XXXI CONGRESO LATINOAMERICANO DE HIDRÁULICA
MEDELLÍN-COLOMBIA OCTUBRE 01-04 2024

Imputación de datos ambientales: marco basado en aprendizaje automático con restricciones físicas

Martina POU[1], Marcos PASTORINI[2], Rafael RODRÍGUEZ[1], Lorena ETCHEVERRY[2], Alberto CASTRO[2] y Angela GORGOGLIONE[1]

[1] Instituto de Mecánica de los Fluidos (IMFIA), Facultad de Ingeniería, Universidad de la República, Uruguay
email: mpou@fing.edu.com

[2] Instituto de Computación (INCO), Facultad de Ingeniería, Universidad de la República, Uruguay

RESUMEN

La estimación de datos faltantes en series temporales ambientales es crucial para representar con exactitud procesos naturales a nivel de cuenca. Diversas técnicas, desde imputaciones estadísticas hasta métodos de aprendizaje automático, se han explorado para abordar este problema complejo. En particular, las técnicas de aprendizaje automático supervisado pueden representar de manera efectiva las relaciones no lineales entre variables medidas en estaciones espacialmente distribuidas. En este trabajo se presenta un marco metodológico (*framework*) para la imputación de datos ambientales que incluye datos meteorológicos, hidrológicos y de calidad de agua. Este marco implementa diferentes modelos de aprendizaje automático y restricciones físicas que consideran la alta variabilidad espacial y temporal de las variables imputadas. Los resultados muestran un *framework* que atiende exitosamente al desafío de imputar datos faltantes en varios dominios ambientales. Más del 75% de todos los datos imputados se caracteriza por NSE>0,45 (resultados satisfactorios). El desarrollo y la implementación de este *framework* representan un avance significativo en la gestión de datos ambientales, proporcionando una metodología eficiente y efectiva para enfrentar uno de los desafíos más persistentes en el campo de la ciencia ambiental.

1. Introducción

La estimación de datos faltantes en series temporales ambientales es crucial para representar con exactitud procesos naturales a nivel de cuenca (Gorgoglione et al., 2020). La exactitud en la representación de estos procesos es esencial para la gestión de recursos hídricos, la predicción de eventos extremos y la formulación de políticas ambientales efectivas.

Diversas técnicas, desde imputaciones estadísticas hasta métodos de aprendizaje automático, se han explorado para abordar este problema complejo. Las técnicas estadísticas incluyen métodos como la media, la mediana, o la interpolación lineal, que son simples y fáciles de implementar, pero a menudo carecen de la capacidad para capturar la complejidad de los datos ambientales (Chen et al., 2021). En contraste, los métodos de aprendizaje automático, que abarcan desde algoritmos de regresión hasta redes neuronales profundas, ofrecen una capacidad superior para modelar relaciones complejas y no lineales entre las variables medidas. En particular, las técnicas de aprendizaje automático supervisado pueden representar de manera efectiva las relaciones no lineales entre variables medidas en estaciones espacialmente distribuidas (Chivers et al., 2020). Estas técnicas pueden manejar la complejidad inherente a los datos ambientales, que a menudo presentan una alta variabilidad espacial y temporal.

La integración de datos ambientales de dominios interconectados, como la meteorología, la hidrología y la calidad de agua, es esencial para mejorar la precisión de las técnicas de imputación y para una comprensión más profunda de la dinámica de la calidad del agua a nivel de cuenca. Esto se debe a que los procesos naturales

están intrínsecamente ligados y las interacciones entre diferentes componentes del medio ambiente son cruciales para obtener una representación completa y precisa.

En este trabajo se presenta un marco metodológico (*framework*) para la imputación de datos ambientales que incluye datos meteorológicos, hidrológicos y de calidad de agua. Este marco implementa diferentes modelos de aprendizaje automático y restricciones físicas que consideran la alta variabilidad espacial y temporal de las variables imputadas. Además, este *framework* está diseñado para gestionar de manera efectiva un alto porcentaje de valores faltantes, lo que lo hace particularmente útil en escenarios donde los datos pueden ser escasos o estar incompletos. El enfoque propuesto no solo mejora la precisión de las imputaciones, sino que también proporciona una herramienta robusta para la investigación y la gestión ambiental a nivel de cuenca.

