T. Darras, *et al.,* 2018. Ensemble model to enhance robustness of flash flood forecasting using an Artificial Neural Network: case-study on the Gardon Basin (south-eastern France). *Boletín Geológico y Minero,* 129 (3): 565-578
ISSN: 0366-0176
DOI: 10.21701/bolgeomin.129.3.007

Ensemble model to enhance robustness of flash flood forecasting using an Artificial Neural Network: case-study on the Gardon Basin (south-eastern France)

T. Darras[1,3], A. Johannet[1], B. Vayssade[1], L. Kong-A-Siou[2] and S. Pistre[3]

(1) LGEI, Ecole des Mines d'Alès, 6 Avenue de Clavières, Univ Montpellier, 30319 Alès Cedex
anne.johannet@mines-ales.fr
(2) ENSEGID, 1 Allée François Daguin, 33607 Pessac, France
kongasiou.line@gmail.com
(3) Hydrosciences, Univ Montpellier, 2 Place Eugène Bataillon,
34095 Montpellier Cedex 5, France
severin.pistre@umontpellier.fr

ABSTRACT

During the last few decades neural networks have been increasingly used in hydrological modelling for their fundamental property of parsimony and of universal approximation of non-linear functions. For the purpose of flash flood forecasting, feed-forward and recurrent multi-layer perceptrons appear to be efficient tools. Nevertheless, their forecasting performances are sensitive to the initialization of the network parameters. We have studied the cross-validation efficiency to select initialization providing the best forecasts in real time situation. Sensitivity to initialization of feed-forward and recurrent models is compared for one-hour lead-time forecasts. This study shows that cross-validation is unable to select the best initialization. A more robust model has been designed using the median of several models outputs; in this context, this paper analyses the design of the ensemble model for both recurrent and feed-forward models.

Key words: cross validation, flash flood, forecasting, model selection, neural network.

Modelo de "ensemble" para incrementar la robusted de la predicción de avenidas utilizando redes neuronales artificiales: aplicación a la cuenca Gardon (sureste de Francia)

RESUMEN

En las últimas décadas el uso de redes neuronales en la modelización hidrológica ha aumentado, debido a su propiedad fundamental como aproximador universal y parsimónico de funciones no lineales. En el campo de la previsión de inundaciones, los perceptrones de alimentación directa (feed-forward) y de tipo multicapa recurrentes (recurrent multilayer) han confirmado su eficiencia. Sin embargo, la capacidad predictiva depende de los parámetros de inicialización del sistema neuronal. La eficacia del método de validación cruzada para seleccionar las condiciones óptimas de inicialización que conducen a las mejores predicciones ha sido analizada. La dependencia de la inicialización en los modelos de retroalimentación y de multicapa recurrente ha sido comparada para las predicciones con antelación de una hora. Nuestro trabajo demuestra que la validación cruzada no permite la selección de la mejor inicialización. Un modelo más robusto ha sido desarrollado gracias al uso de la mediana de los resultados de varios modelos; en ese contexto, este trabajo analiza la estructura de los meta modelos tanto para los sistemas basados en redes retroalimentadas como para aquellos basados en redes multicapa recurrentes.

Palabras clave: riada de ciclo rápido, previsión, redes neuronales artificiales, selección de modelos, validación cruzada.

VERSIÓN ABREVIADA EN CASTELLANO

Introducción

Las riadas rápidas son un tipo de respuestas de pequeñas cuencas hidrográficas a fenómenos de lluvias de intensidad importante. En los últimos 20 años, las riadas de ciclo rápido (flash flood) han ocasionado más de cien muertos y varios billones de euros de daños en el sureste de Francia. La previsión de estos fenómenos es por consiguiente un desafío científico que plantea cuestiones sociales y económicas de importancia. Sin embargo, con nuestros conocimientos actuales, los procesos hidrológicos que ocasionan estas riadas rápidas no han sido completamente comprendidos, lo cual impide el desarrollo de modelos físicos eficaces. Por esta razón las redes neuronales, que no requieren hipótesis físicas sobre el funcionamiento del sistema, sino solo una base de datos, representan una herramienta prometedora. En los últimos años, el modelo particular de los «perceptrones multicapa» ha sido cada vez más utilizado en todos los dominios científicos debido a sus capacidades de aproximación universal y de parsimonia. Este tipo de modelo ha demostrado su eficacia para la identificación de la relación entre la lluvia y el caudal. La capacidad de este perceptron multicapa en la previsión de las riadas rápidas ha sido también demostrada. En esta perspectiva, nuestro trabajo propone una metodología de concepción de un modelo neuronal más robusto de previsión de riadas rápidas.

Métodos

Como las redes neuronales pertenecen a la categoría de sistemas estadísticos, es necesaria una base de datos lo más exhaustiva posible. Esta base de datos debe incluir los valores de las variables de entrada y de salida que representen los diferentes comportamientos observables en la cuenca hidrográfica estudiada. Habitualmente, dos tipos de perceptron multicapa pueden ser utilizados: un modelo directo, no recurrente, y un modelo recurrente («feed forward» y«recurrent multilayer» en inglés, respectivamente). El primero utiliza las observaciones de caudal colectadas precedentemente como información del estado de la cuenca. El segundo no utiliza estas medidas, pero representa la dinámica del sistema mediante la inyección de los valores obtenidos en la salida del modelo en la entrada, haciendo bucles de retroalimentación. Estos modelos recurrentes son por consiguiente dinámicos y podrían teóricamente ser inestables, pero esta inestabilidad nunca se ha observado. En nuestro trabajo, la eficacia de los dos tipos de modelos ha sido comparada.

La concepción propuesta para el perceptron multicapa consiste principalmente en la selección del modelo mediante la definición del número de variables de entrada y del número de neuronas internas. Este número determina mecánicamente el número de parámetros del modelo, es decir, su nivel de complejidad.

Teniendo en cuenta el carácter no lineal del modelo, los parámetros se calculan gracias a un aprendizaje que minimiza el error cuadrático calculado entre los valores de predicción de la salida del modelo y los valores de caudal observados. La eficacia de generalización del modelo corresponde a su capacidad de predecir el comportamiento de un fenómeno de riada rápida que no esté presente en la fase de aprendizaje. La base de datos se divide en dos subconjuntos, uno que será utilizado en la fase de aprendizaje, y otro que servirá para analizar la capacidad de predicción del modelo en la fase de test. Dada la diversidad de las riadas, varios eventos presentes en la base de datos, entre los más importantes o representativos, se utilizarán de forma secuencial solamente en la fase de test. Así, cuatro eventos han sido sucesivamente eliminados de la base de datos de aprendizaje.

Dadas las incertidumbres considerables de la base de datos, el modelo debe ser concebido de forma rigurosa para que no sufra de sobre-aprendizaje debido al dilema sesgo-varianza. Ha sido demostrado que el error cometido por el modelo durante la fase de aprendizaje no puede ser considerado como un buen estimador del error de generalización, ya que durante este aprendizaje el modelo se puede especializar en la realización especifica del ruido del sistema estudiado. Este fenómeno se llama sobreajuste («overfiting» en inglés). El riesgo de sobreajuste aumenta con la complejidad del modelo. Para limitar este fenómeno se recomienda el uso de métodos de regulación que minimicen la varianza de la totalidad del test.

El primero de estos métodos es «la validación cruzada», que consiste en probar la eficacia del modelo en situaciones de validación sobre varios conjuntos de datos, y calcular una nota de validación como la media de los errores sobre todos los conjuntos de validación utilizados sucesivamente. En nuestro trabajo, hemos utilizado este método para determinar la complejidad óptima del modelo (número de entradas y número de neuronas internas). El segundo método de regulación es «el paro anticipado», que consiste en parar el aprendizaje antes que haya sobre-aprendizaje. En este caso, se necesita un subconjunto diferente de los subconjuntos de datos de aprendizaje y de test. Este subconjunto de «paro anticipado» es denominado frecuentemente en la literatura «subconjunto de validación». En nuestro estudio, los dos métodos, «paro anticipado» y «validación cruzada» han sido utilizados simultáneamente. Para especializar el modelo en las riadas impor-

tantes, la validación cruzada ha sido realizada solamente con los eventos más importantes de la base de datos; en estas condiciones, la llamamos «validación cruzada parcial».

