IAHR AIIH
XXIV CONGRESO LATINOAMERICANO DE HIDRÁULICA
PUNTA DEL ESTE, URUGUAY, NOVIEMBRE 2010

PREDICCIÓN DE CAUDALES MEDIOS MENSUALES EN RÍOS DE COLOMBIA USANDO RNA Y MARS

Velásquez N., Sanchez J., Carvajal L.F.
Universidad Nacional de Colombia, Sede Medellín, Colombia, nvelasqg@unal.edu.co, joanysanchezmolina@gmail.com, lfcarvaj@bt.unal.edu.co

RESUMEN:

Se implementan dos metodologías para el pronóstico de caudales medios mensuales en ríos de Colombia: (i) MARS ("Multivariate Adaptative Regresion Splines"), y (ii) RNA (Redes Neuronales Artificiales). El método no paramétrico de MARS y el método RNA, permiten incorporar la persistencia hidrológica y la influencia de fenómenos macroclimáticos como el ENSO. Las predicciones se hicieron durante el período 1996-2007 para seis ríos de Colombia, los cuales son importantes en la generación de energía eléctrica en Colombia: Nare, Porce, Miel, Magdalena en Betania, Guavio y Batá. Se evalúan los dos métodos de predicción usando intervalos de pronóstico de 1 mes, para horizontes de pronóstico de 12 meses. El desempeño de los modelos se cuantifica mediante medidas del error en el horizonte de validación, la bondad del pronóstico se cuantifica a partir del error cuadrático medio, dependiendo del mes en el que se inicia el pronóstico y de la ventana de predicción. Los resultados indican ganancias importantes en la capacidad de predicción en comparación con métodos tradicionales usados en hidro-climatología.

ABSTRACT:

In this work two flow forecasting methodologies has been implemented in different Colombian rivers: (i) MARS ("Multivariate Adaptative Regresion Splines"), and (ii) ANN (Artificial Neural Networks). The no parametric method MARS and the RNA method, allow to incorporate the hidrology persistence and the influence of macroclimate phenomena, like the ENSO. The forecasting was made during the period 1996-2007, in six Colombian rivers that are very important to the hydroelectric generation in Colombia: Nare, Porce, Miel, Magdalena in Betania, Guavio and Batá. The evaluation of the two forecasting methods was made using forecasting intervals of 1 month, for horizons of 12 months. The performance of the two models is quantify using the error in the horizon of the validation, and the goodness of the forecasting is quantify from the mean square error, that depends on the beginning forecast month and the length of the forecast window. Results show an important advance in forecast compare to the traditional methods in hydro-climatology.

PALABRAS CLAVES:

Predicción, MARS, RNA, Caudales medios mensuales.

KEY WORDS

Forecast, MARS, ANN, Monthly mean flows

INTRODUCCIÓN

Colombia, debido a su ubicación geográfica (entre los 4° Sur y los 12° Norte) posee una alta variabilidad climática, influenciada principalmente por el paso de la ZCIT, a la cual se debe la bimodalidad del clima en la mayor parte de su territorio, a esto se suma la influencia de la corriente de vientos del chorro del Chocó, que transporta aire cargado de humedad desde el Pacifico hacia el interior del territorio colombiano (Poveda 1998, Poveda y Mesa 2000). Se ha encontrado que el ENSO ejerce una fuerte influencia en el clima Colombiano, afectando la intensidad con que fenómenos tales como el Chorro del Chocó se presentan, ejerciendo efectos positivos o negativos de acuerdo a la fase en la que se encuentra el ENSO (Poveda 1998, Poveda y Mesa 2000).

Ya que más del 60% de la energía en Colombia es producida a partir de fuentes aprovechamientos hidroeléctricos, es de gran importancia conocer el comportamiento estacional que poseen los caudales de los ríos que se encuentran involucrados en tal generación, y cuya variabilidad se ve afectada directamente por el comportamiento del clima. Dado lo anteriormente descrito es fundamental, para las empresas del sector eléctrico Colombiano, conocer como se podrían comportar los caudales de los ríos en diferentes ventanas de tiempo (máximo 1 año).

Para realizar predicciones de caudales medios mensuales en diferentes ríos Colombianos se emplearon dos metodologías altamente no lineales, las cuales permiten usar como variables predictoras no sólo los registros históricos de caudales, si no además variables macro climáticas tales como el SOI (índice de Oscilación del Sur), el Chorro del Chocó y el MEI (Índice Multivariado del ENSO), dando así la posibilidad de introducir un mayor sentido físico en la predicción de los caudales ya que se está introduciendo la variabilidad aportada por tales variables. Las metodologías usadas son MARS (Polinómios de regresión multivariados y adaptativos) (Friedman 1991), y RNA (Redes Neuronales de Retro Propagación del Error) (Y. Le Cunn 1988).