El enfoque propuesto no solo mejora la precisión de las imputaciones, sino que también proporciona una herramienta robusta para la investigación y la gestión ambiental a nivel de cuenca. Al integrar múltiples fuentes de datos y aplicar técnicas avanzadas de modelado, este *framework* ofrece una solución innovadora para uno de los desafíos más persistentes en la ciencia ambiental.

2. Área de estudio

El área de estudio es la cuenca del río Santa Lucía Chico (Figura 1). Esta cuenca abarca una superficie de 2570 km² y se encuentra ubicada en la zona centro-sur de Uruguay, específicamente en el departamento de Florida. Es una región de gran relevancia a nivel nacional debido a que en ella se encuentra el principal reservorio que abastece de agua potable aproximadamente a la mitad de la población uruguaya. Además, esta cuenca es fundamental para diversas actividades agrícolas e industriales que se desarrollan en la región (Vilaseca et al., 2023).

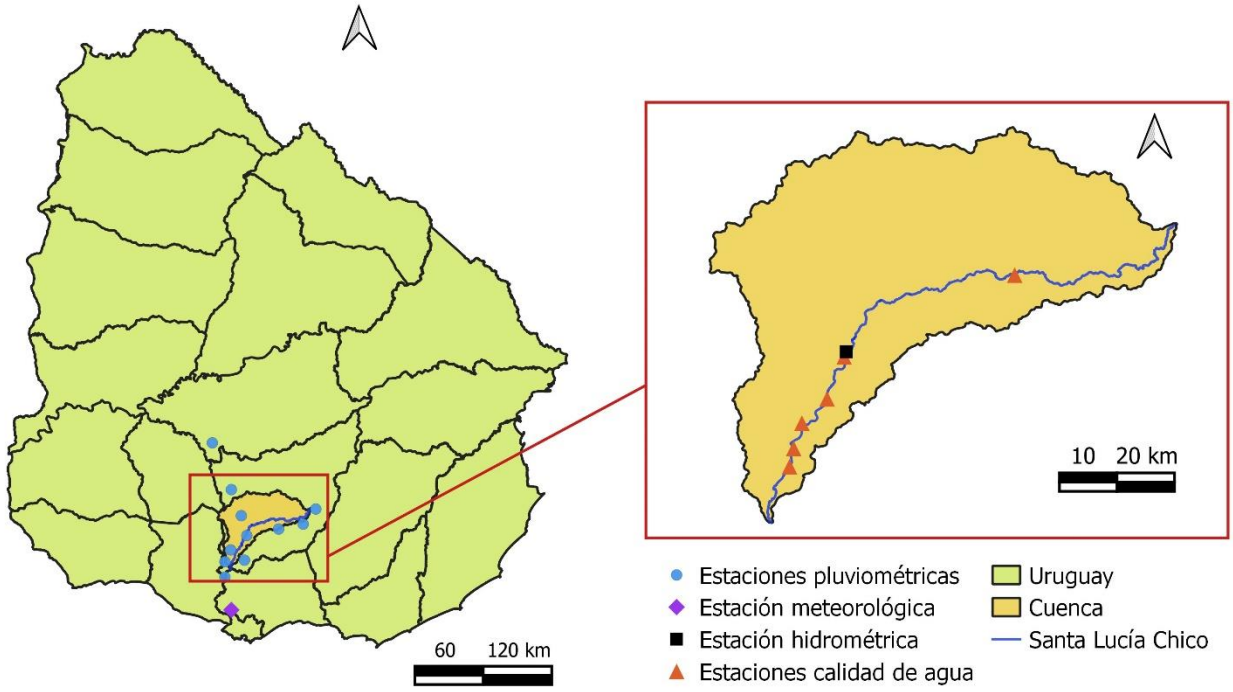

Figura. 1. Zona de estudio y ubicación de las estaciones de monitoreo.