En las primeras etapas del aprendizaje, es necesario inicializar los parámetros de los modelos de forma aleatoria. Como la eficacia del modelo es sensible a esta inicialización, proponemos igualmente la inicialización del modelo usando la validación cruzada.

Resultados y discusión

La validación cruzada se usa habitualmente para la selección del modelo; parece lógico por consiguiente su uso para la selección de la mejor inicialización, ya que este método conduce a la varianza minima sobre el conjunto de eventos utilizado para el aprendizaje. Para evaluar la eficacia de este método, se ha hecho el aprendizaje sobre 200 modelos recurrentes y 200 modelos no recurrentes, entre los cuales la única diferencia es la inicialización. La nota de validación cruzada ha sido calculada (media de los errores de predicción para todos los eventos utilizados en la validación cruzada). En una segunda fase, cada uno de los modelos ha sido clasificado en función de la nota obtenida en orden decreciente. Para cada una de las 200 inicializaciones, la calidad de la previsión de los dos tipos de modelos (recurrentes y no recurrentes) ha sido evaluada para los 4 eventos de mas intensidad de la base de datos (eventos n° 19, 23, 26 y 27). Se observa que los modelos que corresponden a las mejores notas en validacion cruzada no son en general los mejores modelos de previsión para los eventos mas intensos de la base de datos, ya sean modelos recurrentes o no recurrentes. Por consiguiente, en el contexto de las riadas rápidas de cuencas pequeñas, la validación cruzada no es un sistema de selección apropiado.

Para evitar esta dificultad, se propone el uso de un meta modelo compuesto de un cierto número de modelos generados mediante inicializaciones diferentes y que tiene como resultado el cálculo para cada paso de tiempo de la mediana de cada uno de los modelos individuales que lo componen. Seis meta modelos han sido analizados para cada tipo de arquitectura (recurrente y no recurrente). La mediana, med_i (i = 5, 10, 20, 30, 40 y 50) ha sido calculada respectivamente a partir de 5, 10, 20, 30, 40 y 50 modelos, que varían solamente en el número de miembros de cada conjunto. Los 200 modelos recurrentes y los 200 modelos no recurrentes concebidos anteriormente han sido empleados utilizando la técnica de "bootstrap": 1000 extracciones con reposición de meta modelos (i=5, 10, 20, 30, 40 and 50) han sido realizadas. La calidad de predicción de los meta modelos ha sido evaluada estadísticamente sobre estas repeticiones.

Comparado con un modelo simple, la técnica de meta modelos contribuye a una reducción significativa de la variabilidad en la salida para los dos tipos de modelos (recurrentes y no recurrentes). El número óptimo de modelos individuales que constituyen el meta modelo ha sido estimado en 10 para los dos tipos de estructuras.

Introduction

This study focuses on the ability of artificial neural networks to perform robust flash flood predictions. As one of the most important natural hazards in France, especially in the Mediterranean region (Llasat *et al.,* 2010, 2014; Price *et al.,* 2011), flash flooding requires efficient forecasts to limit social and economic damage (Borga *et al.,* 2011; Huet *et al.,* 2003). This phenomenon is a hydrological reaction of high-slopes and relatively small watersheds, up to a few hundreds of km^2 (Gaume *et al.,* 2009), to intense and localized rainfall (Ayral, 2005; Garambois *et al.,* 2014; Marchandise, 2007). Resulting flows reach thousands of $m^3.s^{-1}$ with response times of only a few hours (Montz and Gruntfest, 2002). Without rainfall prediction, the water level forecasts are limited by the flood response time. Physical processes involved in this hydrological reaction are complex and misunderstood, thus leading to forecasting difficulties using physical-based models. Hence, black-box modelling using machine learning approaches, such as neural networks, is proving to be an alternative (Abrahart and See, 2007; Dawson and Wilby, 2001; Toukourou *et al.,* 2011; Kong-A-Siou *et al.,* 2012, 2014). For this reason, this study was funded by the Central Service of Hydrometeorology and Flood Forecasting (SCHAPI in French) of the French Ministry of Ecology in order to deliver real time warnings using its web service (http://www.vigicrues.gouv.fr/). Regarding huge noises and uncertainties concerning the estimation of rainfall and measurements of water levels in such dangerous flood events, a special care was taken on bias-variance trade-off (Geman *et al.,* 1992) which can impediment a good real time generalization, on a never-observed event. It is thus mandatory to rigorously apply regularization methods. Bornancin-Plantier *et al.* (2011) shows that, using early stopping and model selection based on cross validation, the principal hyper-parameter to consider was the initial

parameter distribution before training. On the basis of this study, an investigation of the sensitivity of both predictors: recurrent and feed-forward to this initialization of the models was performed.

The paper is organized into five sections: after the introduction, the watershed of interest is described, and then a presentation of the state of the art of flood forecasting using neural networks follows. The fourth section presents the results and discussion, and the conclusion is presented in section 5.

Area of study

The *Gardon d'Anduze* is a Cevenol and Mediterranean watershed of some 545 km^2. The maximum difference of elevation in the catchment is over than 1,000 m, which is a feature that promotes the great velocity of the flows (Moussa *et al.,* 2007). The rainfall is measured in the catchment area by six rain gauge stations: *Anduze*, *Mialet*, *Saumane*, *Soudorgues*, *Saint-Roman-de-Tousque* and *Barre-des-Cévennes*. The first three also have water level recorders (Fig. 1).

The water level forecast is achieved at *Anduze*. The data used consists of (*i*) rainfall from the six rain gauge stations, and (*ii*) the water level at the *Anduze* station. The study period goes from 1992 to 2008 with an hourly time-step until 2002 that then becomes a five-minute time-step for both rainfall and water level data. For operational needs, the time series must be sampled at half-hour time steps. Re-sampling has thus been achieved on both periods, inducing supplementary noise for the former one.

Event-based modelling is the most relevant approach because of the specific processes involved during such events. Therefore, events have been selected from the whole time series. The criterion proposed by Artigue *et al*. (2012) to select events was used. When a rainfall threshold of 100 mm in 48 hours is measured in at least one of the six stations, an event is extracted. This criterion is chosen to avoid false warnings (induced by heavy rain without flooding). 51 events were thus selected. Table 1 presents the maximum, minimum, median and standard deviation of the maximum water level and the cumulative rainfall. Starting from this set, the most intense events are selected based on the rain. In this way, 19 events amongst the 51 were considered as intense

