

Comparación del filtro de Kalman discreto con el filtro de conjuntos para pronóstico de caudales horarios en el río Huaynamota, Nayarit, México

Comparison of discrete and ensemble Kalman filter for hourly streamflow forecasting in Huaynamota River, Nayarit, México

Ildefonso Narváez-Ortiz¹, ORCID: <https://orcid.org/0000-0002-4988-8886>

Laura Alicia Ibáñez-Castillo², ORCID: <https://orcid.org/0000-0001-9287-655X>

Ramón Arteaga-Ramírez³, ORCID: <https://orcid.org/0000-0001-9459-3588>

Mario Vázquez-Peña⁴, ORCID: <https://orcid.org/0000-0003-2084-7420>

Carlos Cántora-González⁵, ORCID: <https://orcid.org/0000-0001-5204-4361>

¹Universidad Autónoma Chapingo, estudiante del Doctorado en Ingeniería Agrícola y Uso Integral del Agua, Chapingo, México, ildenar@gmail.com

²Profesora-investigadora, Universidad Autónoma Chapingo, posgrado en Ingeniería Agrícola y Uso Integral del Agua, Chapingo, México, México, libacas@gmail.com

³Profesor-investigador, Universidad Autónoma Chapingo, posgrado en Ingeniería Agrícola y Uso Integral del Agua, Chapingo, México, México, arteagar@correo.chapingo.mx

⁴Profesor-investigador, Universidad Autónoma Chapingo, posgrado en Ingeniería Agrícola y Uso Integral del Agua, Chapingo, México, México, mvazquezp@correo.chapingo.mx

⁵Profesor-investigador, Universidad Autónoma Chapingo, posgrado en Ingeniería Agrícola y Uso Integral del Agua, Chapingo, México, México, carlos.cintora@gmail.com

Autora para correspondencia: Laura Alicia Ibáñez-Castillo, libacas@gmail.com

Resumen

La asimilación de datos integrada para el pronóstico de caudales puede brindar flexibilidad y reducción de errores sistemáticos en los modelos. En este trabajo se evalúan la capacidad predictiva del filtro de Kalman discreto, filtro de Kalman de conjuntos y su integración, utilizando registros horarios de caudal de las estaciones Chapalagana y Platanitos ubicadas sobre el río Huaynamota, región hidrológica 12. La cuenca se

ubica al noroeste de la república mexicana, y se comparte entre los estados de Durango, Nayarit, Zacatecas y Jalisco. Para el análisis se utilizaron series con 1 360 registros horarios del año 2017 comprendidos entre el 2 de agosto a las 9:00 horas hasta el 28 de septiembre a las 0:00 horas. Se realizaron pronósticos a 1, 2, 3, 4, 5 y 6 pasos hacia adelante, combinados con tamaños de conjunto de 5, 8, 10, 20, 50 y 100 miembros utilizando caudales de la estación Platanitos como variable exógena. El ajuste entre la serie observada y las pronosticadas se estimó mediante el coeficiente de Nash-Sutcliffe y la raíz del cuadrado medio del error para determinar que el filtro de Kalman discreto alcanza mejor ajuste y actualización con base en el tiempo de retraso entre series. El filtro de Kalman de conjuntos genera un suavizado de la serie pronosticada, y al igual que la integración de filtros aumenta el efecto de desplazamiento de la serie pronosticada. El filtro de Kalman discreto alcanza ajuste superior a ARX y a la combinación ARX-DKF.

Palabras clave: filtro de Kalman, conjuntos, modelos autorregresivos, pronósticos de caudales a corto plazo.

Abstract

Integrated data assimilation for flow forecasting can provide flexibility and reduce systematic errors in the models. In this work we evaluate the predictive capacity of the discrete Kalman filter, ensemble Kalman filter, and its integration, using hourly flow records from Chapalagana and Platanitos stations located on the Huaynamota river, hydrological region

12. The basin is located in the northwest of the Mexican Republic and is shared between the states of Durango, Nayarit, Zacatecas, and Jalisco. For the analysis, series with 1360 data from 2017 were used, from August 2nd at 9:00 a.m. to September 28th at 0:00 a.m. Forecasts were evaluated at 1, 2, 3, 4, 5, and 6 steps forward, combined with set sizes of 5, 8, 10, 20, 50, and 100 members using measurements at the Platanitos station as an exogenous variable. The fit between observed and predicted series was estimated using the Nash-Sutcliffe coefficient and the mean square root of the error to determine that the discrete Kalman filter achieves better fit and update based on the time delay between series. The Ensemble Kalman filter generates smoothing of the predicted series, and the integration of filters increases the displacement effect of the predicted series. The discrete Kalman filter achieves superior adjustment to ARX and the ARX-DKF combination.

Keywords: Kalman filter, ensembles, autoregressive models, short-term streamflow forecasting.

Recibido: 19/03/2019

Aceptado: 23/11/2020

Introducción

La asimilación de datos para la estimación de estados es esencial en las aplicaciones de pronóstico en hidrología (Liu *et al.*, 2012), pues mediante la actualización cíclica que llevan a cabo es posible disminuir los errores que se acumulan en los modelos (Clark *et al.*, 2008; Maxwell, Jackson, & McGregor, 2018). El filtro de Kalman de conjuntos (EnKF) (Evensen, 1994) es uno de los algoritmos ampliamente utilizado como método recursivo en hidrología (Liu & Gupta, 2007; Maxwell *et al.*, 2018; Sun, Seidou, Nistor, & Liu, 2016) es una extensión del filtro de Kalman discreto (Kalman, 1960) para tratar sistemas dinámicos no lineales (Evensen, 1994; Evensen, 2003), y se ha utilizado, entre otros, para el pronóstico de caudales (Maxwell *et al.*, 2018), evapotranspiración (Zou, Zhan, Xia, Wang, & Gippel, 2017) y humedad del suelo (Brandhorst, Erdal, & Neuweiler, 2017; Meng, Xie, & Liang, 2017), además se ha integrado con modelos hidrológicos distribuidos como TopNet, Hydrotel y MGB-IPH (Abaza, Anctil, Fortin, & Turcotte, 2015; Clark *et al.*, 2008; Quiroz, Collischonn, & Paiva, 2019).

En los sistemas de cuenca hidrográfica se puede identificar la relación entre los caudales medidos en diferentes posiciones y un punto de interés, a lo que se denomina función de respuesta (Valdés, Mejía, & Rodríguez-Iturbe, 1980). Con base en esta relación se pueden hacer pronósticos a corto plazo que se actualizan constantemente y de forma indefinida mientras el fenómeno bajo estudio persista.

En este trabajo se evalúa la capacidad de pronosticar del filtro de Kalman discreto (DKF) (Kalman, 1960), el filtro de Kalman de conjuntos (EnKF) (Evensen, 1994) y la integración del filtro de Kalman discreto (DKF) con el filtro de Kalman de conjuntos (EnKF) para el pronóstico de uno hasta seis pasos hacia delante de los caudales del río Huaynamota en la estación Chapalagana. El DKF es implementado para estimar el vector de estado (Hidrograma Unitario Instantáneo, HUI) y el EnKF hace estimaciones escalares de caudal. La integración se hace mediante la ecuación de estados del EnKF.

Materiales y métodos

El área de estudio corresponde a la delimitación aguas arriba a partir de la estación Chapalagana sobre el río Huaynamota (río Chapalagana o río Atengo), se ubica al noroeste de la república mexicana y se comparte entre los estados de Durango, Nayarit, Zacatecas y Jalisco (INEGI, 2010) (Figura 1); geográficamente se ubica entre los $-104^{\circ} 33' 34.16''$ y los $-103^{\circ} 27' 29.84''$ de longitud, y entre los $23^{\circ} 28' 50.05''$ y los $21^{\circ} 23' 57.62''$ de latitud; tiene un área de $12\,075.7\text{ km}^2$; la altitud máxima es de $3\,147\text{ msnm}$ (metros sobre el nivel del mar) y la mínima en la estación Chapalagana es de 219 msnm ; el tiempo de concentración es de 39.88 horas; la precipitación media anual es de 707 mm y la temperatura media anual de $17.9\text{ }^{\circ}\text{C}$ (SMN, 2019).