Para evaluar los errores de ambas metodologías se usó el RMSE (error cuadrático medio). El RMSE se calcula sobre las predicciones hechas a partir de los diferentes meses del año, esto con el fin de observar como la época del año afecta la calidad de las predicciones, debido a los diferentes ciclos que se presentan a lo largo del mismo.

Como objetivo general se busca realizar predicción de caudales medios mensuales en diferentes ríos Colombianos a una ventana de 12 meses, y evaluar el desempeño de las predicciones realizadas mediante el RMSE.

Entre los objetivos específicos se tiene:

- Describir la variabilidad del clima de Colombia, y su importancia en la predicción de caudales medios mensuales.

- Usar los métodos MARS y RNA para realizar predicciones en una ventana de 12 meses, en los ríos Nare, Porce, Miel, Magdalena en Betania, Guavio y Batá.

MARCO TEÓRICO

A continuación se explican los fundamentos teóricos de los métodos MARS y RNA, por lo que ambos han sido aplicados para realizar predicciones sobre los mismos ríos Colombianos.

Método MARS

En particular MARS es una implementación de técnicas propuestas por Friedman (1991) para resolver problemas de regresión. Tales técnicas tienen como objetivo principal predecir valores de una variable continua dependiente o de salida, a partir de un grupo de variables independientes llamadas comúnmente variables predictoras. El método MARS es un método de regresión no paramétrica y no lineal, que está basado en una generalización de particionamiento recursivo, y ha sido utilizado en gran cantidad de áreas del conocimiento.

El modelamiento no paramétrico se caracteriza porque no aproxima una función única en todo el dominio, sino que ajusta una función con varias funciones paramétricas simples, generalmente polinomios de bajo orden, definidas sobre una subregión del dominio (ajuste paramétrico por tramos) ó ajusta una función simple para cada valor de la variable (ajuste global). Los parámetros en cada región también son hallados por mínimos cuadrados.

La función de aproximación ajustada puede ser de la forma:

$$\hat{f}(x) = \sum_{m=1}^{M} a_m \; B_m(x) \qquad [1]$$

donde M es el número de subregiones, a_m son los coeficientes de la función y $B_m(x)$ es una función base dada por:

$$B_m(x) = \prod_{k=1}^{K_M} H\left[S_{KM} \; \left(x\,(k,m) - t_{KM}\right)\right]_{+} \qquad [2]$$

donde H es una función de paso que va desde $k=1$ hasta k_M, el número de divisiones resultantes en la función básica, S_{KM} es igual a ± 1 según sea la división derecha ó izquierda respectivamente, t_{KM} es el nudo o partición de la variable y x es la variable predictora. El signo $(+)$ como subíndice de dicha expresión significa que sólo se toma como resultado cuando el argumento de la función es positivo, de lo contrario se hace igual a cero.

En Friedman (1991) se encuentran los aspectos relacionados con la modelación no paramétrica y la computación adaptativa. Además allí se presenta completamente el algoritmo de ajuste de MARS.

En general, MARS, intenta superar las limitaciones de la modelación no paramétrica y el particionamiento recursivo, planteando algunas generalizaciones a los procedimientos, por ejemplo, garantizando modelos continuos y derivadas continúas. El modelo de ajuste de MARS se puede escribir de la forma:

$$\hat{f}(x) = a_0 + \sum_{k_m=1} f_i(x_i) + \sum_{k_m=2} f_{ij}(x_i, x_j) + \sum_{k_m=3} f_{ijk}(x_i, x_j, x_k) + \ldots\ldots \qquad [3]$$

MARS utiliza el criterio de falta de ajuste, LOF para la selección del modelo, que esta definido como el máximo número de funciones básicas M_{max}. Una función para LOF esta definida como lo muestra la ecuación [4].

$$LOF\ (\hat{f}_M) = GCV\ (M) = \frac{1}{N} \sum_{i=1}^{N} \frac{\left[y_i - \hat{f}_M(x_i)\right]^2}{\left[1 - \dfrac{C(M)}{N}\right]^2} \qquad [4]$$

El criterio $GCV(M)$ es el promedio de los residuos al cuadrado afectado por un factor de penalización, el cual se presenta en el denominador. La función C(M), que es una función de costo, esta dada por la ecuación [5].

$$C(M) = traza(B(B^T B)^{-1} B^T) + 1 \qquad [5]$$

Donde B es la matriz de datos de las M funciones básicas ($B_{ij} = B_i(x_j)$).

