3rd Specialty Conference on Disaster Prevention and Mitigation
3e Conférence spécialisée sur la prévention et la mitigation des désastres naturels

Montréal, Québec
May 29 to June 1, 2013 / *29 mai au 1 juin 2013*

Prediction of Daily Discharge at Bakel (Senegal) using Multiple Linear Regression, Kalman Filer and Artificial Neural Networks

L.Sun[1], I.Nistor[1],O.Seidou[1],S.Sambou[2],C.Kebe[3],S.Tamba[4]
[1] Department of Civil Engineering, University of Ottawa
[2] DakarFaculté des Sciences et Techniques, Université Cheikh Anta DIOP,Senegal
[3] Ecole Supérieure Polytechnique de DAKAR, Senegal
[4] Ecole Polytechnique de THIES,Senegal

Abstract: Daily discharge prediction is critical for dam operation and flood prevention. In operational hydrology, simple and robust techniques are sought to provide one to seven days forecasts of the discharge at key stations. In this paper, Multiple Linear Regression (including regular MLR and Stepwise regression), Back Propagation Artificial Neural Network (BP ANN) and linear Kalman Filter (KF) have been used to predict the daily discharge of the Senegal River at Bakel (Senegal, West Africa) at one to seven days lead. Inputs are the discharges at three upstream stations (Oualia, Gourbassi and Manantali). The Root-Mean-Square Error (RMSE) and Nash–Sutcliffe efficiency coefficient (NS) were used as assessment criteria. For BP ANN, the 1988-2002 measured daily discharge was used for learning process while 2003-2006 was used for validation. The whole time series from 1988 to 2006 was used in KF as no learning process is used for this method. The regression coefficients of MLR were updated based on the whole series also. All three methods provided satisfactory performances with Nash–Sutcliffe efficiency coefficients greater than 0.8 for a lead time of up to 7 days. Both KF and MLR (mainly Stepwise) outperformed BP ANN. The best prediction delay for all methods is two to three days. Kalman filter seems more stable when the prediction delay increases.

1 Background

Short term discharge prediction is the basis of efficient reservoir management, which involves plenty of key issues like irrigation, flood prevention, power generation as well as navigation. If we don't consider anthropogenic influence, the discharge at one point of the river mainly consist two parts: water from the upstream stretches (including tributaries) and runoff concentration from the watershed. Most traditional empirical methods tend to determine the relationship between upstream and downstream discharge directly while rainfall runoff hydrological models more likely to consider the contribution of whole watershed.

Compared to process based hydrological models, mathematical or statistical methods try to build a transfer function between the key variable and somehow arbitrarily slected inputs, without any consideration for the physical processes that links these variables. Among them, regression methods and artificial neural networks methods gained enough attention in past few years. The Kalman filter is another alternative that continuously updates the parameters of a model as new data becomes available. In this paper, we compared the application of these three methods in a short term (1 to 7 days ahead) prediction problem at the Bakel station on the Segenal River. The input variables are the observed flows at 3 upstream stations (including two uncontrolled stations on tributaries: Oualia on Bakoye, Gourbassi on Falémé and one station Manantali, which is controlled by the dam of Manantali), see Figure 1.

The watershed of the Bakel station (Figure 1), divided between Senegal, Mali, Mauritania and Guinea, and covers an area of about 289,000 km^2. It provides nearly all water in the Senegal river. The time series used in this work was obtained from the Senegal River Management Agency (OMVS: Organisation de Mise en Valeur du Fleuve Senegal) database. The time series at each stations covers the period from 1988 (date where the Manantali dam was put in operation) to 2008.. The statistical characteristics including mean discharge M, standard deviation σ, the coefficient of variation Cv and coefficient of skewness Cs are shown in Table 1.