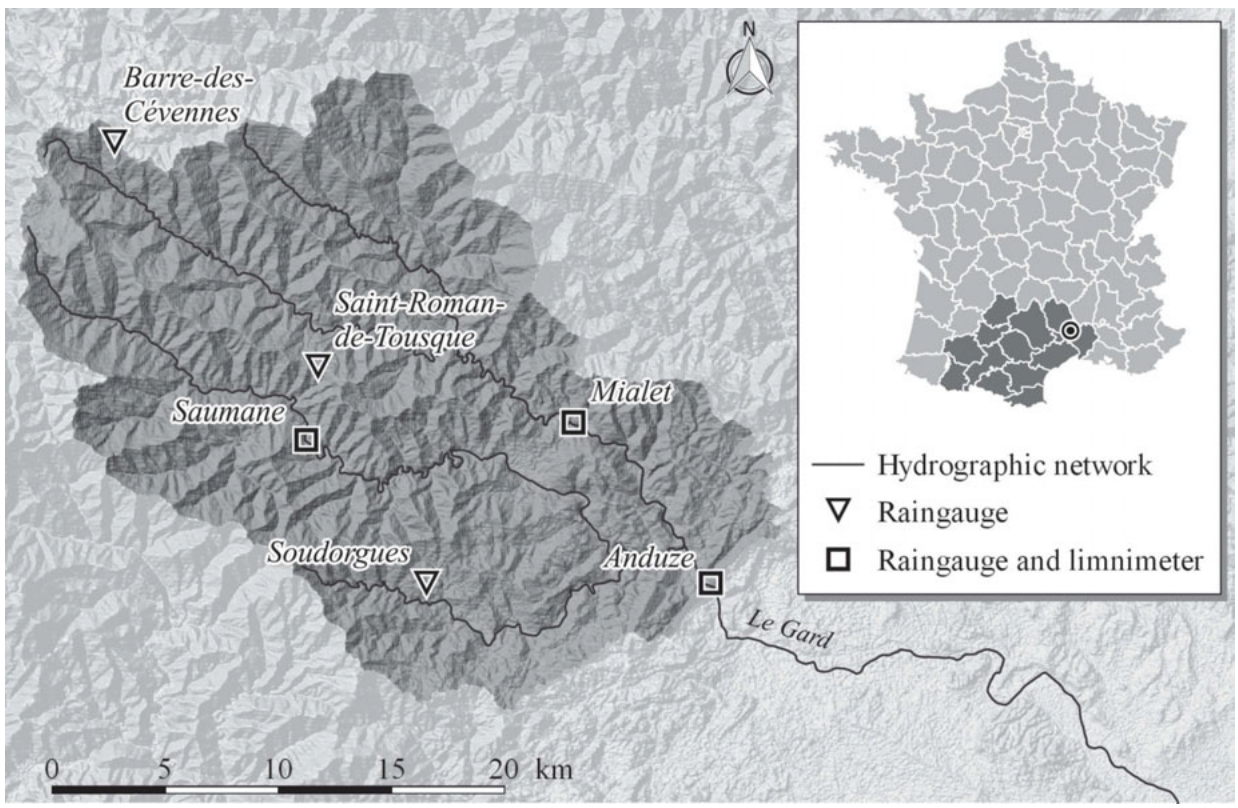

Figure 1. Location of rain gauges and water level recorders in the Gardon d'Anduze catchment area.
Figura 1. *Localización de los pluviómetros y estaciones de aforo en la cuenca vertiente del Gardon d'Anduze.*

events. The most intense event reached 9.88 m while the second one reached only 6.49 m, which represents almost a third less than the most intense event. The spatial and temporal rainfall distribution is generally different from one event to another. The system response can thus be complex and generates one or several flood peaks.

It should be noticed that measurements providing such data are strongly affected by errors and uncertainties: (*i*) the measurement of water levels is affected by variations of the bed during the flood, (*ii*) the accuracy of the rain gauges is assessed at between 10% and 20% (Marchandise, 2007), (*iii*) a lack of representativeness of these data relative to the studied processes is also possible as intense events can occur between two rain gauges and thus would not be measured at all. Water level is preferred to discharge because it is usually assumed that an error of 30% can affect the peak value of discharge (due to the transformation of level in the discharge: rating curve).

In order to evaluate the model performances in real time situations, several events, called test-sets, were discarded from the training database. Because this study is focused on flash floods we selected the most intense events as test-sets. To take into account different configurations of events, 4 events were used as test-sets: 19, 23, 26 and 27. It should be noted that event 19, the most intense event in the database, has not been measured in real time because of damage to the water level recorder. The limnigraph has been estimated by modelling approaches (SIEE, 2004).

As will be explained hereafter, early stopping is used. A dedicated set, called a stop-set, is considered separately from the database. During training, the chosen quality criteria are calculated on both the training set and the stop set. When it is observed that the quality criteria always improves on the training set but worsens on the stop set, meaning that the model begins to be unable to generalise, training is stopped. Early stopping is thus a method for preventing the model to train too much. Another event has thus to be discarded from the database to constitute the stop set. Toukourou *et al*. (2011), proposed an elegant way to select this event, it would be the best predicted event of the database, which implies that it is well accorded with the behaviour of the rest of the database. This selection is reached through cross-validation. Event 13 was selected in this way.

database		
Statistics	Maximum water level (m)	Cumulative rainfall (mm)
Maximum	9.88	403
Minimum	1.00	75
Median	3.16	178
Standard deviation	1.53	74

Table 1. Maximum, minimum, median and standard deviation of the maximum water level and cumulative rainfall of the events database

Tabla 1. *Máximo, mínimo, mediana y desviación estándard para el nivel máximo de agua y lluvia acumulada, para los eventos presentes en la base de datos.*

Methods

The chosen model is based on the multi-layer perceptron, because it displays the two properties of universal approximation and parsimony, the latter being due to the nonlinearity relative to model inputs and parameters (Barron, 1993). The multi-layer perceptron is a feed-forward or a recurrent neural network with one hidden layer of n_h hidden neurons and one output neuron. The hidden layer neurons are non-linear and apply a sigmoid function. On the other hand, the output neuron is linear; its output is equal to the weighted sum of its inputs. This specific architecture provides the property of universal approximation (Hornik *et al*., 1989). For further details on neural networks, the interested reader is referred to Dreyfus (2005).

Based on Artigue *et al.* (2012) a specific architecture was proposed to deal with huge rainfall that should induce saturations of sigmoids, purely linear functions that are added to the model (Fig. 2). Bornancin-Plantier *et al*. (2011) showed that this particular architecture diminishes the sensitivity of the model to the initialization of parameters. Nevertheless as the study of Bornancin-Plantier *et al.* (2011) was only performed on the feed-forward models, we propose to extend this study to the recurrent model in this paper.

Feed-forward neural network

The feed-forward neural network using linear and non-linear parts implements the following equation:

$$y(k+l)=g_{LIN}(\mathbf{u}(k), \ldots \mathbf{u}(k-1), \ldots, \mathbf{u}(k-m_{lin}+1);\theta)+ g_{NN}(y^p(k), y^p(k-1), \ldots, y^p(k-r_{nn}); \mathbf{u}(k), \mathbf{u}(k-1), \ldots, \mathbf{u}(k-m_{nn}+1);\theta) \quad (1)$$

where y is the estimated water level, k is the discrete time, y^p is the measured water level, u is the vector of

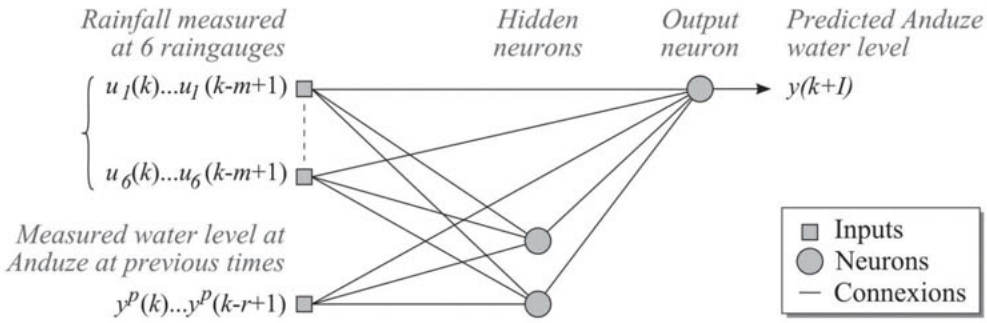

Figure 2. Multi-layer perceptron with linear links (notations will be explained in 3.1; *r* is the order of the model).
Figura 2. Perceptrón multicapa con unions lineales (la notación se explica en el apartado correspondiente; *r* es el orden del modelo).

exogenous variables, r_{nn} is the order of the model, m_{lin} is the width of the sliding widow of rainfall information of the linear part, m_{nn} is the width of the sliding widow of rainfall information of the non-linear part, I is the lead time of forecasts, θ is the matrix of parameters and g_{LIN} and g_{NN} are respectively the linear and the non-linear functions implemented by the feed-forward neural network. It can be noticed that sliding time windows of the linear and non-linear parts were chosen equally.

