DESARROLLO DE UN MODELO DE MACHINE LEARNING PARA EL PRONÓSTICO METEOROLÓGICO DE PRECIPITACIÓN A ESCALA SUBESTACIONAL EN PERÚ

Brayan Urbina[1], Ken Takahashi[1]

[1] Instituto Geofísico del Perú (IGP), Lima, Perú

Palabras clave: Machine Learning, precipitación, downscaling, subestacional

Citar como Urbina, B., & Takahashi, K. (2023). Desarrollo de un modelo de Machine Learning para el pronóstico meteorológico de precipitación a escala subestacional en Perú. *Boletín científico El Niño*, Instituto Geofísico del Perú, vol. 10 n.° 09, págs. 13-17.

Resumen

Este trabajo muestra los avances de la aplicación del Machine Learning (ML) para predecir las precipitaciones en Perú a escala subestacional con hasta seis semanas de anticipación. Para ello, se utiliza el proyecto Subseasonal to Seasonal (S2S) como insumo para aplicar el enfoque de *downscaling* y mejorar la precisión y la resolución espacial de las predicciones teniendo como referencia el producto PISCOp V2.1 del SENAMHI. Luego de procesar los datos, se realiza un Análisis de Componentes Principales y se desarrollan modelos de regresión lineal múltiple para predecir los principales componentes observados utilizando los preestablecidos a partir de NCEP CFSv2 como predictores para cada día de pronóstico. Los resultados muestran un buen desempeño de las predicciones hasta 5 días de anticipación, especialmente en la región amazónica, pero la habilidad de pronóstico disminuye más allá de los cinco días debido a la complejidad de los factores que influyen en las precipitaciones y debido también a la simplicidad del modelo de regresión lineal. Se plantea la posibilidad de incorporar variables adicionales relacionadas con las oscilaciones de Madden-Julian (MJO, por sus siglas en inglés) en futuras versiones del modelo para ampliar el horizonte de pronóstico.

1. Introducción

El Perú tiene un territorio heterogéneo que comprende la Amazonía, la cordillera de los Andes y una costa árida. Está influenciado por una diversidad de fenómenos meteorológicos y climáticos, como El Niño y, particularmente, por fenómenos subestacionales tropicales, en escalas de algunas semanas, como la oscilación Madden-Julian en la atmósfera y las ondas Kelvin ecuatoriales oceánicas. Esto crea una oportunidad importante para la implementación de servicios climáticos con importantes beneficios sociales y económicos significativos, en particular para aquellos asociados con la modelización hidrológica y agrometeorológica, los cuales requieren pronósticos con plazos de entrega de datos de hasta algunas semanas.

Para una buena simulación y pronóstico numérico de los sistemas atmosféricos tropicales subestacionales es necesario representar adecuadamente la fuerte interacción con la convección. Sin embargo, los Modelos de Circulación General (MCG) tergiversan la precipitación convectiva y la temperatura de la superficie del mar en el océano Pacífico oriental (Seager et al., 2019), y no resuelven adecuadamente

las montañas de los Andes para sus aplicaciones. Si bien la reducción de escala utilizando modelos oceánicos-atmosféricos regionales de alta resolución podría ser una forma de superar este problema, la necesidad de conjuntos de múltiples miembros y múltiples modelos hace que este enfoque no sea práctico. Por lo tanto, la Organización Meteorológica Mundial recomienda aprovechar los datos de pronóstico subestacionales fácilmente disponibles de los MCG y utilizar modelos de inteligencia artificial/aprendizaje automático como una forma hábil y altamente eficiente de reducir los pronósticos para las aplicaciones de servicios climáticos.

En tal sentido, el objetivo de este trabajo es utilizar el Machine Learning (ML) para predecir las precipitaciones en el territorio peruano a escala subestacional con hasta 6 semanas de anticipación.