Método RNA

Las redes neuronales se componen básicamente de tres tipos diferentes de estructuras, las cuales son: La micro-estructura es la estructura de la neurona como tal, es decir la función de transferencia aplicada a cada neurona. La meso-estructura consta de la forma de organización y funcionamiento de la red y la macro-estructura aparece cuando se consideran diferentes redes para dar solución a un sistema complejo.

Entre los diferentes tipos de meso-estructuras se encuentra la de "retropropagación" junto con otras como la red multicapa con alimentación hacia adelante, las redes de capa singular conectadas lateralmente y de capa singular ordenada topológicamente, organizándose éstas meso-estructuras en dos grandes categorías, una en la que el entrenamiento se realiza de manera asistida y otra en la que el entrenamiento es no asistido. Para éste caso se ha empleado el algoritmo de "retropropagación", (la cual se basa en la red perceptrón propuesta por Rosenblatt (1958)) el cual como su nombre lo indica se encarga de tomar el error cuadrático medio obtenido a la salida de cada iteración y a partir de éste actualizar las conexiones de la red con el fin de obtener un menor error en la siguiente iteración. Debido a que el algoritmo necesita comparar sus resultados con datos conocidos, ésta estructura pertenece al conjunto de redes entrenadas de manera asistida.

El esquema de la red de retropropagación fue propuesto por Werbos en 1974, la ¡Error! No se encuentra el origen de la referencia. muestra la arquitectura de una red de "retropropagación", la cual consta, al igual que una arquitectura típica de otras redes neuronales, de una capa de entrada, capas ocultas y una capa de salida. Se visualiza los pesos de conexión en los vínculos de todas las neuronas de una capa con las neuronas de la capa contigua, además la arquitectura posee un parámetro que se encuentra conectado a cada neurona de la capa oculta denominado "bias" o sesgo, el cual ayuda a estabilizar la red durante la etapa de aprendizaje y es de valor unitario, el único parámetro que cambia es el peso de la conexión entre los sesgos y las diferentes neuronas.

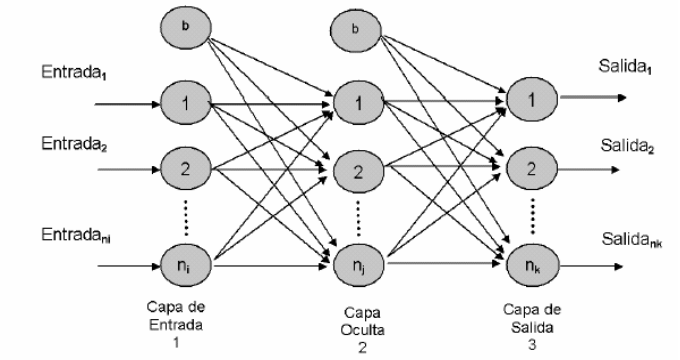

Figura 1-Estructura típica de una red neuronal artificial.

En la ¡Error! No se encuentra el origen de la referencia. se puede apreciar que las capas no requieren tener el mismo número de neuronas, inclusive si posee más de una capa oculta. Lo anterior sugiere que la capa de entrada tiene un número *I* de neuronas, una capa oculta *n1* de *J*

neuronas, una capa oculta *n2* de *H* neuronas, y así sucesivamente hasta la capa la cual tiene un número *K* de neuronas.

Los datos de entrada de la red consisten en un conjunto de datos, de parámetros o bien de un patrón singular. El número de neuronas depende del problema que se vaya a resolver y de la cantidad de datos que se tengan disponibles.

Dentro de cada una de las neuronas de la capa oculta los datos se ajustan mediante funciones de transformación, estas funciones generalmente varían su dominio entre -1 y 1, y en ocasiones entre 0 y 1. A su vez existen funciones continuas y discontinuas, en la Figura 2 se observan las funciones de las transformaciones más típicas.

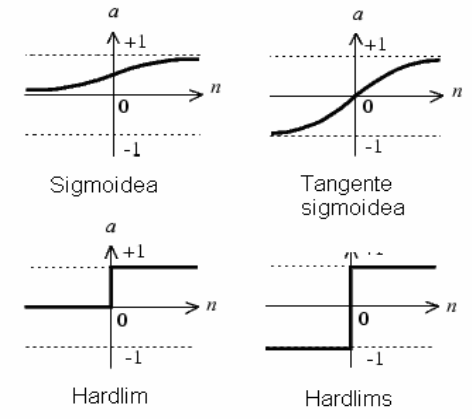

Figura 2-Funciones de transformación.

