2021, Instituto Mexicano de Tecnología del Agua
Open Access bajo la licencia CC BY-NC-SA 4.0
(https://creativecommons.org/licenses/by-nc-sa/4.0/)

DOI: 10.24850/j-tyca-2021-06-06

Articles

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

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

Ildefonso Narváez-Ortiz[1], ORCID: https://orcid.org/0000-0002-4988-8886

Laura Alicia Ibáñez-Castillo[2], ORCID: https://orcid.org/0000-0001-9287-655X

Ramón Arteaga-Ramírez[3], ORCID: https://orcid.org/0000-0001-9459-3588

Mario Vázquez-Peña[4], ORCID: https://orcid.org/0000-0003-2084-7420

Carlos Cíntora-González[5], ORCID: https://orcid.org/0000-0001-5204-4361

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

2021, Instituto Mexicano de Tecnología del Agua
Open Access bajo la licencia CC BY-NC-SA 4.0 (https://creativecommons.org/licenses/by-nc-sa/4.0/)

[2]Professor-researcher, Universidad Autónoma Chapingo, posgrado en Ingeniería Agrícola y Uso Integral del Agua, Chapingo, México, México, libacas@gmail.com

[3]Professor-researcher, Universidad Autónoma Chapingo, posgrado en Ingeniería Agrícola y Uso Integral del Agua, Chapingo, México, México, arteagar@correo.chapingo.mx

[4]Professor-researcher, Universidad Autónoma Chapingo, posgrado en Ingeniería Agrícola y Uso Integral del Agua, Chapingo, México, México, mvazquezp@correo.chapingo.mx

[5]Professor-researcher, Universidad Autónoma Chapingo, posgrado en Ingeniería Agrícola y Uso Integral del Agua, Chapingo, México, México, carlos.cintora@gmail.com

Corresponding author: Laura Alicia Ibáñez Castillo, email: libacas@gmail.com

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.

2021, Instituto Mexicano de Tecnología del Agua
Open Access bajo la licencia CC BY-NC-SA 4.0
(https://creativecommons.org/licenses/by-nc-sa/4.0/)

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.

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

2021, Instituto Mexicano de Tecnología del Agua
Open Access bajo la licencia CC BY-NC-SA 4.0
(https://creativecommons.org/licenses/by-nc-sa/4.0/)

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.

Received: 19/03/2019

Accepted: 23/11/2020

2021, Instituto Mexicano de Tecnología del Agua
Open Access bajo la licencia CC BY-NC-SA 4.0
(https://creativecommons.org/licenses/by-nc-sa/4.0/)

Introduction

Data assimilation for estimating states is essential to applications of hydrologic forecasting (Liu *et al*., 2012) since, by cyclic updating, errors accumulated in the models can be reduced (Clark *et al*., 2008; Maxwell, Jackson, & McGregor, 2018). The ensemble Kalman filter (ENKF) (Evensen, 1994) is one of the algorithms widely used as a recursive method in hydrology (Liu & Gupta, 2007; Maxwell *et al*., 2018; Sun, Seidou, Nistor, & Liu, 2016). It is an extension of the discrete Kalman filter (Kalman, 1960) for treating non-linear dynamic systems (Evensen, 1994; Evensen, 2003) and has been used, among others, to forecast streamflows (Maxwell *et al*., 2018), evapotranspiration (Zou, Zhan, Xia, Wang, & Gippel, 2017) and soil moisture (Brandhorst, Erdal, & Neuweiler, 2017; Meng, Xie, & Liang, 2017). Moreover, it has been integrated with distributed hydrological models such as TopNet, Hydrotel, and MGB-IPH (Abaza, Anctil, Fortin, & Turcotte, 2015; Clark *et al*., 2008; Quiroz, Collischonn, & Paiva, 2019).

In hydrographic basin systems, the relationship between flows measured at different positions and a point of interest can be identified, what is called a response function (Valdés, Mejía, & Rodríguez-Iturbe,

2021, Instituto Mexicano de Tecnología del Agua
Open Access bajo la licencia CC BY-NC-SA 4.0
(https://creativecommons.org/licenses/by-nc-sa/4.0/)

1980). Based on this relationship, it is possible to make short-term forecasts that are updated constantly and indefinitely while the phenomenon under study persists.

In this study, we evaluated the forecasting capacity of the discrete Kalman filter (DKF) (Kalman, 1960), the ensemble Kalman filter (EnKF) (Evensen, 1994), and the integration of DKF with EnKF for forecasting one to six steps forward of the Huaynamota River flow at the Chapalagana station. The DKF is implemented to estimate the state vector (Instantaneous Unit Hydrograph – IUH) and the EnKF makes scale estimation of flow. The integration is carried out using the EnKF equation of states.

Materials and methods

The study area is delimited upriver from the Chapalagana station on the Huaynamota River (Chapalagana River or Atengo River), located in northwestern Mexico and shared by the states of Durango, Nayarit, Zacatecas, and Jalisco (INEGI, 2010) (Figure 1). Geographically, it is between -104° 33’ 34.16” and -103° 27’ 29.84” W and between 23° 28’ 50.05” and 21°23’57.62” N. It has an area of 12 075.7 km^2. The altitude

2021, Instituto Mexicano de Tecnología del Agua
Open Access bajo la licencia CC BY-NC-SA 4.0
(https://creativecommons.org/licenses/by-nc-sa/4.0/)

at the highest point is 3147 m and at the lowest point, at the Chapalagana station, is 219 m. The concentration type is 39.88 hours. The mean annual precipitation is 707 mm and the mean annual temperature is 17.9 °C (SMN, 2019).

2021, Instituto Mexicano de Tecnología del Agua
Open Access bajo la licencia CC BY-NC-SA 4.0
(https://creativecommons.org/licenses/by-nc-sa/4.0/)

