



HAL
open science

Vers une meilleure prise en compte des comportements multi-échelles des hydrosystèmes complexes par les modèles à réseaux de neurones : application aux crues éclair

Bob Saint Fleur

► To cite this version:

Bob Saint Fleur. Vers une meilleure prise en compte des comportements multi-échelles des hydrosystèmes complexes par les modèles à réseaux de neurones : application aux crues éclair. Ingénierie de l'environnement. IMT - MINES ALES - IMT - Mines Alès Ecole Mines - Télécom, 2021. Français. NNT : 2021EMAL0015 . tel-04055304

HAL Id: tel-04055304

<https://theses.hal.science/tel-04055304v1>

Submitted on 2 Apr 2023

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

**THÈSE POUR OBTENIR LE GRADE DE DOCTEUR
DE L'INSTITUT MINES-TELECOM (IMT) –
ÉCOLE NATIONALE SUPÉRIEURE DES MINES D'ALÈS (IMT MINES ALÈS)**

En Sciences de la terre et de l'eau

**École doctorale GAIA – Biodiversité, Agriculture, Alimentation, Environnement, Terre, Eau
Portée par l'Université de Montpellier**

Unité de recherche : Hydrosociences Montpellier

***« Vers une meilleure prise en compte des
comportements multi-échelles des hydrosystèmes
complexes par les modèles à réseaux de neurones :
application aux crues éclair »***

Présentée par Bob E. SAINT FLEUR

Le 16 décembre 2021

Sous la direction de Madame Anne JOHANNET et de Monsieur Séverin PISTRE

Encadrée par Monsieur Guillaume ARTIGUE

Devant le jury composé de

Monsieur Didier GRAILLOT, Directeur de recherche, École des Mines de Saint-Etienne

Rapporteur

Monsieur Didier JOSSELIN, Directeur de recherche, CNRS-Laboratoire UMR ESPACE

Rapporteur

Madame Danièle VALDÈS-LAO, Maîtresse de conférences, UPMC

Examinatrice

Monsieur Roger MOUSSA, Directeur de recherche, INRAE

Examineur

Monsieur Guillaume ARTIGUE, Maître assistant, IMT Mines Alès

Encadrant

Madame Anne JOHANNET, Professeure, IMT Mines Alès

Directrice de thèse

Monsieur Séverin PISTRE, Professeur, HSM – Université de Montpellier

Co-directeur, invité



Remerciements

Je tiens à remercier en premier lieu Madame Anne Johannet et Monsieur Séverin Pistre, la directrice et le co-directeur de cette thèse, d'abord pour avoir eu l'idée de la proposer, ensuite pour l'implication, la rigueur, la pédagogie et du sérieux dont ils ont fait preuve durant ces trois années, sans quoi cette expérience n'aurait pas eu lieu. Je les remercie également d'avoir tenu le même rôle durant mes six mois de stages qui ont précédé la thèse.

Je tiens également à remercier Monsieur Guillaume Artigue, l'encadrant de cette thèse, pour son accompagnement assidu, son implication, sa rigueur et son engagement, investi tout au long de ces trois années qui ont vu se dérouler cette expérience. Cette dernière n'aurait pas non plus son sens sans ces efforts fournis.

Je souhaite exprimer toute ma gratitude à Messieurs Didier Graillot et Didier Josselin d'avoir accepté d'être rapporteurs de ces travaux de thèse. C'est également avec cette même gratitude que je remercie Madame Danièle Valdès et Monsieur Roger Moussa d'avoir examiné ce travail, lui qui présidait aussi mes deux comités de thèse. De même que pour Monsieur Roger Moussa, je tiens à remercier Messieurs Pierre Roussel et Nicolas Massei d'avoir accepté de participer à ces comités de thèse, leur patience, leur implication, leur pédagogie et leurs échanges sur les réseaux de neurones m'ont été d'une très grande utilité dans ce travail.

Je remercie le SCHAPI (Service Central d'Hydrométéorologie et d'Appui à la Prévision des Inondations), les Services de Prévision des Crues du Grand Delta et Météo France d'avoir rendu disponibles et accessibles les données hydrométéorologiques sur lesquelles se sont basées les expériences développées dans cette thèse. Ce remerciement va également de manière spéciale à Monsieur Brunot Janet pour son soutien collaboratif du côté du SCHAPI.

Je remercie Monsieur Dominique Bertin et l'entreprise Geonosis pour ses efforts sur le développement et la mise à notre disposition de l'outil RnfPro sur lequel nos différentes expériences de modélisation hydrologique sont réalisées.

Je remercie les cadres et responsables des laboratoires du centre CREER, de l'équipe HYTAKE, de HydroSciences Montpellier et de la Maison des Sciences de l'eau à Montpellier, d'IMT Mines Alès, en particulier Monsieur Marc Vinches, Monsieur Eric Servat, Madame Valérie Borrell Estupina, Monsieur Christian Pratt, Monsieur Jean-Pierre Bricquet, Madame Catherine Balieu, Monsieur Pierre Genthon, Monsieur Pierre-Alain Ayrat et Madame Sandra Soares-Frazao de l'université catholique de Louvain d'avoir contribué d'une manière ou d'une autre à rendre possible mon admission au master HYDRE à l'université de Montpellier, puis la réalisation de ma thèse à IMT Mines Alès.

D'une manière plus privée, je tiens à exprimer ma gratitude et ma reconnaissance envers Dieu de m'avoir permis de connaître la vie sur terre, ma chère maman Madame Louise Désarmes, mon défunt père Monsieur Emile Saint Fleur, mes deux frères et sœur Poquelin, Berwick et Minerve Saint Fleur, ma fille Myriam Emeline Saint Fleur et sa mère Lumanie Noël pour leur soutien de toutes sortes. Je remercie du même coup Madame Olena Didenko pour sa proximité et ses divers soutiens à ma personne durant une bonne partie de la période de ma thèse.

Mes remerciements amicaux vont également aux différents cadres du laboratoire, en particulier les dames Catherine Balieu, Marine Bertrand, Aurelie Hinsberger, Murielle Avezac, Sandrine Bayle, André Sauvetre, Noemie Fayol, Rosario Spinelli, Sylvie Spinelli, Veronica Artigue etc ... pour les différents moments de partage et de convivialité qu'on a vécus ensemble durant ces trois années de thèse. Je remercie du même coup mes proches amis et/ou ex-collègues doctorants Cagri Alperen, Oumar El-

Farouk, Philippe Lionel Ebengue, Nicolas Akil, Maria Molina, Salma Sadkou, Ghita Serrhini-Naji, Antoine Fricard, Fabienne-Emilie Errero, Emilie Marchand, Sébastien Pinel ainsi que les collègues de l'ATHEMA pour avoir contribué d'une manière ou d'une autre à rendre mon séjour aussi intéressant qu'il y était.

Résumé

Les crues éclair font partie des risques naturels les plus dévastateurs dans le monde. Selon Jamali *et al.* (2020), elles sont responsables de près de 84 % des décès causés par les désastres naturels. Si l'inondation par une crue classique peut être anticipée avec un certain délai, une crue éclair est bien plus rapide et localisée, et donc bien plus difficile à prévoir. Tel est le cas des régions méditerranéennes en France. Face à cette problématique, les institutions chargées de la prévision des crues et des inondations ont besoin d'informations de qualité et de modèles performants afin d'optimiser leurs réponses. Du fait que les pluies génératrices de ces crues éclair sont très hétérogènes aussi bien dans le temps que dans l'espace, en sus du caractère fondamentalement non-linéaire de la relation pluie-débit, leur prévision reste un défi très sérieux. Depuis maintenant trois décennies, les réseaux de neurones ont prouvé leur efficacité pour résoudre des problèmes complexes et non linéaires, en particulier les relations pluie-débits dans diverses situations hydrologiques. Au sein de ces types de modèles, le *Deep Learning* en tant qu'une méthode d'apprentissage qui s'applique principalement aux réseaux de neurones profonds, s'est montré particulièrement performant dans de nombreuses disciplines. Cependant, du fait de leur caractère de boîtes noires, qui nous paraît plutôt un avantage compte tenu du manque de connaissance sur certains processus hydrologiques, l'intérêt de leur application est parfois mis en doute.

Pour cette raison ce travail a appliqué des réseaux de neurones profonds à la prévision des crues éclair avec trois principaux objectifs : le premier objectif vise l'interprétation des paramètres des couches profondes des trois types de perceptrons généralement utilisés en hydrologie : statique, dirigé, récurrent. Pour ce faire ce travail a suivi deux étapes : (i) extraire et traiter les valeurs des paramètres des modèles optimisés en utilisant la méthode « *Knowledge eXtraction (KnoX)* » proposée par Kong A Siou *et al.* (2013) ; (ii) interpréter ces paramètres grâce à une analyse comparative de ces informations avec les données caractérisant certains processus hydrologiques du bassin versant ; cette partie a été publiée dans les actes de la conférence internationale ITISE 2018, pour le volet prévision, et dans un chapitre de livre de la série LNCS (Lecture Notes in Computer Sciences) pour l'interprétation des paramètres. Le second objectif est de prendre en compte la spatialisation des précipitations et les effets d'échelle associés grâce à une modélisation profonde par bassins emboîtés. Ceci a permis d'allonger l'échéance de prévision qui avait été réalisée par Artigue (2012) de 2 à 3 heures, grâce à l'introduction d'information physiques dans le modèle ; ce travail a été soumis au journal NHSS. Le troisième objectif, en suivant une approche duale du premier objectif, propose d'injecter de l'information dans les paramètres des couches profondes des modèles afin d'améliorer leur performance en prévision. Si ce dernier objectif n'a pas été complètement atteint on note cependant qu'il apporte des bénéfices pour les réseaux récurrents avec une amélioration des résultats de prévision significative. Ce travail a mis, une fois de plus, en évidence la qualité de la méthode de sélection de modèles, qui permet non seulement de réaliser des prévisions performantes sur les événements les plus intenses de la base de données, mais aussi d'aboutir à une architecture profonde parcimonieuse permettant, grâce à la règle d'apprentissage de levenberg-Marquardt d'apprendre les couches profondes sans observer d'atténuation du gradient.

Abstract

Flash floods are among the most devastating natural hazards in the world. According to Jamali et al. (2020), they are responsible for nearly 84% of deaths from natural disasters. While flooding from a conventional flood can be anticipated with some lead time, a flash flood is much more rapid and localized, and therefore much more difficult to predict. This is the case in the Mediterranean regions of France. Faced with this problem, the institutions in charge of flood forecasting need quality information and efficient models in order to optimize their responses. Because the rainfall that generates these flash floods is very heterogeneous both in time and space, in addition to the fundamentally non-linear nature of the rainfall-flow relationship, forecasting them remains a very serious challenge. For three decades now, neural networks have proven their efficiency in solving complex and non-linear problems, in particular rainfall-flow relationships in various hydrological situations. Within these types of models, Deep Learning as a learning method that is mainly applied to deep neural networks has proven to be particularly successful in many disciplines. However, because of their black box character, which seems to us rather an advantage considering the lack of knowledge on some hydrological processes, the interest of their application is sometimes questioned.

For this reason, this work has applied deep neural networks to flash flood forecasting with three main objectives: the first objective aims at interpreting the parameters of the deep layers of the three types of perceptrons generally used in hydrology: static, directed, recurrent. To do so, this work followed two steps: (i) extracting the parameter values of the optimized models using the "Knowledge eXtraction (KnoX)" method proposed by Kong A Siou et al. (2013); (ii) interpreting these parameters through a comparative analysis of this information with data characterizing some hydrological processes of the watershed; this part has been published in the proceedings of the ITISE 2018 international conference, for the forecasting component, and in a book chapter of the LNCS (Lecture Notes in Computer Sciences) Springer series for the interpretation of the parameters. The second objective is to take into account the spatialization of precipitation and the associated scale effects through deep nested basin modelling. This allowed to extend the forecast lead time that had been achieved by Artigue (2012) from 2 to 3 hours, thanks to the introduction of physical information in the model; this work has been submitted to the NHSS journal. The third objective, following a dual approach of the first objective, proposes to inject information into the deep layer parameters of the models in order to improve their forecasting performance. If this last objective has not been completely achieved, we note however that it brings benefits for recurrent networks with a significant improvement of forecasting results.

This work has highlighted, once again, the quality of the model selection method, which allows not only to make efficient forecasts on the most intense events of the database, but also to achieve a parsimonious deep architecture allowing, thanks to the Levenberg-Marquardt training rule, to train the deep layers without observing any vanishing gradient.

Sommaire

INTRODUCTION	1
CHAPITRE I : PROBLEMATIQUE DES CRUES ECLAIR ET QUESTIONS SCIENTIFIQUES	4
1 LA PROBLEMATIQUE DES CRUES ECLAIR	4
1.1 ENJEUX SOCIETAUX DES CRUES ECLAIR	5
1.2 GESTION INSTITUTIONNELLE	6
1.3 DESCRIPTION DES CRUES ECLAIR	7
1.4 DONNEES CARACTERISTIQUES	11
2 QUESTIONS SCIENTIFIQUES	15
3 CONCLUSION	15
CHAPITRE II : ETAT DE L'ART/ MATERIELS ET METHODES	17
1 ETAT DE L'ART SUR LA MODELISATION DES CRUES ECLAIR	17
1.1 MODELISATION HYDROLOGIQUE : SIMULATION ET/OU PREVISION	17
1.2 GRANDES ETAPES DE MISE EN ŒUVRE D'UN MODELE	21
1.3 CRITERES D'EVALUATION DES MODELES	22
1.4 À PROPOS DE LA MODELISATION HYDROLOGIQUE SUR LES CRUES ECLAIR	25
1.5 CONCLUSION	26
2 LES RESEAUX DE NEURONES ET LES MODELES PROFONDS POUR LA MODELISATION DES PROCESSUS DYNAMIQUES NON-LINEAIRES	27
2.1 MODELISATION BOITE NOIRE/NON PARAMETRIQUE	27
2.2 LES RESEAUX DE NEURONES	27
2.3 APPRENTISSAGE DE MODELES DYNAMIQUES PAR RESEAUX DE NEURONES	37
2.4 METHODES SPECIFIQUES DES MODELES PROFONDS (<i>DEEP LEARNING</i>)	51
2.5 APPLICATION DES RN A LA PREVISION HYDROLOGIQUE	52
2.6 EXTRACTION DES PARAMETRES DES MODELES PROFONDS AVEC LA METHODE <i>KNOWLEDGE EXTRACTION (KNOX)</i>	55
2.7 INTERPRETATION PHYSIQUE DES MODELES PROFONDS	55
CONCLUSION DU CHAPITRE SUR LES RESEAUX DE NEURONES ET LES MODELES PROFONDS	56
CHAPITRE III : SITE D'EXPERIMENTATION	58
1 BASSIN VERSANT DU GARDON A MIALET	58
1.1 CARACTERISTIQUES TOPOGRAPHIQUES ET HYDROGRAPHIQUES	58
1.2 CARACTERISTIQUES GEOLOGIQUES	59
1.3 OCCUPATION DES SOLS	60
2 DONNEES DISPONIBLES	60
2.1 PRESENTATION DE LA BASE DE DONNEES	61
2.2 TRAITEMENT DES DONNEES	61
3 CONCLUSION	69
CHAPITRE IV : TOWARDS A BETTER CONSIDERATION OF RAINFALL AND HYDROLOGICAL SPATIAL SCALES BY A DEEP NEURAL NETWORK MODEL TO IMPROVE FLASH-FLOODS FORECASTING. CASE STUDY ON THE GARDON BASIN, FRANCE	71
INTRODUCTION DU CHAPITRE	71
1. INTRODUCTION	1
2. MATERIAL AND METHODS	3

2.1 STUDY AREA-----	3
2.2 DATABASE-----	3
2.3 ANN PRINCIPLES AND DESIGN -----	3
3. RESULTS -----	8
3.1 USE OF THE DATABASE -----	8
3.2 MODEL SELECTION -----	9
3.3 FORECASTING RESULTS -----	9
4. DISCUSSION -----	9
5. CONCLUSIONS -----	10
REFERENCES-----	11
POINTS DE VALORISATION SCIENTIFIQUE DE LA PUBLICATION -----	84
CHAPITRE V : KNOWLEDGE EXTRACTION (KNOX) IN DEEP LEARNING: APPLICATION TO THE GARDON DE MIALET FLASH FLOODS MODELLING -----	85
INTRODUCTION DU CHAPITRE-----	85
1 INTRODUCTION-----	1
2 MATERIALS AND METHODS -----	2
2.1 STUDY AREA: LOCATION AND GENERAL DESCRIPTION-----	2
2.2 DATABASE -----	2
2.3 ARTIFICIAL NEURAL NETWORK-----	2
3 RESULTS -----	5
4 INTERPRETATION -----	8
5 CONCLUSION -----	8
6 AKNOWLEDGEMENT-----	8
7 REFERENCES -----	8
POINTS DE VALORISATION DE LA PUBLICATION -----	94
CHAPITRE VI : DEEP MULTILAYER PERCEPTRON FOR KNOWLEDGE EXTRACTION: UNDERSTANDING THE GARDON DE MIALET FLASH FLOODS MODELLING-----	95
INTRODUCTION DE LA PUBLICATION-----	95
1 INTRODUCTION-----	1
2 MATERIAL AND METHODS -----	1
2.1 STUDY AREA: LOCATION AND GENERAL DESCRIPTION-----	1
2.2 DATABASE -----	2
2.3 ARTIFICIAL NEURAL NETWORKS -----	2
2.4 EXTRACTING INFORMATION: KNOX METHOD -----	3
3 RESULTS -----	4
3.1 CHOICE OF VARIABLES-----	4
3.2 MODEL SELECTION -----	4
3.3 DISCHARGE ESTIMATION -----	5
3.4 CONTRIBUTIONS OF INPUT VARIABLES-----	5
3.5 RESULTS: CONTRIBUTIONS AS A FUNCTION OF TIME WINDOWS-----	6
3.6 RESULTS: EFFECTS OF THE BIAS -----	7

4	DISCUSSION	8
4.1	SELECTING A MODEL TYPE FOR PHYSICAL KNOWLEDGE EXTRACTION	8
4.2	RESPONSE TIME AND CONTRIBUTIONS	9
4.3	BIAS INPUT IMPORTANCE	9
5	CONCLUSIONS AND PERSPECTIVES	9
6	ACKNOWLEDGEMENTS	9
7	REFERENCES	10
	POINTS DE VALORISATION DE L'ARTICLE	105
	CHAPITRE VII : INITIALISATION DES COUCHES PROFONDES AVEC LA CORRELATION CROISEE	107
1	INTRODUCTION	107
2	APPROCHE EXPERIMENTALE	107
2.1	ARCHITECTURE GENERIQUE	107
2.2	SELECTION DES HYPERPARAMETRES DE L'ARCHITECTURE	108
2.3	CHOIX DES METHODES DE REGULARISATION ET DE LEURS HYPERPARAMETRES	111
3	RESULTATS ET INTERPRETATION	112
3.1	COMPLEXITE DES MODELES	112
3.2	TYPE D'INITIALISATION ET METHODES DE REGULARISATION	114
3.4	VISUALISATION DES PREVISIONS SUR L'EVENEMENT DE TEST	117
4	DISCUSSION	123
	CONCLUSION DU CHAPITRE ET PERSPECTIVES	123
	CONCLUSION GENERALE ET PERSPECTIVES	125
	RÉFÉRENCES BIBLIOGRAPHIQUE	128
	ANNEXES	144

Liste des figures

Figure 1 : Occurrence climatique observée de pluies quotidiennes d'au moins 200 mm en 24 heures sur la période de 1971-2020. Src : METEO-France, édition du 29/03/2021	4
Figure 2. : Structure spatiale moyenne des champs de pluie au sol de l'épisode du 8-9 septembre 2002 dans le Gard (un système méso-échelle avec un cumul maximal de 700 mm, des intensités supérieures à 10 mm/h), radar Bollène. Traits du centre aux contours : cumul sur 5 – 15 – 30 – 60 min [tirée de Boudevillain et al., 2009]	12
Figure 3: Neurone formel	28
Figure 4: Réseau de neurones multicouche avec une couche cachée et une couche de sortie d'un seul neurone	30
Figure 5: Modèle peu profond et modèle profond.	31
Figure 6: Perceptron à deux couches avec une seule couche cachée	31
Figure 7: Modèle statique	33
Figure 8: Modèle postulé récurrent.	34
Figure 9: Schéma du passage du processus physique au prédicteur optimal (avec bu , bs et bp : les différentes sources des bruits)	34
Figure 10 : Modèle postulé (a) et prédicteur associé (b) à l'hypothèse bruit de sortie (Johannet (2011))	35
Figure 11: Modèle récurrent	36
Figure 12 : Modèle postulé (a) et prédicteur associé (b) à l'hypothèse bruit d'état (Johannet (2011))	36
Figure 13: Modèle dirigé	37
Figure 14: Schéma de l'asservissement de la fonction de coût durant l'apprentissage. La contre-réaction apparaît sous la forme du retour de $J(k)$ vers l'opérateur de différentiation calculant l'erreur- $J(k)$. J_c est la valeur de consigne de la fonction de coût c'est-à-dire 0. $J(k)$ est la valeur de J pour l'exemple k ; cette valeur est renvoyée vers l'entrée par la rétroaction, puis soustraite à la valeur de consigne pour délivrer l'erreur. Les deux blocs qui suivent ont pour fonction de calculer le gradient de l'erreur, puis le nouveau jeu des paramètres et la nouvelle valeur de la fonction de coût.	43
Figure 15: Réseau simple de type « chain rule ». On note x l'entrée, c_i le paramètre arrivant au neurone i et y la sortie du réseau. Les neurones 1 et 2 sont non-linéaires avec une sigmoïde et le neurone de sortie est linéaire ($y = v$).	43
Figure 16: Illustration du dilemme biais variance	45
Figure 17: Illustration de l'utilisation de l'arrêt précoce). Gauche (sans l'arrêt-précoce) ; Droite (avec l'arrêt précoce)	48
Figure 18: Localisation du bassin versant du Gardon de Mialet (à reproduire soi-même)	58
Figure 19: Caractéristiques topographiques et hydrographiques du bassin versant de Gardon de Mialet	59
Figure 20: Carte géologique simplifiée du bassin versant de Gardon de Mialet	59
Figure 21: Occupation de l'espace du bassin versant de Gardon de Mialet. Src : Corine Land_cover 2016	60
Figure 22: Courbe de tarage de la station hydrométrique de Mialet.	61
Figure 23: Durée des épisodes pluvieux retenus	63
Figure 24: Cumul des pluies des différents événements sur les trois pluviomètres sous formes de boîtes à moustache incluant la médiane ; BDC est Barre-des-Cévennes, SRDT est Saint-Roman-de-Tousque	63
Figure 25: Intensités maximales des différents épisodes pluvieux sur les trois pluviomètres sous forme de boîte à moustache; BDC est Barre-des-Cévennes, SRDT est Saint-Roman-de-Tousque	64

<i>Figure 26: Débits de pointe (a) et distribution des débits spécifiques (b) des différents épisodes pluvieux</i>	65
<i>Figure 27: Hydrogrammes de quelques épisodes de la base de données</i>	65
<i>Figure 28: Test de double cumul entre les pluviomètres ; BDC est Barre-des-Cévennes, SRDT est Saint-Roman-de-Tousque sur les épisodes sélectionnés (1992-2019)</i>	66
<i>Figure 29: Répartition sous forme de nuages de points des observations mi- horaires des pluies entre les trois pluviomètres, incluant leur droite de régression et leur coefficient de corrélation linéaire simple ; BDC est Barre-des-Cévennes, SRDT est Saint-Roman-de-Tousque (1992-2019)</i>	66
<i>Figure 30: Corrélations croisées entre les stations pluviométriques ; BDC pour Barre-des-Cévennes, SRDT pour Saint-Roman-de-Tousque ; P.moy. pour la pluie moyenne arithmétique</i>	67
<i>Figure 31: Corrélations croisées pluie-débits entre les différentes stations au niveau du bassin versant de Gardon de Mialet. Les valeurs médianes (noir) et moyennes (orange) sont calculées sur les 58 événements. BDC est Barre-des-Cévennes, SRDT est Saint-Roman-de-Tousque.</i>	68
<i>Figure 32: Map and location of the Gardon de Mialet basin and its measurement stations.</i>	16
<i>Figure 33 : Architecture générique du perceptron multicouche utilisé</i>	107
<i>Figure 34: Repère de sélection de la profondeur historique par la corrélation croisée</i>	109
<i>Figure 35 : Comparaisons des valeurs des paramètres en fonction de la profondeur de la fenêtre temporelle d'entrée de 20 pas de temps. Graphe de gauche : valeurs brutes ; graphe de droite : valeurs absolues.</i>	110
<i>Figure 36 : Score de la validation croisée (persistance) en fonction de l'hyperparamètre de la modération des poids</i>	112
<i>Figure 37 : Scores de validation croisée en fonction du nombre de neurones de la couche ξ pour le modèle dirigé (a), et le récurrent (b).</i>	113
<i>Figure 38 : Prévion des crues en fonction des différents types d'initialisation avec la combinaison de l'arrêt précoce et la modération des poids</i>	117
<i>Figure 39 : Prévion des crues en fonction des différents types d'initialisation régulée par l'arrêt précoce seul</i>	119
<i>Figure 40 : Prévion des crues en fonction des différents types d'initialisation régulée par la modération des poids seule</i>	121
<i>Figure 41 : Comparaison de la variabilité des prévisions des crues en fonction du type d'initialisation et de la méthode de régularisation. Le débit observé (vert, trait plein) ; écart-type : l'initialisation aléatoire (noir, trait plein) ; initialisation et fixation par corrélation croisée totale (bleu, trait plein et continu) ; initialisation et fixation par corrélation croisée partielle (orange, trait plein et continu)</i>	122

Liste des tableaux

Tableau 1 : Principales caractéristiques de la base de données (BDC est Barre-des-Cévennes, SRDT est Saint-Roman-de-Tousque) -----	63
Tableau 2 : Descriptif des relations entre les variables pluviométriques et hydrométriques ; BDC est Barre-des-Cévennes, SRDT est Saint-Roman-de-Tousque -----	69
Tableau 3 : Dimensions des profondeurs des fenêtres temporelles des variables d'entrée-----	108
Tableau 4 : dimensionnement des couches profondes au niveau des modèles dirigés -----	113
Tableau 5 : dimensionnement des couches profondes au niveau des modèles récurrents -----	114
Tableau 6 : Performance en validation croisée des modèles avec l'utilisation conjointe des deux méthodes de régularisation (arrêt précoce et modération des poids) -----	115
Tableau 7 : Variation de performance ApMp par rapport à l'arrêt précoce seul, APs -----	116
Tableau 8 : Gain ou déficit de performance de la combinaison à la modération des poids seule-----	116
Tableau 9 : Performance en test des modèles obtenus avec l'utilisation conjointe des deux méthodes de régularisation (arrêt précoce et modération des poids) -----	118
Tableau 10 : Gain ou déficit de performance en test de la combinaison par rapport à l'arrêt précoce seul -----	120
Tableau 11 : Gain ou déficit de performance en test de la combinaison par rapport à la modération des poids seul-----	120

Liste des annexes

Annexe 1: Résumé des 58 évènements -----	145
Annexe 2: Hydrogrammes des épisodes de la validation croisée de la période 1992-1999-----	146
Annexe 3: Hydrogrammes des épisodes de la validation croisée de la période 2000-2019-----	147
Annexe 4: Hydrogrammes des épisodes hors validation croisée de la période 1992-1999 -----	148
Annexe 5: Hydrogrammes des épisodes hors validation croisée de la période 2000-2009 -----	149
Annexe 6: Hydrogrammes des épisodes hors validation croisée de la période 2010-2019 -----	150
Annexe 7: Les épisodes de Test (a) et d'Arrêt (b)-----	151

Liste des publications

- [page 85]. Bob Saint Fleur, Guillaume Artigue, Anne Johannet, Séverin Pistre. Knowledge Extraction (KnoX) in Deep Learning: Application to the Gardon de Mialet Flash Floods Modelling. *ITISE 2019 - International Conference on Time Series and Forecasting*, Sep 2019, Granada, Spain. ([hal-02462501](https://hal.archives-ouvertes.fr/hal-02462501))
- [page 95]. Saint Fleur B.E., Artigue G., Johannet A., Pistre S. (2020) Deep Multilayer Perceptron for Knowledge Extraction: Understanding the *Gardon de Mialet* Flash Floods Modeling. In: Valenzuela O., Rojas F., Herrera L.J., Pomares H., Rojas I. (eds) *Theory and Applications of Time Series Analysis*. ITISE 2019. Contributions to Statistics. Springer, Cham. https://doi.org/10.1007/978-3-030-56219-9_22
- [page 71]. *Towards a better consideration of rainfall and hydrological spatial scales by a deep neural network model to improve flash-floods forecasting. Case study on the Gardon basin, France. Soumise au journal NHESS lors du dépôt définitif du manuscrit.*

Liste des abréviations

Abréviation	Plein texte
BDC	Barre-des-Cévennes
Mia	Mialet
SRDT	Saint-Roman-de-Tousque
P _{SRDT}	Le pluviomètre à Saint-Roman-de-Tousque
Artificial neural network (ANN) (Réseau de neurones artificiels)	
CNN	<i>Convolutional Neural Network</i>
GAN	<i>Generative Adversarial Network</i>
MLP ou PMC	<i>Multilayer Perceptron</i> ou Perceptron Multicouche
RNN	<i>Recurrent Neural Network</i>
Notations personnalisées et variables	
iCP	Initialisation avec la corrélation croisée partielle
iCT	Initialisation avec la corrélation croisée totale
iRnd	Initialisation aléatoire (des paramètres du modèle)
fCP	Fixation des paramètres avec la corrélation croisée partielle
fCT	Fixation des paramètres avec la corrélation croisée totale
Xcor.	Corrélation croisée
Mwd	<i>Memory Windows of the cross correlation</i> ($R_{xy} \geq 0.2$)
Fwd	<i>Full window on the cross correlogram</i> ($R_{xy} \geq 0$)
APs	Arrêt précoce seul (utilisé tout seul)
Mps	Modération des poids (utilisée seule)
ApMp	Utilisation conjointe de l'arrêt précoce et de la modération des poids
C _{ji}	Paramètres liant le neurone i au neurone j vers la sortie
R _{xy} (parfois C _{xy})	Corrélation croisée entre les variables x et y
Q ^p ou Y ^p	Débit observé au niveau du processus
Q ^s ou Y ^s ou \hat{y}	Débit simulé par le modèle optimisé ou postulé
k	Compteur discrétisant du temps et des éléments des vecteurs
n _r ou w	Dimension d'un fenêtrage glissant sur un vecteur à l'entrée d'un modèle
T _r ou T.Rep	Temps de réponse
Q _{obs}	Débit observé utilisé au niveau des modèles dirigés
Q _{rec}	Débit simulé aux instants écoulés réutilisé en entrée par bouclage au niveau des modèles récurrents (appelé également ordre du réseau)
b _p - b _s - b _u	Bruits liés (resp.) au processus, à la simulation ou aux variables d'entrées (pp.34)
Critères de qualité et notations courantes	
C _p	Critère de persistance en test
SPPD ou S _{PPD}	<i>Synchronous percentage of the peak discharge</i>
P _d ou P _D	<i>Peak delay</i>
S _p	Score persistance sur la validation croisée
S _n	Nash sur la validation croisée
Fig.	Figure (ex. Fig. 5 --> Figure #5)
Éq.	Equation (ex. Éq. 15 --> l'équation N° 15)
m.a.s.l	<i>Meter above the sea level</i>
Institutions	
HSM	<i>Hydrosciences Montpellier</i>
LGEI	<i>Laboratoire de Génie de l'Environnement Industriel (Alès, France). Devenu CREER (Centre de Recherche et d'Enseignement en Environnement et en Risques) en 2020</i>
SCHAPI	<i>Service Central d'Hydrométéorologie et d'Appui à la Prévision des Inondations</i>
SPC	<i>Services de Prévisions des Crues</i>

INTRODUCTION

Motivations

Les crues éclair et enjeux

Les crues éclair constituent l'une des menaces naturelles les plus dangereuses dans le monde. Elles sont souvent définies comme des crues soudaines qui se produisent sur des bassins versants de petite taille. Elles sont caractérisées par des débits de pointes élevés en des temps de montée très courts, et elles sont en général assez brèves (Gaume et Borga, 2008). Les régions méditerranéennes, en particulier les régions du nord-ouest de la Méditerranée, dont le sud de la France, sont souvent confrontées à ces phénomènes. En effet, elles ont été sérieusement touchées ces dernières décennies (Gaume *et al.*, 2009; Llasat *et al.*, 2010, 2013; Giuntoli *et al.*, 2012). Ces menaces hydrométéorologiques, quand elles se manifestent et se croisent aux enjeux socio-démographiques, touchent aussi bien la vie humaine que les biens matériels. A titre d'exemple : les événements de Nîmes en 1988, de Vaison-la-Romaine en 1992, de l'Aude en 1999, du Gard en 2002, du Var en 2010, les Alpes-Maritimes en 2020 ont causé presque chacun plus d'une vingtaine de morts avec des dégâts estimés parfois à plus d'un milliard d'euros. A travers tous ces événements malheureux dont la liste pourrait être allongée, les principaux éléments qui reviennent sont la rapidité et l'amplitude des crues qui ne laissent que peu de temps de réaction aux services concernés. Ces événements hydrologiques particulièrement dangereux sont en général les produits d'épisodes pluvieux intenses, souvent très localisés et difficiles à anticiper avec pertinence.

Les besoins et les limites actuelles

Au regard des enjeux humains et financiers qui sont ainsi exposés aux menaces de ces crues éclair, les autorités publiques et scientifiques doivent trouver des moyens efficaces et efficients afin de limiter les impacts de ces risques sur la société. D'un point de vue pratique, cela passe par la disposition d'informations sur la potentielle manifestation de la menace ; les services météorologiques et hydrologiques sont les principaux concernés à ce niveau. Ensuite, la disposition au préalable d'un plan de gestion de crises du côté des pouvoirs publics en cas de manifestation de cette menace. En France, à côté de Météo France et des entités de recherche, le SCHAPI (Service Central d'Hydrométéorologie et d'Appui à la Prévision des Inondations) et le SPC (Service de Prévision de Crues) sont directement impliqués dans la prévision de ce genre de menace. En effet, ils sont chargés de produire des prévisions de crues et de les communiquer au grand public sous formes cartes de vigilances accessibles sur le site Vigicrues (<http://www.vigicrues.gouv.fr/>). Cet exercice de prévision reste une tâche scientifiquement très délicate du fait que, hormis le caractère fondamentalement non-linéaire de la relation pluie-débit, les pluies qui sont à l'origine de ces crues éclair sont difficilement prévisibles avec les moyens dont on dispose actuellement. De plus, non seulement les processus physiques gouvernant le comportement du système hydrologique sont en pratique très mal connus, les données relatives à leurs observations sont également très bruitées car difficiles à mesurer correctement (Lumbroso et Gaume, 2012; Kazuhiko *et al.*, 2021). En ce sens, malgré les nombreux efforts déployés à travers diverses approches de modélisation hydrologique, une caractérisation pertinente de ces phénomènes-là demeure une préoccupation sociétale et scientifique majeure.

Apports potentiel des réseaux de neurones et du *Deep Learning*

Au cours de ces dernières décennies, les réseaux de neurones, en tant que modèles à apprentissage statistique, se sont montrés très efficaces pour la modalisation hydrologique. Apparentés à des modèles du type boîte noire, ils présentent l'avantage de pouvoir reproduire le fonctionnement d'un système naturel sans connaissance a priori sur les processus physiques sous-jacents. En effet,

appliqués à la prévision des crues éclair en absence de prévision de pluies sur des bassins méditerranéens, ils se sont révélés très performants. Ils sont donc proposés comme une approche complémentaire voire une alternative à la caractérisation de la problématique de ces phénomènes naturels. Cependant, l'aspect « boîte noire », dépourvue de paramètre que l'on puisse comprendre ou interpréter, est perçu comme une limitation, car il empêche de comprendre le modèle et peut parfois induire un manque de confiance. Cependant, ne pas avoir besoin d'explicitier la structure du modèle est un intérêt majeur pour bon nombre de systèmes non conçus par l'intelligence humaine et, pour certains, complètement inconnus. C'est le principal avantage qui fonde les travaux réalisés dans cette thèse, qui applique la modélisation statistique au domaine de l'environnement. Toujours dans une perspective d'amélioration de l'existant, divers travaux ont été menés en vue d'élucider les contenus des modèles dits boîtes noires. Citons entre autres (Johannet *et al.*, 2008, Kong A Siou, 2011 ; Taver, 2014). Certains de ces travaux ont permis de décrire divers aspects de la dynamique hydrogéologique d'un aquifère karstique (cas du Lez, Baget), des systèmes naturels susceptibles d'être assimilés également à des boîtes noires. Les méthodes ainsi développées ont permis d'élucider des réalités qui n'avaient jamais été expliquées jusque-là. C'est entre autres sur cette même lancée que s'appuie la présente étude. Elle permet une meilleure prise en compte des comportements multi-échelles des bassins versants soumis aux crues éclair à travers les modèles à réseaux de neurones et l'approche du *Deep Learning*. Cette dernière, étant une méthode d'apprentissage généralement basée sur des réseaux de neurones profonds, offre la possibilité de modéliser des relations très complexes.

Objectifs de la thèse et plan des contributions apportées

Ce travail consiste dans un premier temps à apprécier la performance du *Deep Learning* appliqué à la prévision des crues éclair. Pour cela, nous appliquons un réseau de neurones profonds aux crues éclair d'un bassin versant cévenol (le bassin versant du Gardon à Mialet) sur lequel on compte allonger les échéances des prévisions préalablement réalisées. Ceci concerne les travaux de Artigue (2012) ayant atteint un horizon de prévision de crue de 2 heures avec un réseau de neurones peu profond sur le Gardon de Mialet à Mialet. Le nouveau modèle profond est postulé de telle sorte qu'il puisse exploiter l'influence de la spatialisation des pluies sur les crues éclair, étant donné que les pluies en question sont généralement très hétérogènes dans le temps et dans l'espace. Pour ce faire, nous avons postulé un modèle qui s'aligne sur la configuration spatiale des pluviomètres installés sur le bassin versant à travers une succession de couches de neurones cachés imbriquées de l'amont vers l'aval. Le modèle sélectionné est testé sur les deux épisodes de crue les plus importants de la base de données disponible de 1992 à 2008, dont l'épisode de septembre 2002 qui a fait plus de 20 victimes dans le Gard. Le *Deep Learning* étant actuellement une approche particulièrement efficace, il est attendu que son application en respectant la configuration spatiale des pluies sur le bassin versant puisse permettre non seulement de disposer d'une prévision de crue à une plus longue échéance, mais que l'on puisse également établir un lien entre sa configuration (la complexité du modèle) et la réalité hydrologique du bassin versant. Par la suite, toujours dans une approche de *Deep Learning*, nous appliquons les trois principaux types de perceptrons couramment utilisés en hydrologie pour la modélisation de ces crues éclair. Le but poursuivi dans cette deuxième vague d'expériences est d'accéder aux informations des paramètres internes de ces modèles neuronaux profonds pour en apprécier l'analogie aux informations connues du processus hydrologique du bassin versant. Cette démarche vise à apporter plus d'éclairage au sujet du caractère de boîte noire de ces modèles non-paramétriques. L'accès aux informations du modèle est réalisé en utilisant la méthode **Knowledge eXtraction** (KnoX) présentée par Kong A Siou *et al.* (2013). Après l'extraction des informations des couches profondes des différents modèles, nous réalisons une comparaison de l'expression des différentes variables explicatives des crues éclair observés sur le bassin versant. Nous poursuivons les analyses de ces informations extraites en les comparant aux données telles que le poids spatial relatif des pluviomètres, la corrélation croisée pluie

débit ainsi que le temps de réponse. Ces expériences constituent les résultats présentés au niveau des chapitres V et VI. Dans le but d'éprouver la pertinence de ces expériences, nous appliquons la démarche inverse qui consiste à introduire des informations du processus physique aux paramètres des couches profondes du modèle neuronal appliqué à ces crues éclair. Cette introduction d'informations se réalise principalement par l'initialisation de ces différents paramètres par données de la corrélation croisée. Entre temps, l'idée de les laisser s'ajuster par l'apprentissage ou de les verrouiller à l'avance s'impose et offre la possibilité d'analyser l'évolution de la performance des différents modèles, qu'ils soient appris ou fixés. Ces dernières expériences sont présentées au niveau du *Chapitre VII* de ce manuscrit.

Contexte expérimental de la thèse

Ces différentes expériences sont menées au niveau du centre CREER d'IMT Mines Alès et au sein de l'Unité Mixte de Recherche HydroSciences Montpellier. Le bassin versant du Gardon à Mialet a été choisi comme site expérimental. Le choix de ce bassin versant repose sur l'existence et sur l'accès à une base de données hydrométéorologiques couvrant près de 30 ans d'observation, ainsi qu'une assez bonne connaissance de ce bassin versant grâce à différents travaux de thèse dont celui de Artigue (2012).

Plan du manuscrit

Dans ce manuscrit, nous commençons par présenter, au travers du premier chapitre la problématique des crues éclair, les conditions favorables à leur manifestation, leurs enjeux sur la société ainsi que les difficultés liées à la disponibilité de leurs données caractéristiques. Dans le deuxième chapitre, nous présentons les principes et quelques avancés scientifiques dans le domaine de la modélisation hydrologique en général, et des crues éclair en particulier. Nous y présenterons également les réseaux de neurones et le *Deep Learning* comme outils et approche de modélisation hydrologique, ainsi que leur principe de mise en œuvre. Au niveau du troisième chapitre, nous présentons le bassin versant du Gardon de Mialet ainsi que les données qui vont être utilisées dans ce travail. Les *Chapitres IV à VII* présentent les principaux résultats des diverses approches en commençant par le modèle profond spatialisé sur les pluies, ensuite l'extraction des données des modèles neuronaux, puis l'interprétation des données extraites, et enfin l'introduction des informations dans un modèle profond appliqué aux crues éclair de ce bassin versant.

1 La problématique des crues éclair

Les crues éclair sont définies comme des crues à apparition soudaine, difficiles à prévoir, avec des temps de montée rapides et des débits spécifiques élevés (AGI, Glossary of Geology, American Geological Institute Washington, DC, 1972 ; IAHS, 1974; Montz et Gruntfest, 2002). Bien que leur occurrence soit souvent liée à des épisodes pluvieux intenses, d'autres évènements tels que la rupture d'un barrage naturel ou artificiel, la fonte soudaine d'un glacier suite à une éruption volcanique ou encore un tsunami, sont autant de conditions favorables à leur apparition. En général, il est question d'épisodes pluvieux intenses qui peuvent être localisés sur à peine un ou deux petits bassins versants, ou parfois plus étendus, donnant lieu à une crue éclair très locale (Marcos *et al.*, 2016). Le caractère soudain de ces crues et l'amplitude des niveaux d'eau qu'elles entraînent les rendent particulièrement dangereuses et de nature à ne laisser que peu de temps de réaction aux populations exposées (Georgakakos, 1986). En général, les conditions favorables aux pluies provoquant les crues éclair sont particulières, voire propres à des régions bien spécifiques dans le monde. Par exemple, le caractère localisé et intense de ces pluies peut dépendre de la topographie locale sous certaines conditions atmosphériques. C'est la raison pour laquelle il n'y a que peu de régions dans le monde où elles sont fréquentes. En France, ce sont plutôt les régions du pourtour méditerranéen, en particulier celles du flanc sud des Cévennes, qui sont les plus exposées, en lien avec la fréquence et l'intensité des épisodes pluvieux qui s'y produisent (Figure 1). Ces caractéristiques météorologiques ont conduit à retenir la région cévenole comme site d'expérimentation dans le cadre de cette étude. Face au risque d'inondation que représentent les crues éclair, menaçant les vies et les biens, leur anticipation ainsi que des mesures de gestion de crise sont nécessaires afin d'en minimiser l'impact.

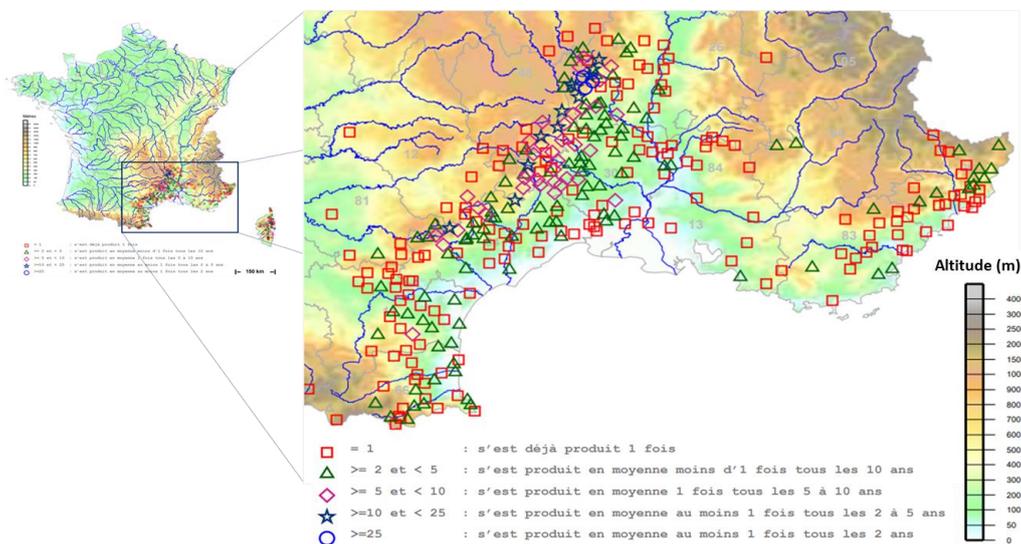


Figure 1 : Occurrence climatique observée de pluies quotidiennes d'au moins 200 mm en 24 heures sur la période de 1971-2020. Src : METEO-France, édition du 29/03/2021

Dans la suite de ce document, nous soulignons ce que ces menaces représentent en termes d'enjeux sociétaux, ainsi que les réponses institutionnelles qui sont apportées de manière générale d'abord, puis en France particulièrement. Ensuite, nous abordons ce qui a été réalisé en termes d'efforts de description de ces crues particulières, leurs données caractéristiques et enfin quelques approches de modélisation tendant vers une prévision opérationnelle.