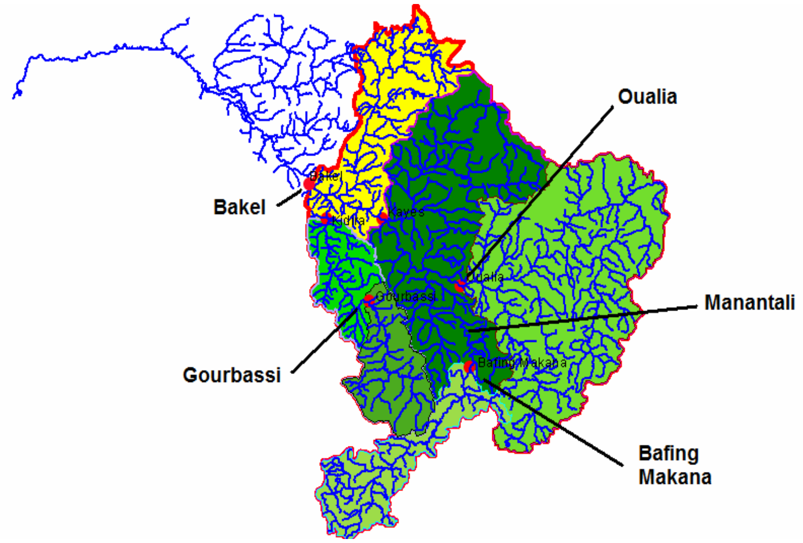

Figure 1: Senegal River Basin upstream from Bakel with sub-basins.

Table 1 Statistics of stream flows for each station

Stations	Rivers	Mean discharge (m^3/s)	Standard deviation σ	Coefficient of variation Cv	Coefficient of skewness Cs
Oualia	Bakoye	77	41	1.03	0.54
Gourbassi	Falémé	76	31	0.65	0.41
Manantali	Bafing	214	77	1.22	0.36
Bakel	Senegal	446	167	1.12	0.37

2 Methods

2.1 The Multi-layered Perceptron Networks

An Artificial Neural Network can be defined as a complex compound of interconnected elementary computation units (the neurons). Neurons are organized into groups or layers and can be connected in different ways. It is the topology of connections between neurons that defines the architecture of network. The task is defined by the network designer and is shown as examples with a set of input values and a set of desired outputs. Then the network must learn these examples and be able to provide correct answers for other unknown entries. Learning is the procedure for estimating the parameters of the network itself to satisfy a given performance criterion. It is performed according to an algorithm specific to the network architecture.

There are several types of ANN. The Multiple Layer Perceptron (MLP) is the most commonly used one, especially in nonlinear regression problems. MLP comprises one or more hidden layers with the activation function of sigmoid type and an output layer.

The hidden layer neurons receive information from the neurons (or units) of layer c - 1 and are connected to neurons of layer c + 1. There is no connection between neurons of the same layer. Each neuron in the output layer performs a non-linear function of the inputs of the network. For more details on MLP, see Hornik, Stinchcombe et al(1989), Eberhart and Dobbins(1990). The determination of the parameters of the MLP network is obtained from a supervised learning mode.

In the mode of learning, the outputs must adapt the ANN parameters to minimize the difference between the output of the system and of the model. The learning technique used in most MLP networks is back propagation (Hintont et al. 1986, Rumelhart 1986, Hinton et al. 2002).

In the application of ANN, the basin was considered having three inputs and one output. The inputs consist average daily flows observed at three hydrometric stations: Oualia and Gourbassi located on tributaries of uncontrolled Bakoye, Faleme station and Manantali dam at the outlet of the same name. The outlet station Bakel is used as the output of the system. Among them Manantali is subject to the management rules of Manantali dam (see Figure 1).

The model has been written in the form of

[1] $Q_B(t)=f(Q_G(t-h),Q_O(t-h),Q_M(t-h))$

where

Q_B, Q_G, Q_O and Q_M denote the discharge at Bakel, Gourbassi, Oualia and Manantali respectively, f (•) is a non-linear function, t is the time of prediction and the h is the prediction delay (days).

The structure and parameter identification of a MLP is to set the number of entries, the number of hidden layers, the number of neurons per layer coated and determining an algorithm to learn parameters (synaptic weights and biases) of network. In this study, the MLP network has three inputs and one output. However, there is no procedure to determine the number of hidden layers and of neurons in these layers. After several attempts, a MLP network with 3 input neurons, one hidden layer comprising four neurons and an output layer of one neuron(donated as MLP [3 4 1]) was used (see Figure 3).

