionan con cambios de temperatura en la superficie del océano Indico. [29] Gissila T. et al. demostraron la relación que existe entre la temperatura de la superficie del océano Indico y las lluvias de verano en Etiopía. Asimismo, [14] Chan et al, comprobaron que existe una relacion directa entre el dipolo que se genera en el oceano Indico y las precipitaciones en la Cuenca del Plata y el sur de Brasil.

En Argentina, varios autores han abordado esta problemática. Por ejemplo, [30] Oliveri P., detectó en su tesis de licenciatura que la advección de calor y humedad provenientes de los anticiclones semipermanentes del Atlántico Sur y del Pacífico Sur, influyen de forma relevante sobre la temperatura y la precipitación en escalas estacionales en Argentina.

4. Aporte esperado al finalizar el proyecto

El aporte de esta tesis de maestría es la posibilidad de contar con modelos no lineales que puedan predecir la precipitación con un mes de antelación en diferentes regiones del Gran Chaco argentino. Actualmente el pronóstico en escalas mensuales está poco desarrollado. Los modelos dinámicos, o sea basados en ecuaciones para la atmósfera, tienen en general poca eficiencia para predecir la lluvia [31][32][33] y, por lo tanto, los métodos estadísticos dan la oportunidad de modelar precipitación para usar el conocimiento de su comportamiento en el pasado. En este trabajo, se deriva un esquema mensual de pronóstico de precipitación para la región del Gran Chaco utilizando metodologías estadísticas no lineales.

5. Transferencia de los resultados

El grupo de trabajo que dirige la Marcela H. González (codirectora de esta tesis), está consolidado y trabaja conjuntamente con el Servicio Meteorológico Nacional (SMN) en la elaboración de los pronósticos trimestrales de precipitación y temperatura cada mes en reuniones de consenso realizadas para tal fin desde 2007. Además, ha firmado acuerdos específicos como, por ejemplo: en diciembre de 2015 por resolución Nº3174 del CD de la FCEN UBA, un Convenio Marco de Cooperación Académica entre la Facultad de Ingeniería de la Universidad del Comahue y la FCEN UBA donde este grupo de trabajo se compromete a la cooperación mutua para el desarrollo de modelos estadísticos de pronóstico de precipitación y temperatura para la región del Comahue, con el objetivo de mejorar la operatividad de las presas hidroeléctricas; una carta de intención de cooperación mutua con la Comisión Regional del Río Bermejo (COREBE) firmada en setiembre de 2008 entre CIMA (CONICET-UBA) y COREBE y otra que actualmente se encuentra en trámite para renovar la anterior.

Esto prueba que no solo los modelos (sino tambien todos los scripts que realizan la descarga, limpieza y calculo de predictores) resultantes de esta tesis de maestría pueden ser transferidos y utilizados por instituciones nacionales y provinciales para mejorar los pronósticos y orientar a las actividades que dependen altamente del clima

6. Objetivos de la investigación

6.1. Objetivo general

Valiéndose de datos meteorológicos de dominio público (Ver sección 7.1), esta tesis de maestría, plantea demostrar que es posible utilizar varias variables atmosféricas y oceánicas para predecir a mediano plazo (pronóstico mensual) la precipitación acumulada para la región de Gran Chaco en Argentina. El modelo utilizará los datos de las variables de reanálisis (archivos NC) de un mes y utilizarlos para predecir las lluvias de la región del mes siguiente.

6.2. Objetivos específicos

1. Regionalizar el área de estudio del Gran Chaco Argentino
2. Definir predictores para la precipitación en cada mes del año y en cada región
3. Elaborar modelos de predicción estadística para cada mes del año
4. Determinar la performance de dichos modelos

7. Hipótesis

Es posible desarrollar un modelo de machine learning que encuentre predictores entre las variables de superficie terrestre y atmósfera que puedan ser usadas para generar modelos no lineales para pronosticar a mediano plazo (un mes) las precipitaciones de la región de Gran Chaco

