MARN
Ministerio de Medio Ambiente y Recursos Naturales

Aplicación de técnicas de modelación hidrológicas para pronóstico a corto plazo en El Salvador. Caso de estudio: Cuenca del Rio Torola.

José Rodolfo Valles León
Dirección Observatorio Ambiental, Gerencia de hidrología

RESUMEN

En este documento se pretende explorar técnicas de modelación hidrológica para el pronóstico de caudales en El Salvador. Dichos modelos establecen relaciones entre variables hidrometeorológicos tales como evapotranspiración, precipitación y caudal. Entre las herramientas ocupadas están los modelos auto-correlativos, Regresión Lineal Múltiple (MLR), Modelo de Árbol (M5-MT) y las Redes Neuronales Artificiales (ANN) con diferentes funciones de activación.

La primera parte de la investigación se enfoca en encontrar las variables hidrometeorológicas pasadas que brindan mayor información a los caudales pronosticados en la cuenca del Rio Torola, usando las características físicas de la cuenca y herramientas estadísticas como el Coeficiente de Correlación (CoC) y la Información Mutua Promedio (AMI). Una vez definido las variables de entrada del modelo, se evaluaron los resultados de las diferentes herramientas de modelación hidrológica ocupando funciones de evaluación tales como el Error Medio (ME), Eficiencia Nash-Sutcliffe (NSE), Coeficiente de determinación (R^2), Raíz del Error Cuadrático de la Media (RMSE) y su valor normalizado (NRMSE).

Los resultados muestran que, para pronóstico hidrológico con 1 hora de tiempo de ventaja, la arquitectura de Redes Neuronales Artificiales (ANN) con función de activación hiperbólica reproduce de forma precisa los caudales observados. Sin embargo, existen algunas subestimaciones del modelo, los cuales son debido a la falta de información de precipitación horaria en la parte alta y media de la cuenca. Adicionalmente, se evaluaron modelos de pronósticos para 2 y 3 horas de tiempo de ventaja, los cuales aumentan su error con respecto al caudal observado de 1 hora de tiempo de ventaja, pero se mantienen dentro de un rango muy bueno en su desempeño.

Palabras claves: *Pronostico, Redes Neuronales Artificiales, Modelo de Árbol, Regresión Lineal Múltiple, tiempo de ventaja*

INTRODUCCION

Los modelos hidrológicos son ampliamente ocupados para evaluar condiciones en la cuenca y para conocer futuros estados, los cuales son ocupados como pronósticos hidrológicos. Dichos pronósticos, permiten aumentar el tiempo de ventaja para poder actuar de mejor manera ante amenaza de inundaciones, sequia u otro fenómeno hidrometeorológico.

Los modelos basados en datos (Data-Driven) se pueden ocupar como modelos hidrológicos, los cuales solamente requieren información histórica y muy poca información de las características físicas de la cuenca. Diversos autores mencionan que estos modelos han tenido mucho éxito en brindar pronósticos a corto plazo con solamente relacionar precipitación y caudal observada en la cuenca.

El Salvador posee este tipo de información hidrometeorológica, teniendo registro de precipitación y caudal horario desde el año 2005. Debido a lo mencionado anteriormente, el objetivo de esta investigación es de explorar distintas herramientas de modelación hidrológica basado en datos para pronóstico de crecidas en la cuenca del Rio Torola. Estos modelos no han sido explorados en la región por lo que su implementación significaría un avance hacia técnicas de modelación modernas y más precisas.

METODOLOGIA

La primera fase del estudio está basada en el análisis de los datos hidrometeorológicos (lluvia y nivel) ubicado dentro o en los alrededores de la cuenca. En este análisis se identificaron valores atípicos dentro de la serie histórica con el objetivo de obtener una serie de datos de calidad para modelación hidrológica. Posterior a la selección de datos, la serie resultante fue transformada a datos horarios, el cual está justificado en el hecho de que el tiempo de concentración promedio de la cuenca es de 8 horas hasta el punto de control ubicado en la estación Osicala.