2. Metodología

Para esta investigación utilizamos uno de los modelos miembros del proyecto Subseasonal to Seasonal (S2S; Vitart et al., 2017) como insumo para nuestro modelo de ML. De esta manera, el enfoque utilizado para este sistema de pronóstico es el *downscaling*, donde tendremos la variable de lluvias pronosticadas desde el NCEP CFSv2 (National Center for Environmental Prediction-Climate Forecast System Version 2) en una resolución más baja para luego transformarlas a una escala local (Perú) de mayor exactitud de pronóstico y mayor resolución espacial.

En resumen, se representará la variabilidad de las precipitaciones observadas y pronosticadas por NCEP CFSv2 para el territorio peruano mediante principales componentes y se desarrollarán modelos de regresión lineal múltiple para predecir cada componente principal observada y cada tiempo de pronóstico (lead) usando los principales componentes pronosticados por NCEP CFSv2 como predictores.

2.1 Fuentes de datos

Los datos de precipitación considerados como objetivo de pronóstico (predictando) fueron obtenidos a través del producto de precipitación grillado PISCOp V2.1 (*peruvian interpolated data of* SENAMHI'S *Climatological and Hydrological Observations*) de alta resolución del SENAMHI, que utiliza un algoritmo basado en métodos de interpolación a partir de datos de pluviómetros en Perú e información satelital para convertirla en información de alta resolución (Aybar et al., 2020). Este producto proporciona registros de lluvias en todo el territorio nacional a intervalos regulares diarios. Además, como fuente de previsibilidad (predictores), se utilizaron datos de precipitación del pronóstico subestacional de hasta 6 semanas del modelo climático océano-atmósfera global del NCEP CFSv2 (Saha et al., 2014). Los datos obtenidos tienen resolución espacial de 1° × 1° y con resolución temporal diaria, y abarcan el periodo entre 1999 y 2021 en todos los datos.

2.2 Procesamiento de datos

Como se aprecia en el flujo de la Figura 1, el primer paso fue el procesamiento de los datos. Para ello, se realizó (1) la descarga de los datos en el periodo del caso de estudio para luego seleccionar el dominio de interés, (2) se calcularon las anomalías con el periodo base climatológico 1999-2010 y (3) se utilizó el filtro pasa-banda Butterworth (Butterworth, 1930) de orden 1 y periodos de corte entre los 20 y 60 días. Esto último se realizó para obtener la señal subestacional tanto de las precipitaciones pronosticadas como de las observadas, de manera que el ajuste del modelo de ML se pueda enfocar mejor en los patrones relevantes en esta escala.

El segundo paso fue hacer un Análisis de Componentes Principales (Wold et al., 1987), donde se seleccionan los componentes más representativos de la variabilidad, tanto para los predictores como los predictandos. Esto reduce la dimensionalidad de los datos y disminuye a su vez el tamaño del modelo necesario, lo cual es importante para evitar que el modelo se sobreajuste a patrones espurios.

Posteriormente, además de hacer la normalización estándar, se guardan los patrones espaciales de los Principales Componentes (PCs) o EOFs de PISCOp V2.1. Finalmente, el *dataset* se divide en dos grupos. El primer grupo de datos se utiliza para el entrenamiento, abarcando los períodos 1999-2010

y 2015-2017, mientras que el segundo grupo de datos se utiliza para validar el modelo, abarcando el periodo 2018-2021. En el caso de los datos de PISCOp V2.1, se consideran los primeros 13 PCs, los cuales explican el 90 % de la varianza.

2.3 Desarrollo del método

Una vez preparados los datos (según la Figura 1), se procede a realizar la selección de predictores mediante un *backward elimination* (Ferri et al., 1994) de los PCs calculados previamente. Este proceso de selección consiste en entrenar primero un modelo de regresión lineal múltiple con todos los posibles predictores para, seguidamente, hacer una evaluación del modelo tentativo mediante la métrica RMS (*root mean square*) y luego comparar con el modelo que resulta de quitar uno de los predictores, repitiendo esto último con cada predictor y eliminando la versión del modelo con peor desempeño según la métrica seleccionada. Esto se repite hasta finalmente obtener la mejor combinación de predictores.