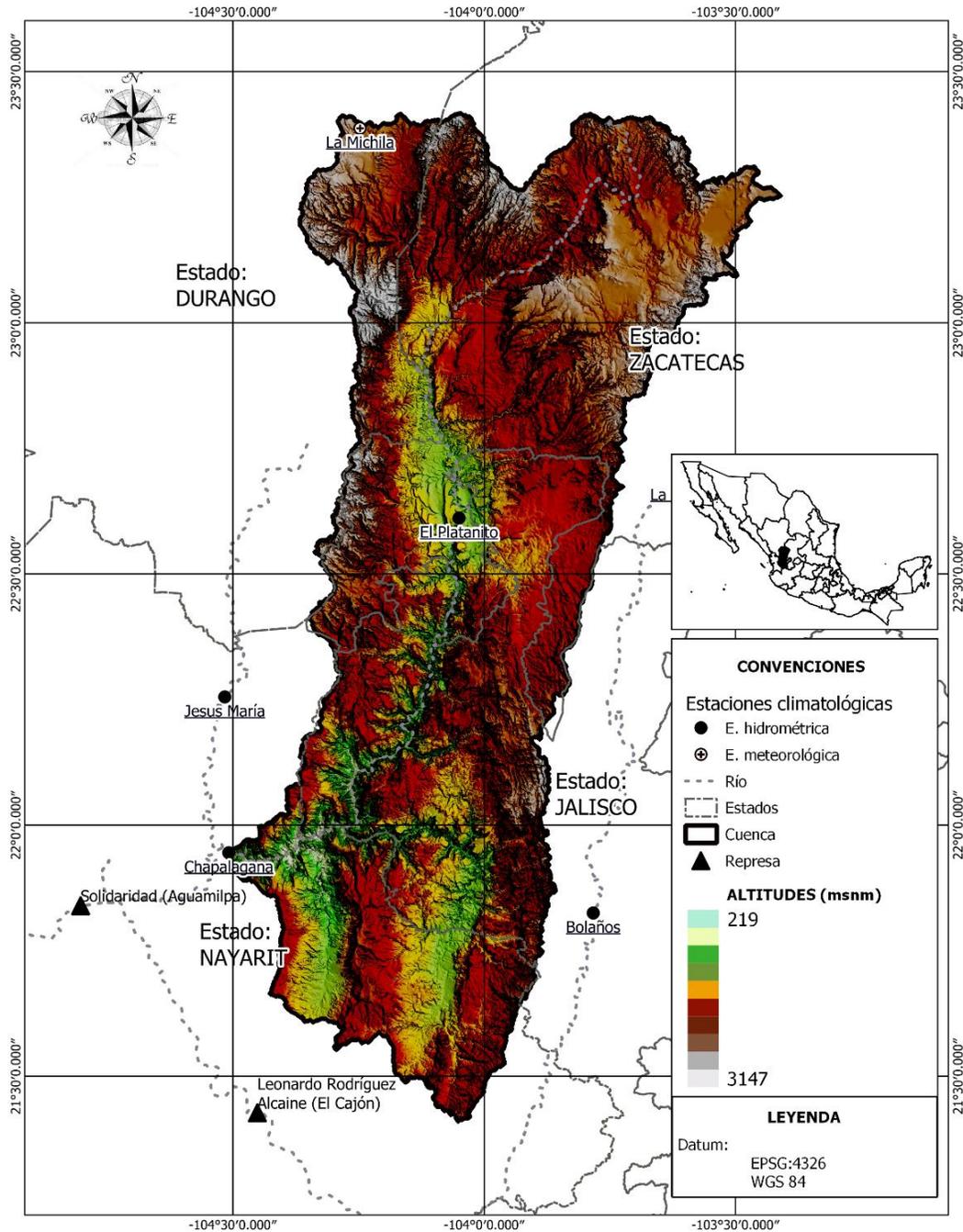


Figura 1. Localización del área de estudio.

La principal estructura hidráulica con fines de generación de electricidad que se encuentra aguas abajo del área de estudio y sobre el río Lerma-Santiago Pacífico es la represa Solidaridad, comúnmente conocida como Aguamilpa, ubicada en los $-104^{\circ} 48' 10.55''$ de longitud y los $21^{\circ} 50' 22.74''$ de latitud, con una capacidad de 5 540 millones de m^3 , para generar 960 MW de electricidad (Conagua, 2008). La presa Aguamilpa se encuentra a unos 90 km del océano Pacífico, donde el río Santiago desemboca frente a las costas del estado de Nayarit.

Se evaluó la capacidad predictiva del algoritmo del filtro de Kalman discreto (DKF) (Kalman, 1960); el filtro de Kalman de conjuntos (EnKF) (Evensen, 1994); la integración del DKF y EnKF (DKF-EnKF), y el modelo autorregresivo con variable exógena de primer orden (AR(1,1)), para realizar el pronóstico a 1, 2, 3, 4, 5 y 6 horas (pasos L) hacia adelante de caudales horarios en la estación Chapalagana, con base en los caudales de la estación Platanito, localizada aguas arriba, como variable exógena. Se utilizaron series horarias de caudal del año 2017 entre el 2 de agosto a las 9:00 horas hasta el 28 de septiembre a las 0:00 horas, para un total de 1 360 registros suministrados por la CFE (Comisión Federal de Electricidad). Para el análisis de sensibilidad del tamaño del conjunto se utilizaron 5, 8, 10, 20, 50 y 100 miembros que se combinaron con los seis pasos.

La implementación del DKF, EnKF, DKF-EnKF y ARX(1,1) se hizo mediante rutinas en R (R Core Team, 2020), que generan los pronósticos

a seis pasos con DKF y ARX, y con 36 combinaciones entre pasos por tamaños de conjunto en EnKF y DKF-EnKF. El DKF se implementó para estimar el vector de estado que equivale a la función de respuesta de la cuenca o hidrograma unitario instantáneo (HUI) (Valdés *et al.*, 1980) y el EnKF realiza estimaciones escalares del caudal (Evensen, 2009), es decir, con el DKF se estiman los valores correspondientes a las columnas del HUI que se pasan a multiplicar por los valores de caudal; por su parte, el EnKF estima directamente valores de caudal. En los tres casos con filtros se consideraron las últimas observaciones de cada variable, es decir, un retraso autorregresivo (Valdés *et al.*, 1980). El modelo ARX se implementó de forma recursiva con base en una fracción de serie con 100 registros.

La cantidad adecuada de miembros en los conjuntos de EnKF y DKF-EnKF se determinó mediante análisis de sensibilidad, con base en la raíz del cuadrado medio del error (RMSE) (Quiroz *et al.*, 2019). La generación del ruido blanco (simulación de Monte Carlo) se hizo con el paquete *mvtnorm* (normal multivariante y distribución *t*) (Genz & Bretz, 2009). En los tres algoritmos evaluados, la varianza Q se supone constante (Simon, 2001) de valor cero (Morales-Velázquez, Aparicio, & Valdes, 2014) y R con valor cercano a cero (0.01) para dar flexibilidad a la convergencia del algoritmo (Welch & Bishop, 2006).

El ajuste de las series pronosticadas se evaluó mediante el coeficiente de Nash-Sutcliffe (Nash & Sutcliffe, 1970) y la raíz del cuadrado medio del error (RMSE) (Morales-Velázquez *et al.*, 2014). Los

supuestos de normalidad de errores del filtro de Kalman se verificaron mediante gráficos comparados con la curva normal estandarizada (González-Leiva, Ibáñez-Castillo, Valdés, Vázquez-Peña, & Ruiz-García, 2015). Los valores atípicos y su ubicación se determinó con base en los residuales estandarizados (Cryer & Chan, 2008).

Filtro de Kalman discreto

Es un algoritmo que permite la identificación de sistemas dinámicos lineales como un estimador óptimo de estados mediante un proceso recursivo (Kalman, 1960) (Figura 2). El DKF sirve de base para los algoritmos que tratan sistemas no lineales (Welch & Bishop, 2006).

Ecuación de estado

$$x_k = Ax_{k-1} + Bu_{k-1} + w_{k-1}$$

Ecuación de medición

$$z_k = Hx_k + v_k$$

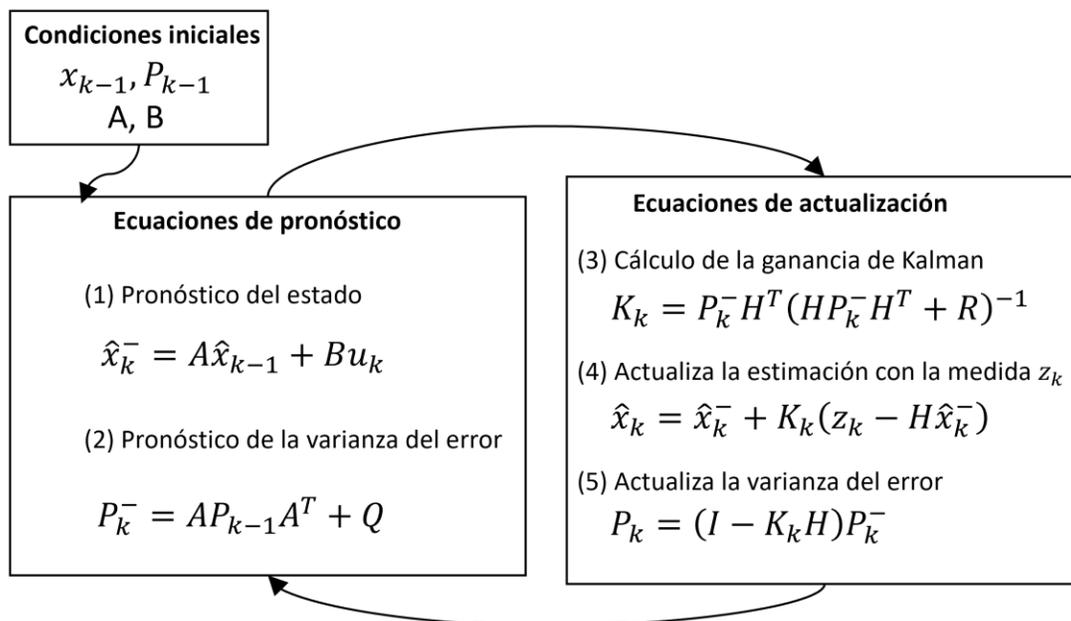


Figura 2. Algoritmo del filtro de Kalman discreto (Welch & Bishop, 2006).