La segunda fase contempló el análisis de interdependencia entre variables en donde se definieron que variables hidrometeorológica brindan mayor cantidad de información beneficiosa para el pronóstico de los caudales con 1, 2 y 3 horas de tiempo de ventaja o anticipo. Para lo anterior, se ocuparon herramientas estadísticas tales como el Coeficiente de Correlación (CoC) y la Información Mutua Promedio (AMI, por sus siglas en ingles).

El Coeficiente de Correlación es un número que mide la dependencia lineal de una variable Y con respecto a otra variable X. Este número varía entre -1 y 1, en donde un valor igual a 0 indica no relación lineal, un valor de -1 indican una relación lineal negativa y un valor de 1 indica una relación lineal positiva. Por otra parte, la Información Mutua Media (AMI, por sus siglas en inglés) es una medida de la información que puede transferir un conjunto de datos con relación a otro (Shannon 1948). Abebe and Price (2003) menciona que el AMI puede ser usado para encontrar tanto correlaciones lineales como no lineales, debido a que emplea un conjunto de teorías que no están limitadas a una función. Un valor de AMI cercano a 0 indica que una variable X no tiene información suficiente que transmitir a una variable Y.

Posterior al análisis de interdependencia entre variables, los datos se dividieron en tres conjuntos de datos con diversos objetivos. La serie de entrenamiento corresponde a la mayoría de datos con el cual se determina los parámetros del modelo mediante optimización. El conjunto de datos para validación cruzada es una serie de datos que permiten controlar la etapa de entrenamiento con el fin de evitar el sobreajuste del modelo a los datos de entrenamiento. El tercer tipo de dato es de verificación del modelo para evaluar el desempeño del modelo con una serie de datos que el modelo no ha visto antes.

Una vez definido las variables que brindan información beneficiosa al pronóstico hidrológico, se ocuparon modelos auto-correlativos, regresión lineal múltiple (MLR), modelo de árbol (MT) y Redes Neuronales Artificiales (ANN) para pronosticar caudales con 1, 2 y 3 horas de tiempo de ventaja.

El modelo auto-correlativo de referencia conocido como Naïve es un modelo que permite comparar técnicas de pronostico sofisticados, con el fin de establecer un mínimo requerido de desempeño. Dicho modelo no requiere parámetros y solamente ocupa información de caudales pasados como muestra la siguiente ecuación:

$$Y_{t+i} = Y_t$$

La Regresión Lineal Múltiple (MLR por sus siglas en inglés) es un modelo que permite pronosticar una variable dependiente con respecto a dos o más variables independiente mediante la siguiente forma:

$$Y_{t+1} = \beta_o + \sum \beta_p \cdot X_{pi}$$

En donde β_o y β_p son los parámetros del modelo y X_{pi} son las diferentes variables de entrada.

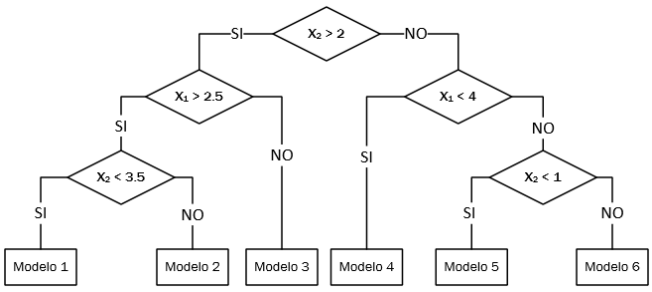

Figura 1: *Estructura del modelo de árbol M5 de dos variables de entrada (x) y separados en 6 modelos de regresiones lineales*