El producto final del modelo es la anomalía de precipitación subestacional a paso diario sobre todo el territorio peruano en la grilla de PISCOp V2.1, la cual se construye mediante la suma de los PCs pronosticadas multiplicadas por los patrones espaciales de los EOFs correspondientes.

3. Resultados

Siguiendo el esquema de la Figura 1, después de obtener el modelo seleccionado (*predictive model*), este se utiliza para realizar predicciones con el *dataset* independiente de prueba (*test data*) y ver cuál es la habilidad del modelo comparando los PCs con lo observado según lo calculado con el producto PISCOp V2.1 para ese periodo. Luego se hace la reconstrucción de los mapas de precipitación utilizando los EOFs almacenados en la etapa de procesamiento.

En resumen, el flujo metodológico de la Figura 1 funciona para cada *lead* (día de pronóstico desde una fecha de referencia) de pronóstico del NCEP CFSv2. Por lo tanto, podemos obtener predicciones de lluvias teóricamente con hasta 43 días de anticipación.

La Figura 2 resume el rendimiento para cada *lead* mediante la correlación obtenida entre los PCs pronosticados del modelo de regresión lineal (modelo de *downscaling*) y los datos observados de PISCOp V2.1 en la data de prueba (*test data*). De acuerdo con esta figura, los PCs 1, 3 y 6, enfocados principalmente en la región amazónica y parte de la sierra oriental, tienen el mejor rendimiento en el eje de los *leads*. Las PCs 2, 5 y 7 están más enfocadas en la sierra y en algunas zonas de la Amazonía, aunque pierden rendimiento a los pocos días de pronóstico (3 *leads* en promedio).