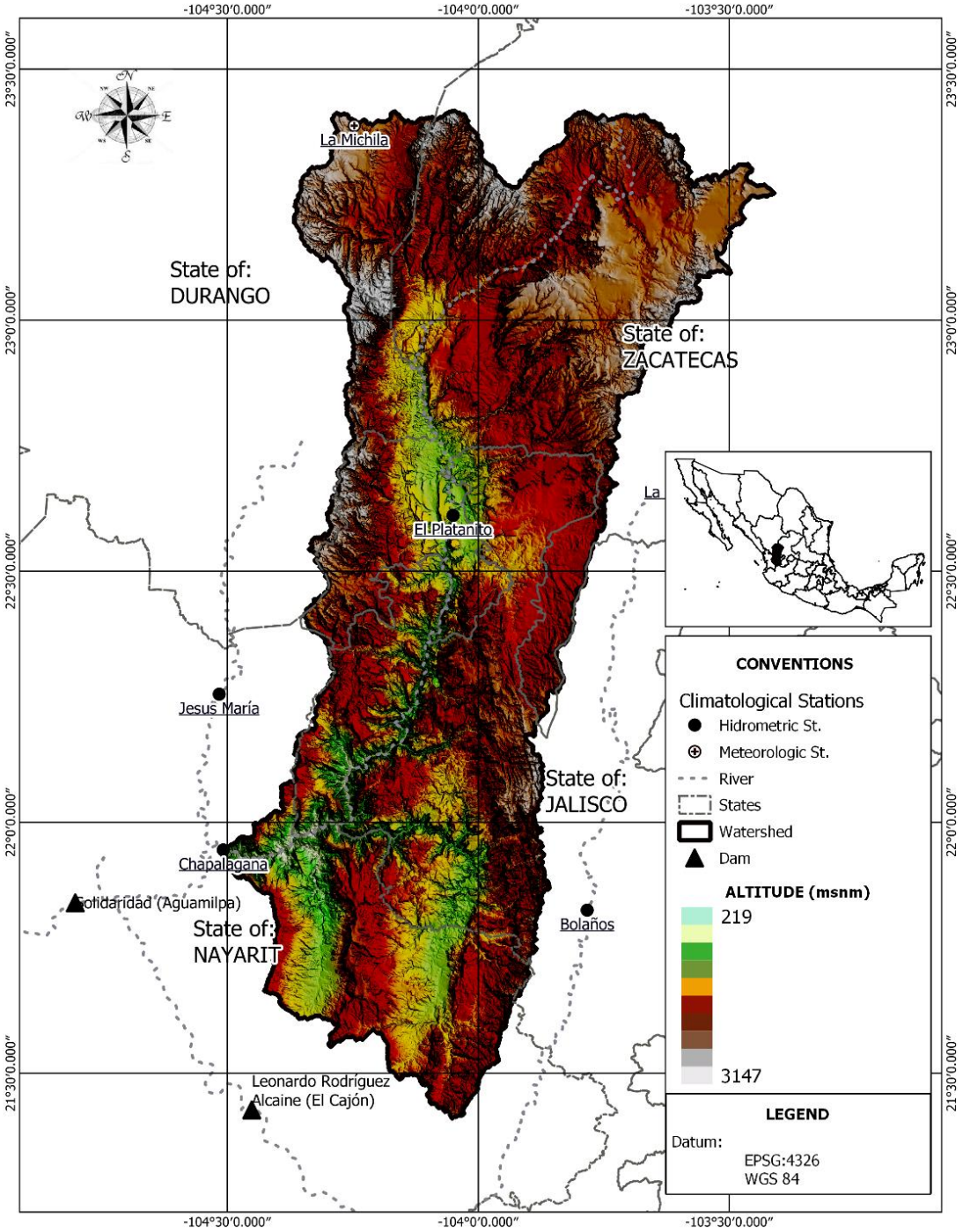

Figure 1. Location of the study area.

2021, Instituto Mexicano de Tecnología del Agua
Open Access bajo la licencia CC BY-NC-SA 4.0
(https://creativecommons.org/licenses/by-nc-sa/4.0/)

The main hydraulic structure for generating electricity found downriver from the study area and on the River Lerma-Santiago Pacific is the Solidaridad dam, commonly known as Aguamilpa, located at -104°48'10.55" W and 21°50'22.74" N. It has a capacity of 5 540 million m^3 and generates 960 MW of electricity (Conagua, 2008). The Aguamilpa dam is approximately 90 km from the Pacific coast of Nayarit where the Santiago River empties.

We evaluated the predictive capacity of the Discrete Kalman Filter (DKF) (Kalman, 1960), the Ensemble Kalman Filter (EnKF) (Evensen, 1994), the integration of DKF and EnKF (DKF-EnKF), and the autoregressive model with a first-order exogenous variable (AR(1,1)) to forecast at 1, 2, 3, 4, 5, and 6 hours (L steps) forward of hourly flows at the Chapalagana station, based on the flows at the Platanito station, located upriver, as the exogenous variable. We used hourly flow series from 2017 between August 2 at 9:00 hours and September 28 at 0:00 hours, for a total of 1360 registers provided by the Comisión Federal de Electricidad (CFE). For analysis of set size sensitivity, we used 5, 8, 10, 20, 50, and 100 members, which were combined with the six steps.

DKF, EnKF, DKF-EnKF, and ARX (1,1) were implemented by means of routines in R (R Core Team, 2020), which generate the forecasts in six steps with DKF and ARX, and with 36 combinations between steps by set size in EnKF and DKF-EnKF. DKF was implemented to estimate the state vector that is equivalent to the response function of the basin or Instantaneous Unit Hydrograph (IUH) (Valdés *et al*., 1980), and EnKF

2021, Instituto Mexicano de Tecnología del Agua
Open Access bajo la licencia CC BY-NC-SA 4.0
(https://creativecommons.org/licenses/by-nc-sa/4.0/)

makes scale estimations of the flow (Evensen, 2009). That is, DKF estimates the values corresponding to the IUH columns that are multiplied by the flow values, while EnKF estimates flow values directly. In the three cases with filters, the final observations of each variable are considered, that is, an autoregressive delay (Valdés *et al*., 1980). The ARX model was implemented recursively based on a fraction of a 100-register series.

The suitable number of members in the EnKF and DKF-EnKF sets was determined with a sensitivity analysis, based on the mean square root error (MSRE) (Quiroz *et al*., 2019). White noise (Monte Carlo simulation) was generated with mvtnorm software (multivariant and normal t distribution) (Genz & Bretz, 2009). In the three algorithms evaluated the Q variance was assumed to be constant (Simon, 2001), that is, zero (Morales-Velázquez, Aparicio, & Valdes, 2014), and R to be near-zero (0.01) to confer flexibility to the convergence of the algorithm (Welch & Bishop, 2006).

The fit of the forecasted series was evaluated with the Nash-Sutcliffe coefficient (Nash & Sutcliffe, 1970) and the mean square root error (MSRE) (Morales-Velázquez *et al*., 2014). The assumptions of error normality of the Kalman Filter were verified using graphs that were compared with the normal standardized curve (González-Leiva, Ibáñez-Castillo, Valdés, Vázquez-Peña, & Ruiz-García, 2015). Atypical values and their location were determined based on standardized residuals (Cryer & Chan, 2008).

2021, Instituto Mexicano de Tecnología del Agua
Open Access bajo la licencia CC BY-NC-SA 4.0
(https://creativecommons.org/licenses/by-nc-sa/4.0/)

Discrete Kalman Filter

The Discrete Kalman Filter is an algorithm that allows the identification of linear dynamic systems as an optimal estimator of states using a recursive process (Kalman, 1960) (Figure 2). DKF serves as a basis for algorithms that deal with non-linear systems (Welch & Bishop, 2006).

2021, Instituto Mexicano de Tecnología del Agua
Open Access bajo la licencia CC BY-NC-SA 4.0
(https://creativecommons.org/licenses/by-nc-sa/4.0/)

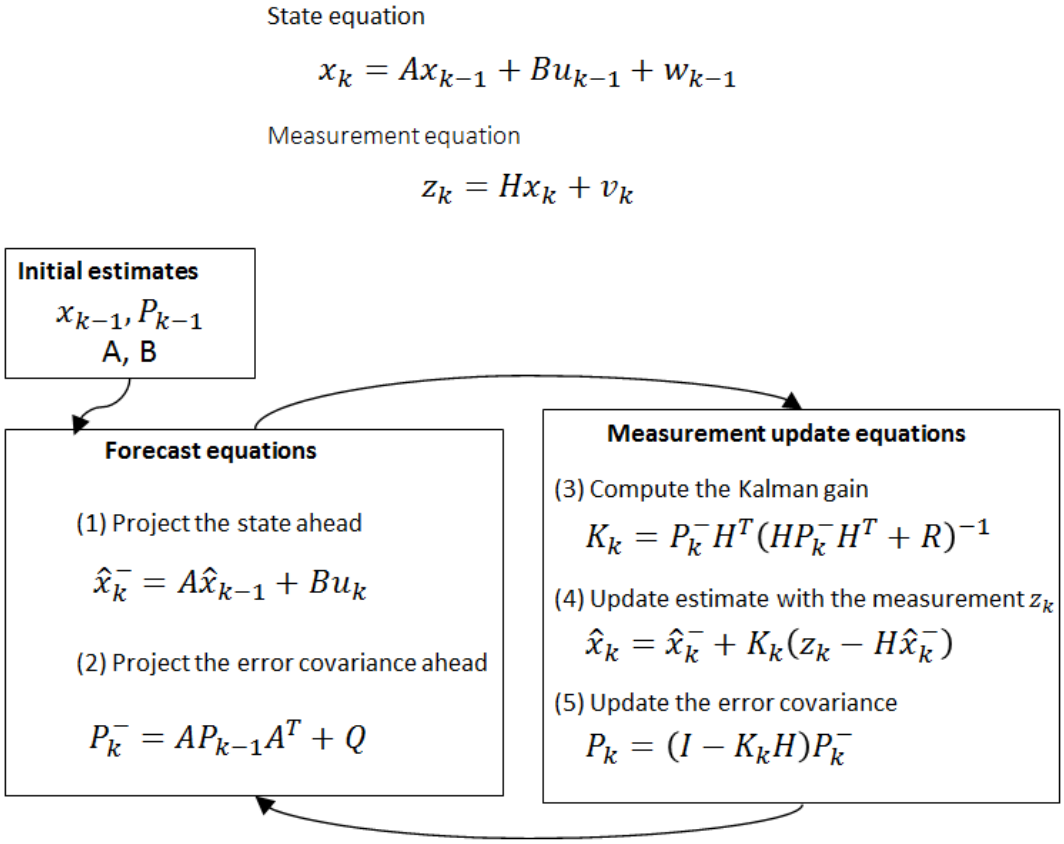

Figure 2. Discrete Kalman filter algorithm (Welch & Bishop, 2006).