El modelo de árbol M5 mostrado en Figura 1 (Quinlan 1992; Frank et al. 2000) combina las características de clasificación y regresión lineal, bajo la suposición que la dependencia funcional no es constante en todo el campo de variables y que se puede aproximar a pequeños sub-campos (Solomatine and Xue 2004)

Las Redes Neuronales Artificiales (ANN por sus siglas en inglés) es una técnica de Machine Learning muy ocupado en problemas de predicción numérico y de clasificación (Solomatine and Xue 2004). La arquitectura del modelo se muestra en Figura 2 donde se puede observar las capas de entrada (input), oculta (Hidden) y las de salida (Output). Estas capas están conectadas por medio de líneas, las cuales son parámetros del modelo. En la capa oculta se encuentra los nodos o neuronas, las cuales reciben la información de la capa de entrada y evalúa una función de activación que puede ser sigmoidea, hiperbólica o lineal. El valor de cada neurona es transferido a la capa de salida y combinado con las otras neuronas para brindar una salida numérica.

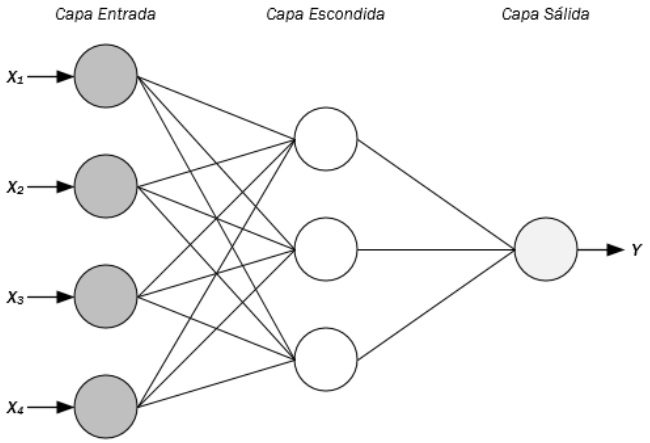

Figura 2: *Ejemplo de Red Neuronal Artificial con 4 variables entrada, 3 neuronas y 1 variable de salida*

Los resultados de todos los modelos mencionados anteriormente fueron evaluados mediante inspección visual de caudales observados y simulados, así como el cálculo del Error Medio (ME), Eficiencia Nash-Sutcliffe (NSE), Coeficiente de determinación (R^2), Raíz del Error Cuadrático de la Media (RMSE) y su valor normalizado (NRMSE).

CASO DE ESTUDIO

Para este estudio, se ha seleccionado la cuenca del Rio Torola, la cual tiene un área de 908 km² hasta la estación Osicala, en el departamento de Morazán, El Salvador. La cuenca nace en el departamento de La Paz (Honduras) y desemboca hasta el cauce del Río Lempa en El Salvador.

La información disponible incluye 3 años de datos horarios (01 de enero 2005 hasta el 31 de diciembre de 2007) de caudal, precipitación y evapotranspiración potencial. Los datos de caudal fueron calculados mediante una curva de descarga y nivel registrado en la estación Osicala. La Precipitación Media Areal fue calculada mediante polígonos de Thiessen usando 3 estaciones de precipitación ubicada en los alrededores de la cuenca (Ver Figura 3). La evapotranspiración horaria areal fue obtenida usando las tablas de relación entre la

evapotranspiración potencial y la elevación desarrolladas por el Servicio Nacional de Estudio Territoriales (SNET 2005), las cuales están basado en la fórmula de Hargreaves.