Recurrent neural network

The recurrent model was designed for this study. As proposed by Artigue *et al.* (2012) a supplementary input: the cumulative rainfall from the beginning of the event was added in order to represent the soil moisture indication which was provided by the measured water level in the previous feed-forward model. This soil moisture indication appears to have a significant impact on flash flood geneses and evolutions (Anctil et al., 2008; Cosandey and Robinson, 2000; Le Lay and Saulnier, 2007; Nikolopoulos *et al.,* 2011; Tramblay *et al.,* 2010). Cross correlation (rainfall, water level), for each gauge station, was used to define an interval of investigation of temporal window width. The best temporal window width was then selected using partial cross validation (as defined by Toukourou *et al*., 2011). Persistency was chosen as the performance criteria (presented in the next section). The results provided the following temporal window widths (Table 2). For the cumulative rainfall inputs in both the non-linear and linear parts only the current time step *k* was taken into account. In the recurrent neural network using linear and non-linear parts, the output can be expressed as (using the same notations as previously):

$$y(k+I)=g_{LIN}(\mathbf{u}(k), \mathbf{u}(k-1), ..., \mathbf{u}(k-m_{lin}+1);\theta)+g_{NN}(y(k+I-1), y(k+I-2), ..., y(k+I-r_{nn}); \mathbf{u}(k), \mathbf{u}(k-1), ..., \mathbf{u}(k-m_{nn}+1);\theta) \quad (2)$$

Performance criteria

In order to assess the performance of models, three criteria are used: R^2, persistency and the synchronous percentage of pic discharge (S_{PPD}).

The Nash-Sutcliffe efficiency (Nash and Sutcliffe, 1970), or R^2, is the most commonly used criterion in hydrology:

$$R^2 = 1 - \frac{\sum_{k=1}^{n}\left(y_p^k - y^k\right)^2}{\sum_{k=1}^{n}\left(y_p^k - \overline{y_p}\right)^2} \quad (3)$$

The nearest to 1 the Nash-Sutcliffe efficiency is, the better the results are. Nevertheless this criterion can reach good values even if the model proposes bad forecasts (Moussa, 2010). To avoid this problem, persistency is used.

$$C_p = 1 - \frac{\sum_{k=1}^{n}\left(y_p^{k+I} - y^{k+I}\right)^2}{\sum_{k=1}^{n}\left(y_p^{k+I} - y_p^k\right)^2} \quad (4)$$

Persistency (Kitanidis and Bras, 1980) provides information on the prediction capability of the model compared to the naive forecast. The naive forecast postulates that the output of the process at time step $k+I$ (where I is the lead time of forecast) is the same as the value at time k. The nearest to 1 the persistence efficiency is, the better the results are. A positive result means that model prediction is better than the naive prediction.

The synchronous percentage of the peak dis-

	Input variables	Initial temporal window widths	Intervals of investigation	Selected temporal window widths
Nonlinear part	Rainfall at *Anduze*	*k* to *k*-4	*k* to *k*-6	*k* to *k*-4
	Rainfall at *Mialet*	*k* to *k*-4	*k*-7 to *k*-18	*k* to *k*-9
	Rainfall at *Saumane*	*k* to *k*-4	*k*-7 to *k*-15	*k* to *k*-8
	Rainfall at *Soudorgues*	*k* to *k*-4	*k*-7 to *k*-18	*k* to *k*-13
	Rainfall at Saint-Roman-de-Tousque	*k* to *k*-4	*k*-10 to *k*-18	*k* to *k*-16
	Rainfall at Barre-des-Cévennes	*k* to *k*-4	*k*-10 to *k*-21	*k* to *k*-13
	Water level at *Anduze*	*k* to *k*-2	-	-
	Cumulative rainfall	-	*k* to k-3	k
	Recurrent outputs	-	-	*k* to *k*-2
Linear part	Rainfall at *Anduze*	*k* to *k*-4	*k* to *k*-6	*k* to *k*-1
	Rainfall at *Mialet*	*k* to *k*-4	*k*-7 to y-18	*k* to *k*-8
	Rainfall at *Saumane*	*k* to *k*-4	*k*-7 to y-15	*k* to *k*-15
	Rainfall at *Soudorgues*	*k* to *k*-4	*k*-7 to y-18	*k* to *k*-18
	Rainfall at Saint-Roman-de-Tousque	*k* to *k*-4	*k*-10 to y-18	*k* to *k*-18
	Rainfall at Barre-des-Cévennes	*k* to *k*-4	*k*-10 to *k*-21	*k* to *k*-11
	Water level at *Anduze*	*k* to *k*-2	-	-
	Cumulative rainfall	-	*k* to k-3	k

Table 2. Temporal windows widths of variables.
Tabla 2. *Anchuras de las ventanas temporales de las variables.*

charge: S_{PPD} (Artigue *et al.*, 2012) is a relevant criterion to assess the flash flood forecasting performance of a model on the peak discharge. It shows the forecast quality at the peak discharge through the ratio between the observed and forecast discharges at the observed peak discharge moment.

$$S_{PPD} = 100 \frac{y^{k_{max}}}{y_p^{k_{max}}} \quad (5)$$

where k_{max} is the instant of the observed peak discharge.

Regularization methods

Generalization is a major challenge for neural networks. Geman *et al*. (1992) showed that the training error is not a good estimator of the generalization error (*i.e.* the error in validation or test). The primary consequence of the bias-variance trade-off is overfitting. The model yields excellent results on the training set; however the results on the validation or test sets are rather poor. To overcome this drawback, regularization methods must be used (Kong-A-Siou *et al.,* 2011, 2012); such methods are essentially aimed at reducing model variance. The first method is cross-validation and was proposed in Stone (1974); it allows the choice of the optimal model complexity (*i.e.* number n_h of hidden neurons and selection of input variables). A second regularization method, early stopping, consists of stopping the training algorithm before experiencing overtraining (Sjöberg *et al.*, 1995); it necessitates an additional dataset, denoted as a stop set that is independent of both the training and test sets. Both early stopping and cross-validation methods have been applied simultaneously in this study. In order to specialise the model on the behaviour of intense events, the cross-validation is performed only on intense events where it is then called "partial cross-validation".

Results

Inefficiency of cross-validation to select the best initialisation

The use of partial cross-validation in order to select

the best initialization for operational prediction was assessed on both recurrent and feed-forward models. To reach this goal 200 initializations were done. The partial cross-validation score (S_{PCV}) was chosen in regard to both R^2 and persistence efficiency. The same 200 initializations were then run as in operational conditions on the four test-sets (most intense events). Figures 3 and 4 show the 200 partial cross-validation scores of the respectively feed-forward and recurrent models, using C_P, ranked in decreasing order; the rank is the same with the persistency criterion. The performance of each model, corresponding to each initialization is also drawn. It appears that the partial cross-validation is unable to select the best initialization for operational forecasting: the S_{PVC} decreases from the best score to the worst one. At the same time, the C_P scores for the test events evolve without structure around a horizontal straight line; they are not diminished. Consequently, the selection of a single initialization would lead to a non-robust model.