La ecuación de estado tiene como entrada dos series horarias (n) de caudal de la estación Chapalagana y Platanitos. La matriz A ($n \times n$) y B ($n \times 1$) relacionan el estado del tiempo $k - 1$ con el estado actual en el tiempo k . Debido a que en el curso del río no se encuentran estructuras

hidráulicas, como represas que alteren de forma repentina los volúmenes de caudal, en este trabajo no se considera variable de control, por lo tanto sólo se tiene la matriz A, que se supone constante a lo largo del proceso. La matriz H en la ecuación de medición se compone por un vector fila de $1 \times n$ que contiene la última observación de cada variable de entrada. El valor pronosticado z_k se obtiene mediante la ecuación de medición, multiplicando la matriz H por el vector de estado x_k^- ($n \times 1$). A continuación se presenta la matriz A que relaciona el estado anterior con el actual:

$$A = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \quad (1)$$

La nueva medición para el tiempo k (Q_k) se utiliza para actualizar el algoritmo según los cambios que se estén presentando en el sistema; el valor pronosticado z_k (corresponde a un valor de caudal, tamaño p) lleva implícito el error de la medición w_{k-1} ; la ecuación de estado contiene el error del proceso v_k , los cuales deben cumplir el supuesto de normalidad:

$$w_k = N(0, Q) ; v_k = N(0, R) \quad (2)$$

Con las ecuaciones de estado y medición, el algoritmo inicia un ciclo que se repite de forma indefinida. En el tiempo $k - 1$ hace la estimación a priori (pronóstico) de los estados, que son actualizados (estimación a

posteriori) en el tiempo k . Los estados se toman como función de respuesta de la cuenca y con la estimación a posteriori se hace el pronóstico en el tiempo $k + 1$. Este ciclo se repite de forma indefinida realizando el pronóstico del tiempo $k + 1$ con base en la matriz H y el vector de estado actualizados hasta el tiempo k .

Filtro de Kalman de conjuntos

El filtro de Kalman de conjuntos (EnKF) es un estimador subóptimo, que se basa en simulaciones de Monte Carlo para la estimación del error estadístico (Evensen, 1994; Gillijns *et al.*, 2006; Rafieeiniasab, Seo, Lee, & Kim, 2014). Se supone distribución normal de los errores y las estimaciones se hacen con base en conjuntos que agrupan q valores generados al azar bajo distribución normal. Al igual que el DKF tiene dos grupos de ecuaciones: análisis y predicción. En la Figura 3, las dos primeras ecuaciones tienen sus equivalentes en el DKF, y la última corresponde al promedio de los miembros que se supone como mejor pronóstico. El segundo grupo, de forma general, consta de cuatro ecuaciones, donde se estiman errores y probabilidades que son insumo para calcular la ganancia de Kalman (Gillijns *et al.*, 2006).

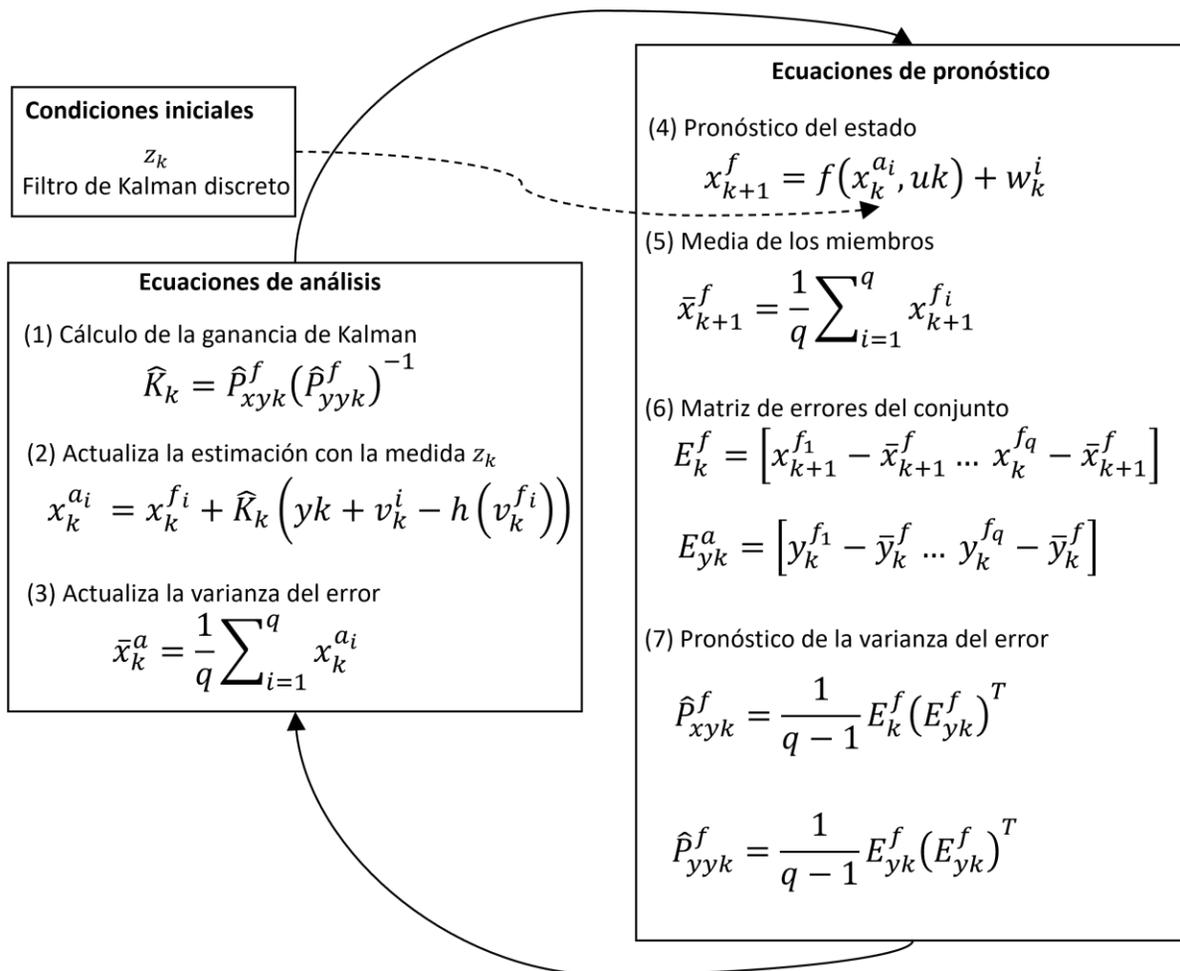


Figura 3. Algoritmo del filtro de Kalman de conjuntos (Gillijns *et al.*, 2006).

Los errores w_k^i y v_k^i corresponden al ruido que contiene el proceso y la medición, respectivamente; se suponen como ruido blanco con media cero y varianza Q y R (Figura 3), de igual forma que en la Ecuación (2).

Para el pronóstico de caudales, la función de respuesta (x_{k+1}^f) del EnKF se estableció incorporando ruido blanco w_k^i a los estados previos x_{k-1}^{ai} :

$$x_{k+1}^f = x_k^{ai} + w_k^i \quad (3)$$

El ruido en las mediciones se genera agregando a la medición en el tiempo k , q desviaciones con distribución normal:

$$y_k^i = y_k + v_k^i \quad (4)$$

El vector y_k^i de $1 \times q$ corresponde a las q mediciones ruidosas; y_k es la medición en el tiempo k , y el subíndice i representa la cantidad de miembros que corresponde a $i = 1, 2, \dots, q$. Es decir, se generan q valores al azar bajo distribución normal, que se suman al valor observado, y resulta en un vector de tamaño q con el valor observado como media. Se supone que a mayor cantidad de miembros mejor será el ajuste del pronóstico debido a que se puede obtener mejor estimación de la distribución de probabilidades (Leutbecher, 2019); sin embargo, aumentar la cantidad de miembros implica mayor esfuerzo computacional, por tanto se debe hacer un análisis de sensibilidad para determinar la cantidad que obtenga los mínimos errores con esfuerzo

computacional aceptable. Para estudios hidrológicos se sugiere entre 50 y 300 miembros por conjunto (Gillijns *et al.*, 2006; Quiroz *et al.*, 2019).

Filtro de Kalman discreto y filtro de Kalman de conjuntos

La integración del DKF y EnKF se hizo mediante la ecuación de estado del algoritmo del EnKF. El DKF genera un pronóstico para el tiempo $k + 1$ (\hat{Q}_{k+1}), a partir del cual se determina la diferencia con respecto al estado previo de EnKF (x_{k-1}^{ai}), que finalmente se suma al estado previo x_{k-1}^{ai} , además se le agrega ruido blanco mediante simulación de Monte Carlo w_k^i :

$$x_{k+1}^f = x_{k-1}^{ai} + (\hat{Q}_{k+1} - x_{k-1}^{ai}) + w_k^i \quad (5)$$

Modelo autorregresivo de primer orden (ARX(1,1)) y DKF

El modelo autorregresivo de primer orden, también conocido como modelo de Markov, es una de las primeras aproximaciones para el estudio de series de tiempo y se basa en la autocorrelación que se presenta dentro de la misma serie de datos (Box, Jenkins, Reinsel, & Ljung, 2016; Bras & Rodríguez-Iturbe, 1985). Algebraicamente se describe de la siguiente forma:

$$y_{k+1} = \sum_{i=0}^{na} \alpha_i y_{k-i} + \sum_{j=0}^{nb} \beta_j \gamma_{k-j} + e_{k+1} \quad (6)$$

Donde y_{k+1} corresponde al valor pronosticado; α_i y β_j son los parámetros del modelo; y_k y γ_k corresponden a la variable de entrada y variable exógena, respectivamente. La estimación de los parámetros se hace por mínimos cuadrados y se requiere una sección de serie de al menos 50 registros. En el modelo ARX(na , nb), na y nb representan los retrasos autorregresivos que se utilizan en cada variable. El modelo ARX

estima los parámetros α y β , que son incorporados en la matriz A del algoritmo DKF (Figura 2), y la actualización se aplica al caudal para el pronóstico en el siguiente ciclo.

Resultados y discusión

En total se evaluaron 36 combinaciones resultado de seis pasos (L) por seis tamaños de conjunto. Los tres algoritmos se ejecutaron con 1 360 registros horarios de caudal provenientes de las estaciones Chapalagana y Platanitos, para obtener pronósticos de caudal en la estación Chapalagana.