8. Plan de Trabajo

Esta sección del documento explica cómo se planea alcanzar el objetivo, abordado el tema desde distintos puntos de visa: Los datos, la metodología de trabajo, las métricas o criterios de éxito y un cronograma de trabajo

8.1. Los Datos

Este trabajo de maestría utiliza fuentes de datos diversas, las cuales se detallan a continuación:

8.1.1. Datos de variables meteorológicas y oceánicas a nivel global

El [34] proyecto NCEP/NCAR reanalysis utiliza tecnología de punta para recolectar datos y combinarlos con registros de satélite y radar y así generar una base de datos en puntos de retículo con resolución de 2,5°. Actualmente cuentan con datos históricos desde el año 1948 hasta la fecha para varias variables meteorológicas, de las cuales en esta tesis solo se utilizarán las siguientes:

1. Presión atmosférica a diferentes niveles de presión
    a. 200 Hpa niveles altos (HGT200)
    b. 500 Hpa niveles medios (HGT500)
    c. 1000 Hpa niveles bajos (HGT1000)
2. Temperatura de la superficie del mar (SST)
3. Cantidad de agua precipitable en la columna atmosférica (TCW)
4. Componente zonal del viento (dirección Oeste-Este) en capas bajas (U850)
5. Componente meridional del viento (dirección Sur-Norte) (V850)

Cada uno de los siete archivos mencionados contiene los datos de la variable que representa en una matriz de tres dimensiones:

1. **Latitud:** Esta se define como la medida angular de la distancia entre un punto y el Ecuador, sea hacia el norte o hacia el sur. Esta dimensión de los datos de entrada cuenta con 73 valores de latitud
2. **Longitud:** Se define como la medida angular de la distancia entre un punto y el Meridiano de Greenwich, sea hacia el este o hacia el oeste. Esta dimensión de los datos de entrada está dividida en 144 valores de longitud.
3. **Tiempo:** Una medición diaria entre el 01-01-1948 y el 31-12-2019, los datos anteriores a 1979 no serán tenidos en cuenta, debido a que solo los datos desde 1979 al presente son confiables gracias a la utilización de satélites espaciales.

8.1.2. Datos de estaciones meteorológicas

Metadatos de 34 estaciones meteorológicas correspondientes al área de Gran Chaco, Argentina y sus alrededores. La fuente de datos es el Servicio Meteorológico Nacional y el Instituto Nacional de Tecnología Agropecuaria, Las estaciones cercanas a la región de Gran Chaco se incluyen con el objeto de obtener mayor precisión en la zona de frontera de la región.

1. ID de la estación (idOMM)
2. Nombre de la estación
3. Datos de geolocalización de la estación (latitud, longitud, elevación)