La cuenca del río Santa Lucía Chico no solo es crucial por su papel en el suministro de agua potable, sino también por su contribución a la economía local y nacional. En términos agrícolas, la región es conocida por la producción de cultivos como soja, maíz y trigo, así como por la cría de ganado bovino y ovino. Estas actividades dependen en gran medida de la disponibilidad y calidad del agua de la cuenca.

Desde el punto de vista industrial, la cuenca alberga varias industrias que requieren grandes cantidades de agua para sus procesos productivos, incluyendo plantas procesadoras de alimentos y bebidas, así como fábricas de productos químicos y textiles. La interacción entre estas actividades económicas y los recursos hídricos de la

cuenca plantea importantes desafíos en términos de gestión sostenible y protección ambiental (DINAMA y JICA, 2011).

Además, la cuenca del río Santa Lucía, en la cual la cuenca del Santa Lucía Chico está ubicada, es un área de interés ecológico. Alberga una diversidad de ecosistemas acuáticos y terrestres que sostienen una variedad de flora y fauna, incluyendo especies endémicas y en peligro de extinción. La conservación de estos ecosistemas es vital para mantener la biodiversidad y los servicios ecosistémicos que proporcionan, tales como la regulación del clima, la filtración de agua y el hábitat para la vida silvestre (MVOTMA, 2017).

El estudio y la gestión de la cuenca del río Santa Lucía Chico son esenciales para garantizar un equilibrio entre el desarrollo económico y la conservación de los recursos naturales. Las investigaciones y los proyectos enfocados en esta cuenca pueden ofrecer valiosos conocimientos y herramientas para la gestión integrada de los recursos hídricos, contribuyendo así a la sostenibilidad ambiental y al bienestar de la población local y nacional.

3. Datos disponibles

Los datos utilizados pueden agruparse en 3 grupos: *i)* hidrológicos: caudal (Q) [m^3/s] y nivel (h) [m] con datos diarios medidos en una estación desde 1971 hasta 2020; *ii)* meteorológicos: diarios de precipitación (P) [mm] en nueve pluviómetros desde 1980 hasta 2020, temperatura del aire media (TA_{med}) [$ºC$ ($24h$)], temperatura del aire máxima (TA_{max}) [$ºC$], temperatura del aire mínima (TA_{min}) [$ºC$], humedad relativa (HR) [%], radiación solar (RS) [W/m^2], heliofanía (Hel) [hs], evapotranspiración Penman (ET) [mm] y velocidad del viento (VV) [$2m/km/24h$] en la estación meteorológica más cercana con datos de 2013 a 2020 y muy pocos datos faltantes; *iii)* calidad del agua: se consideraron seis estaciones desde 2004 hasta 2020 con datos mensuales, en las cuales se mide fósforo total (FT) [$\mu g/L$], nitrógeno total (NT) [mg/L], ion nitrato (NO_3^-) [mg/L], ion nitrito (NO_2^-) [mg/L], ion amonio (NH_4^+) [mg/L], ion fosfato (PO_4^{3-}) [$\mu g/L$], solidos totales (ST) [mg/L], solidos suspendidos totales (SST) [mg/L], turbidez ($Turbidez$) [NTU], temperatura (T) [$ºC$], oxígeno disuelto (OD) [mg/L], demanda bioquímica de oxígeno (DBO) [mg/L], clorofila-a ($Clo-A$) [$\mu g/L$], potencial de hidrógeno (pH) y conductividad ($Cond$) [$\mu S/cm$].

Se pudo apreciar que el porcentaje de datos faltantes es de sólo 9% para las variables hidrológicas. Para las variables meteorológicas, el porcentaje de datos faltantes es menor al 9%. Las variables de calidad del agua muestran un porcentaje de datos faltantes que varía entre 57% y 66%.