The state equation has as entry data two-hourly flow series (n) from the Chapalagana and Platanitos stations. Matrix A ($n\ x\ n$) and B ($n\ x\ 1$) relate the state of time $k-1$ with the current state of time k. Because in the river course there are no hydraulic structures such as dams that would suddenly alter flow volumes, in this study a control variable was not considered, and only matrix A, which was assumed constant throughout

2021, Instituto Mexicano de Tecnología del Agua
Open Access bajo la licencia CC BY-NC-SA 4.0
(https://creativecommons.org/licenses/by-nc-sa/4.0/)

the process, was used. Matrix H is the measurement equation is composed of a vector row of $1\ x\ n$ that contains the last observation of each entry variable. The predicted value z_k is obtained with the measurement equation, multiplying matrix H by the state vector x_k^- $(n\ x\ 1)$. Following, matrix A, which relates the previous state to the current state, is presented.

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

The new measurement for time k (Q_k) is used to update the algorithm according to the changes that are occurring in the system, and the predicted value z_k (corresponding to a flow value of size p) implicitly carries the measurement error w_{k-1}, and the state equation contains the process error v_k, which must satisfy the assumption of normality:

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

A cycle that repeats indefinitely begins with the state and measurement equations of the algorithm. At time $k-1$ it makes an *a priori* estimate (forecast) of the states that are updated (*a posteriori* estimation) in time k. The states are taken as the basin's response function, and with the *a posteriori* estimation, the forecast in time $k+1$ is made. This cycle

2021, Instituto Mexicano de Tecnología del Agua
Open Access bajo la licencia CC BY-NC-SA 4.0
(https://creativecommons.org/licenses/by-nc-sa/4.0/)

is repeated indefinitely, forecasting time $k+1$ based on matrix H and the updated state vector up to time k.

Ensemble Kalman Filter

The Ensemble Kalman Filter (EnKF) is a suboptimal estimator based on Monte Carlo simulations for estimating statistical error (Evensen, 1994; Gillijns *et al*., 2006; Rafieeinasab, Seo, Lee, & Kim, 2014). Normal distribution of errors is assumed, and the estimations are made based on ensembles that group q values generated at random under normal distribution. Like DKF, it has two groups of equations: analysis and prediction. In Figure 3, the first two equations have their equivalents in DKF, and the latter corresponds to the average of the members, which is assumed to be a better forecast. The second group, in general, comprises four equations where errors and probabilities, which are inputs to calculate the Kalman gain, are estimated (Gillijns *et al*., 2006.

2021, Instituto Mexicano de Tecnología del Agua
Open Access bajo la licencia CC BY-NC-SA 4.0
(https://creativecommons.org/licenses/by-nc-sa/4.0/)

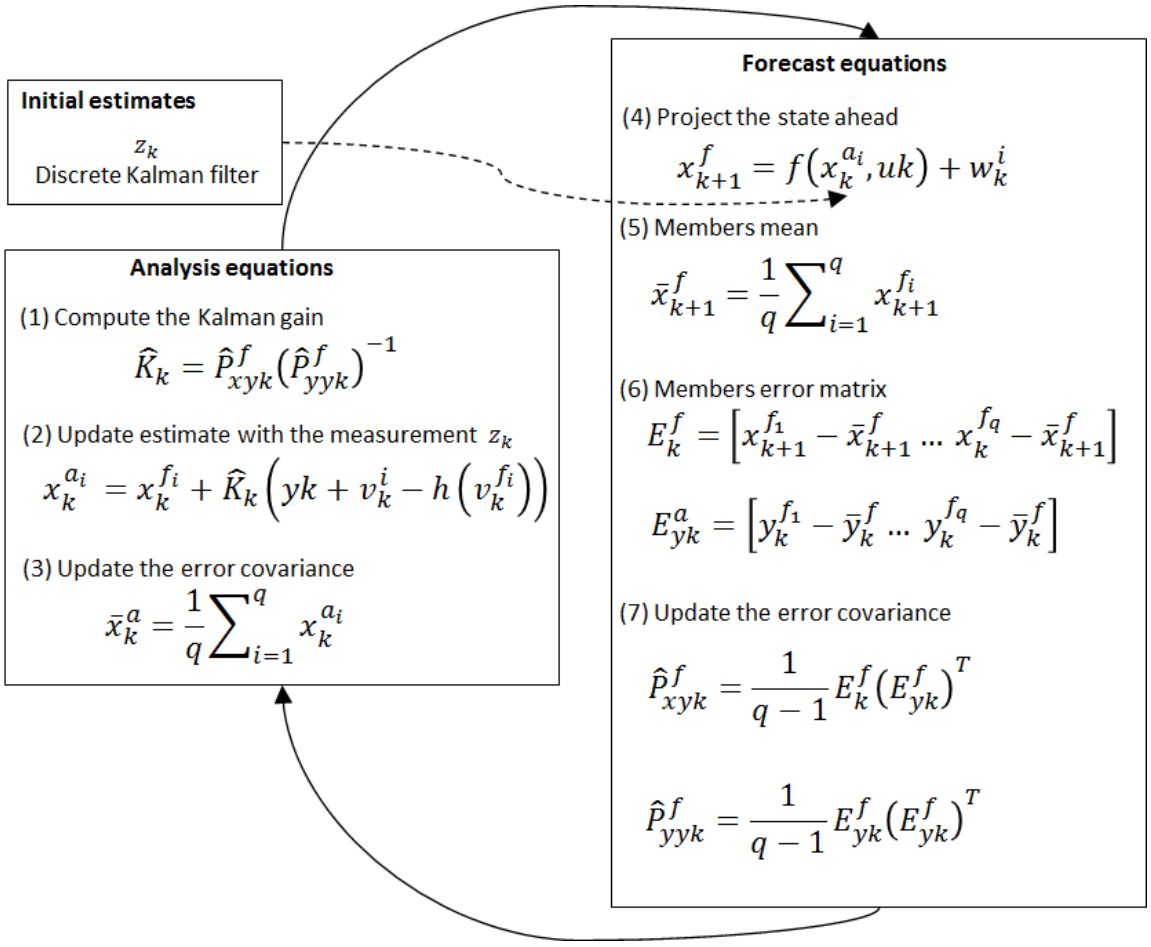

Figure 3. Algorithm of the ensemble Kalman filter (Gillijns *et al*., 2006).

The errors w_k^i and v_k^i correspond to the noise the process and the measurement contains, respectively, and are assumed as white noise with mean zero and variance Q and R (Figure 3), as in equation 2. For the flow forecast, the response function (x_{k+1}^f) of EnKF was established incorporating white noise w_k^i into the previous states x_{k-1}^{ai}:

2021, Instituto Mexicano de Tecnología del Agua
Open Access bajo la licencia CC BY-NC-SA 4.0
(https://creativecommons.org/licenses/by-nc-sa/4.0/)

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

Noise in the measurements is generated by adding the measurement in time k, q deviations with normal distribution.

$$y_{k}^{i} = y_{k} + v_{k}^{i} \quad (4)$$

Vector y_{k}^{i} of $1\ x\ q$ corresponds to q noisy measurements, y_{k} is the measurement in time k, and the subindex i represents the number of members corresponding to $i = 1,2,\ldots,\ldots,q$. That is, q values are generated at random under a normal distribution and are added to the observed value, resulting in a q-sized vector with the observed value as the mean. It is assumed that a larger number of members will be the fit of the forecast because a better estimation of the distribution of probabilities can be obtained (Leutbecher, 2019). However, increasing the number of members implies greater computational effort. Therefore, an analysis of sensitivity should be done to determine the number that obtains a minimum of errors with acceptable computational effort. For hydrological studies, between 50 and 300 members per set are suggested (Gillijns *et al*., 2006; Quiroz *et al*., 2019).

2021, Instituto Mexicano de Tecnología del Agua
Open Access bajo la licencia CC BY-NC-SA 4.0
(https://creativecommons.org/licenses/by-nc-sa/4.0/)

Discrete Kalman filter and ensemble Kalman filter

DKF and EnKF were integrated using the state equation of the EnKF algorithm. DKF generates a forecast for time $k+1$ ($\hat{Q}_{k+1}$), with which the difference relative to the previous state of EnKF (x_{k-1}^{ai}) is determined and which is finally added to the previous state x_{k-1}^{ai}. Moreover, white noise is added using Monte Carlo simulation w_k^i.