En la Figura 4 se muestra el análisis de sensibilidad que permite identificar el tamaño adecuado de los conjuntos para los algoritmos de EnKF (línea punteada) y DKF-EnKF (línea continua), mediante la relación entre el valor de RMSE a medida que aumenta la cantidad de miembros por conjunto. El conjunto con cinco miembros presenta el RMSE más alto en cada uno de los pasos L ; cuando se tienen 8 y 10 miembros se dan variaciones con ligera tendencia a disminución, y los conjuntos con 20, 50 y 100 miembros reflejan convergencia y estabilidad; por tanto, en el

caso de la serie bajo estudio para optimizar el ajuste y la carga computacional se pueden usar 20 miembros. La combinación DKF-EnKF en los seis pasos L tienen mayores valores de RMSE que EnKF; las diferencias se hacen cada vez más notables en pasos más grandes, sin embargo, la convergencia se alcanza a partir de 20 miembros por conjunto.

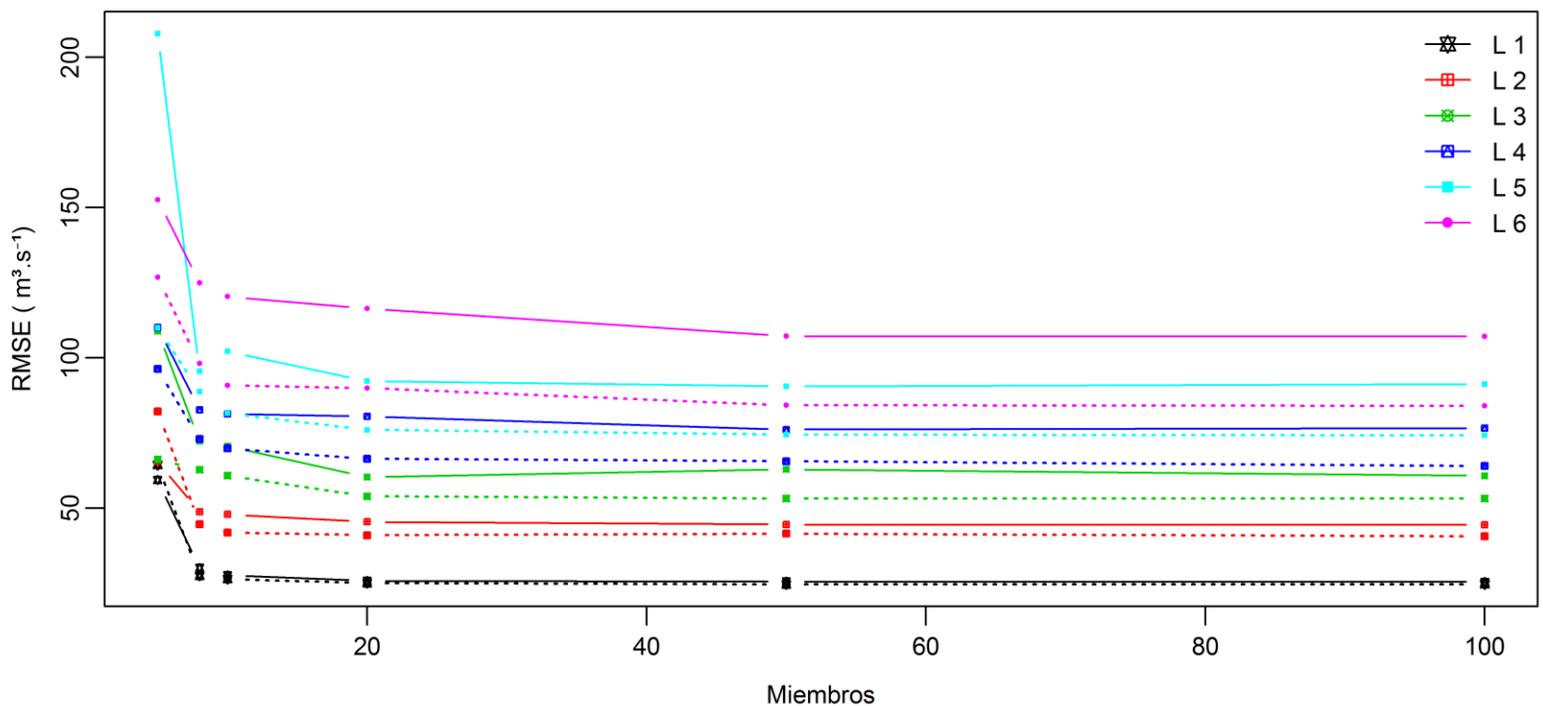


Figura 4. Raíz del cuadrado medio del error con diferentes tamaños de conjunto.

La importancia de la definición de la cantidad adecuada de miembros radica en obtener el mejor ajuste posible sin agregar carga computacional que no mejore el comportamiento del algoritmo. Cuando se tienen conjuntos grandes, la carga computacional es alta; por el contrario, si se usan conjuntos pequeños se puede perder ajuste entre la serie observada y la pronosticada (Gillijns *et al.*, 2006; Quiroz *et al.*, 2019).

Los resultados que se presentan son con base en 20 miembros por cada conjunto. En la Tabla 1 se tienen los indicadores estadísticos de ajuste de la serie observada contra las series pronosticadas.

Tabla 1. Resumen de estadísticos para la aplicación de DKF, EnKF, DKF-EnKF y ARX-DKF.

	Índice	L1	L2	L3	L4	L5	L6	DEp
DKF	RMSE	24.76	39.55	50.46	59.21	67.20	74.49	18.36
	Nash-Sutcliffe	0.9869	0.9666	0.9457	0.9252	0.9037	0.8816	0.0393
	Media	161.55	161.71	161.39	161.57	160.74	160.67	0.45
	DEs	217.58	216.33	213.69	212.47	209.96	208.30	3.58
EnKF	RMSE	25.08	41.00	53.99	66.40	76.04	89.93	23.67
	Nash-Sutcliffe	0.9866	0.9641	0.9378	0.9060	0.8767	0.8275	0.0587
	Media	160.65	160.55	160.88	160.15	160.93	161.27	0.38
	DEs	216.48	215.81	217.25	216.67	218.17	219.18	1.23
DKF-EnKF	RMSE	25.83	45.45	60.29	80.57	92.24	116.41	32.89

	Nash-Sutcliffe	0.9858	0.9559	0.9225	0.8616	0.8185	0.7110	0.1013
	Media	160.28	159.51	159.78	160.34	161.64	162.09	1.03
	DEs	216.21	217.95	219.67	226.07	229.10	239.75	8.84
ARX-DKF	RMSE	26.07	44.11	59.70	72.60	83.23	91.65	24.71
	Nash-Sutcliffe	0.9855	0.9586	0.9243	0.8882	0.8532	0.8222	0.0625
	Media	163.32	165.27	166.79	168.01	169.71	170.52	2.72
	DEs	219.90	222.08	224.70	226.56	228.20	228.34	3.42

RMSE: raíz del error cuadrático medio

Nash-Sutcliffe: índice de Nash

Media: promedio

DEs: desviación estándar en cada serie pronosticada

DEp: desviación estándar entre los seis pasos por cada algoritmo

El promedio y desviación estándar de la serie observada son 160.95 y 216.67, respectivamente

La implementación del modelo ARX(1,1) y el modelo ARX-DKF con actualización recursiva obtienen similares resultados debido a que los parámetros de ARX se multiplican por los valores de caudal en el tiempo k , operación que se ejecuta de manera matricial dentro de la ecuación de pronóstico de estado del algoritmo del DKF. Los resultados que se presentan se refieren únicamente al modelo ARX-DKF.

Para el paso $L = 1$, los algoritmos con filtro de Kalman generan pronósticos con coeficientes de Nash-Sutcliffe (NS) superiores a 0.9855, que es ligeramente mejor a lo encontrado en los pronósticos de los caudales en la presa Ángel Albino Corzo (NS = 0.9774) (Morales-Velázquez *et al.*, 2014). Sin embargo, a medida que el paso L se hace

más grande, se presentan diferencias crecientes entre los algoritmos. El DKF se mantiene con el mejor ajuste, seguido del EnKF y ARX-DKF y, por último, la integración DKF-EnKF con un descenso pronunciado hasta 0.7110 en el paso $L = 6$ (Tabla 1). El RMSE se puede interpretar con unidades en m^3/s y presenta un comportamiento de ajuste similar al evidenciado por NS. Tanto NS como RMSE demuestran que el menor error se logra con DKF seguido de EnKF, ARX-DKF, y DKF-EnKF. En términos de NS y RMSE, se puede advertir que los algoritmos DKF y EnKF generan pronósticos con mejor ajuste respecto al modelo ARX-DKF; además, el empleo del DKF no requiere de largos tramos de serie para el entrenamiento, lo que brinda mayor simplicidad del modelo.

De acuerdo con el resumen estadístico de la Tabla 1, la media y la desviación estándar en los seis pasos tienden a sobreestimarse con diferentes magnitudes respecto a la serie observada. La mayor estabilidad de la media la tiene el algoritmo de EnKF con desviación estándar (DEp) de 0.38, con leve diferencia con DKF, que tiene desviación de 0.45 entre los seis pasos, mientras que ARX-DKF y DKF-EnKF tienen mayor desviación estándar de 1.03 y 2.72, respectivamente. Por otra parte, la dispersión representada por la desviación estándar entre las series pronosticadas (DEp) confirma que EnKF guarda mayor similitud con la serie observada en los seis pasos. El aparente mejor ajuste de los pronósticos de EnKF se debe al efecto de desplazamiento respecto a la observada sin mayor cambio en los picos máximos o mínimos de tal manera que no hay efecto notable en la media y desviación estándar.