En éste caso se ha empleado la función tansigmoidea para la transformación de la información en las neuronas de la capa oculta, en la ecuación [6] se presenta la manera de operar de tal función.

$$f(x) = \frac{2}{(1+e^{-2x})} - 1 \qquad [6]$$

Se tiene entonces que el valor de entrada a la función, es dado por los valores de entrada a la red, los cuales son asignados mediante las neuronas de la capa de entrada, y por los pesos de las conexiones lo cuales se actualizan en cada iteración con el fin de optimizar la red. De esta manera la entrada de datos a cada neurona de la capa oculta queda dada por la siguiente ecuación.

$$Entrada = \sum_{i=1}^{n} w_i x_n \qquad [7]$$

Para implementar el algoritmo de retropropagación del error a través de la red durante el aprendizaje, se toma el error cuadrático medio para cuantificar el aprendizaje, definido como:

$$E = \frac{\sum_{i=1}^{n} (Q_t - Q_r)^2}{n} \qquad [8]$$

Donde Q_t es el caudal obtenido por la red y $Q?$ es el caudal observado. El error depende del valor observado y el obtenido por la red, el observado se le entrega a la red como un dato para que ella realice el aprendizaje, a esto se le llama aprendizaje asistido, el cual es el modo de aprendizaje más usado para éste tipo de problemas.

El algoritmo de retropropagación básicamente busca minimizar el error representado por la ecuación [8] actualizando los pesos de las conexiones entre las diferentes capas. Para ello se usa el factor de momento, la tasa de aprendizaje, los pesos de las conexiones de la iteración anterior y el error obtenido en cada iteración. El algoritmo de actualización de factores de ponderación ó de pesos, funciona de acuerdo como se muestra en la ecuación [9]

$$dx_i = ?\%dx_{i-1} + ?\%(1-?\%)de/dx_i \qquad [9]$$

Donde *dx* es la derivada de los pesos de conexión, *de* es la derivada del error al final de cada iteración, ?º es el factor de momento, y ?ºes la tasa de aprendizaje de la red. Al incluir el factor de momento se ayuda a la red a no caer en valores que son óptimos locales, de forma que sea capaz de encontrar óptimos globales. La definición de la tasa de aprendizaje es crucial para el desempeño de la red, ya que si éste valor es muy pequeño el entrenamiento necesitará de una gran cantidad de iteraciones para reducir su error, pero si es muy alto es probable que se obtenga un mal entrenamiento.

METODOLOGÍA

Como se ha mencionado en los objetivos, la predicción de series de caudales medios mensuales se realizó sobre los siguientes ríos: Nare, Porce, Miel, Magdalena en Betania, Guavio y Batá, con un promedio de longitud de registro de 47 años, teniendo en cuenta que las series finalizan en Diciembre del 2007.

En ambos modelos el periodo de calibración va desde la fecha de inicio de registro hasta enero de 1996, teniéndose así 12 años para el período de validación. En el caso de las Redes Neuronales el período de calibración está comprendido entre la fecha de inicio de las series y diciembre de 1995, mientras que en el caso de la metodología MARS la calibración se hace de forma dinámica, es decir la calibración siempre estará comprendida entre el mes de inicio de la serie de tiempo de caudales y el mes anterior al mes de inicio de las predicciones. Los restantes valores se toman para la validación de los modelos y comparar los resultados obtenidos por las predicciones con lo observado durante el período en cuestión.