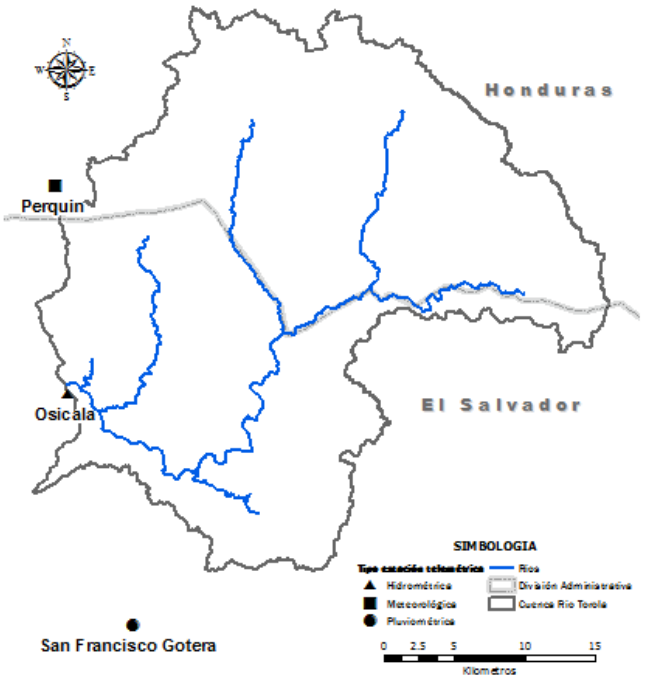

Figura 3: *Mapa de ubicación de la cuenca del Rio Torola hasta el punto de control en Osicala*

La Figura 3 detalla el área de la cuenca hasta el punto de control ubicado en la estación hidrométrica Osicala y las estaciones de precipitación. Es de hacer notar, que no existe información de precipitación horaria en la parte media y alta de la cuenca, lo cual es una limitante en el proceso de obtener una muy buena descripción de la lluvia en la cuenca.

ANALISIS DE INTERDEPENDENCIA ENTRE VARIABLES

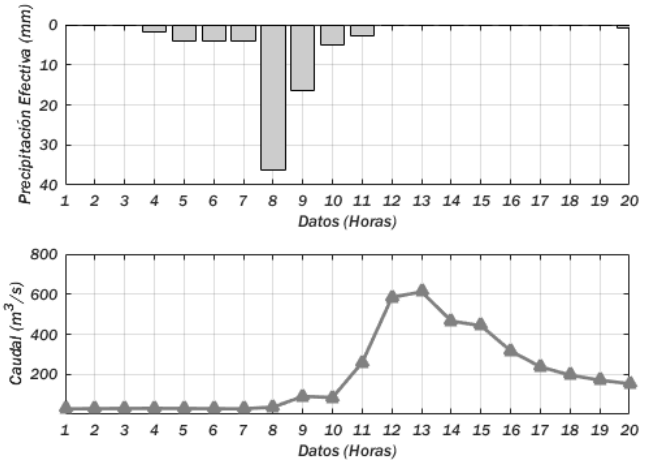

Figura 4: *Precipitación efectiva (mm) y caudal observado (m^3/s) en Rio Torola*

El primer problema es identificar cuantas variables de precipitación y caudal pasados se deben incluir en el modelo de pronóstico. Este proceso es normalmente desarrollado por un análisis de las características físicas de la cuenca con el objetivo de identificar su tiempo de respuesta, el cual es el tiempo entre el evento de lluvia y el caudal pico en el punto de control Osicala. Un ejemplo del análisis desarrollado se muestra en la Figura 4 en donde se puede observar que el tiempo de reacción de la cuenca es aproximadamente de 4 a 5 horas, en promedio. Por lo tanto, es recomendable ocupar precipitación de hasta 4 horas de retraso como variable de entrada.

De igual forma, un análisis fue realizado usando el Coeficiente de Correlación (CoC) y la Información Mutua Promedio (AMI) con el fin de encontrar cuantas variables pasadas de precipitación efectiva y caudal brindan mayor información al caudal pronosticado con 1 hora de tiempo de ventaja. Figura 5 muestra el resultado de dicho análisis, en donde se observa que la cuenca posee un tiempo de 4 horas de respuesta en promedio, debido a que la precipitación el mayor valor de correlación al caudal pronosticado con 1 hora de ventaja. Un análisis similar fue realizado con los caudales pasados para evaluar la autocorrelación que existe con el caudal pronosticado t+1. El resultado muestra que los caudales pasados brindan la mayor cantidad de información y correlación al caudal futuro.