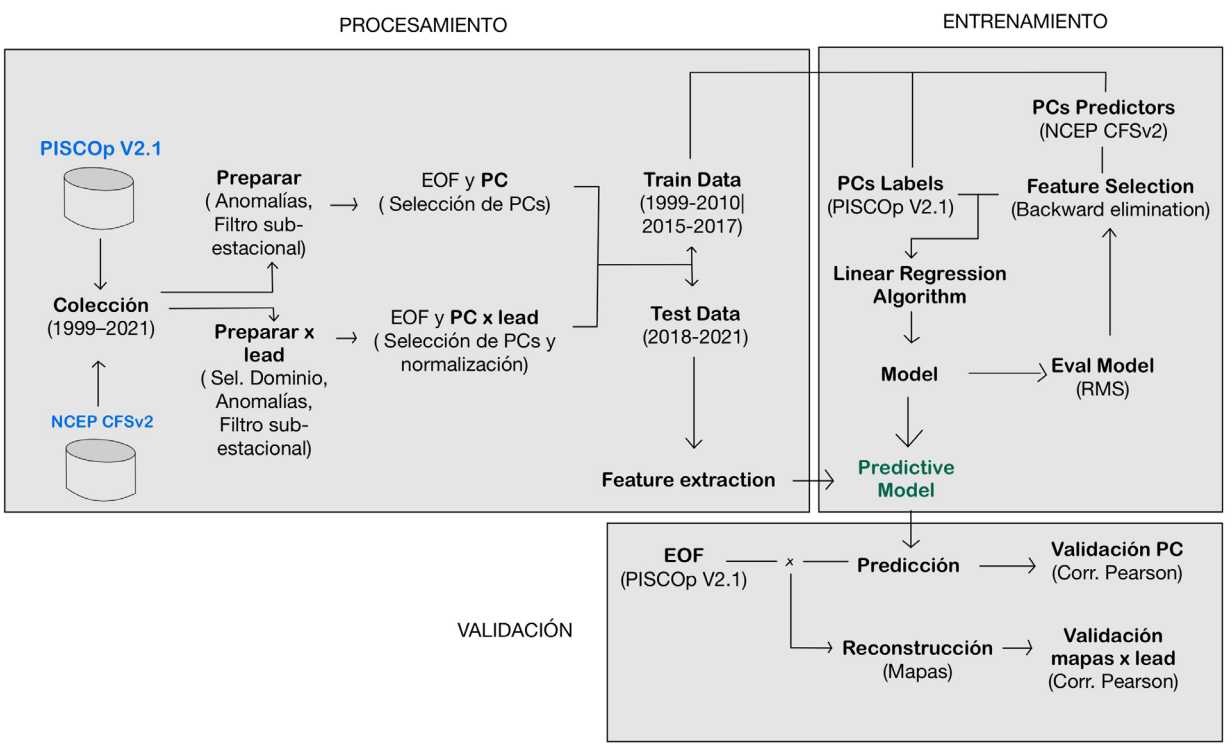


Figura 1. Diagrama de flujo del modelo de regresión lineal para cada *lead*. Se muestran los pasos involucrados en el procesamiento, entrenamiento y validación.