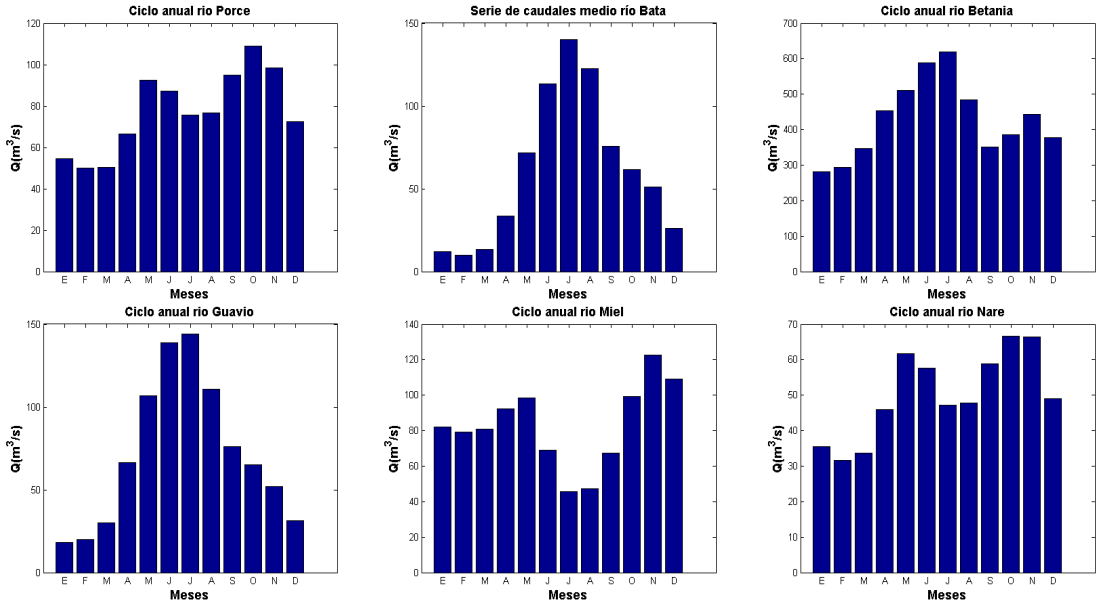

Figura 3-Ciclo anual de los diferentes ríos trabajados.

En la Figura 3 se presenta el ciclo anual de los diferentes ríos que se han estudiado. Se puede observar que para algunos ríos el ciclo es unimodal mientras que para otros, es de carácter bimodal. La presencia de la bimodalidad se da en gran parte por el paso que hace la ZCIT sobre la zona de los Andes Colombianos en el transcurso del año hidrológico. Se tiene entonces que 4 de los 6 ríos analizados presentan un comportamiento bimodal (Porce, Betania, Miel y Nare), lo cual los hace más variables, y por ende se dificulta en mayor medida la predicción. En la Figura 4, se muestra como ejemplo la serie de caudales medios del río Nare, el cual en la Figura 3 presenta un ciclo bimodal muy marcado.

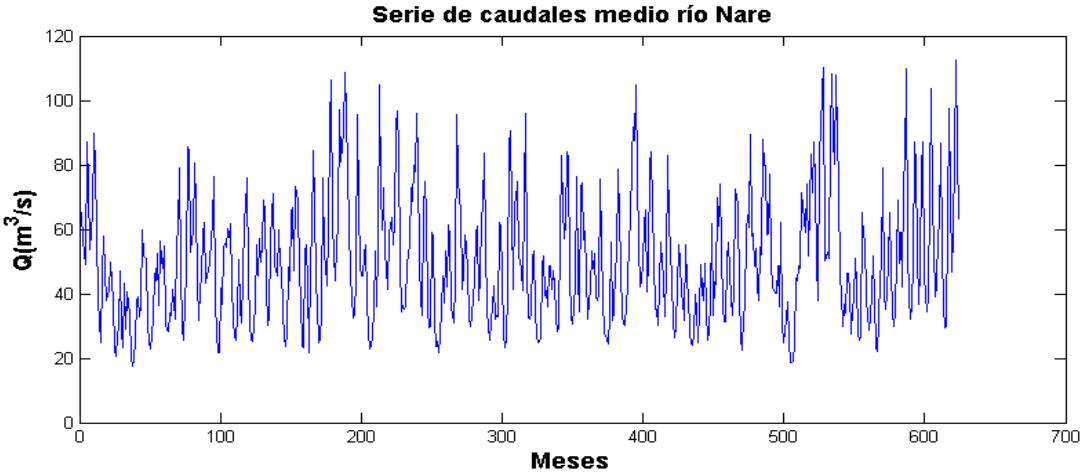

Figura 4-Serie de tiempo de caudales medios mensuales del río Nare.

Además de las series de caudales medios mensuales se han empleado como datos de entrada diferentes series de variables macroclimáticas, ya que éstas ayudan a la comprensión de la variabilidad de los caudales en las diferentes regiones de Colombia. Basados en el hecho de que el comportamiento de éstas variables se encuentra correlacionado con la variabilidad del clima colombiano en diferentes rezagos temporales, debido a factores tales como la ubicación geográfica y mecanismos de tipo oceánico y artmosférico como el Fenómeno del Niño Oscilación del Sur. Se realizaron ensayos, empleando las variables en diferentes rezagos. En la Tabla 1, se presentan las correlaciones de las variables más influyentes en la predicción, con los caudales medios mensuales de los ríos, se muestra únicamente el valor de correlación máximo obtenido y el rezago en que éste es obtenido. Como puede apreciarse la gran mayoría de los ríos presentan correlaciones significativas con las variables macro climáticas empleadas, se observa también que la variable Choco presenta el promedio de las correlaciones más altas y en muchas de las corrientes se encuentran las correlaciones más altas al tener la serie del Chocó rezagada 4 meses, por lo que es de gran utilidad al ser utilizada como variable predictora tanto en la metodología MARS como en RNA.