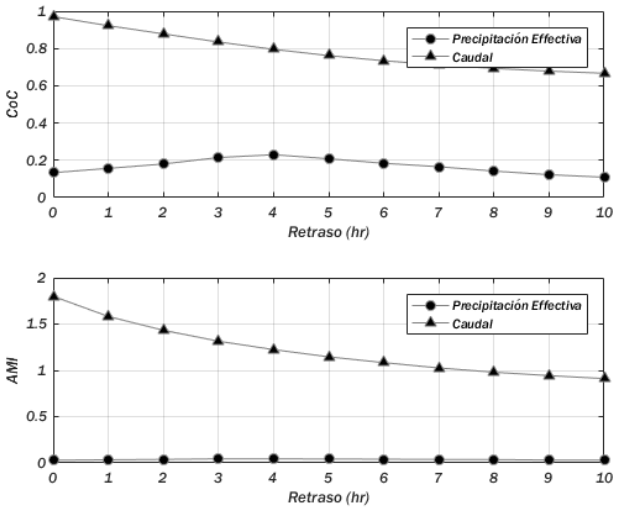

Figura 5: *Coeficiente de Correlación (CoC) e Información Mutua Promedio (AMI) entre las variables de entrada y el caudal pronosticado con 1 hora de tiempo de ventaja*

En base a los resultados obtenidos anteriormente, se puede generar una función de caudal pronosticado con 1 hora de tiempo de ventaja, la cual tendría la siguiente forma:

$$Q_{t+1} = f(RE_t, RE_{t-1}, RE_{t-2}, RE_{t-3}, RE_{t-4}, Q_t, Q_{t-1}, Q_{t-2})$$

En donde RE es la precipitación efectiva y Q es el caudal en el punto de medición de la estación hidrométrica Osicala.

La función de predicción del caudal t+1 requiere 8 variables de precipitación efectiva (RE) y los caudales anteriores (Q), sin embargo, debe mantenerse un número óptimo de variables de entrada debido a que se corre el riesgo de sobre-parametrizar el modelo hidrológico. Debido a lo anterior, se simplificó la función mostrada en la siguiente ecuación, usando una media móvil entre las 5 variables de precipitación efectiva ($MARE_t$)

$$Q_{t+1} = f(MARE_t, Q_{t-1}, Q_{t-2})$$

PREPARACIÓN DE LOS DATOS

Una vez definido las variables de entrada que se usaran para generar los diferentes modelos, se deben dividir la serie en tres tipos de datos:

- **Entrenamiento:** consiste en la mayor cantidad de los datos disponibles y con el cual se entrena las diferentes herramientas de modelación para encontrar sus parámetros.
- **Validación cruzada:** esta serie de datos es usada con el fin de evaluar el desempeño del modelo y evitar el sobreajuste de los parámetros en la etapa de entrenamiento.
- **Verificación:** en esta serie de datos se evalúa el desempeño del modelo con una serie de datos que no ha visto anteriormente.

La serie total horaria ocupada en esta investigación fue de 26,264 datos correspondiente a los años hidrológicos 2005, 2006 y 2007. Las propiedades estadísticas de cada una de las series, se describe en la siguiente tabla:

Tabla 1: *Propiedades estadísticas de los tres tipos de datos*

	Entrenamiento	Validación cruzada	Verificación
Cantidad (Horas)	13,109	7,276	5,879
Máximo (m^3/s)	706.53	611.83	549.25
Mínimo (m^3/s)	0.01	0.74	1.16
Promedio (m^3/s)	20.49	37.15	30.07
Desviación estándar (m^3/s)	39.27	58.11	51.46

RESULTADOS Y DISCUSION

Diferentes herramientas de modelación hidrológica fueron evaluadas tales como Regresión Lineal Múltiple (MLR), modelo de árbol (M5-MT) y Redes Neuronales Artificiales (ANN) con dos funciones de activación: sigmoidea e hiperbólica. Lo anterior con el objetivo de explorar cada una de ellas para pronosticar caudales con 1, 2 y 3 horas de tiempo de ventaja.