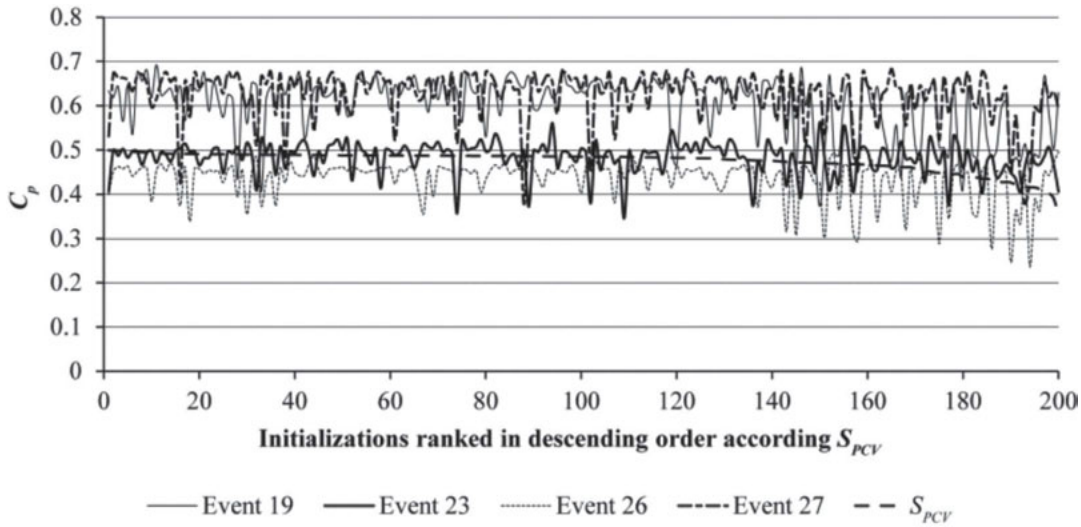

Figure 3. Persistence criterion (C_P) of the 200 feed-forward models tested on the four test-sets classed in decreasing order regarding partial cross-validation scores (S_{PCV}).
Figura 3. *Criterio de persistencia (C_P) de los 200 modelos "feed-forward" chequeados sobre los cuatro conjuntos de test en orden decreciente con respecto a la puntuación de la validación cruzada parcial (S_{PCV}).*

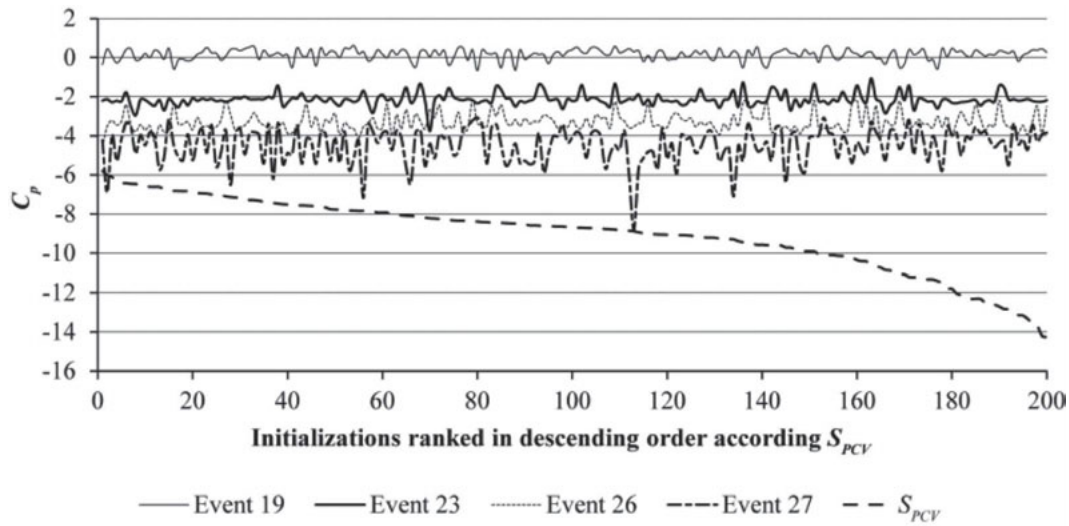

Figure 4. Persistence criterion (C_P) of the 200 recurrent models tested on the four test-sets classed in decreasing order regarding partial cross-validation scores (S_{PCV}).
Figura 4. *Criterio de persistencia (C_P) de los 200 modelos "recurrente" chequeados sobre los cuatro conjuntos de test en orden decreciente con respecto a la puntuación de la validación cruzada parcial (S_{PCV}).*

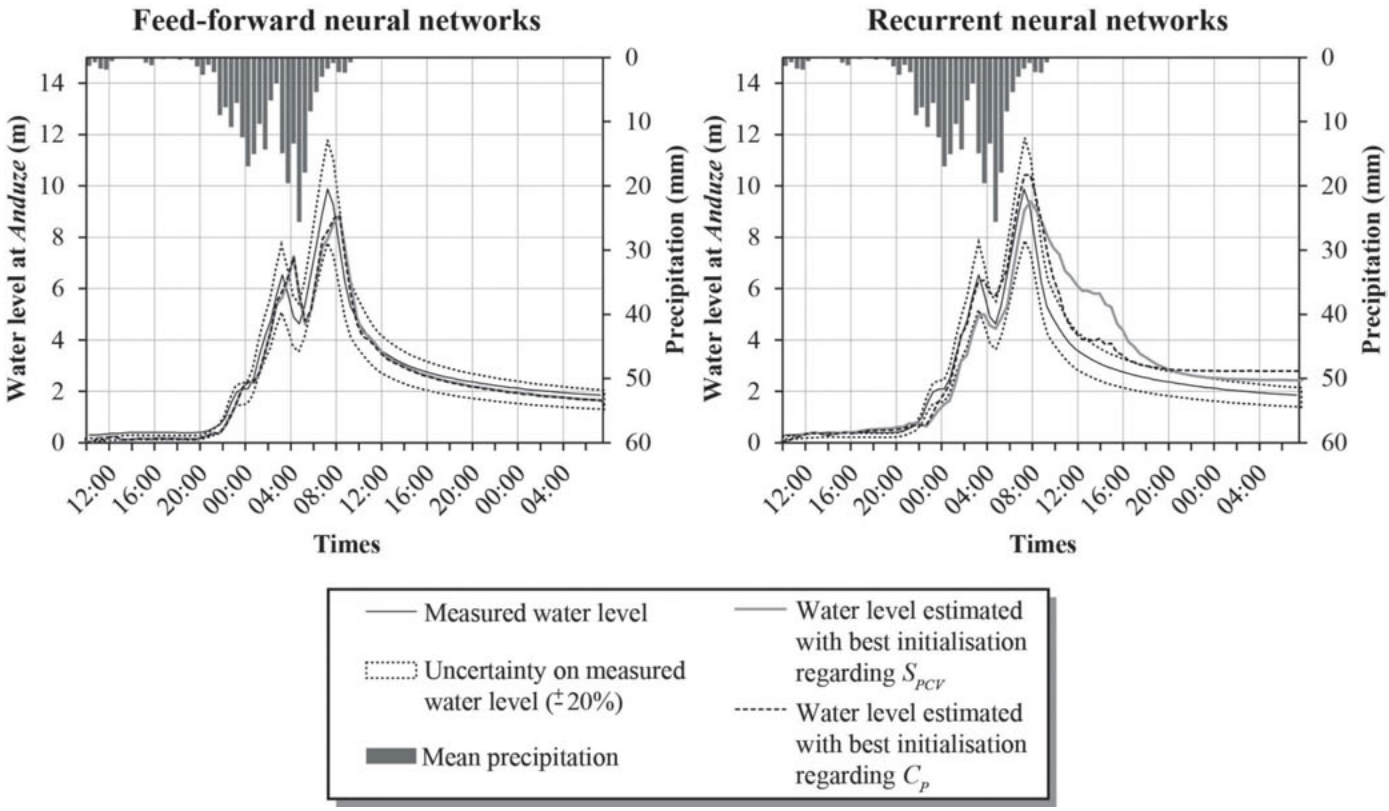

Figure 5. Hydrographs of observed water level, and estimated water level with best initialisations regarding SPCV and CP on test-set of the flash flood event that occurred on 8 and 9 September 2002 (event 19).
Figura 5. *Hidrogramas del nivel del agua observado y nivel estimado con la mejor inicialización de acuerdo a SPCV y CP sobre el conjunto de datos test del evento de avenida que ocurrió entre el 8 y 9 de septiembre de 2002 (evento 19).*