Tabla 1.- Correlaciones cruzadas entre los caudales de los ríos

Corriente	Choco	MEI	SOI
Bata	-0.76 (rezago -4)	-0.19 (rezago -5)	-0.16 (rezago 6)
Betania	-0.49 (rezago -4)	-0.33 (rezago -3)	0.25 (rezago -6)
Porce	0.59 (rezago 0)	-0.52 (rezago 0)	0.45 (rezago 0)
Miel	0.47 (rezago -5)	-0.46 (rezago 0)	0.41 (rezago 0)
Nare	0.55 (rezago 0)	-0.51 (rezago 0)	0.46 (rezago 0)
Guavio	-0.65 (rezago -4)	-0.19 (rezago -5)	0.15 (rezago -5)

RESULTADOS Y ANÁLISIS

El desempeño de los modelos es evaluado para todas las series mediante el error cuadrático medio (RMSE), el cual está definido como se muestra a continuación en la ecuación [10]

$$\%RMSE = \frac{\sqrt{\frac{1}{n}\sum (Qobservado - Qpredicho)^2}}{\overline{Q}observado} \times 100 \qquad [10]$$

En las Figura 5 y 6, se muestran los errores obtenidos por esta metodología, para ambos métodos y para todos los ríos.

Para los diferentes ríos se compara el RMSE promedio de las predicciones hechas con inicio en cada uno de los meses de todos los años en el período de validación de las metodologías, con una ventana de predicción de doce meses.

Así por ejemplo se quiere evaluar el RMSE de predecir julio de cualquier año comenzando en diciembre del año anterior, el criterio lo que hace es establecer una diferencia entre el valor real y predicho del caudal de julio de ese año, los valores de las predicciones de todos los julios con mes de inicio diciembre en todo el período de validación son almacenados en un vector. En la expresión 10, n representa es el número de años de validación de los modelos, que para nuestro caso es igual a 12, que es igual al número de julios predichos con mes de inicio diciembre. El procedimiento anterior es análogo para todas las predicciones con mes de inicio enero, febrero, marzo, etc. Es importante tener en cuenta que el período de calibración se va actualizando a medida que se va pasa de un mes de inicio a otro. En la figura 4 se ilustra lo planteado anteriormente.

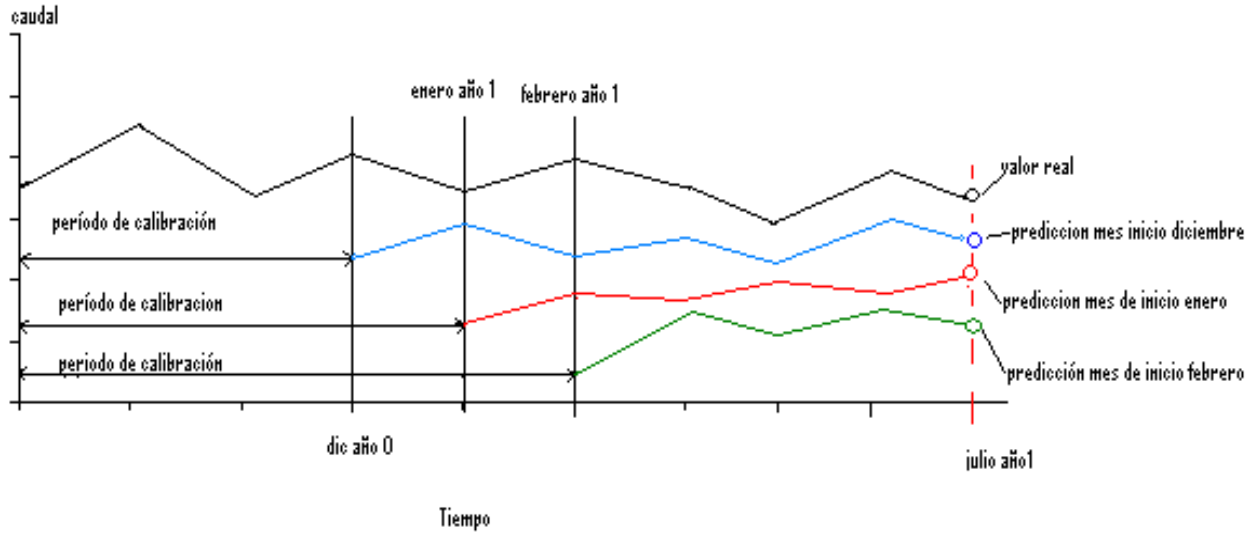

Figura 4- actualización de los períodos de calibración para cada mes de inicio.