1.1 Enjeux sociétaux des crues éclair

Les inondations constituent l'un des risques naturels les plus dévastateurs dans le monde. Selon Jamali *et al.* (2020), elles sont responsables de près de 84 % des décès causés par les désastres naturels. Si l'inondation par une crue classique peut être anticipée avec un certain délai, une crue éclair est bien plus rapide et localisée, et donc bien plus difficile à prévoir. A travers le monde, le nombre de victimes, voire leurs recensements, est grandement lié au niveau de développement économique et à l'étendue du territoire du pays concerné. À titre d'exemple, entre 1996 et 2017, 75 000 inondations ont eu lieu aux Etats-Unis (Ahmadalipour et Moradkhani, 2019), faisant 278 morts alors que plus de 28 000 crues éclair ont eu lieu entre 2007-2015 (Gourley *et al.*, 2017). En Chine, entre 2000 et 2018, plus de 16 000 morts par les crues éclair ont été enregistrées, soit 74% des décès causés par les crues (Ministry of Water Resources of China, 2018). Dans le monde, les dégâts matériels accompagnant ces catastrophes ont été estimés en 2013 à plus de 50 milliards de dollars (Wasko et Sharma, 2017).

En Europe, Gaume *et al.* (2009) ont inventorié environ 550 crues éclair majeures sur le pourtour méditerranéen entre 1950 et 2006, dont 236 en France. Par exemple, en Espagne, en 1996, 86 morts ont été constatés dans un camping des Pyrénées espagnoles suite à une crue éclair (IPCC, 2012). Les crues éclair sont donc si dangereuses qu'un seul événement peut produire de très lourdes pertes. Trigo *et al.* (2016) ont inventorié une liste des 20 épisodes pluvieux les plus meurtriers au Portugal, en identifiant celui du 25 au 26 novembre 1967 comme étant le plus meurtrier, avec plus de 500 morts. Le 25 septembre 1962, une crue éclair s'est produite au nord-est de l'Espagne, entraînant plus de 800 morts en moins de 5 heures (Gaume *et al.*, 2009; Marcos *et al.*, 2016). Selon Destro *et al.* (2018), Ngo *et al.* (2018), il faut attribuer une très grande part de la responsabilité des dégâts causés par les crues éclair à leur soudaineté et à leur amplitude. Considérant l'évolution du contexte climatique, ces menaces pourraient s'accroître dans les prochaines décennies. Selon Alfieri *et al.* (2017), sous l'influence combinée du réchauffement climatique, de la croissance démographique et de son influence sur l'occupation des sols, les dégâts matériels annuels pourraient atteindre plus de 100 milliards de dollars d'ici 2050 dans certaines régions de la Chine.

À côté de ce que représente la menace elle-même, un autre facteur non-négligeable est l'appréhension des populations concernées par le phénomène en termes de risque. Diakakis *et al.* (2018) ont réalisé une étude à propos de la perception des populations de la partie est de la Méditerranée, spécialement la Grèce, sujette aux crues éclair, sur les risques d'inondation associés. Les principaux aspects abordés concernaient le risque d'inondation en soi, la mitigation, l'importance accordée aux mesures de protection, la confiance dans les institutions, la vigilance aux annonces de crues et les actions de protections relatives. Les données recueillies révèlent que, selon les expériences des répondants, les risques d'inondation sont classés en troisième position derrière les séismes et les feux de forêt. Une grande majorité des répondants estime que le risque est en train d'augmenter, principalement à cause des facteurs anthropiques. Il est révélé également qu'il peut exister une faible confiance dans les autorités, associée à un faible niveau de connaissance des mesures de protection et de vigilance, aussi bien en termes de préparation que de protections individuelles face aux inondations. D'autres facteurs tels que l'âge et l'expérience des inondations ont été soulignés comme très influents sur la perception des populations. Les personnes âgées sont plus sensibilisées au risque, probablement par l'expérience. Bien que cette étude ait été menée en Grèce, il est possible que ses conclusions soient comparables pour une bonne part du reste du monde. Ceci pourrait permettre de déduire qu'une bonne part des décès causés par les crues peut être liée à la perception des populations au risque que représente les crues. Notons ici qu'une étude menée en 2014 par (Marcos *et al.*, 2016) sur l'évolution des crues en Catalogne entre 1981 et 2010 révèle que, 60% des 110 décès enregistrés sur 19% des 23 épisodes catastrophiques sont survenus lorsque des personnes tentaient de traverser les rues inondées ou des rivières en crues. Ruin *et al.* (2017) rapportent que, durant les crues qui s'abattaient sur le Colorado

en 2013 et la Côte d'Azur en 2015, beaucoup de décès auraient pu être évités si les victimes s'étaient comportées plus raisonnablement. Les auteurs soulignent également que, hormis la possibilité de la sous-estimation des risques, les préoccupations individuelles quotidiennes, que ce soit pour des raisons sociales, telle que le travail ou la récupération des enfants à l'école, sont autant de facteurs modulant le comportement des gens. Dans ce dernier cas, il est évoqué la possible minimisation du risque par certaines institutions publiques, telles que les entreprises ou les écoles. Bien que toutes ces raisons incombent en grande part à la responsabilité des populations, on ne peut pas exclure d'éventuels manques d'efficacité des mesures prises ou proposées par les autorités, en raison ou non d'un manque d'organisation ou d'anticipation (performance des prévisions). Ce sont, quoi qu'il en soit, autant de raisons pour que l'implication sociale et institutionnelle face à ces risques soit le plus large possible.

La France, plus spécialement au niveau du pourtour méditerranéen, et de la région cévenole en particulier, reste le pays qui enregistre le plus de crues éclair en Europe, bien qu'elle ne soit pas celui comptant le plus de victimes (Price *et al.*, 2011). À titre d'exemple, un important épisode méditerranéen s'est produit au pied des Cévennes en septembre 2002, faisant plus de 20 morts et près de 1,2 Milliards d'euros de dégâts (Huet *et al.*, 2003). Durant la dernière décade du mois de novembre 2019, 11 personnes ont péri lors de deux épisodes méditerranéens qui se sont produits dans le Sud-Est de la France. Plus récemment, au début du mois d'octobre 2020, la tempête Alex fait plus de 20 morts et disparus sur les Alpes-Maritimes, département déjà frappé, mais sur le littoral, en octobre 2015. Les dégâts matériels associés à ces deux événements ont été estimés à près d'un milliard d'euros. Face à ces risques majeurs, la réponse des autorités doit se concentrer sur plusieurs aspects.

1.2 Gestion institutionnelle

En amont de la gestion des situations de crise et dans le but d'améliorer la compréhension des crues éclair et de faciliter leur prévision ainsi qu'une réduction des dommages sur la société à travers le monde, divers grands projets de recherche ont vu le jour comme par exemple les projets HYMEX et HYDRATE en Europe (Borga *et al.*, 2011), FLOODSCALE, ANR-Flash et BVNE, auxquels le CREER d'IMT Mines Alès a participé, en France, le CFFSE en Chine (Yuan, Xu and Arulrajah, 2017) ou encore FLASH aux Etats-Unis (Gourley *et al.*, 2017). Ces projets ont pour principal but la mise à disposition de données d'observations des crues éclair, afin d'en faciliter, entre autres, la compréhension, la simulation et la prévision.

Concernant la gestion des crises, en France, les principales institutions qui sont concernées par les inondations sont : (i) Météo France, chargée de fournir des prévisions de pluie ; (ii) le SCHAPI (Service Central d'Hydrométéorologie et d'Appui à la Prévision des Inondations), créé en 2003 par les pouvoirs publics français en appui aux Services de Prévision des Crues avec une couverture nationale sur les crues ; (iii) les Services de Prévision de Crues (SPC), chargés de l'hydrométrie locale des cours d'eau et de la prévision des crues, qui sont affectés à des bassins ou sous-bassins hydrologiquement cohérents. Les principales missions du SCHAPI consistent à produire, à partir des informations fournies par Météo France et des informations hydrométriques et prévisions produites par le SPC, des cartes de vigilance de crues sur les cours d'eaux règlementaires à destination du grand public. Les informations fournies dans ces cartes de vigilances, actualisées au moins deux fois par jour incluent quatre niveaux de vigilances colorés du vert (pas de vigilance) au rouge (risque de crue majeure) et sont mises à la disposition du public sur le site www.vigicrues.gouv.fr. Depuis peu, des alertes plus locales sont fournies par le service « Vigicrues Flash » qui consiste en une détection automatique du risque de crue soudaine à deux niveaux : risque de crue forte et risque de crue très forte, avec une actualisation toutes les 5 minutes.

1.3 Description des crues éclair

1.3.1 Caractéristiques climatiques

Les climats les plus favorables aux épisodes pluvieux intenses ou diluviens, générateurs de crues éclair, sont les climats océaniques, méditerranéens, équatoriaux, tropicaux et, à une moindre fréquence, les climats continentaux et de montagne. Les épisodes cévenols, très connus en France, sont quant à eux influencés par le climat méditerranéen. Ils se produisent majoritairement en automne ou en fin d'été.

1.3.2 Caractéristiques météorologiques

Les événements pluvieux à l'origine des crues éclair obéissent à des conditions météorologiques particulières, puisqu'elles ne se produisent que dans des endroits bien spécifiques à travers le monde. Au Vietnam, de telles pluies se produisent dans les régions montagneuses au Nord-ouest du pays (Tien Bui *et al.*, 2012). En France, c'est le plus souvent au niveau du flanc sud-est des montagnes des Cévennes que ces épisodes surviennent (Figure 1). Ils sont souvent orageux et sont caractérisés par leur forte intensité et leur faible étendue spatiale. Les conditions favorables doivent permettre une convection profonde, par soulèvement orographique, dynamique, ou orographique et dynamique. Un exemple typique est celui des épisodes cévenols. Bien que ce terme-ci soit réservé aux phénomènes météorologiques particuliers intervenant principalement dans les Cévennes et leurs environs, il est fréquent de le retrouver appliqué improprement aux phénomènes similaires se produisant sur les reliefs des arrière-pays méditerranéens. On préférera alors le terme d'épisode méditerranéen. Il s'agit d'orages très violents et localisés, s'accompagnant de pluies diluviennes qui provoquent dans la plupart des cas des inondations. Selon Jacq (1996), pour être qualifiés de diluviens, ces épisodes pluvieux doivent donner des cumuls d'au moins 200 mm en de 24h. Divers travaux (Gaume, 2002; Borrell, 2004; Ayrat, 2005; Marchandise, 2007; Coustau, 2011; Artigue, 2012) réalisent des descriptions très détaillées des processus de formation des pluies à l'origine des crues éclair. Au regard de la zone concernée par ce travail, les épisodes pluvieux concernés appartiennent à la catégorie des épisodes méditerranéens, dont certains sont des épisodes cévenols.

Pour compléter la description de ces épisodes pluvieux, nous nous référons à Artigue (2012) qui récapitule les trois grands contextes météorologiques favorables aux épisodes cévenols :

- La présence d'un forçage orographique

Ce forçage intervient dans le soulèvement d'une masse d'air douce et humide en provenance du quadrant Sud-est. Les précipitations produites peuvent dépasser 400 mm en 24h, avec des intensités dépassant rarement les 50mm/h. A titre d'exemple, en octobre 2008, un épisode pluvieux cumulait 485 mm en 24h à Mialet (30) avec une intensité maximale d'environ 50 mm/h sur l'ensemble du massif cévenol. En octobre 2021, Villefort (48) fut touché par un cumul de plus de 450mm en moins de 24h avec des lames horaires ne dépassant pas 70mm. Certains de ces événements peuvent toutefois présenter des intensités plus fortes comme à Valleraugue en 2020 (près de 700mm en 12 heures, avec plusieurs lames horaires de plus de 100mm).

- La genèse d'une structure convective à régénération arrière

Dans le cadre de ces événements, les éléments orographiques peuvent jouer un rôle, mais ne sont pas les déclencheurs du phénomène. L'élément déclencheur est le plus souvent la mise en place d'une alimentation suivant une structure convergente au sol et divergente en altitude. Ceci nécessite un contexte météorologique assez particulier qui correspond à l'approche d'un profond thalweg associé à un front froid qui s'isole parfois en cut-off, générant un flux rapide du secteur sud présentant une courbure cyclonique. Ces orages à propagation rétrograde font bien souvent parfois partie des systèmes convectifs de méso-échelle (MCS). Ils peuvent donner lieu

à des systèmes convectifs stationnaires ou à déplacement lent qui peuvent être multicellulaires rétro-régénératifs, quasi-linéaires (*Quasi linear convective systems* ou *QLCS*) ou encore orages dits en « V », dont la pointe concentre les plus fortes précipitations.

Ces orages peuvent donner plusieurs centaines de millimètres de cumuls de pluie, avec des intensités pouvant dépasser 100 mm en une heure. Ils durent à peine quelques heures dans la plupart des cas. C'est ce même type d'orages qui s'était produit au niveau de l'Aude en novembre 1999 où plus de 600 mm de pluie étaient enregistré en 36 heures dont 112 mm en une heure à Lézignan-Corbières. C'est aussi le cas du Gard en septembre 2002 où près de 700 mm de pluie tombaient en 24 heures dont près de 140 en une heure à Anduze. En 2021, le sud-ouest de l'agglomération nîmoise a aussi été frappé par ce type d'orage, avec des cumuls de près de 300mm en quelques heures.

- La combinaison des deux contextes précédents

Le soulèvement de la masse d'air chaud et humide peut être causé autant par la confrontation d'un front froid ou non que par la présence d'un relief sur le trajet de cette dernière. Les deux conditions précédentes n'étant pas exclusives, elles peuvent se combiner ou se succéder sur quelques jours pour produire des séquences pluvio-orageuses dont l'intensité est comparable aux deux premiers cas.

En France, pour les trois cas, ces événements se produisent de manière préférentielle en automne, car à la fin de l'été, les masses d'air polaires s'immiscent de plus en plus au sud, générant une augmentation de la probabilité d'un conflit de masses d'air pouvant conduire aux épisodes décrits plus haut.

1.3.3 Genèse des crues éclair

1.3.3.1 Genèse des écoulements dans le versant

Nous nous limiterons ici au cas des Cévennes, soit celui qui nous intéresse dans ces travaux. Nous avons donc, en règle générale, des pentes fortes, un réseau hydrologique bien développé et un substratum quasi-imperméable, le tout dans un contexte très peu artificialisé (pas ou très peu de terrains agricoles ou urbanisés).

Schématiquement, on peut résumer la genèse des écoulements en utilisant quatre termes :

- L'intensité de la pluie,
- La capacité d'infiltration globale (en surface et en subsurface immédiate) du sol,
- La conductivité hydraulique latérale du sol,
- La saturation des sols.

Il existe alors plusieurs cas de figure. Dans les trois premiers, on observe du ruissellement en surface, combiné ou non à des écoulements de subsurface. Notons que l'existence de ce ruissellement de surface n'est pas corroborée par la littérature ou par les observations in situ dans les Cévennes (Artigue, 2012).

- L'intensité de la pluie dépasse la capacité d'infiltration et la fraction infiltrée ne dépasse pas la conductivité hydraulique latérale : dans ce cas, du ruissellement de surface peut apparaître dans un processus appelé ruissellement hortonien (Horton, 1933). Il s'agit d'une saturation par le haut et la fraction ruisselée est égale à la différence entre l'intensité pluvieuse et la fraction infiltrée. Le sol est par ailleurs en capacité d'évacuer latéralement la fraction infiltrée ce qui évite une saturation complète du sol hydrologique.

- L'intensité de la pluie dépasse la capacité d'infiltration et la fraction infiltrée dépasse la conductivité hydraulique latérale : dans ce cas, du ruissellement de surface peut apparaître en deux temps. Dans un premier temps, le ruissellement est de type hortonien (saturation par le dessus) puis, dans un second temps, une fois le sol totalement saturé, un ruissellement de saturation par le dessous apparaît, dit hewlettien (Cappus, 1960; Hewlett *et al.*, 1967). On passe donc du cas de figure précédent à un ruissellement de toute la pluie tombée lorsque le sol devient totalement saturé.
- L'intensité de la pluie ne dépasse pas la capacité d'infiltration mais elle dépasse la conductivité hydraulique latérale du sol. Ce dernier finit par se saturer entièrement, annulant ainsi la capacité d'infiltration et autorisant le ruissellement de surface. On parle encore ici de ruissellement hewlettien.
- L'intensité de la pluie ne dépasse pas la capacité d'infiltration ni la conductivité latérale. La totalité de l'écoulement s'effectue en subsurface. Ce cas de figure est le plus adaptée à ce que l'on observe au niveau des bassins cévenols (Artigue, 2012).

L'ensemble de ces cas ne prend pas vraiment en compte la dynamique dans l'intensité des précipitations ou dans la conductivité latérale. Pourtant, plusieurs auteurs suggèrent un transfert de subsurface brutal au sein du versant, sous la forme d'une onde de pression, étant selon eux le seul à même d'expliquer les observations géochimiques (Gaume, 2002) et la réponse presque instantanée aux plus fortes intensités en pied de versant (Musy, 2009; Musy *et al.*, 2014). Dans ce cas, l'infiltration serait maximale, le sol se saturerait progressivement avec un écoulement de subsurface classique (dépendant de la conductivité hydraulique latérale du sol) puis, à la faveur d'une forte intensité pluvieuse, un transfert rapide vers le pied de versant se ferait par transfert de pression (effet piston), induisant une conductivité apparente temporairement très élevée (Hewlett et Hibbert, 1963; Rasmussen *et al.*, 2000; Musy *et al.*, 2014).

1.3.3.2 Influence de la géologie

La géologie constitue l'un des facteurs les plus influents sur les réponses hydrologiques des bassins versants. Suivant les travaux de (Vannier *et al.*, 2016), on peut décrire quatre principaux facteurs d'influence: (i) l'impact direct du cheminement souterrain de l'eau sur son temps de transit, influençant le temps de réponse du bassin versant; (ii) la nature de l'interface sol-roche mère qui est le siège des écoulements préférentiels, ce qui influence la réactivité du bassin versant; (iii) la perméabilité de la roche mère qui influence la fraction ruisselée sur le bassin versant; (iv) le stockage à l'échelle du bassin, ce qui détermine les écoulements de base et l'humidité du sol qui représentent une sorte de seuil sur la réponse du bassin en situation de crues. Compte tenu de la nature de la roche-mère, son épaisseur altérable sera plus ou moins importante. Par exemple, l'altération des formations granitiques et de certaines formations sédimentaires ont tendance à donner lieu à un sol assez profond propice aux infiltrations et aux écoulements de subsurface. Quant aux formations de type schisteux, l'altération peut être plus faible et conduit généralement à un sol de faible épaisseur. Ces formations sont réputées être dotées d'une assez faible capacité de stockage de l'eau. Sous l'influence des fortes pentes, c'est l'écoulement latéral en subsurface qui prédomine à travers une matrice parcourue de fissures et de macropores, particulièrement dans le cas des Cévennes.

Vannier *et al.* (2016) ont également analysé l'apport de la prise en compte de la couche des roches altérées pour la simulation des crues rapides ou lentes sur les bassins versants cévenols à travers des modèles à base physique. Ils ont remarqué que dépendamment de la géologie, deux comportements hydrologiques se distinguent. Sur les roches cristallines, telles que les granites et les gneiss, l'intégration des informations physiques descriptives ont permis d'améliorer considérablement la simulation des crues, qu'elles soient lentes ou rapides. En revanche, au niveau des formations

géologiques schisteuses, l'intégration des informations a été jugée bénéfique, mais insuffisante pour simuler la dynamique des crues observées. Il y émergeait l'hypothèse de l'existence d'un processus mal connu et donc mal représenté associé à la structure feuilletée des schistes.

1.3.3.3 Influence de la topographie

À côté de la forte intensité des pluies qui la provoque, certaines caractéristiques des bassins versants jouent un rôle important dans la réponse hydrologique. En ce sens, Marcos *et al.* (2016) attestent qu'outre son exposition aux épisodes pluvieux intenses, l'affinité de la région méditerranéenne aux crues éclair est également liée à son littoral montagneux qui favorise des écoulements torrentiels concentrés sur des petits bassins versant pentus et faiblement perméables. En effet, l'un des facteurs aggravants de la transformation de la pluie diluvienne en crues subites est la taille du bassin versant. Ceci traduit la capacité de ces derniers à concentrer toute l'eau précipitée dans leur réseau hydrographique assez rapidement, ils ont donc un temps de réponse relativement bref (Montz et Grunfest, 2002). À l'inverse, l'occurrence d'une pluie localisée sur un très grand bassin versant est susceptible de produire une onde de crue qui pourrait être atténuée par non seulement la grande taille du bassin versant, mais aussi la faible pente et la longueur importante du cours d'eau drainant ce dernier. Selon Gaume *et al.* (2009) tous les bassins réputés soumis aux crues rapides sont de tailles généralement petites ou modérées, allant à peine de quelques dizaines à quelques centaines de kilomètres carrés. Dans le cas des versants cévenols, il existe une combinaison entre fortes pentes des versants, faible capacité de stockage liée à une faible profondeur de la roche mère schisteuse, forte densité de drainage et fréquence des pluies de fortes intensités qui est particulièrement favorable à l'occurrence de crues éclair (Ayrat, 2005 ; Marchandise, 2007 ; Artigue *et al.*, 2012; Garambois *et al.*, 2014).

1.3.3.4 Influence de l'occupation des sols

Schématiquement, l'influence de l'occupation des sols peut être prise en compte en considérant trois niveaux : les végétations sauvages, les cultures et les zones artificialisées (Fritsch, 1995). L'influence de la végétation sauvage est liée principalement à sa nature, à son étendue à travers sa canopée et sa densité. Elle intervient en particulier dans l'interception d'une partie de la pluie qui tombe tout en favorisant une meilleure infiltration de l'eau dans le sol (Hewlett et Hibbert, 1967; Bosch et Hewlett, 1982). Plusieurs travaux dont Hewlett *et al.* (1977), Lavabre *et al.* (1991) ont prouvé l'influence positive de la présence forestière sur les bassins versants. Cette fraction d'eau susceptible d'être interceptée est variable et dépend essentiellement de la structure de l'épisode pluvieux en présence, de la quantité de pluie, de sa durée et de la saison. Plus l'intensité et la quantité de la pluie sont importantes, plus l'interception végétale est négligeable (Hewlett et Bosch, 1984; Meunier, 1996), le volume total intercepté étant limité. Par son influence positive sur l'infiltration et la saturation du sol, la présence de la végétation joue un rôle favorable à l'allongement du temps de réponse hydrologique de manière significative par rapport à un bassin versant qui en est dépourvu. Cette contribution s'explique entre autres par la réduction des effets *splash*, ou imperméabilisation précoce du sol en surface, que provoquerait la dissipation de l'énergie cinétique des gouttelettes d'eau au contact du sol nu.

Les cultures, quant à elles, ont une influence assez différente de celle de la végétation. Si la végétation sauvage intervient à travers son volume dans l'intérêt de l'infiltration, les cultures ont une influence contraire en facilitant, bien que ce soit indirectement, l'imperméabilisation des sols et la diminution des interception (Schofield et Ruprecht, 1989). Selon Cosandey (2003), les pratiques agricoles mécanisées favorisent le ruissellement par l'imperméabilisation des sols suite aux tassements fréquents provoqués par le passage des engins agricoles.

L'urbanisme et les activités touristiques interviennent à deux niveaux sur l'aggravation des impacts des crues éclair. Le premier niveau concerne le processus de la production de la crue elle-même puisqu'ils

participent à l'imperméabilisation de l'espace, ce qui a pour conséquence d'augmenter la vitesse et la fraction ruisselée de la pluie tout en diminuant l'infiltration (Hollis, 1975). Le second niveau, concernant le danger lui-même, se réfère à la présence des vies et des biens aux endroits éventuellement exposés à ces crues.

1.3.3.5 Influence des aménagements anthropiques (barrages, digues)

Les infrastructures telles que les barrages et les digues peuvent avoir un double effet sur les crues en général, et les crues éclair en particulier. La première catégorie d'effets peut être bénéfique dans la mesure où elles permettent soit d'écarter l'onde de crue si elle est produite en amont du barrage, soit de la contenir dans le lit de la rivière dans le cas des digues bien dimensionnées. En même temps, un barrage ne sera d'aucune utilité si l'onde de crue est produite en aval de celui-ci et les effets d'un mauvais dimensionnement de l'ouvrage sont particulièrement dangereux en cas de rupture. La situation peut être tout aussi catastrophique dans le cas de la rupture d'une digue, sans oublier que la digue peut servir d'obstacle à l'évacuation de l'eau en arrière de celle-ci dans certaines situations.

1.3.3.6 Influence de l'état hydrique initial du sol

L'importance de l'état hydrique initial du sol a été mise en évidence à travers la plupart des premiers travaux de modélisation du processus hydrologique. Identifié comme l'une des principales variables influençant la capacité d'infiltration des sols, il est intégré dans la plupart des modèles visant la caractérisation du ruissellement ou de l'écoulement au niveau des bassins versant (Green et Ampt, 1911; Horton, 1933; Philip, 1957; Morel-Seytoux, 1978; Diskin et Nazimov, 1995). Étant également influencé par le climat, il joue un rôle déterminant dans la réaction hydrologique d'un bassin versant (Nikolopoulos *et al.*, 2011). Au niveau des versants cévenols, la plupart des épisodes importants de crues se produisent en automne. En fin d'été, il est fréquent que l'état hydrique très sec des bassins conduise à des réponses hydrologiques plus faibles, même pour une forte pluie. D'ailleurs, après avoir analysé la base de données HYDRATE, (Gaume *et al.*, 2010) ont constaté que même en automne, il faut parfois plusieurs centaines de millimètres de pluies pour générer des crues majeures au niveau des bassins méditerranéens végétalisés.

1.3.3.7 Conclusions

Sur la base des éléments évoqués dans cette section, on peut admettre que les principales conditions hydrologiques favorables aux crues éclair au niveau des bassins cévenols se résument à leur petite taille, leurs fortes pentes, leur densité de drainage élevée, la faible épaisseur de leurs sols ainsi que, bien sûr, leur localisation dans une région où les pluies diluviennes peuvent se produire. La genèse des écoulements est assez complexe et présente encore un certain nombre d'inconnues. A l'échelle du bassin par exemple, la transformation pluie-débit est éminemment non-linéaire, mais ponctuellement, aux plus fortes intensités et durant les transferts les plus rapides vers le cours d'eau, on peut observer une transformation presque linéaire de la pluie en débit à un moment donné de l'épisode, avec une fraction de la pluie transitée quasi immédiatement à la rivière qui tend à se rapprocher de 100% (Bessiere, 2008; Artigue *et al.*, 2012).

1.4 Données caractéristiques

Outre le caractère dangereux des crues éclair, elles sont en général caractérisées par leur faible fréquence. Compte tenu de cette faible occurrence, il est difficile de disposer d'une base de données suffisamment fournie qui permettrait une caractérisation exhaustive de leur manifestation. C'est pour cette raison que le choix de la zone d'étude concerne les Cévennes, là où la fréquence de ces phénomènes est assez conséquente. Ce problème d'observation se fait sentir à deux niveaux : l'un dans la quantité de données, l'autre dans leur qualité. Dans les sections suivantes, nous présentons les différents types de données concernées, les stratégies de collecte de ces dernières, ainsi que les facteurs naturels et artificiels qui en façonnent les caractéristiques.

1.4.1 Problématique de la collecte et de traitement des données

Les principales données caractéristiques des épisodes de crue utilisées dans ce travail sont la pluviométrie et l'hydrométrie. Outre les difficultés associées à la disponibilité de ces données, leur acquisition peut se faire de plusieurs manières dont les plus courantes sont : l'installation préalable de pluviomètres au sol, les radars pluviométriques, les méthodes hybrides (pluviomètre et radar), les limnimètres associés ou non à des courbes de tarage.

1.4.1.1 Réseaux pluviométriques seuls

Les mesures effectuées par les pluviomètres sont ponctuelles. Elles peuvent être ensuite spatialisées par des méthodes d'interpolation telles que le krigeage ou les polygones de Thiessen (Delhomme, 1978; Creutin et Obled, 1982; Creutin *et al.*, 1985). Cette spatialisation des précipitations est d'autant plus représentative que la densité des pluviomètres au sol est importante. Comme la plupart des mesures physiques, les mesures pluviométriques ne sont pas à l'abri d'incertitudes même dans le cas des pluies les moins intenses. Dans le cas des pluies intenses et localisées, les incertitudes peuvent devenir très importantes si la densité du réseau de pluviomètres est faible (Villarini *et al.*, 2008). En effet, il reste possible que la cellule la plus intense d'un orage se trouve entre deux pluviomètres, induisant une mesure peu représentative au niveau de ceux-ci. Hormis ces situations défavorables, certaines incertitudes peuvent provenir de l'instrument de mesure lui-même ou de son environnement immédiat. Les incertitudes de l'appareillage sont mentionnées et détaillées par le fabricant et sont généralement inférieures à celles issues de l'environnement de mesure ou de la spatialisation des pluies. Compte tenu de ces situations, les erreurs de mesures des pluviomètres sont quasiment certaines et peuvent même atteindre 20 % sur les cumuls événementiels (Marchandise, 2007).

1.4.1.2 Radar pluviométriques seuls

Comme mentionné précédemment, outre les caractères intense et localisé des pluies qui sont à l'origine des crues éclair, leur intensité peut être très variable d'un point à l'autre dans l'espace. A titre d'exemple, on peut citer les orages en V ou celui du 8 au 9 septembre 2002 (Figure 2) dans le Gard au sud de la France (Boudevillain *et al.*, 2009).

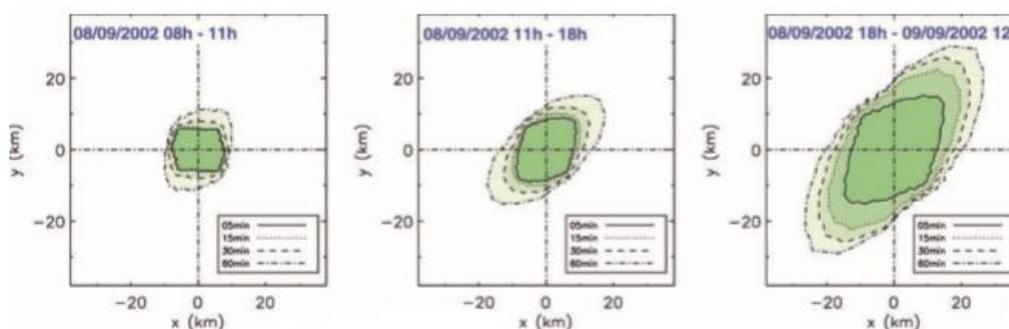


Figure 2. : Structure spatiale moyenne des champs de pluie au sol de l'épisode du 8-9 septembre 2002 dans le Gard (un système méso-échelle avec un cumul maximal de 700 mm, des intensités supérieures à 10 mm/h), radar Bollène. Traits du centre aux contours : cumul sur 5 – 15 – 30 – 60 min [tirée de Boudevillain *et al.*, 2009]

Sur les images de la Figure 2, on peut voir que l'épisode était non seulement très pluvieux (> 700 mm), mais aussi qu'il ne s'est pratiquement pas déplacé ou dispersé entre 8h et 18h le 8 septembre 2002. L'étendue de la trace des cumuls au sol reste inférieure à 10 km de diamètre durant toute cette période, et persiste au-delà de 24 heures.

Dans ces situations, pour un pluviomètre statique placé au sol, il peut s'avérer difficile de garantir une collecte satisfaisante de l'information pluvieuse représentative de ce genre d'épisodes. C'est pour cette raison qu'est mise en avant l'utilité de la technologie radar dont l'application remonte au début des années 1980 en hydrologie, pouvant fournir des informations complémentaires sur la structure spatiale de l'épisode. Les radars pluviométriques sont capables de fournir non seulement des données sur la quantité de pluie qui tombe, mais également de fournir une meilleure indication spatiale de l'épisode pluvieux. Ils peuvent également être combinés aux technologies de prévision numérique du temps (PNT) pour une meilleure prévision du développement des cellules orageuses et de leur trajectoire (OMM, 2011). Leur principe de fonctionnement consiste en l'émission d'ondes électromagnétiques dans l'atmosphère qui sont réfléchies par les précipitations. Les faisceaux réfléchis sont ensuite enregistrés, traités puis convertis en cumul de pluie par intervalle de temps. Diverses techniques dont Bellon et Austin (1984), Delrieu *et al.* (1988) existent pour la transformation des réflectivités en données de lame d'eau précipitée.

Les données pluviométriques ainsi produites sont caractérisées par une résolution de l'ordre du kilomètre carré et des rayons d'action jusqu'à 200 km. Elles ne sont donc pas exemptes d'incertitudes et d'erreurs dont les principales sources sont la présence d'obstacles physiques pouvant biaiser la réflectivité de l'onde (Tabary *et al.*, 2002), la variabilité de la nature des météores et des conditions atmosphériques, les phénomènes d'évaporation et la désagrégation des gouttelettes de pluies au cours de leur chute ou encore, la faible densité du réseau de pluviomètres leur servant de base d'étalonnage. Ce sont autant de raisons pour lesquelles, quand elles existent, les informations au sol peuvent malgré tout être plus fiables en absence d'entraves flagrantes à la mesure. Selon Dupasquier *et al.* (2000), sans corrections, les erreurs entre les données radars et lames pluviométriques peuvent dans certains cas atteindre plus de 30 %. Diverses techniques ont été mises au point afin de réduire les incertitudes dont sont sujettes les données radars (Testud *et al.*, 2000; Pellarin *et al.*, 2002). En somme, relativement, les radars apportent une information essentielle sur la structure du champ précipitant, ce qui est impossible à obtenir avec les pluviomètres ponctuels.

1.4.1.3 Combinaison de pluviomètres et de radar

Les méthodes hybrides consistent en la combinaison d'observations de pluviomètres au sol et des radars. Les pluviomètres servent alors de référence d'étalonnage pour les observations radars en utilisant un facteur correctif déterminé à partir de ces derniers. Les approches de corrections peuvent être l'application d'un facteur multiplicatif uniforme (ex. (Wilson et Brandes, 1979; Koistinen et Pahukka, 1984), d'une correction non linéaire uniforme ou une approche géostatistique par une technique de cokrigage simplifié (Krajewski, 1987; Delrieu *et al.*, 1988), par exemple. Cette approche peut se révéler plus efficace avec un réseau de pluviomètres au sol suffisamment bien disposé. En effet, elle faciliterait non-seulement une meilleure représentativité des données mesurées, mais également l'évolution de la structure spatiale de la masse nuageuse générant la pluie.

1.4.1.4 Hauteurs d'eau et débits

Généralement, les informations de débits sont issues de la conversion des chroniques de hauteur d'eau en chroniques de débit à travers une courbe de tarage préalablement établie. La courbe de tarage étant une fonction reliant la hauteur d'eau à son débit, elle est réalisée, en basses et parfois en moyennes eaux, par jaugeage du débit pour différentes hauteurs du cours d'eau. Ce jaugeage étant difficile à réaliser pour les hautes eaux, on procède souvent par extrapolation pour ces dernières. Compte tenu du différentiel de vitesse propre aux cours d'eau au droit de leur section, de la non-uniformité de leur section en travers, on peut être amené à construire la courbe de tarage en plusieurs parties. Il existe également de nos jours diverses techniques permettant de mesurer le débit de manière quasi-directe en utilisant un capteur ultra-sonique (Agence de l'eau Artois-Picardie), ou d'une

technologie radar (Costa *et al.*, 2006; Le Coz *et al.*, 2010; Zolezzi *et al.*, 2011; Dramais *et al.*, 2014). Ces technologies sans contact utilisant des capteurs situés généralement hors de l'eau en hauteur, permettent de mesurer conjointement la hauteur et ou la section mouillée, la vitesse de l'eau en surface à partir de laquelle la vitesse moyenne est déduite, puis en déterminer le débit de manière continue.

Comme pour les données de pluie, il existe diverses sources d'incertitudes accompagnant les données hydrométriques. Les situations telles que la présence de la végétation saisonnière, la présence d'obstacles temporaires tel que les débris charriés par l'eau, le curage des cours d'eau sont autant de cas où la hauteur de l'eau peut être modifiée sans pourtant affecter le débit correspondant si la courbe de tarage n'est pas adaptée. Ces situations infligent un caractère temporaire et de caducité aux courbes de tarages établies surtout dans les cours d'eau naturels. Ces incertitudes, lorsqu'elles sont ignorées peuvent prendre des proportions très importantes sur les débits déterminés par extrapolations des courbes de tarages, étant donné le caractère quadratique de la relation qui lie la hauteur d'eau à son débit. Selon Marchandise (2007), il est illusoire d'espérer des incertitudes inférieures à 20% pour les volumes et les débits alors que les erreurs sur les pics de crues peuvent approcher les 30% pour les débits extrêmes.

1.4.2 Fréquence de mesure et d'échantillonnage

La pluviométrie étant un phénomène qui est caractérisé par une forte variabilité spatio-temporelle (Villarini *et al.*, 2008), la réalisation de sa observation suffisamment réaliste est une tâche particulièrement difficile. La mesure pluviométrique étant l'enregistrement d'un cumul de pluie sur un intervalle de temps fixé, lequel cumul est ensuite attribué à la fin de cette intervalle-là. Par conséquent, plus cet intervalle est grand, plus la structure temporelle de la pluie risque d'être affectée. Villarini et al (2008) ont mené une étude sur les incertitudes liées à la fréquence des mesures réalisées par les pluviomètres. Ils ont montré qu'en-deçà d'une certaine fréquence de mesures, ces incertitudes sont proportionnelles à l'intervalle de mesure (Steiner *et al.*, 2003). Il s'y était avéré qu'une fréquence de l'ordre de 5 minutes permettait une représentation acceptable de l'observation pluvieuse. L'inverse n'est toutefois pas sans conséquence, car avec une fréquence d'observation trop élevée, on encourt un risque d'observer trop de bruit dans la donnée. Pour ce qui est de la combinaison des données des pluviomètres et des radars, divers auteurs dont (Bell *et al.*, 1990; Fabry *et al.*, 1994; Steiner, 1996; Nystuen, 1998; Jordan *et al.*, 2000; Villarini *et al.*, 2008) ont procédé à l'estimation des erreurs en fonction de la disposition spatiale des pluviomètres et de l'intervalle de mesures.

Ceci est tout aussi valable pour les mesures hydrométriques. En effet, s'il s'agit d'un cumul pour les données de pluies, c'est généralement une moyenne dans le cas des données de hauteurs d'eau sur l'intervalle considéré. Pour une fréquence trop élevée, les vaguelettes de surface risquent d'être traitées comme vraie hauteur d'eau alors qu'elles ne sont que des bruits. A l'inverse, une fréquence trop faible risque de manquer des données telles que l'instant du pic de la crue ou d'autres évolutions intermédiaires importantes.

Quant au rééchantillonnage visant l'utilisation de la donnée, il y a lieu de tenir compte de l'objectif lié à l'utilisation de la donnée et des contraintes opérationnelles. Selon le théorème de Nyquist-Shannon, aucune information n'est perdue lors de l'échantillonnage si la fréquence utilisée est au moins le double de la fréquence la plus élevée du signal (Shannon, 1998). Sur cette base, on préférera une donnée dont la fréquence d'observation serait un bon compromis entre le nombre de valeurs et la conservation de l'information contenue dans le signal en question. Dans le cas des crues au niveau du Gardon, une fréquence d'échantillonnage de l'ordre de 20 à 30 minutes est acceptable (Toukourou, 2009).

2 Questions scientifiques

Plusieurs auteurs ont constaté qu'outre le caractère dévastateur des crues éclair, ajouté à la faible documentation les concernant, leur anticipation reste un problème majeur tant sur le plan scientifique que technique. Pour citer Estupina Borrell *et al.* (2005), la rareté de ces phénomènes amplifie la difficulté de leur analyse statistique, y compris la calibration des modèles hydrologiques associés. Il est également un fait bien connu que la plupart des modèles à base physique peuvent produire une simulation acceptable de ces événements, mais très peu arrivent à y produire de la prévision utilisable en conditions opérationnelles (Borrell, 2004; Estupina Borrell *et al.*, 2005), notamment parce que les prévisions météorologiques ne permettent pas de disposer des données de pluie avec une qualité suffisante aux échelles de temps et d'espace de ces phénomènes.

Les principales sources de difficulté sont liées à la variabilité spatiale des précipitations, des processus hydrologiques, eux-mêmes parfois mal définis, induisant une nature non-linéaire de la relation pluie-débit. En effet, hormis les précipitations, les aspects physiques du terrain ont une influence non-négligeable sur ces crues. L'implication de ces derniers éléments est en général mal décrite pour la grande majorité des bassins versants.

Les approches statistiques ont en revanche montré qu'elles étaient performantes grâce à leur caractère prédictif sans prévision de leurs entrées et grâce au fait qu'elles ne font pas d'hypothèse sur les processus mis en jeu. Cependant, compte tenu de leur caractère de « boîte noire », ces approches ont parfois dû faire face à des réticences. Cependant, vue la démocratisation de l'intelligence artificielle et les performances qu'elle atteint, en particulier pour les modèles à réseaux de neurones et le *Deep Learning* (Lecun *et al.*, 2015; Shen, 2018), cette approche a gagné en popularité. Le *Deep Learning* en hydrologie et l'interprétation physique de ces modèles de type boîte noire sont en revanche encore peu répandus.

C'est donc dans l'objectif d'appliquer des modèles de type *Deep Learning* aux crues éclair que ces travaux ont été menés. En particulier, l'extraction d'informations physiques du modèle ou, à l'inverse, l'introduction de telles informations dans celui-ci a été pratiquée, afin d'améliorer la modélisation de ces phénomènes et de tempérer l'image de boîte noire dont ils pâtissent parfois. Ainsi, un modèle profond sera mis en œuvre sur un bassin cévenol et, de ce modèle, seront extraites des caractéristiques physiques, notamment sur la répartition spatiale des pluies. Dans un second temps, de l'information physique en provenance du processus de la crue éclair sera introduite dans le modèle afin d'apprécier son comportement en simulation ou en prévision sur le bassin versant. On exploitera la polyvalence et la performance de l'approche du *Deep Learning* pour tenter de décrire physiquement une part de la transformation de la pluie intense en crue éclair. Ces informations permettront d'améliorer la modélisation de ces crues et donc leur prévision.

3 Conclusion

Dans ce chapitre, nous avons présenté l'essentiel de la problématique des crues éclair en évoquant le risque qu'elles font peser sur la société, les difficultés liées à leur caractérisation et à leur anticipation. Nous avons également souligné les difficultés liées à la disponibilité et à la qualité des données les caractérisant en vue de la réalisation d'un modèle d'anticipation le plus performant possible. Ceci a amené à constater que les données dont on dispose sont souvent très bruitées et peuvent induire des biais. Mais elles ne sont pas les seules sources de biais : spatialisation des pluies, influence des conditions hydriques initiales des sols, impact de la géologie sur les écoulements, etc. rendent également difficile la compréhension physique de la genèse des crues éclair. Dans la littérature, divers travaux ont mis en évidence les limites de la plupart des modèles basés sur les processus, ce qui peut être interprété comme une exigence d'alternatives plus efficaces, tout au moins complémentaires.

L'utilisation des modèles statistiques s'impose aujourd'hui comme une option crédible. Parmi ces modèles, le *Deep Learning* est de plus en plus utilisé pour la résolution de problèmes variés, en sciences de l'environnement ou dans d'autres domaines. Cependant, son utilisation en hydrologie n'est pas aussi répandue qu'elle ne l'est pour des domaines tels que l'imagerie, ou les sciences médicales. Au vu de sa performance multidisciplinaire, son application à la modélisation des crues éclair peut ouvrir de nombreuses possibilités d'amélioration des simulations et prévisions.

1 Etat de l'art sur la modélisation des crues éclair

1.1 Modélisation hydrologique : simulation et/ou prévision

1.1.1 Généralités

Selon l'OMM (1992), un modèle hydrologique peut se définir comme une représentation mathématique d'un système hydrologique. Cette représentation mathématique prend donc la forme d'une fonction $f(\cdot)$, telle que $y = f(x)$ où y est la grandeur physique à représenter et x la variable qui l'explique. Dans un modèle pluie-débit, la pluie constitue l'une des principales variables qui explique le débit. D'autres variables telles que la température ou l'évapotranspiration, combinées entre elles à l'aide de paramètres, dont la représentativité physique est plus ou moins importante, sont autant d'éléments qui modulent le comportement hydrologique du bassin versant. La réalisation d'un modèle hydrologique se base sur une série de combinaisons d'éléments dont les principaux ont fait l'objet d'une assez cohérente classification. Selon Gaume (2003), Borrell (2004), Mathevet (2005), Marchandise (2007), Artigue (2012), etc., les principales composantes d'un modèle hydrologique sont :

- Les variables d'entrées

Les variables d'entrées sont les données qui vont être utilisées par le modèle pour produire une sortie. D'une manière générale, elles constituent les signaux ou impulsions qui font réagir le système naturel. Par conséquent, ce sont les variables explicatives du phénomène. En hydrologie, ce sont principalement les précipitations, la température, l'évapotranspiration ou encore les apports extérieurs.

- Les variables d'état

Les variables d'état décrivent l'évolution de l'état du système naturel. Pour un système hydrologique, les variables d'état peuvent être la piézométrie ou encore l'humidité du sol.