S_{PPD}		Without ensemble	med_5	med_{10}	med_{20}	med_{30}	med_{40}	med_{50}
Event 19	Minimum	67.7	74.4	77.7	78.3	79.4	79.8	80.0
	Maximum	90.3	89.1	86.8	84.1	82.8	82.8	82.4
	Delta	22.6	14.7	9.1	5.8	3.4	3.0	2.3
Event 23	Minimum	75.3	76.7	77.3	77.7	77.9	77.9	77.9
	Maximum	85.0	83.2	80.4	79.4	79.0	79.6	78.8
	Delta	9.7	6.5	3.1	1.7	1.2	1.7	0.9
Event 26	Minimum	63.7	71.3	71.7	76.9	77.7	78.1	78.1
	Maximum	82.2	80.1	79.5	79.3	79.3	79.1	79.0
	Delta	18.5	8.7	7.8	2.4	1.6	1.0	0.9
Event 27	Minimum	79.6	85.0	86.8	87.3	87.4	87.5	87.6
	Maximum	97.1	95.2	90.9	90.3	88.8	88.6	88.3
	Delta	17.5	10.2	4.2	3.0	1.4	1.0	0.7

Table 3. Minimum, maximum and Delta of the 1000 S_{PPD} provided by each med_5, med_{10}, med_{20}, med_{30}, med_{40} and med_{50} and the 200 initializations feed-forward models for each test-set.
Tabla 3. *Mínimo, máximo y Delta de los 1000 S_{PPD} proporcionado por cada med_5, med_{10}, med_{20}, med_{30}, med_{40} and med_{50} y las 200 inicializaciones de los modelos feed-forward para cada conjunto de test.*

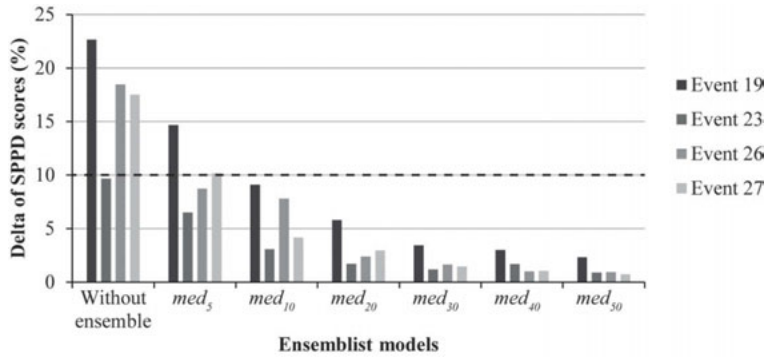

Figure 6. Feed-forward model. Evolution of the 1000 S_{PPD} Delta obtained for each test set versus the number of models in the ensemble. The threshold of acceptability is shown by a dashed line.
Figura 6. *Modelo "feed-forward". Evolución de 1000 S_{PPD} Delta obtenido para cada conjunto de test frente al número de modelos en el conjunto. El umbral de aceptación se muestra por una línea discontínua.*

S_{PPD}		Without ensemble	med_5	med_{10}	med_{20}	med_{30}	med_{40}	med_{50}
Event 19	Minimum	64.8	72.8	75.0	77.4	77.5	77.7	78.4
	Maximum	88.5	85.7	84.0	82.8	82.6	82.0	81.9
	Delta	23.7	12.9	9.0	5.4	5.1	4.3	3.5
Event 23	Minimum	61.9	63.1	63.9	64.7	64.5	65.1	65.2
	Maximum	73.8	70.6	68.8	68.3	68.1	67.4	67.2
	Delta	11.9	7.5	4.9	3.6	3.5	2.4	2.1
Event 26	Minimum	44.8	46.1	49.2	50.2	51.1	51.6	51.9
	Maximum	63.2	58.8	56.8	56.5	55.6	55.5	55.4
	Delta	18.3	12.7	7.6	6.3	4.4	4.0	3.5
Event 27	Minimum	63.9	64.4	64.8	65.6	65.7	65.7	65.9
	Maximum	74.9	72.1	70.4	69.8	68.4	68.3	68.2
	Delta	11.0	7.7	5.6	4.2	2.7	2.6	2.2

Table 4. Minimum, maximum and Delta of the 1000 S_{PPD} providing by each med_5, med_{10}, med_{20}, med_{30}, med_{40} and med_{50} and the 200 initializations recurrent models for each test-set.
Tabla 4. *Mínimo, máximo y Delta de los 1000 S_{PPD} proporcionados por cada med_5, med_{10}, med_{20}, med_{30}, med_{40} y med_{50} y las 200 inicializaciones recurrentes para cada conjunto de datos.*

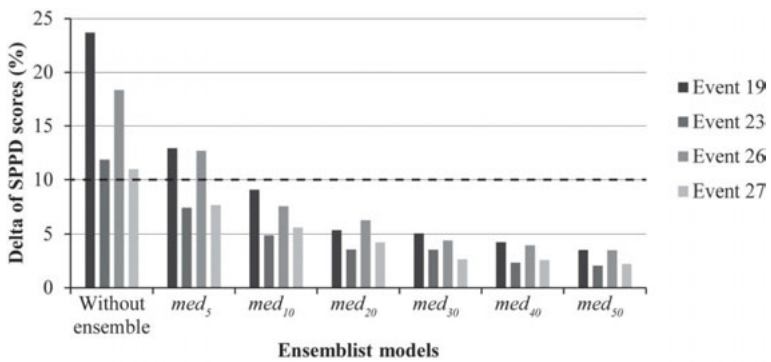

Figure 7. Recurrent model. Evolution of the 1000 S_{PPD} Delta obtained for each test set versus the number of models in the ensemble. The threshold of acceptability is shown by a dashed line.
Figura 7. *Modelo recurrente. Evolución de los 1000 S_{PPD} Delta obtenidos para cada conjunto de test frente al número de modelos en el conjunto. El umbral de aceptabilidad se muestra por una línea discontinua.*