4. Metodología

Se implementó un *framework* de imputación de datos faltantes que cuenta de cuatro fases (Figura 2): en la Fase 1, se aplican restricciones físicas que consideran la variabilidad espacial y temporal, la correlación y los rangos de variación de las variables consideradas. La Fase 2, selecciona variables de ayuda para entrenar los modelos. En la Fase 3, se entrenan y testean los modelos y, finalmente, en la Fase 4 se generan los sets de datos imputados.

Una descripción más detallada del marco metodológico se puede encontrar en Pastorini et al. (2024).

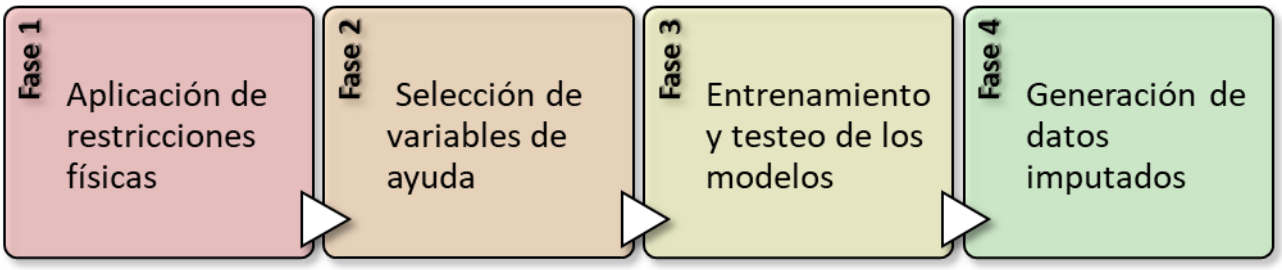

Figura. 2. Conceptualización del marco de imputación.

Fase 1: se consideran rangos de valores permitidos para las variables para evitar imputaciones irreales. Se calculan matrices de correlación de las variables de Pearson, Spearman y Kendall con límite de |0,5| y se consideran también las correlaciones entre variables presentes en la literatura.

Las dependencias espaciales consideran la posición de las estaciones de monitoreo. Por ejemplo, el caudal de una estación será influenciado por la precipitación de las estaciones aguas arriba. Además, las correlaciones espaciales representan el peso que se le otorga a las variables de ayuda de acuerdo a cuan alejadas están de la estación cuya variable se está imputando, según el método de Ponderación Inversa de la Distancia (*IDW*, por sus siglas en inglés):

$$Y_m = \frac{\sum_{i=1}^{n} Y_i d_{mi}^{-k}}{\sum_{i=1}^{n} d_{mi}^{-k}} \quad (1)$$

donde Y_m es la observación en la estación m, n es el número de estaciones, Y_i es la observación en la estación i, d_{mi} es la distancia entre m e i, y k es el exponente que generalmente varía entre 1 y 6. En este estudio, se asumió $k = 2$.

Se utiliza la media móvil ponderada exponencialmente (*EWMA*) para contemplar la variabilidad temporal:

$$EWMA(Y_n) = \frac{\sum_{i=0}^{t}(1-\alpha)^i Y_{n-1-i}}{\sum_{i=0}^{t}(1-\alpha)} \quad (2)$$

$$\alpha = \frac{2}{t+1} \quad (3)$$

donde Y es una serie temporal, n es el número de observaciones, α es el peso asignado, y t es la ventana temporal seleccionada basada en la variable a imputar: dos meses para las variables de calidad del agua, una semana para las variables hidrométricas y un día para las variables climáticas. Este método otorga mayor peso a las observaciones más recientes.

Cada una de estas restricciones se implementa como una variable artificial (columna) que se agrega al conjunto de datos original, creando, de esta manera, una nueva variante del conjunto de datos.

Fase 2: se consideran como variables de ayuda a aquellas que tuvieran menos del 50% de los datos faltantes y son temporalmente imputados con regresión lineal. Todas las variables se reducen a escala mensual para coincidir a las variables de calidad de agua. Además, se utiliza una normalización *min-max* para igualar la importancia de las variables con diferentes escalas. Finalmente, los datos fueron tabulados, resultando en un conjunto de datos donde cada punto (fila) incluía información temporal (en forma de nuevas variables artificiales, es decir, la columna EWMA). De esta manera, los datos eran independientes entre sí.