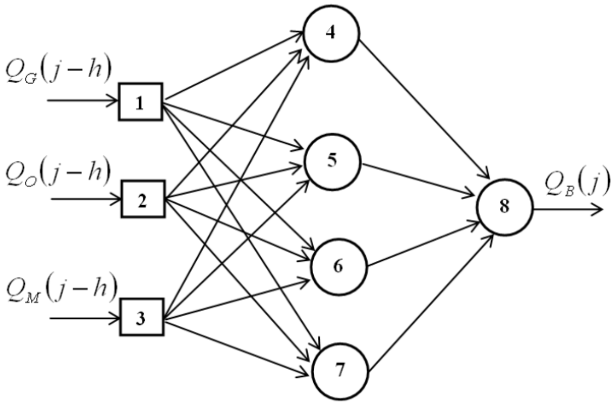

Figure 3: Model structure for the Bakel Station discharge prediction MLP network

The proposed MLP structure was used to predict the discharge at Bakel with delay h of 1 to 7 days. For each prediction delay h, the learning process A and validation process V over the shown periods were constituted as Equation 2 and Equation 3:

[2] $A=[Q_G(t)\ Q_O(t)\ Q_M(t)\ ;\ Q_G(t+1)\ Q_O(t+1)\ Q_M(t+1)\ ;\ \ldots\ ;Q_G(t+N-h)\ Q_O(t+N-h)\ Q_M(t+N-h)]$

[3] $V=[Q_B(t+h)\ ;\ Q_B(t+h+1)\ ;\ \ldots\ ;\ Q_B(t+h+N-1)]$

where N is the size of the database, and the delay h=1,2,…,7.

For each step of predicting a network MLP [3 4 1] is identified with the corresponding learning database.

2.2 Kalman filter

The Kalman Filter is a minimum variance estimation in the special case when the system is a linear stochastic dynamical system. The system could be described by the coupled equations:

[4] $x_{k+1}=M_kx_k+B_ku_k+\square_k$

[5] $y_k^o=H_kx_k+\square_k$

Where x is the state vector with dimension of n, y is the observation vector with a dimension of m. u_k here stands for the forcing term, whose coefficient B_k is often set to zero to simplify the notation as its determination does not affect the estimation process(Drécourt 2003). M_k is the linear state operator at time k, H_k is the observation operation at time k. η_k and ε_k are two independent sequences of zero-mean white noises with covariance of Q_k and R_k respectively.

Kalman filter is very flexible when it comes to implementation as the variables interested can be allocated either in state vector or observation plus multi combinations with the parameters. Here we assume the Bakel station discharge Q_b can be represented by a linear combination of historic three upstream flows and a constant,

[6] $Q_b(t)=k_1*Q_o(t-h)+k_2*Q_g(t-h)+k_3*Q_m(t-h)+k_4$

where $Q_b(t)$ is the discharge of Bakel at time t, $Q_o(t-h)$, $Q_g(t-h)$ and $Q_m(t-h)$ are the discharges of Oualia, Gourbassi and Manantali at the time t-h, h=1,…,7 day(s) is the delay or the prediction period, k_i, i = 1, …, 3 are coefficients of each upstream discharge while k_4 is the constant term .

Based on this relationship, we can set the Kalman filter with the state vector

[7] $x=(Q_b,k_1,k_2,k_3,k_4)$

and the observation

[8] $y=Q_b$

Thus we have the model matrix

[9]
```
M(:,:,t)=[0 Qo(t) Qg(t) Qm(t) 1;
          0   1     0     0   0;
          0   0     1     0   0;
          0   0     0     1   0;
          0   0     0     0   1];
```

and the observation matrix

[10]   H=[1 0 0 0 0]

Here we put the coefficients k_1, k_2, k_3 and k_4 in the state vector rather than the model because we would need the prior knowledge of the coefficients if we put them in the model matrix and this procedure can only be implemented offline and the coefficients vary with the length variation of the time series, making the offline calculation hard to implement.

One issue in Kalman filter model implementation is the determination of model error covariance matrix Q and observation error covariance matrix R, also known as the model noises. The existing methods either have too many assumptions (Mehra 1970) or too complicated to apply (Odelson, Rajamani et al. 2006). A commonly accepted perspective is a satisfactory result would be hard to obtain no matter how superior the method is without prior knowledge of the specific problem. Some researches show that the filtered result might not be as sensitive to the error covariance, especially the measurement error R, as people thought (Bergman and Delleur 1985). Thus the manual trial of Q and R is adopted in this study.

3 Results

The Root Mean Square Error (RMSE) and Nash-Sutcliffe coefficient (NS) are selected to evaluate the performance of three methods. Notice only validation period results have been analysed in BP ANN (1998-2002 for learning, 2003-2006 for validation). The RMSE and NS for each methods as well as a combination of Stepwise Regression and Kalman filter (STWKF) have been presented in Figure 4 and Figure 5.