- Les variables de sorties

Ces variables constituent la grandeur à modéliser ou à expliquer par le système artificiel. Dans le cadre de la modélisation hydrologique, il s'agit le plus souvent du débit ou de la hauteur d'eau à l'exutoire du bassin versant. La variable en sortie peut également être l'évolution du niveau piézométrique, de l'humidité du sol, du niveau d'eau dans un barrage etc. Dans les modèles inverses, les variables d'entrée décrites plus haut peuvent se retrouver en sortie.

- Les paramètres

Les paramètres constituent des éléments propres du modèle assurant la manière dont les variables sont combinées pour expliquer la sortie désirée. Leur typologie et leur nature sont définies en fonction du type de modèle en question. Leur valeur peut être fixée par des mesures in situ, ou déterminées par un processus itératif de calibration, automatisé ou non.

- Les conditions initiales, conditions aux limites

Les conditions initiales permettent de situer l'état du système avant le début de la période modélisée. La plupart du temps, il s'agit simplement de la première valeur des variables d'état. Quant aux conditions aux limites, il s'agit de variables qui décrivent les conditions d'interaction de l'hydrosystème

modélisé avec les hydrosystèmes voisins et qui peuvent avoir une influence sur lui (marées, hauteur d'un lac, apports ou pertes, etc.)

1.1.2 Types de modélisation

Les modèles peuvent être ou non à vocation prédictive. Leur fonctionnement est donc différent selon cette vocation.

1.1.2.1 Approche non prédictive

Dans ce cas, la simulation consiste à reproduire le fonctionnement observé du système naturel. On détermine donc la grandeur à modéliser à l'instant k , à partir des variables explicatives prises jusqu'à ce même instant avec ou sans un historique.

On a donc :

$$y(k) = f(x_{k-r}, \dots, x_k) \quad (\text{Éq. 1})$$

Avec r la profondeur historique considéré sur les variables entrées x ; k , le temps discret ; $y(k)$, la sortie du modèle ; $f(\cdot)$, la fonction réalisée par le modèle.

1.1.2.2 Approche prédictive

La prévision consiste à estimer les variables de sortie au-delà du dernier instant d'observation. Elle fait intervenir un horizon de prévision non-nul (h_p).

$$y(k + h_p) = f(x_{k-r}, \dots, x_k) \quad (\text{Éq. 2})$$

Avec r la profondeur historique considérée sur les variables entrées x ; k le temps discret ; h_p l'horizon de prévision du modèle ; $y(k + h_p)$, la sortie du modèle ; $f(\cdot)$ la fonction réalisée par le modèle.

Dans un processus de prévision, le modèle doit anticiper l'évolution de la variable de sortie. Naturellement, sans information sur les valeurs futures des variables d'entrées, seul le temps de réponse et/ou l'inertie du processus de transfert peuvent fournir des résultats utilisables sur des horizons courts. Il existe donc des approches sans prévision des entrées mais aussi des approches avec prévisions des entrées. Dans ce cas, on utilise la plupart du temps des prévisions issues de modèles météorologiques. Pour les grands bassins et les longues échéances, la résolution des modèles météorologiques peut être assez grossière. Pour les bassins de taille plus réduite et à court terme, elle doit être élevée afin de résoudre la convection à l'origine des épisodes pluvieux. Ces prévisions sont certes en amélioration constante mais, la plupart du temps, elles restent d'une précision insuffisante aux échelles de temps et d'espace nécessaires (quelques dizaines de kilomètres carrés et quelques heures). En l'absence de prévisions issues de modèles météorologiques, il est possible de considérer des pluies futures nulles, égales à la moyenne des dernières pluies ou encore constantes par rapport à la dernière pluie. Ces prévisions rudimentaires sont la plupart du temps erronées, mais elles permettent de projeter dans le futur quelques-uns des cas de figure possibles.

À des fins de recherche uniquement, des approches basées sur des pluies prévues dites « parfaites » peuvent aussi être mises en œuvre. Cela consiste à utiliser, *a posteriori*, les observations de pluie (le terme « parfait » est alors assez inadapté, puisque les pluies observées ne sont pas parfaites...) comme prévision de précipitations. Il s'agit là de distinguer, dans la chaîne de prévision, l'erreur issue de la prévision météorologique de l'erreur issue du modèle hydrologique et de son initialisation. Cette démarche n'a aucune forme d'utilité en conditions opérationnelles (temps réel).

1.1.3 Catégorisation des modèles

Il reste difficile de fournir une classification exhaustive des modèles hydrologiques, car selon l'approche envisagée, un modèle peut appartenir à plusieurs classes en même temps. Cependant, certains travaux tels que (Borrell, 2004 ; Artigue, 2012) proposent une catégorisation bien détaillée des modèles hydrologiques. Entretemps, dans le cadre de ce travail, nous considérons deux angles de vue orientés vers la prise en compte ou non des processus physiques et l'échantillonnage des données pour proposer deux grands groupes.

1.1.3.2 Prise en compte des processus hydrologiques

Selon la considération des processus hydrologiques par les modèles, Perrin (2000) propose trois groupes :

- Les modèles empiriques ou statistiques

Les modèles statistiques se rapportent à des boîtes noires (Clarke, 1973). Ils n'intègrent aucune approche sur les processus physiques de la transformation de la pluie en débit, en supposant que cette dernière est indirectement expliquée dans la relation entre les observations en entrée et celles faites à l'exutoire du bassin versant. Ils peuvent représenter le comportement hydrologique global du bassin versant par une simple fonction mathématique paramétrée entre les variables d'entrée (pluies, températures, ETP ...) et sorties (débits, ETP, ...). Le traitement réalisé sur les données observées est de type déterministe. Il repose sur l'adaptation des paramètres intrinsèques de ladite fonction mathématique, sans que ces derniers n'aient à priori une signification physique. Ces modèles sont appréciés pour leur efficacité, leur robustesse, leur polyvalence et leur universalisme moyennant la disponibilité d'une bonne base de données. Les modèles à réseaux de neurones, qui sont utilisés dans cette étude, font partie de cette catégorie. Ces derniers sont présentés au niveau de la section 2 de ce chapitre.

- Les modèles conceptuels

Les modèles conceptuels font référence à une représentation simplifiée du système physique naturel. Ils assimilent la plupart du temps le bassin versant à un ensemble de réservoirs connectés. Le débit observé à l'exutoire est donc une résultante intégrant les influences des forçages atmosphériques (pluie, évapotranspiration, température, ...), anthropiques (prélèvement, transfert entre bassins, etc ...) et les flux d'échanges qui se manifestent entre les différents réservoirs. Les principaux processus physiques interprétés sont l'interception, l'infiltration, l'évapotranspiration, le ruissellement et le stockage. Leurs performances dépendent énormément des conditions initiales matérialisées par le niveau de saturation du bassin qui est traduit par un taux de remplissage d'un réservoir. A l'instar des modèles statistiques, les paramètres du modèle conceptuel n'ont pas forcément une signification physique au sens propre, même s'ils représentent, grossièrement, certaines caractéristiques comme la capacité d'infiltration, la conductivité hydraulique ou encore la capacité de stockage. Les modèles du type GR pour Génie Rural (Edijatno *et al.*, 1999; Perrin *et al.*, 2007) sont des exemples typiques de modèles conceptuels. Un autre exemple bien connu des modèles appliqués aux bassins versants comportant des glaciers est celui du modèle HBV (Bergström, 1995). Le modèle SCS pour *Soil Conservation Services* est également un modèle conceptuel beaucoup plus simple où le bassin versant est considéré comme un réservoir unique dont le niveau monte et descend sous l'influence des précipitations et l'écoulement à l'exutoire. Les seuls paramètres considérés sur ces derniers sont la transmissivité et la capacité du réservoir.

- Les modèles à base physique

Les modèles à base physique reposent sur une représentation la plus fidèle possible des processus physiques gouvernant la transformation de la pluie en débit. Ils font appel à diverses lois physiques dont les plus populaires sont les familles des équations de Saint-Venant pour les écoulements à faible profondeur, les relations de Manning pour les écoulements dans les rivières, la loi de Darcy pour les écoulements souterrains en milieu saturé ou encore l'équation de Richards pour les écoulements en milieu non-saturé.

La mise en œuvre pratique de ces modèles se heurte à la grande complexité des processus physiques liés au dynamisme d'un bassin versant. Cette complexité est liée non seulement à la connaissance des différents processus, mais aussi à l'accès aux données les décrivant (Beven, 1989). Cette contrainte amène à considérer des hypothèses pas forcément vérifiables pour les processus mal connus, ce qui donne lieu à la nécessité d'avoir un nombre conséquent de paramètres (Payraudeau, 2002). Théoriquement, ces modèles ne nécessitent pas de calibration : les mesures *in situ* peuvent être intégrées directement. Dans les faits, la variabilité spatiale et temporelle de certaines caractéristiques hydrauliques et hydrologiques des bassins ne permet pas de fixer des paramètres sans avoir à les ajuster. Le postulat de la compréhension intégrale des phénomènes à l'œuvre dans la réponse hydrologique reste assez théorique, en particulier dans les crues éclair.

1.1.3.3 *Considération sur la discrétisation temporelle*

- Modèles évènementiels

Les modèles évènementiels utilisent une base de données constituée d'événements définis suivant l'objectif visé. Ces événements peuvent être des épisodes d'étiages, de sécheresses, de crues ou de pluies définis à partir d'un seuil donné. Dès lors ce seuil défini, la durée de l'épisode est variable. Un modèle évènementiel donné est spécialisé et réservé au type d'événements pour lequel il a été conçu.

- Modèles continus

Certains modèles visent une représentation du comportement hydrologique des bassins versants sur une longue période incluant une succession de situations hydrométéorologiques différentes. Ces modèles se veulent la représentation d'un processus continu sur un intervalle de temps relativement long. Ils peuvent être utilisés pour modéliser l'évolution d'une variable hydrologique sur plusieurs années (Boughton et Droop, 2003). Bien qu'il n'y ait pas une limite temporelle bien définie catégorisant les modèles continus, la plupart des travaux réalisés dans ce sens concernent des périodes allant d'une saison à plusieurs années.

1.1.3.4 *Autres critères de classification*

D'autres critères de catégorisation de modèles existent. Selon le fait que l'on considère l'intégralité du bassin versant ou qu'on le divise en sous-bassins, le modèle peut être global, semi-distribué ou distribué. Sood et Smakhtin (2015) ont réalisé une revue assez générale sur les modèles globaux. Singh (1995) présente également un ensemble de critères de classification des modèles globaux et distribués. Topmodel (Beven, 1997) et MARINE (Borrel, 2004) sont des modèles distribués à base physique qui reposent sur la distribution des paramètres topographiques qui gouvernent la réponse hydrologique du bassin versant.

La plupart des modèles hydrologiques sont déterministes. L'introduction d'une variable aléatoire dans un modèle le rend stochastique. Cette variable aléatoire peut se référer à une grandeur méconnue ou mal connue mais nécessaire au fonctionnement du modèle.

1.1.4 Critères d'appréciation des modèles hydrologiques

Selon l'INRAE (2021), le pouvoir prédictif d'un modèle hydrologique performant peut dépendre de quatre qualités essentielles et non-exclusives :

- Généralisable

La capacité de généralisation d'un modèle est l'une des caractéristiques les plus recherchées en hydrologie. On attend d'un modèle, produit pour un bassin versant, qu'il soit également performant sur d'autres bassins versants qui lui sont physiquement semblables. Bien qu'il soit une cause perdue d'espérer que des bassins versants soient identiques, certaines considérations physiques, géographiques, climatiques et/ou météorologiques peuvent permettre le regroupement de bassins par leurs similitudes. On attend donc du modèle qu'il soit capable de s'adapter à des bassins versants variés.

- Fiable

Selon Yang et Parent (1996) la fiabilité d'un modèle hydrologique est une mesure de l'adéquation entre sa réalisation et ce qui est attendu par le modélisateur. Celle-ci peut se faire par mesure de l'erreur dans l'absolu sur les données disponibles et de manière relative dans une perspective d'extrapolation. Le modèle fiable est capable d'être performant dans diverses situations, ce qui rejoint la notion de robustesse.

- Robuste

Selon Ouarda et Ashkar (1998), statistiquement la robustesse fait référence à la capacité d'un modèle à maintenir sa performance dans diverses situations et dans des conditions non optimales au regard de sa conception. On s'attend alors à ce que le modèle robuste reste opérationnel en dehors de sa zone de « confort ». Selon Kuczera (1982), la robustesse d'un modèle hydrologique est fonction de deux propriétés : résistance (Matalas et Fiering, 1977) et efficacité. La résistance d'un modèle est matérialisée par sa capacité à performer sur des événements extrêmes sans pertes de performance. Quant à l'efficacité, elle conditionne la performance d'un modèle résistant. Selon Kuczera (1982) l'identification d'un estimateur résistant ne garantit pas forcément une performance acceptable, puisqu'une variation de ce dernier sur les données disponibles, aussi faible soit-elle, peut être considérée comme exagérée dans la pratique. Il est évoqué qu'un manque d'efficacité peut être lié à la représentativité des données, tout comme l'architecture du modèle. D'où la nécessité de viser la parcimonie dans la conception des modèles (Tukey, 1961; Box *et al.*, 2015)

- Stable

On peut définir la stabilité d'un modèle hydrologique par la stabilité de ses variables internes ou externes. Lorsque des entrées finies sont appliquées au modèle, ce dernier ne génère pas de signaux qui divergent. On peut élargir ce concept à la génération d'oscillations plus ou moins rapides ou plus ou moins fortes. Ces oscillations sont vues comme une première marque de l'instabilité.

1.2 Grandes étapes de mise en œuvre d'un modèle

1.2.2 Choix du modèle

La question du choix d'un modèle s'aborde à trois niveaux. Le premier est la prise en compte de l'objectif de modélisation. Le choix du modèle sera différent selon que l'on souhaite aborder la question des crues ou des basses-eaux, la fonte glaciaire, les apports en eau ou encore la définition de débits caractéristiques. Le deuxième niveau concerne le type de modèle. Ce choix repose en général sur la préférence du modélisateur et ses compétences, bien qu'il devrait surtout reposer

sur les performances des différents types de modèle dans une situation se rapprochant de l'objectif de modélisation. Il n'existe néanmoins pas d'approche préétablie pour ce type de choix. Le troisième niveau repose sur la recherche de la complexité optimale du modèle désigné au regard du nombre de paramètres et de la quantité de données disponibles. Selon Blöschl et Grayson (2001), il existe une relation non-linéaire assez informative entre l'évolution de la complexité du modèle, la quantité des données et la performance du modèle. Selon cette relation, pour une quantité de données disponible, la performance du modèle passe par un maximum en faisant varier progressivement le nombre de paramètres du modèle. Le nombre de paramètres pour lequel la performance du modèle est maximale correspond à la complexité optimale du modèle postulé. Quant aux données, leur qualité et leur quantité influencent sensiblement la performance du modèle. Plus on améliore leur quantité, plus la performance du modèle peut augmenter avant d'atteindre un seuil. Au-delà de ce seuil, sans augmentation de la complexité du modèle, aucune information significative supplémentaire ne saurait être apportée par le modèle.

1.2.3 Calage et/ou apprentissage

Les termes « calage » et « apprentissages » sont essentiellement similaires. Cependant, le premier est plus répandu et plus adapté aux modèles à base physique, alors que le second fait plus référence aux modèles statistiques. L'étape du calage concerne l'ajustement des valeurs des paramètres qui lient les variables du modèle en vue de la réduction d'une différence entre simulation et observation, donc du biais. Cet ajustement peut être réalisé manuellement suivant un processus « essai-erreur », ou de manière automatique en faisant intervenir des algorithmes d'optimisation et des supports de calculs appliqués sur une fonction mesurant l'erreur entre valeur simulée et valeur observée du processus (Madsen, 2000). Il existe une panoplie de méthodes et d'algorithmes d'optimisation. Compte tenu du nombre souvent élevé des paramètres de certains modèles, le domaine décrit par la fonction d'erreur correspondant est une surface discontinue, non-dérivable, contenant plusieurs zones de convergences et de minima locaux. Ces derniers traduisent l'existence de plusieurs jeux de paramètres offrant des solutions partielles au problème en question. Ils constituent en revanche de véritables pièges pouvant limiter la performance de certaines méthodes d'optimisation (Luenberger et Ye, 1984).

La recherche de la solution optimale fait donc intervenir deux grandes tendances : l'une est purement mathématique incluant des processus itératifs avec des méthodes de calcul directes ou de gradients ; l'autre, moins populaire, est probabiliste. Les méthodes les plus populaires de la première tendance sont celles se basant sur le gradient de la fonction de l'erreur en fonction des paramètres. Swann (1969) propose une revue sur l'optimisation non-linéaire. Les méthodes de type probabiliste (Duan *et al.*, 1992) font souvent intervenir des approches d'équifinalités (Beven, 1993) ou de vraisemblance (Beven et Binley, 1992).

1.2.4 Validation

L'étape de validation consiste à apprécier la capacité du modèle dimensionné à satisfaire l'objectif de la modélisation, c'est-à-dire sa capacité à reproduire le comportement du bassin versant à partir des données qui n'ont pas été utilisées lors de la phase de calage. Cette étape se réalise à deux niveaux. Le premier niveau concerne le calage même des paramètres. Le second niveau, correspondant à la validation, consiste à appliquer le jeu de paramètres du modèle calé sur un jeu de données qui n'a jamais été utilisé pour le calage. Cette étape permet d'apprécier la robustesse du modèle.

1.3 Critères d'évaluation des modèles

Dans la grande majorité des cas, les qualités d'un modèle sont appréciées en évaluant l'écart entre ses sorties et les observations. Dans le cadre d'un modèle hydrologique, cette évaluation se base sur l'écart entre le débit simulé et le débit observé à l'exutoire du bassin versant. Au regard du type de problème

en présence et de l'objectif poursuivi, il existe plusieurs angles d'approches pour cette évaluation, donnant lieu à plusieurs critères de qualité.

Dans le cadre de ce travail de modélisation de crues éclair, nous considérons trois angles d'observation : l'erreur globale sur l'hydrogramme observé, les erreurs d'amplitude et de temps sur les pics de crues ainsi que la capacité du modèle à faire mieux que la prévision naïve dans un contexte de prévision.

1.3.2 Critère de Nash-Sutcliffe (Nash et Sutcliffe, 1970)

Le critère de Nash-Sutcliffe (Nash et Sutcliffe, 1970), appelé souvent critère de Nash ou le NSE (pour *Nash and Sutcliffe Efficiency*), est le plus populaire des critères utilisés en modélisation hydrologique. Il est du type R^2 . Il se calcule comme suit.

$$Nash \text{ ou } R^2 = 1 - \frac{\sum_{k=1}^n (Q_k - \hat{Q}_k)^2}{\sum_{k=1}^n (Q_k - \bar{Q})^2} \quad (\text{Éq. 3})$$

$$-\infty < Nash \leq 1$$

Avec Q_k , le débit observé à l'instant discret k ; \hat{Q}_k , le débit simulé pour l'instant discret k ; \bar{Q} , le débit moyen observé durant les n observations ; k , le temps discret ; n , le nombre de valeurs.

Plus le Nash est proche de 1, plus la performance du modèle est bonne. Cependant, cette performance reste relative à la complexité du problème en question. Dans le cadre de la modélisation des crues éclair, un Nash à partir de 0.8 peut être signe d'une performance acceptable (Moussa *et al.*, 2007).

Il y a trois repères d'appréciation pour ce critère :

- Si **Nash = 1**, cela indique que le modèle n'a produit aucun écart par rapport à ce qui a été observé. Le débit simulé correspond parfaitement au débit observé. Mathématiquement, le numérateur (l'erreur quadratique moyenne) du second terme de l'équation s'annule. Cette situation ne peut être qu'idéale dans la majorité des cas, mais compte-tenu des incertitudes sur la mesure du débit, elle ne peut pas traduire une modélisation parfaite.
- Si **Nash = 0**, cela indique que le modèle réalise l'équivalent du simple calcul de la moyenne des n débits observés. Mathématiquement, le second terme de l'équation vaudrait l'unité, annulant ainsi cette dernière.
- Si **Nash < 0**, cela indique que le modèle est incapable de faire mieux que le simple calcul de la moyenne des débits observés. Il n'y a alors aucun intérêt d'utiliser un tel modèle.

Le critère de Nash comporte cependant une limite bien connue : (i) un épisode pluvieux très long tend à fournir une moyenne assez faible, donnant ainsi un dénominateur relativement grand qui tend à diminuer l'importance des écarts entre débit observé et simulé ; (ii) une bonne prise en compte des valeurs de début crue et de celles de la récession peut masquer l'importance d'un éventuel grand écart sur les pics de crue. Ces deux limitations sont à prendre en considération dans l'évaluation de la performance des modèles par ce critère. C'est l'une des raisons pour lesquelles une analyse des pics est nécessaire.

1.3.3 Critères de pic

Les critères de pic se basent autant sur les aspects quantitatifs des pics que sur leur synchronisation. Ce critère comporte trois variantes :

- Le **PPD** (*Percentage of the Peak Discharge*) est le rapport du pic de la crue simulée à celui de la crue observée sans considération sur un éventuel écart temporel.
- Le **SPPD** (*Synchronous Percentage of the Peak Discharge*) ou pourcentage de débit simulé au moment du pic de la crue observée. Contrairement au PPD, il permet d'apprécier la capacité du modèle à produire un débit de pointe à l'instant voulu.

$$\text{SPPD}(\%) = 100 * \frac{\hat{Q}_{k|Q^{max}}}{Q^{max}} \quad (\text{Éq. 4})$$

Avec Q^{max} , la valeur du pic du débit observé ; $\hat{Q}_{k|Q^{max}}$, la valeur du débit simulé à l'instant du pic du débit observé.

Plus le SPPD est proche de 100 %, plus le modèle est correct sur les pics de crue, à condition que le modèle ne produise pas un PPD très élevé avec un décalage significatif, qu'il faut donc évaluer.

- Le **P_D** (Peak delay). Ce critère évalue la différence temporelle entre le pic simulé et le pic de crue observé sans tenir compte de l'aspect quantitatif. Il vaut :

$$P_D = k_{\hat{Q}^{max}} - k_{Q^{max}} \quad (\text{Éq. 5})$$

Avec $k_{\hat{Q}^{max}}$, l'instant du pic de la crue simulée ; $k_{Q^{max}}$, l'instant du pic de la crue observée.

Plus le **P_D** est proche de zero (0), plus le modèle est correct sur les pics de crue.

Ces trois critères sont complémentaires et permettent de situer la performance du modèle sur les pics en amplitude comme en temps.

1.3.4 Critère de persistance

Parmi les simulations qui donnent lieu à des critères de Nash satisfaisants, il en est une particulière que l'on appelle la prévision naïve et qui consiste à reproduire la valeur observée à l'horizon de prévision voulu. Une telle prévision n'apporte pourtant aucune information à l'utilisateur. C'est pour cette raison que Kitanidis et Bras (1980) ont proposé le critère de persistance (Éq.6). Ce critère permet de vérifier la différence entre le résultat du modèle et la prévision naïve. Il se base sur une comparaison de la variabilité du débit anticipé par rapport à celle du débit observé à l'horizon de prévision considéré. Plus ces valeurs se rapprochent, plus le modèle est proche de la prévision naïve.

$$C_p = 1 - \frac{\sum_{k=1}^n (Q_{k+h} - \hat{Q}_{k+h})^2}{\sum_{k=1}^n (Q_{k+h} - Q_k)^2} \quad (\text{Éq. 6})$$

$$-\infty < C_p \leq 1$$

Avec Q_{k+h} , le débit observé à l'horizon de prévision h au-delà de l'instant k ; \hat{Q}_{k+h} , le débit simulé pour l'horizon de prévision h au-delà de l'instant k ; Q_k , le débit observé à l'instant k ; k , l'instant d'observation considérée ou disponible.

Comme pour le critère de Nash, il y a lieu de considérer trois repères d'appréciation pour ce critère :

- Si $C_p = 1$, cela indique que le modèle a réalisé une prévision parfaite. Le numérateur s'annule, il n'y a eu aucun écart entre le débit prévu et le débit effectivement observé.
- Si $C_p = 0$, cela indique que le modèle réalise une prévision équivalente à la prévision naïve.
- Si $C_p < 0$, cela indique que la prévision réalisée par le modèle n'est pas meilleure qu'une prévision naïve.

1.4 À propos de la modélisation hydrologique sur les crues éclair

En raison des spécificités des crues éclair que nous avons déjà eu l'occasion de décrire (variabilité spatiale et temporelle des pluies élevée, conditions de mesures difficiles, non linéarité et non stationnarités diverses dans les processus, rareté des événements, bruit élevé dans les données), leur bonne représentation par des modèles est bien souvent entravée. Pour la prévision, le couplage hydrométéorologique, en sus des incertitudes liées à la prévision météorologique elle-même, présente un niveau de difficulté supplémentaire.

Cette approche de couplage qui permet d'intégrer les prévisions de pluies dans les modèles hydrologiques a été explorée par différents auteurs. En général, le couplage des modèles météorologiques à des modèles hydrologiques fournit des performances assez intéressantes sur les bassins de grandes tailles où les crues sont généralement lentes. À titre d'exemple, Bartholmes et Todini (2005) ayant utilisé cette approche sur le bassin du Pô en Italie, ont pu modéliser même à près de 10 jours à l'avance, plus 80 % du pic de crue. Sans compter le fait que la grande taille du bassin pallie l'effet de la variabilité spatiale et temporelle des précipitations, toute onde de crue localisée, aussi importante soit-elle, est sujette à un amortissement avant d'atteindre l'exutoire. Un tel scénario ne peut être espéré au niveau d'un petit bassin versant. Alfieri *et al.* (2011) ont testé l'utilisation des prévisions météorologiques de pluies pour la prévision des crues sur le Gardon d'Anduze. Les résultats obtenus ont fait état d'une sous-estimation des crues importantes. Les conclusions de cette étude ont porté sur la robustesse des modèles et la nécessité de tester cette méthode sur plusieurs événements. En 2001, Dolciné *et al.* (2001) avaient pu montrer que l'utilisation des pluies futures obtenues par télédétection pouvait améliorer la prévision des crues sur le Gardon d'Anduze. Cependant, à cause de la faible importance de la crue qui servait de test en condition opérationnelle, l'approche n'a pas pu être jugée généralisable. Toutefois, les auteurs ont pu souligner que les informations apportées par les prévisions de pluies n'étaient pas très utilisées pour les courts horizons de prévisions.

Dans la littérature, diverses approches ont été adoptées pour modéliser les crues éclair. En 2005, Borrell *et al.* (2005) ont utilisé le modèle distribué à base physique MARINE dans la modélisation de crues éclair. Ils ont obtenu des performances acceptables en prévision en temps réel, quoique moins bonnes qu'en simulation. En 2010, Braud *et al.* (2010) ont utilisé les modèles distribués CVN et MARINE pour simuler la crue de septembre 2002 dans le Gard. Ils ont pu identifier la pluie comme la principale variable qui gouverne la dynamique des pics de crue tout en appréciant l'influence de diverses autres composantes physiques du bassin versant. Cependant, les conclusions ont recommandé plus d'observation sur le terrain afin d'améliorer l'explication des processus sous-jacents. En 2007, Moussa *et al.* (2007) ont utilisé le modèle MODSPA, un modèle distribué à base physique, pour simuler les processus rapides et lents sur le bassin versant du Gardon d'Anduze. Ils en ont obtenu des performances assez élevées avec des Nash proches de 0.9.

Divers auteurs dont Toukourou (2009), Artigue (2012), Darras (2015) ont réalisé des modèles de prévisions des crues éclair à partir des modèles statistiques de type réseaux de neurones sur les bassins versant des Gardons de Mialet et d'Anduze. Kong A Siou (2011) a également utilisé des réseaux de neurones dans la prévision des crues dans l'aquifère karstique du Lez. Ils ont obtenu des performances très satisfaisantes avec des Nash au-delà de 0.9 et des persistances utiles jusqu'au temps de réponse des bassins versant en question. C'est également ce type de modèle qui est utilisé dans le cadre de la présente étude. À travers les différents travaux consultés, on constate que les modèles à réseaux de neurones ont l'avantage d'être fondamentalement prédictifs du fait qu'ils peuvent réaliser des prévisions de crue en l'absence de pluies futures jusqu'au temps de réponse du bassin versant. Par ailleurs, étant donnée la faible connaissance des processus mis en jeu dans la genèse des écoulements, l'absence d'hypothèse *a priori* sur ce fonctionnement est un avantage indéniable de ces modèles.

1.5 Conclusion

La réalisation d'un modèle hydrologique, que ce soit en prévision ou en simulation, repose sur la définition claire d'un objectif de modélisation, sur la disponibilité des données nécessaires, sur les compétences du modélisateur, sur ses connaissances sur les types de modèles et sur son affinité pour un type de modèle en particulier. Cependant, quoique les approches de modélisations puissent être différentes, les modèles sont soumis aux mêmes jugements. Leur robustesse, leur persistance, leur fiabilité sont autant de qualités recherchées pour tout type de modèle hydrologique. Les crues éclair étant particulières, que ce soit par leur faible occurrence ou observation, la faible représentativité des données qui les décrivent, leur caractère fortement non-linéaire, elles demeurent des phénomènes difficiles à cerner. Comme bon nombre de processus naturels, les mécanismes physiques qui gouvernent les processus hydrologiques sous-jacents sont mal connus. Les modèles statistiques, en particulier les réseaux de neurones, ont montré depuis des décennies qu'ils étaient très performants dans la modélisation hydrologique, en particulier sur les crues éclair. Outre leur caractère opérationnel, c'est l'une des principales raisons du choix de ce type de modèle dans le cadre de ce travail. Ce choix est également renforcé par l'expertise développée au niveau du laboratoire du centre CREER de l'IMT Mines Alès depuis les années 2000.

Dans la suite de ce document, nous présentons de manière détaillée les réseaux de neurones ainsi que les grands principes de leur mise en œuvre en tant que modèle hydrologique. Ensuite, nous abordons l'approche du *Deep Learning* appliquée en hydrologie, avec un bref survol des principaux travaux réalisés à partir de cette dernière.

2 Les réseaux de neurones et les modèles profonds pour la modélisation des processus dynamiques non-linéaires

2.1 Modélisation boîte noire/non paramétrique

La notion de boîte noire fait référence à un type de modèle qui représente des processus en n'utilisant que les données issues de ces processus ; ces modèles sont aussi appelés « non paramétriques ». Dans un modèle paramétrique le concepteur postule une structure au modèle, par exemple une équation physique ou une chaîne de relation entre événements, puis estime ou mesure les paramètres ; dans la modélisation non paramétrique, la structure du modèle (par exemple un système linéaire du premier ordre) ainsi que l'estimation de ses paramètres, sont calculées à partir des seules données. De nombreuses méthodes ont été étudiées pour concevoir des modèles de type boîtes noires, elles ont été présentées par Sjöberg *et al.* (1995) et Nerrand *et al.* (1993) dans un cadre conceptuel structuré et cohérent. Parmi ces méthodes, on trouve les réseaux de neurones artificiels, dont nous allons présenter les propriétés principales justifiant leur utilisation dans de nombreux domaines, en particulier pour la modélisation des processus dynamiques non-linéaires.

Concernant les réseaux de neurones, que ce soit pour mieux les comprendre ou les rendre plus performants, plusieurs travaux se sont penchés sur la question de comment intégrer ou extraire de la connaissance dans ou depuis ces modèles. On parle alors selon les méthodes utilisées de boîtes grise, ou de boîtes transparentes. Johannet (2011), citant les travaux de Oussar et Dreyfus (2001), souligne la possibilité de réaliser un modèle de type boîte grise en représentant la part du processus bien connue sous la forme d'un modèle de connaissance (ex. les équations différentielles) et la part moins bien connue avec un modèle de type boîte noire. Par ailleurs, deux types d'approches peuvent permettre d'éclairer les boîtes noires neuronales : (i) l'extraction d'information sur les traitements du modèle à partir des paramètres du modèle (méthode KnoX (Kong-A-Siou *et al.*, 2013); (ii) la visualisation de signaux cachés par le processus (boîtes transparentes, Johannet *et al.* (2008)).

2.2 Les réseaux de neurones

2.2.1 Présentation générale

Introduit par McCulloch et Pitts (1943), le concept de « réseau de neurones formels » fait référence à une fonction mathématique inspirée du fonctionnement des neurones biologiques. Ce concept s'applique à des problèmes de classification et d'identification de fonctions sous une approche purement mathématique. Pour éviter toute confusion avec l'approche dite « neuromimétique », qui vise à mieux comprendre les neurones et systèmes nerveux centraux biologiques, le terme *formel ou artificiel* fut associé à ces formulations mathématiques. Ainsi les termes « réseaux de neurones formels » et « réseaux de neurones artificiels » se substituent sans confusion, tout comme le sont « neurones formels » et « neurones artificiels ». Cependant, par simplicité, dans la suite de ce document nous utiliserons les termes « neurones » et « réseau de neurones » indifféremment, en omettant le terme « artificiel » ou « formel ».

Les réseaux de neurones constituent une famille de modèles non-linéaires paramétrés. Considérés comme des modèles de type *boîte noire*, ils ont la capacité de représenter un système ou fonctionnalité en n'utilisant que les informations d'entrées et sorties du dit système, sans considération approfondie du processus physique ou des relations sous-jacentes. Dans ce chapitre, nous présentons une description de la structure et du fonctionnement des réseaux de neurones dans une première section. Dans une seconde section, nous présentons le type de réseaux qu'on utilise ainsi que leurs propriétés qui justifient leur choix dans notre travail. Dans la troisième section, nous présentons l'approche de leur conception. Ensuite, nous abordons les *réseaux de neurones profonds*, souvent appelés « Deep

Learning ». Puis nous présenterons l'état de l'art sur l'application des réseaux de neurones classiques et **profonds** dans la modélisation des crues éclair. Du fait de la capacité intrinsèque des modèles profonds à extraire des caractéristiques ou à calculer des prétraitements, seront également abordées les approches développées autant pour extraire de l'informations des modèles que pour y introduire des connaissances à *priori* par un autre moyen que l'apprentissage.

2.2.2 Neurones formels

Un neurone formel est un opérateur mathématique qui effectue une transformation généralement non-linéaire, d'une ou plusieurs variables, dites d'entrée en une ou plusieurs variables en sortie. Cette transformation non-linéaire se réalise suivant deux opérations : le calcul d'une somme pondérée des variables d'entrée, appelée **potentiel**, suivi de l'application d'une fonction non-linéaire, appelée **fonction d'évaluation**, délivrant la **sortie**. La connexion entre deux neurones ou la connexion d'une variable à un neurone est assurée par un coefficient modulant l'importance numérique de l'information à propager. Les coefficients de pondération sont appelés, par inspiration biologique, les coefficients synaptiques du neurone, ou paramètres du modèle, ils peuvent avoir des valeurs négatives pour inhiber une valeur, ou positives pour l'activer. Une fois la structure du modèle et les fonctions d'activation posées, l'ensemble des coefficients du réseau de neurones constitue le jeu de paramètre qui le définit. Le potentiel du neurone est noté \mathcal{V} , ses paramètres \mathbf{C} , et sa sortie y . Schématiquement, on présente le neurone formel comme sur la Figure 3.

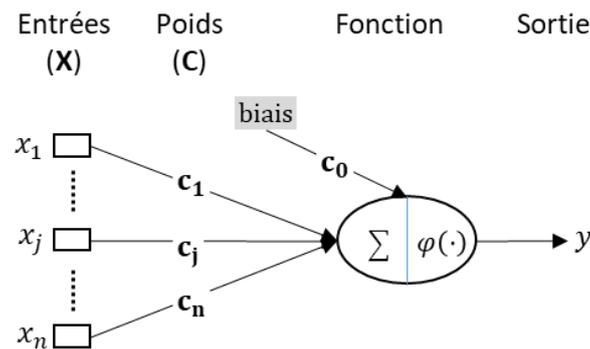


Figure 3: Neurone formel

Le neurone formel est donc un opérateur qui effectue les opérations suivantes.

$$y = \varphi(v) \text{ tel que } v = \sum_{j=0}^n c_j x_j + C_0 \quad (\text{Éq. 7})$$

Avec :

- $\{x_j, 0 \leq i \leq n; i \in \mathbb{N}^+\}$, les composantes du vecteur des variables d'entrée,
- $\{c_j, 0 \leq i \leq n; i \in \mathbb{N}^+\}$, les paramètres du neurone,
- c_0 , le paramètre associé à l'entrée constante ou « biais ». On note que l'entrée constante permet au neurone de délivrer une sortie non nulle lorsque toutes ses entrées sont égales à 0,
- $\varphi(\cdot)$, la fonction d'activation.

Il existe trois grandes catégories de fonctions d'activation :

- Les fonctions d'activations de type seuil : la fonction de Heaviside ou à seuil binaire [0 ou 1].

$$\varphi(v) = \begin{cases} 1 & \text{si } v > 0 \\ 0 & \text{si } v \leq 0 \end{cases} \quad (\text{Éq. 8})$$

La fonction dite « sigmoïde » qui joue le rôle d'un seuil mais qui est dérivable afin de pouvoir réaliser les calculs d'apprentissage sur les réseaux multicouches.

$$\varphi(v) = \frac{1}{1 + e^{-\alpha v}}, \text{ ou } \varphi(v) = \tanh\left(\frac{v}{2}\right) = \frac{1 - e^{-\alpha v}}{1 + e^{-\alpha v}} \quad (\text{Éq. 9})$$

Où α définit la pente de la sigmoïde à l'origine. Pour des valeurs très élevée de cette pente, la fonction sigmoïde se rapproche d'une fonction seuil, c'en est donc un facteur de saturation.

Ces fonctions (seuil et sigmoïdes) sont très utilisées pour des problèmes de classification lorsque le modèle doit fournir une décision.

On note que la dérivée des deux types de sigmoïdes tend vers 0 lorsque le potentiel atteint des valeurs suffisamment grandes (en fonction de la pente α). Cette propriété de toutes les sigmoïdes va entraîner des limites sur l'apprentissage des réseaux profonds (section 2.4.1).

- Les fonctions d'activations de type linéaire : elles appliquent un gain (Éq. 10) et sont souvent utilisée en sortie du modèle.

$$\varphi(v) = av \quad (\text{Éq. 10})$$

où a est le gain.

- Les fonctions non linéaires en partie dérivable, par exemple les fonctions ReLU, dont la version la plus simple est l'identité pour les valeurs positives du potentiel et la valeur nulle pour les potentiels négatifs (Éq.11).

$$\varphi(v) = \begin{cases} 1 & \text{si } v \geq 1 \\ v & \text{si } 0 < v < 1 \\ 0 & \text{si } v < 0 \end{cases} \quad (\text{Éq. 11})$$

2.2.3 Architectures et modèles de réseaux de neurones

Un réseau de neurones est une structure obtenue par l'interconnexion de plusieurs neurones selon un graphe, ou une **architecture**, prédéterminé. Dans le modèle historiquement le plus utilisé : le perceptron multicouche, les neurones sont agencés par couches qui se succèdent de l'entrée vers la sortie (Figure 4). Une couche est un ensemble de neurones qui reçoivent les mêmes informations et qui ne sont pas connectés entre eux ; chaque neurone délivre sa sortie : soit à la couche suivante, soit en sortie du modèle.

On distinguera l'architecture qui est le schéma général du graphe, du **modèle** qui une instance de cette architecture, complètement définie et permettant de calculer des sorties. Le modèle doit donc être complètement explicité : définition des hyperparamètres (nombre de couches, nombres de neurones par couches), types de fonction d'évaluation et hyperparamètres de ces dernières, Le modèle comprend également les valeurs des paramètres liant les neurones (coefficients synaptiques).

L'architecture d'un réseau multicouches comporte une couche des variables d'entrée (qui n'est pas une couche de neurones contrairement à ce qu'écrivent certains auteurs car une entrée seule n'effectue aucune transformation, ni aucun apprentissage), une succession de couches cachées et la couche de sortie du réseau. On note qu'il y a deux catégories de neurones dans un réseau :

- Les neurones cachés dont les sorties constituent des variables d'entrée pour les neurones de la couche suivante, ils sont appelés « cachés » car ils ne sont pas les sorties du réseau : on ne connaît pas quelle doit être la valeur de leur sortie.
- Les neurones de sortie pour lesquels on connaît ce que doit être la valeur désirée, par exemple la sortie d'un processus mesuré sur le terrain, ou la classe de l'objet présenté en entrée.

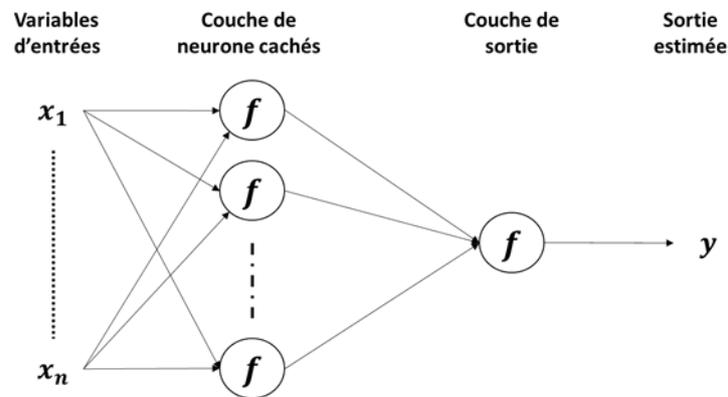


Figure 4: Réseau de neurones multicouche avec une couche cachée et une couche de sortie d'un seul neurone

Si l'on se limite aux modèles **supervisés**, c'est à dire pour lesquels on connaît ce que doit être la valeur désirée des sorties du modèle, on peut se baser sur deux critères pour classer les types de réseaux de neurones :

- **Selon la complexité de la fonction à réaliser** : on trouve des réseaux à une seule couche de neurones non linéaires, qui est alors une couche de sortie ; ce sont les modèles « historiques » (Minsky *et al.*, 2017) dont les performances sont limitées aux fonctions linéaires. Concernant les réseaux inspirés du perceptron (défini à la section 2.2.4) à une couche, mais ayant plusieurs neurones non linéaires, ces derniers permettent d'aborder des fonctions beaucoup plus complexes (Rumelhart et Mc Clelland, 1968), et si l'on a plus d'une couche cachée on parle alors de modèles profonds (Lecun *et al.*, 2015). Nous verrons plus loin dans cette partie à quelle problématique les modèles profonds apportent une solution.
- **Selon la prise en compte de l'aspect temporel du processus** : plusieurs possibilités existent pour prendre en compte le caractère dynamique du système d'étude ; ceci conduit à s'interroger sur le rôle du bruit dans le choix de l'architecture (Nerrand *et al.*, 1993).

Il existe une grande variété d'autres types d'architectures (Voir Fjodor Van Veen, 2016) mais nous nous limitons aux architectures inspirées du perceptron multicouche, les plus utilisés, qui vont nous permettre de poser le cadre de ce travail.

2.2.3.1 Profondeur des architectures

La notion de profondeur est liée au nombre de couches cachées constituant le réseau de neurones (Figure 5). En d'autres termes, plus il y a de couches cachées dans un modèle, plus ce modèle est profond. La couche la plus profonde est celle qui est directement connectée aux variables d'entrée. D'un point de vue fonctionnel, le rôle essentiel de cette couche profonde est de réaliser un prétraitement sur les variables d'entrée pour en extraire l'information pertinente, qui sera transformée dans les couches superficielles, pour conduire par exemple à une décision. Ainsi, des caractéristiques permettant d'effectuer la reconnaissance d'une image sont extraites par les couches

profondes, tandis que l'association de ces caractéristiques au code de l'image est réalisée par les couches superficielles qui effectuent une « association ».

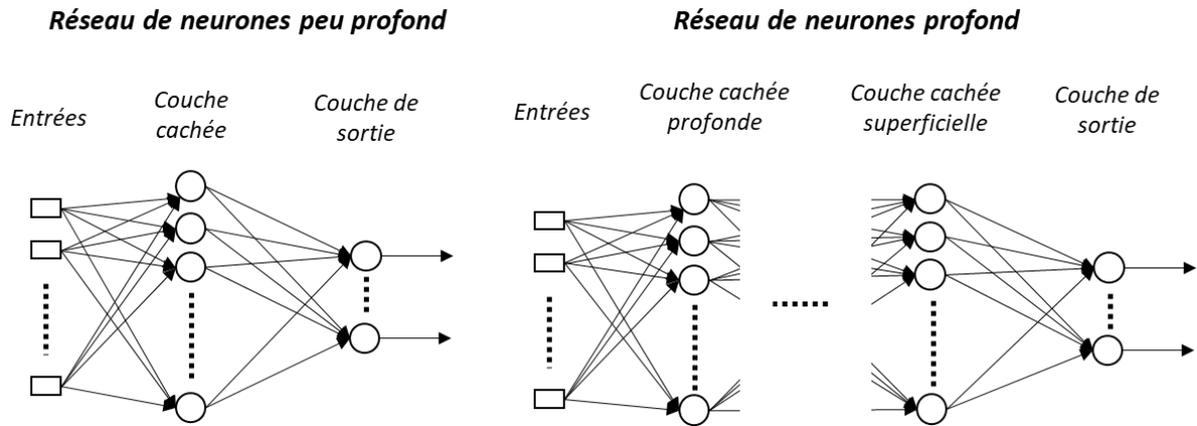


Figure 5: Modèle peu profond et modèle profond.

Un réseau peu-profond contient une seule couche cachée non linéaire. Un réseau profond en contient plusieurs. Les modèles profonds sont souvent appelés « *Deep Learning* », sous-entendant que l'apprentissage est profond, ce qui n'est pas toujours le cas. Cette appellation masque aussi le fait qu'il s'agit de réseaux de neurones, et non d'un autre concept.