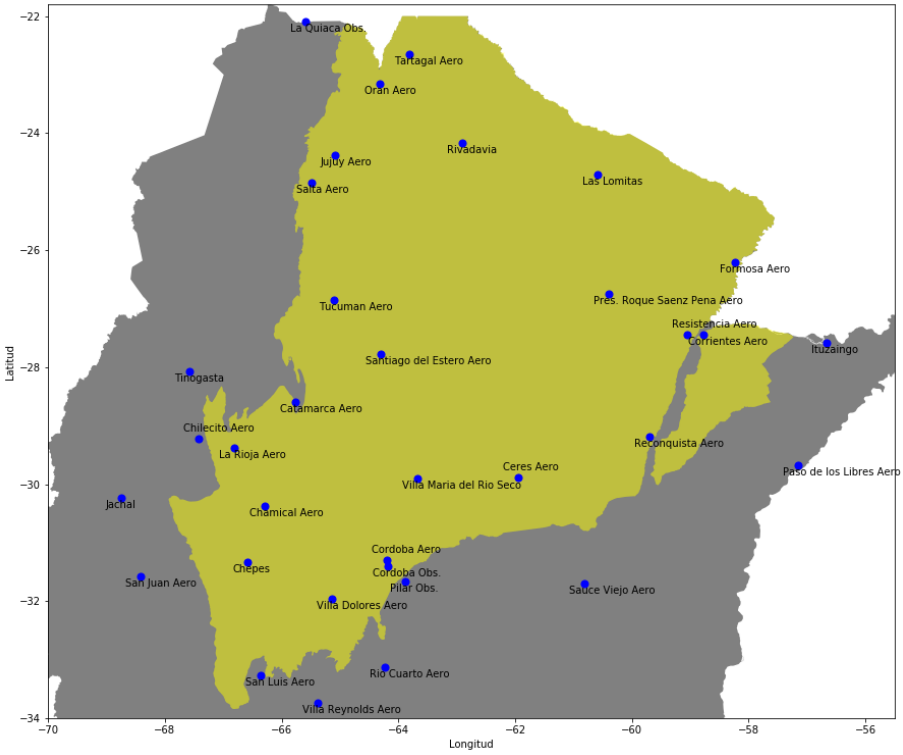


Fig. 1. *Estaciones meteorológicas de Gran Chaco, Argentina, y alrededores*

8.1.3. Datos observados de precipitación mensual acumulada.

Precipitaciones acumuladas medidas en milímetros para cada estación, datos históricos desde el año 1961 a la fecha, sin embargo, por lo explicado anteriormente, solo de utilizará el período desde 1979 hasta la actualidad.

8.1.4. Shapefiles

El formato Shapefile (SHP) es un formato de archivo informático propietario de datos espaciales desarrollado por la compañía ESRI, quien crea y comercializa software para Sistemas de Información Geográfica. Se utilizaron archivos shapefile para los polígonos de Argentina y de Gran chaco a fin de localizar sobre ellos las estaciones meteorológicas

8.2. Metodología de trabajo

Para lograr los objetivos propuestos en este plan, se proponen las siguientes actividades:

8.2.1. Actividad 1: Generación de la base de datos

Se utilizarán datos en formato NetCDF para las variables meteorológicas y oceánicas a nivel global según lo detallado en el ítem 8.1.1

Se generarán archivos de precipitación mensual para 34 estaciones meteorológicas (detalladas en la sección 8.1.2). Cabe destacar que, utilizando la ubicación de las estaciones en el mapa, el territorio que abarcan será dividido en pequeñas regiones utilizando polígonos de Voronoi, de esta manera se garantiza que cada locación del área de estudio se corresponda sólo con su estación meteorológica más cercana

8.2.2. Actividad 2: Regionalizar el área de estudio

Utilizando las precipitaciones mensuales se procederá a agrupar las estaciones meteorológicas utilizando una red neuronal no supervisada SOM (mapa auto-organizativo, por sus siglas en ingles). De esta metodología surgirán varios grupos bien diferenciados, cada uno capturado por una neurona de la red.

8.2.3. Actividad 3: Definición de predictores

Utilizando los datos de las variables meteorológicas en una matriz bidimensional de 73 x 144, se construirán las series temporales para cada mes del año en el período 1979 hasta la actualidad, en cada punto de reticulado. Además, para el mismo período se construirán las series de precipitación promedio areal para cada grupo determinado en la actividad 2.

Para cada mes, la serie de precipitación representativa de cada grupo se correlacionará con las series de variables meteorológicas en el mes previo y se obtendrán campos espaciales de correlación para cada mes del año, para cada variable meteorológica y para cada grupo. En ellos se determinarán las áreas con correlación significativa (95% de confianza) utilizando un test normal. Cuando un punto supere un cierto umbral de correlación, la correlación se tomará como significativa, y cuando varios puntos aledaños posean una correlación significativa, el área en cuestión se marcará como significativa.

A posteriori se definirán como predictores a las variables promediadas en dichas zonas que resultaron significativas y que además respondan a un fenómeno físico que las sustente.

8.2.4. Actividad 4: Diseño de los modelos

Se alimentarán uno o mas modelos de machine learning utilizando los predictores definidos con variables meteorológicas en un mes que se correlacionaron significativamente con las lluvias del mes siguiente y que fueron definidos en la actividad 3.