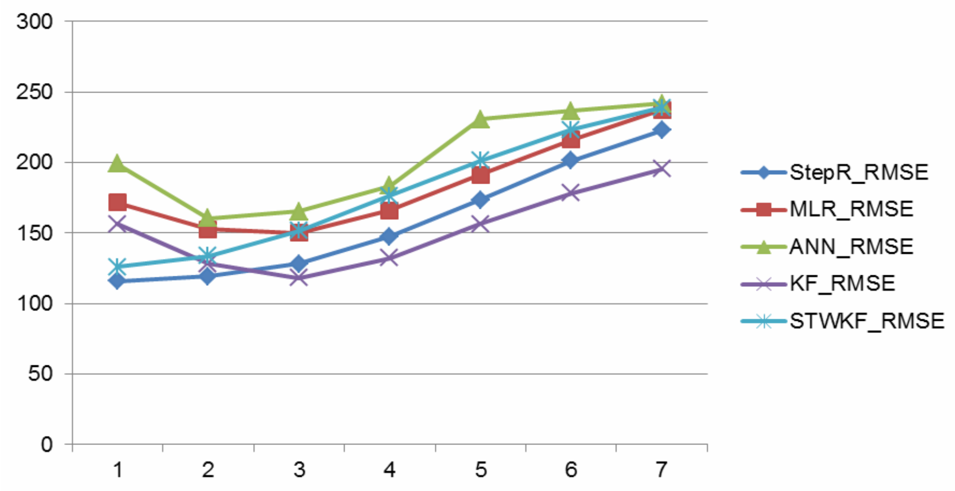

Figure 1: RMSE versus prediction delay h for different methods (1998-2008) (StepR_RMSE, MLR_RMSE，ANN_RMSE, KF_RMSE and STWKF_RMSE stand for the RMSE of Stepwise Regression, regular Multiple Linear Regression, ANN, Kalman filter and Stepwise Regression+Kalman filter)

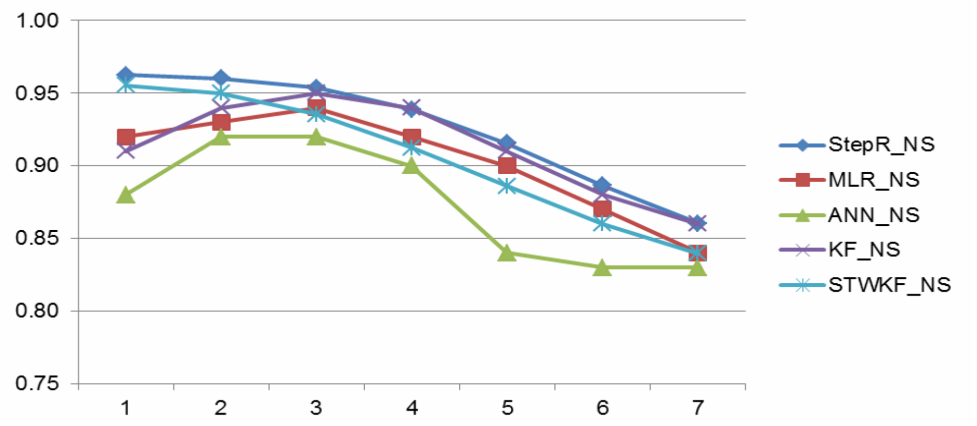

Figure 5: NS versus prediction delay h for different methods (1998-2008) (StepR_NS, MLR_NS, ANN_NS, KF_NS and STWKF_NS stands for the NS of Stepwise Regression, regular Multiple Linear Regression, ANN, Kalman filter and Stepwise Regression+Kalman filter)

3.1 Stepwise Regression

Stepwise Regression is applicable especially when the goal is to produce a predictive model that is parsimonious and accurate. When Stepwise Regression is used some form of validation analysis is necessary. Cross validation is used in this study. To do cross validation we randomly split the data set into a 75% training sample and a 25% validation sample. The training sample was used to develop the model. The effectiveness was assessed on the validation sample.

Compare the R^2 of the 25% validation sample to the R^2 of the 75% training sample: if the shrinkage (R^2 of the later to the former) is 2% or less, we conclude that validation is successful. Take the case of prediction delay h=7 days, the R^2 for the learning sample and validation sample is 0.983 and 0.972 respectively. The shrinkage rate is thus 1.12%, which means the application of Stepwise Regression here is legitimate.

The comparison for the prediction of different delays between regular Multiple Linear Regression and Stepwise Regression is shown in Figure 4 and Figure 5. Frankly, the multiple linear regression results are good enough for operational prediction work but Stepwise Regression outperforms it in all delays.