Las diferencias entre los pronósticos generados se hacen notables de forma gráfica, y para una comparación detallada en la Figura 5 y Figura 6 se presentan dos avenidas a diferentes pasos.

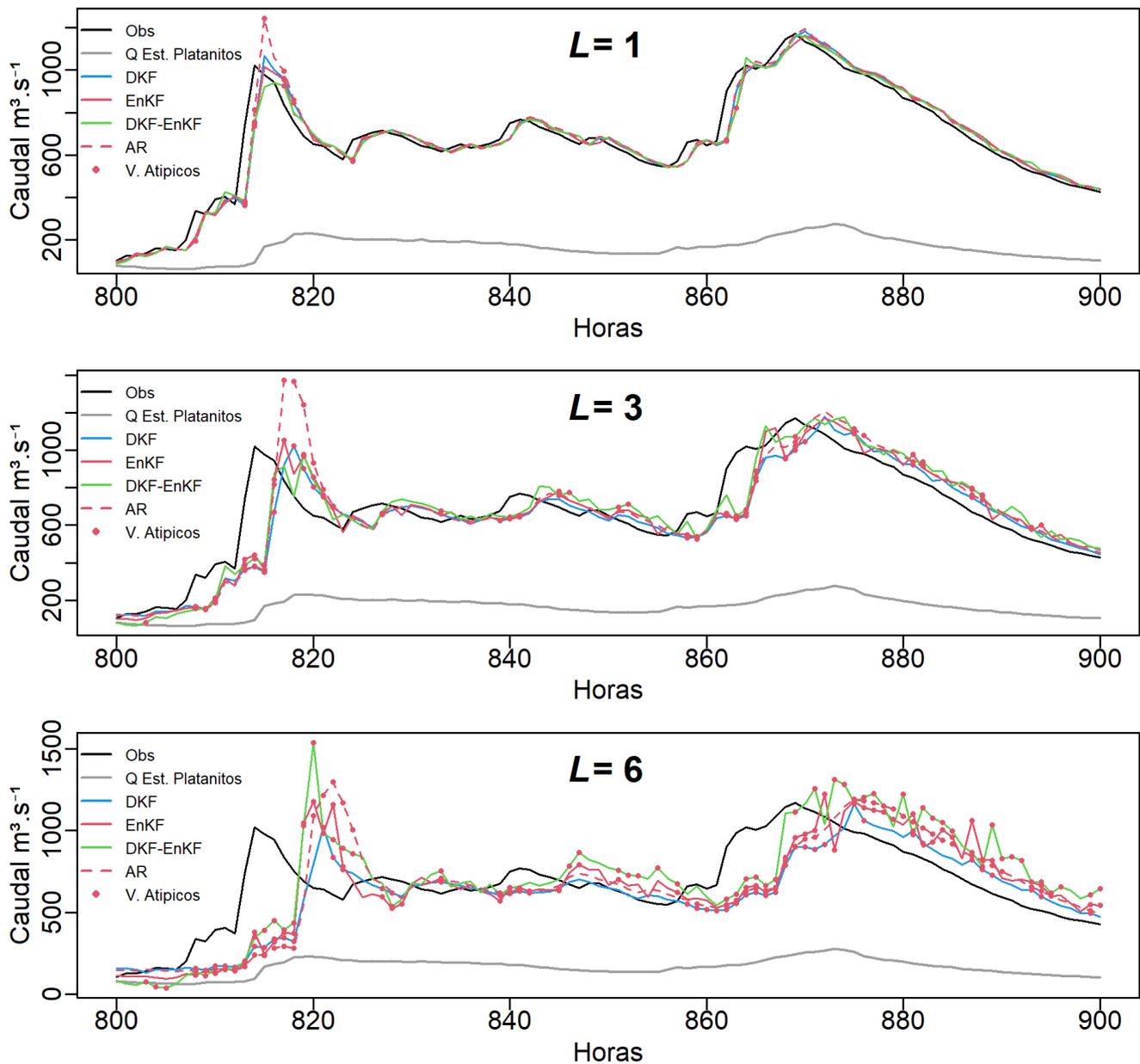


Figura 5. Caudal observado y pronosticado con DKF, ENKF y DKF-EnKF (avenida del 4/9/2017 16:00 h al 8/9/2017 20:00 h).

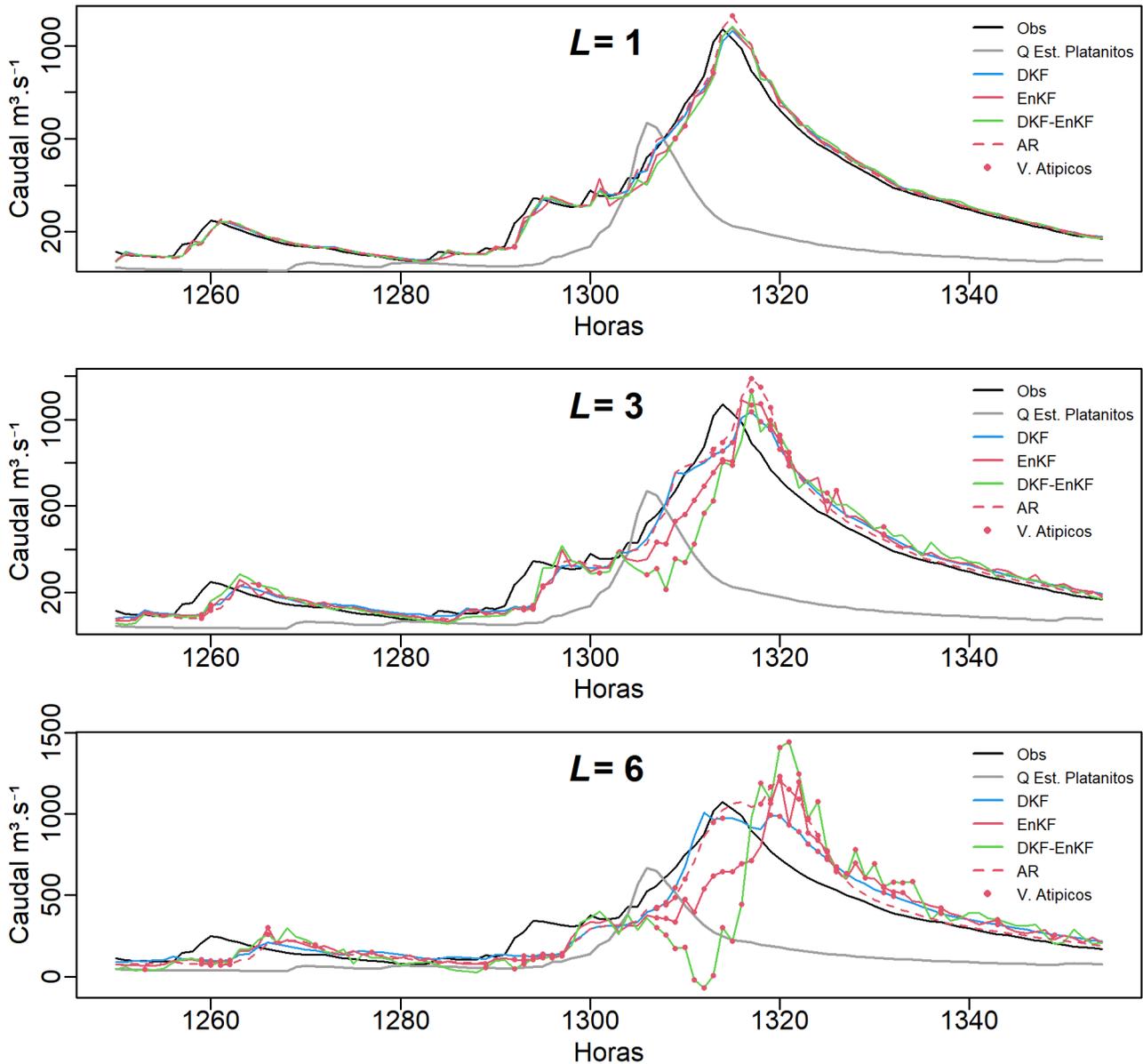


Figura 6. Caudal observado y pronosticado con DKF, En KF y DKF-EnKF (avenida del 23/09/17 10:00 h al 28/09/17 00:00 h).

Los pronósticos para el tiempo $k + L$ se desplazan a un valor similar a L debido a que la actualización incorpora la última medición correspondiente al tiempo k ; en otras palabras, a mayor distancia entre el conjunto de datos medido y el momento $k + L$ pronosticado se obtiene disminución progresiva del ajuste entre el paso $L = 1$ a $L = 6$, y se manifiesta como desplazamiento de la serie pronosticada sobre la serie observada. No obstante, en el evento mostrado en la Figura 6, en el pronóstico generado por DKF y ARX el desplazamiento es casi nulo; esto se debe a que el retraso entre el pico del 25/09/2017 18:00 en la estación Platanitos y el pico del 26/09/17 02:00 en la estación Chapalagana es de ocho horas. Es decir, DKF estima y actualiza el hidrograma unitario instantáneo (HUI) con base en registros de caudal de la estación Platanitos del tiempo actual (k), que en la estación Chapalagana se mostrarán ocho horas después, como consecuencia del tiempo de concentración. La proximidad del tamaño del retraso y el paso $L = 6$ permite que la actualización se haga con registros correlacionados, disminuyendo el desplazamiento y dando mejor ajuste en el pico máximo. Por su parte, ARX genera el pronóstico con similar ajuste que DKF, pero con un segmento de serie de mayor extensión, es decir, el modelo DKF representa mayor sencillez para implementarlo.