2.2.3.2 Complexité des modèles

La complexité du modèle est une notion importante. Elle est généralement définie comme le nombre de paramètres libres du réseau (Gallinari et Cibas, 1999). Ainsi un neurone unique à n_e entrées a une complexité de n_e . Le réseau de neurones de la Figure 4 comportant n_e entrées et n_c neurones cachés a une complexité telle que :

$$c = n_e n_c + n_c = n_c (1 + n_e)$$

2.2.4 Prise en compte de la complexité de la fonction à réaliser : le perceptron multicouche et ses propriétés fondamentales

Le perceptron multicouche (Rumelhart *et al.*, 1986; Bishop, 1995) est un type particulier d'architecture de réseaux de neurones où, dans sa forme originale: (i) il comporte au moins une couche cachée de neurones non-linéaires, une couche de neurones de sortie, et des entrées recevant les variables d'entrée ; (ii) toutes les variables d'entrée sont connectées à tous les éléments de la couche suivante ; (iii) tous les neurones d'une couche sont connectés à tous les neurones de la couche suivante, ainsi de suite jusqu'à la sortie ; (iv) la couche de sortie peut comporter des neurones linéaire ou non-linéaire.

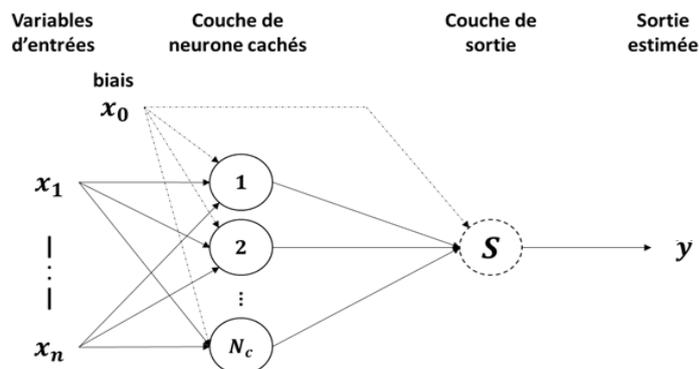


Figure 6: Perceptron à deux couches avec une seule couche cachée

La sortie est calculée comme suit :

$$y = f \left(c_{s0} x_0 + \sum_{j=1}^{N_c} \left[c_{sj} \varphi \left(\sum_{i=0}^n c_{ji} \cdot x_i + x_0 \right) \right] \right) \quad (\text{Éq. 12})$$

Avec :

- y la valeur réalisée par le modèle à partir des variables \mathbf{x} ,
- $f(\cdot)$, la fonction réalisée sur la sortie du réseau de neurones,
- $\varphi(\cdot)$, la fonction d'activation des neurones cachés,
- c_{s0} , le paramètre reliant l'entrée constante (le biais) et la sortie du réseau,
- c_{sj} , le paramètre reliant le neurone caché j à la sortie du réseau,
- c_{ji} , le paramètre reliant la variable d'entrée i au neurone caché j .

Ce type d'architecture, peut être profond ou non. Le perceptron multicouche a été appliqué à des problèmes de classification et d'estimation de fonctions (Trenn, 2008), mais ces fonctionnalités premières ont récemment été étendues à de nombreuses autres fonctionnalités comme les systèmes de recommandation (par exemple du visionnage d'un film), de compression d'information, de cryptage, de conduite automatique, Les perceptrons multicouches sont très populaires dans la modélisation des problèmes non-linéaires. Cette popularité est due à deux propriétés fondamentales : l'approximation universelle et la parcimonie.

2.2.4.1 Approximation universelle

Selon Hornik *et al.* (1989), cette propriété peut s'énoncer comme suit :

« Toute fonction bornée suffisamment régulière peut être approchée uniformément, avec une précision arbitraire, dans un domaine fini de l'espace de ses variables, par un réseau de neurones comportant une couche de neurones cachés en nombre fini, possédant toute la même fonction d'activation bornée, et un neurone de sortie linéaire ».

Elle a été démontrée par Cybenko (1989) et Funahashi (1989). Elle exprime qu'un perceptron à une couche non linéaire et une couche de sortie linéaire peut identifier toute fonction différentiable. C'est cette propriété qui justifie l'utilisation du perceptron multicouche dans l'identification des systèmes dynamiques et non linéaires, tels les systèmes hydro(géo)logiques. Cette propriété n'est cependant qu'un théorème qui démontre l'existence de la solution du problème, elle n'indique pas la méthode selon laquelle les paramètres du modèle-solution seront déterminés.

2.2.4.2 Parcimonie

En 1993, Barron (1993) démontre une seconde propriété très intéressante : **la parcimonie**. Elle stipule qu'un approximateur non-linéaire en ses paramètres nécessite en général moins de paramètres qu'un approximateur linéaire en ses paramètres pour un même problème donné. Ceci s'explique par le fait que ce type de modèles linéaires voient leur nombre de paramètres croître exponentiellement avec l'augmentation des variables, alors que ceci accuse une croissance linéaire pour les modèles non-linéaires en leurs paramètres. Donc, on peut augmenter le nombre de variables du modèle non-linéaire sans le sur-paramétrer pour autant.

2.2.5 Prise en compte du temps

Lorsque les variables d'entrée et de sortie sont des échantillons de signaux temporels, on doit s'interroger sur le type de processus à identifier. Si ce dernier possède un état interne qui dépend du temps, on parle d'un système dynamique. Dans le cas contraire le système est dit statique.

2.2.5.1 Modèle statique

Un modèle statique ne reçoit en entrée que des forçages externes, autrement appelés « variables exogènes ». C'est un filtre à réponse impulsionnelle finie ; cela implique que ses sorties n'évoluent pas lorsque ses entrées exogènes sont constantes (Figure 7).

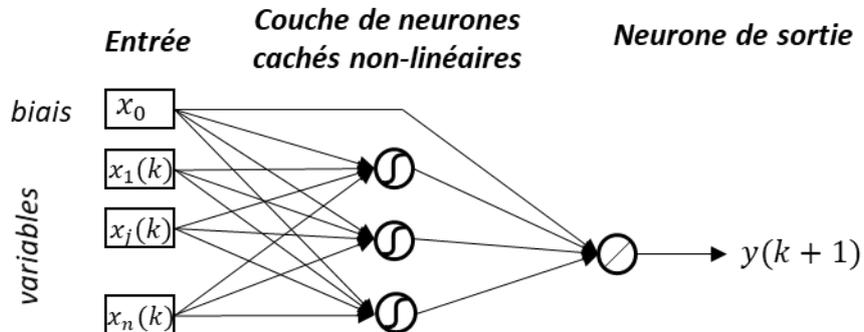


Figure 7: Modèle statique

On peut exprimer la fonction réalisée par ce type de modèle par l'équation 7.

$$y(k) = g_{rn}(\mathbf{x}(k), \dots, \mathbf{x}(k - n_r + 1), \mathbf{C}) \quad (\text{Éq. 13})$$

Avec :

- $y(k)$ la sortie du modèle à l'instant discret k ,
- \mathbf{x} , le vecteur des variables d'entrée, incluant l'entrée constante,
- n_r , la profondeur historique considérée sur les variables d'entrée,
- g_{rn} , la fonction non-linéaire réalisée par le réseau de neurones,
- \mathbf{C} , la matrice des paramètres du modèle.

2.2.5.2 Modèle dynamique

Dans la théorie des systèmes, on représente la dépendance du processus au temps sous forme d'un modèle dit récurrent, qui comporte des bouclages qui renvoient les sorties du modèle vers ses entrées. Cette représentation d'un processus physique dynamique par un modèle est appelée « modèle postulé ». Cette démarche conduit à distinguer les deux types d'entrées qui alimentent le modèle postulé :

- Les entrées exogènes qui traduisent les évolutions des forçages ou des variables d'environnement.
- Les entrées récurrentes qui correspondent aux sorties du modèle aux pas de temps précédents rebouclées vers les entrées ; le nombre de pas de temps pris en compte est appelé l'ordre du modèle.

Le modèle postulé du processus dynamique peut alors être représenté comme en Figure 8.

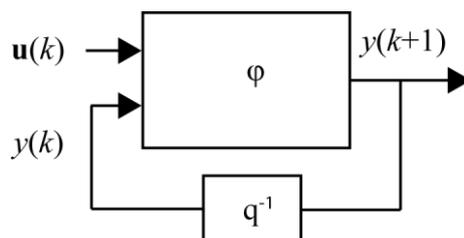


Figure 8: Modèle postulé récurrent.

Le modèle postulé reçoit en entrée les variables exogènes à l'instant k (vecteur $\mathbf{u}(k)$) et la sortie précédente $y(k)$; sa sortie est $y(k + 1)$; q^{-1} est l'opérateur retard.

Le modèle postulé est une proposition de fonctionnement du processus que l'on souhaite représenter (Figure 9), cependant, il est important, dès que l'on s'intéresse à des processus réels, de considérer comment opèrent les bruits et perturbations majeures sur le processus dynamique que l'on souhaite identifier. Nerrand *et al* (1993) ont proposé une méthode intéressante guidant l'explicitation de comment les bruits ou perturbations s'appliquent et permettant d'introduire trois catégories de prédicteurs neuronaux.

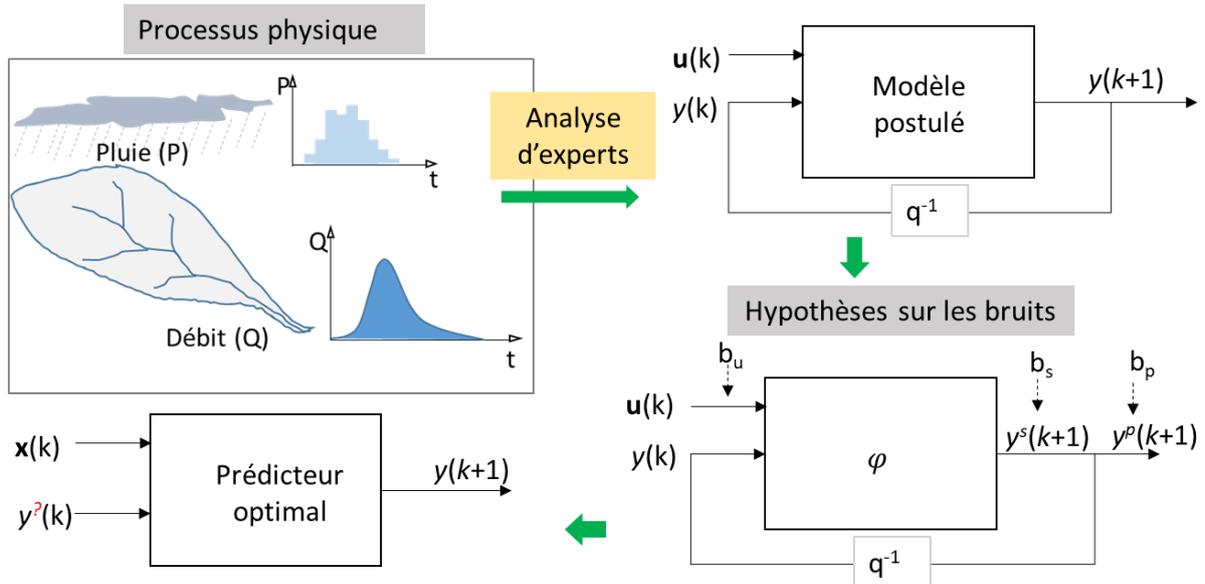


Figure 9: Schéma du passage du processus physique au prédicteur optimal (avec b_u , b_s et b_p : les différentes sources des bruits)

Considérons par exemple les bruits issus des capteurs qui mesurent les variables d'entrée ou de sortie. Il existe principalement deux possibilités. Les perturbations peuvent s'appliquer :

- Sur la sortie du modèle postulé, on parle alors de bruit de sortie, par exemple le capteur de sortie subit une dérive.
- Sur l'état du modèle postulé. La perturbation sur l'état peut provenir de deux origines : des perturbations sur les entrées, qu'elles soient exogènes ou pas ; ou une fonction postulée incorrecte.

Selon que les incertitudes ou bruits opèrent principalement selon une hypothèse ou l'autre, deux types de modèle postulés peuvent être choisis.

2.2.5.3 Hypothèse bruit de sortie

On suppose ici que le bruit intervient de manière prépondérante sur la mesure de la sortie du processus, il peut s'agir par exemple d'un défaut du capteur. Ce bruit est représenté de manière additive à la sortie du modèle postulé comme représenté en Figure 10.a. Dans cette représentation la sortie du modèle postulé s'exprime comme :

$$y^p(k + 1) = y^s(k + 1) + b(k + 1) \quad (\text{Éq. 14})$$

Ou en explicitant $y^s(k + 1)$:

$$y^p(k + 1) = \varphi(y^s(k), \dots, y^s(k - r + 1), \mathbf{u}(k), \dots, \mathbf{u}(k - w + 1)) + b(k + 1) \quad (\text{Éq. 15})$$

Où :

- $\mathbf{u}(k)$, est le vecteur des variables exogènes au temps discret k ,
- $y^s(k + 1)$, est la valeur de l'état du modèle postulé au temps discret $k+1$,
- $y^p(k + 1)$, est la sortie mesurée sur le processus physique à $k+1$,
- q^{-1} , est la notation du retard unité qui appartient à la boucle de rétroaction bouclant les sorties vers les entrées,
- $\varphi(\cdot)$, est la fonction physique effectuée par le processus, représentée dans le modèle postulé,
- $b(k + 1)$, est le bruit non mesuré affectant le processus au temps discret $k+1$.

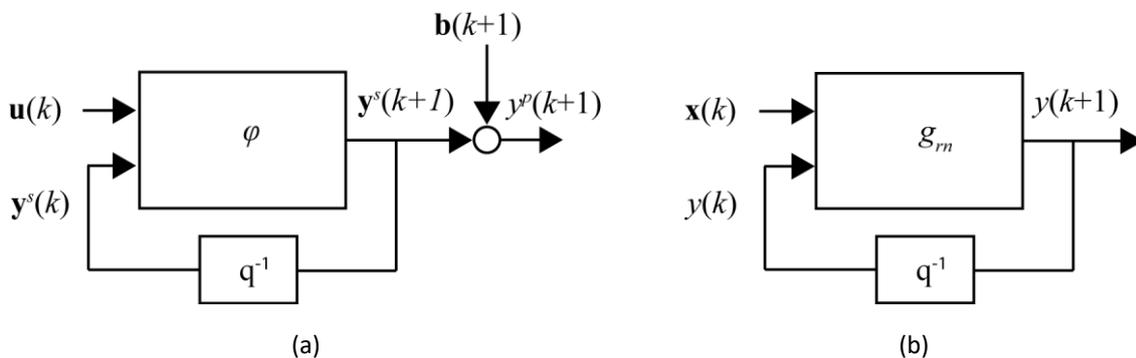


Figure 10 : Modèle postulé (a) et prédicteur associé (b) à l'hypothèse bruit de sortie (Johannet (2011))

Le prédicteur neuronal (Figure 10.b.) fait intervenir les notations suivantes :

- Sa sortie est notée : $y(k + 1)$,
- Le vecteur de ses entrées exogènes est noté $\mathbf{x}(k)$,
- Son entrée récurrente est $y(k)$,
- La fonction non linéaire réalisée par le modèle neuronal est notée g_{rn} .

La relation entrée-sortie du prédicteur s'écrit ainsi :

$$y(k + 1) = g_{rn}(y(k), \dots, y(k - r + 1); \mathbf{x}(k), \mathbf{x}(k - 1) \dots, \mathbf{x}(k - n_r + 1); \mathbf{C}) \quad (\text{Éq. 16})$$

Nerrand *et al.* (1993) ont décrit que le prédicteur neuronal optimal pour représenter le modèle postulé est représenté en Figure 10.b. On note que ce modèle neuronal comporte un bouclage de sa sortie vers son entrée ; en effet, si la sortie mesurée sur le processus est entachée de bruit, il vaut mieux appliquer en entrée : non pas la sortie mesurée, mais la sortie estimée par le modèle lui-même.

Ce type de modèle est dit récurrent. Comme il prend en compte les valeurs précédentes de sa sortie calculée, son horizon de prévision est limité à l'horizon de prévision des variables exogènes. C'est ce type de modèle qui est utilisé pour les prévisions à longue échéance, par exemple pour le changement climatique. Les variables exogènes appliquées peuvent être des prévisions ou des scénarii d'évolution. Si l'on ne dispose pas de prévision des variables exogènes, alors son horizon de prévision est limité à un pas de temps.

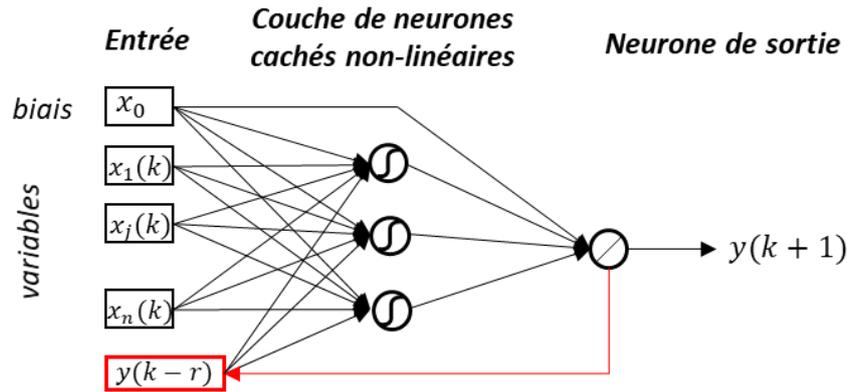


Figure 11: Modèle récurrent

2.2.5.4 Hypothèse bruit d'état

Dans l'hypothèse *bruit d'état*, les perturbations opèrent sur l'état du modèle postulé (Figure 12.a). La relation entrées-sortie du modèle postulé s'écrit alors :

$$y^p(k+1) = \varphi(y^p(k), \dots, y^p(k-r+1), \mathbf{u}(k), \dots, \mathbf{u}(k-w+1)) + b(k+1) \quad (\text{Éq. 17})$$

Où :

- $y^p(k+1)$ est la grandeur observée du processus, sortie de cette représentation du modèle postulé ; $b(k+1)$ est un bruit agissant sur l'état à l'instant $k+1$,
- φ est une fonction,
- $\mathbf{u}(k)$ le vecteur des variables exogènes, r et w sont des entiers positifs.

Cette hypothèse correspond au modèle d'un processus NARX (*Nonlinear AutoRegressive exogenous* ou modèles autorégressifs non-linéaire à entrées exogènes).

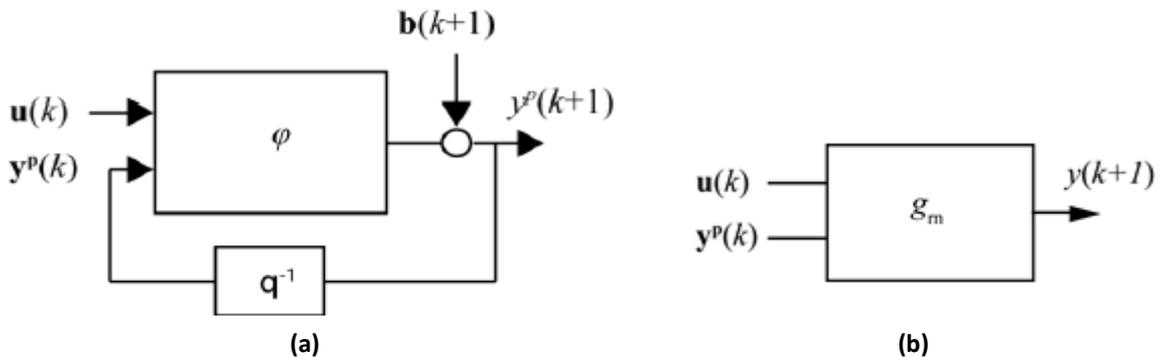


Figure 12 : Modèle postulé (a) et prédicteur associé (b) à l'hypothèse bruit d'état (Johannet (2011))

Compte tenu de l'expression du modèle postulé, comme son état est bruité, il est alors plus intéressant d'appliquer en entrée du prédicteur les variables observées du processus. Le prédicteur neuronal optimal correspondant à cette hypothèse n'est donc pas un modèle récurrent ; ses entrées comportent les variables exogènes et les sorties observées du processus ($y^p(k)$).

Le prédicteur neuronal optimal associé s'écrit donc :

$$y(k+1) = \varphi_{rn}(y^p(k), \dots, y^p(k-w+1); \mathbf{x}(k), \mathbf{x}(k-1) \dots, \mathbf{x}(k-n_r+1); \mathbf{C}) \quad (\text{Éq. 18})$$

Avec :

- $y(k)$, la sortie du modèle à l'instant discret k ,

- \mathbf{x} , le vecteur des variables d'entrée, incluant le biais,
- n_r , la profondeur historique considérée sur les variables d'entrée,
- g_{rn} , la fonction non-linéaire réalisée par le prédicteur,
- \mathbf{C} , la matrice des paramètres du modèle.

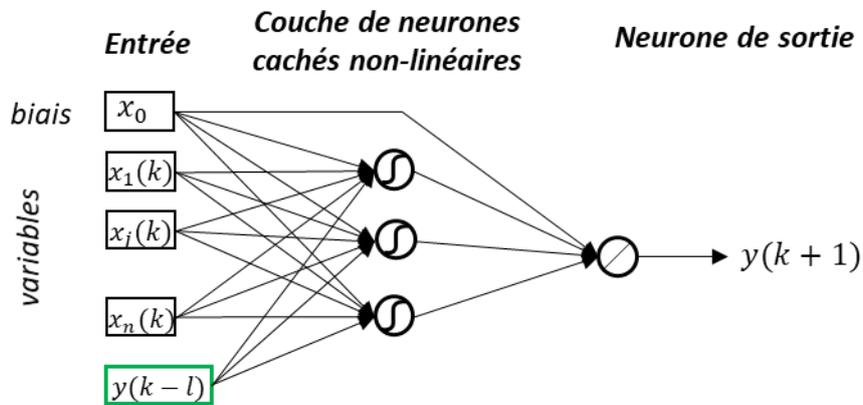


Figure 13: Modèle dirigé

On appelle ce prédicteur « modèle dirigé » car ses entrées sont dirigées par les valeurs observées. Les prédicteurs inspirés de cette hypothèse sont en général très performants puisqu'une grande partie de l'information nécessaire à leur fonctionnement leur est apportée par la variable d'état observée. En effet, le modèle n'utilise les variables exogènes que pour déterminer l'incrément à ajouter aux observations passées de la sortie du processus.

On pourra remarquer que ce modèle est de type statique : il ne reçoit en entrées que des variables exogènes ou mesurées sur le processus. Néanmoins, comme la sortie mesurée sur le processus qui est appliquée en entrée permet de représenter l'impact des variables exogènes pour le passé lointain, on comprend aisément que ce modèle permet d'avoir des vecteurs de variables exogènes de dimension temporelle réduite., en comparaison avec le modèle purement statique. Par ailleurs, il a été montré par Taver *et al.* (2015) que ce modèle est particulièrement adapté pour les bassins non stationnaires car la variable $y^p(k)$ qu'il reçoit en entrée, équivalente une variable d'état, lui apporte une information totalement mise à jour sur l'état du bassin.

2.3 Apprentissage de modèles dynamiques par réseaux de neurones

L'apprentissage d'un réseau de neurones est le calcul de ses paramètres \mathbf{C} . Étant un modèle de type boîte noire, lorsque les paramètres du réseau de neurones sont calculés, ils définissent également la « fonction » réalisée par le modèle. L'apprentissage est réalisé à partir d'une base d'exemple $\{\mathbf{x}^k, \mathbf{y}^k, k \in \mathbb{N}\}$, où \mathbf{x}^k est le vecteur d'entrée du modèle et \mathbf{y}^k le vecteur de sortie.

Dans la littérature, on distingue trois types d'apprentissage : l'apprentissage *supervisé*, l'apprentissage *non-supervisé* et l'apprentissage par *renforcement*. Le premier type suppose l'existence d'un *professeur* qui connaît quelle doit être la sortie \mathbf{y}^k , on l'appelle « sortie désirée », pour tous les exemples présentés en entrée. Concernant le second type, il n'existe pas de professeur qui puisse définir quelle doit être la sortie désirée, comme par exemple dans le cas du *clustering* ; les paramètres du modèle sont alors calculés en fonction, par exemple, d'une distance, ou une fonction de ressemblance, entre la sortie et certains autres exemples. Le troisième type, par *renforcement*, utilise un signal issu d'une fonction capable d'évaluer la satisfaction de la sortie du modèle par rapport à un objectif de performance ; si la sortie du modèle est satisfaisante, l'apprentissage renforce la probabilité

de cette sortie et au contraire si cette sortie n'est pas satisfaisante, l'apprentissage pénalise cette sortie en diminuant sa probabilité d'occurrence (Sutton et Barto, 2018).

Dans le cadre de ce travail, disposant d'informations mesurées sur le processus physique étudié nous nous limiterons à la présentation de l'apprentissage dans le cadre supervisé.

2.3.1 Principes de l'apprentissage statistique supervisé

À partir d'une base de données comportant les exemples décrivant le processus étudié au cours du temps discret k :

- $k = kT$, où T est la période d'échantillonnage que l'on prend égale à 1 sans perdre en généralité, avec $k \in \mathbb{N}^+$, et $k = 1, \dots, N$,
- $\mathbf{x}(k)$ est le vecteur des variables explicatives mesurées à k ,
- $\mathbf{y}(k)$ est le vecteur des sorties du processus au cours du temps.

On note également :

- $\hat{\mathbf{y}}(k) = f(\mathbf{x}(k), \mathbf{C})$, la valeur calculée par le modèle, qui est la valeur estimée de la sortie par le modèle au temps discret k .

Le principe de l'apprentissage supervisé est de minimiser l'erreur de sortie du modèle sur la base de données d'observation.

2.3.1.1 Fonction de coût

La fonction d'erreur, ou de coût, la plus couramment utilisée est la fonction des moindres carrés qui s'écrit comme suit :

$$J(\mathbf{C}(k)) = \frac{1}{2} \sum_{k=1}^N [\mathbf{y}(k) - \hat{\mathbf{y}}(k)]^2 = \sum_{t=1}^N [\mathbf{y}(k) - f(\mathbf{x}(k), \mathbf{C})]^2 \quad (\text{Éq. 19})$$

Où $J(\mathbf{C})$ est la fonction de coût, calculée pour la matrice des paramètres $\mathbf{C}(k)$; $\mathbf{x}(k)$, $\mathbf{y}(k)$ et $\hat{\mathbf{y}}(k)$ tels que définis dans la section précédente.

C'est cette fonction que nous utiliserons pour minimiser de manière itérative les paramètres du modèle. À cette fin on doit définir :

- La sous partie de la base de données qui est utilisée pour minimiser la fonction de coût ; on l'appelle l'ensemble d'apprentissage,
- Le nombre de présentations de la base de données utilisée pour l'apprentissage,
- Un critère d'arrêt de l'apprentissage,
- Une valeur d'initialisation des paramètres,
- La règle de calcul itératif des paramètres à partir de l'initialisation, que l'on appelle règle d'apprentissage.

2.3.1.2 Règles d'apprentissage

Une des raisons du choix des réseaux de neurones est le fait qu'ils effectuent une transformation non-linéaires des entrées vers les sorties. Ce caractère important du modèle fait que la fonction de coût évoquée ci-dessus n'est pas quadratique dans l'espace des paramètres ; il n'existe donc pas, en général

de solution à l'équation 19. Pour obtenir une solution, sans garantie que cette solution soit optimale, les méthodes d'apprentissage sont le plus souvent fondées sur la minimisation de la fonction de coût par une méthode de gradient : $\nabla J(\mathbf{C})$. Selon la nature du problème : ses non-linéarités et le nombre (souvent élevé) des paramètres utilisés, l'espace des paramètres peut comporter de très nombreux minima locaux dont certains s'approchent du minimum global ; ce dernier correspondant à la solution optimale recherchée.

Les méthodes basées sur le gradient sont les plus utilisées ; certaines prennent en compte l'entière de la base d'apprentissage (appelée alors *Gradient Descent*), d'autres procèdent par sous-ensembles, réduisant du même coup le temps de calcul de chaque itération quand le nombre de paramètres est très grand (appelée alors *Stochastic Gradient Descent*), d'autres accélèrent la rapidité de convergence du modèle vers le minimum (règles incluant un *Momentum term*) (Zhang *et al.*, 2018). Toutes les méthodes dérivées du gradient sont dites « du premier ordre » puisqu'elles n'utilisent que la dérivée première pour calculer le gradient.

Les méthodes dites du second ordre utilisent les dérivées secondes et sont réputées plus efficaces (Maier *et al.*, 2010). La méthode de Levenberg-Marquardt, que nous utilisons dans ce travail, est une méthode du second ordre qui utilise une approximation de la matrice des dérivées secondes (matrice Hessienne) à partir du produit des dérivées premières. Nous l'appliquons à l'ensemble de la base d'apprentissage, mais elle peut être appliquée à un sous ensemble glissant, donnant alors un caractère adaptatif au modèle.

- Les méthodes du premier ordre

La méthode du premier ordre la plus simple modifie les paramètres comme suit :

$$\Delta \mathbf{C}(k) = -\mu \nabla J(\mathbf{C}(k)) \quad (\text{Éq. 20})$$

Où $\Delta \mathbf{C}(k)$ est le vecteur d'incrément à appliquer au vecteur des paramètres $\mathbf{C}(k)$; $\mu \in \mathbb{R}$ est appelé le pas du gradient ; il doit être positif et inférieur à un pour favoriser la convergence de la règle.

- Les méthodes du second ordre

Les méthodes du second ordre font intervenir la matrice Hessienne \mathbf{H} , matrice des dérivées secondes de la fonction de coût par rapport aux paramètres. Selon Press *et al.* (1992), si le vecteur courant des paramètres est à une distance suffisamment faible du minimum considéré, noté Ω , on peut faire l'approximation de considérer que la fonction coût $J(\mathbf{C})$ est quadratique.

Ainsi, dans la zone de l'hyperespace des paramètres dont le minimum local est Ω , considérons que l'application de l'incrément $\delta \mathbf{c}$ conduit minimum Ω . On a alors :

$$J(\mathbf{C} + \Delta \mathbf{C}) = J(\Omega) = J(\mathbf{C}) + \nabla J(\mathbf{C})\Delta \mathbf{c} + \frac{1}{2} \Delta \mathbf{C}^T \mathbf{H}(\mathbf{C}) \Delta \mathbf{C} \quad (\text{Éq. 21})$$

Où T, indique la transposition matricielle.

Comme Ω est un minimum de J , les dérivées s'annulent en ce point ; on a donc à l'itération k de l'apprentissage :

$$\frac{\partial J(\Omega)}{\partial \mathbf{C}} = \mathbf{0} = \nabla J(\mathbf{C}(k)) + \mathbf{H}(\mathbf{C}(k))\Delta \mathbf{C}(k) \quad \rightarrow \quad \mathbf{C}(k+1) = \mathbf{C}(k) - \mathbf{H}^{-1} \nabla J(\mathbf{C}(k)) \quad (\text{Éq. 22})$$

On obtient donc une expression de règle de modification des paramètres qui est proche de la règle de gradient (équation 20) :

- L'incrément $\Delta \mathbf{C}(k)$ est calculé comme le produit d'un facteur (\mathbf{H}^{-1} ici) multipliant le terme $-\nabla \mathbf{J}(\mathbf{C}(k))$.
- Le terme $\mathbf{H}^{-1}(k)$ n'est cependant pas constant puisque qu'il dépend des termes de second ordre de la fonction de coût à l'itération k de l'apprentissage

Mais, selon Maier et Dandy (2000), l'utilisation des méthodes du second-ordre pose des difficultés pratiques d'application : (i) le calcul de l'inverse de la matrice hessienne à chaque itération nécessite l'inversion d'une matrice de dimension n_c^2 si n_c est le nombre de paramètres, cela augmente significativement le temps de calcul quand le nombre de paramètres augmente. Ces règles ne sont donc applicables qu'à des modèles comportant un nombre peu élevé de paramètres (environ 500 à 1000) ; (ii) il n'y a aucune garantie que la matrice hessienne soit toujours définie positive, ce qui peut faire évoluer improprement l'algorithme. Ces deux obstacles de la méthode de Newton ont été levés par l'algorithme de Levenberg-Marquardt présenté au niveau de la section 2.3.4.1 (Golden, 1996).

2.3.1.3 Règle de la rétropropagation

La rétropropagation est une règle de gradient de plus grande pente qui peut être déclinée en gradient total (appliquée à tout ou plusieurs des exemples en une fois) ou en gradient partiel (à un seul exemple). Elle propose une solution élégante pour calculer les dérivées de la fonction de coût par rapport aux paramètres dans les réseaux multicouches (Rumelhart *et al.*, 1986). Elle est donc très utilisée en tant que telle, mais aussi dans des règles du second ordre comme la règle de Levenberg Marquard. Comme nous mobiliserons ces calculs un peu plus loin pour expliquer le problème du gradient évanescent, nous la décrivons ici pour un gradient partiel afin de simplifier les notations, et sans perdre en généralité, pour un seul neurone de sortie délivrant la sortie $\hat{y}(k)$, pour l'exemple k . On notera $y_h(k)$ les sorties calculées des neurones cachés, sans ambiguïté, puisqu'il n'existe pas de valeur mesurée pour ces neurones, par définition.

Considérons que l'on présente au réseau l'exemple \mathbf{x} , au temps k : $\mathbf{x}(k)$.

Phase 1 : En premier lieu, et comme pour les autres règles, on calcule la valeur de la sortie du neurone de sortie : $\hat{y}(k)$; c'est ce que l'on appelle la « propagation »

Pour ce faire on calcule successivement tous les potentiels $v_h(k)$, puis toutes les sorties des neurones cachés $y_h(k)$, puis la sortie du réseau $\hat{y}(k)$, depuis l'entrée vers la sortie, on distingue les neurones cachés (indice h , comme *hidden neuron* en anglais), du neurone de sortie, et les entrées d'indice j .

Le réseau calcule donc dans l'ordre :

1. Les potentiels et sorties des neurones cachés : $y_h(k) = f(v_h(k))$,
2. Le potentiel et la sortie du neurone de sortie : $\hat{y}(k) = f(v(k))$,
3. L'incrément des coefficients liés au neurone de sortie par la règle du gradient de plus grande pente.
4. L'incrément des coefficients de tous les neurones cachés.

Phase 2 : la rétropropagation

On peut donc voir la rétropropagation comme la propagation de valeurs depuis la sortie vers les entrées au travers d'un réseau de neurones linéaires.

1. Calcul, en commençant par la sortie et vers l'entrée, Δc_{ij} : incréments aux paramètres c_{ij} liant les neurones j aux neurones i :

- si le paramètre c_j est lié au neurone de sortie :

$$\Delta c_j(k) = -\mu r(k) y_j(k) \quad (\text{Éq. 23})$$

Avec

$$r(k) = (y(k) - \hat{y}(k)) f'(v(k))$$

Où f' désigne la dérivée du potentiel du neurone de sortie par rapport au paramètre c_j

- si le paramètre c_{ij} connecte deux neurones cachés j et i :

$$\Delta c_{ij}(k) = -\mu r_i(k) y_j(k) \quad (\text{Éq. 24})$$

Avec

$$r_i(k) = \sum_{\{h\}} r_h(k) c_{hi} f'(v_i)$$

Où h est l'indice des neurones sur lesquels agit le neurone i , et $f'(v_i)$ désigne la dérivée du potentiel du neurone caché i par rapport au paramètre c_{hj} .

Les deux principales limites à l'utilisation de la rétropropagation sont : (i) elle converge lentement et tout particulièrement lorsque la fonction de coût a des dérivées faibles (plateaux) dans l'hyperespace des paramètres.

La solution est très sensible à l'initialisation, au départ de l'apprentissage.

2.3.1.4 Règle de Levenberg-Marquardt

La méthode de Levenberg-Marquardt (Levenberg, 1944; Marquardt, 1963) est une technique très populaire utilisée dans l'optimisation des problèmes non-linéaires utilisant les fonctions coût des moindres carrés. La règle de Levenberg-Marquardt propose des solutions aux trois difficultés suivantes :

- Le calcul des dérivées secondes qui est long en temps de calcul,
- Le bon conditionnement de la matrice Hessienne,
- La lenteur de convergence de la rétropropagation.

La première considération est la suivante : on remarque en faisant le calcul des dérivées secondes de la fonction de coût quadratique en fonction du vecteur des paramètres, que ce calcul comporte deux termes : le premier qui est le produit des dérivées premières et le second qui inclut les dérivées secondes. Or ce second terme est multiplié par le facteur : $(y(k) - \hat{y}(k))$ qui mesure l'écart entre la sortie du modèle et la sortie optimale. Ainsi, autour du minimum de la fonction de coût que l'on cherche à approcher : Ω , cet écart est par construction le plus faible possible. Il est vraisemblablement dû à des erreurs de mesures ou à des approximations aléatoires et donc les écarts se compensent entre eux, rendant ce terme très faible. Ainsi, lorsque que le modèle est suffisamment proche du minimum optimal, le terme comportant les dérivées secondes est faible et peut-être négligé.

La deuxième considération est pratique : dans le « paysage » que forme la fonction de coût dans l'hyperespace des n_c paramètres il y a de nombreux minima locaux, parfois pour des valeurs de coût élevées. Il est donc habile de commencer une première phase de minimisation par une règle de gradient de plus grande pente (Éq.20), afin de s'approcher de la cuvette contenant le minimum Ω . Puis, dans une seconde phase, lorsque le modèle s'approche du fond de la cuvette et que les dérivées

premières deviennent faibles, il est intéressant d'utiliser les termes d'ordre 2 et de passer alors sur l'incrément de deuxième ordre de l'équation 22.

Pour passer de la première phase à la seconde phase, Levenberg et Marquardt ont proposé d'ajouter un terme rendant la matrice Hessienne diagonale dominante afin d'une part de faciliter son inversion, et d'autre part de contrôler le passage de la première phase à la seconde phase.

On introduit une matrice \mathbf{D} qui s'écrit ainsi :

$$\mathbf{D} = \mathbf{H}\lambda\mathbf{I} \quad (\text{Éq. 25})$$

Avec \mathbf{D} la matrice qui remplace la matrice Hessienne, \mathbf{I} la matrice identité et λ : un hyperparamètre scalaire à ajuster.

En s'inspirant de l'équation 22, les incréments apportés aux paramètres s'expriment alors selon l'équation 26.

$$\mathbf{C}(k + 1) = \mathbf{C}(k) - \mathbf{D}^{-1}\nabla\mathbf{J}(\mathbf{C}) \quad (\text{Éq. 26})$$

Au démarrage de l'apprentissage, l'hyperparamètre λ doit être choisi grand, d'une part, comme indiqué précédemment pour rendre la matrice \mathbf{D} diagonale dominante, et d'autre part car ce paramètre λ intervient dans l'équation 24 sous la forme de son inverse, et λ joue donc un rôle équivalent à celui du pas de gradient de la règle de gradient de plus grande pente. λ^{-1} doit donc être petit. Puis quand l'apprentissage progresse et que le vecteur des paramètres se rapproche du minimum $\mathbf{\Omega}$, on diminue l'hyperparamètre λ , de manière, en diminuant les termes diagonaux de la matrice \mathbf{D} , à prendre en compte les termes du second ordre. Dans cette seconde phase, le rôle de la matrice \mathbf{D}^{-1} peut être vu comme celui d'un pas du gradient variable. La direction de l'incrément est donnée par le gradient et l'amplitude de l'incrément par \mathbf{D}^{-1} .

Le passage d'une phase à l'autre se fait de manière automatique en augmentant ou diminuant λ à chaque itération k de l'apprentissage :

- Si $J(\mathbf{C}(k + 1)) < J(\mathbf{C}(k))$, l'optimisation continue, on s'approche du minimum et on peut diminuer λ : $\lambda(k + 1) = \frac{\lambda(k)}{10}$,
- Si $J(\mathbf{C}(k + 1)) > J(\mathbf{C}(k))$, on s'éloigne du minimum et il faut alors augmenter l'hyperparamètre λ : $\lambda(k + 1) = 10\lambda(k)$.

La règle de Levenberg-Marquardt ajuste ainsi automatiquement la valeur du « pas du gradient » pour passer du premier ordre au second ordre.

Le lecteur aura noté que rien ne permet d'assurer que $\mathbf{\Omega}$ soit le minimum global de la fonction de coût ; c'est pourquoi en pratique nous verrons que nous adoptons une démarche s'inspirant des méthodes « évolutionnistes » en testant plusieurs initialisations des paramètres avant l'apprentissage afin, dans les faits, d'augmenter la possibilité de s'approcher d'un des minimas proches de $\mathbf{\Omega}$.

2.3.2 Stabilité et performance de l'apprentissage : le problème gradient « évanescent »

Si l'on considère l'équation 20, il apparaît clairement que le calcul de l'apprentissage est un processus qui comporte une boucle de rétroaction de manière à minimiser la fonction de coût (Figure 14).

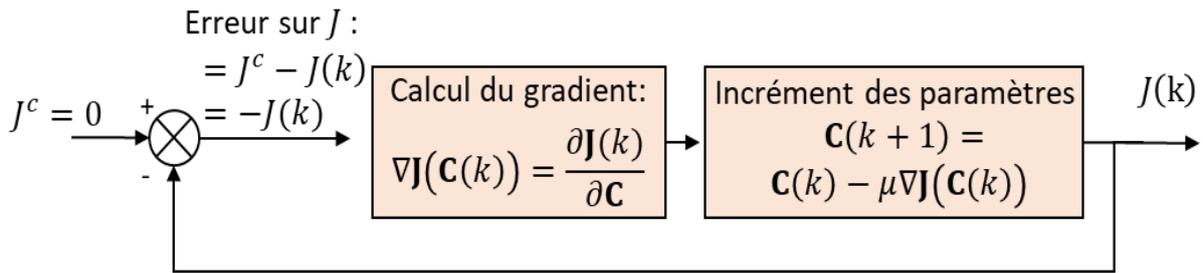


Figure 14: Schéma de l'asservissement de la fonction de coût durant l'apprentissage. La contre-réaction apparaît sous la forme du retour de $J(k)$ vers l'opérateur de différentiation calculant l'erreur- $J(k)$. J^c est la valeur de consigne de la fonction de coût c'est-à-dire 0. $J(k)$ est la valeur de J pour l'exemple k ; cette valeur est renvoyée vers l'entrée par la rétroaction, puis soustraite à la valeur de consigne pour délivrer l'erreur. Les deux blocs qui suivent ont pour fonction de calculer le gradient de l'erreur, puis le nouveau jeu des paramètres et la nouvelle valeur de la fonction de coût.

Comme pour tout système bouclé, il est donc possible que l'apprentissage diverge : la valeur de certains paramètres tend vers l'infini. C'est pour cette raison qu'il est indiqué que la partie proportionnelle du gain de cette boucle (le pas de gradient) soit inférieure à 1, même si cette condition de stabilité n'est pas suffisante et même parfois contredite en pratique.

Par ailleurs, du fait des saturations intrinsèques aux fonctions sigmoïdes (Éq.9), on observe que les incréments appliqués aux paramètres diminuent grandement depuis la sortie vers l'entrée lorsque l'on utilise la règle de la rétropropagation, utilisée également pour calculer les dérivées premières de la fonction de coût par rapport aux paramètres pour la règle de Levenberg-Marquardt. C'est ce que l'on appelle le problème du « gradient évanescent » (*vanishing gradient* en anglais).

Pour expliquer ce phénomène nous pouvons représenter un réseau particulièrement simple composé d'une chaîne de neurones comprenant deux neurones cachés sigmoïdes et un neurone de sortie linéaire. Nous calculons l'incrément appliqué à ses trois paramètres au cours d'un calcul d'apprentissage grâce à la rétropropagation (Figure 14) en utilisant les équations de la section 0 (2.3.1.3 Règle de la rétropropagation).

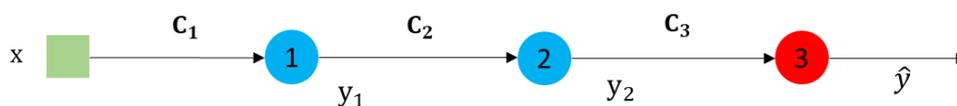


Figure 15: Réseau simple de type « chain rule ». On note x l'entrée, c_i le paramètre arrivant au neurone i et y la sortie du réseau. Les neurones 1 et 2 sont non-linéaires avec une sigmoïde et le neurone de sortie est linéaire ($\hat{y} = v$).

Nous allons calculer les incréments aux trois paramètres c_1, c_2, c_3 en mobilisant les équations 23 et 24.

1. Calcul de la propagation :

$$\hat{y} = c_3 y_2, \text{ (fonction d'évaluation linéaire)}$$

$$y_2 = f(c_2 y_1) ; \text{ donc en remplaçant } y_2 \text{ par son expression : } \hat{y} = c_3 y_2 = c_3 f(c_2 y_1),$$

$$y_1 = f(c_1 x) ; \text{ donc en remplaçant } y_1 \text{ par son expression :}$$

$$\hat{y} = c_3 f(c_2 y_1) = c_3 f(c_2 f(c_1 x)) \quad (\text{Éq. 27})$$

2. Calcul de la rétropropagation

À partir de l'équation 23 pour le neurone de sortie :

$\Delta c_3 = -\mu r y_2$, avec $r = (y(k) - \hat{y}(k))$, car la dérivée de la fonction identité vaut 1,

$\Delta c_2 = -\mu r_2 y_1$, avec $r_2 = r c_2 f'(v_3)$, dans ce cas simple, h ne comprend que le neurone 3 (de sortie),

$\Delta c_1 = -\mu r_1 x$, avec $r_1 = r_2 c_1 f'(v_2)$, si l'on remplace r_2 par son expression : $r_1 = r c_2 f'(v_3) c_1 f'(v_2)$, puis r par son expression :

$$\Delta c_1 = -\mu r c_2 f'(v_3) c_1 f'(v_2) x \quad (\text{Éq. 28})$$

Sur l'équation 28, on note alors que l'incrément au paramètre c_1 comporte en facteur autant de produits de la dérivée de la sigmoïde qu'il y a de couches de neurones sigmoïdes. Or nous avons souligné lors de la présentation de la sigmoïde que sa dérivée pouvait être rapidement très faible lorsque le potentiel atteint de grandes valeurs. Ainsi le gradient est atténué par la dérivée de la sigmoïde autant de fois qu'il traverse de couches sigmoïdes. Ce phénomène est appelé le « gradient évanescent » car c'est en fait la rétropropagation de gradient de la fonction de coût qui cause ce phénomène. Le terme de correction devient donc extrêmement faible et il est impossible d'apprendre sur les couches les plus profondes du réseau.