Fase 3: se corren diversos modelos de imputación que se pueden agrupar en modelos univariados: *Inverse Distance Weighting (IDW);* y multivariados: *Ridge Regressor (RR), TheilSen Regressor (TR), Huber Regressor (HR), Bayesian Ridge Regressor (BRR), Support Vector Regressor (SVR), K-Nearest Neighbors Regressor (KNNR), Random Forest Regressor (RFR) y Multivariate Imputation by Chained Equations (MICE)*. Se utiliza la validación cruzada de 10 pliegues para elegir el mejor modelo. Para evaluar el desempeño de los modelos, se calculan y comparan los valores de la eficiencia de Nash-Sutcliffe (NSE), usada como función objetivo, el sesgo porcentual (PBIAS) y la eficiencia Kling-Gupta (KGE). La Tabla 1 resume las calificaciones de rendimiento definidas para cada métrica de evaluación (NSE, PBIAS, KGE). Estas métricas se derivan de estudios publicados anteriormente (Chen et al., 2017; Moriasi et al., 2015; Rodríguez et al., 2021).

Tabla 1. Calificaciones de rendimiento definidas para cada métrica de evaluación.

Rating	Variables hidrométricas y climáticas	Variables de Calidad de agua físicas	Variables de Calidad de agua químicas
NSE			
Muy bueno	NSE > 0,80	NSE > 0,80	NSE > 0,65
Bueno	0,70 < NSE ≤ 0,80	0,70 < NSE ≤ 0,80	0,50 < NSE ≤ 0,65
Satisfactorio	0,50 < NSE ≤ 0,70	0,45 < NSE ≤ 0,70	0,35 < NSE ≤ 0,50
Insatisfactorio	NSE ≤ 0,50	NSE ≤ 0,45	NSE ≤ 0,35
PBIAS			
Muy bueno	| PBIAS | < 5	| PBIAS | < 10	| PBIAS | < 15
Bueno	5 ≤ | PBIAS | < 10	10 ≤ | PBIAS | < 15	15 ≤ | PBIAS | < 20
Satisfactorio	10 ≤ | PBIAS | < 15	15 ≤ | PBIAS | < 20	20 ≤ | PBIAS | < 30
Insatisfactorio	| PBIAS | ≥ 15	| PBIAS | ≥ 20	| PBIAS | ≥ 30
KGE			
Satisfactorio / Bueno	KGE ≥ -0,41	KGE ≥ -0,41	KGE ≥ -0,41
Insatisfactorio	KGE < -0,41	KGE < -0,41	KGE < -0,41

Fase 4: Se sigue un proceso iterativo donde se comienza con las variables con menos datos faltantes. Una vez que todas las variables fueron imputadas, obtuvimos el conjunto de datos completo final. El resultado del *framework* es el mejor par de conjunto de datos y modelo.

5. Resultados

La selección del mejor modelo se basó en el mayor valor de NSE, siendo esa la función objetivo. Además, se utilizaron PBIAS y KGE para determinar la exactitud del marco de imputación. El set de datos de salida incluye una serie de datos aumentada para todas las variables climáticas, hidrológicas y de calidad de agua con una frecuencia mensual.

En las Figuras 3, 4 y 5 se muestra una representación de *box-plot* del desempeño de los distintos dominios en términos de NSE, PBIAS y KGE, respectivamente. Se puede observar que más del 75% de todos los datos imputados se caracteriza por NSE>0,45 (resultados satisfactorios). En particular, el NSE mínimo calculado para las variables meteorológicas es de 0,72, lo que significa que todas las imputaciones pueden considerarse buenas (33%) y muy buenas (66%). Para las variables hidrométricas, el NSE siempre es >0,97, mostrando el rendimiento muy bueno del marco propuesto. El rendimiento tiende a disminuir al tratar con variables de calidad del agua. Más del 78% de las variables físicas de calidad de agua se caracterizan por NSE>0,45 (resultados satisfactorios), y más del 66% de las variables químicas de calidad de agua alcanzan NSE>0,35 (resultados satisfactorios). Para ambos dominios, más del 91% de los datos imputados tienen NSE>0, lo que significa que, para casi todas las imputaciones de calidad de agua, el marco propuesto es mejor que la media utilizada como función imputadora.