3.2 ANN

Back Propagation (BP) algorithm is used in the ANN model. Problems such as trapping into local minima and slow convergence might arise as the initial weights for the network are random and technologies like global search techniques have been suggested to evade the drawbacks of BP(Chang, Lin et al. 2012). As a basic study to assess the power of ANN, these technologies have not been considered here while they could be a potential topic in future.

Different implementations of BP ANN could have slightly different results while basic identical patterns can still be recognized. The best prediction delay is two to three days though it is hard to determine whether two days or three days is the perfect delay as the computation load is huge and large number of implementation is unacceptable. With the increase of the prediction delay, the effectiveness of ANN will decline.

3.3 Kalman filter

The implementation of Kalman filter is much easier than the development of an ANN model. It is found that Kalman filter has the best performance also when prediction delay h=2 days to 3 days. In fact, all

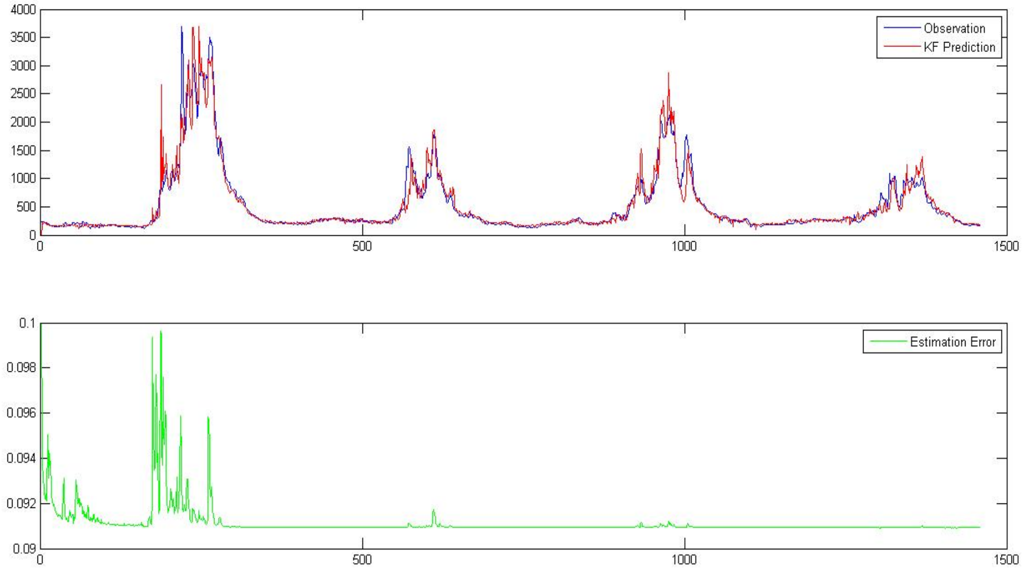

Figure 6: Kalman filter prediction results for prediction delay h=2 days (In the upper figure, blue is the observation, red is the prediction; in the lower figure, green is the estimation error)

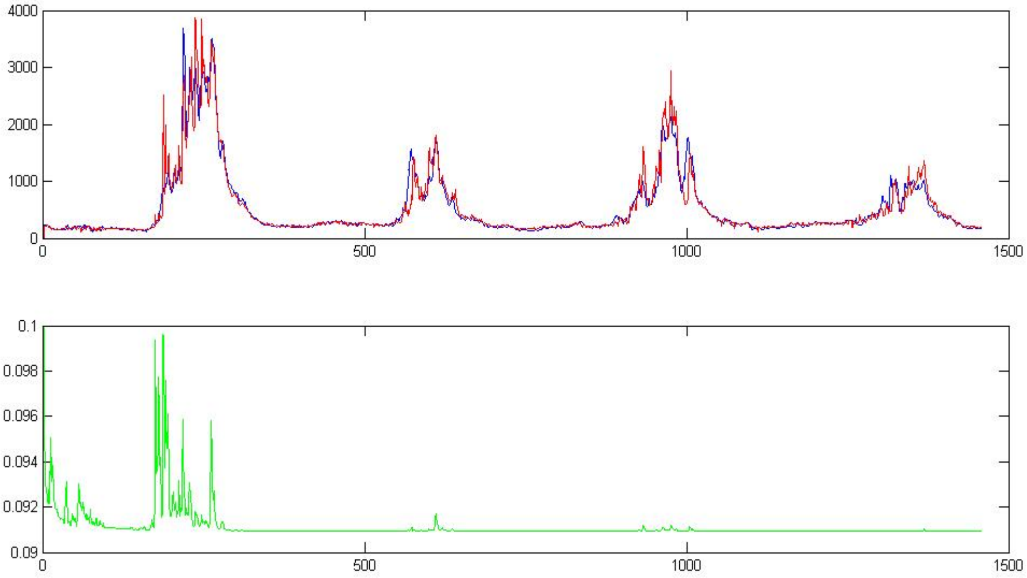

Figure 7: Kalman filter prediction results for prediction delay h=3 days. (In the upper figure, blue is the observation, red is the prediction; in the lower figure, green is the estimation error)