Ce phénomène a été identifié par de nombreux auteurs dès 1994 (Bengio *et al.*, 1994; Hochreiter et Schmidhuber, 1997; Pascanu *et al.*, 2013). Plusieurs stratégies ont été adoptées pour pallier ce problème : la multiplication du pas du gradient à la traversée de chaque couche linéaire (Bertin et Lambert, 1993; Johannet *et al.*, 1994), l'ajout de neurones linéaires comme « shunts » de chaque couche cachée (Schmidhuber, 1992; Hochreiter et Schmidhuber, 1997; Hinton *et al.*, 2006; Vincent *et al.*, 2008; Cho *et al.*, 2014), l'utilisation de fonctions d'activation non-saturable (Chandar *et al.*, 2019), la réduction du gradient (Pascanu *et al.*, 2013), ou sur la direction du gradient (Kanuparthi *et al.*, 2019). D'autres méthodes utilisent le principe d'orthogonalité et d'unité sur les modèles récurrents profonds (Arjovsky *et al.*, 2015; Jing *et al.*, 2017; Mhammedi *et al.*, 2017; Lezcano-Casado et Martínez-Rubio, 2019; Maduranga *et al.*, 2019). Ces principes veulent que les valeurs propres des matrices de poids entre les couches cachées soient abaissées à l'unité pour éviter l'explosion du gradient, ou augmentées à l'unité pour en éviter l'évanescence. D'autres méthodes utilisent la matrice Hessienne (Martens et Sutskever, 2011), ou des méthodes agissant sur les poids (Doya, 1993; Glorot and Bengio, 2010; Mishkin and Matas, 2016; Xie et al., 2018 ; Jing et al., 2017). D'ailleurs, Pascanu et al (2013) ont réalisé une synthèse très utile à propos de toutes ces méthodes-là.

Dans ce travail, afin de diminuer la sensibilité à l'évanescence du gradient, nous chercherons à éviter de faire l'apprentissage sur les couches les plus profondes en initialisant les paramètres des couches profondes par la corrélation croisée entre les variables d'entrée et la variable de sortie.

2.3.3 Le surajustement et le dilemme biais-variance

Grace à la propriété d'approximation universelle, les réseaux de neurones sont capables d'apprendre toute fonction non-linéaire, pourvu que la base d'apprentissage comprenne l'information nécessaire. Cependant, à l'exception des bases de données issues de sorties de modèles, les bases de données issues de processus réels, d'une part comportent des incertitudes et bruits, qui ne sont pas toujours maîtrisés, et d'autre part ces bases de données ne permettent pas toujours de couvrir tout l'espace des états. En particulier certains processus, comme en météorologie, d'une part ne sont pas mesurables avec les dimensions spatiales et temporelles qui seraient nécessaires, et d'autre part n'ont pas encore permis de recenser les situations les plus rares.

Geman *et al.* (1992) ont exprimé les conséquences de ce constat sous la forme du « dilemme biais-variance » : les bases de données à disposition ne permettent pas toujours d’approcher parfaitement la fonction de régression théorique grâce à l’apprentissage, réalisé à partir des sous ensemble de cas mesurés. D’une part, s’il existe un bruit incompressible dans le jeu de données, ce dernier ne pourra être supprimé, et dans le meilleur des cas la variance de ce bruit se trouvera inchangée dans l’erreur de la fonction de coût après l’apprentissage. D’autre part l’apprentissage conduit le modèle à se spécialiser sur l’expression particulière du bruit de son ensemble d’apprentissage et cette spécialisation est d’autant plus forte que le modèle est complexe ; c’est ce que l’on appelle le « surajustement ».

Le surajustement peut être mesuré sur un ensemble différent de l’ensemble d’apprentissage : l’ensemble de test. L’ensemble de test est un sous ensemble de données non utilisées pour calculer les paramètres ; ce sont donc des données « nouvelles » pour le modèle et qui visent à évaluer la capacité de généralisation du modèle à des situations inconnues.

Geman *et al.* (1992) ont exprimé que l’erreur calculée sur l’ensemble d’apprentissage est un mauvais approximateur de l’erreur calculée sur l’ensemble de test, dite erreur de généralisation, et que la différence entre les deux erreurs augmente avec la complexité du modèle (le nombre de paramètres libres). Or la capacité de généralisation à des exemples inconnus, à partir d’un apprentissage, est le principal intérêt des réseaux de neurones. Il est donc particulièrement important de maximiser la capacité de généralisation. Geman *et al.* (1992) indiquent qu’il existe un optimum de complexité pour lequel la capacité de généralisation est maximale :

- Un modèle de trop faible complexité, trop simple, par exemple un modèle linéaire, ne peut apprendre correctement la fonction demandée, et génère alors un fort « biais »,
- Un modèle trop complexe ne parvient pas à généraliser correctement et délivre une sortie comprenant de grandes oscillations que l’on appelle la « variance »
- Un modèle avec une complexité adaptée parvient à faire une balance entre les deux termes et à optimiser ses capacités de généralisation.

Le dilemme biais-variance est illustré sur la Figure 16 qui présente comment évoluent schématiquement les valeurs du biais et de la variance, au fur et à mesure de l’augmentation de la complexité du modèle.

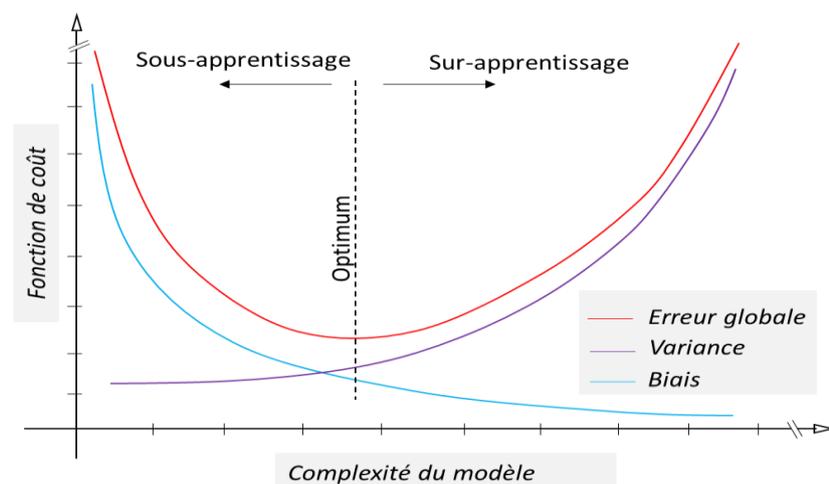


Figure 16: Illustration du dilemme biais variance

Pour maximiser les capacités de généralisation, on a recours à des techniques dites de « régularisation » dont l'objet est de diminuer la variance en généralisation. Elles sont présentées ci-après.

2.3.4 Méthodes de régularisation

Il existe dans la littérature deux catégories de méthodes de régularisation : les méthodes de régularisation dites « passives » qui ne modifient pas les modèles, et celles dites « actives » qui ont un impact sur le modèle lui-même. Pour évaluer les performances du modèle en généralisation, il est nécessaire de diviser l'ensemble de données disponibles en deux sous ensemble : l'ensemble d'apprentissage (déjà évoqué) et l'ensemble de test. Ceci se fait dès le début du travail de conception du modèle car l'ensemble de test ne doit pas être utilisé, en aucune façon, pour la conception du modèle.

2.3.4.1 Méthodes de régularisation passives

- Validation croisée

Cette méthode a été proposée en 1976 par Stone (1976), par la suite Picard and Cook (1984) parvenaient à proposer une extension de celle-ci à l'évaluation de la capacité prédictive des modèles de régression. Elle permet de disposer de la meilleure complexité pour un modèle, donc de disposer du meilleur jeu de paramètres vis-à-vis du problème en question via les données. Elle est appréciée à travers un score dit « score de validation croisée » calculé comme sur l'équation 29. Cette méthode est particulièrement utile lorsque les données sont non stationnaires et qu'il est donc impossible d'évaluer la validité d'un modèle sur un sous-ensemble de la base de données trop réduit.

La validation croisée se réalise en divisant l'ensemble de données constituant l'ensemble d'apprentissage en K sous-ensembles où, tour à tour, l'apprentissage est réalisé sur $K - 1$ sous-ensembles et l'évaluation des performances en validation sur l'ensemble mis de côté. Après chaque apprentissage, on calcule le score de qualité choisi sur l'ensemble de validation. Après les K apprentissage on peut donc calculer un score de qualité sur l'ensemble de la base disponible : le score de validation croisée (Éq.29).

$$S_v = \frac{1}{K} \sum_{j=1}^K S_N(j) \quad (\text{Éq. 29})$$

Avec :

- $S_N(j)$: le critère de qualité calculé sur le sous-ensemble j ,
- S_v , le score de validation croisée.

Kong A Siou *et al.* (2012) ont illustré l'apport de la validation croisée comme méthode de régularisation pour la modélisation de l'hydrosystème karstique du Lez.

- Modèles d'ensemble

Il est bien connu que le perceptron multicouche est sensible à l'initialisation de ses paramètres avant l'apprentissage (Dreyfus, 2004). Ce phénomène est particulièrement important pour la modélisation des crues rapides qui nécessitent un pas de temps assez fin (la demi-heure). Il a été mis en évidence par Darras *et al.* (2014) qu'il était même impossible de sélectionner une initialisation qui permettrait de délivrer des prévisions acceptables pour les quatre événements de crue majeurs de son bassin d'étude, le Bassin de Mialet qui est également notre bassin d'étude. Dans sa thèse, Bornancin-Plantier (2013) a mis en évidence que l'influence sur le Bassin versant d'Anduze, de l'initialisation du modèle

était prépondérante par rapport à celle de tous les autres hyperparamètres (Bornancin-Plantier, 2013). Pour pallier ce problème, Darras *et al.* (2018) ont proposé d'utiliser un « modèle d'ensemble ». Cette approche consiste à considérer plusieurs modèles qui délivrent leurs sorties de manière synchrone, chaque modèle étant initialisé avec une initialisation différente. La dispersion de ces sorties peut être assez importante et Darras *et al.* (2018) ont proposé de calculer la sortie du modèle d'ensemble en prenant à chaque pas de temps la médiane des m membres de l'ensemble :

$$\hat{y}^M(k) = \text{Mediane}_j[\hat{y}^j(k)] \quad (\text{Éq. 30})$$

Avec les mêmes notations que précédemment et "Mediane_j" représentant le calcul de la médiane sur les membres de l'ensemble notés j . La médiane est utilisée car elle est un meilleur estimateur de l'espérance mathématique que la moyenne pour les ensembles ne suivant pas une distribution gaussienne.

Des études spécifiques au bassin considéré doivent être faites pour déterminer le nombre maximum de membres dans l'ensemble. On peut considérer que l'utilisation d'un modèle d'ensemble est une méthode de régularisation car il réduit significativement la variance de la sortie.

2.3.4.2 Méthodes de régularisation actives

Les méthodes de régularisation actives interviennent durant l'apprentissage et conduisent donc à des modèles qui généralisent mieux ou qui sont moins complexes. Elles sont utilisées soit pour contraindre les paramètres, soit pour arrêter l'apprentissage à son optimum. La régularisation active est basée sur l'hypothèse que le sur-apprentissage commence au moment où le nombre de paramètres commence à devenir excessif (Kong A Siou *et al.*, 2012). Cette excessivité, que ce soit en terme numérique ou normatif, est prohibitif pour le modèle. Bien que d'après Bartlett (1997), une des façons d'apprécier l'effet des grandes valeurs des paramètres est la question de saturation des sigmoïdes. Quand la valeur d'un paramètre est trop élevée, la fonction sigmoïde qui l'utilise peut devenir saturée et atteint son potentiel maximal (soit ± 1). Cette fonction n'étant plus dérivable, ne participe plus à l'apprentissage. Les paramètres concernés par ce problème représentent un poids mort dans le jeu de paramètres. L'une des manières de prévenir ce phénomène est de réduire la valeur de ce paramètre. C'est pour cela que Geman *et al.* (1992) introduisait la pénalisation des poids qui consiste à appliquer une pénalité à la fonction coût. Une autre technique très populaire consiste à arrêter l'apprentissage de manière prématurée pour éviter que les paramètres n'aient des valeurs trop importantes. Cette dernière, connue sous le terme « arrêt précoce » a été proposée par Sjöberg *et al.* (1995).

- Modération des poids (*Weight decay*)

Il a été montré que la capacité du modèle à généraliser dépend plus de la norme des paramètres que de leur nombre (Bartlett, 1997). La modération des poids consiste ainsi à limiter l'amplitude des paramètres lors de l'apprentissage. Elle s'opère par ajout d'un terme proportionnel à la norme des paramètres dans la fonction de coût. La fonction de coût s'exprime alors (Éq.31).

$$J^{mp}(\mathbf{C}) = (1 - \gamma) \cdot J + \gamma \|\mathbf{C}\|^2 \quad (\text{Éq. 31})$$

Avec :

- γ , l'hyperparamètre gérant la proportionnalité entre le \mathbf{E}_{MSE} et la valeur des paramètres \mathbf{c} ,
- J , la fonction de coût, comme précédemment définie (Éq.19).
- $\|\mathbf{C}\|$, normes des paramètres du modèle.

La détermination de l'hyperparamètre γ peut se faire aussi bien de manière statistique (MacKay, 1992) que de manière heuristique (Dreyfus *et al.*, 2008). Toutefois, un choix mal adapté de l'hyperparamètre

γ peut dégrader significativement la performance du modèle, comme lui être sans effet. Les auteurs ont bien entendu souligné que cette approche, bien qu'elle ne garantisse pas toujours une amélioration de la performance du modèle, peut éviter le sur-apprentissage pour les modèles dont la complexité, *i.e.* le nombre de paramètres, devient trop importante.

Toukourou *et al.*, (2009) ont comparé la régularisation par modération des poids pour la prévision des crues à Anduze, en conjonction avec la validation croisée, sans noter d'amélioration significative avec l'ajout du terme de modération.

- Arrêt précoce

Proposée par Sjoberg *et al.* (1995), l'arrêt précoce est utilisé pour arrêter l'apprentissage. L'idée sous-jacente consiste à arrêter l'apprentissage lorsque les paramètres commencent à prendre des valeurs trop importantes. Pour ce faire, un sous-ensemble dit « d'arrêt », disjoint de ceux de l'apprentissage et de celui du test/validation, doit être défini. Pour appliquer cette technique, la fonction de coût J est calculée simultanément sur le sous-ensemble d'arrêt et celui de l'apprentissage. Au cours du processus d'apprentissage, on note qu'à partir d'un certain nombre d'itérations, la fonction de coût calculée sur le sous-ensemble d'arrêt se met à augmenter. On interprète ce phénomène comme la manifestation de la difficulté du modèle à généraliser sur un ensemble de données inconnues : l'ensemble d'arrêt. Il convient donc d'arrêter le processus d'apprentissage au minimum de la fonction coût sur l'ensemble d'arrêt et de retenir le jeu de paramètres correspondant à ce minimum. Cette méthode est très utilisée en modélisation hydrologique (Coulibaly *et al.*, 2000). La Figure 17, illustre comment intervient l'arrêt précoce pour arrêter l'apprentissage.

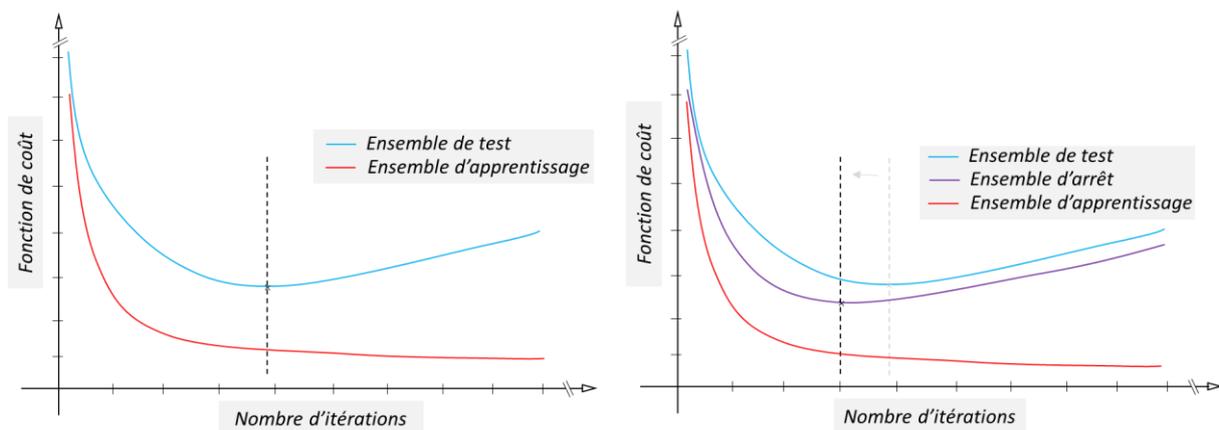


Figure 17: Illustration de l'utilisation de l'arrêt précoce). Gauche (sans l'arrêt-précoce) ; Droite (avec l'arrêt précoce)

Comme le soulignent Bowden *et al.* (2002) et Toukourou *et al.* (2011), le sous-ensemble d'arrêt doit être choisi de telle sorte qu'il soit représentatif de la base de données et de l'objectif poursuivi. Toukourou *et al.* (2011) ont proposé de choisir comme sous-ensemble d'arrêt le sous-ensemble de validation croisée qui produit le meilleur score de validation croisée. Ainsi le couple : « ensemble d'apprentissage et ensemble d'arrêt » est le couple le plus cohérent.

2.3.5 Conception du modèle à réseau de neurones

La mise en œuvre d'un modèle statistique repose sur la disponibilité d'une base de données suffisamment représentative du phénomène à modéliser. Les variables mesurées doivent être disponibles pour représenter l'ensemble de l'espace des états.

2.3.5.1 Sélection des variables

La sélection des variables consiste à choisir parmi les variables disponibles, celles qui ont une explication pertinente à fournir à la description du phénomène. Appelées variables explicatives, elles peuvent être identifiées par estimation de leur relation avec la variable à expliquer, et/ou par l'expertise du modélisateur. De nombreuses techniques de sélection de variables d'entrées sont utilisées, Bowden *et al.* (2005) font appel aux ressorts suivants :

- Connaissances à priori du système faisant intervenir l'expertise du modélisateur (Campolo *et al.*, 1999). Bien que cette méthode soit très appréciée, elle reste tout de même assez subjective. C'est l'une des raisons pour lesquelles Maier et Dandy (1997) suggèrent une combinaison de cette approche à des approches de type statistique.
- Les analyse corrélatoires entre les variables peuvent mettre en évidence les relations temporelles entre les différentes variables. Ces techniques sont très utiles pour mieux connaître les variables significatives des processus dynamiques. Elles ont été largement utilisées par Maier *et Dandy* (1997) et Haugh *et Box* (1977).

$$C_{xy}(k) = \frac{Cov(x_i, y_{i+k})}{\sigma_x \sigma_y} = \frac{\frac{1}{m} \sum_{i=1}^{m-k} (x_i - \bar{x})(y_{i+k} - \bar{y})}{\sigma_x \sigma_y} \quad (\text{Éq. 32})$$

Avec :

- $C_{xy}(k)$, Corrélation simple entre la variable x et y
 - k , le décalage temporel,
 - $Cov(x_i, y_{i+k})$, la covariance entre les variables,
 - $\sigma_x \sigma_y$, les écarts-types des variables x et y ,
 - m , la troncature appliquée sur la série temporelle, souvent égale au tiers de la donnée.
- L'extraction et considération des connaissances apportées par les variables dans le réseau de neurone entraîné. Cette approche se base essentiellement sur l'analyse de sensibilité entre variables explicatives et expliquées. Elle a été largement utilisée par Maier *et al.* (1998), Schleiter *et al.* (1999) et Liong *et al.* (2000).
 - L'approche heuristique (Bowden *et al.*, 2005) consiste en la sélection du meilleur groupe de variables en procédant étape par étape ou par l'utilisation d'un réseau de neurones pour déterminer le meilleur jeu de variables et leur dimensionnement. Cette méthode qui a été utilisée par nombreux auteurs dont (Jain *et al.*, 1999; Tokar *et Johnson*, 1999) est réputée très gourmande en temps de calculs. Parmi les méthodes heuristiques, on compte La validation croisée, qui est une méthode très efficace. Toukourou *et al.*, (2011) ont prouvé qu'on peut sélectionner les variables par la validation croisée.

À côté de ces techniques très connues, Bowden *et al.* (2005) proposent deux autres techniques de sélection des variables lorsque celles-ci sont trop nombreuses pour être testées chacune. Ces techniques sont par conséquent très utiles dans les cas où l'utilisation de l'intégralité des variables disponibles risque de conduire au sur-apprentissage. Elles visent à ne considérer que les variables les plus informatives parmi un groupe de variables qui auraient eu les mêmes informations à apporter au modèle.

- L'utilisation de l'algorithme noté PMI (*Partial Mutual Information*), introduite par Sharma (2000), est fondée sur la théorie de l'information. Elle peut fournir une mesure de la dépendance partielle ou additionnelle que peut apporter un groupe particulier de variables d'entrées au model existant.
- L'utilisation d'une méthode non-supervisée de *clustering*, par exemple les cartes auto-organisées (*SOM : self Organized map*) introduites par Kohonen (1982).

Il convient de souligner que l'utilisation d'une de ces approches n'en n'exclut pas une autre, car plusieurs travaux utilisent une combinaison de plusieurs d'entre elles pour sélectionner les variables. Dans le cadre de ce travail nous mobilisons également plusieurs méthodes : (i) la méthode experte, (ii) l'analyse corrélatoire, et (iii) la validation croisée. On utilise également la corrélation croisée (Éq.27) pour sélectionner l'historique des variables. Cette dernière option, utilisée par Mangin (1984) pour apprécier l'effet mémoire d'un système hydrogéologique, permet d'appréhender non seulement la force de la relation entre les variables, mais également la persistance de cette relation dans le temps.

2.3.5.2 Sélection d'un modèle neuronal

La sélection d'un modèle pertinent s'applique sur une architecture et vise à déterminer tous les hyperparamètres relatifs à cette architecture, à son apprentissage, aux méthodes de régularisation. Contrairement à ce que l'on peut lire dans la littérature, il existe une méthode systématique pour déterminer tous ces hyperparamètres grâce à la validation croisée.

La méthode suivie dans ce travail est la suivante :

- 1) Sélection des variables d'entrées candidates.
- 2) Définition des sous-ensembles d'apprentissage, de test et d'arrêt ; sélection des sous-ensembles de validation croisée. Concernant le choix de l'ensemble de test, et contrairement à ce qui est parfois écrit dans la littérature (« les réseaux de neurones sont prisonniers de leur ensemble d'apprentissage »), nous choisissons l'ensemble le plus emblématique de l'objectif de modélisation en ensemble de test : la crue la plus forte pour la prévision des crues, l'étiage le plus sévère pour la prévision de la ressource (Akil *et al.*, 2021). L'ensemble de test peut comporter plusieurs sous-ensembles, par exemple plusieurs crues, afin d'évaluer plus largement les performances du modèle du fait des non stationnarités des processus et des variables.
- 3) Choix du type d'architecture, basé sur l'expertise du modélisateur et sur les résultats de la sélection. Ce choix dépend de l'objectif poursuivi par rapport aux propriétés connues d'une architecture par rapport à un autre. Par exemple, il est connu qu'un réseau récurrent représente mieux la dynamique du processus qu'un réseau statique. En même temps, un réseau statique permet une expression plus claire de la contribution relative d'une série de variables explicatives à la description d'un phénomène (Saint Fleur *et al.*, 2019).
- 4) Utilisation de la validation croisée pour sélectionner le modèle de plus faible variance. Toukourou *et al.* (2009) ont utilisé la validation croisée pour sélectionner non-seulement la dimension des fenêtres temporelles des variables d'entrées, mais également le nombre de neurones cachés. Cette approche permet de sélectionner tous les hyperparamètres en un seul plan d'expérience si le plan n'est pas trop volumineux. Autrement il faut sélectionner les

hyperparamètres en séquence, par exemple les variables d'entrées et leur profondeur temporelle, puis l'ordre du modèle, puis la complexité.

Dans certaines situations, soit pour des raisons de temps de calcul ou pour mieux s'accorder à l'objectif de la modélisation, on peut être amené à réaliser la validation croisée sur une partie bien distincte de la base de données. Cette démarche est alors une « validation croisée partielle » (Toukourou, 2009). Dans le cadre de cette étude pour les mêmes raisons ci-dessus, en raison du nombre important des éléments de la base de données, seules les gros épisodes pluvieux ont été retenus pour la validation croisée. Ce choix permet non seulement de gagner du temps, mais également de rester le plus proche possible de la nature du problème à résoudre.

- 5) Réaliser un apprentissage sur l'ensemble de la base d'apprentissage, puis évaluer les performances du modèle sur l'ensemble de test.

2.4 Méthodes spécifiques des modèles profonds (*Deep Learning*)

Le *Deep Learning* (Lecun *et al.*, 2015) ou, de manière littérale, l'apprentissage profond, s'applique aux réseaux de neurones profonds. Comme mentionné au niveau de la section (2.2.3.1 *Profondeur des architectures*), les modèles profonds sont des réseaux de neurones. Partant du traitement automatique des images, au traitement du langage et du texte manuscrit, son champ d'application s'élargit. Certains auteurs associent les performances remarquables des modèles profonds à leur aptitude à traiter de manière efficiente d'importantes quantités de données (Deng *et al.*, 2009), à la possibilité d'opérer avec une architecture plus versatile (He *et al.*, 2016) et à la haute performance des algorithmes d'apprentissage qu'ils utilisent (Goyal *et al.*, 2015). Selon Marr (2020) le *Deep Learning* a fait exploser les applications de l'Intelligence artificielle. Shen (2018) souligne d'ailleurs deux principales fonctionnalités explicites du *Deep Learning* : (i) la conception de modèles plus précis, et permettant de traiter de plus grands volumes de données avec une assistance humaine assez faible ; (ii) l'exploration des données, doté d'une capacité de découverte de fonctions au-delà des connaissances et capacités actuelles.

2.4.1 Apprentissage de modèles profonds

L'apprentissage des modèles profonds se révèle beaucoup plus long voire plus complexe que celui d'un réseau de neurones classique pour deux raisons. D'une part nous avons déjà évoqué la difficulté liée au gradient évanescent ; d'autre part le nombre de paramètres augmente exponentiellement avec la profondeur du réseau. C'est la raison pour laquelle il devient plus difficile d'atteindre le minimum global de la fonction de coût, ou un de ses minima dégénérés, lors de l'apprentissage. Il existe des méthodes d'optimisation visant l'optimum global dans l'espace des paramètres (Zhang *et al.*, 2018), mais ces méthodes exigent des moyens de calcul très puissants compte tenu du nombre élevé des paramètres en présence. On a évoqué dans les sections précédentes, que les méthodes d'optimisation les plus utilisées sont celles se basant sur le gradient de la fonction coût. Dans le cas du *Deep Learning*, la plupart des algorithmes d'optimisations utilisés actuellement sont des dérivées de ces méthodes-là. Les plus connues de ces techniques sont le *Stochastic Gradient Descent* (SGD), Adagrad (Duchi *et al.*, 2011), RMSProp (Tieleman and Hinton, 2012), Adam (Kingma et Ba, 2014). Puisqu'elles ne sont pas à l'abri des pièges des minima locaux, il existe des techniques alternatives d'aplanissement de l'hyperplan des paramètres (Zhang *et al.*, 2015; Chaudhari *et al.*, 2017; Zhang et Brand, 2017) leur permettant de converger vers le minimum global. Il existe également des adaptations apportées par diverses études pour limiter le problème du gradient évanescent. Parmi ces solutions, il y a l'utilisation

de la fonction ReLU, ou l'intégration d'options d'oubli et de mémoire variable dans l'architecture du réseau récurrent. Cette dernière solution donna lieu aux types de modèles récurrents appelés LSTM ou « Long-Short-terme-Memory », pourvus de mémoire à plus long terme. L'avantage de la fonction ReLU provient du fait qu'elle réduit les possibilités des neurones profonds de traiter les signaux inhibés, ou les sorties négatives

Dans le cadre de ce travail, c'est la méthode de Levenberg-Marquardt qui est utilisée pour optimiser nos modèles profonds car même si ces modèles sont profonds le nombre de paramètre est compatible avec l'utilisation de cette méthode.

2.5 Application des RN à la prévision hydrologique

Les premières utilisations des réseaux de neurones à la modélisation hydrologique remontent à environ trois décennies. Au cours des années 2000, Maier *et al.* (2000) réalisent un aperçu assez large des différentes approches et avancées réalisées.

2.5.1 Utilisation du modèle Perceptron multicouche (MLP) à la prévision des crues

Considéré comme un puissant outil de régression non-linéaire, les perceptrons multicouches ont été longtemps utilisés pour des problèmes de classifications (Rojas, 2017) et d'identification de fonctions (Trenn, 2008). Il existe une grande quantité de travaux utilisant le perceptron multicouche pour la modélisation hydrologique, incluant l'estimation de variables non mesurées, par exemple l'évapotranspiration (Tabari et Talaee, 2013), la prévision de la sécheresse (Ali *et al.*, 2017) ou la prévision des crues (Araujo *et al.*, 2011) ou les nappes (Kong A Siou *et al.*, 2011 ; Sahu *et al.*, 2020). Ils sont également utilisés dans la prévision des crues éclair (Toukourou *et al.*, 2011 ; Artigue *et al.*, 2012, Darras *et al.*, 2014).

Quant aux crues éclair, à proprement parler, les travaux sont plus récents. Les études qui se penchent sur ces problématiques en utilisant les réseaux de neurones sont peu nombreux et les démarches peu explicites. Maier et Dandy (2000) soulignent que dans la plupart des publications existantes dans ce domaine, le processus de modélisation est en général mal décrit, ce qui risque de compromettre l'optimalité des résultats et la cohérence des comparaisons faites entre approches ou modèles. Piotrowski *et al.* (2006) ont utilisé diverses approches dont un perceptron multicouche (MLP) pour faire de la prévision des crues éclair à partir de l'historicité des pluies et du coefficient de ruissellement. Dans cette étude, une piètre performance a été constatée pour le perceptron, alors que Toukourou *et al.* (2011) soulignent que celle-ci semblait provenir d'un problème de sur-paramétrage puisqu'aucune méthode de régularisation n'y avait été utilisée. Divers autres points soulevés concernent des conclusions qui sont plus optimistes que réalistes. Sahoo *et al.* (2006) et Dinu *et al.* (2017), ayant utilisé diverses approches pour la modélisation statistique des crues éclair, trouvent des résultats qui paraissent excellents. Cependant, le fait que les bassins versants considérés étaient en général assez petits (moins de 70 km²), et que les observations étaient très limitées (quelques épisodes pluvieux), ou les méthodes de régularisations ne sont pas correctement appliquées, il existait un grand risque que ces conclusions optimistes soient biaisées. Toutefois, il existe diverses études dont Toukourou *et al.* (2009, 2011), Johannet (2011), Artigue *et al.* (2011, 2012), Darras *et al.* (2014, 2015) qui utilisent des bases de données assez conséquentes (plus de 20 ans) et bien décrites sur des bassins réputés pour les crues éclair au sud de la France, qui utilisent des méthodes de régularisation très efficaces et très connues (Giustolisi et Laucelli, 2005) appliquées sur des modèles perceptrons multicouches. Les différents résultats obtenus montrent que les réseaux de neurones en particulier les MLP étaient très performants pour la prévision des crues éclair.

2.5.2 Utilisation d'autres réseaux profonds dans la prévision hydrologique

Selon Shen et Lawson (2021), la période 2017 – 2020 représente une époque où le *Deep Learning* confirme ses capacités dans le domaine hydrologique. Allant des systèmes hydrologiques les plus simples, au plus complexes tels que les bassins karstiques ou ceux comportant des glaciers. Les principaux modèles de *Deep Learning* utilisés en hydrologie sont : le Perceptron multicouche (MLP), les réseaux récurrents (RNN), les Long-short-Term-Memory (LSTM) et les réseaux convolutionnels (CNN). Ces modèles diffèrent principalement par leur architecture et leurs hyperparamètres (nombre de couches cachés, le nombre de neurones, le type de fonction d'activation, la méthode d'optimisation etc...).

2.5.2.1 Les Long-Short-Term-Memory (LSTM) et RNN

Les réseaux récurrents (Rumelhart *et al.*, 1986; Werbos, 1988; Elman, 1990) sont des systèmes dynamiques qui ont fait leur preuve dans la modélisation des données séquentielles et des séries temporelles. Selon LeCun *et al.* (2015) leur apprentissage se révèle difficile puisqu'ils sont très sensibles au problème du gradient évanescent. Néanmoins ils ont été appliqués de manière très efficace pour la modélisation des données séquentielles, par exemple à la reconnaissance du langage (Mikolov *et al.*, 2010) ou des textes (Mikolov *et al.*, 2013). Grâce aux divers travaux d'amélioration de leur processus d'apprentissage (Hinton *et al.*, 2012) et de leur architecture (El Hahi et Bengio, 1996; Hochreiter et Schmidhuber, 1997), ces modèles voient leur application toucher la plupart des domaines impliquant des processus dynamiques tels que l'hydrologie (Coulibaly *et al.*, 2001; Chiang *et al.*, 2004).

Conçus pour disposer d'une mémoire sur une plus longue durée que celle des réseaux récurrents classiques (Jaeger, 2012) et d'une fonction d'oubli des informations non utiles (Gers *et al.*, 2000), les Long Short-Term-Memory (Hochreiter et Schmidhuber, 1997) sont des modèles récurrents spécialement conçus pour éviter la difficulté du gradient évanescent. Les *Long-Short-Term-Memory* excellent dans divers domaines tels que la modélisation du langage (Sundermeyer *et al.*, 2012), la reconnaissance vocale (Graves, 2012). Cependant ces modèles sont en général extrêmement complexes et comptent fréquemment des dizaines de couches et des milliers de neurones.

En hydrologie, ils sont considérés comme étant assez performants par plusieurs auteurs. Fang *et al.* (2021) ont utilisé un modèle Long Short-Term-Memory dans la modélisation hydrodynamique. Ils ont constaté que le LSTM pouvait reproduire avec une grande fidélité l'humidité du sol obtenue à travers la mission SMAP (*Soil Moisture Active Passive*) aux Etats-Unis. Zhang *et al.* (2018) l'ont utilisé pour la prévision du niveau d'une nappe pour l'irrigation au niveau du district de Hetao en Chine. Kratzert *et al.* (2018) l'utilisent dans la prévision journalière du débit sur deux bassins versant dont un est sous l'influence des neiges, puis en comparer la performance à celle d'un couplage de deux modèles à base physique SSMAM et Snow-17. Il en résulte que le LSTM surpasse sur tout point le modèle physique qui a longtemps été utilisé dans la région concernée. Li *et al.*, (2021) l'utilisent pour la modélisation des crues intenses sur des bassins versants à Houston au Texas. Comparés avec ceux du modèle à base physique GSSHA (*Gridded Surface Subsurface Hydrologic Analysis*), les résultats du LSTM ont été considérés meilleurs en termes de performance en prévision et en temps de calibration. Le constat est le même pour les épisodes courts, longs ou intenses, par exemple le cas de l'ouragan Harvey en 2017.

Cependant, même si les résultats du LSTM paraissent convaincants, il convient de souligner que peu de comparaisons existent avec les perceptrons multicouches, beaucoup moins complexes. On peut citer à ce propos deux travaux pour lesquels le perceptron multicouche, soit est équivalent au modèle LSTM (Nérot, 2018), soit le surpasse grandement (Jeannin *et al.*, 2021). Compte tenu de l'impact environnemental du numérique le Perceptron multicouche apparaît alors comme bien plus vertueux.

2.5.2.2 Les réseaux convolutionnels (CNN) et les Generative Adversarial Networks (GAN)

Les réseaux convolutionnels (LeCun *et al.*, 1998; Lecun *et al.*, 2015) calculent les convolutions entrées-sortie grâce à la duplication du modèle alimenté par des entrées décalées dans le temps. Les convolutions effectuent une succession de traitements s'apparentant au filtrage ; ils sont en général suivis de couches supplémentaires regroupant les caractéristiques semblables extraites des variables. Ils ont été originellement appliqués au traitement numérique des images (He *et al.*, 2016). Depuis des décennies, leur application touche les traitements des données séquentielles (Sejnowski et Rosenberg, 1987), les problèmes de classification (Cui *et al.*, 2016) et plus récemment ceux de la prévision des séries temporelles (Borovykh *et al.*, 2017; Bai *et al.*, 2018).

Qian *et al.* (2019) ont utilisé un CNN et un GAN pour augmenter la vitesse de calcul d'un modèle à base physique utilisant le SWE (*Shallow Water Equation*) pour la prévision de l'inondation urbaine sur la rivière Austin au Texas. Utilisant les critères d'erreur quadratique moyenne et le ratio du pic, ils ont constaté que ces modèles pouvaient être très efficaces pour la prévision en temps réel des crues du fait de leur temps de calcul bien moindre que celui du modèle à base physique.

Sur la comparaison déjà citée (Jeannin *et al.*, 2021) le modèle CNN a des performances très légèrement inférieures à celles du perceptron multicouche. Cependant sa complexité est largement supérieure ce qui en fait un modèle peu vertueux.

2.5.3 Intercomparaisons des modèles profonds en hydrologie

En balayant différents travaux mettant en comparaison la performance des principales architectures profondes en hydrologie, il apparaît sans surprise que certains auteurs accordent plus d'intérêts à l'architecture qu'ils maîtrisent le mieux, ceci risque de rendre les conclusions peu fiables. Cependant, certains travaux réalisés avec un recul acceptable, ou mobilisant plusieurs équipes spécialisées chacune dans un modèle différent, offrent des comparaisons assez objectives sur les principales architectures. Müller *et al.* (2021) ont analysé la performance d'un LSTM, d'un MLP, un Simple RNN et d'un CNN comme les principales architectures utilisées en *Deep Learning* pour la prévision journalière du niveau des nappes en Californie. Se basant sur les techniques d'optimisation « RBF : *Radial basis functions and Stochastic sampling* » et le « *Gaussian Process* », ils en ont analysé les performances suivant les critères tels que la RMSE, les valeurs finales des paramètres, la convergence des optimisateurs des paramètres, le temps de calcul, performance en validation ainsi que leur performance en prévision. Les résultats obtenus montrent que le MLP, qui était le plus simple des modèles, avait la meilleure performance et le meilleur temps d'optimisation. Le RNN a été le plus mauvais avec notamment un temps de calcul plus long, ainsi que des problèmes de divergences répétés durant la rétropropagation. Le CNN affichait une bonne performance sur la prévision mais il est très gourmand en temps de calcul. Le LSTM a été reconnu le moins robuste avec la plus forte dispersion en sortie. Atmaja et Akagi, (2020) ont utilisé un MLP profond de 5 couches en vue d'adresser le problème d'exigence de puissance de calcul par les modèles profonds sur les tâches telles que le traitement du son, du langage et des images. La comparaison qui en suivait avait impliqué un LSTM et un CNN pour le même niveau de complexité. Les résultats montrent que le MLP était plus performant que le LSTM et CNN qui est en dernière position. À la lumière d'une bonne partie des travaux réalisant ce genre de comparaisons, il peut être constaté que bon nombre de conclusions s'accordent pour reconnaître que les perceptrons méritent leur forte popularité dans la modélisation des phénomènes dynamiques non-linéaires tels que l'hydrologie. Pour plus d'information sur l'utilisation des réseaux de neurones ou du *Deep Learning* en hydrologie, on peut consulter (Maier *et al.*, 2000; Maier *et al.*, 2010) ou Shen *et al.* (2018).

Par ailleurs, vraisemblablement dû au déficit de formation à l'intelligence artificielle dans le domaine de l'hydrologie, les études ne sont pas toujours réalisées rigoureusement, en utilisant les méthodes de régularisation adaptées. Ceci conduit donc soit à une surestimation des résultats (le modèle est sélectionné en prenant en compte l'ensemble de test), soit à une sous-estimation (pas de méthode de régularisation).

Très peu d'études mesurent la qualité des modèles en utilisant un critère adapté à la prévision, par exemple le critère de persistance. Or le critère de Nash peut être excellent sur un modèle qui se contente de délivrer la prévision naïve. Il n'est donc pas adapté à la problématique des crues.

2.6 Extraction des paramètres des modèles profonds avec la méthode *Knowledge eXtraction (KnoX)*

La méthode KnoX (*Knowledge eXtraction*) a été proposée par Kong-A-Siou *et al.* (2013) dans le but de rendre exploitables les valeurs des paramètres des réseaux de neurones. Elle a été utilisée pour extraire des connaissances au sein d'un perceptron multicouche sous forme d'informations physiquement interprétables sur l'hydrodynamique de l'hydrosystème karstique du Lez. Cette méthode a permis d'identifier non seulement les zones contributives du système karstique, mais la proportion de leur contribution à la source du Lez. Ces informations ainsi extraites furent confrontées aux informations obtenues par les techniques de traçage physiques disponibles et furent jugées cohérentes au fonctionnement du système. Cette méthode a été par la suite appliquée par Taver *et al.* (2014), Johannet *et al.* (2014) sur les bassins karstiques du Lez (Hérault, France) et du Baget (Ariège, France) pour déterminer l'origine des eaux et leur temps de transfert dans ces hydrosystèmes. Dans cette étude, nous, Saint Fleur *et al.* (2019, 2020) avons utilisé cette même approche pour extraire les connaissances d'un modèle neuronal appliqué aux crues éclair du bassin versant de Mialet (Gard, France). Cette méthode ainsi que les informations qu'elle a permis d'extraire sont présentées au niveau des *Chapitres V* et *VI* de ce document.

2.7 Interprétation physique des modèles profonds

Les questionnements sur la signification des paramètres internes des réseaux de neurones en général, le *Deep Learning* en particulier remontent à des décennies. Ce sont des questionnements assez légitimes. Dans cette perspective, Chen *et al.* (2018) identifient cinq principales approches d'interprétation des valeurs des paramètres internes de ces modèles : (i) visualisation directe des fonctions d'activations; (ii) Analyse corrélatoire entre données d'activation et les données d'entrées; (iii) conception des réseaux de manière spécifique aux domaines pour l'explication des caractéristiques émergentes; (iv) rétropropagation de la pertinence des caractéristiques; (v) utilisation des connaissances physiques du problème pour exposer les caractéristiques apprises par le modèle profond. Dans le domaine hydrologique, des efforts ont été faits sur l'interprétation des réseaux de neurones, notamment sur l'importance relative des variables explicatives. Cependant, une bonne partie de ces études concernait principalement les réseaux de neurones classiques du type *feedforward*. Ce n'est que récemment que d'autres études ont mis le cap sur l'interprétation de quelques modèles profonds du type LSTM, en hydrologie.

En effet, Li *et al.* (2021) utilisant un modèle LSTM dans la modélisation des crues intenses à Houston au Texas, ont pu, grâce aux valeurs des paramètres interne du modèle, identifier les pluviomètres les plus informatifs parmi les 153 disponibles. Ceci a permis une meilleure sélection des variables, entraînant la réduction de la redondance de l'information, un modèle plus parcimonieux et une convergence plus rapide du modèle. Kratzert *et al.* (2019), à travers l'étude présentée plus haut, ont

analysé le contenu des cellules de mémoire des LSTM qu'ils ont utilisés en vue d'en apprécier la signification physique. Ils ont donc pu remarquer que ces contenus-là corroboraient au comportement hydrodynamique du réservoir ainsi que de l'influence de la saisonnalité sur le bassin versant en question.

Conclusion du chapitre sur les réseaux de neurones et les modèles profonds

Nous avons présenté dans ce chapitre les réseaux de neurones en tant qu'approche de modélisation statistique, avec un focus mis sur les réseaux de neurones profonds. Après avoir défini les concepts clés liés à cette approche, les principales architectures utilisées en hydrologie, ainsi que les démarches de conception d'un modèle depuis la disposition des données jusqu'à l'évaluation du modèle, nous avons évoqué la question de boîte noire ainsi que les possibilités d'accéder ou d'influencer son contenu. Nous avons vu, à travers les différents travaux de modélisation hydrologique depuis le début des années 1990, que les réseaux de neurones ont prouvé qu'ils étaient une alternative très efficace à la modélisation hydrologique physique quand le fonctionnement du bassin est mal connu ou quand les temps de calcul du modèle physique sont trop longs. Cette observation est valable aussi bien pour la prévision des crues éclair, que ce soit dans un bassin versant de surface, karstique ou sous l'influence des glaciers.

Quant au *Deep Learning*, dont l'application en hydrologie est aussi ancienne qu'elle l'est pour les domaines de l'imagerie ou du traitement du langage, les quelques travaux qui ont été menés dans ce domaine prouvent qu'elle est pleine de potentiel pour adresser de nombreux problèmes liés à l'eau. Dans ce chapitre, nous avons survolé les principales applications en hydrologie en général, et sur la prévision des crues particulier. On a pu remarquer qu'une bonne part de ses utilisations ne concerne pas directement la modélisation de la crue proprement dite, mais le plus souvent comme une méthode intermédiaire d'extraction de caractéristiques de la donnée. On a aussi remarqué que les principales architectures utilisées en hydrologie étaient Perceptron multicouche (ou MLP : *MultiLayer Perceptron*), le *long-Short-Term-Memory* (LSTM) et les réseaux récurrents (ou RNN : *Recurrent Neural network*), puis les réseaux de neurones convolutionnels (ou CNN : *Convolutional Neural Networks*).