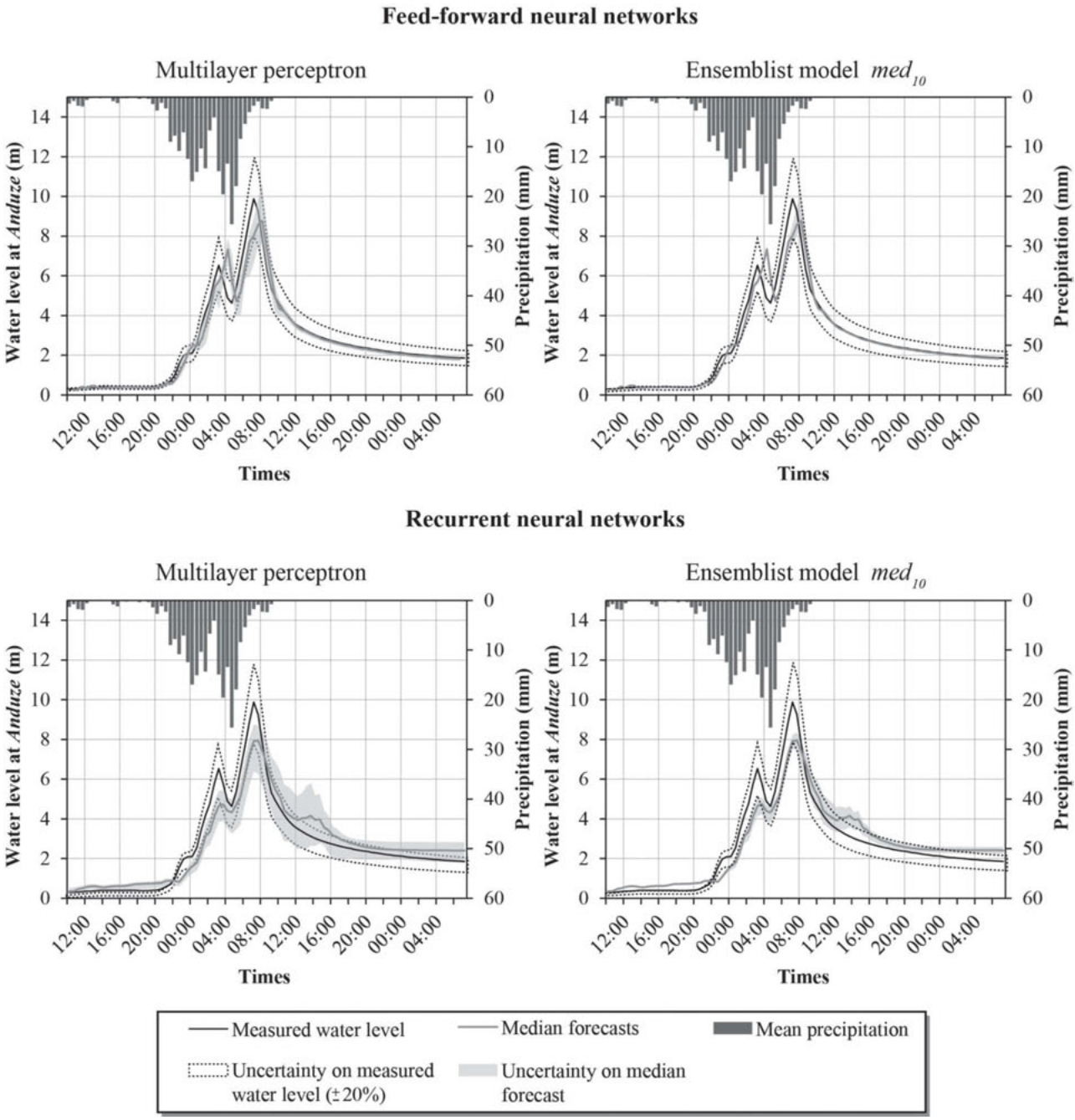

Figure 8. Medians and uncertainties of forecasts of feed-forward and recurrent multi-layer perceptron and ensemblist models med_{10} of the flash flood event that occurred on 8 and 9 September 2002 (event 19).

Figura 8. *Medianas e incertidumbres de los pronósticos del perceptron multi-capa recurrente y feed-forward y modelos de conjunto med_{10} del evento de avenida que tuvo lugar entre el 8 y 9 de septiembre de 2002 (evento 19).*

Figure 5 shows the hydrographs of estimated water level of the flash flood event that occurred on 8 and 9 September 2002 (*i.e.* event 19) obtained from model using the best initialisation regarding the S_{PCV} and the model using the best initialisation regarding the C_P on test-set. These hydrographs are drawn for feed-forward and recurrent models. This figure highlights the difference between best models *a priori* in generalisation, selected thanks to the partial cross-validation and the actual best models in generalisation. Differences appear significant, especially with the recurrent model.

Towards a more robust model

Because of the inefficiency of cross-validation to select the best initialization among the 200 tested, Bornancin-Plantier (2011) proposed using an "ensemble" strategy. This method is similar to the Ensemble Methods from Dietterich (2000). The number of models to take into account needed to be optimized: in order to limit investigations, only 6 possibilities were stated: 5, 10, 20, 30, 40 and 50 different models. Models designed from the median of each time-step output using respectively 5, 10, 20, 30, 40 and 50 models are hereafter celled: med_5, med_{10}, med_{20}, med_{30}, med_{40} and med_{50}. In order to be independent from the random procedure, 200 models for each test-set were used.

From these 200 models, 1,000 random draws, without reset for each number of models, were made. In this way, 1,000 forecasts for each med_i, (i=5, 10, 20, 30, 40 and 50) were obtained. Then, minimum, maximum and delta between the maximum and the minimum of S_{PPD} (called Delta hereafter) for the 1,000 forecasts were calculated. S_{PPD} is used in this part of the study because of its operational interest. The Delta highlights the variability of the forecasts.

Table 3 shows the minimum, maximum and delta of the 1000 S_{PPD} provided by the ensemble models for each med_i, (i=5, 10, 20, 30, 40 and 50), for each test-set, for the feed-forward models and Table 4 for the recurrent models. The minimum, maximum and Delta of the S_{PPD} provided by the 200 initializations are also noted. The evolution of the 1,000 S_{PPD} Delta obtained for each ensemble category is shown in Figure 6.

In order to select the optimal number of models to bring in the ensemble, we arbitrarily consider that the acceptable variation of SPPD must be inferior to 10% of the maximum value. This choice implies that the Delta of the SPPD must be inferior to 10% (remember that the SPPD is lower or equal to 1). Giving the threshold of D of SPPD, the med10 model is kept for the both feed-forward and recurrent models. It can be noted that this quality criterion is not really significant for the recurrent models that are very inefficient.

As can be seen in both Figures 6 and 7 the behaviours of models are very different from one event to another. This explains the difficultly in flash flood forecasting: each event is very different from another and thus induces a great dependency on the model selection. This sensitivity is reduced by the ensemble approach.

Figure 8 shows the medians and uncertainties of forecasts of feed-forward and recurrent multi-layer perceptron and ensemblist models med10 of flash flood event 19. A significant reduction of uncertainty is observed between classic multi-layer perceptron approach and the ensemblist in both feed-forward and recurrent models. This reduction is particularly important with the recurrent model for this event 19.

Conclusions

Reliable forecasting of flash floods is a difficult task to perform because of the importance of noise and uncertainties. The variability of the forecast may be important depending on the initialization of the network parameters. Moreover, the well-known cross-validation method appeared to be inefficient for selecting the best initialization regarding generalization performances. A more robust model was then studied for both recurrent and feed-forward multi-layer perceptron in an ensemble framework in calculating the median output of the number of models to determine. Experiments showed that the number of models to take into account using this method is low and is the same for feed-forward or recurrent model. Feed-forward models show good performances allowing the population to be efficiently warned when recurrent models are not as efficient as the previous ones. This difference is due to the lack of good informative inputs in the recurrent models such as the water level input used in the feed-forward models. Based on the feed-forward model, a real time forecasting prototype was designed and provided to the SCHAPI Warning Service. An in-depth learning inspired approach based on this study (Schmidhuber, 2015) could be interesting in order to better initialise the model.

Acknowledgements

The authors would like to thank the Grand Delta Flood Forecasting Services (SPCGD) for the availability of data. We also thank Dominique Bertin (with the Geonosis Company) for the development and the constant improvements the software RNF Pro and Miguel Lopez-Ferber for the Spanish translation of the abstract and abridged Spanish version. And finally, special thanks go to Bruno Janet from the Central Service of Hydrometeorology and flood forecasting support (SCHAPI) for its support for this study.

References

Abrahart, R. J. and See, L. M. 1997. Neural network modelling of non-linear hydrological relationships, *Hydrological Earth System Sciences*, 11 (5), 1563–1579.

Anctil, F., Lauzon, N. and Filion, M. 2008. Added gains of soil moisture content observations for streamflow predictions using neural networks, *Journal of Hydrology*, 359(3-4), 225–234.

Artigue, G., Johannet, A., Borrell, V. and Pistre, S. 2012. Flash flood forecasting in poorly gauged basins using neural networks: case study of the Gardon de Mialet basin (southern France), *Natural Hazards Earth System Sciences*, 12 (11), 3307–3324.

Ayral, P. A. 2005. Contribution à la spatialisation du modèle opérationnel de prévision des crues éclair ALHTAÏR, Université de Provence Aix-Marseille.

Barron, A. R. 1993. Universal approximation bounds for superpositions of a sigmoidal function, *IEEE Transactions on Information Theory*, 39 (3), 930–945.

Borga, M., Anagnostou, E. N., Blöschl, G. and Creutin, J. D. 2011. Flash flood forecasting, warning and risk management: the HYDRATE project, *Environmental Science Policy*, 14 (7), 834–844.