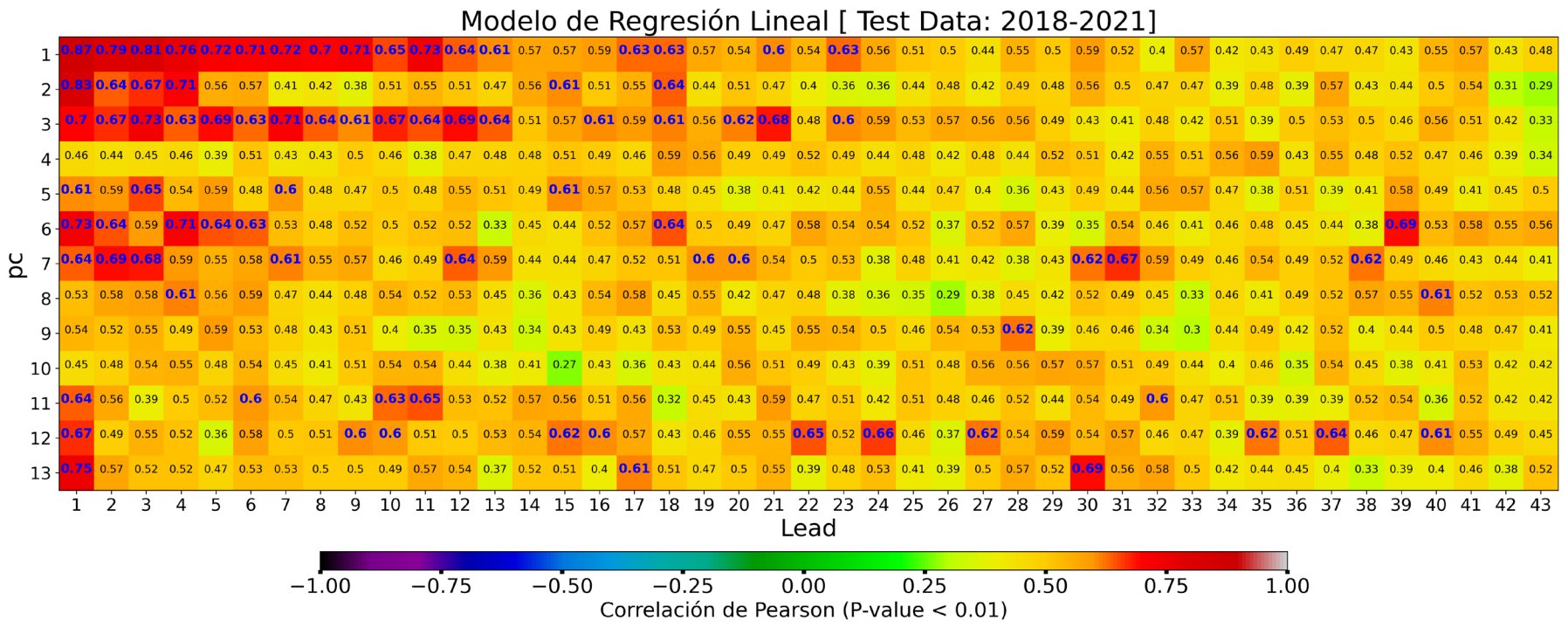
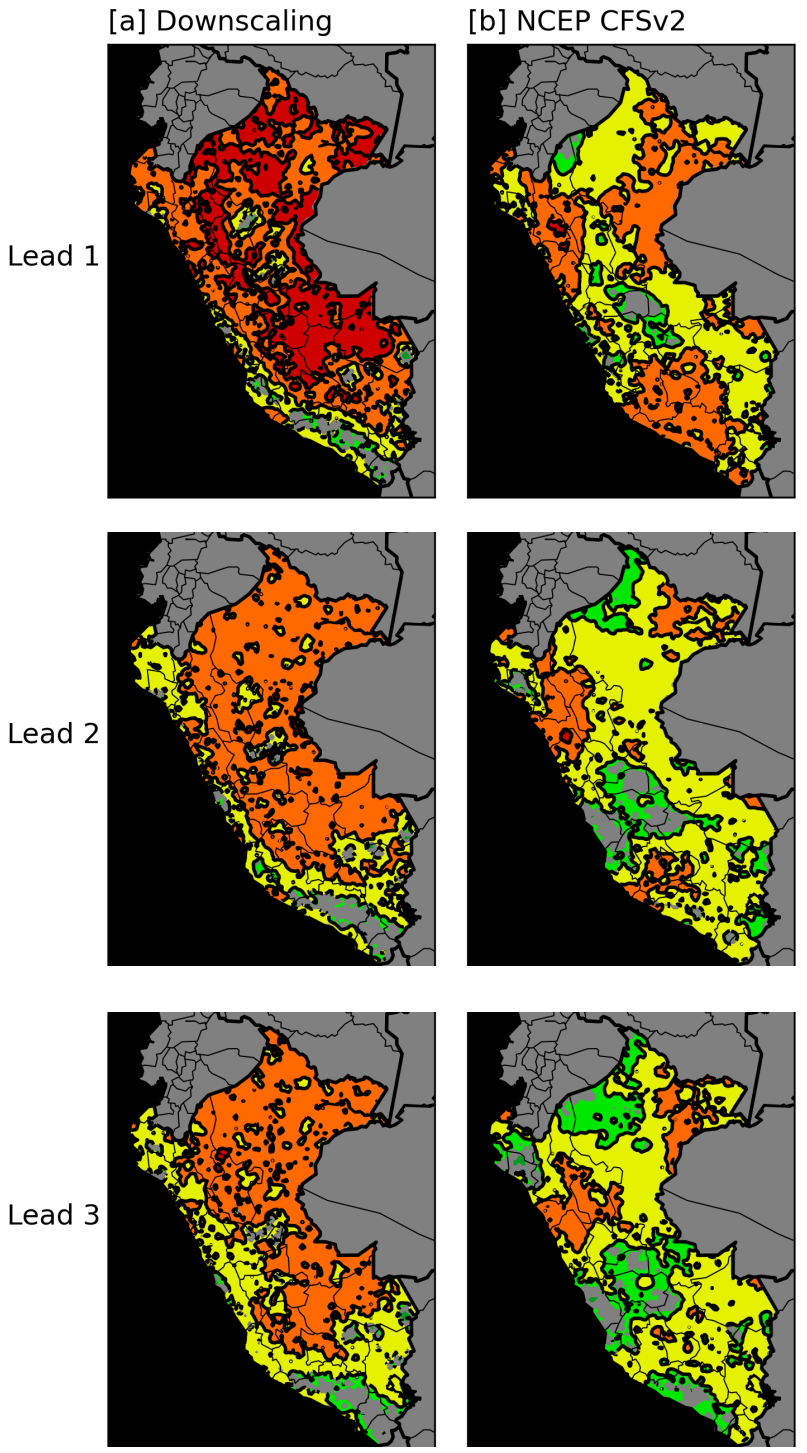


Figura 2. Tabla de valores de los coeficientes de correlación de los PCs versus cada *lead* de pronóstico. Se resaltan en morado los valores por encima de 0.6.