Malgré la grande performance du *Deep Learning*, sa polyvalence et sa popularité, elle n'échappe pas à des critiques mettant en évidence certaines faiblesses. Selon Zhang *et al.* (2017), les modèles profonds ont parfois tendance à mémoriser des détails et des bruits contenus dans la donnée d'apprentissage. Arplt *et al.* (2017) ont également souligné une préférence de traitement du *Deep Learning* orientée vers les motifs les plus simples de la donnée. D'autres critiques mettent encore plus l'accent sur l'aspect de boîte noire auquel les réseaux de neurones en général n'échappent pas. Cette critique, étant fondée sur le nombre très élevé des paramètres des modèles profonds, n'aurait nullement allégé ce que la notion de boîte noire représente pour un simple réseau de neurones. Dans la pratique, bien qu'il soit considéré comme un obstacle à la capacité de généralisation d'un modèle neuronal, le sur-paramétrage est une situation fréquente et pas forcément préjudiciable aux modèles profonds. Selon certains auteurs, cet avantage leur est attribué principalement grâce à la performance des algorithmes de régularisation utilisés, leur flexibilité, leur polyvalence ainsi que la non-linéarité liée aux fonctions constituant leur couche de neurones profonds. On a également souligné la plupart des approches proposées pour atténuer l'influence du problème du gradient évanescent lié à ces modèles grâce à la fonction ReLU et les propriétés d'oubli des LSTM. L'autre inconvénient majeur auquel les modèles profonds font face est que leur efficacité dépend d'un grand volume de données.

Au niveau du chapitre suivant, nous présentons la zone d'expérimentation, en l'occurrence le bassin versant du Gardon de Mialet, ainsi que la base des données qui seront utilisées pour cette étude. Les modèles que nous utilisons sont des perceptrons ne comportant que trois à quatre couches. Nous nous limitons pour l'instant à ce nombre de couches car les complexités optimales des modèles concernant ce bassin versant indiquent que c'est suffisant.

1 Bassin versant du Gardon à Mialet

Le Gardon de Mialet est un cours d'eau tributaire du Gardon d'Anduze. En considérant la station hydrométrique de Mialet comme exutoire, il draine une surface de 220 km² dont la majeure partie se trouve en Lozère en amont puis dans le Gard en aval (Figure 18).

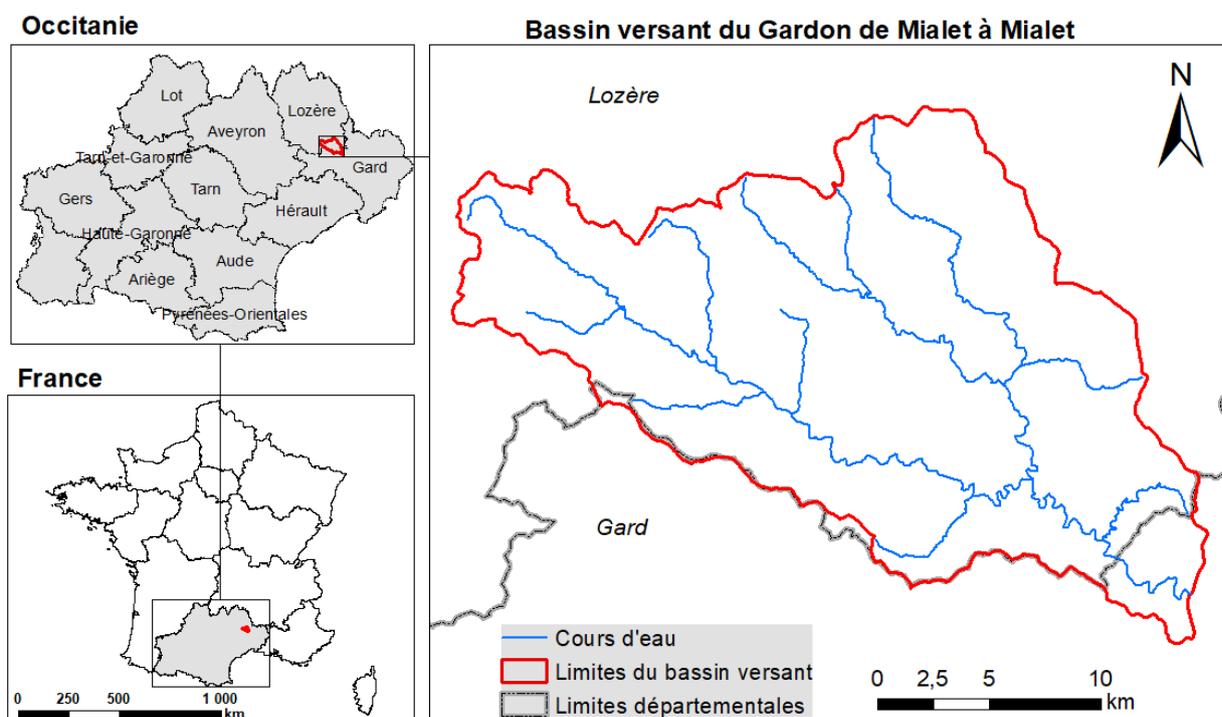


Figure 18: Localisation du bassin versant du Gardon de Mialet (à reproduire soi-même)

Ce cours d'eau résulte de la confluence du Gardon de Sainte-Croix sur sa rive droite et du Gardon de Saint-Martin-de-Lansuscle en rive gauche. Son choix comme bassin versant expérimental repose sur plusieurs raisons dont les principales sont : (i) sa localisation au cœur d'une zone où les épisodes pluvieux intenses sont fréquents, notamment les épisodes cévenols (Voir *Chapitre I*); (ii) l'existence d'une base de données hydrométéorologiques assez riche et sur une durée suffisante (1992 – 2019) ; (iii) l'existence de plusieurs études, dont celle de Artigue (2012), sur ses caractéristiques hydrologiques. Les caractéristiques hydrologiques et climatiques de ce bassin correspondent en tout point à celles qui ont été décrites dans le premier chapitre.

1.1 Caractéristiques topographiques et hydrographiques

D'un point de vue topographique, le gradient d'altitude sur ce bassin est très marqué et les altitudes s'échelonnent entre 170 m à l'exutoire et 1170 m.s.m à son point culminant. Selon une orientation Nord-ouest – Sud-est, il est long 38 km et large de 19 km pour une pente moyenne de 33% (Figure 19). Le rectangle équivalent qui lui correspond est long de 35.26 km et large de 6.24 km.

Compte tenu de la dominance d'une formation géologique faiblement perméable (Ayrat, 2005), d'un substratum pratiquement affleurant, des pentes fortes des versants et la forte pluviosité de la zone, ce bassin versant présente une densité de drainage très élevée. Ceci le rend très réactif aux

sollicitations pluvieuses avec des temps de réponses très courts, compris entre deux et quatre heures (Artigue *et al.*, 2012).

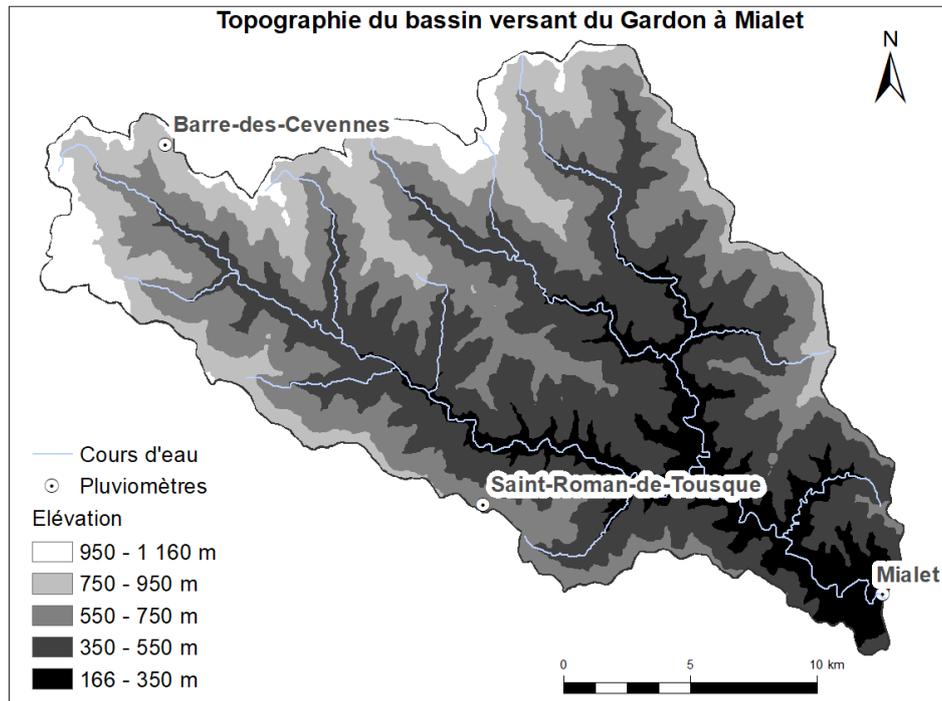


Figure 19: Caractéristiques topographiques et hydrographiques du bassin versant de Gardon de Mialet

1.2 Caractéristiques géologiques

La Figure 20 présente la carte géologique du bassin versant du Gardon de Mialet dont les descriptions sont reportées juste après.

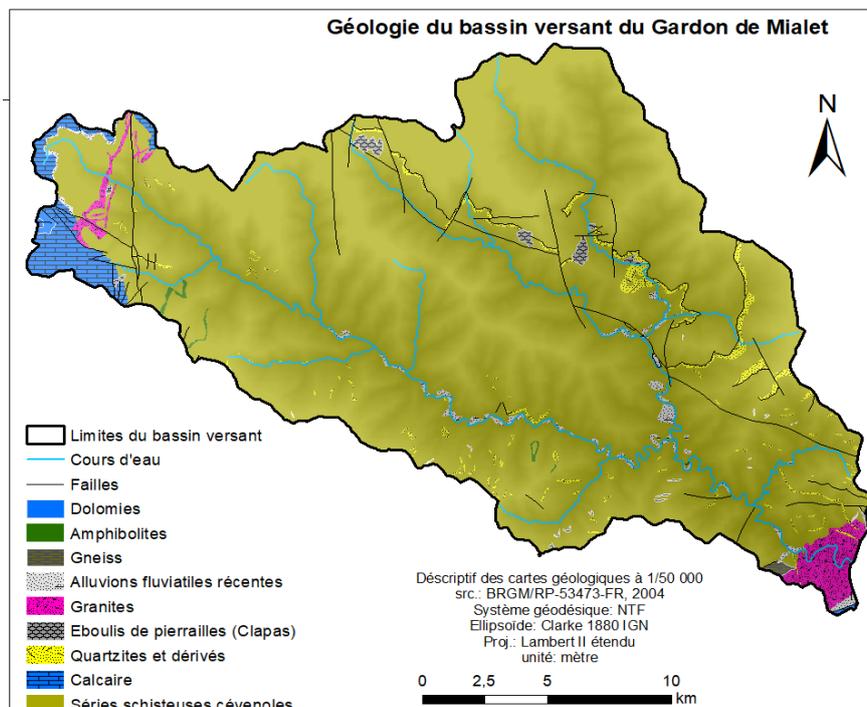


Figure 20: Carte géologique simplifiée du bassin versant de Gardon de Mialet

Ce bassin versant, comme pour la plupart des bassins des Cévennes, est dominé par une formation géologique métamorphique constituée essentiellement de schistes, micaschiste et gneiss à 95%, ainsi qu'une petite partie occupée par une formation sédimentaire par endroits karstique (Calcaires, dolomies et grès) et de formations détritiques du Trias ainsi que des alluvions du Quaternaire granite à l'exutoire (Ayrat, 2005).

L'homogénéité de la lithologie de ce bassin versant est très marquée, surtout avec l'affleurement de ces roches métamorphiques sur plus de 90 % de la surface de ce dernier (Figure 20). Cette configuration donne lieu à un sous-sol peu poreux et pratiquement imperméable. D'un versant à l'autre, le pendage, la schistosité et la fracturation impose, comme décrit dans le premier chapitre de ce manuscrit, la dominance d'un écoulement de de subsurface à la rencontre entre les terrains superficiels altérés et le substratum. Les sols y sont généralement peu profonds et sont repartis entre des rankers, des lithosols et des sols bruns (Ayrat, 2005). Les plus profonds sont aussi les plus rares et se concentrent au fond des vallées et des zones d'accumulations en pied de versant.

1.3 Occupation des sols

La couverture du sol du bassin versant (Figure 21) est majoritairement (80%) occupée par des forêts constituées essentiellement de châtaigniers et de résineux en haut des versants, puis de garrigues boisées et de chênes verts en aval (Ayrat, 2005). Le reste est partagé entre les espaces cultivés et tâches urbaines. Les zones cultivées sont constituées essentiellement de vergers en aval, puis de prairies le long des cours d'eau. Les zones pastorales ainsi que les taches urbaines situées principalement au sommet ne couvrent qu'une infime partie du bassin versant.

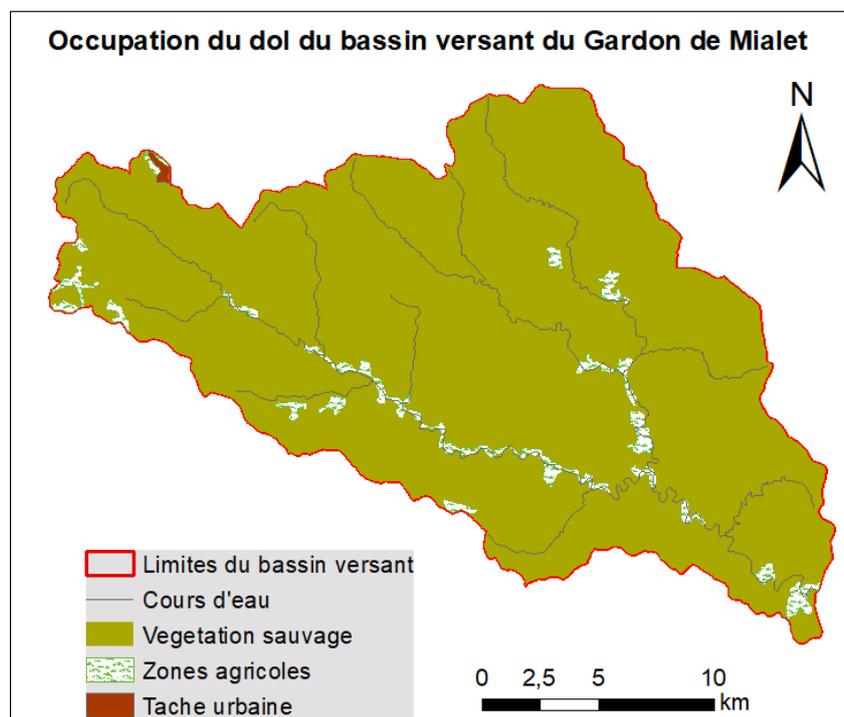


Figure 21: Occupation de l'espace du bassin versant de Gardon de Mialet. Src : Corine Land_cover 2016

2 Données disponibles

Dans cette partie, nous présentons les données hydrométéorologiques qui sont utilisées dans le cadre de ce travail, incluant un bref détail sur les traitements effectués et les approches d'extraction des événements appliquées.

2.1 Présentation de la Base de données

Les données disponibles dans le cadre de ce travail ont été fournies par le SCHAPI dans le cadre du projet BVNE en 2006. Cette base de données qui s'étendait alors de 1992 à 2008 a ensuite été complétée jusqu'en 2019 dans le cadre de ce travail. Elle est constituée d'observations sur trois stations pluviométriques et une station limnimétrique (Figure 19). Les stations pluviométriques sont situées à Barre-des-Cévennes, Saint-Roman-de-Tousque et Mialet, où le poste pluviométrique est doublé d'une station limnimétrique pour laquelle une courbe de tarage existe. Au cours du temps, la fréquence d'échantillonnage a évolué, en passant d'un pas de temps horaire avant 2002, à 5 minutes sur le reste de la période. Lors du travail de Artigue (2012) sur les crues éclair sur ce bassin versant, les données ont été ré-échantillonnées à 30 minutes, conformément aux contraintes opérationnelles du Service de Prévision des Crues.

2.2 Traitement des données

Le traitement des données dans cette étude se fait en vue de s'intéresser dans un premier temps à la problématique de la prévision statistique des crues éclair, puis à l'application du *Deep Learning* sur ces modèles en vue d'intégrer des informations physiques dans le modèle et d'en améliorer l'apprentissage. Les principales opérations portaient donc sur la sélection des épisodes pluvieux les plus importants de cette base de données.

2.2.1 Transformation des hauteurs en débit

Les données de hauteur d'eau ont été utilisées pour déterminer les débits correspondants en appliquant les courbes de tarages fournies par le SPC pour la station de débit à Mialet (Figure 22).

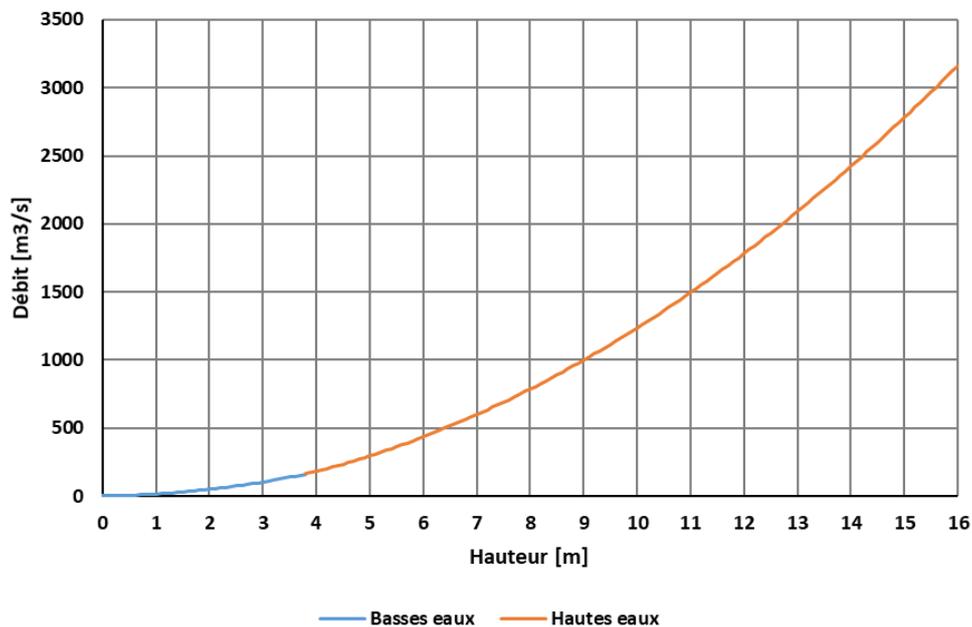


Figure 22: Courbe de tarage de la station hydrométrique de Mialet.

Cette courbe de tarage comporte deux parties suivant la hauteur d'eau :

- En dessous de $Heau = 3.83$ m (basses eaux) :

$$Q_e = 10.024 * h^2 + 3.9707 * h + 2.17348$$

- À partir de $Heau = 3.83$ m (hautes eaux) :

$$Q_e = 12.154 * h^2 + 5.470 * h - 34.59$$

Où Q_e est le débit et h la hauteur d'eau.

Comme la plupart des courbes de tarages, le débit est une fonction quadratique de la hauteur d'eau. Ceci signifie que les incertitudes sur les hauteurs d'eau sont amplifiées lors de la conversion en débit.

2.2.2 Rééchantillonnage des données

Pour ramener les données horaires en données semi-horaires, Artigue (2012) avait adopté une approche d'interpolation qui lui permettait de préserver le cumul. C'est cette méthode qui a été utilisée. Pour les données de hauteurs d'eau, l'application de l'interpolation par moyenne simple convenait très bien. Quant aux données collectées à cinq minutes d'intervalle, le principe du cumul a été adopté sur les pluies tout en positionnant l'instant d'échantillonnage retenu à la fin des six unités écoulées. Il importe de souligner que ces rééchantillonnages, quoi qu'ils soient utiles, ne sont pas sans impact sur la morphologie des événements. En effet, il existe un risque réel de modification des pics des événements aussi bien au niveau de leur amplitude que de leur superposition dans le temps dont un risque de décalage artificiel jusqu'à 30 minutes. Face à ce risque, qui aurait eu une plus grande répercussion sur les informations de débits, Artigue (2012) optait pour la conservation des valeurs maximales sur les 30 dernières minutes à la place des moyennes qui les écraseraient. Ce choix, malgré le fait qu'il comporte le risque d'un décalage du pic d'au plus 25 minutes, permet d'une part de garder une certaine cohérence avec l'objectif de prévision des événements intenses que sont les crues éclair et d'autre part de garantir la causalité du processus d'échantillonnage. Il est en effet indispensable que le calcul de prévision puisse être effectué sans nécessiter les informations des données futures.

2.2.3 Approche d'extraction et de sélection des événements

Dans le cadre de ce travail, une approche de modélisation événementielle est adoptée. Il faut donc extraire les événements. Pour une partie de cette étude, la base de données de 58 événements extraits par Artigue (2021) est utilisée. Cette base de données s'étendait jusqu'en 2008 seulement. Pour une autre partie de ces travaux, la période 2009-2019 a été valorisée. Le critère choisi est le dépassement d'un cumul de 120mm en 48h sur au moins deux des trois pluviomètres, à condition que l'ensemble des données soient disponibles sur toutes les stations pour l'épisode. Ce cumul est supérieur à celui choisi par Artigue (2012) afin de limiter la quantité d'événements. En conservant le seuil de 100mm (Artigue, 2012), on aurait disposé de plus de 100 événements, ce qui aurait constitué un problème pour les temps de calculs et n'aurait probablement pas été nécessaire à optimiser le modèle. Nous obtenons ainsi, par le plus grand des hasards, 58 événements aussi. Nous avons considéré qu'un événement débutait avant les premiers millimètres de pluies et se terminait à la fin de la partie rapide de la courbe de décrue. Une description rapide de ces événements est disponible au niveau de l'Annexe 1, les hydrogrammes correspondants sont regroupés par décade et suivant qu'il fasse partie ou non des sous-ensembles de la validation croisée au niveau des Annexe 2 à Annexe 7.

Sur le tableau résumant les différents épisodes sélectionnés (Annexe 1), les trois couleurs marquent trois périodes de la base de données d'où sont tirés les épisodes : pas de temps horaire converti en 30 minutes en vert (1992-2002), pas de temps de 5 minutes converti en 30 minutes jusqu'en 2008 en jaune (données déjà extraites par Artigue, 2012), puis jusqu'en 2019 en gris avec cette même résolution. Outre l'identification des événements, les numérotations ici jouent également un rôle d'archive en vue de respecter les premiers codes utilisés depuis les travaux de Ayrat (2005).

Selon les données enregistrées sur les trois pluviomètres du bassin versant sur cette période, on constate que les épisodes pluvieux cumulaient des précipitations qui dépassaient les 600 mm, si on considère le cas de la station de Saint-Roman-de-Tousque, en novembre 2011. Bien que cet épisode

fût le plus pluvieux, il n'a pas donné lieu à la crue la plus intense. Cette dernière fut provoquée par l'épisode de septembre 2002 avec 832 m³/s. La base de données ainsi ré-échantillonnée est présentée dans l'Annexe 1 et le Tableau 1 en donne les principales caractéristiques.

Tableau 1 : Principales caractéristiques de la base de données (BDC est Barre-des-Cévennes, SRDT est Saint-Roman-de-Tousque)

58 événements	Cumul à BDC (mm)	Cumul à SRDT (mm)	Cumul à Mialet (mm)	Débit de pointe (m ³ /s)	Débit spécifique (m ³ /s/km ²)	Date
Maximum (pluie)	508	606	429	430	1,95	02/11/2011
Maximum (débit)	70	303	346	833	3,79	08/09/2002
Maximum	508	606	429	833	3,79	
Médiane	166	203	175	156	0,71	
Moyenne	187	220	197	200	0,91	N/A
Minimum	0	88,5	56	24	0,11	
Ecart-Type	92	83	86	154	0,70	

2.2.4 Description des événements retenus

Au niveau de cette section, nous présentons les caractéristiques de l'ensemble des épisodes de la base de données sur 1992-2019, en commençant par leur durée, ensuite leurs intensités maximales, les cumuls totaux pour finir à la distribution des débits spécifiques. La base de données sur 1992-2008 est déjà extensivement décrite dans Artigue (2012).

2.2.4.1 Durée

Les épisodes sont majoritairement brefs puisque la plupart d'entre eux durent moins de cinq jours. La classe la plus représentée est d'ailleurs celle dont la durée est comprise entre deux et trois jours (Figure 23). Ces observations sont cohérentes avec la brièveté des réponses aux pluies intenses de ce bassin.

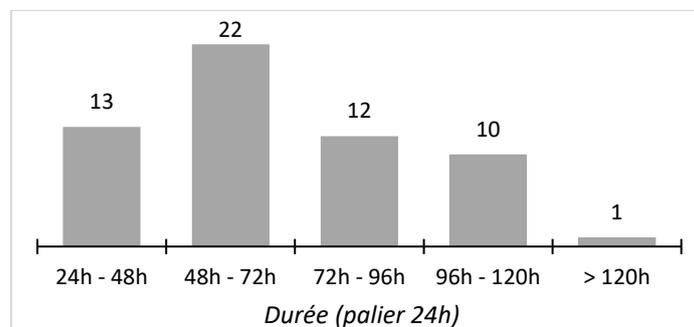


Figure 23: Durée des épisodes pluvieux retenus

2.2.4.2 Cumuls

La répartition des cumuls de pluie des événements sélectionnés sont présentés sur la Figure 24.

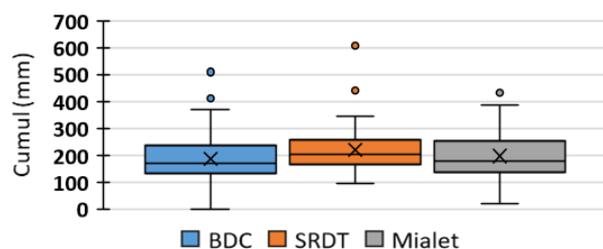


Figure 24: Cumul des pluies des différents événements sur les trois pluviomètres sous formes de boîtes à moustache incluant la médiane ; BDC est Barre-des-Cévennes, SRDT est Saint-Roman-de-Tousque

Le critère de sélection adopté a permis de ne disposer que des événements dont le cumul total moyen minimal était assez important pour susciter une interrogation sur l'ampleur de la réponse chez l'utilisateur final.

On peut remarquer que sur les 58 événements sélectionnés, un seul ne présente aucune donnée enregistrée sur le pluviomètre qui se trouve à Barre-des-Cévennes. Il s'agit de celui du 3 novembre 2017. Ceci conforte le constat selon lequel la variabilité spatiale des pluies peut être très importante. Hormis cette exception, on voit que la station de Saint-Roman-de-Tousque est celle qui enregistre le plus haut cumul de pluie et montre la dispersion la moins importante des trois stations. Elle est suivie par Barre-des-Cévennes puis Mialet en termes de cumul, mais la dispersion des données sur celles-ci reste quasiment identique. Le cumul médian est au moins 170 mm sur le pluviomètre le moins pluvieux. Deux épisodes pluvieux sortent du lot par le haut avec des cumuls situant entre 400 et 606 mm. Il s'agit de ceux du 2 novembre 1994 (411 mm sur Barre-des-Cévennes), du 20 janvier 1996 (439 mm sur Saint-Roman-de-Tousque) et du 2 novembre 2011.

2.2.4.3 Intensités

Les intensités moyennes maximales semi-horaires entre les épisodes, représentées par les croix (X) sur la Figure 25 varient entre 10 et 20 mm. Cette grandeur qui traduit l'intensité des pluies qui s'abattent sur la zone est naturellement très variable. Sur la même figure, on peut identifier quelques intensités semi-horaires exceptionnelles entre 50 et 80 mm sur les stations Saint-Roman-de-Tousque et Mialet respectivement. Sur les trois pluviomètres, pris d'amont vers l'aval et en dehors des valeurs exceptionnelles susmentionnées, les valeurs maximales des 25% des intensités les plus élevées se rapprochent respectivement de 20, de 30 et de 50 mm en 30 minutes. Quant aux intensités médianes, elles ne dépassent pas 12 mm en 30 minutes pour l'ensemble des trois pluviomètres. La station de Saint-Roman-de-Tousque est celle qui enregistre l'intensité minimale la plus importante, bien que celle-ci ne dépasse pas 11,5 mm. Si les intensités croissent de l'amont vers l'aval, on a bien constaté que c'est en partie l'inverse pour les cumuls pluviométriques enregistrés. (Figure 24). En effet, Mialet est la station montrant le plus faible cumul maximal et pourtant, elle enregistre l'intensité pluviométrique la plus forte de la base de données.

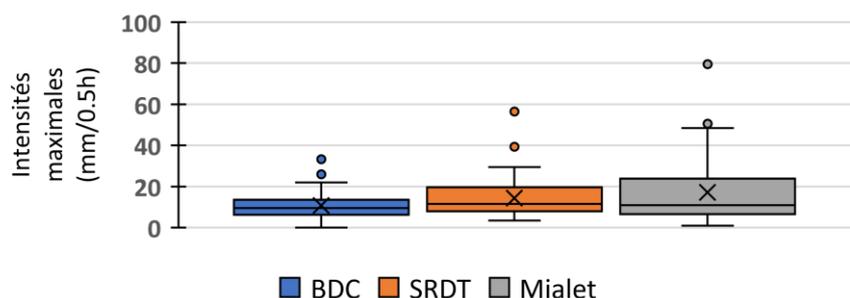


Figure 25: Intensités maximales des différents épisodes pluvieux sur les trois pluviomètres sous forme de boîte à moustache; BDC est Barre-des-Cévennes, SRDT est Saint-Roman-de-Tousque

2.2.4.4 Débits

Quant aux débits de pointe (Figure 26.a.), on peut noter que la proportion des débits en dessous de 200 m³/s est plus importante. Le débit de pointe médian des épisodes sélectionnés est à peu près 150 m³/s, celui-ci ne vaut donc que les trois quarts du débit de pointe moyen. L'analyse des débits spécifiques (Figure 26.b.) permet de mettre en évidence la réaction du bassin versant en termes de capacité de production de débit par unité de surface. En effet, on peut constater qu'environ un tiers des épisodes pluvieux font état d'une production au-delà d'un m³/s/km². On considère généralement

que les crues sont particulièrement intenses lorsqu'elles produisent plus d'un mètre cube par seconde et par kilomètre carré. Cela concerne 19 épisodes sur 58.

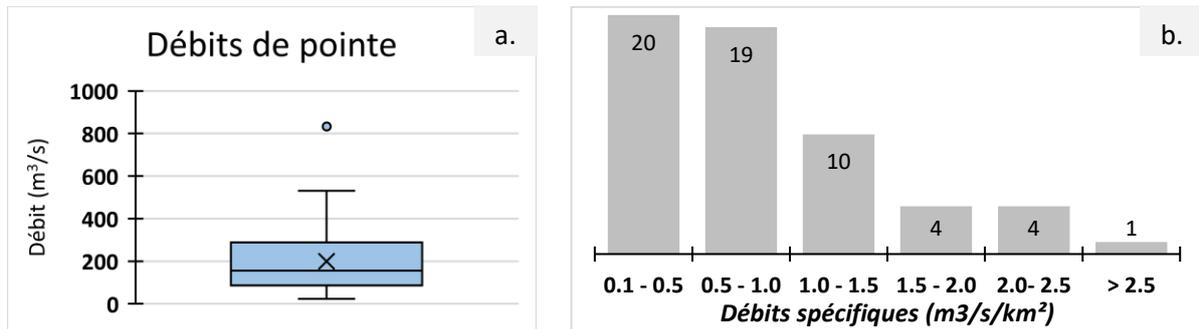


Figure 26: Débits de pointe (a) et distribution des débits spécifiques (b) des différents épisodes pluvieux

2.2.5 Événements type

Les événements sélectionnés sont de formes assez variées. Cette diversité de formes est tout aussi présente au niveau des pluies que sur la réponse hydrologique du bassin versant. Les autres hydrogrammes sont disposés au niveau des *Annexe 2* à *Annexe 7*.

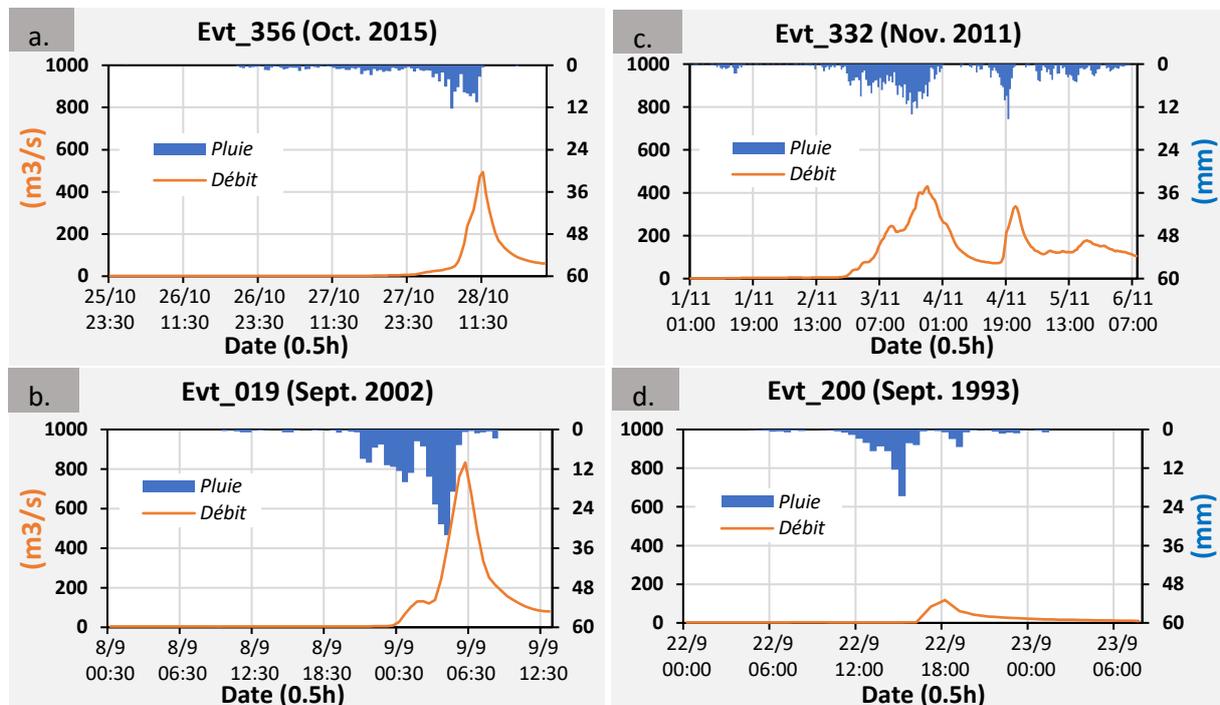


Figure 27: Hydrogrammes de quelques épisodes de la base de données

Parmi les 58 épisodes retenus, la Figure 27 présente quatre événements correspondant aux principales situations hydrologiques observées dans la base de données. Sur ces événements, on observe l'une ou l'autre des situations suivantes : (i) Une faible pluie qui génère un fort débit de pointe (Figure 27.a) ; (ii) Une pluie exceptionnelle grande avec un débit de pointe exceptionnellement grand, comme celui de septembre 2002 (Figure 27.b) ; (iii) Un épisode de pluie avec plusieurs pics de crues (Figure 27.c) ; (iv) Un épisode comportant une forte pluie et un faible débit de pointe (Figure 27.d). A travers ces différents cas, on peut constater qu'une forte pluie ne génère pas automatiquement une grande crue et vice-versa, dépendamment de l'état hydrique au moment où cette dernière se produit. Parallèlement, deux pics de pluies ne donnent pas forcément deux pics de crues, que ce soit en termes d'amplitude ou de positionnement temporel. Ces quelques exemples, quoique limités, se révèlent déjà

suffisants pour se faire une idée du caractère non-linéaire de la relation entre les pluies et le débit sur ce bassin versant.

2.2.6 Relations et corrélations entre variables

2.2.6.1 Test de double cumul et analyse corrélatoire linéaire simple

Afin de détecter d'éventuelles erreurs de mesure pluviométriques, un test des doubles cumuls a été réalisé et présenté au niveau de la Figure 28.

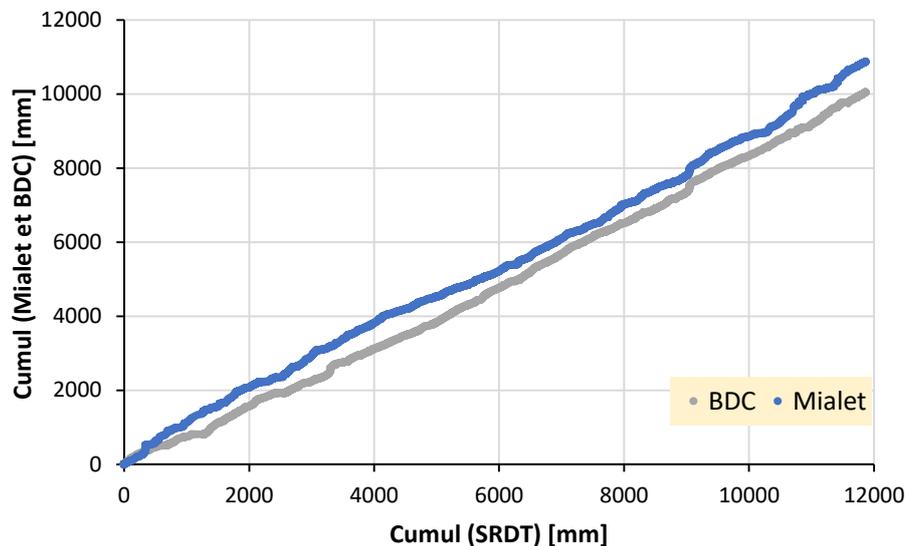


Figure 28: Test de double cumul entre les pluviomètres ; BDC est Barre-des-Cévennes, SRDT est Saint-Roman-de-Tousque sur les épisodes sélectionnés (1992-2019)

On ne constate pas de discontinuité majeure pouvant traduire un quelconque problème significatif avec la mesure pluviométrique.

En vue d'apprécier la structure des différentes pluies qui sont enregistrées au niveau du bassin versant, une analyse de la corrélation linéaire des mesures de précipitation entre les trois stations a été menée (Figure 29). En analysant les graphiques de cette figure, on voit que malgré la petite taille du bassin versant, les informations pluvieuses enregistrées au niveau des pluviomètres circonscrits à ce dernier sont très faiblement corrélées au pas de temps d'une demi-heure. Ce constat concorde assez bien à la description qui a été faite sur les caractéristiques météorologiques de cette zone au niveau du *Chapitre I*.

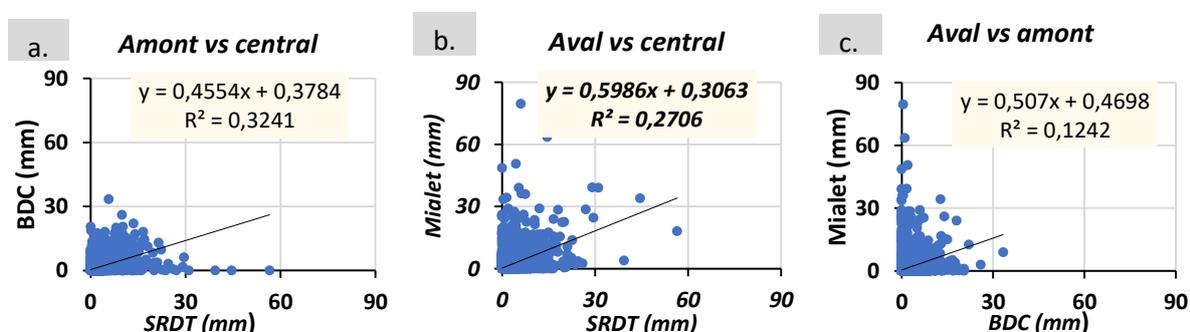


Figure 29: Répartition sous forme de nuages de points des observations mi-horaires des pluies entre les trois pluviomètres, incluant leur droite de régression et leur coefficient de corrélation linéaire simple ; BDC est Barre-des-Cévennes, SRDT est Saint-Roman-de-Tousque (1992-2019)

Malgré la proximité géographique des trois pluviomètres (d'amont vers l'aval, 19 km et 16 km autour de celui du milieu), la dispersion des précipitations enregistrées d'une station à l'autre est très importante. Cette disparité peut avoir deux causes principales : la faible probabilité que les valeurs enregistrées soient synchrones, étant entendu que la taille des objets météorologiques déclenchant les épisodes est plus faible que la distance qui sépare les pluviomètres, d'une part, et d'autre part la différence climatique entre l'amont et l'aval, l'aval étant par exemple plus favorable aux fortes intensités. Bien que faible, l'homogénéité est d'ailleurs plus grande entre les deux pluviomètres amont que lorsque l'aval est impliqué. Ce constat est également conforté en observant les coefficients de détermination qui sont de 0.38, 0.32 et 0.15 pour respectivement Saint-Roman-de-Tousque – Barre-des-Cévennes, Saint-Roman-de-Tousque - Mialet et Barre-des-Cévennes - Mialet.

2.2.6.2 *Corrélations croisées entre les données pluviométriques*

Pour apprécier la dynamique des épisodes pluvieux et la réaction correspondante du bassin versant, nous nous basons sur les corrélations croisées entre les 3 pluviomètres (Figure 30) d'une part et avec le débit observé à l'exutoire de ce dernier d'autre part (Figure 31). Cette analyse devrait permettre non-seulement d'avoir une première description de la réaction du bassin versant à l'impulsion du signal pluvieux, mais aussi d'apprécier les relations entre les données pluviomètres. Toutefois, ces analyses corrélatoires étant de nature linéaire, elles ne sauraient être suffisantes pour permettre d'expliquer exhaustivement la réaction hydrologique globale du bassin versant.

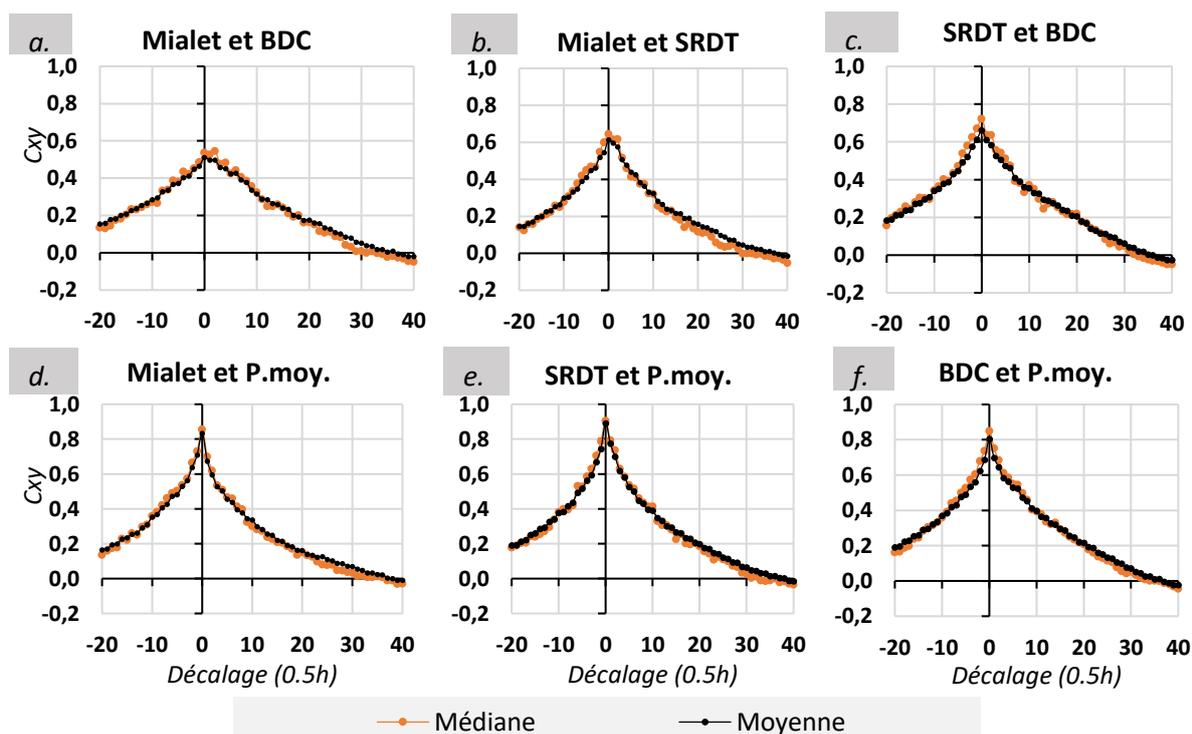


Figure 30: *Corrélations croisées entre les stations pluviométriques ; BDC pour Barre-des-Cévennes, SRDT pour Saint-Roman-de-Tousque ; P.moy. pour la pluie moyenne arithmétique*

Outre les constats faits à partir des tests de corrélations linéaires simples, qui réapparaissent ici, on voit que plus les stations sont distantes, moins les données qu'elles enregistrent se ressemblent. Si les épisodes pluvieux étaient tous homogènes et uniformes sur l'entièreté du bassin versant, on pourrait s'attendre à ce que ces valeurs se rapprochent logiquement de l'unité pour un décalage nul. Sous cet angle, on voit que les deux stations les plus proches et se trouvant dans la partie élevée du bassin versant (SRDT et BDC) présentent un corrélogramme assez symétrique au décalage zéro (Figure 30.c). La valeur maximale y est également la plus élevée des trois combinaisons. En revanche, quand la

station aval (Mialet) est concerné (Figure 30.a&b), la symétrie du corrélogramme disparaît, et un écart à semblable à celui d'un temps de réponse se produit. Cet écart du pic du corrélogramme est, sans surprise, d'autant plus marqué que ces pluviomètres sont distants. Cela souligne ici encore le caractère localisé que peuvent avoir ces épisodes pluvieux sur des bassins versants même aussi petits que celui du Gardon de Mialet. Quant à la relation des observations pluvieuses aux pluies moyennées du bassin versant (Figure 30.d-f), on voit que les corrélogrammes sont tous pratiquement symétriques au décalage zéro. Cependant, les maxima des corrélogrammes vont du côté de la station du milieu (SRDT) à celle du sommet (BDC) en passant par celle se trouvant en aval (Mialet). Ce constat pourrait traduire que les pluies qui tombent au sommet du bassin versant ont moins de similitudes que celles tombées plus en aval. Ceci peut être lié aux valeurs des cumuls qui y sont moins élevées, ainsi que leur nombre en termes de différence. Cependant, en parlant de corrélation croisée pluie-pluie, il ne saurait exister de relation de cause à effet entre deux pluviomètres, même s'ils sont proches. On ne peut donc pas considérer la partie descendante des corrélogrammes comme une sorte d'inertie à proprement parler.