La validación del proceso de imputación fue notable, mostrando resultados en general muy buenos y buenos en términos de las calificaciones PBIAS y KGE.

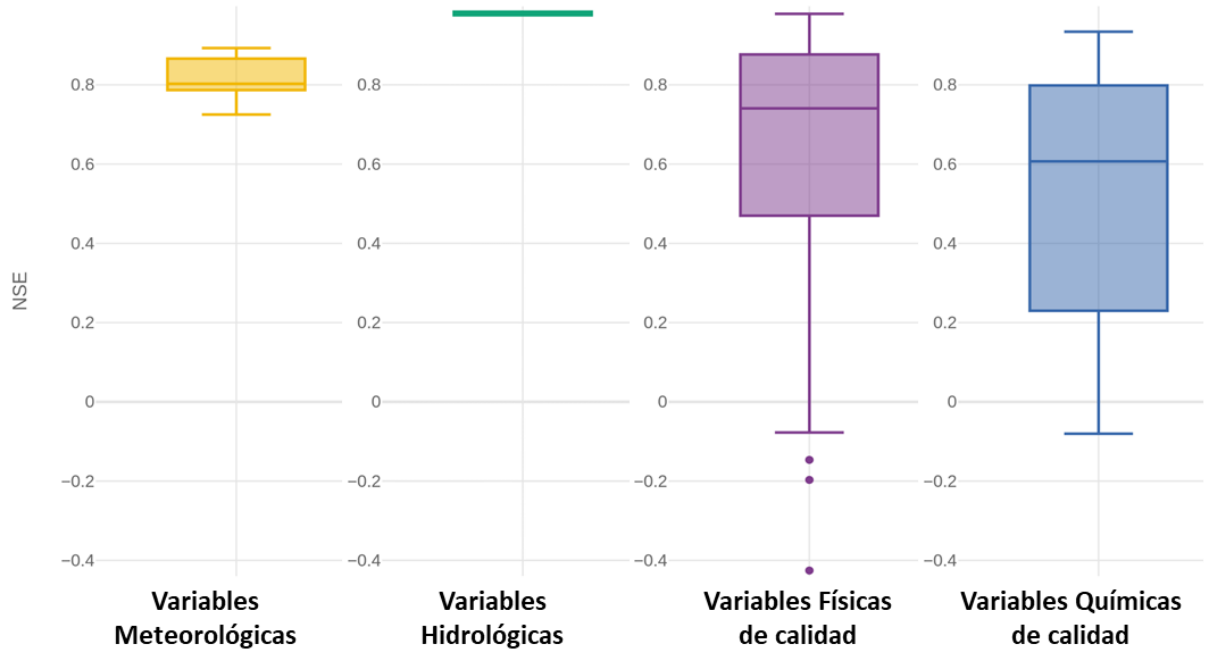

Figura. 3. Box-plots del desempeño de NSE para las variables meteorológicas, hidrométricas, de calidad de agua físicas y químicas.