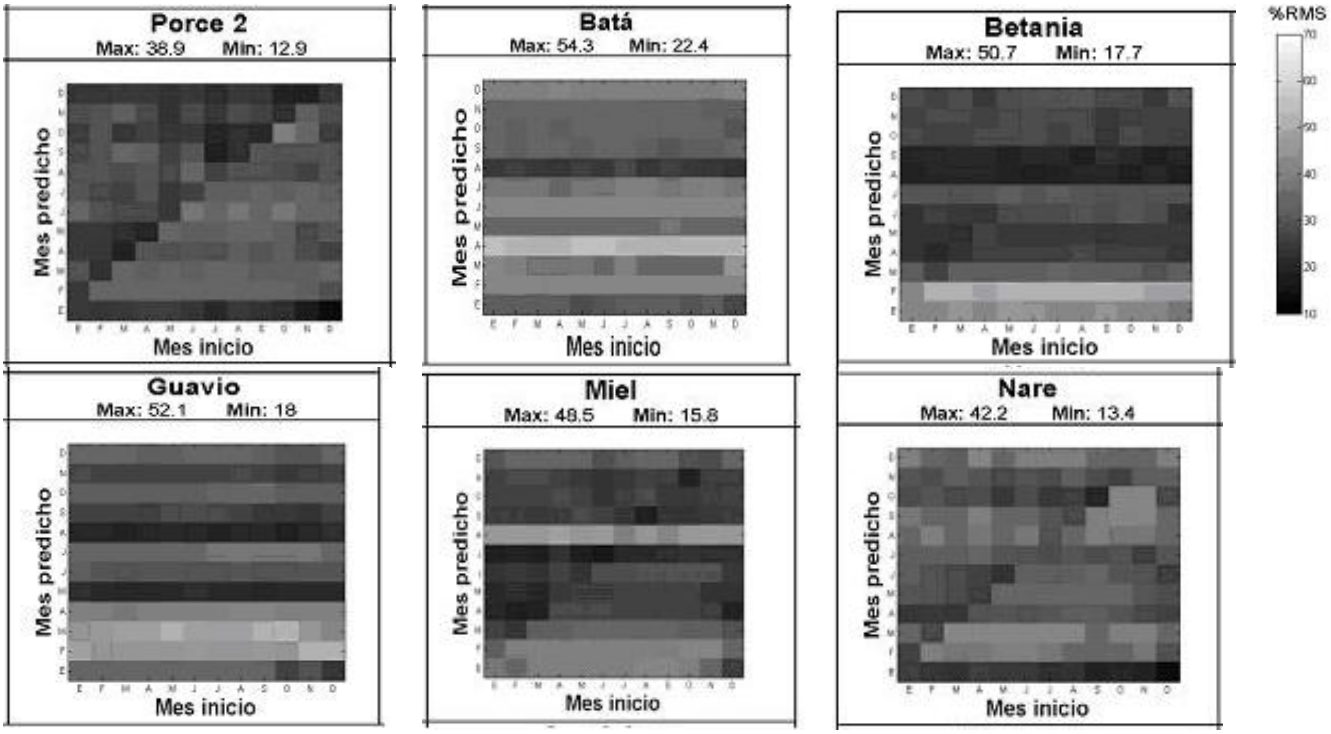

Figura 5- Error cuadrático medio para los diferentes ríos usando RNA

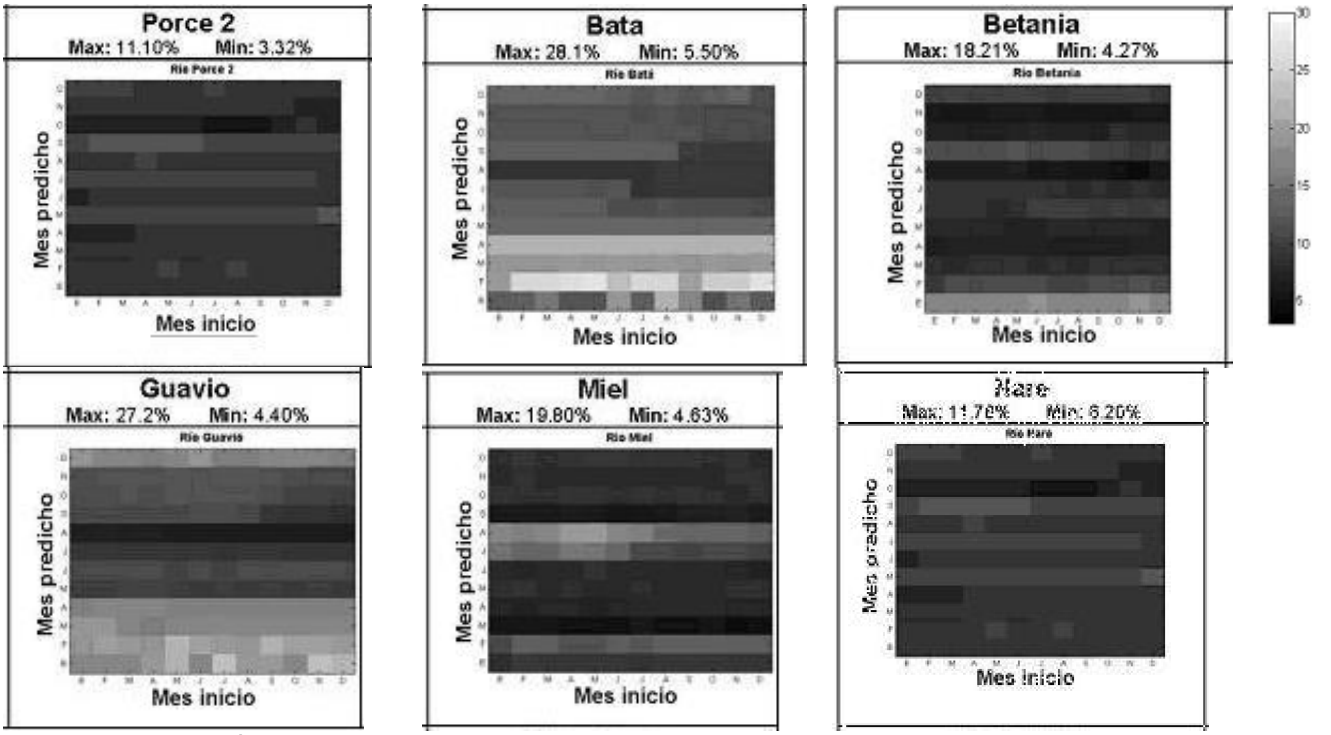

Figura 6- Error cuadrático medio para los diferentes ríos usando MARS.

En términos generales durante las validaciones el método MARS presenta un desempeño mejor que el mostrado por el método RNA, a pesar de esto los mayores errores se siguen presentando para ambos métodos en los mismos ríos, así mismo sucede con las corrientes que presentan los menores errores. Observando las Figura 5 y 6 se encuentra que el comportamiento de los errores para ambos métodos varía, lo cual indica que cada método posee sensibilidades diferentes y esto en gran parte puede ser debido a la manera de operar de cada uno.

Se encuentran bandas de meses en los que indiferente del mes de inicio de predicción los errores se presentan de manera homogénea, por lo que son meses de más fácil predicción, y esto se encuentra asociado con que son meses de caudales y variabilidad más baja.

Ambos modelos se han calibrado con las variables macroclimáticas mencionadas anteriormente, pero no se han empleado datos de lluvia, debido a la no disponibilidad de los mismos. Igualmente las variables empleadas generan visibles mejoras en el desempeño de ambos modelos, ya que mediante su uso se da una mejor comprensión de los procesos físicos que se llevan a cabo en las cuencas de cada una de las corrientes.

CONCLUSIONES