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

First-order autoregressive model (ARX(1,1)) and DKF

The first-order autoregressive model, also known as the Markov model, is one of the first approximations for the study of time series. It is based on the autocorrelation that appears in the same data series (Box, Jenkins,

2021, Instituto Mexicano de Tecnología del Agua
Open Access bajo la licencia CC BY-NC-SA 4.0
(https://creativecommons.org/licenses/by-nc-sa/4.0/)

Reinsel, & Ljung, 2016; Bras & Rodríguez-Iturbe, 1985). Algebraically, it is described in the following form:

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

where y_{k+1} is the predicted value, α_i and β_j are the model parameters, and y_k and γ_k are the entry variable and the exogenous variable, respectively. The parameters are estimated by least squares and require a section of series of at least 50 registers. In the model, ARX(na, nb), na and nb represent the autoregressive delays that are used in each variable. The ARX model estimates parameters α and β, which are incorporated into matrix A of the DKF algorithm (Figure 2), and updating is applied to the flow for the forecast in the following cycle.

Results and discussion

We evaluated a total of 36 combinations resulting from six steps (L) by six set sizes. The three algorithms were executed with 1360 hourly flow

2021, Instituto Mexicano de Tecnología del Agua
Open Access bajo la licencia CC BY-NC-SA 4.0
(https://creativecommons.org/licenses/by-nc-sa/4.0/)

registers from the Chapalagana and Platanitos stations to obtain flow forecasts at the Chapalagana station.

Figure 4 shows the sensitivity analysis that allows identification of the suitable set size for the EnKF (dotted line) and DKF-EnKF (solid line) algorithms through the relationship between the MSRE value as the number of set members increases. The set with five members has the highest MSRE in each of the L steps. When there are 8 and 10 members, the variations have a slight tendency to diminish, and the sets with 20, 50, and 100 members reflect convergence and stability. Therefore, in the case of the series under study, to optimize the fit and the computational load, 20 members can be used. The combination DKF-EnKF in the six L steps has higher MSRE values than EnKF, and the differences are increasingly notable in larger steps. However, convergence is reached as of 20 members per set.

2021, Instituto Mexicano de Tecnología del Agua
Open Access bajo la licencia CC BY-NC-SA 4.0
(https://creativecommons.org/licenses/by-nc-sa/4.0/)

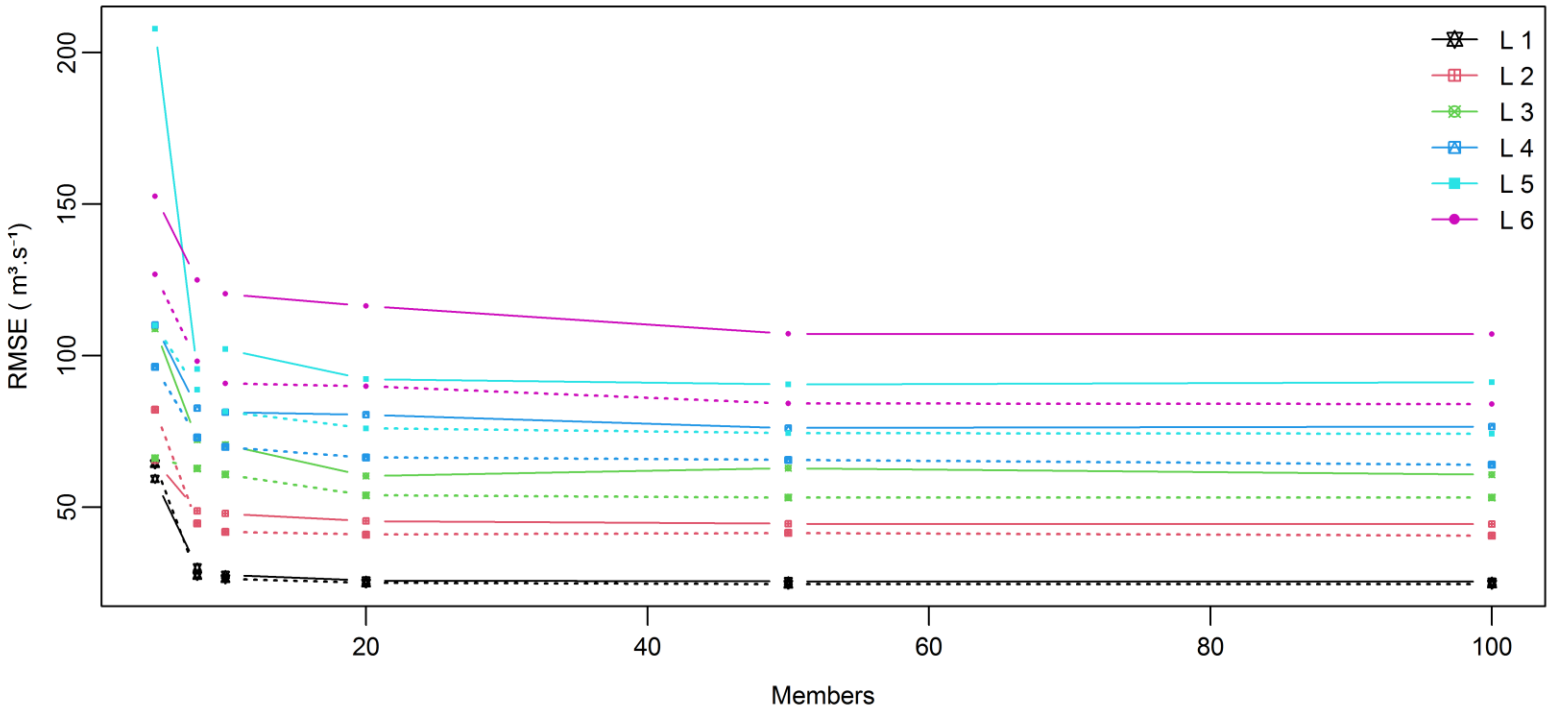

Figure 4. Root mean square error with different ensemble sizes.

The importance of defining the suitable number of members lies in obtaining the best possible fit without adding a computational load that does not improve the behavior of the algorithm. When sets are large, computational load is high, but if small sets are used, fit between the observed and predicted series can be lost (Gillijns *et al*., 2006; Quiroz *et al*., 2019).

The results we present are based on 20 members per set. Table 1 presents the statistical indicators of fit of the observed series against the predicted series.

2021, Instituto Mexicano de Tecnología del Agua
Open Access bajo la licencia CC BY-NC-SA 4.0
(https://creativecommons.org/licenses/by-nc-sa/4.0/)

Table 1. Summary of statistics for application of para la DKF, EnKF, DKF-EnKF, and ARX-DKF.

	Index	L1	L2	L3	L4	L5	L6	SDs
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
	Mean	161.55	161.71	161.39	161.57	160.74	160.67	0.45
	SDp	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
	Mean	160.65	160.55	160.88	160.15	160.93	161.27	0.38
	SDp	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
	Mean	160.28	159.51	159.78	160.34	161.64	162.09	1.03
	SDp	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
	Mean	163.32	165.27	166.79	168.01	169.71	170.52	2.72
	SDp	219.90	222.08	224.70	226.56	228.20	228.34	3.42

MSRE: mean square root error

Nash-Sutcliffe: Nash index

Mean: Average

SDp: standard deviation in each predicted series

SDs: the standard deviation between the six steps for each algorithm

The average and standard deviation of the observed series are 160.95 and 216.67, respectively.

2021, Instituto Mexicano de Tecnología del Agua
Open Access bajo la licencia CC BY-NC-SA 4.0
(https://creativecommons.org/licenses/by-nc-sa/4.0/)

Implementation of the ARX(1,1) model and the ARX-DKF model with recursive updating obtain similar results because the parameters of ARX are multiplied by flow values in time k, an operation that is executed in a matrix manner in the forecast equation of the DKF algorithm. The results presented refer only to the ARX-DKF model.