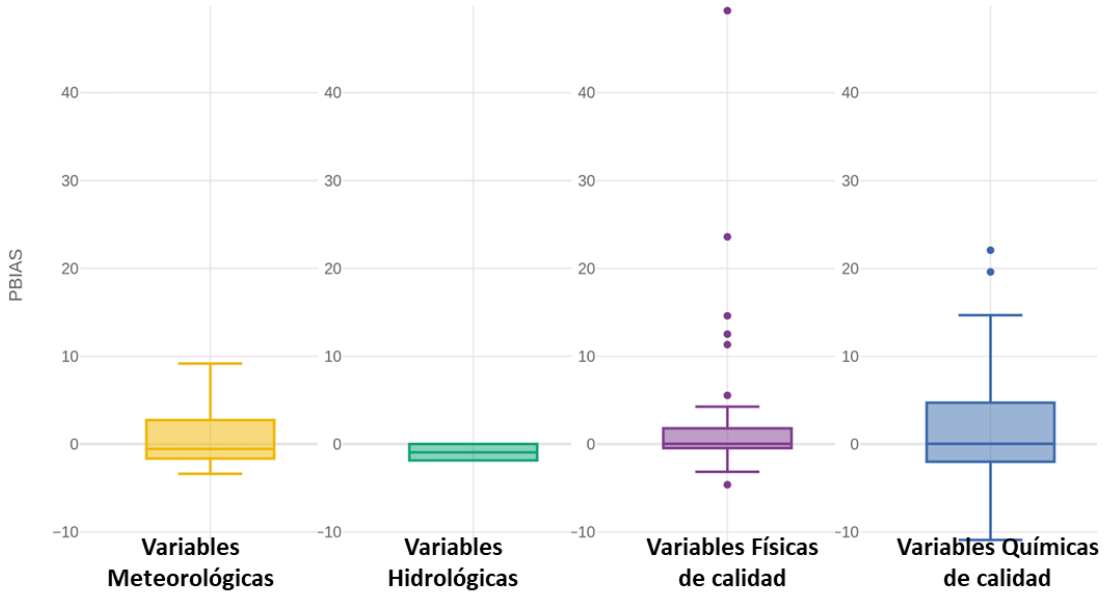

Figura. 4. Box-plots del desempeño de PBIAS para las variables meteorológicas, hidrométricas, de calidad de agua físicas y químicas.

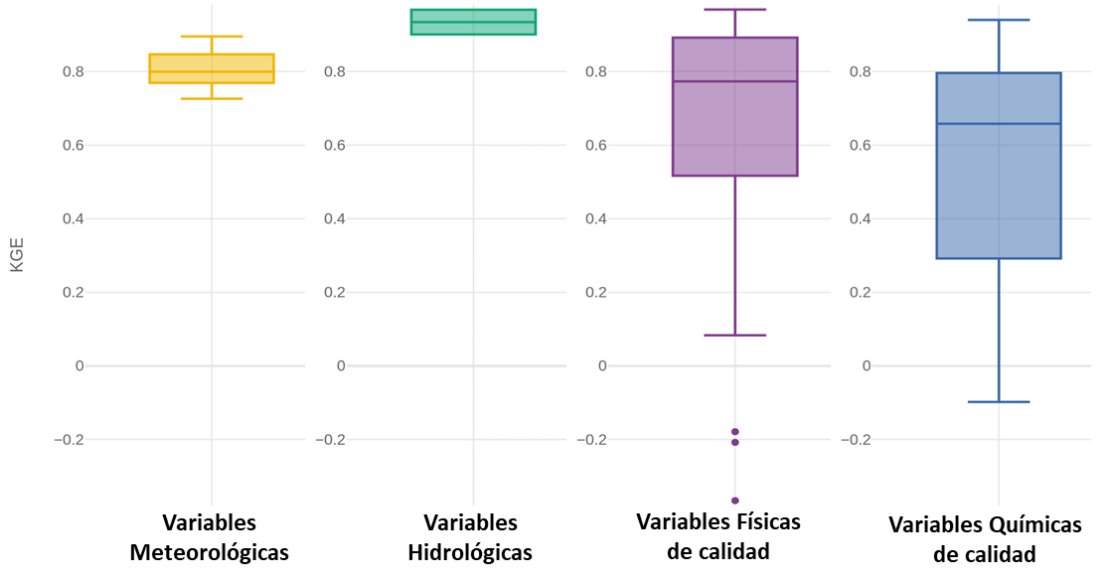

Figura. 5. Box-plots del desempeño de KGE para las variables meteorológicas, hidrométricas, de calidad de agua físicas y químicas.