En éste trabajo se ha realizado un ejercicio de predicción usando dos metodologías no lineales diferentes, la primera de ellas MARS, y la segunda RNA. Se usó para la predicción, diferentes variables macro climáticas, todas ellas representativas de la variabilidad del ENSO.

Los resultados muestran que el uso de éste tipo de metodologías presenta ganancias en la predicción hecha a una ventana de 12 meses. Así mismo el uso de variables macro climáticas genera mejoras en la calidad de las predicciones.

Se ha encontrado un mejor desempeño en el uso de la metodología MARS, lo cual no indica que el método RNA deba ser desechado, ya que en éste campo se pueden realizar una gran cantidad de avances, y éste ante otros métodos se obtienen resultados competentes.

Bibliografía

Poveda, G., (1998). "Retroalimentación dinámica entre el fenómeno El Niño- Oscilación del Sur y la hidrología de Colombia", *Tesis Ph.D., Universidad Nacional de Colombia*.

Poveda, G., O. J. Mesa, L. F. Carvajal, C. D. Hoyos, J. F. Mejía, L. A. Cuartas y A. Pulgarín. (2002). "Predicción de caudales medios mensuales en ríos colombianos usando métodos no lineales". *Meteorología Colombiana*, 6, 101-110.

Friedman, J. H. (1991). "*Multivariate Adaptive Regression Splines"*, Stanford University. Stanford, CA.

Y. Le Cunn. (1988) "A Theorical Framework for Back-Propagation", Proceeding of the 1988 Conectionist Models Summer School, pag. 21-28.

Rosenblatt F (1958). The Perceptron: A probabilistic model for information storage and organization in the brain. *Psychological Review* 65: 386-408.

Werbos, P.J. (1974/1994), The Roots of Backpropagation, NY: John Wiley & Sons. Harvard Ph.D. thesis, *Beyond Regression.*