For step $L = 1$, the algorithms with the Kalman filter generate forecasts with Nash-Sutcliffe coefficients (NS) above 0.9855, which is slightly better than that found in the flow forecasts at the Ángel Albino Corzo dam (NS = 0.9774) (Morales-Velázquez *et al*., 2014). However, in the measure that step L becomes larger, increasing differences appear between the algorithms. DKF remains with the best fit, followed by EnKF and ARX-DKF, and lastly by the integration DKF-EnKF, with a pronounced decrease to 0.7110 in step $L = 6$ (Table 1). MSRE can be interpreted with units in m^3/s, and has adjustment behavior similar to that evidenced by NS. Both NS and RMSE demonstrate that the smallest error is achieved by DKF, followed by EnKF, ARX-DKF, and DKF-EnKF. In terms of NS and RMSE, it can be seen that the DKF and EnKF algorithms generate forecasts with a better fit than the ARX-DKF model. Moreover, the use of DKF does not require long sections of series for training, meaning more model simplicity.

According to the statistical summary in Table 2, the mean and standard deviation in the six steps tend to overestimate in different magnitudes, relative to the observed series. The most stable algorithm was that of EnKF, with a standard deviation of 0.38, slightly different from

2021, Instituto Mexicano de Tecnología del Agua
Open Access bajo la licencia CC BY-NC-SA 4.0
(https://creativecommons.org/licenses/by-nc-sa/4.0/)

that of DKF, which has a deviation of 0.45 between the six steps, while ARX-DKF and DKF-EnKF have higher standard deviations of 1. 03 and 2.72, respectively. The dispersion represented by the standard deviation among the forecasted series (Dep) confirms that EnKF is more similar to the observed series in the six steps. The apparent better fit of the EnKF forecasts is due to the effect of displacement relative to the observed data with little change in the maximum or minimum peaks, in such a way that there were no notable effects on the mean or standard deviation.

The differences among the generated forecasts become notable in graphic form, and for a detailed comparison, Figure 5 and Figure 6 present two floods at different steps.

2021, Instituto Mexicano de Tecnología del Agua
Open Access bajo la licencia CC BY-NC-SA 4.0
(https://creativecommons.org/licenses/by-nc-sa/4.0/)

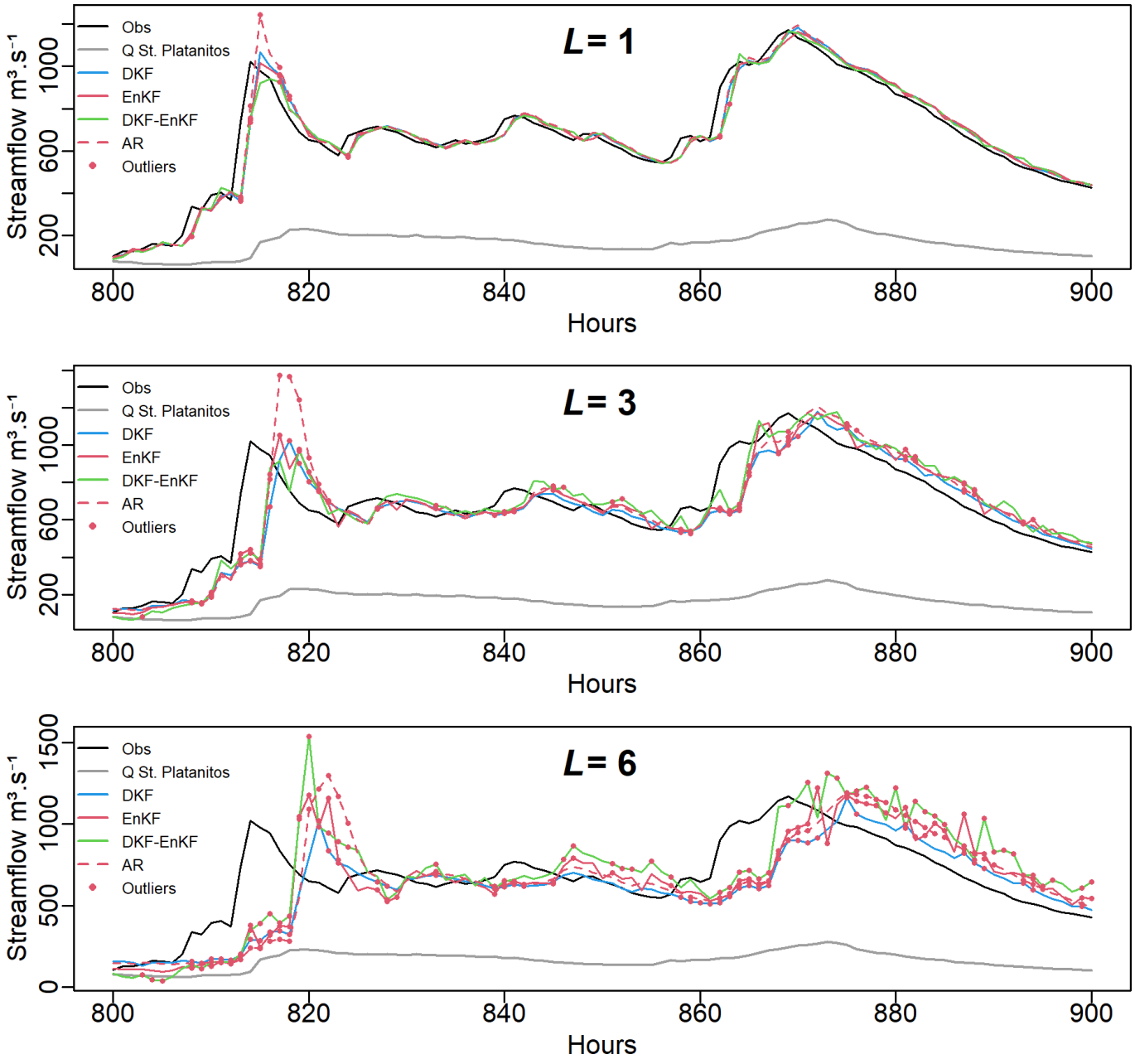

Figure 5. Observed flows and flows forecasted with DKF, EnKF, and DKF-EnKF (flood 4/9/2017 16:00 h to 8/9/2017 20:00 h).

2021, Instituto Mexicano de Tecnología del Agua
Open Access bajo la licencia CC BY-NC-SA 4.0
(https://creativecommons.org/licenses/by-nc-sa/4.0/)