Tabla 2: *Cálculo del error entre los caudales observados y simulados en la etapa de verificación para pronóstico con 1 hora*

	ME (m^3/s)	NSE (-)	RMSE (m^3/s)	NRMSE (-)	R^2 (-)
Naïve	-0.01	0.94	12.44	0.24	0.97
MLR	-0.17	0.95	11.16	0.22	0.97
MT	0.21	0.95	11.15	0.22	0.97
ANN Sigmoidea	0.05	0.94	11.23	0.25	0.96
ANN Hiperbólica	0.04	0.95	10.99	0.21	0.97

Los modelos resultantes fueron evaluados mediante diversas fórmulas evaluación de desempeño, tales como el Error Medio (ME), la Raíz del Error Medio Cuadrático (RMSE), el valor normalizado del RMSE, el Coeficiente de Eficiencia de Nash-Sutcliffe (NSE) y el coeficiente de determinación (R^2). Para más información acerca de las debilidades y fortalezas de cada una de estas funciones de errores, se recomienda verificar referencias bibliográficas en Gupta et al. (1998)

Tabla 2 muestra los cálculos de los errores del modelo en replicar los caudales observados con 1 hora de tiempo de ventaja para cada una de las herramientas mencionada anteriormente. Se puede observar que cada herramienta de modelación hidrológica ocupada brinda mejores resultados que el modelo de referencia auto regresivo Naive. Sin embargo, la Red Neuronal Artificial (ANN) con función de activación hiperbólica muestra los mejores resultados. Por ejemplo: el valor normalizado del RMSE es de 0.21, lo cual indica un muy buen desempeño del modelo, ya que es menor al 0.5.

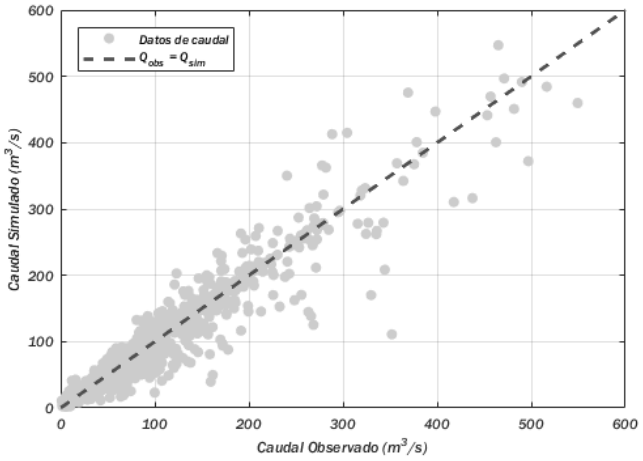

Figura 6: *Gráfica de dispersión entre datos observados y simulados en la etapa de verificación para pronóstico de 1 hora*

Figura 6 muestra un diagrama de dispersión entre los caudales medidos en la estación Osicala y los caudales simulados usando una Red Neuronal Artificial con función de activación hiperbólica. Se puede observar que el modelo brinda una representación aceptable de caudales bajos y altos en la cuenca. Para ilustrar lo anterior, se brinda el caso del caudal observado de 456 m^3/s, el cual el modelo reproduce 456.8 m^3/s. Por el contrario, se observan algunos eventos perdidos o mal representados. Por ejemplo, el caudal observado de 351.9 m^3/s fue subestimado por el modelo, ya que brinda una salida de 111.0 m^3/s.