La Figure 31 présente les corrélations croisées entre les différents pluviomètres et le débit, ainsi qu'entre les pluies moyennes et calculées par la méthode des polygones de Thiessen, et enfin l'autocorrélation du débit.

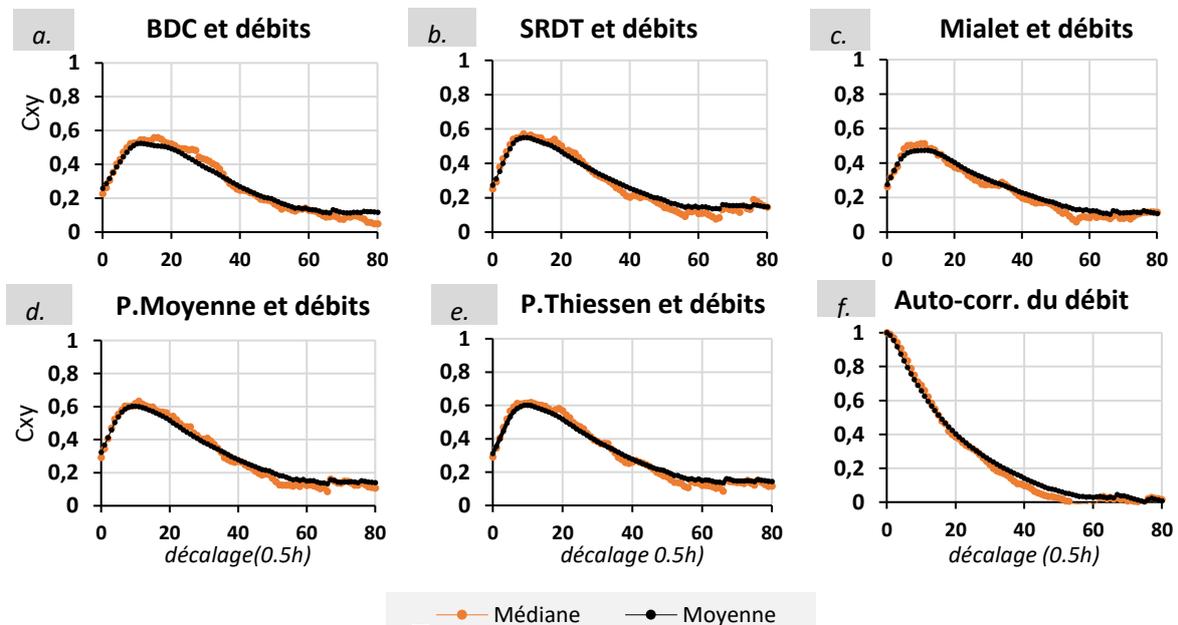


Figure 31: Corrélations croisées pluie-débits entre les différentes stations au niveau du bassin versant de Gardon de Mialet. Les valeurs médianes (noir) et moyennes (orange) sont calculées sur les 58 événements. BDC est Barre-des-Cévennes, SRDT est Saint-Roman-de-Tousque.

Au niveau des graphiques de la Figure 31.a-c., sont présentés les corrélogrammes croisés pluie-débit pour les pluviomètres de l'amont vers l'aval. Les corrélogrammes correspondants indiquent globalement que les pluies enregistrées au niveau des différentes stations peuvent fournir une bonne explication concernant le débit mesuré à l'exutoire. Quant au corrélogramme croisé pluie-débit utilisant les pluies moyennes (Figure 31.d.&e.), que ce soit par la moyenne arithmétique ou par la méthode de Thiessen, l'information apportée à l'explication du débit est légèrement plus importante que celle de n'importe lequel des trois pluviomètres. Ceci semble traduire l'utilité de l'ensemble des trois pluviomètres pour bien décrire le comportement hydrologique du bassin versant, en tout cas dans sa composante linéaire. Comme indiqué à la Figure 31.f, l'autocorrélation du débit qui traduirait l'inertie du bassin versant, indique qu'en moyenne, l'effet mémoire du système lors des crues

sélectionnées ne dépasse que rarement 24 heures. Ces différentes corrélations croisées (C_{xy}) sont résumées dans le *Tableau 2*.

Tableau 2 : Descriptif des relations entre les variables pluviométriques et hydrométriques ; BDC est Barre-des-Cévennes, SRDT est Saint-Roman-de-Tousque

Éléments		Corrélation croisée (Médiane/Moyenne)				
		<i>P. Thiessen</i>	<i>P.moy.</i>	<i>BDC</i>	<i>SRDT</i>	<i>Mialet</i>
BDC	<i>Cxy</i>	0.86/0.81	0.85/0.80		0.72/0.66	0.55/0.51
	<i>Décalage (0.5h)</i>	0/0	0/0		0/0	2/0
SRDT	<i>Cxy</i>	0.96/0.95	0.90/0.89	0.72/0.66		0.64/0.61
	<i>Décalage</i>	0/0	0/0	0/0		0/0
Mialet	<i>Cxy</i>	0.78/0.74	0.86/0.83	0.55/0.51	0.64/0.61	
	<i>Décalage</i>	0/0	0/0	2/0	0/0	
Debit	<i>Cxy</i>	0.62/0.60	0.63/0.60	0.56/0.52	0.57/0.55	0.51/0.47
	<i>Décalage</i>	11/10	11/10	15/11	9/9	9/7
T. Rep.	Heure	5.5/5	5.5/5	7.5/5.5	4.5/4.5	4.5/3.5

On constate que de l'amont vers l'aval, la corrélation entre les précipitations et le débit à l'exutoire ne suit pas une distribution linéaire. En effet, elle est plus marquée pour la station du milieu (Saint-Roman-de-Tousque) avec un score de de 0.57, contre 0.56 pour la plus éloignée (Barre-des-Cévennes) et 0.51 pour Mialet située en aval (Tableau 2). Quant aux pluies moyennes, elles sont mieux corrélées avec le débit (0.62). En se basant sur le corrélogramme des pluies moyennes, le temps de réaction moyen du bassin versant est de quatre à cinq heures selon les données qui constitue cette base de données. Cette valeur reste cohérente avec celle qu'avait trouvée Artigue (2012).

3 Conclusion

Après avoir détaillé les caractéristiques du bassin du Gardon à Mialet et décrit le plus précisément possible les données qui y sont disponibles, on peut mesurer toute la complexité et la non-linéarité de la relation entre pluie et débit. En effet, les différents types d'épisodes de pluie et les réponses qui y sont associées sont autant d'éléments montrant qu'il est très délicat de dégager des tendances de comportement suffisamment robustes. Considérant le nombre suffisamment conséquent d'épisodes pluvieux intenses sur cette base de données, ainsi que l'efficacité dont fait preuve l'approche de modélisation statistique sur ces phénomènes ces dernières décennies, on peut être convaincu que le choix des prédicteurs neuronaux pour la prévision des crues éclair sur ce bassin versant reste un choix fondé et soutenu. Dans la suite de ce travail, l'accent sera mis sur la description des processus physiques sous-jacents à travers l'approche du *Deep Learning*.

Chapitre IV : Towards a better consideration of rainfall and hydrological spatial scales by a deep neural network model to improve flash-floods forecasting. Case study on the Gardon basin, France

Introduction du chapitre

Dans ce chapitre, nous proposons une amélioration des performances existantes en prévision des crues éclair sur le bassin-versant du Gardon à Mialet à l'aide d'un réseau de neurones profond. Cette particularité permet de prendre en compte la variabilité spatiale de la pluie et des processus hydrologiques. A partir d'une architecture postulée discriminant trois zones pour les trois pluviomètres :

- on observe la qualité de la prévision, en particulier vis-à-vis des prévisions déjà existantes sur ce bassin et avec cette base de données (Artigue *et al.*, 2012),
- on observe l'aboutissement de la sélection du modèle afin d'analyser la prise en compte implicite des caractéristiques physiques du bassin.

L'architecture est postulée de telle sorte que l'on mette le modèle sur le chemin des caractéristiques physiques du bassin, mais l'introduction de connaissance reste limitée à ce point. Ces expériences font l'objet d'une publication sous forme d'article soumise au journal NHESS. Les résultats et les points de valorisation de cet article sont présentés en fin de chapitre.

Towards a better consideration of rainfall and hydrological spatial scales by a deep neural network model to improve flash-floods forecasting. Case study on the Gardon basin, France.

Bob E. Saint Fleur¹, Sam Alier¹, Emilien Lassara¹, Antoine Rivet¹, Guillaume Artigue¹, Séverin Pistre¹, Anne Johannet¹

¹ HydroSciences Montpellier, Univ. Montpellier, IMT Mines Ales, IRD, CNRS, 6 av. de Clavières, 30100 Ales, France

Correspondence to: Guillaume Artigue (guillaume.artigue@mines-ales.fr)

Abstract. Flash floods frequently hit the Mediterranean regions and cause numerous fatalities and heavy damage. Their forecast is still a challenge because of the poor knowledge of the processes involved and because of the difficulty to forecast heavy convective rainfall. In any case, early warning remains a strong need. In this study, the authors propose to build a deep artificial neural network for flash flood forecasting, allowing, by its specific architecture, to take better account of the spatial variability and the scales of the rainfall as well as the hydrological responses. For this purpose, a database of 58 heavy rainfall events extracted from 16 years of hydrometeorological observations on a well-studied basin in Southern France is applied to train a deep recurrent neural network. After training and rigorous optimization using several well-known regularization methods, the results are of twofold: first, the improvement of the lead-time from two hours to three hours show that the forecasts are suitable for an operational use; second, the model selection process converged towards an architecture that considers some of the known physical processes of the basin.

1. Introduction

Flash floods are rapid and intense floods that occur on small to middle-sized basins (Gaume et al., 2009). Such flows can reach thousands of cubic meters per second with response times of a few hours only (Montz and Grunfest, 2002). Especially in the Mediterranean regions, they are one of the most destructive natural hazards, and frequently cause numerous fatalities and heavy damage (Llasat et al., 2010; Price et al., 2011; Llasat et al., 2014). For example, the sole event of September 9th 2002, which affected, among others, the Gard (Southern France) including the basin of the *Gardon de Mialet*, caused 22 casualties and about 1.2 billion € of damages (Gaume and Bouvier, 2004). The lack of knowledge about the physical processes involved in the generation of these floods is not only a scientific challenge but also a societal challenge, given their socioeconomic and environmental impacts (Gaume and Bouvier, 2004; Llasat et al., 2010). Most of the time, these flood events are the response to heavy rainfall from convective systems, with a very high variability in space and time (Ayrat, 2005; Garambois et al., 2014; Marchandise, 2007). The hydrodynamic behavior of the hydrosystems that are exposed to intense rainfall events depends as well on soil moisture as geology, tectonics, elevation, and land use (Anctil et al., 2008; Nikopoulos et al., 2011). Soil moisture content estimation at the watershed scale has proven beneficial for discharge prediction (Kitanidis and Bras, 1980; Parajka et al., 2006, Woodridge et al., 2003). Nevertheless, the associated measurements accuracy is highly dependent on field measurement techniques; also, they provide relative spatial and temporal distributions (Gabriel et al., 2007; Lauzon et al., 2004) rather than an absolute information. In addition, the accuracy of meteorological forecasts is often insufficient at the time and space scales needed (Tripathy et al., 2021), e.g., hourly time step (or under) and about 100 sq.km.

These forecasts are however most of the time needed as inputs for hydrological models to produce forecasts at a useful lead-time, especially when transfer times are short, typically on small to middle-sized basins.

Consequently, the involved nonlinear flow production processes are also subject to high spatial and temporal variabilities, resulting from the combination of the rainfall variability and of the variability of the field conditions. In such a context, the behavior of a basin submitted to a rainfall episode is difficult to describe accurately, which can sometimes highlight the lack of representativeness of the physically based models' parameters. Therefore, one can question the relevance of using physically based models on such issues. However, among the actions that can be implemented to protect persons and goods in case of crisis, early warning and forecasting remains one of the most relevant tools for authorities (Borga et al., 2011).

Artificial neural networks (ANN) are now well known to be able to model hydrological systems (Dawson and Wilby, 2001; Shen, 2018). Based on the causal relationship between inputs (mostly rainfall) and outputs (level or discharge), they do not need any physical hypothesis. Thus, because they are not trapped by any assumed relationship, they can, if the forecast horizon remains lower than the response time of the basin, produce forecasts of their outputs without any forecast of their inputs. In a context where rainfall forecasts are lacking and where the non-linear physical processes are poorly described, they can play a key role in flash flood forecasting. Moreover, the increase of calculation capabilities now allows exploring more easily complex architectures, especially using several hidden layers, thus designing deep neural networks. Even if deep learning use in the field of hydrology is only becoming popular these years, the deep learning approach has the potentiality to bring many responses to the challenges linked to hydrology (Sahu et al., 2020). Regarding the versatility, the multidisciplinary and the high performance of deep learning on many complex and non-linear tasks, it can be considered as a good candidate to improve flash floods forecasting. For these reasons, we use a deep recurrent ANN as one of the main types of Multilayer Perceptron used in hydrology (Zulifqar Ali et al., 2017; Tabari and Talaei, 2013; Araujo et al., 2011).

In this paper, we focus on the *Gardon de Mialet* basin (France) at the *Mialet* gauge station, which is frequently subjected to flash floods. The *Gardon de Mialet* river belongs to the watch map provided by the web service of the Central Service of Hydrometeorology and Flood Forecasting (SCHAPI in French) of the French Ministry of Ecology, designed to provide real time warning and forecasts (<http://www.vigicrues.gouv.fr/>). It has already been studied (Artigue et al., 2012) because it generates important risks for the population of vacationers camping along the river and for the *Anduze* town located downstream. Readers interested in learning more about the *Gardon de Mialet* basin can refer to (Artigue et al., 2012).

The purpose of this work, compared to previous studies, is to represent hydrological processes within a deep recurrent neural network model in order to forecast discharges at the outlet of the *Gardon de Mialet* basin. The design explicitly considers the role of elevation and distance to the outlet, through nested basins, to better represent the different spatial and temporal scales involved in the processes generating the responses to the rainfalls. This allows increasing usefully the forecast horizon from 2h to 3h. This should allow authorities to better warn populations: evacuate campers and the most vulnerable houses and interrupt traffic on the potentially flooded roads.

After presenting in Section 2 the basin, the database and the tools and methods necessary for the proper design of an ANN model, Sect. 3 describes the results, both in terms of deep models obtained, and in terms of forecast quality. Section 4, discussion, puts these results in perspective of the operational needs of flood forecasting and of the increase of the forecasting lead time. The conclusion summarizes the key points of the article and addresses new perspectives.

2. Material and methods

2.1 Study area

The *Gardon de Mialet* basin is located in Southern France, in the *Cévennes* range. Its surface is 220 sq.km while its elevation ranges from 170 to 1170 m.a.s.l. and its slopes overpass 30% on average (Fig. 1). It is mostly covered in forests and the underground is quasi exclusively made of micaschists. The climate is typically Mediterranean, with an increasing alteration to a mountain climate with elevation. Heavy rainfall, mostly in autumn after long summer droughts, often reach hundreds of millimeters per day on this basin, with a very high variability in space and time. For example, in September 2002, 340 mm were observed at the outlet of the basin while on the crest, there was only 68 mm. In October 2008, it was 485mm in *Mialet* versus 17mm in *Barre-des-Cévennes*.

In Fig.1, three measurement stations are presented. Lined up on the south-western side of the basin, they are three rain gauges at *Barre-des-Cévennes* (upstream, 930 m.a.s.l.), *Saint-Roman-de-Tousque* (650 m.a.s.l.) and at *Mialet* (outlet, 170 m.a.s.l) with a discharge station in *Mialet*. Geostatistical work on the *Gardon d'Anduze* basin (545 sq. km.), that includes the *Gardon de Mialet* basin, has shown that six rain gauges can describe rainfall variability with a sufficient accuracy (Obled et al., 2008). Since the *Gardon de Mialet* basin area represents about 40% of the *Gardon d'Anduze* basin area, it seems acceptable to consider that three rain gauges scattered over the basin are at least able to provide rainfall measurements of equivalent quality.

2.2 Database

The database extracted by Artigue et al. (2012) contains data from 1992 to 2008 at an hourly time step up to 2001 and a five minutes' time step after. These data are converted to a 30 minutes' time step, which corresponds to the operational time step of the local flood forecasting service "SPC Grand Delta". Following the idea of producing forecasts close to operational conditions, the criteria for extracting events in this database is based on cumulative rainfall: a threshold of 100 mm within sliding 48 hours on at least one of the rain gauges has been chosen. This allows avoiding false positive forecasts by considering the events for which the basin showed a small response while rainfall was significant enough to enjoin the forecaster to question a hydrological model. The result of this extraction is 58 events, concatenated in a single time series of 13,611 examples (couples of input vector and associated measured discharge at the same time).

These events are the same used in a previous study and have already been extensively described (Artigue et al., 2012). Using the same database allows assessing the potential enhancement brought by the present paper.

The main features of the database are presented in Tab.1. Up to 523 mm has been recorded in six days, which corresponds to one of the longest events. In half an hour (sampling rate), rainfall can reach 62 mm (*Mialet* rain gauge) and discharge can exceed 800 m³/s, which corresponds to 3.7 m³/s/km².

The response time of the basin was studied in (Artigue et al., 2012), depending on the event and on the location of the rain, it is included in the interval 1h-3.5h (*Mialet* rain gauge), 2.5h-4.5h (*Saint-Roman-de-Tousque* rain gauge), 1h-3.5h (*Barre-des-Cévennes* rain gauge). These values will therefore constrain the forecast horizons of models that do not take rainfall forecasts data as input in their forecasting process.

2.3 ANN principles and design

2.3.1 Definitions

A neuron is a mathematical operator that calculates a weighted sum of its inputs, called the potential of the neuron, and that applies to this potential a non-linear function to give an output. Neurons can be combined in a network among which the

multilayer perceptron (MLP) is widely used in hydrology (Fig. 2). In Fig. 2, each arrow represents a parameter linking either a variable to a neuron or a neuron to another neuron. These parameters are the weights used to compute the potential of the neuron.

The MLP has the universal approximation property (Hornik et al., 1989). It states that this kind of model, with one or more hidden non-linear layer, is able to approximate any differentiable function with an arbitrary accuracy. Their accuracy should increase with the number of hidden neurons. Another important property of the MLP states that it is more parsimonious than others statistical models linear with respect to their parameters, especially when the number of variables increases (Barron, 1993). These properties make the model particularly suitable for hydro(geo)logical non-linear processes (Jeannin et al., 2021). Several works pointed out the use of Long-Short-Term-Memory models (LSTM), a particular type of deep recurrent neural network model, for hydrologic tasks (Fang et al., 2021; Zhang et al., 2018; Kratzert et al., 2018; Li et al., 2021) with some interesting results. Recent studies such as Muller et al. (2021) and Atmaja et al. (2020) compared MLP networks to other types of networks such as LSTM or convolutional neural networks and concluded to the superiority of MLP for the targeted tasks.

2.3.2 ANN training and overfitting

The training of ANN consists in calculating the parameters' vector so that a cost function (usually based on least squares), measuring the error between simulated and desired outputs, is minimized (Rumelhart et al., 1986). This minimization is performed using iterative training rules; in our case, the second order algorithm of Levenberg-Marquardt (Levenberg, 1944; Marquardt, 1963). For a deep ANN, the training can be exposed to the well-known problem of vanishing or exploding gradient (Bengio et al., 1994; Hochreiter and Schmidhuber, 1997). This problem can be described as an exponential lessening or amplification of the parameters and/or the gradient values, during the training process, through the network layers (due to back-propagation). Through years of study, many strategies have been proposed to mitigate this problem. Bertin and Lambert (1993), and Johannet et al. (1994) proposed to apply a factor to the gradient step when crossing a layer; Schmidhuber (1992) and Hochreiter and Schmidhuber (1997) proposed to add a linear neuron as a shunt at each hidden layer. Chandar et al. (2019), proposed to use a non-saturable activation function and Kanuparthi et al (2019) proposed a method to change the direction of the gradient. There are many other methods to counterbalance this effect and most of them were described in Pascanu et al. (2013).

Training is performed on a set of data called "training set" which represents the major part of the database. When that step is over, the model performance is assessed on the "test set", different and never seen by the model. The test set is also not used for model design. It is extracted from the database at the beginning of the process and is only used at the end to assess the model by calculating the "generalization" error.

While iterations progress in the training step, the cost function is being minimized, but the model can tend to fit both the signal and of the noise carried by the data. In the meantime, bias is minimized and variance may be led to increase leading to overfitting on the test set. This is called the "bias-variance dilemma" and was first described by Geman et al. (1992). A model that over adapted itself to the training set may not be able to generalize the performance obtained on this set to another. Consequently, measuring the error calculated on the training set may not be a relevant estimator of the generalization error. In a similar way, a too simple model will not be able to adapt to the signal and will produce a high bias, while a too complex model will adapt to the noise existing in the training set data, leading to a high variance in test. In the case of ANN, the complexity can be measured by the number of free parameters.

In order to counterbalance these effects, the design of ANN must be performed using regularization methods as presented in Kong A Siou et al. (2011a, 2011b) and highlighted herein the next section.

2.3.3 Regularization methods

Three types of regularization methods have been used in this study: early stopping, cross-validation and ensemble model. They are described in this section.

- **Early stopping:** when using early stopping, the number of training iterations must be limited in order to avoid overtraining. To stop training before overtraining occurs, another independent set of data is introduced, called the “stop set” (improperly called “validation set” in literature). The cost function is measured on this set during training: when the cost function becomes minimal on the stop set, training must be stopped to avoid overfitting. Sjöberg et al. (1995) show that this method diminishes the variance and can thus be considered as a regularization method. This regularization method is called “early stopping”.
- **Cross-validation:** because of the bias-variance dilemma, as the error calculated on the training set does not allow assessing the generalization error, Stone (1974) introduced the “cross-validation” method that allows measuring the quality of the generalization error on the database used for training. It consists in dividing the training set into N subsets, performing N training, each on $N-1$ subsets, and then calculating the generalization error on the remaining subset. This calculation is repeated for all the subsets in order to obtain the generalization error on all the subsets. The cross-validation score is calculated based on the generalization errors of the N subsets. This cross-validation score, which is a relevant estimator of the generalization error, can be for example either the average or a median of the N scores obtained for all the subsets. Toukourou et al. (2011) proposed an adaptation of this method called “partial cross-validation” in which only a certain type of events (in this case flash flood events) belonging to the training set are used for cross-validation. This allows specializing the estimation of the generalization error on this type of events.
- **Ensemble model:** it is well known that the ANN model’s output might be sensitive to parameters’ initialization during the training step (Dreyfus 2005). To overcome that sensitivity, Darras et al. (2014) proposed to calculate the output of the model as the median, at each time step, of a sufficient number of models (members of the ensemble) that differ by their random initializations. As recommended by this author for a similar basin, we choose 10 members in the ensemble. Moreover, Kong-A-Siou et al. (2015) and Akil et al. (2021), showed that this was a way to consider a part of the modeling uncertainty. In the perspective of the development of an operational tool, this would mean that the uncertainty related to parameters’ initialization could be represented.

2.3.4 Variable selection

Variable selection is a critical stage of the neural network model design. Variables can be selected using several methods.

- **Cross-correlation:** it is possible to perform cross-correlations between inputs and the output, and to choose the input variables with the highest correlations with the output. .
- **Cross-validation:** the cross-validation score is calculated for several models having increasing complexity. The complexity can be increased by adding inputs variable, or increasing the number of neurons. Among all these models of different complexity, the one with the lowest cross-validation score is chosen because it is the one with the best generalization. If two models have equivalent cross-validation scores then the model with the lowest complexity is chosen. The model selection using cross validation is a very long process. Kong A Siou et al. (2011) proposed to

reduce this time by performing a two-step selection: first the selection of candidate variables through cross-correlation calculation to diminish their number, and then the fine-tuning of the optimal set of variables with cross-validation. It should be stressed that model selection by cross-validation takes into account the non-linearity of the model, whereas this is not the case for cross-correlation. Cross-validation is thus more relevant.

2.3.5 Model design: postulated architecture and generic architecture

In practice, a model architecture is postulated, based on the expertise and knowledge of the modeler, and then optimized. Regarding modeling of a dynamical process, that is the case of this study, there are several types of multilayer perceptron depending on how the dynamic character of the basin is considered (Nerrand et al., 1993). Three of them are considered:

- The **static model**: it only uses exogenous variables as inputs. It thus needs a great amount of information to be able to calculate the current state of the system. Also, time has no functional role played in it.
- The **feedforward model**: it uses exogenous variables and the previous observed outputs up to the instant of simulation as inputs. This type of model can be used when being sure that observed output values would be available in real-time conditions, or when the output noise is considered lower than the state noise (Nerrand et al., 1993).
- The **recurrent model**: it uses exogenous variables and the previous simulated outputs as inputs. This type of model can be used when the availability of the observed output values is not guaranteed or when the state noise is considered to be lower than the output noise (Nerrand et al., 1993). This model's lead time is limited to the response time of the system, unless a forecast of the exogenous variables is provided.

In this study, the recurrent model is preferred, especially because of the low reliability of the measurements transmission for the observed discharge during flash flood events, thus in extreme conditions. The output noise could thus be extremely important. Besides, as the purpose of this paper is also to find a representation of the process that diminishes the state noise, it is more relevant to use a model in which this noise has a strong impact. The generic equation of such a model is presented in Eq. (1).

$$y(k+l) = f_{NN}(y(k), \dots, y(k-r+1)); \mathbf{x}(k), \dots, \mathbf{x}(k-n+1); \mathbf{W} \quad (1)$$

Where $y(k)$ is the output of the model; f_{NN} is the function implemented by the model; $\mathbf{x}(k)$ the vector of exogenous variables; \mathbf{W} is the vector of parameters; l the lead time; r is the order of the model; n is the length of the vector of the sliding time windows of exogenous inputs.

For the generic architecture of the model, as Artigue et al. (2012) did, we chose to build a model divided into two parts: one is linear and the other not. We also use a deeper architecture:

- Regarding the non-linear part, it contains three deep cascading hidden layers: one for each rain gauge. The upstream rain gauge (*Barre-des-Cévennes*) and its cumulative rainfall since the beginning of the event, feed the deeper hidden layer. The output of this layer, together with the mid-basin rain gauge (*Saint-Roman-de-Tousque*) and its cumulative rainfall since the beginning of the event, feed another hidden layer. The output of this hidden layer, together with the downstream rain gauge (*Mialet*) and its cumulative rainfall since the beginning of the event, feed the “downstream” hidden layer. This last layer is then connected to the output linear neuron. Each hidden layer receives the simulated output of the model at previous time. This architecture thus describes three distinct “cascading” parts of the basin.
- Second, a linear part is added to the shallow output layer, dedicated to representing the linear relation between rainfall and discharge, which is not most of the involved processes, but still an important part during very high flows

(Gaume et al., 2002). This linear part is composed of three input windows, one for each rain gauge, directly connected to the linear neuron of the output.

The resulting postulated architecture is presented in Fig. 3.

2.3.6 Model selection

Based on the postulated architecture, a rigorous process of model selection must be applied to define all the hyper-parameters of the model. To this end, as suggested by Dreyfus (2005), cross validation is performed on the postulated models of increasing complexity until the cross-validation score stabilizes or decreases. Finally, the chosen model is the one having the best score of cross-validation over all the complexity trials. In the present case:

- $n_{BDC}, n_{SRDT}, n_{MIA}$, window length of the Barre des Cévennes, Saint-Roman de Tousque and Mialet rain gauges (green in Fig. 3),
- $n_{C_BDC}, n_{C_SRDT}, n_{C_MIA}$, window length of the three cumulative rainfalls from previous rain gauges (yellow in Fig. 3),
- $n_{BDC_L}, n_{SRDT_L}, n_{MIA_L}$, window length of the three rainfalls from previous rain gauges applied to the linear part of the model (orange in Fig. 3),

the number of hidden neurons for each layer:

- $h_{BDC}, h_{SRDT}, h_{MIA}$, (blue in Fig. 3),

and the order of the model:

- $r_{BDC}, r_{SRDT}, r_{MIA}$, (pale green in Fig. 3).

All these hyper-parameters are thus accurately defined taking into account the nonlinear behavior of the model..

2.3.7 Training

For each training, early stopping is implemented so that overtraining is avoided. 100 epochs are left to find a minimum to the cost function calculated on the stop set, which has been sufficient for each experience of this study.

2.3.8 Performance assessment

The results are assessed using several criteria.

- Nash criterion: NSE (Nash and Sutcliffe Efficiency; Nash and Sutcliffe, 1970)

The first is the Nash criterion NSE , or R^2 , which is described in Eq. (2) and which ranges from $-\infty$ to 1, 1 being perfect and 0 being a simulation equivalent to the simulation providing the average of observed discharges.

$$NSE = 1 - \frac{\sum_{k=1}^j (y_p(k) - y(k))^2}{\sum_{k=1}^j (y_p(k) - \bar{y}_p)^2} \quad (2)$$

Where $y_p(k)$ is the observed (process) output at the discrete time k , \bar{y}_p is the average of the j observed outputs of the considered dataset, $y(k)$ the forecast value at the discrete time k .

- Persistency criterion: C_p

In order to compare the forecast of the model with a naïve forecast, the persistence criterion C_p (Kitanidis and Bras, 1980) is presented in (3). The naïve forecast consists in forecasting at a future time the same value as at the present

time. It is a good forecast if the forecast horizon is small compared to the evolution dynamics of the system. The persistency criterion also ranges from $-\infty$ to 1, 1 being perfect and 0 being a simulation equivalent to a naïve forecast. We use this criterion in this study as a cross-validation score.

$$C_p = 1 - \frac{\sum_{k=1}^j (y_p(k+l) - y(k+l))^2}{\sum_{k=1}^j (y_p(k+l) - y_p(k))^2} \quad (3)$$

Where the notations are as described for the Nash criterion, except that l is the lead-time of the forecast.

If the purpose of modelling is floods, as it is the case in this paper, Artigue et al. (2012) proposed three criteria to analyse the performance of the model: the *SPPD* (4) the *PPD* (5), and the *PD* (6). They allow focusing on the peak, taking into account the timing.

- Synchronous Percentage of the Peak Discharge: *SPPD*

$$SPPD = 100 \frac{y(k_{peak_obs})}{y_p(k_{peak_obs})} \quad (4)$$

Where the notations are as described for the two previous formula except that k_{peak_obs} is the instant of the observed flood peak.

- Percentage of the Peak Discharge: *PPD*

The *PPD* is simply the ratio between the two flood peaks: the predicted peak at time k_{peak_pred} divided by the observed peak at time k_{peak_obs} .

$$PPD = 100 \frac{y(k_{peak_pred})}{y_p(k_{peak_obs})} \quad (5)$$

- Peak Delay: *PD*

The Peak Delay is the lag between the observed and simulated peaks. It is negative when the simulated peak occurs before the observed peak.

$$PD = k_{peak_obs} - k_{peak_pred} \quad (6)$$

With the same notations as in previous lines.

3. Results

The presented results are of two kind: first the results of the model selection and their analysis and second the hydrological results on the four tested events, displayed through hydrograms and criteria.

3.1 Use of the database

Among the 58 available flood events, 53 are used for training, one for early stopping and four events for testing. Among the 53 events used for training, 17 events are kept for a partial cross validation and are selected because of their intensity (more than $1 \text{ m}^3 \cdot \text{s}^{-1} \cdot \text{km}^{-2}$ for the specific peak discharge). Two of the test events selected are the top two most intense ones: the first one which reached $819 \text{ m}^3 \cdot \text{s}^{-1}$ peak flow (event number 19, September 2002) and the second one, $454 \text{ m}^3 \cdot \text{s}^{-1}$ peak flow (event number 13, September 2000). This allows assessing the ability of the model to generalize to unknown events with greater magnitudes. The two other test events are lower and have a very different shape from the previous ones. They are specifically used to extent the assessment of the ability of the model to perform well on large variety of events. These events are numbered 302 ($200 \text{ m}^3 \cdot \text{s}^{-1}$, November 2003) and 8 (two peaks, 119 and $269 \text{ m}^3 \cdot \text{s}^{-1}$, November 1996). It is important to underline that

the test events are not used neither for training, nor for cross-validation, nor for stopping, nor for model selection. They are extracted from the database at the beginning of the process and are only used at the end to evaluate the model.

3.2 Model selection

The models have been selected from the postulated generic architecture following the method described in Section 2.3.6. The resulting six models designed for each lead-time (18 models in total) are presented in Tab. 2.

It is interesting to note that the sizes of the sliding time windows of rainfall (n_{BDC} , n_{SRDT} , n_{MIA}) are decreasing while the rain gauge that feeds them is closer to the outlet of the basin for the five first lead times. This trend is less clear but still exists for the linear part of the model (n_{BDC_L} , n_{SRDT_L} , n_{MIA_L}). On the other hand, the cumulative rainfall (n_{C_BDC} , n_{C_SRDT} , n_{C_MIA}), used to represent soil moisture, shows no significant difference from a rain gauge to another. The order is generally more important downstream as well as the number of neurons in the hidden layers. For the higher lead-times, it is more difficult to find a meaning to the result of the selection process, which might be related to the fact that input variables are not fully explanatory anymore.

In order to improve the understanding of these elements, the plots presented in Fig. 4 highlight the evolution of the main hyper parameters as a function of the lead-times, per rain gauge.

3.3 Forecasting results

The performance criteria for the four tested events are presented for each of the six lead times in Tab. 3. They are compared to the scores obtained on the same events by Artigue et al, 2012.

The two most intense test events forecasts are plotted in Fig. 5 and the two lower events in Fig.6.

One can note that the persistence criterion is lower for the shorter lead-times, which is consistent and usually observed, as the naive forecast is more relevant when the lead-time is short. It is thus more difficult to provide a better forecast than the naive forecast for short lead times. Nevertheless, for the four events, the scores show a clear added value of the present model, especially for the $k+3$ et $k+4$ lead times. Above $k+4$, the performance decreases, especially in terms of PD and $SPPD$, but the forecast still brings a perfectible but exclusive information for an operational use.

For the intense events, the plots report that the lower persistence criteria of the shorter lead times do not affect the added value of the forecast. Up to $k+3$, the model performs very well on peaks and on discharge rise. Beyond, at $k+4$, the performances remain satisfactory but for the event 19, the rise is a little late. Beyond again, the model is too late for the event 19 whereas it still performs well for the event 13 at $k+5$. At $k+6$, even on the event 13, the forecast is difficult to value.

For the other two events, the performance is a little lower, which could be due to the fact that the model has been selected with intense events and that also the complex events are more difficult to simulate. One can note that despite this higher underestimation of the peak value, the dynamics of the events keeps being represented correctly, even on double peaks and longer events.

Compared to Artigue et al. (2012), the present study shows a clear enhancement, especially on the PPD and $SPPD$ values, which are the most important for crisis management.

4. Discussion

First of all, one can see, as shown in the “results” section that ANN show their ability to model and to forecast flash floods, without any forecast of their inputs, within the response time of the basin. This ability particularly suits the issue of flash flood forecasting for which early warning needs are very high. However, the lead time unsurprisingly remains limited to a

few hours, as the performance significantly decreases while the lead time reaches five to six time-steps (2h30 to 3h maximum). The results presented in the previous chapter are thus considered as good, taking into account the difficulty of the task: important uncertainties on past rainfall and flow data, no future rainfall used. Contrarily to what is sometimes written in the literature, the model remarkably well proves its ability to generalize to extreme values since two of the events used for test are the two most intense of the database and were never used for training nor for model selection. The two other events, lower, led the model to operable forecasts, although they were more complex.

For ease of comparison with the previous work (Artigue et al., 2012), the scores obtained by the previous study on the same events with the same database are reminded. It can be noted that they are generally worse except for the lowest lead-times. This shows that considering the spatial variability of rainfall and hydrological processes through “cascaded” basins increases the performance of the model.

The selected architecture also shows several very interesting elements, which confirm the implicit inclusion of the underlying physics in the database. First, the length of the sliding time windows, which is interpreted as, as more or less, the responses times, is logically higher for the upstream rain gauges than for the downstream ones. This is observed as well on the linear parts as on the non-linear ones. Indeed, it can be logically noticed that for the rain gauge farthest from the outlet, older data are needed as the response time of the discharge to this rain gauge is higher, compared to the rain gauge near the outlet. Indeed, in this case, the selection process adapted the model to this logic, showing that it has captured the underlying physics in the data. That observation is the logical consequence of two choices: on the one hand the consideration of a model taking into account physical considerations on the spatialization of processes and secondly the implementation of the rigorous process of model selection and regularization. This allows foreseeing good prospects for "grey box" type modelling using deep models.

Second and similarly, in a very consistent way, when the lead-time increases, the length of the rainfalls sliding time windows tends to decrease, whether about discrete or cumulative rainfall. The model selection thus excluded the oldest values that become less and less physically explanatory to the output. Again, the selection process shaped the model as a function of a physical reality. On the same study area, Toukourou et al. (2011) shown a similar result of the model selection process regarding the length of the sliding time windows and the lead time, but the model was not divided to consider the spatial variability of rainfall and processes.

Third and similarly again, the order is lower for the upstream part. As it is the number of previous estimated values used by the model, for the same reasons as the previous points, the selection process led to reduce the length of these inputs in order to adapt the response time toward downstream and in the future.

In terms of complexity, the number of neurons in each layer increases as it represents a more downstream part of the basin. It could be interpreted as the need, for the model, to process more and more information or more complex information, as getting closer to the outlet.

5. Conclusions

The purpose of this study was to enhance flash flood forecasting on a much-studied basin but whose processes are not well known, by taking into account the spatial variability and scales of rainfall and hydrological response. To achieve this goal, a deep recurrent ANN has been designed and optimized to consider the different space and time scales of the response of the whole basin, implementing a kind of “cascading” sub basins.

The model selection process, devoted to avoid overfitting, was very relevant as the selected model is consistent with the supposed physical or hydrological behavior of the basin. The model, constrained by high-level knowledge, thus succeeded

in learning the underlying physical relationships in the database. In addition, he managed to generalize what he had learned to the two most intense events in the database, themselves belonging to the test set.

The use of the Levenberg-Marquardt algorithm with a relatively parsimonious model does not seem to make it sensitive to the vanishing gradient issue, even if the model has three nonlinear hidden layers.

However, these conclusions could be strengthened by experimenting that kind of generic architecture on other basins especially: (i) smaller, in order to understand from which basin size this method can be relevant; (ii) similar and bigger, in order to challenge the method on different or even on more complex basins, for example karst basins. Besides, the selection process proposed in this study is generic and could be used for any ANN model on any basin.

6. Acknowledgments:

Authors would like to warmly thank the Flood Forecasting Service “SPC Grand Delta” and SCHAPI for providing data. Authors also warmly thank Dominique Bertin from the Geonosis company for the creation and the constant enhancement of the RNF Pro software.

References

- Akil, N., Artigue, G., Savary, M., Johannet, A. and Vinches, M.: Uncertainty Estimation in Hydrogeological Forecasting with Neural Networks: Impact of Spatial Distribution of Rainfalls and Random Initialization of the Model. *Water*, 13(12), 1690, <https://doi.org/10.3390/w13121690>, 2021.
- Ancil, F., Lauzon, N. and Filion, M.: Added gains of soil moisture content observations for streamflow predictions using neural networks, *J. Hydrol.*, 359(3-4), 225–234, doi:10.1016/j.jhydrol.2008.07.003, 2008.
- Araujo, P., Astray, G., Ferrerio-Lage, J.A., Mejuto, J.C., Rodriguez-Suarez, J.A. and Soto, B.: Multilayer perceptron neural network for flow prediction, *Journal of Environmental Monitoring*, 13(1), pp. 35–41. <https://doi.org/10.1039/C0EM00478B>, 2011.
- Artigue, G., Johannet, A., Borrell, V. and Pistre, S.: Flash flood forecasting in poorly gauged basins using neural networks: case study of the Gardon de Mialet basin (southern France). *NHESS*, 12, 3307-3324. doi:10.5194/nhess-12-3307-2012, 2012.
- Atmaja, B. T. and Akagi, M.: Deep Multilayer Perceptrons for Dimensional Speech Emotion Recognition, Asia-Pacific Signal and Information Processing Association Annual Summit and Conference, APSIPA ASC 2020 – Proceedings, 2020.
- Ayral, P.-A.: Contribution à la spatialisation du modèle opérationnel de prévision des crues éclair ALHTAÏR, Thèse de l’Université de Provence Aix-Marseille., 2005. Ayral, P.-A.: Contribution to the spatialization of the operational flash flood forecasting model ALHTAÏR, PhD, Université de Provence Aix-Marseille, 2005.
- Barron, A.R.: Approximation bounds for superpositions of a sigmoidal function. In Proceedings of the IEEE International Symposium on Information Theory—Proceedings, San Antonio, TX, USA, 17–22 January 1993. 930–945, 1993.
- Bengio, Y., Simard, P. and Frasconi, P.: Learning Long-Term Dependencies with Gradient Descent is Difficult, *IEEE Transactions on Neural Networks*, 5(2), pp. 157–166. doi: 10.1109/72.279181, 1994.
- Bertin, D. Lambert, J. Contribution à la modélisation du système karstique : application à la prédiction des débits d’eau souterrains, Rapport de stage d’initiation à la recherche de l’Ecole des Mines d’Alès. Alès, France, 1993.
- Borga, M., Anagnostou, E. N., Blöschl, G. and Creutin, J.-D.: Flash flood forecasting, warning and risk management: the HYDRATE project, *Environ. Sci. Policy*, 14(7), 834–844, doi:10.1016/j.envsci.2011.05.017, 2011.

Chandar, A.P.S., Sankar, C., Vorontsov, E., Kahou, S., Bengio, Y. Towards non-saturating recurrent units for modelling long-term dependencies, 33rd AAAI Conference on Artificial Intelligence, AAAI 2019, 31st Innovative Applications of Artificial Intelligence Conference, IAAI 2019 and the 9th AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2019, 33(01), pp. 3280–3287. doi: 10.1609/aaai.v33i01.33013280, 2019.

Darras, T., Johannet, A., Vayssade, B., Long-a-Siou, L., Pistre, S. Influence of the Initialization of Multilayer Perceptron for Flash Floods Forecasting: How Designing a Robust Model. In International Work-Conference on Time Series 2014, Springer: Granada, Spain, p. 13, 2014

Dawson C. W. and Wilby R.: An Artificial Neural Network Approach to Rainfall Runoff Modelling, *Hydrological Science*, Vol. 43, No. 1, 1998, pp. 47-66.

Dreyfus, G. *Neural Networks, Methodology and Applications*, Springer: Berlin, Germany, p. 509, 2005.

Fang, Z., Wang, Y., Peng, L., Hong, H. Predicting flood susceptibility using LSTM neural networks, *Journal of Hydrology*, 594, p. 125734. <https://doi.org/10.1016/j.jhydrol.2020.125734>, 2021.

Gabriel G. Katul, Amilcare Porporato, Edoardo Daly, A. Christopher Oishi, Hyun-Seok Kim, Paul C. Stoy, Jehn-Yih Juang, and Mario B. Siqueira. On the spectrum of soil moisture from hourly to interannual scales. *Water Resources Research*, vol. 43, W05428, doi:10.1029/2006WR005356, 2007

Garambois, P. A., Larnier, K., Roux, H., Labat, D. and Dartus, D.: Analysis of flash flood-triggering rainfall for a process-oriented hydrological model, *Atmospheric Res.*, 137, 14–24, doi:10.1016/j.atmosres.2013.09.016, 2014.

Gaume, E. (2002) *Eléments d’analyse sur les crues éclair*. Université du Québec.

Gaume, E. and Bouvier, C.: Analyse hydro-pluviométrique des crues du Gard et du Vidourle des 8 et 9 septembre 2002, *Houille Blanche*, (6), 99–106, 2004. Hydro-rainfall analysis of the Gard and Vidourle floods of September 8 and 9, 2002, *Houille Blanche*, (6), 99-106, 2004.

Gaume, E., Bain, V., Bernardara, P., Newinger, O., Barbuc, M., Bateman, A., Blaškovičová, 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.: A compilation of data on European flash floods, *J. Hydrol.*, 367(1–2), 70–78, doi:10.1016/j.jhydrol.2008.12.028, 2009.

Geman, S., Bienenstock, E., Doursat, R. Neural Networks and the Bias/Variance dilemma. *Neural Comput.*, 4, 1–58, 1992.