Dado que el objetivo final es comparar los resultados de previsión de nuestro modelo de *downscaling* con el producto de NCEP CFSv2, se obtienen correlaciones en todo el Perú. La Figura 3 muestra que el mejor rendimiento en el área de estudio se obtiene para el modelo *downscaling*, sobre todo en la mayor parte de la región amazónica hasta el *lead* 5 de pronóstico. También se aprecia un resultado por encima del valor de correlación positivo (>0.25) para la sierra del Perú y parte de la costa central y norte, mientras que el producto NCEP CFSv2 no tiene una habilidad adecuada de pronóstico (correlación por debajo de 0.25) en la zona centro del Perú. Por otro lado, la habilidad en el sur disminuye rápidamente desde el primer día de pronóstico (*lead* 1) y en gran parte de la zona norte, excepto para un área cercana a la región La Libertad.

Entonces, podemos concluir que la variabilidad de los patrones de EOFs localizada en región amazónica del producto PISCOp V2.1 responde adecuadamente y de forma lineal a algunos PCs predichos por el modelo climático, al menos para los primeros *leads* de pronóstico. Esto, presumiblemente, se debe a que algunos mecanismos responsables de las precipitaciones en esa región son bien reproducidos por el producto NCEP CFSV2. Por último, ninguno de los modelos resuelve bien el pronóstico de lluvias subestacionales, para ninguna región, más allá de los 5 días de pronóstico.



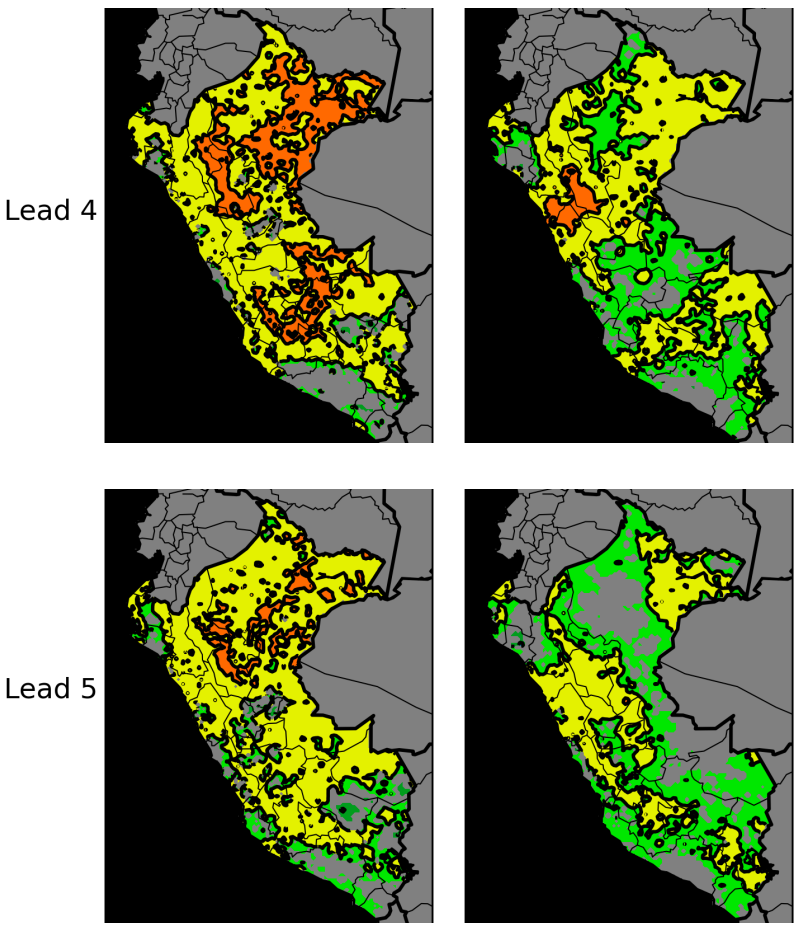


Figura 3. Mapas de correlación de Pearson hasta el *lead 5* (día cinco de pronóstico): entre el Producto PISCOp V2.1 y la reconstrucción de las predicciones del modelo de regresión lineal (izquierda) y, lo mismo que el anterior, pero evaluando con los pronósticos de precipitación directos y previamente regrillados del producto NCEP CFSv2 (derecha).