Bornancin-Plantier, A., Johannet, A., Borrell Estupina, V., Roussel-Ragot, P. and Dreyfus, G. 2011. Conception de modèles de prévision des crues éclair par apprentissage artificiel, in EGU2011-1794, 2011, vol. 13.

Cosandey, C. and Robinson, M.: *Hydrologie continentale*, A. Colin., 2000.

Dawson, C. W. and Wilby, R. L. 2001. Hydrological modelling using artificial neural networks, *Progress in Physical Geography*, 25 (1), 80–108.

Diettrich, T. G. 2015. Ensemble Methods in Machine Learning, in Lecture Notes in Computer Science, p. 115, Springer Verlag, New-York. [online] Available from: http://www.eecs.wsu.edu/~holder/courses/CptS570/fall07/papers/Dietterich00.pdf (Accessed 16 June 2015).

Dreyfus, G. 2005. *Neural Networks: Methodology and Applications*, Softcover reprint of hardcover 1st ed. 2005 edition., Springer, Berlin; New York.

Garambois, P. A., Larnier, K., Roux, H., Labat, D. and Dartus, D. 2014. Analysis of flash flood-triggering rainfall for a process-oriented hydrological model, *Atmospheric Research*, 137, 14–24.

Gaume, E., Bain, V., Bernardara, P., Newinger, O., Barbuc, M., Bateman, A., Blaškovi ová, L., Blöschl, G., Borga, M., Dumitrescu, A., Daliakopoulos, I., Garcia, J., Irimescu, A., Kohnova, S., Koutroulis, A., Marchi, L., Matreata, S., Medina, V., Preciso, E., Sempere-Torres, D., Stancalie, G., Szolgay, J., Tsanis, I., Velasco, D. and Viglione, A. 2009. A compilation of data on European flash floods, *Journal of Hydrology*, 367 (1-2), 70–78.

Geman, S., Bienenstock, E. and Doursat, R. 1992. Neural Networks and the Bias/Variance Dilemma, *Neural Computing*, 4 (1), 1–58.

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

Huet, P., Martin, X., Prime, J.-L., Foin, P., Laurain, C. and Cannard, P. 2003. Retour d'expériences des crues de septembre 2002 dans les départements du Gard, de l'Hérault, du Vaucluse, des Bouches-du-Rhône, de l'Ardèche et de la Drôme., Inspection générale de l'Environnement, Paris, France. [online] Available from: http://cgedd.documentation.developpement-durable.gouv.fr/document.xsp?id=Cgpc-OUV00000419 (Accessed 24 March 2015).

Kitanidis, P. K. and Bras, R. L. 1980. Real-time forecasting with a conceptual hydrologic model: 2. Applications and results, *Water Resources Research*, 16(6), 1034–1044.

Kong-A-Siou, L., Johannet, A., Borrell, V. and Pistre, S. 2011. Complexity selection of a neural network model for karst flood forecasting: The case of the Lez Basin (southern France). *Journal of Hydrology*. 403 (3–4), 367–380.

Kong-A-Siou, L., Johannet, A., Valérie, B. E. and Pistre, S. 2012. Optimization of the generalization capability for rainfall–runoff modeling by neural networks: the case of the Lez aquifer (southern France), *Environmental Earth Sciences*, 65 (8), 2365–2375.

Kong-A-Siou, L., Fleury, P., Johannet, A., Borrell Estupina, V., Pistre, S. and Dörfliger N. , 2014. Performance and complementarity of two systemic models (reservoir and neural networks) used to simulate spring discharge and piezometry for a karst aquifer. *Journal of Hydrology*, 519 (D), 3178-3192.

Le Lay, M. and Saulnier, G. M. 2007. Exploring the signature of climate and landscape spatial variabilities in flash flood events: Case of the 8–9 September 2002 Cévennes-Vivarais catastrophic event, Geophysics Res. Letters, 34(13) [online] Available from: http://onlinelibrary.wiley.com/doi/10.1029/2007GL029746/full (Accessed 8 December 2014).

Llasat, M. C., Llasat-Botija, M., Prat, M. A., Porcú, F., Price, C., Mugnai, A., Lagouvardos, K., Kotroni, V., Katsanos, D., Michaelides, S. and others 2010. High-impact floods and flash floods in Mediterranean countries: the FLASH preliminary database, Advances in Geosciences, 23 (23), 47–55.

Llasat, M. C., Marcos, R., Llasat-Botija, M., Gilabert, J., Turco, M. and Quintana-Seguí, P. 2014. Flash flood evolution in North-Western Mediterranean, *Atmospheric Research*, 149, 230–243.

Marchandise, A. 2007. Modélisation hydrologique distribuée sur le Gardon d'Anduze; étude comparative de différents modèles pluie-débit, extrapolation de la normale à l'extrême et tests d'hypothèses sur les processus hydrologiques, Université Montpellier II-Sciences et Techniques du Languedoc. [online] Available from: http://www.ohmcv.fr/Documents/theses/these_marchandise-old.pdf (Accessed 8 December 2014).

Montz, B. E. and Gruntfest, E. 2002. Flash flood mitigation: recommendations for research and applications, *Global Environmental Change Part B Environmental Hazards*, 4 (1), 15–22.

Moussa, R. 2010. When monstrosity can be beautiful while normality can be ugly: assessing the performance of event-based flood models, *Hydrology Sciences Journal*, 55, 1074–1084.

Moussa, R., Chahinian, N. and Bocquillon, C. 2007. Distributed hydrological modelling of a Mediterranean mountainous catchment – Model construction and multi-site validation, *Journal of Hydrology*, 337 (1–2), 35–51.

Nash, Je. and Sutcliffe, J. V., 1970. River flow forecasting through conceptual models part I—A discussion of principles, *Journal of Hydrology*, 10 (3), 282–290.

Nikolopoulos, E. I., Anagnostou, E. N., Borga, M., Vivoni, E. R. and Papadopoulos, A. 2011. Sensitivity of a mountain basin flash flood to initial wetness condition and rainfall variability, *Journal of Hydrology*, 402 (3-4), 165–178.

Price, C., Yair, Y., Mugnai, A., Lagouvardos, K., Llasat, M. C., Michaelides, S., Dayan, U., Dietrich, S., Galanti, E., Garrote, L., Harats, N., Katsanos, D., Kohn, M., Kotroni, V., Llasat-Botija, M., Lynn, B., Mediero, L., Morin, E., Nicolaides, K., Rozalis, S., Savvidou, K. and Ziv, B. 2011. The FLASH Project: using lightning data to better understand and predict flash floods, *Environmental Science Policy*, 14 (7), 898-911.

Schmidhuber, J., 2015. Deep Learning in Neural Networks: An Overview. *Neural Networks,* 61, 85–117.

SIEE, 2004. Validation des relevés hydrométriques de l'événement des 8 & 9 septembre 2002, Direction Départementale de l'Equipement du Gard.

Sjöberg, J., Zhang, Q., Ljung, L., Benveniste, A., Delyon, B., Glorennec, P.-Y., Hjalmarsson, H. and Juditsky, A. 1995. Nonlinear black-box modeling in system identification: a unified overview, *Automatica*, 31 (12), 1691–1724.

Stone, M. 1974. Cross-Validatory Choice and Assessment of Statistical Predictions, *Journal of the Royal Statistical Society Series B Methodology*, 36 (2), 111–147.

Toukourou, M., Johannet, A., Dreyfus, G. and Ayral, P.-A. 2011. Rainfall-runoff modeling of flash floods in the absence of rainfall forecasts: the case of "Cévenol flash floods," *Applied Intelligence*, 35 (2), 178–189.

Tramblay, Y., Bouvier, C., Martin, C., Didon-Lescot, J.-F., Todorovik, D. and Domergue, J.-M. 2010. Assessment of initial soil moisture conditions for event-based rainfall–runoff modelling, *Journal of Hydrology*, 387 (3–4), 176–187.

Recibido: septiembre 2016
Revisado: diciembre 2016
Aceptado: enero 2017
Publicado: septiembre 2018