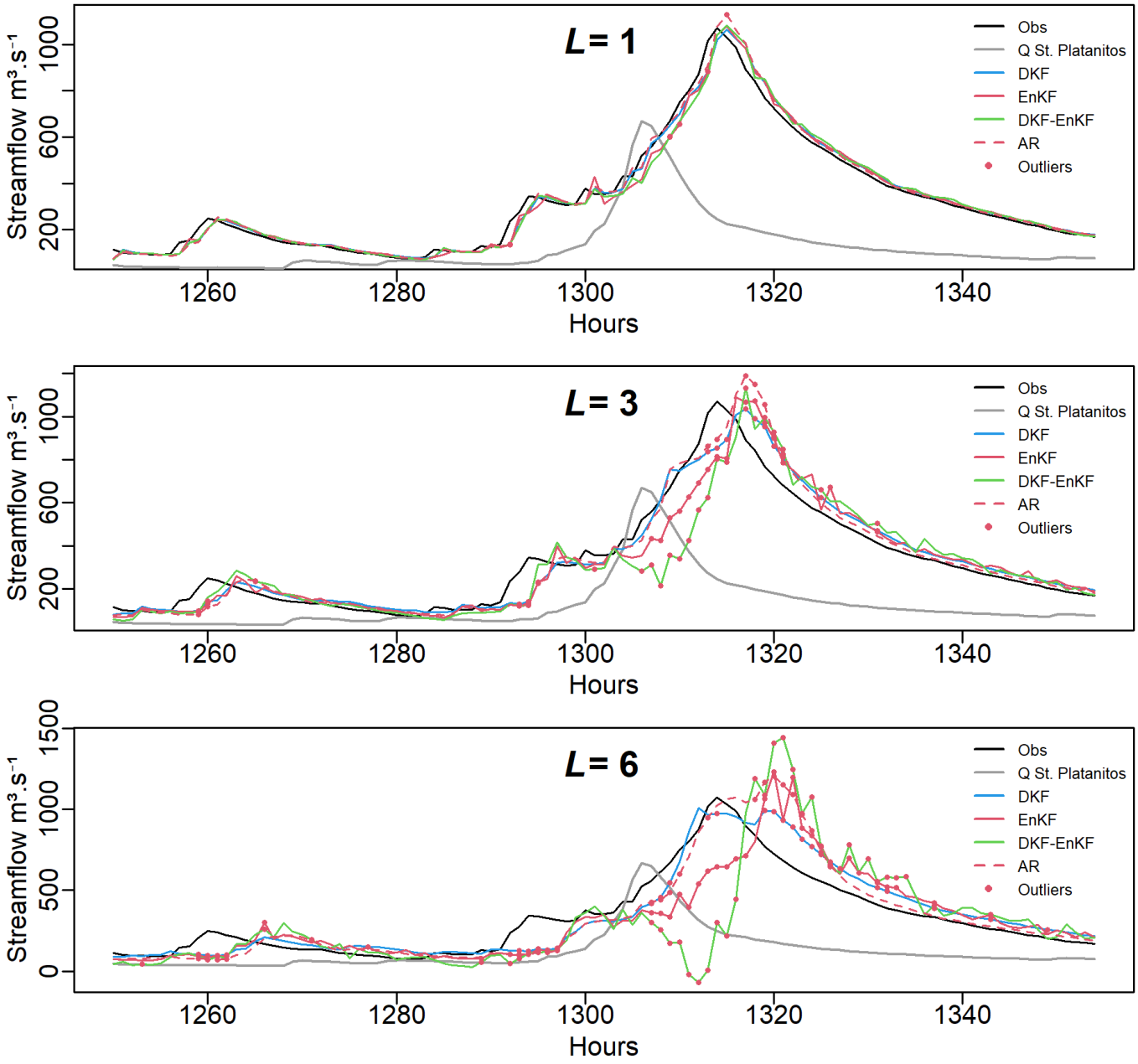

Figure 6. Observed flow and flow forecasted with DKF, EnKF, and DKF-EnKF (Flood 23/09/17 10:00 h to 28/09/17 00:00 h).

2021, Instituto Mexicano de Tecnología del Agua
Open Access bajo la licencia CC BY-NC-SA 4.0
(https://creativecommons.org/licenses/by-nc-sa/4.0/)

The forecasts for time $k+L$ displaced a value similar to L because updating incorporates the last measurement corresponding to time k. In other words, with larger distances between the measured data set and the forecasted $k+L$ moment, there is a progressive decrease in the fit between step $L=1$ and $L=6$, manifested as displacement of the predicted over the observed series. Nevertheless, for the event shown in Figure 6, with the forecasts generated by DKF and ARX, displacement is almost null. This is because the delay between the peak on September 25, 2017, at 18:00 at the Platanitos station, and the peak on September 26, 2017, at 02:00 at the Chapalagana station is eight hours. That is, DKF estimates and updates the Instantaneous Unit Hydrograph (IUH) based on flow registers from the Platanitos station at the current time (k), which at the Chapalagana station appeared 8 hours later as a consequence of concentration time. The proximity of the delay magnitude and step $L=6$ allows that updating can be done with correlated registers, decreasing displacement and conferring better fit at the maximum peak. ARX generates the forecast with a fit similar to that of DKF but with a longer series segment; that is, implementation of the DKF model is simpler.

The EnKF and DKF-EnKF algorithms did not obtain improvements because EnKF makes scale estimations, and delay between series is not included. Unlike DKF, the forecast generated by DKF-EnKF is inverted in such a way as to generate a minimum peak (Hour 1311, 25/09/2017 23:00) due to updating with the maximum flow peak presented at the Platanitos station (Hour 1306, 25/09/2017 18:00) that the EnKF

2021, Instituto Mexicano de Tecnología del Agua
Open Access bajo la licencia CC BY-NC-SA 4.0
(https://creativecommons.org/licenses/by-nc-sa/4.0/)

algorithm tries to correct. A possible adequation is the implementation of EnKF for estimating states instead of scales that can also add capacity for non-linear treatment, as well as integration of the Kalman Filter with distributed simulation models, as proposed by Rafieeinasab *et al*. (2014). Modeling in Kalman simplifies the parameters, which include the distributed model Sacramento Soil Moisturing Accounting (SAC), and finally compares its ensemble Kalman filters with the result of this model. The work of Rafieeinasab *et al*. (2014) is quite interesting from two standpoints: 1) The work compares two ensemble Kalman filters, the simple EnKF and that of maximum likelihood (MELF); 2) It considers an entire system that relates flow forecasting with soil moisture, which is seen as a balance that considers mean rainfall in the basin and a mean evaporation potential in the basin. Conceptually, our model is simpler because it only simulates data from measured flows. However, constructive criticism of the results is the following: their basin is smaller (435 km^2) than ours (12 076 km^2). The order of the size of their flows is small (100 to 200 m^3/s), but their error (MSRE) is very high, possibly reaching 45 m^3/s. In our study, the flows are in the order of 200 to 1 100 m^3/s. Nevertheless, the maximum MSRE is 116 m^3/s in the worst case.

Forecasts can achieve a better fit by dynamically incorporating time delay between series, treating variables with different types of observation and proportions (Meng *et al*., 2017), and incorporating variables that affect time delays, such as location and direction of precipitation events and prior soil moisture condition in the basin. Precipitation, atmospheric pressure, temperature, and relative humidity

2021, Instituto Mexicano de Tecnología del Agua
Open Access bajo la licencia CC BY-NC-SA 4.0
(https://creativecommons.org/licenses/by-nc-sa/4.0/)

are variables that can contribute to better fits in the forecast. Moreover, given that the flows measured in the upper part of the basin are proportionally smaller due to the smaller capture area, it is convenient to include parameters that help equate flow magnitudes in the model. Also, it is determinant to have registers from weather stations distributed throughout the area of the basin to define and update the response function of the basin (IUH) (González-Leiva *et al*., 2015).

The dispersion between the observed series and the predicted series (Figure 7) shows algorithm similarity. In the measure that L increases, the value of the slope (in the linear equation, Figure 7) decreases; that is, in general, the predicted values are underestimated, relative to the observed value.

2021, Instituto Mexicano de Tecnología del Agua
Open Access bajo la licencia CC BY-NC-SA 4.0
(https://creativecommons.org/licenses/by-nc-sa/4.0/)