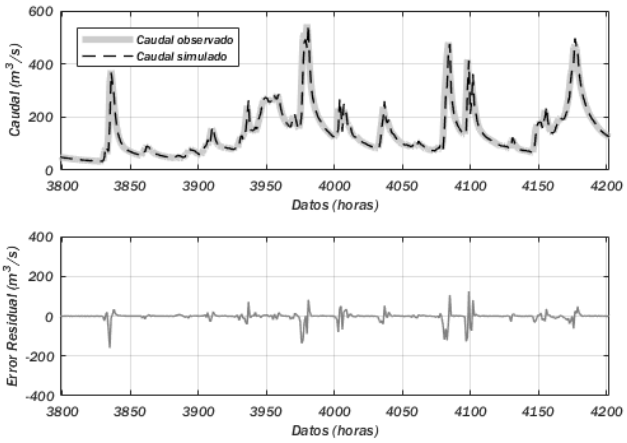

Figura 7: *Fragmento de los caudales observados y simulados en la etapa de verificación del modelo para pronostico con 1 hora de tiempo de ventaja*

Figura 7 muestra el error residual entre el caudal observado y el simulado para un pequeño fragmento en el conjunto de datos de verificación. Se puede apreciar la buena reproducción en los caudales pico y en el volumen de agua. No obstante, los errores residuales cercanos a 200 m^3/s es producido por una mala representación de la lluvia areal que alimenta el modelo, ya que se generó datos de lluvia con solo 3 estaciones telemétricas, de las cuales ninguna está ubicada dentro de la cuenca. Esta situación genera que el modelo sea más auto-correlativo en los periodos donde la lluvia areal calculada sea cercano a 0 y el caudal de la cuenca pase de un régimen bajo a alto.

Esta limitante en la descripción de la lluvia areal promedio puede ser superada con más información de precipitación en la cabecera y en parte media de la cuenca, ya que el modelo tendría datos de precipitación de estaciones beneficiosas para el cálculo del caudal en el punto de control Osicala.

Usando el mismo procedimiento descrito anteriormente, se desarrollaron modelos de redes neuronales con función hiperbólica para pronósticos hidrológicos con 2 y 3 horas de tiempo de ventaja. Las variables de entrada y los números de nodos ocupados para cada uno de estos modelos son mostrados en la Tabla 3 en donde los valores de precipitación efectiva (MARE), es el acumulado de las ultimas 5, 4 y 3 horas para los pronósticos de 1, 2 y 3 horas, respectivamente.

Tabla 3: *Resumen de los modelos de Redes Neuronales con función hiperbólica*

Variables de entrada	Variable salida	Neuronas
$MARE_{5t}$, Q_t, Q_{t-1}, Q_{t-2}	Q_{t+1}	2
$MARE_{4t}$, Q_t, Q_{t-1}	Q_{t+2}	2
$MARE_{3t}$, Q_t, Q_{t-1}	Q_{t+3}	2

Tabla 4 detalla un resumen del cálculo del error en los modelos de pronóstico hidrológico para los tiempos de ventaja de 1, 2 y 3 horas usando Redes Neuronales con función de activación Hiperbólica.

Tabla 4: *Cálculo del error en el modelo de pronóstico con 2 y 3 horas de tiempo de ventaja*

	ME (m^3/s)	NSE (-)	RMSE (m^3/s)	NRMSE (-)	R^2 (-)
Q_{t+1}	0.04	0.95	10.99	0.21	0.97
Q_{t+2}	0.06	0.86	18.60	0.36	0.93
Q_{t+3}	-0.05	0.80	23.09	0.45	0.89

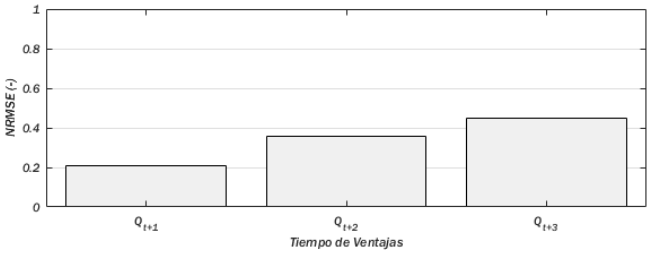

Figura 8: *Gráfica de barras entre el tiempo de ventaja del pronóstico hidrológico y el valor del NRMSE*