Por otra parte, los algoritmos de EnKF y DKF-EnKF no obtienen mejoras debido a que EnKF hace estimaciones de escalares y no se incluye el retraso entre series. A diferencia de DKF, el pronóstico generado por

DKF-EnKF se invierte de tal forma que genera un pico mínimo (hora 1311, 25/09/2017 23:00), que se debe a la actualización con el pico máximo de caudal que presenta la estación Platanitos (hora 1306, 25/09/2017 18:00) y que el algoritmo de EnKF trata de corregir. Una posible adecuación es la implementación de EnKF para la estimación de estados en lugar de escalares que también puede agregar capacidad para tratamiento de no linealidad además de la integración del filtro de Kalman con modelos de simulación distribuidos, como lo proponen Rafieeinasab *et al.* (2014). En realidad, para modelar en Kalman, simplifica los parámetros que incluye el modelo distribuido Sacramento Soil Moisturing Accounting (SAC) y al final compara sus filtros de Kalman de conjuntos con los resultados de ese modelo. El trabajo de Rafieeinasab *et al.* (2014) es muy interesante desde dos puntos de vista: 1) el trabajo compara dos filtros de Kalman de conjuntos, el sencillo (ENKF) y el de máxima verosimilitud (MELF); 2) considera todo un sistema que relaciona pronóstico de caudales con humedad del suelo y dicha humedad está vista desde un balance que considera una lluvia media en la cuenca, una evaporación potencial media en la cuenca. Desde el punto de vista conceptual, nuestro modelo es más sencillo, porque sólo realiza la asimilación de datos de los caudales medidos. Sin embargo, una crítica constructiva a sus resultados es la siguiente: su cuenca es pequeña (435 km²) comparada con nuestra cuenca (12 076 km²). El orden del tamaño de sus caudales es pequeño (entre 100 y 200 m³/s), pero su error (RSME) es muy alto, pudiendo llegar a ser de 45 m³/s. En nuestro trabajo, los

caudales son del orden de 200 a 1 100 m³/s y sin embargo el máximo RSME es 116 m³/s en el peor caso.

Los pronósticos pueden alcanzar mejor ajuste con la incorporación dinámica del tiempo de retraso entre series, tratamiento de variables con distintos tipos de observación y proporciones (Meng *et al.*, 2017), e incorporación de las variables que afectan el tiempo de retraso, como la ubicación y dirección de los eventos de precipitación, y la condición de humedad antecedente del suelo de la cuenca. La precipitación, presión atmosférica, temperatura y humedad relativa son variables que pueden ayudar a tener mejores ajustes en el pronóstico. Asimismo, dado que los caudales que se miden en la parte alta de la cuenca son proporcionalmente menores debido a la menor área de captación disponible, es conveniente incluir en el modelo parámetros que ayuden a equiparar las magnitudes de los caudales. Además, es determinante tener registros de estaciones meteorológicas distribuidas en el área de la cuenca para definir y actualizar la función de respuesta de la cuenca (HUI) (González-Leiva *et al.*, 2015).

La dispersión entre la serie observada y las series pronosticadas (Figura 7) muestran similitud entre los algoritmos. A medida que se aumenta L , el valor de la pendiente (en la ecuación lineal, Figura 7) va disminuyendo, es decir, que en general los valores pronosticados son subestimados respecto al valor observado.

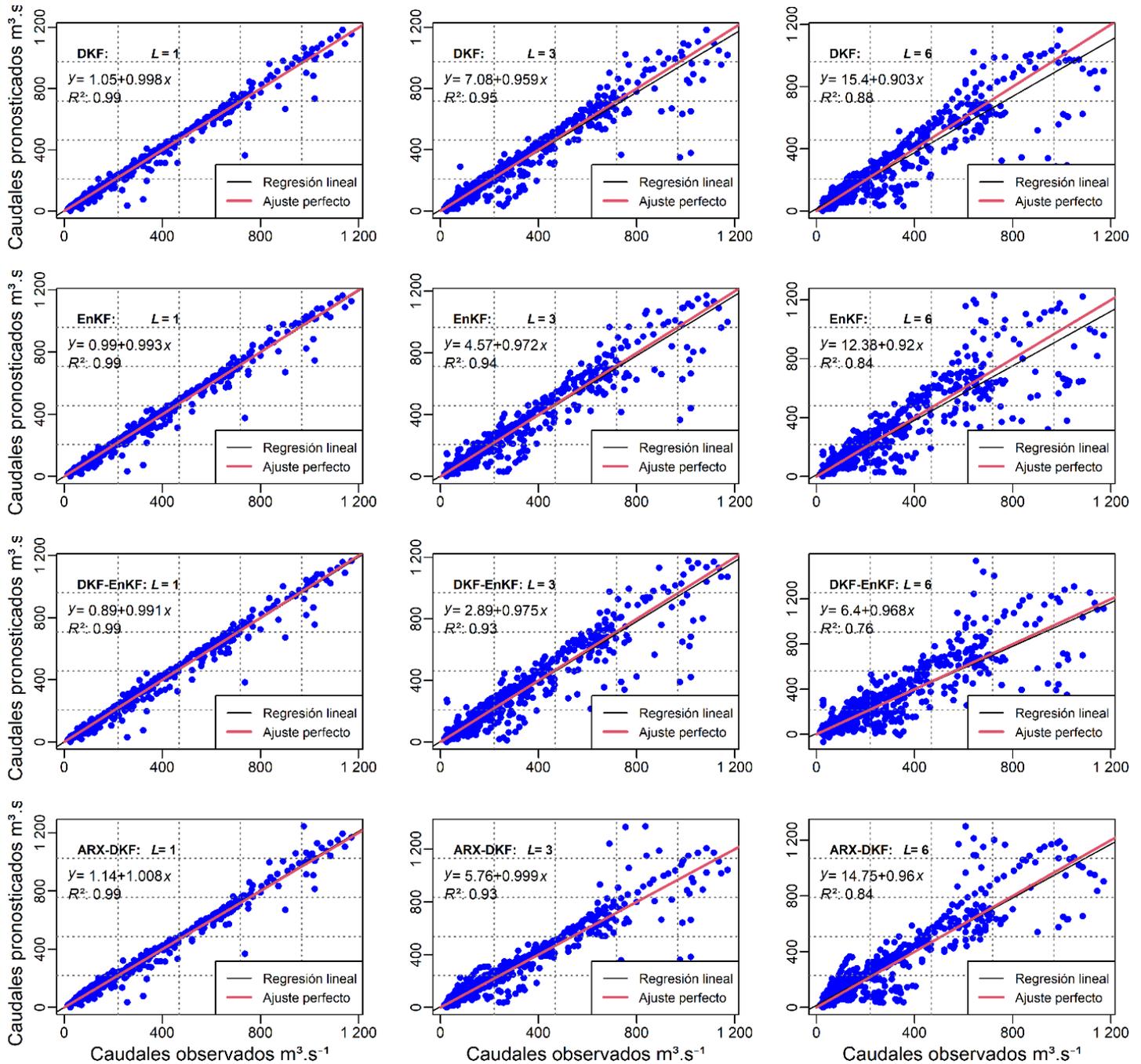


Figura 7. Caudales observados *versus* pronosticados.

Cuando $L = 1$, el ajuste lineal entre los valores observados y pronosticados tiene el coeficiente de determinación de 0.99, que es el más alto y similar entre los cuatro algoritmos (Tabla 2). En los pasos $L = 2$ en adelante, la dispersión aumenta debido al distanciamiento del conjunto de datos utilizado para el pronóstico. El DKF mantiene el mejor ajuste en los seis pasos, con marcada similitud con EnKF hasta $L = 3$. Cuando $L = 6$ DKF tiene el coeficiente de determinación de 0.88, que es el más alto. El comportamiento del ajuste lineal es congruente con el valor de NS mostrado en la Tabla 1. Los puntos alejados de la línea central corresponden a pronósticos donde el caudal de la estación Chapalagana presenta cambios bruscos para conformar una avenida. Dichos puntos generan pronósticos con magnitudes atípicas. En las horas $k = 813$ y $k = 814$ se presentan las mayores desviaciones respecto al valor observado y equivalen a la etapa inicial de conformación de la avenida que se muestra en la Figura 5.

Tabla 2. Coeficientes de determinación del caudal observado contra los pronosticados.

	L1	L2	L3	L4	L5	L6
DKF	0.99	0.97	0.95	0.93	0.90	0.88
EnKF	0.99	0.96	0.94	0.91	0.88	0.84

DKF-EnKF	0.99	0.96	0.93	0.87	0.84	0.76
ARX-DKF	0.99	0.96	0.93	0.90	0.87	0.84

A fin de verificar la normalidad de los residuales generados a partir de los pronósticos del DKF, EnKF y DKF-EnKF, en la Figura 8 se muestran los histogramas y el ajuste de la curva normal para cada algoritmo en los diferentes pasos. Es importante considerar que es ideal que los residuales tengan distribución normal, pero en pocos casos se obtiene así, por tanto se admite una aproximación (Chatfield, 2001).