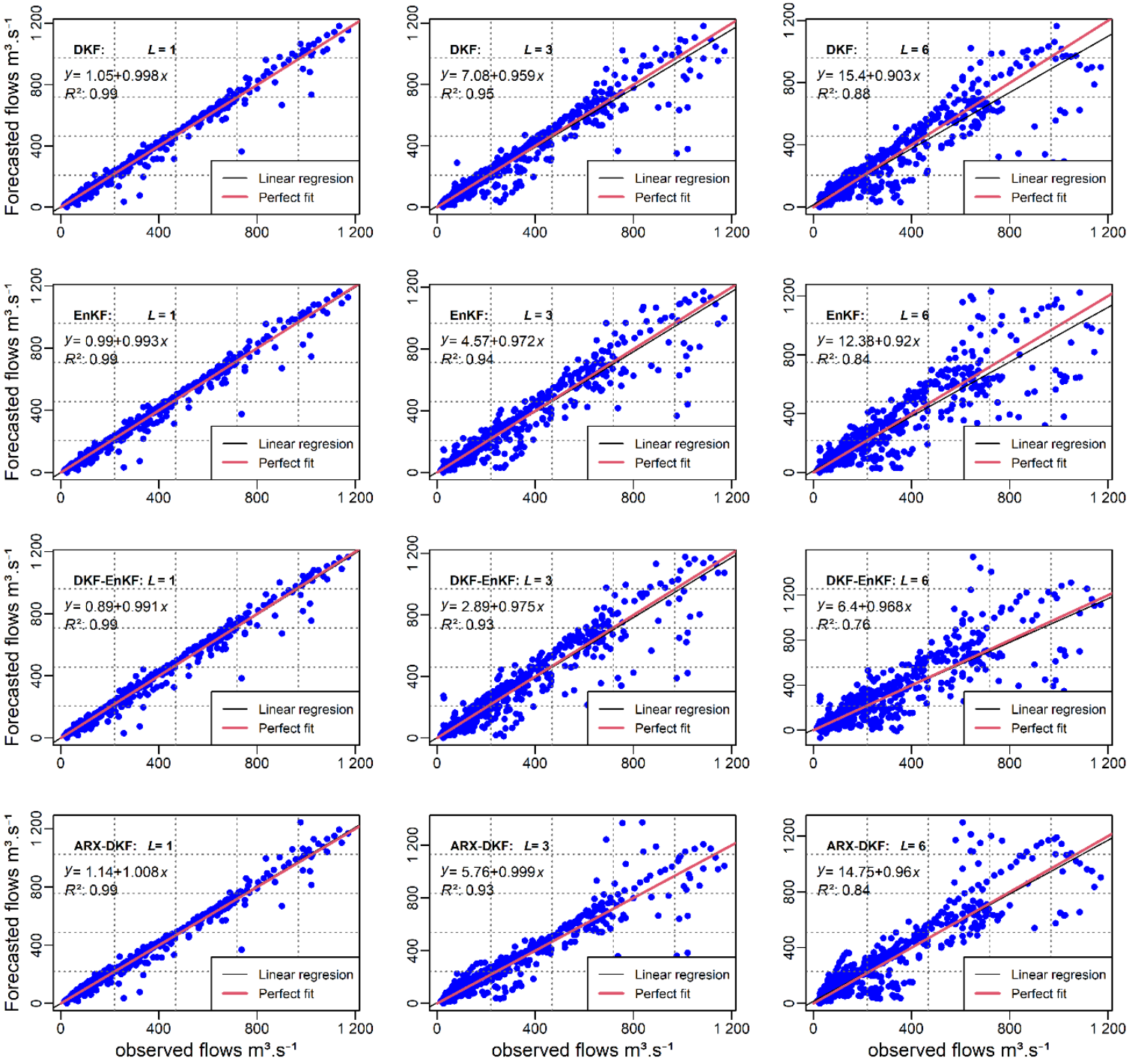

Figure 7. Observed vs. forecasted flows.

2021, Instituto Mexicano de Tecnología del Agua
Open Access bajo la licencia CC BY-NC-SA 4.0
(https://creativecommons.org/licenses/by-nc-sa/4.0/)

When $L = 1$, the linear fit between observed and predicted values has a coefficient of determination of 0.99, which is higher and similar in the four algorithms (Table 2). From steps $L = 2$ onward, dispersion increases due to the distancing of the data set used for the forecast. DKF maintains the best fit in the six steps, with notable similarity to EnKF up to $L = 3$. When $L = 6$, DKF has the coefficient of determination of 0.88, which is higher. The behavior of the linear fit is congruent with the NS value shown in Table 1. The points that are far away from the central line correspond to forecasts in which flow at the Chapalagana station undergoes brusque changes when conforming the flood. Said points generate forecasts with atypical magnitudes. At $k = 813$ and $k = 814$ hours, the highest deviations relative to the observed value occur, equivalent to the initial stage of flood conformation, which is shown in Figure 5.

Table 2. Coefficients of determination of the observed flow against the forecasted flow.

	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

2021, Instituto Mexicano de Tecnología del Agua
Open Access bajo la licencia CC BY-NC-SA 4.0
(https://creativecommons.org/licenses/by-nc-sa/4.0/)

To verify the normality of the generated residuals from the DKF, EnKF, and DKF-EnKF forecasts, Figure 8 shows the histogram and the fit of the normal curve for each algorithm in the different steps. It is important to consider that it is ideal that the residuals have a normal distribution, but this occurs in few cases. Therefore, an approximation is admitted (Chatfield, 2001).

2021, Instituto Mexicano de Tecnología del Agua
Open Access bajo la licencia CC BY-NC-SA 4.0
(https://creativecommons.org/licenses/by-nc-sa/4.0/)

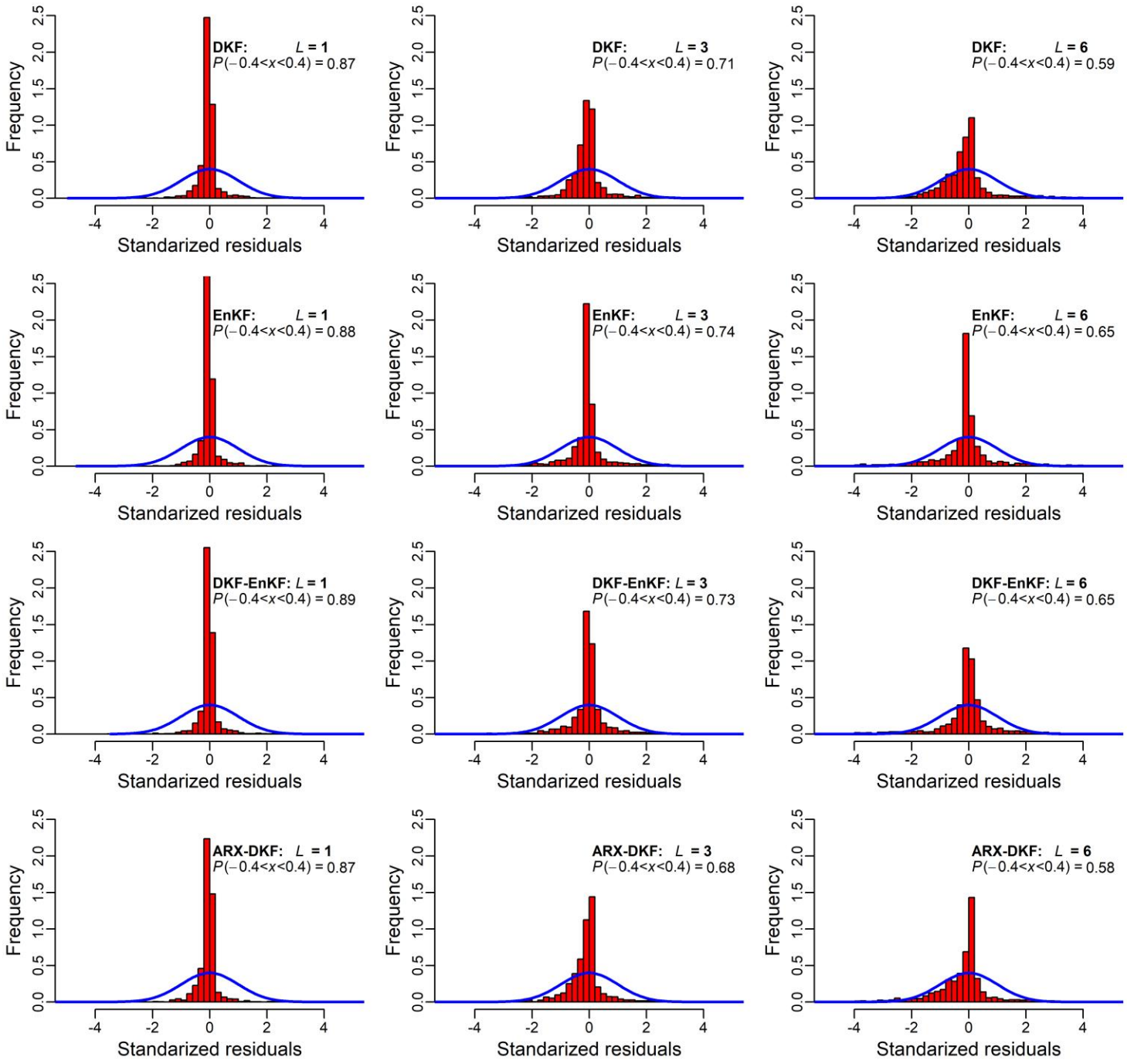

Figure 8. Histogram of DKF residuals.

2021, Instituto Mexicano de Tecnología del Agua
Open Access bajo la licencia CC BY-NC-SA 4.0
(https://creativecommons.org/licenses/by-nc-sa/4.0/)

The symmetric distribution with higher forecast concentration in the central part can be seen. Of the total estimations in the range of -0.4 to 0.4 of the standardized residuals, $L = 1$ groups 86.03% to 87.3%, while $L = 6$ groups between 56.67 and 63.02%. Assuming that there is a near-normal distribution in the residuals, standardization was done to determine the presence of atypical values, higher and lower than the standard deviations for 1 and 6 steps, resulting from the DKF, EnKF, and DKF-EnKF forecasts (Figure 9). The atypical values are located in the ascending and descending parts of the peaks and are shown as red dots in Figures 5 and 6. In accord with the statistical indexes of Table 1, in the measure that step L becomes larger, atypical values appear more frequently. Step $L = 1$ presents a similarity that becomes increasingly distorted until it reaches a range of 55 between DKF and DKF-EnKF in step $L = 6$.