Se utilizarán modelos de redes neuronales artificiales, y opcionalmente, algún otro modelo (XGBoost o Support Vector Regression) a fin de establecer una comparación y realizar un ensamble

En todos los casos, se buscará para cada modelo propuesto el mejor conjunto de hiperpárametros, es decir, aquellos valores que maximizan la performance del modelo, para ello se utilizarán técnicas de optimización vistas en la maestría como grid-search u optimización bayesiana.

Los modelos predictivos estimarán el valor de la precipitación mensual (regresión), estos valores de lluvia predichos serán luego discretizados en 3 categorías:

1. Precipitación por debajo de lo normal (1er tercil)
2. Precipitación normal (2do tercil)
3. Precipitación por encima de lo normal (3er tercil)

8.2.5. Actividad 5: Medida de la performance de los modelos

Los modelos generados se entrenarán utilizando 30 años de datos (1979 – 2008) y se evaluaran contra un set de datos de Testing formado por los años (2009 - 2019). Se evaluará la eficiencia de los modelos comparándolos con la precipitación observada.

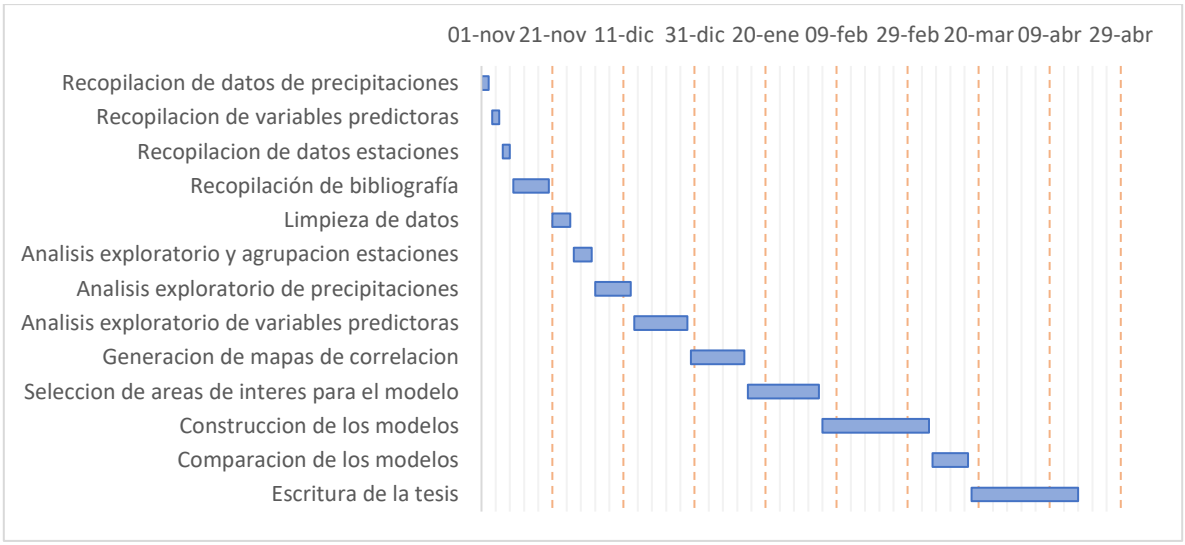
Para todos los modelos generados se evaluará la varianza explicada.

La precipitación observada también se clasificará en las 3 categorías con terciles para poder construir tablas de contingencia de precipitación pronosticada y observada. En función de ellas podrá evaluarse la probabilidad de detección, el error general, la relación de falsa alarma [35]

Se evaluará la posibilidad de calcular otras métricas adicionales como: la Matriz de confusión luego de discretizar, las curvas ROC y AUC, error rate, precisión, recall, F-Score, etc.

8.3. Cronograma de Trabajo

El siguiente diagrama de Gantt muestra una estimación de los tiempos del proyecto. Algunas de las siguientes actividades ya se encuentran en marcha al momento de presentar este documento.



Fíg. 2. Diagrama de Gantt del proyecto de tesis

8.4. Disponibilidad de infraestructura

Se detallan a continuación los distintos elementos de infraestructura que le dan soporte a esta tesis.