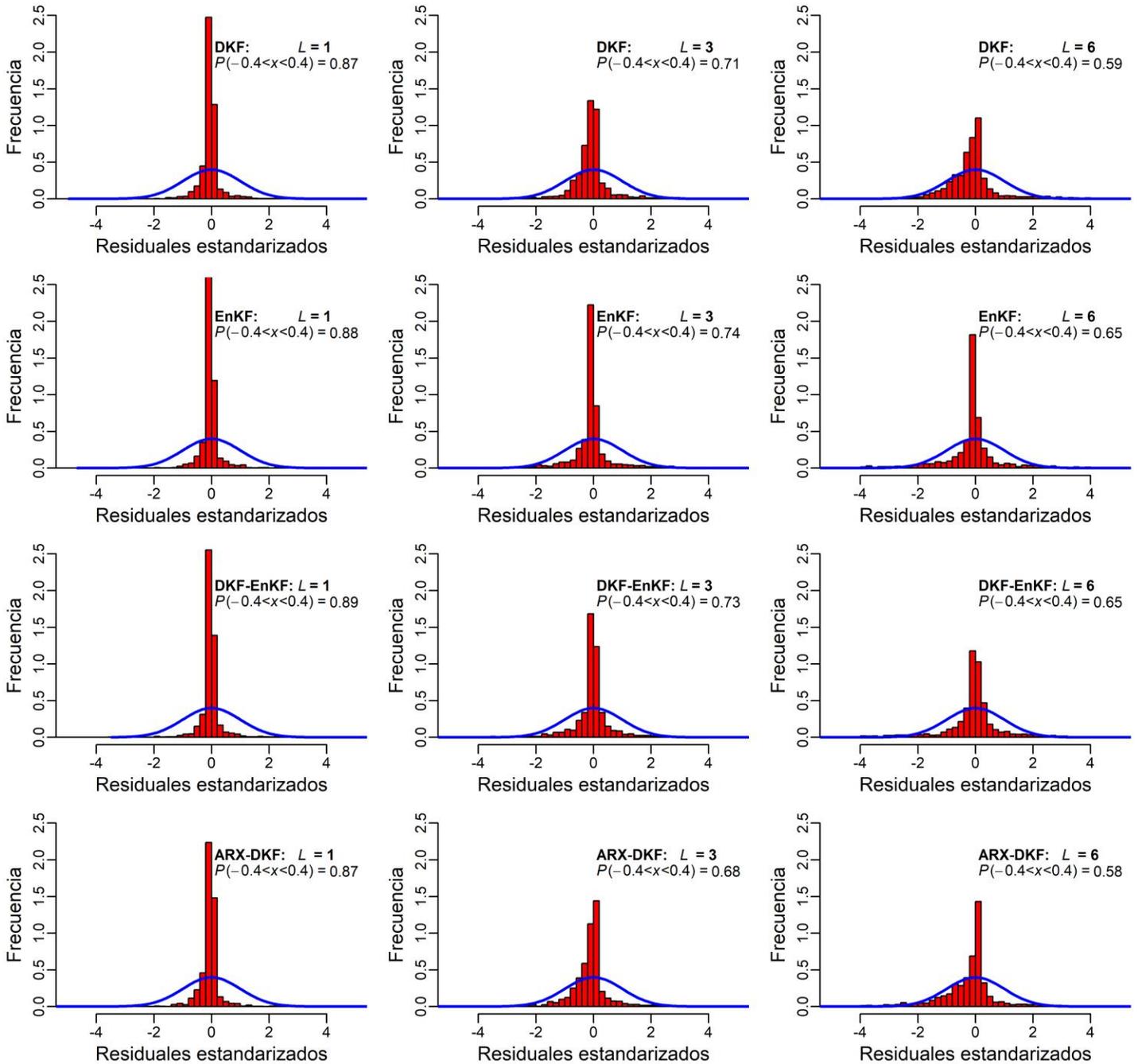


Figura 8. Histograma de residuales del DKF.

Se puede notar distribución simétrica con mayor concentración de pronósticos en la parte central. En el rango entre -0.4 y 0.4 de los residuales estandarizados, $L = 1$ agrupa entre 86.03 y 87.3 % y en $L = 6$ se agrupan entre el 56.67 y 63.02 % del total de las estimaciones. Suponiendo que se tiene distribución normal aproximada en los residuales, se hizo la estandarización para determinar la presencia de valores atípicos mayores y menores a dos desviaciones estándar a partir del cero. En la Figura 9 se agrupan los residuales estandarizados para 1 y 6 pasos, resultado de los pronósticos realizados con DKF, EnKF y DKF-EnKF. Los valores atípicos se ubican en los ascensos y descensos de los picos, y se muestran como puntos rojos en la Figura 5 y Figura 6. En concordancia con los índices estadísticos de la Tabla 1, a medida que el paso L se hace más grande, los valores atípicos tienen mayor frecuencia. En el paso $L = 1$ se presenta similitud que se va distorsionando hasta tener un rango de 55 entre DKF y DKF-EnKF en el paso $L = 6$.

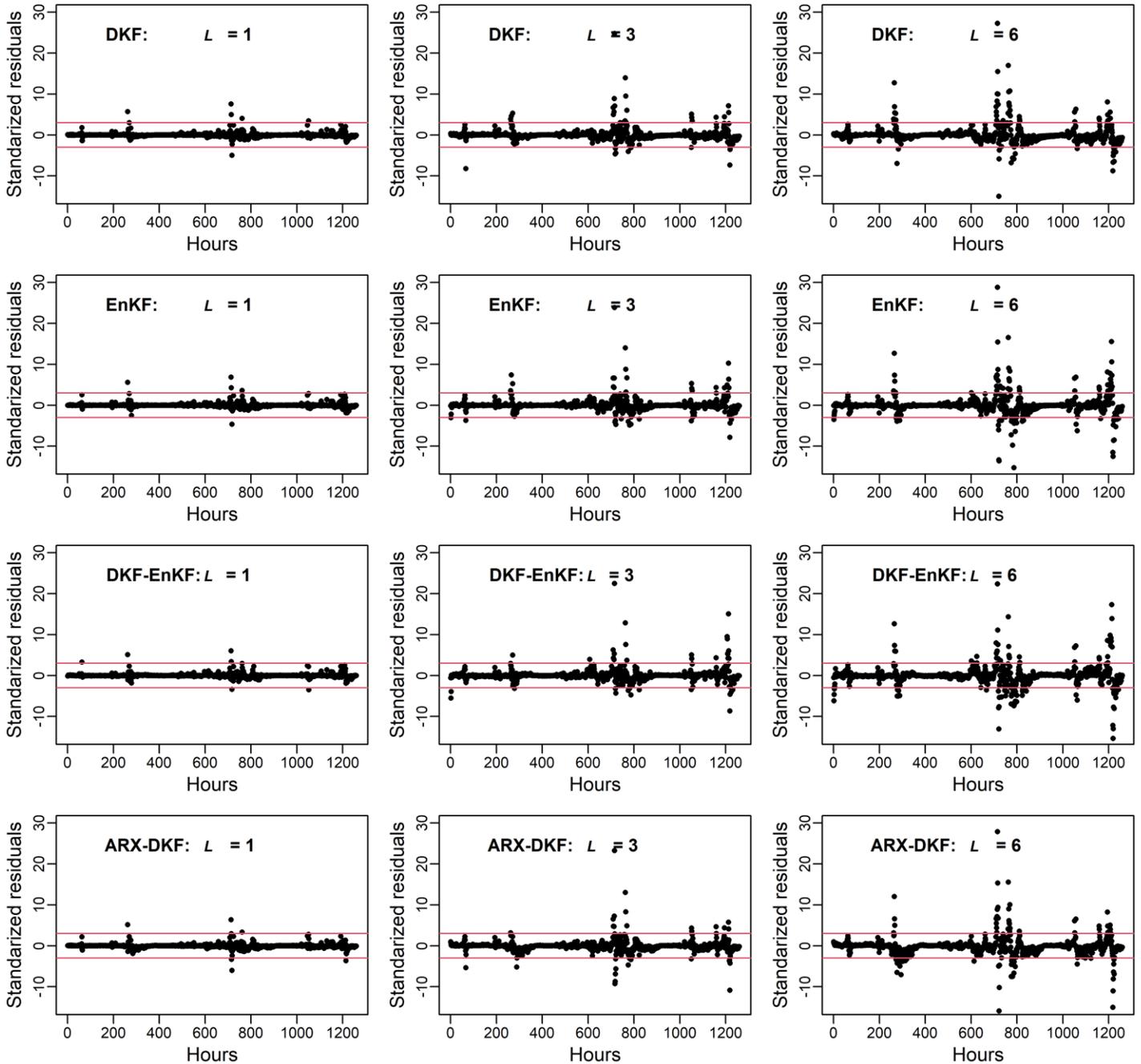


Figura 9. Residuales estandarizados.

La presencia de datos atípicos está estrechamente relacionada con el nivel de ajuste. Como se muestra en la Tabla 3, DKF presenta la menor cantidad de valores atípicos asociados con la capacidad que tienen para gestionar los retrasos entre las series (Tabla 1). Por el contrario, DKF-EnKF, con excepción del paso $L = 1$, tiene la mayor presencia de valores atípicos debido a la aplicación de ganancia por el DKF y en secuencia por el EnKF que proyecta de forma ampliada los cambios bruscos en el caudal; asimismo, el pico mínimo que se muestra en la Figura 7 favorece al aumento de valores atípicos.

Tabla 3. Cantidad de valores atípicos por algoritmo y paso.