En la Figura 8 se puede observar que, aumentando el tiempo de ventaja del pronóstico hidrológico, el desempeño del modelo se ve disminuido. El resultado anterior era esperado debido a que es la precipitación observada en las estaciones telemétricas está saliendo del sistema y para aumentar el tiempo de ventaja del pronóstico se debe recurrir a lluvia pronostica. Adicionalmente, el valor normalizado del RMSE de 0.45 representa un desempeño del modelo entre muy bueno y bueno, de acuerdo a (D. N. Moriasi et al. 2007)

CONCLUSIONES Y RECOMENDACIONES

El presente documento pretende evaluar nuevas técnicas de modelación hidrológica en la región con el objetivo de pronosticar caudales con 1, 2 y 3 horas de tiempo de ventaja. La modelación hidrológica está basada en técnicas de Machine Learning, el cual es muy ocupado a nivel mundial para resolver diversos problemas relacionado con el agua. Entre las herramientas ocupadas en este documento se puede mencionar modelo auto-correlativo, regresión lineal múltiple, Modelo de árbol y Redes Neuronales Artificiales.

La predicción del caudal con 1 hora de tiempo de ventaja brinda muy buenos resultados debido a que reproduce bien los caudales altos y bajos. Sin embargo, se observó subestimación de algunos eventos, los cuales pueden deberse a una pobre descripción de la lluvia areal, la cual alimenta el modelo. Dicha precipitación es obtenida con la información disponible de 3 estaciones telemétricas, las cuales no están ubicado dentro de la cuenca.

Los pronósticos hidrológicos en la cuenca con 2 y 3 horas de tiempo de ventaja, brindan resultados aceptables y que se encuentran dentro del rango muy bueno del valor normalizado del RMSE (0.5), de acuerdo a lo propuesto por D. N. Moriasi et al. (2007). Adicionalmente, se pudo comprobar que, aumentando el tiempo de ventaja del pronóstico, se aumenta el error del modelo. Lo anterior debido a que, a mayor tiempo de ventaja, la precipitación en la cuenca va saliendo del sistema. Debido a lo anterior, se prefirió mantener el pronóstico hidrológico hasta 3 horas de tiempo de anticipo o ventaja.

Los modelos basados en datos (Data-Driven) permiten obtener información útil de predicción a corto plazo con solamente información histórica de variables hidrometeorológicas. Estas técnicas de modelación son muy novedosas en la región y los resultados de este estudio nos permiten dar un primer acercamiento a su uso y beneficios, con el objetivo de ser replicados en otras cuencas para obtener pronósticos hidrológicos a corto plazo y con lluvia observada.

REFERENCIAS

Abebe A, Price R (2003) Managing uncertainty in hydrological models using complementary models.

D. N. Moriasi, J. G. Arnold, M. W. Van Liew, et al (2007) Model Evaluation Guidelines for Systematic Quantification of Accuracy in Watershed Simulations.

Frank E, Chui C, Witten I (2000) Text categorization using compression models.

Gupta HV, Sorooshian S, Yapo PO (1998) Toward

improved calibration of hydrologic models: Multiple and noncommensurable measures of information. Water Resour Res. doi: 10.1029/97WR03495

Quinlan J (1992) Learning with continuous classes.

Shannon C (1948) A mathematical theory of communication, Part I, Part II.

SNET (2005) Balance hídrico integrado y dinámico en El Salvador. 109.

Solomatine D, Xue Y (2004) M5 model trees and neural networks: application to flood forecasting in the upper reach of the Huai River in China.