8.4.1. Hardware

Todo el preprocesamiento de los datos, generación de mapas, generación de predictores, entrenamiento y validación de los modelos de machine learning, etc. Será realizado en una notebook doméstica. Se utilizará una laptop Lenovo Yoga con procesador Intel Core I7 séptima generación.

8.4.2. Software

Todo el software, lenguajes de programación, y librerías utilizadas en el presente trabajo son gratuitas y de libre acceso público. Se detallan a continuación los mismos:

1. Python 3.7: Lenguaje de programación
2. Pandas: Manipulación de datos y data frames
3. Numpy: Manipulación de números, matrices y vectores
4. GeoPanads: Georreferenciación y manipulación de los shape files
5. Matplotlib: Librería gráfica, generación de mapas de predictores
6. GeoVoronoi: Calculo de polígonos de Voronoi
7. Xlrd: Manipulación de archivos en formato Excel
8. cv2: Librería gráfica, detección de superposición de capas y formas en imágenes
9. netCDF4: lectura y manipulación de archivos en formato netCDF
10. Tensorflow y Keras: Generacion de modelos de redes neuronales
11. Otros paquetes

8.4.3. Otros

Se cuenta además de lo mencionado anteriormente con conexión a internet de alta velocidad, acceso a diversas fuentes de información como Wikipedia y repositorios de papers y libros, almacenamiento en la nube, etc.

9. Referencias

[1] A. Mary, “Deterministic chaos , fractals , and quantumlike mechanics in atmospheric flows,” no. 21, 1989.

[2] C. J. C. Reason, “Subtropical Indian Ocean SST dipole events and southern South African rainfall,” vol. 28, no. 11, pp. 2225–2227, 2001.

[3] D. Karoly and V. Dayton, *Meteorology in Southern Hemisphere*. 1999.

[4] J. E. Oliver and J. J. Hidore, “Climatology, an introduction,” *Merrill Publ. Co.*, 1984.

[5] J. P. Peixdto, “Physics of climate,” 1984.

[6] J. R. Holton and R. Dmowska, *El Niño, La Niña, and the southern oscillation*. Academic press, 1989.

[7] M. L. Salby, *Physics of the Atmosphere and Climate*. 2012.

[8] A. Rolla and M. Gonzalez, “Some precipitation patterns that affect agricultural practices in the planes of Buenos Aires (Argentina),” *Agric. Res. Updat.*, vol. 22-cap 8, pp. 209–238.

[9] K. E. Trenberth and T. J. Hoar, “The 1990-1995 El Nifio-Southern Oscillation event: Longest on record,” vol. 23, no. 1, pp. 57–60, 1996.

[10] C. F. Ropelewski and M. S. Halpert, “Global and Regional Scale Precipitation Patterns Associated with the El Niño-Southern Oscillation.” 1987.

[11] C. Vera, G. Silvestri, V. Barros, and A. Carril, “Differences in El Nino Response over the Southern Hemisphere,” no. 1999, pp. 1741–1753, 2004.

[12] A. M. Grimm, “Interannual climate variability in South America : impacts on seasonal precipitation , extreme events , and possible effects of climate change,” pp. 537–554, 2011.

[13] N. H. Saji and P. N. Vinayachandran, “A dipole mode in the tropical Indian Ocean,” vol. 401, no. September, pp. 360–363, 1999.

[14] S. Chan, S. Behera, and T. Yamagata, “Indian Ocean Dipole influence on South American rainfall : Climatic impacts of Indian Ocean dipoles, El Nino-Southern oscillation, and their interaction with the monsoon systems in the Asia-Oceania region,” *Geophys. Res. Lett.*, vol. 35, no. 14, 2008.

[15] T. Andréa and T. Ambrizzi, “Can Indian Ocean SST anomalies influence South American rainfall ?,” pp. 1615–1628, 2012.

[16] V. E. Kousky, “Precipitation and atmospheric circulation anomaly patterns in the South American sector,” vol. 3, 1988.