	L1	L2	L3	L4	L5	L6
DKF	13	41	63	84	96	112
	1.03 %	3.25 %	5.00 %	6.67 %	7.62 %	8.89 %
EnKF	15	41	70	99	112	143
	1.19 %	3.25 %	5.56 %	7.86 %	8.89 %	11.35 %
DKF-EnKF	14	46	66	105	132	167
	1.11 %	3.65 %	5.24 %	8.33 %	10.48 %	13.25 %
ARX-DKF	15	42	62	87	107	137
	1.19 %	3.33 %	4.92 %	6.90 %	8.49 %	10.87 %

Conclusiones

El DKF obtiene el mejor ajuste debido a la estimación de estados que permite el aprovechamiento del retraso entre series como base para la actualización. El EnKF estima valores escalares y genera suavizado con efecto de desplazamiento, que mantiene estables la media y la desviación estándar, pero con ajuste inferior al DKF. De esta manera, la utilización del DKF donde se incorpore de forma recursiva el tiempo de retraso para actualización puede mejorar ajuste en los picos, además, la simplicidad para programar y operar el DKF los hace viables para pronóstico de caudales a corto plazo. El pronóstico de caudales horarios, con todos los filtros de Kalman analizados, es muy bueno con coeficientes de Nash-Sutcliffe muy altos y errores bajos medidos como RSME.

Para mejorar el ajuste de los pronósticos es importante adelantar trabajos de investigación donde se incluya la actualización recursiva del tiempo de retraso entre las series de diferentes estaciones teniendo en cuenta que se presentan variaciones a lo largo del tiempo. Además, se pueden evaluar los pronósticos con tamaños de paso de varias horas, por

ejemplo, agrupar seis horas mediante valores como la media o máximo, a fin de que cada paso equivalga a seis horas y un pronóstico a seis pasos equivalga a 36 horas, de esta forma se da más flexibilidad al modelo y se puede ampliar el periodo de pronóstico.

Agradecimientos

A la Comisión Federal de Electricidad por permitir acceso a su base de datos hidrometeorológica a través del sitio web administrado por el Instituto Nacional de Electricidad y Energías Limpias durante el año 2018.

Referencias

- Abaza, M., Anctil, F., Fortin, V., & Turcotte, R. (2015). Exploration of sequential streamflow assimilation in snow dominated watersheds. *Advances in Water Resources*, 86, 414-424. Recuperado de <https://doi.org/10.1016/j.advwatres.2015.10.008>
- Box, G., Jenkins, G., Reinsel, G., & Ljung, G. (2016). *Time series analysis: Forecasting and control* (5th ed.). New Jersey, USA: Wiley.
- Brandhorst, N., Erdal, D., & Neuweiler, I. (2017). Soil moisture prediction with the ensemble Kalman filter: Handling uncertainty of soil hydraulic parameters. *Advances in Water Resources*, 110(February), 360-370. Recuperado de <https://doi.org/10.1016/j.advwatres.2017.10.022>
- Bras, R., & Rodríguez-Iturbe, I. (1985). *Random functions and hydrology*.

Massachusetts, USA: Addison-Wesley Publishing Company.

Chatfield, C. (2001). Prediction intervals for time-series forecasting. In: Armstrong, J. S. (ed.). *A handbook for researchers and practitioners*. Vol. 30 (pp. 475-494). Springer, U.S. Recuperado de <https://doi.org/10.1007/978-0-306-47630-3>

Clark, M. P., Rupp, D. E., Woods, R. A., Zheng, X., Ibbitt, R. P., Slater, A. G., Schmidt, J., & Uddstrom, M. J. (2008). Hydrological data assimilation with the ensemble Kalman filter: Use of streamflow observations to update states in a distributed hydrological model. *Advances in Water Resources*, 31(10), 1309–1324. Recuperado de <https://doi.org/10.1016/j.advwatres.2008.06.005>

Conagua, Comisión Nacional del Agua. (2008). *Estadísticas del agua en México*. México, DF, México: Secretaría del Medio Ambiente y Recursos Naturales.

Cryer, J. D., & Chan, K.-S. (2008). *Time series analysis with applications in R* (2nd ed.). New York, USA: Springer.

Evensen, G. (2009). *Data assimilation: The ensemble Kalman filter* (2nd ed.). Berlin, Germany: Springer Berlin Heidelberg. Recuperado de <https://doi.org/10.1007/978-3-642-03711-5>

Evensen, G. (2003). The ensemble Kalman filter: Theoretical formulation and practical implementation. *Ocean Dynamics*, 53(4), 343-367. Recuperado de <https://doi.org/10.1007/s10236-003-0036-9>

Evensen, G. (1994). Sequential data assimilation with a nonlinear quasi-

geostrophic model using Monte Carlo methods to forecast error statistics. *Journal of Geophysical Research*, 99(C5), 10143. Recuperado de <https://doi.org/10.1029/94jc00572>

Genz, A., & Bretz, F. (2009). *Computation of multivariate normal and t probabilities*. New York, USA: Springer-Verlag.

Gillijns, S., Mendoza, O. B., Chandrasekar, J., De-Moor, B. L. R., Bernstein, D. S., & Ridley, A. (2006). What is the ensemble Kalman filter and how well does it work? *2006 American Control Conference*. Recuperado de <https://doi.org/10.1109/ACC.2006.1657419>

González-Leiva, F., Ibañez-Castillo, L. A., Valdés, J. B., Vázquez-Peña, M. A., & Ruiz-García, A. (2015). Pronóstico de caudales con filtro de Kalman discreto en el río Turbio. *Tecnología y ciencias del agua*, 6(4), 5-24.

INEGI, Instituto Nacional de Estadística y Geografía. (2010). *Hidrografía*. Recuperado de <https://www.inegi.org.mx/temas/hidrografia/default.html#Descargas>

Kalman, R. E. (1960). A new approach to linear filtering and prediction problems. *Journal of Basic Engineering*, 82(Series D), 35-45.

Leutbecher, M. (2019). Ensemble size: How suboptimal is less than infinity? *Quarterly Journal of the Royal Meteorological Society*, 145(S1), 107-128. Recuperado de <https://doi.org/10.1002/qj.3387>

Liu, Y., Weerts, A. H., Clark, M., Hendricks-Franssen, H.-J., Kumar, S.,

- Moradkhani, H., Seo, D.-J., Schwanenberg, D., Smith, P., van Dijk, A. I. J. M., van Velzen, N., He, M., Lee, H., Noh, S. J., Rakovec, O., & Restrepo, P. (2012). Advancing data assimilation in operational hydrologic forecasting: Progresses, challenges, and emerging opportunities. *Hydrology and Earth System Sciences*, 16(10), 3863-3887. Recuperado de <https://doi.org/10.5194/hess-16-3863-2012>
- Liu, Y., & Gupta, H. V. (2007). Uncertainty in hydrologic modeling: Toward an integrated data assimilation framework. *Water Resources Research*, 43(7), 1-18. Recuperado de <https://doi.org/10.1029/2006WR005756>
- Maxwell, D. H., Jackson, B. M., & McGregor, J. (2018). Constraining the ensemble Kalman filter for improved streamflow forecasting. *Journal of Hydrology*, 560, 127-140. Recuperado de <https://doi.org/10.1016/j.jhydrol.2018.03.015>
- Meng, S., Xie, X., & Liang, S. (2017). Assimilation of soil moisture and streamflow observations to improve flood forecasting with considering runoff routing lags. *Journal of Hydrology*, 550, 568-579. Recuperado de <https://doi.org/10.1016/j.jhydrol.2017.05.024>
- Morales-Velázquez, M. I., Aparicio, J., & Valdes, J. B. (2014). Pronóstico de avenidas utilizando el filtro de kalman discreto. *Tecnología y ciencias del agua*, 5(2), 85-110.
- Nash, J. E., & Sutcliffe, J. V. (1970). River flow forecasting through conceptual models part I — A discussion of principles. *Journal of*

Hydrology, 10(3), 282-290. Recuperado de
[https://doi.org/10.1016/0022-1694\(70\)90255-6](https://doi.org/10.1016/0022-1694(70)90255-6)

Quiroz, K., Collischonn, W., & Paiva, R. C. D. (2019). Data assimilation using the ensemble Kalman filter in a distributed hydrological model on the Tocantins River, Brasil. *RBRH*, 24, 1-15. Recuperado de <https://doi.org/10.1590/2318-0331.241920180031>

R Core Team. (2020). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. Recuperado de <http://www.R-project.org/>. <https://www.r-project.org/>

Rafieeinassab, A., Seo, D. J., Lee, H., & Kim, S. (2014). Comparative evaluation of maximum likelihood ensemble filter and ensemble Kalman filter for real-time assimilation of streamflow data into operational hydrologic models. *Journal of Hydrology*, 519(PD), 2663-2675. Recuperado de <https://doi.org/10.1016/j.jhydrol.2014.06.052>

Simon, D. (2001). Kalman filtering. *Embedded Systems Programming*, June, 72-79.

SMN, Servicio Meteorológico Nacional. (2019). *Sistema de información climática computarizada CLICOM*. Ciudad de México, México: Servicio Meteorológico Nacional. Recuperado de <http://clicom-mex.cicese.mx/malla/index.php>

Sun, L., Seidou, O., Nistor, I., & Liu, K. (2016). Review of the Kalman-type hydrological data assimilation. *Hydrological Sciences Journal*,

61(13), 2348-2366. Recuperado de
<https://doi.org/10.1080/02626667.2015.1127376>

Valdés, J., Mejía, J., & Rodríguez-Iturbe, I. (1980). *Filtros de Kalman en la hidrología: predicción de descargas fluviales para la operación óptima de embalses* (Informe Técnico No. 80-2), Universidad Simón Bolívar, Caracas, Venezuela.

Welch, G., & Bishop, G. (2006). An introduction to the Kalman filter. In: *University of North Carolina at Chapel Hill*, (1), 2-15.

Zou, L., Zhan, C., Xia, J., Wang, T., & Gippel, C. J. (2017). Implementation of evapotranspiration data assimilation with catchment scale distributed hydrological model via an ensemble Kalman Filter. *Journal of Hydrology*, 549, 685-702. Recuperado de <https://doi.org/10.1016/j.jhydrol.2017.04.036>