2021, Instituto Mexicano de Tecnología del Agua
Open Access bajo la licencia CC BY-NC-SA 4.0
(https://creativecommons.org/licenses/by-nc-sa/4.0/)

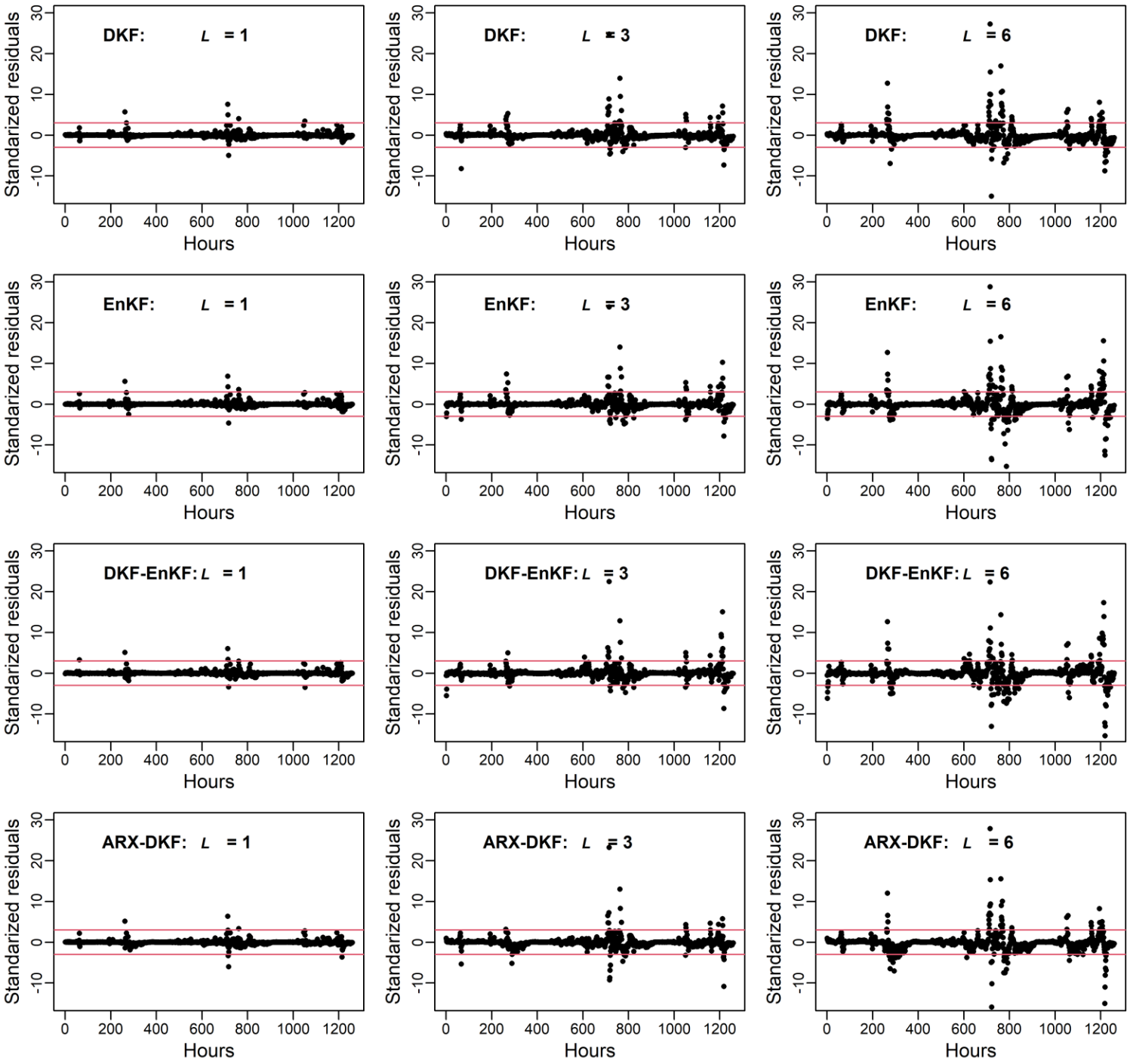

Figure 9. Standardized residuals.

2021, Instituto Mexicano de Tecnología del Agua
Open Access bajo la licencia CC BY-NC-SA 4.0
(https://creativecommons.org/licenses/by-nc-sa/4.0/)

The presence of atypical data is closely related to the level of fit. As shown in Table 3, DKF presents the lowest number of atypical values associated with the capacity for administrating the delays between the series (Table 1). In contrast, DKF-EnKF, except for step $L = 1$, has a greater presence of atypical values due to the application of gain by DKF and, in sequence, by the EnKF, which amply projects the brusque changes inflow. Also, the smallest peak, shown in Figure 7, favors the occurrence of more atypical values.

Table 3. The number of atypical values by algorithm and step.

	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 %

2021, Instituto Mexicano de Tecnología del Agua
Open Access bajo la licencia CC BY-NC-SA 4.0
(https://creativecommons.org/licenses/by-nc-sa/4.0/)

Conclusions

DKF achieves the best fit due to the estimation of states, which allows taking advantage of the delay between series as the basis for updating. EnKF estimates scale values and generates smoothing with a displacement effect that maintains the stability of the mean and the standard deviation but has a fit that is inferior to that of DKF. In this way, the use of DKF where time delay is incorporated recursively for updating can improve fit in the peaks. Moreover, the simplicity of programming and operating the DKF make it feasible for forecasting short-term flows. Forecasting hourly flows with all the Kalman filters analyzed is quite good with very high Nash-Sutcliffe coefficients and very low errors measured as MSRE.

To improve the fit of the forecasts, it is important to conduct studies that include recursive updating of time delay between the series of different stations, taking into account that variations occur throughout the time. Also, forecasts can be evaluated with step sizes of several hours, for example, by grouping six hours using values such as the mean or the maximum so that each step is equivalent to six hours and a forecast at

2021, Instituto Mexicano de Tecnología del Agua
Open Access bajo la licencia CC BY-NC-SA 4.0
(https://creativecommons.org/licenses/by-nc-sa/4.0/)

six steps is equivalent to 36 hours. In this way, the model is more flexible, and the forecasting period can be lengthened.

Acknowledgments

The authors thank the Comisión Federal de Electricidad for permitting access to their hydrometeorological database through the website administered by the Instituto Nacional de Electricidad y Energías Limpias during 2018.

References

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. Recovered from 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. Recovered from 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:

2021, Instituto Mexicano de Tecnología del Agua
Open Access bajo la licencia CC BY-NC-SA 4.0
(https://creativecommons.org/licenses/by-nc-sa/4.0/)

Armstrong, J. S. (ed.). *A handbook for researchers and practitioners.* Vol. 30 (pp. 475-494). Springer, U.S. Recovered from 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. Recovered from 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. Recovered from 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. Recovered from 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. Recovered from https://doi.org/10.1029/94jc00572

2021, Instituto Mexicano de Tecnología del Agua
Open Access bajo la licencia CC BY-NC-SA 4.0
(https://creativecommons.org/licenses/by-nc-sa/4.0/)

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*. Recovered from https://doi.org/10.1109/ACC.2006.1657419

González-Leiva, F., Ibáñez-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*. Recovered from 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. Recovered from 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

2021, Instituto Mexicano de Tecnología del Agua
Open Access bajo la licencia CC BY-NC-SA 4.0
(https://creativecommons.org/licenses/by-nc-sa/4.0/)

opportunities. *Hydrology and Earth System Sciences*, 16(10), 3863-3887. Recovered from 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. Recovered from 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. Recovered from 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. Recovered from 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. Recovered from 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. Recovered from

2021, Instituto Mexicano de Tecnología del Agua
Open Access bajo la licencia CC BY-NC-SA 4.0
(https://creativecommons.org/licenses/by-nc-sa/4.0/)

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. Recovered from http://www.R-project.org/. https://www.r-project.org/

Rafieeinasab, 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. Recovered from 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. Recovered from 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. Recovered from 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

2021, Instituto Mexicano de Tecnología del Agua
Open Access bajo la licencia CC BY-NC-SA 4.0
(https://creativecommons.org/licenses/by-nc-sa/4.0/)

Bolivar, Caracas, Venezuela.

Welch, G., & Bishop, G. (2006). An introduction to the Kalman filter. In: *University of North Carolina at Chapel Hil*, (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. Recovered from https://doi.org/10.1016/j.jhydrol.2017.04.036