these methods work best with the prediction delay h=2 to 3 days and have their accuracy decline with the increase of the delay. The Kalman filter is a recursive method with no learning period .

As shown in Figure 4 and Figure 5, the Kalman filter together with Stepwise Regression are superior to ANN model. But their strength is different: Stepwise Regression tends to work better when the prediction period is as short as two to three days though Kalman performs consistently well as the prediction delay grows. The two days and three days ahead Kalman filter predictions are compared to observations in Figure 6 and Figure 7. As we can see, the prediction agrees with the observations pretty well in both peak periods and base flow periods. Since a common problem in streamflow forecasting is the loss of performance during the flood season, the Kalman filter seems to be a good alternative in accurately predicting streamflow. The Kalman filter outperforms all the models tested in this work at least on this point.

Since the Kalman filter and Stepwise Regression are the best two methods here, we hereby considered combining the two to have even an even better performance. Use Stepwise Regression to choose the coefficients for the possible independent variables in the state function for each prediction lead then update them with a Kalman filter. The effectiveness of this method, namely STWKF, has been displayed in Fig 4 and Fig 5. When the prediction delay is shorter than 3 days, this combined model has the similar power between Kalman filter and Stepwise regression. As the delay increases, the combined model could not compete with either Kalman filter or Stepwise Regression and also even worse than the regular multiple linear regression. However, it is still better than ANN model.

4 Conclusion and Future works

Stepwise Regression and Kalman filter are found to be more accurate and efficient compared to BP ANN model in this study. The BP ANN model requires more parameters setting and computation demanding than Stepwise Regression and Kalman filter though we did not show the comparison here.

The prediction delay for which most models performed best is two to three days. Stepwise Regression was found more powerful when the prediction delay is shorter than three days, while Kalman filter is more stable when the prediction delay is longer than three days and shorter than 7 days.

We also found that the combination of the two best methods, Stepwise Regression and Kalman filter, could not necessarily produce a better model.

This study is based on a very special case with long records of three simultaneous upstream daily discharge records, while this is usually not the common case in real life measurement. Thus a potential direction is to assess different methods performance with fewer inputs. And also, we used the original data as inputs directly, while some scholars suggested that mathematical transformation like using the logarithms of the original data might make the models, especially Kalman filter, work better. Finally, this paper only explored the annual stream flow prediction as a whole but did not analyze the case of wet season and dry season respectively. It could also be part of the future work.

References

Bergman, M. and J. Delleur (1985). "Kalman Filter Estimation and Prediction of Daily Stream Flows:I.Review,Algrorithm and Simulation." *JAWRA Journal of the American Water Resources Association* **21**(5): 815-825.

Chang, Y. T., J. Lin, J. S. Shieh and M. F. Abbod (2012). "Optimization the initial weights of artificial neural networks via genetic algorithm applied to hip bone fracture prediction." *Advances in Fuzzy Systems* **2012**: 6.

Drécourt, J. P. (2003). "*Kalman filtering in hydrological modelling*." Hørsholm, Denmark, DAIHM.

Eberhart, R. C. and R. W. Dobbins (1990). *Neural network PC tools: a practical guide*, Academic Press Professional, Inc.

Hornik, K., M. Stinchcombe and H. White (1989). "Multilayer feedforward networks are universal approximators." *Neural networks* **2**(5): 359-366.

Mehra, R. (1970). "On the identification of variances and adaptive Kalman filtering." *Automatic Control, IEEE Transactions on* **15**(2): 175-184.

Odelson, B. J., M. R. Rajamani and J. B. Rawlings (2006). "A new autocovariance least-squares method for estimating noise covariances." *Automatica* **42**(2): 303-308.

Rumelhart, D. E., G. E. Hintont and R. J. Williams (1986). "Learning representations by back-propagating errors." *Nature* **323**(6088): 533-536.