[17] C. F. Gan, M. A., Kousky, V. E., & Ropelewski, “The South America Monsoon Circulation and Its Relationship to Rainfall over West-Central Brazil,” no. 1998, pp. 47–66, 2004.

[18] M. A. Gan, V. B. Rao, and M. C. L. Moscati, “South American monsoon indices,” vol. 223, no. July 1979, pp. 219–223, 2006.

[19] M. Gonzalez, J. Nery, and V. Barros, “Caracteristicas de la precipitacion en argentina subtropical y Brasil meridional y de la conveccion tropical,” no. 1, pp. 1–5, 1996.

[20] V. Barros and M. Gonzalez, “Climate variability over subtropical South America and the Southamerican monsoon: a review,” vol. 27, pp. 33–57, 2002.

[21] J. Kalnay, E., Mo, K. C., & Paegle, “Large-Amplitude, Short-Scale Stationary Rossby Waves in the Southern Hemisphere: Observations and Mechanistic Experiments to Determine their Origin.” .

[22] J. W. Kidson, “Principal Modes of Southern Hemisphere Low-Frequency Variability Obtained from NCEP – NCAR Reanalyses,” pp. 2808–2830, 1999.

[23] J. A. Marengo *et al.*, “Recent developments on the South American monsoon system,” vol. 21, no. December 2010, pp. 1–21, 2012.

[24] G. E. Silvestri and C. S. Vera, “Antarctic Oscillation signal on precipitation anomalies over southeastern South America,” vol. 30, no. 21, pp. 1–4, 2003.

[25] M. Gonzalez, “Some indicatros of interannual rainfall variability in Patagonia (Argentina),” *Clim. Var. - Reg. Temat. patterns*, vol. 6, pp. 133–161.

[26] E. M. Garbarini, “Algunos indicadores para la predicción estadística de la precipitación estacional en Argentina. Tesis de Licenciatura,” Departamento de Ciencias de la Atmósfera y los Océanos, Facultad de Ciencias Exactas y Naturales, Universidad de Buenos Aires., 2016.

[27] E. M. Garbarini, “The influence of Atlantic High on season rainfall in Argentina,” *Int. J. Climatol. #39*, vol. 12, pp. 4688–4702.

[28] Z. Xiaogu and F. Carsten S, “A Study of Predictable Patterns for Seasonal Forecasting of New Zealand Rainfall,” no. Madden 1976, pp. 3320–3333, 2006.

[29] T. Gissila, E. Black, D. I. F. Grimes, and J. M. Slingo, “Seasonal forecasting of the ethiopian summer rains,” vol. 1358, pp. 1345–1358, 2004.

[30] P. Oliveri, “La influencia de los océanos cercanos sobre la precipitación y temperatura estacionales en Argentina. Tesis de Licenciatura en Ciencias de la Atmósfera, Unversidad de Buenos Aires.,” *Tesis UBA*, 2018.

[31] D. Rostkier-Edelstein *et al.*, “High-resolution forecasts of seasonal precipitation: a combined statistical-dynamical downscaling approach,” *Ann. EGU*, 2010.

[32] G. T. Diro, A. M. Tompkins, and X. Bi, “Dynamical downscaling of ECMWF Ensemble seasonal forecasts over East Africa with RegCM3,” *J. Geophys*, 2012.

[33] L. Yuan, L. Guihua, W. Zhiyong, and H. H., “High-Resolution Dynamical Downscaling of Seasonal Precipitation Forecasts for the Hanjiang Basin in China Using the Weather Research and Forecasting Model,” *J. Appl. Meteorol. Climatol.*, vol. 56, pp. 1515–1535, 2017.

[34] Kalnay et al, “The NCEP/NCAR 40-year reanalysis project,” 1996. [Online]. Available: https://www.esrl.noaa.gov/psd/data/gridded/data.ncep.reanalysis.html.

[35] D. S. Wilks, “Statistical Methods in the Atmospheric Sciences,” vol. 95, no. 449, pp. 344–345, 2014.