6. Conclusiones

En este estudio se implementó un *framework* que atiende exitosamente al desafío de imputar datos faltantes en varios dominios ambientales. Este *framework* combina modelos basados en datos con conocimientos físicos, lo que resulta en imputaciones satisfactorias y precisas. Al integrar datos meteorológicos, hidrológicos y de calidad del agua, este enfoque logra capturar la complejidad y variabilidad inherentes a los procesos naturales a nivel de cuencas hidrográficas.

La integración de dichos datos en modelos a escala de cuencas hidrográficas no solo mejora el rendimiento de las simulaciones y predicciones de la calidad del agua, sino que también permite una comprensión más profunda de las dinámicas ambientales. Esto es crucial para la planificación y gestión de recursos hídricos, ya que proporciona una base sólida para la toma de decisiones en diversas aplicaciones, como la gestión de riesgos de inundaciones, la optimización del uso del agua para la agricultura y la industria, y la protección de ecosistemas acuáticos.

Además, la capacidad del *framework* para manejar altos porcentajes de valores faltantes y su adaptabilidad a diferentes variables ambientales lo convierte en una herramienta robusta y versátil. Esta robustez es especialmente importante en situaciones donde los datos son escasos o están incompletos, lo que es común en estudios ambientales. En conclusión, el desarrollo y la implementación de este *framework* representan un avance significativo en la gestión de datos ambientales, proporcionando una metodología eficiente y efectiva para enfrentar uno de los desafíos más persistentes en el campo de la ciencia ambiental.

Agradecimientos

Este trabajo fue financiado por la Agencia Nacional de Investigación e Innovación (ANII) con el proyecto FMV_3_2022_1_172720.

Referencias

Chen, H., Luo, Y., Potter, C., Moran, P.J., Grieneisen, M.L., Zhang, M. (2017). Modeling pesticide diuron loading from the San Joaquin watershed into the Sacramento-San Joaquin Delta using SWAT. Water Res., 121, 374–385.

Chen, Z., Xu, H., Jiang, P., Yu, S., Lin, G., Bychkov, I., Hmelnov, A., Ruzhnikov, G., Zhu, N., Liu, Z. (2021). A transfer learning-based LSTM strategy for imputing largescale consecutive missing data and its application in a water quality prediction system. J. Hydrol. 602, 126573

Chivers, B.D., Wallbank, J., Cole, S.J., Sebek, O., Stanley, S., Fry, M., Leontidis, G. (2020). "Imputation of missing sub-hourly precipitation data in a large sensor network: A machine learning approach". Journal of hydrology, Vol. 588, pp. 125126.

DINAMA y JICA (2011). Proyecto sobre control de contaminación y gestión de la calidad de agua en la cuenca del río Santa Lucía. Informe final del Proyecto.

Gorgoglione, A., Castro, A., Chreties, C., Etcheverry, L. (2020). "Overcoming data scarcity in earth science". Data, 5 (1), pp. 5.

Moriasi, D.N., Gitau, M.W., Pai, N., Daggupati, P. (2015). "Hydrologic and water quality models: Performance measures and evaluation criteria". Transactions of the ASABE, Vol. 58 (6), pp. 1763-1785.

MVOTMA (2017). Plan Nacional de Aguas, p.130-131. ISBN: 978-9974-658-31-8. http://www.mvotma.gub.uy/politica-nacional-de-aguas/plan-nacional-de-aguas

Pastorini, M., Rodríguez, R., Etcheverry, L., Castro, A., Gorgoglione, A. (2024). Enhancing environmental data imputation: A physically-constrained machine learning framework. Science of The Total Environment 926:171773.

Vilaseca, F., Castro, A., Chreties, C., Gorgoglione, A. (2023). Assessing influential rainfall–runoff variables to simulate daily streamflow using random forest. Hydrological Sciences Journal, 68(12), 1738–1753.