4. Discusión

La limitación en el desempeño del pronóstico de *downscaling* más allá de los 5 días se debe posiblemente a la complejidad del desarrollo de la precipitación, ya que esta está influenciada por diversos factores tanto oceánicos como atmosféricos que no han podido ser adecuadamente capturados por el modelo climático. Asimismo, debemos tener en cuenta que el modelo de regresión lineal, por su misma simplicidad, no puede resolver comportamientos no lineales posiblemente existentes entre los predictores y la variable objetivo.

Por último, considerando que las oscilaciones de Madden-Julian (MJO), en principio, nos ofrecen una oportunidad de predicción a mayor plazo, es posible que la incorporación de variables predictoras adicionales asociadas a estas puedan ampliar el horizonte de pronóstico. Esto se evaluará en las siguientes versiones del modelo de *downscaling*.

Referencias

Aybar, C., Fernández, C., Huerta, A., Lavado, W., Vega, F., & Felipe-Obando, O. (2020). Construction of a high-resolution gridded rainfall dataset for Peru from 1981 to the present day. *Hydrological Sciences Journal*, 65, 770-785.

Butterworth S. (1930). "On the Theory of Filter Amplifiers". Experimental Wireless and the Wireless Engineer, 7, 536-541.
Ferri, F.J., Pudil, P., Hatef, M., & Kittler, J. (1994). Comparative Study of Techniques for Large-Scale Feature Selection. *Pattern Recognition in Practice* IV, 16, 403-413.

Saha, S., Moorthi, S., Wu, X., Wang, J., Nadiga, S., Tripp, P., Behringer, D., Hou, Y., Chuang, H., Iredell, M., Ek, M., Meng, J., Yang, R., Mendez, M. P., van den Dool, H., Zhang, Q., Wang, W., Chen, M., & Becker, E. (2014). The NCEP Climate Forecast System Version 2. *Journal of Climate*, 27, 2185-2208. https://doi.org/10.1175/JCLI-D-12-00823.1

Seager, R., Cane, M., Henderson, N., Lee, D.-E., Abernathey, R., & Zhang, H. (2019). Strengthening tropical Pacific zonal sea surface temperature gradient consistent with rising greenhouse gases. *Nature Climate Change*, 9(7), 517–522. https://doi.org/10.1038/s41558-019-0505-x

Vitart, F., Ardilouze, C., Bonet, A., Brookshaw, A., Chen, M., Codorean, C., Déqué, M., Ferranti, L., Fucile, E., Fuentes, M., Hendon, H., Hodgson, J., Kang, H., Kumar, A., Lin, H., Liu, G., Liu, X., Malguzzi, P., Mallas, I., Manoussakis, M., Mastrangelo, D., MacLachlan, C., McLean, P., Minami, A., Mladek, R., Nakazawa, T., Najm S., Nie, Y., Rixen, M., Robertson A. W., Ruti, P., Sun, C., Takaya, Y., Tolstykh, M., Venuti, F., Waliser, D., Woolnough, S., Wu, T., Won, D-J., Xiao, H., Zaripov, R., & Zhang, L. (2017). The Subseasonal to Seasonal (S2S) Prediction Project Database. *Bulletin of the American Meteorological Society*, 98(1), 163-173. https://doi.org/10.1175/BAMS-D-16-0017.1

Wold, S., Esbensen, K. & Geladi, P. (1987). Principal Component Analysis. *Chemometrics and Intelligent Laboratory Systems*, 2, 37-52. http://dx.doi.org/10.1016/0169-7439(87)80084-9