Hochreiter, S., Schmidhuber, J., Long Short-Term Memory, *Neural Computation*, 9(8), pp. 1735–1780. doi: 10.1162/neco.1997.9.8.1735, 1997.

Hornik, K., Stinchcombe, M., White, H. Multilayer Feedforward Networks are Universal Approximators. *Neural Netw.*, 2, 359–366. 1989.

Jeannin, P.Y., Artigue, G., Butscher, C., Chang, Y., Charlier, J.B., Duran, L., Gill, L., Hartmann, A., Johannet, A., Jourde, H., Kavousi, A., Liesch, T., Liu, Y., Lüthi, M., Malard, A., Mazzilli, N., Pardo-Igúzquiza, E., Thiéry, D., Reimann, T., Schuler, P., Wöhling, T., Wunsch, A. Karst modelling challenge 1: Results of hydrological modelling, *Journal of Hydrology*, Volume 600, 126508, <https://doi.org/10.1016/j.jhydrol.2021.126508>, 2021.

Johannet, A., Mangin, A. and Hulst, D.D.: Subterranean Water Infiltration Modelling by Neural Networks: Use of Water Source Flow, in International Conference on Artificial Neural Networks ICANN 94. Vol II. Sorrento, Italy, pp. 1033–1036, 1994.

Kanuparthi, B., Arpit, D., Kerg, G., Ke, N.R., Mitliagkas, I., Bengio and Y. H-Detach: Modifying the LSTM gradient towards better optimization, 7th International Conference on Learning Representations, ICLR 2019, 2019.

Kitanidis, P. K. and Bras, R. L.: Real-time forecasting with a conceptual hydrologic model: 2. Applications and results, *Water Resour. Res.*, 16(6), 1034–1044, doi:10.1029/WR016i006p01034, 1980.

Kong A Siou, L., Johannet, A., Borrell Estupina, V. and Pistre, S.: 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), pp. 367–380. doi: 10.1016/j.jhydrol.2011.04.015, 2011a.

Kong A Siou, L., Johannet, A., Borrell Estupina, V. and Pistre, S.: 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), pp. 2365–2375. doi: 10.1007/s12665-011-1450-9, 2011b.

Kong-A-Siou, L., Johannet, A., Estupina, V., and Pistre, S.: Neural networks for karst groundwater management: case of the Lez spring (Southern France). *Environmental Earth Sciences*, 74 (12), pp.7617-7632, 2015.

Kratzert, F., Klotz, D., Brenner, C., Schulz and K., Herrnegger: M. Rainfall-runoff modelling using Long Short-Term Memory (LSTM) networks, *Hydrology and Earth System Sciences*, 22(11), pp. 6005–6022. doi: 10.5194/hess-22-6005-2018, 2018.

Lauzon, N., Anctil, F. and Petrinovic, J.: Characterization of soil moisture conditions at temporal scales from a few days to annual, *Hydrol. Process.*, 18(17), 3235–3254, doi:10.1002/hyp.5656, 2004.

Levenberg, K.: A method for the solution of certain non-linear problems in least squares. *Q. Appl. Math.*, 2, 164–168, 1944.

Li, W., Kiaghadi, A. and Dawson, C.N.: High Temporal Resolution Rainfall Runoff Modelling Using Long-Short-Term-Memory (LSTM) Networks, *Neural Computing and Applications*, 33(4), pp. 1261–1278. doi: 10.1007/s00521-020-05010-6, 2021.

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: High-impact floods and flash floods in Mediterranean countries: the FLASH preliminary database, *Adv. Geosci.*, 23(23), 47–55, 2010.

Llasat, M. C., Marcos, R., Llasat-Botija, M., Gilabert, J., Turco, M. and Quintana-Seguí, P.: Flash flood evolution in North-Western Mediterranean, *Atmospheric Res.*, 149, 230–243, doi:10.1016/j.atmosres.2014.05.024, 2014.

Marchandise, A.: 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, *Distributed hydrological modeling on the Gardon d'Anduze, comparative study of different rainfall-flow models, ex-trapolation from normal to extreme and hypothesis testing on hydrological processes* 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), 2007.

Marquardt, D.W.: An Algorithm for Least-Squares Estimation of Nonlinear Parameters. *J. Soc. Ind. Appl. Math.*, 11, 431–441, 1963.

Montz, B. E. and Grunfest, E.: Flash flood mitigation: recommendations for research and applications, *Glob. Environ. Change Part B Environ. Hazards*, 4(1), 15–22, doi:10.1016/S1464-2867(02)00011-6, 2002.

Müller, J., Park, J., Sahu, R., Varadharajan, C., Arora, B., Faybishenko B. and Agarwal, D.: Surrogate optimization of deep neural networks for groundwater predictions, *Journal of Global Optimization*, 81(1), pp. 203–231. doi: 10.1007/s10898-020-00912-0, 2021.

Nash, J. E. and Sutcliffe, J. V.: River flow forecasting through conceptual models part I – A discussion of principles, *J. Hydrol.*, 10, 282–290, 1970.

- Nerrand, O., Roussel-Ragot, P., Personnaz, L., Dreyfus, G. and Marcos, S.: Neural Networks and Nonlinear Adaptive Filtering: Unifying Concepts and New Algorithms. *Neural Comput.*, 5, 165–199, 1993.
- Nikolopoulos, E. I., Anagnostou, E. N., Borga, M., Vivoni, E. R. and Papadopoulos, A.: Sensitivity of a mountain basin flash flood to initial wetness condition and rainfall variability, *J. Hydrol.*, 402(3-4), 165–178, doi:10.1016/j.jhydrol.2010.12.020, 2011.
- Obled, C, Zin, I. and Hingray, B.: « Choix des pas de temps et d’espace pour des modélisations parcimonieuses en hydrologie des crues ». "Choice of time and space steps for parsimonious modeling in flood hydrology", Colloque SHF - Prévisions hydrométéorologiques – pp. 127-134, 2008.
- Parajka, J., Naeimi, V., Blöschl, G., Wagner, W., Merz, R. and Scipal, K.: Assimilating scatterometer soil moisture data into conceptual hydrologic models at the regional scale, *Hydrol Earth Syst Sci*, 10(3), 353–368, doi:10.5194/hess-10-353-2006, 2006.
- Pascanu, R., Mikolov, T. and Bengio, Y.: On the difficulty of training recurrent neural networks, 30th International Conference on Machine Learning, ICML 2013, 2013.
- 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., Nicolaidis, K., Rozalis, S., Savvidou, K. and Ziv, B.: The FLASH Project: using lightning data to better understand and predict flash floods, *Environ. Sci. Policy*, 14(7), 898–911, doi:10.1016/j.envsci.2011.03.004, 2011.
- Rumelhart, D.E., Hinton, G.E. and Williams, R.J.: Learning representations by back-propagating errors’, *Nature*, 323(6088), pp. 533–536. doi: 10.1038/323533a0, 1986.
- Sahu, R. K., Muller, J., Park, J., Varadharajan, C., Arora, B., Faybishenko, B. and Agarwal, D.: Impact of Input Feature Selection on Groundwater Level Prediction From a Multi-Layer Perceptron Neural Network, *Frontiers in Water*, 2, p. 46. doi: 10.3389/frwa.2020.573034, 2020.
- Schmidhuber, J.: Learning to Control Fast-Weight Memories: An Alternative to Dynamic Recurrent Networks, *Neural Computation*, 4(1), pp. 131–139. doi: 10.1162/neco.1992.4.1.131, 1992.
- Shen, C.: A Transdisciplinary Review of Deep Learning Research and Its Relevance for Water Resources Scientists, *Water Resources Research*, 54(11), pp. 8558–8593. doi: 10.1029/2018WR022643, 2018.
- Sjöberg, J., Zhang, Q., Ljung, L., Benveniste, A., Delyon, B., Glorennec, P.-Y., Hjalmarsson, H. and Juditskys, A.: Nonlinear Black-box Modeling in System Identification: A Unified Overview. *Automatica*, 31, 1691–1724, 1995.
- Stone, M.: Cross-Validatory Choice and Assessment of Statistical Predictions (With Discussion). *J. R. Stat. Soc. Ser. B*, 38, 111–147, 1974.
- Tabari, H. and Talaei, P.H.: Moisture index for Iran: spatial and temporal analyses, *Global and Planetary Change*, 100, pp. 11–19., 2013.
- Toukourou, M., Johannet, A., Dreyfus, G. and Ayrat, P.A.: Rainfall-runoff modelling of flash floods in the absence of rainfall forecasts: the case of “Cévenol flash floods”, *J. Appl. Intell.*, 35, 1078–189, 2011.
- Tripathy, S. S., Karmakar, S. and Ghosh, S.: Hazard weather scale for extreme rainfall forecast reduces uncertainty, *Journal of Hydrology*, 14, 100106, <https://doi.org/10.1016/j.wasec.2021.100106>, 2021.
- Wooldridge, S. A., Kalma, J. D. and Walker, J. P.: Importance of soil moisture measurements for inferring parameters in hydrologic models of low-yielding ephemeral catchments, *Environ. Model. Softw.*, 18(1), 35–48, doi:10.1016/S1364-8152(02)00038-5, 2003.

Zhang, J., Zhu, Y., Zhanga, X., Ye M. and Yang, J.: Developing a Long Short-Term Memory (LSTM) based model for predicting water table depth in agricultural areas, *Journal of hydrology*, 561, pp. 918–929, 2018.

Zulifqar Ali, Z., Hussain, I., Faisal, M., Nazir, H.M., Hussain, T., Shad, M.Y., Shoukry, A.M. and Gani, S.H.: Forecasting drought using multilayer perceptron artificial neural network model, *Advances in Meteorology*, Vol. 2017, 5681308. <https://doi.org/10.1155/2017/5681308>, 2017.

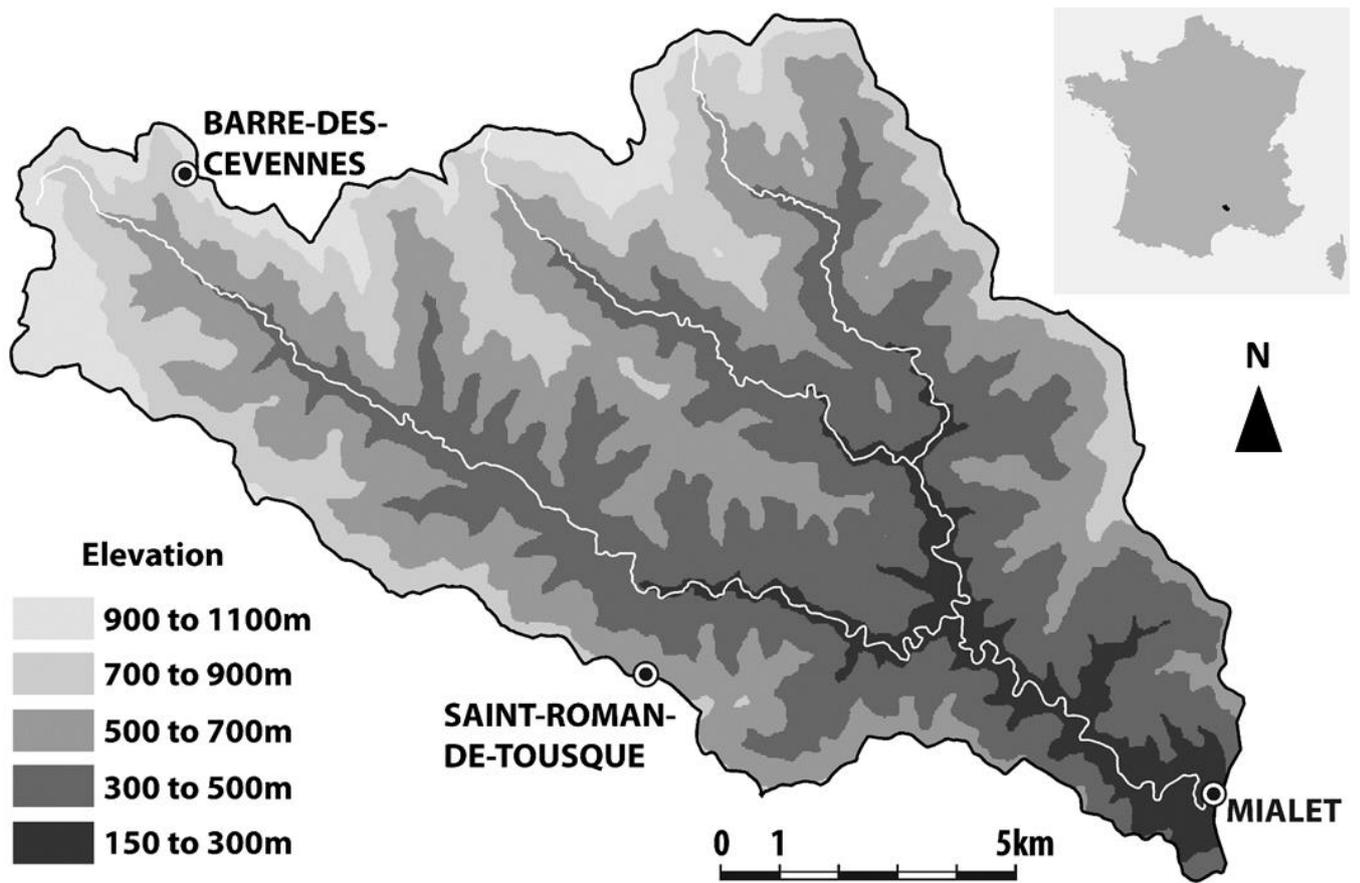


Figure 32: Map and location of the *Gardon de Mialet basin* and its measurement stations.

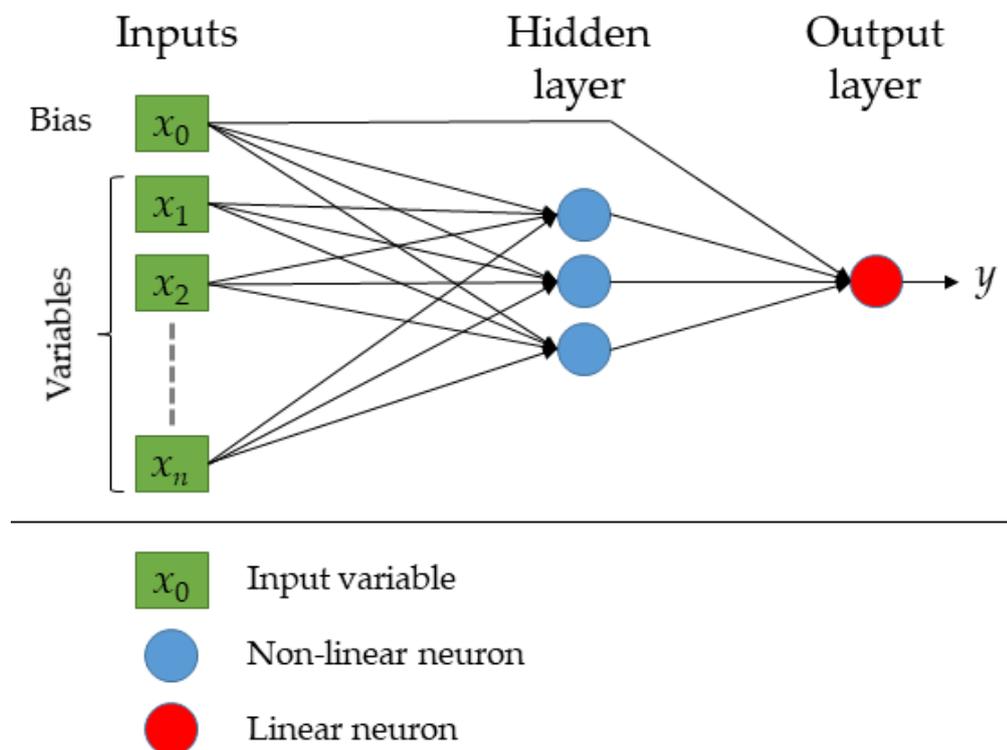


Figure 2: Schematic representation of a multilayer perceptron having one hidden layer where x_j are inputs variables and y the output.

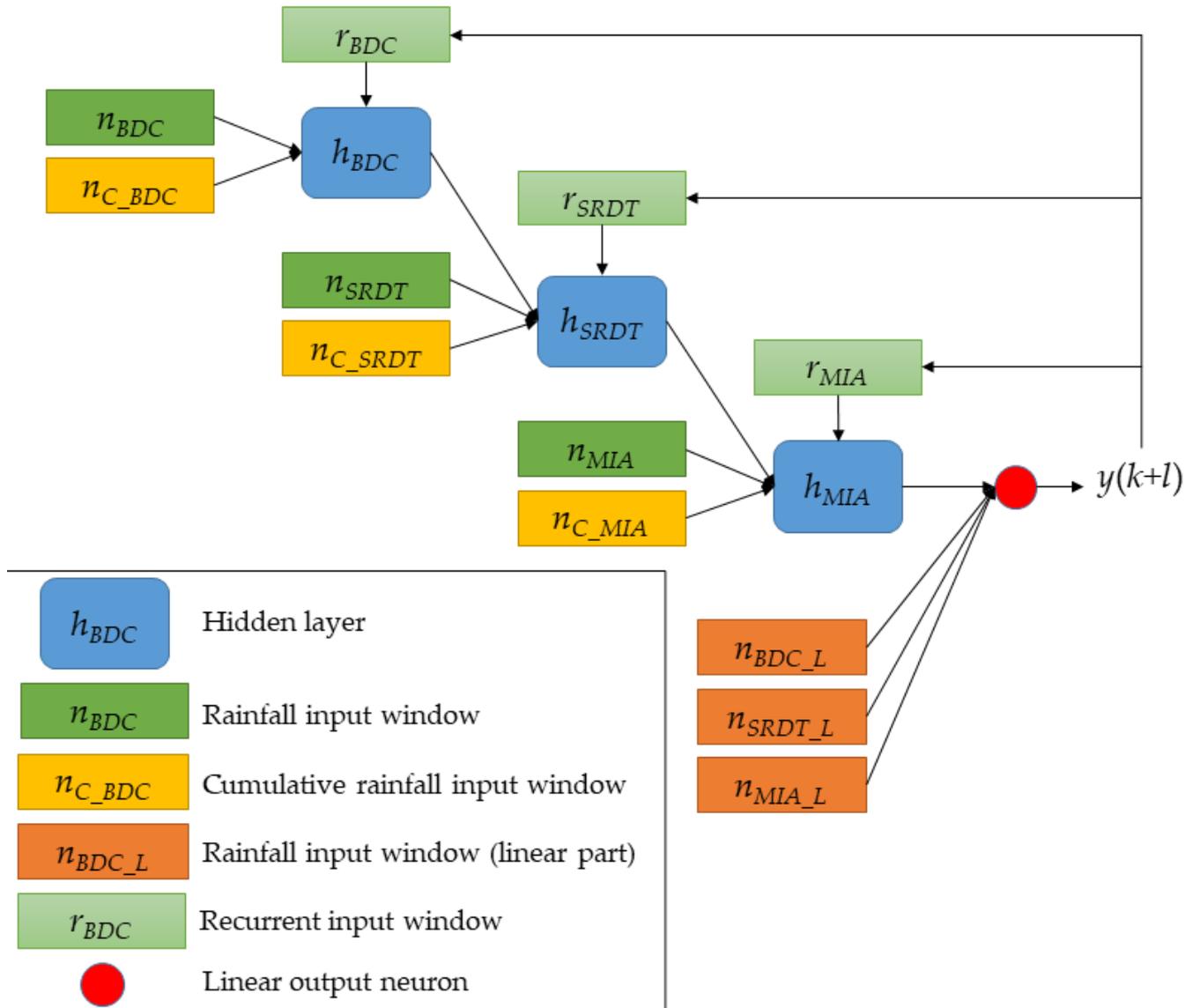


Figure 3: Generic postulated architecture of the developed models in which BDC stands for *Barre-des-Cévennes* rain gauge, SRDT stands for *Saint-Roman-de-Tousque* rain gauge and MIA stands for *Mialet* rain gauge. $y(k+l)$ is the output of the model, a forecast at the l lead time, and k is the discrete time.

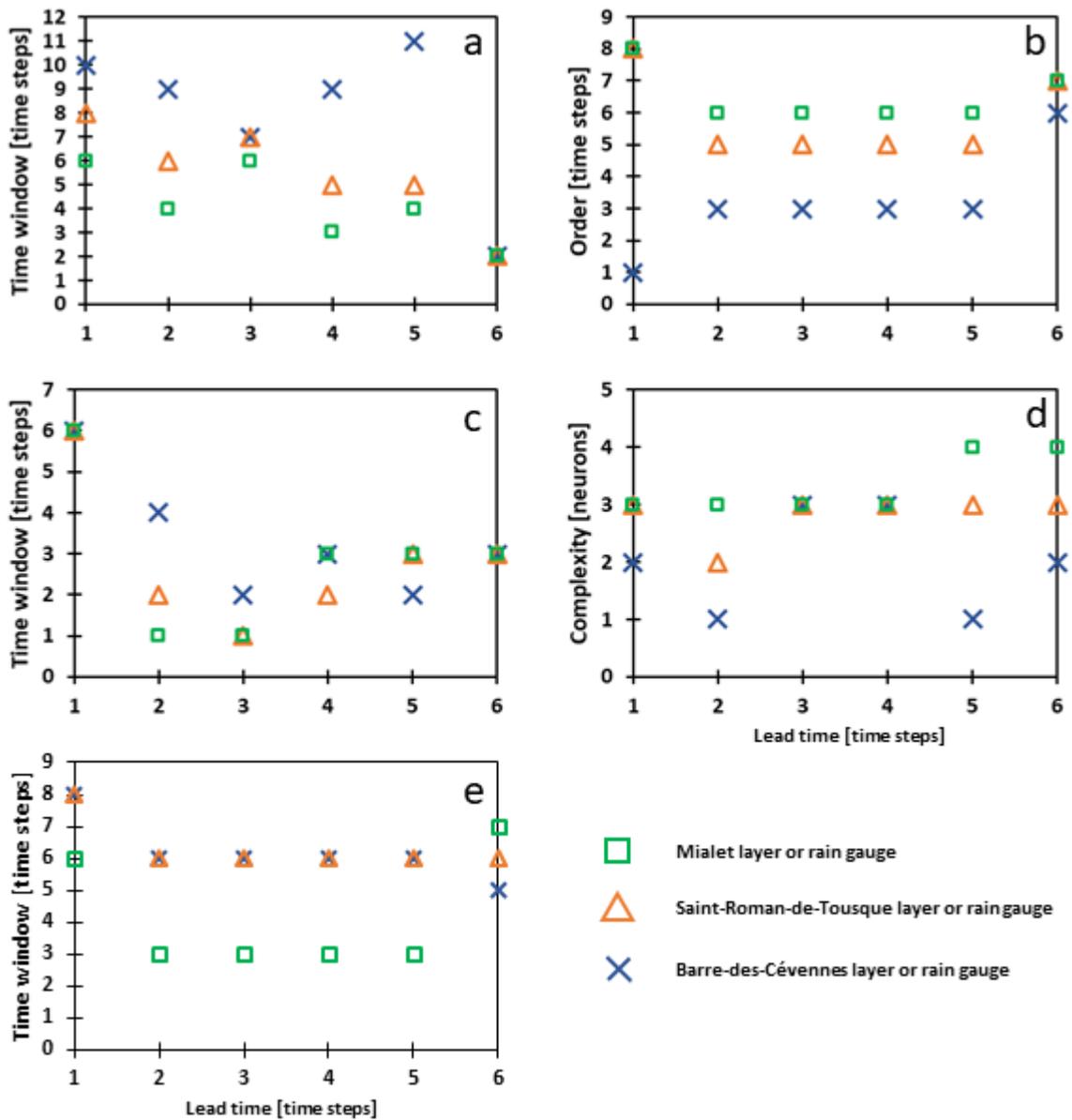


Figure 4: graphic representation of the main elements of the selected architecture as a function of the lead-time; a: time window of rain gauges connected to their hidden layers; b: order for each hidden layer; c: time window of cumulative rainfall values of each rain gauge; d: complexity of each hidden layer; e: time window of rain gauges linearly connected.

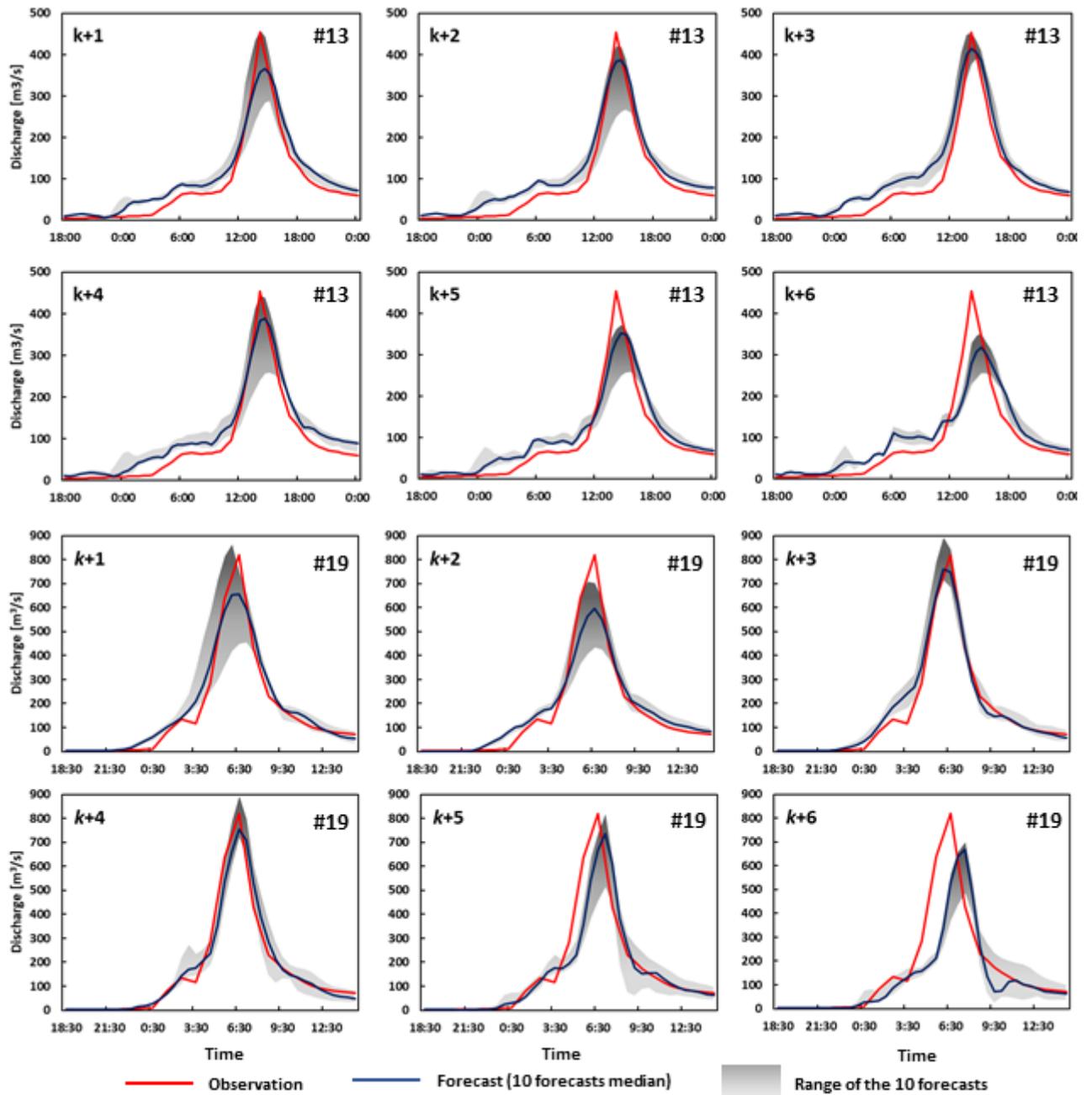


Figure 5: Hydrograms of forecasts for the two most intense events (#13 and #19) and for the six lead times considered. The median and the range of the 10 forecasts from the 10 training initializations are represented.

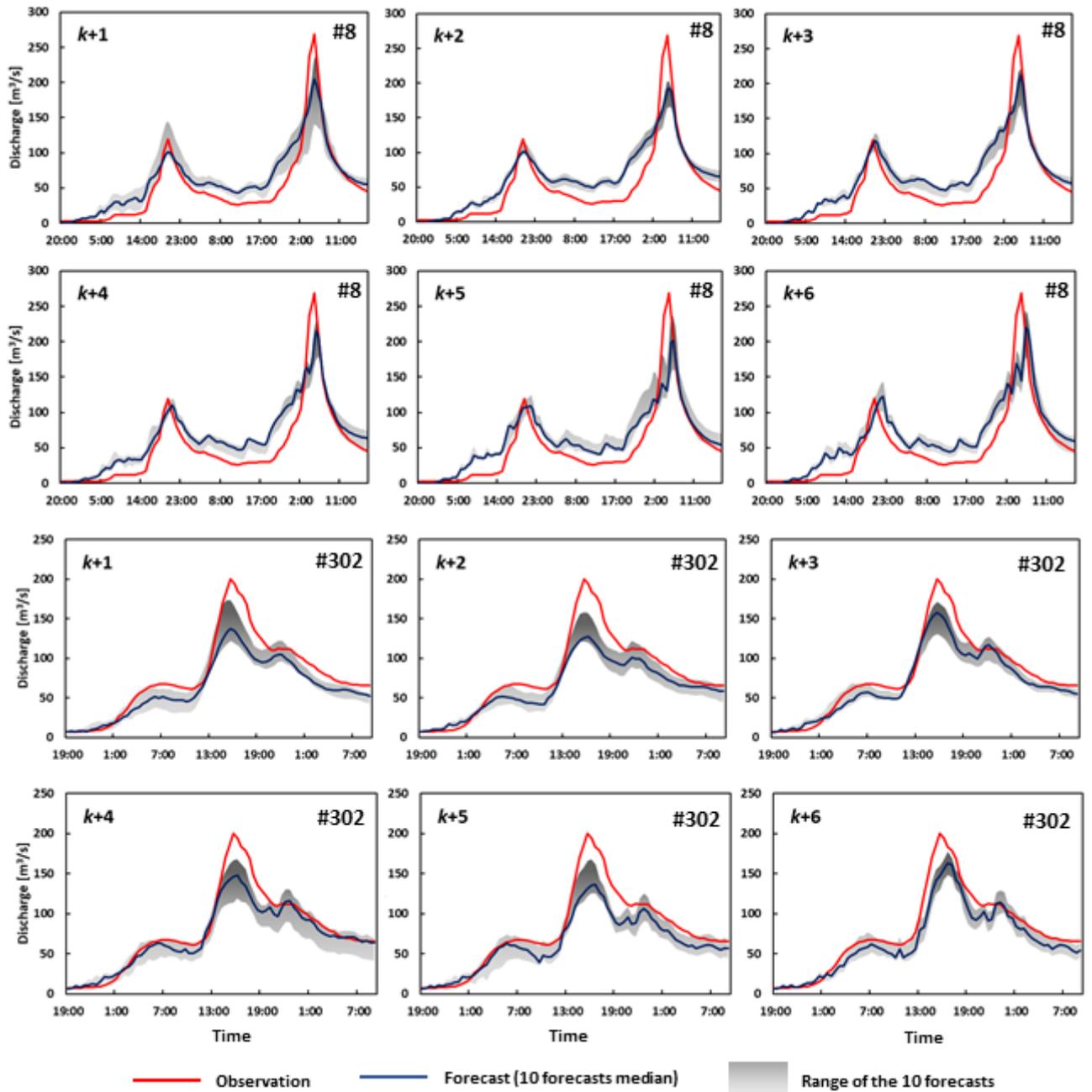


Figure 6: Hydrograms of forecasts for two low and complex events (#8 and #302) and for the six lead times considered. The median and the range of the 10 forecasts from the 10 training initializations are represented.

Table 1: Characteristic values measured by the three rain gauges and the hydrometric station for all the database.

	Rainfall [mm]			Discharge	
	Barre-des-Cévennes	Saint-Roman-de-Tousque	Mialet	Discharge [m ³ /s]	Specific discharge [m ³ /s/km ²]
Maximum (30 min)	33	42	62	819	3.72
Minimum (30 min)	0	0	0	2	0.01
Median (30 min)	0.3	0.3	0.2	-	-
Average (30 min)	1	1.3	1.2	-	-
Maximum (event)	410	523	485	-	-
Minimum (event)	0	1	40	-	-
Median (event)	133.5	164.5	141	29	0.13
Average (event)	143	173	169	43	0.2

Table 2: Results of the optimization of the generic architecture presented in Fig. 3: values of all the hyperparameters for the six lead-times explored. There is one model for each lead-time. All the hyper-parameters correspond to the ones described in Fig. 3 and section 2.3.5.

Hyper-parameters	Lead time					
	$k+1$	$k+2$	$k+3$	$k+4$	$k+5$	$k+6$
n_{BDC}	10	9	7	9	11	2
n_{C_BDC}	6	4	2	3	2	3
r_{BDC}	1	3	3	3	3	6
n_{BDC_L}	8	6	6	6	6	5
h_{BDC}	2	1	3	3	1	2
n_{SRDT}	8	6	7	5	5	2
n_{C_SRDT}	6	2	1	2	3	3
r_{SRDT}	8	5	5	5	5	7
n_{SRDT_L}	8	6	6	6	6	6
h_{SRDT}	3	2	3	3	3	3
n_{MIA}	6	4	6	3	4	2
n_{C_MIA}	6	1	1	3	3	3
r_{MIA}	8	6	6	6	6	7
n_{MIA_L}	6	3	3	3	3	7
h_{MIA}	3	3	3	3	4	4

Table 3: Performance criteria values for the four events in test and for all the lead times explored. The unit for Peak Delay is step times (30 minutes). For ease of comparison, the scores obtained on the same events with the same database on the same basin by Artigue et al. (2012) are shown in brackets and all the scores enhanced by the present study are in bold type.

Event	Criterion	Lead-time					
		<i>k</i> +1	<i>k</i> +2	<i>k</i> +3	<i>k</i> +4	<i>k</i> +5	<i>k</i> +6
13	<i>NSE</i>	0,87 (0,94)	0,89 (0,95)	0,89 (0,93)	0,88 (0,93)	0,91	0,84
	<i>C_p</i>	-1,49 (0,07)	0,41 (0,65)	0,71 (0,82)	0,79 (0,86)	0,89	0,84
	<i>PPD</i>	81% (84%)	85% (83%)	88% (79%)	84% (77%)	80%	72%
	<i>SPPD</i>	79% (81%)	84% (79%)	87% (79%)	84% (75%)	78%	64%
	<i>PD</i>	0 (1)	0 (1)	-1 (1)	0 (1)	1	2
19	<i>NSE</i>	0,89 (0,85)	0,92 (0,92)	0,98 (0,91)	0,92 (0,91)	0,86	0,64
	<i>C_p</i>	-0,07 (-0,25)	0,75 (0,71)	0,96 (0,84)	0,92 (0,89)	0,89	0,77
	<i>PPD</i>	80% (67%)	79% (72%)	96% (74%)	94% (83%)	92%	80%
	<i>SPPD</i>	80% (63%)	79% (69%)	94% (73%)	94% (83%)	81%	63%
	<i>PD</i>	0 (-1)	0 (-1)	-1 (-1)	0 (0)	1	2
302	<i>NSE</i>	0,79 (0,63)	0,76 (0,64)	0,89 (0,57)	0,87 (0,57)	0,79	0,87
	<i>C_p</i>	-8,66 (0,65)	-2 (0,8)	0,34 (0,84)	0,54 (0,79)	0,47	0,75
	<i>PPD</i>	68% (64%)	64% (70%)	79% (72%)	74% (72%)	68%	81%
	<i>SPPD</i>	68% (64%)	63% (70%)	79% (72%)	73% (72%)	66%	74%
	<i>PD</i>	0 (0)	1 (1)	0 (0)	1 (0)	2	2
8	<i>NSE</i>	0,82 (0,80)	0,73 (0,78)	0,76 (0,77)	0,74 (0,81)	0,75	0,71
	<i>C_p</i>	-4,77 (0,64)	-1,36 (0,75)	-0,03 (0,67)	0,3 (0,7)	0,51	0,55
	<i>PPD</i>	76% (70%)	71% (62%)	79% (67%)	80% (65%)	75%	82%
	<i>SPPD</i>	76% (70%)	71% (62%)	75% (63%)	75% (63%)	57%	54%
	<i>PD</i>	0 (0)	1 (0)	1 (1)	1 (2)	2	2

Points de valorisation scientifique de la publication

Un modèle profond a été appliqué à la prévision crues éclair du bassin versant du Gardon de Mialet en utilisant une partie de la base de données présentée au niveau du *Chapitre III*. Le modèle a été testé sur quatre événements dont les deux épisodes de crues les plus importants de la période 1992-2008 sur la zone, soit celui du 28 septembre 2000 et celui du 8 au 9 septembre 2002 qui peut être considéré comme un événement exceptionnel. Les deux autres événements sont sélectionnés compte tenu de leurs caractéristiques assez différentes des deux premiers. Les principaux résultats obtenus à l'issue de ces expériences indiquent que :

- Les réseaux de neurones restent très efficaces pour la prévision des crues éclair. En effet, l'un des objectifs de ces expériences était de prolonger l'échéance de prévision atteinte précédemment par Artigue (2012) de deux à trois heures. L'utilisation d'un modèle profond, prenant en compte la variabilité spatiale des pluies et des temps de réponses sur le même bassin versant et avec la même base de données, a effectivement permis de disposer d'une échéance de prévision utile plus longue (2h30) sur les crues éclair sans information de prévision des pluies, mais aussi d'améliorer les performances.
- L'architecture du modèle profond optimisé sur ce bassin versant évoque une cohérence assez intéressante vis-à-vis de la réalité hydrologique du bassin versant lors des crues éclair.
 - Plus le pluviomètre est éloigné de l'exutoire, plus la profondeur historique utilisée par le modèle augmente. Cette sélection est sensée puisque la pluie qui tombe en amont met plus de temps pour atteindre l'exutoire que celle qui tombe en aval. Cette donnée, liée aux temps de réponses calculés sur le bassin versant pour les différents pluviomètres, a été bien mise en évidence à travers la sélection du modèle.
 - Plus l'échéance de prévision augmente, plus la profondeur de l'historique considéré sur les variables diminue. Ceci traduit classiquement la perte de l'information dans le temps par le système. Les informations éloignées dans le temps perdent graduellement de leur significativité. Conséquemment, la qualité des prévisions réalisées par le modèle ne commence à vraiment se détériorer qu'au-delà de l'échéance équivalente au temps de réponse médian du bassin versant.
 - L'ordre du modèle récurrent, qui fournit une information sur l'état du système est plus important à l'aval qu'en amont. Puisque tout le système évolue vers l'aval, les plus importantes informations s'y trouvent également et pour les mêmes raisons que pour les entrées exogènes, il est physiquement logique que l'ordre des entrées récurrentes plus proches de l'exutoire soit plus grand.
 - La complexité du modèle augmente de l'amont vers l'aval. Puisque l'évolution du système se dirige vers l'aval, les informations à traiter sont également plus importantes, d'où la nécessité d'avoir un nombre de neurones plus conséquent.

A la lumière de ces différents constats, l'utilisation des modèles à réseaux de neurones profonds prenant en compte les variabilités spatiales des pluies et des temps de réponses permet non seulement de disposer d'une meilleure prévision sur ces crues éclair, mais également de tenir compte des réalités physiques observables sur le bassin versant.

Ces résultats paraissent très intéressants sur plusieurs points de vue. D'abord ils permettent de constater que ce type d'approche statistique n'est pas aussi éloignée de la réalité physique que l'on pourrait le penser. Ils permettent également, par l'intégration implicite des caractéristiques physiques du phénomène naturel, d'avoir une meilleure prise en compte des processus générateurs des différentes configurations de ces crues via le *Deep Learning*. On peut, en visant la généralisation de ces conclusions, envisager d'éprouver cette approche sur d'autres bassins de taille et de contexte hydrologique différents que celui de notre bassin versant expérimental.

Chapitre V : Knowledge Extraction (KnoX) in Deep Learning: Application to the Gardon de Mialet Flash Floods Modelling

Introduction du chapitre

Ce chapitre a fait l'objet d'une publication sous forme d'acte de conférence qui a été présentée lors de l'édition 2019 de l'*International Conference on Times Series and Forecasting (ITISE)* à Granada en Espagne. Elle est référencée aux pages 178 – 189 du volume 1 du proceeding accessible via https://itise.ugr.es/ITISE2019_Vol1.pdf. Elle a pour objectif de présenter dans un premier temps l'application des réseaux de neurones à la modélisation des crues éclair au niveau d'un bassin versant cévenol, puis extraire les informations contenues dans le modèle optimisé en vue d'en améliorer leur compréhension et leur utilité. Nous avons, pour cela, utilisé le bassin versant du Gardon de Mialet comme zone expérimentale. Pour accéder aux informations intrinsèques de ces modèles, on a utilisé la méthode « Knowledge eXtraction » ou « KnoX » proposée par Kong A Siou *et al.* (2013), d'où le titre de la publication « **Knowledge Extraction (KnoX) in Deep Learning: Application to the Gardon de Mialet Flash Floods Modelling** ». Ce travail s'ouvre à l'apprentissage profond « *Deep Learning* » car il s'agissait d'interpréter les informations contenues au niveau des paramètres des couches profondes du modèle neuronal.

Knowledge Extraction (KnoX) in Deep Learning: Application to the *Gardon de Mialet* Flash Floods Modelling

Bob E. Saint Fleur ^{1,2}, Guillaume Artigue ¹, Anne Johannet ¹, Severin Pistre ²

¹ IMT Mines Alès, Laboratoire de Génie et de l'Environnement Industriel (LGEI), Alès, France

² Hydrosociences, Univ Montpellier, CNRS, IRD, 34090 Montpellier, France

Corresponding author Guillaume ARTIGUE (guillaume.artigue@mines-ales.fr)

Abstract. Flash floods frequently hit Southern France and cause heavy damages and fatalities. To better protect persons and goods, official flood forecasting services in France need accurate information and efficient models to optimize their decision and policy. Since heavy rainfalls that cause such floods are very heterogeneous, it becomes a serious challenge for forecasters. Such phenomena are typically nonlinear and more complex than classical floods events. That problem leads to consider complementary alternatives to enhance the management of such situations. For decades, artificial neural networks have been very efficient to model nonlinear phenomena, particularly rainfall-discharge relations in various types of basins. They are applied in this study with two main goals: first modelling flash floods on the *Gardon de Mialet* basin; second, extract internal information from the model by using the Knowledge eXtraction method to provide new ways to improve models. The first analysis shows that the kind of nonlinear predictor influences strongly the representation of information: e.g. the main influent variable (rainfall) is more important in the recurrent and static models than in the feed-forward one. For understanding flash floods genesis, recurrent and static models appear thus as better candidates, even if their results are not better.

1 Introduction

In the Mediterranean regions, flash floods due to heavy rainfalls frequently occur and cause numerous fatalities and costly damages. During the last few years, the south of France has been particularly exposed to these catastrophic situations. In such cases, damages can reach more than one billion euros, and, in only one event, there can be more than 20 fatalities [1]. Facing these issues, authorities need reliable forecasts for early warning purposes. Unfortunately, both the short-term rainfall forecasts and the processes leading to the discharge response remain poorly known at the space and time scales required. It is thus difficult to provide forecasts using the traditional coupling between a meteorological model and a physically based hydrological model.

Artificial Neural Networks therefore appear as an alternative paradigm as they are able to provide forecasts of an output (discharge) without making any other hypothesis on the system than the causality between rainfall and discharge. ANN have been applied in a wide variety of domains as they are essentially based on data and training [2]. They appear as particularly suitable for identifying the generating processes in hydrological time series because of their ability to model nonlinear dynamic systems [3,4]. However, due to their statistical origin, it is difficult to associate meaning to their internal parameters and they are rightly considered as black-box models. For this reason and to enhance the understanding of the behavior of the model, several works have been done to bring more transparency in the operating mode and introduced concepts of gray-box and transparent-box models [5,6]. In hydrology, several works have been conducted to make neural networks models more physically meaningful [6, 7, 8].

To be considered as gray-box (or transparent-box), ANN internal information or data must be accessible. In this paper, it will not be discussed *deep learning* itself, but an intermediate method to analyze the meaningful of internal information about neuronal models in hydrology operating on deep models. That method is termed **Knowledge eXtraction** (KnoX), it has been proposed by [7]. It was proved efficient on a fictitious basin, before being applied, by simulation, to estimate contributions and response times of various parts of a karst aquifer: the *Lez* aquifer (Southern France). It was later used by [8] for better apprehend the contributions of surface or underground processes in generation of floods on the Lavallette basin (Southern France).

Several studies were performed on the *Mialet* basin: first [4] showed that flash flood discharge can be forecasted by a multilayer perceptron with reasonable quality up to two-hours lead time; second, [9] showed that the initial value of the neural network parameters in flash floods forecasting has a major impact on the result. The purpose of this work is thus to better understand how the main variables influence the basin's outflow, regarding the model scheme used, in order to diminish the sensitivity of the model to the initialization of its parameters.

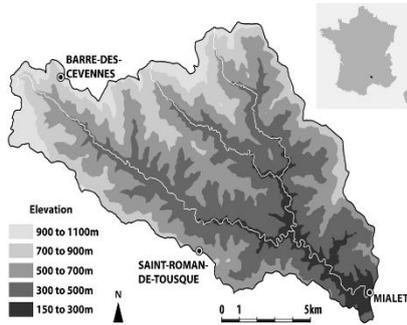
In the next sections, we will briefly present neural networks, their operating principles in hydrology, the deep multilayer perceptron used, as well as a reminder about the KnoX method and the models designed. The focus is

set on a discussion about the behavior of the variable's weights according to the model type used, by applying the KnoX method to extract that information.

2 Materials and methods

2.1 Study area: location and general description

The *Gardon de Mialet* basin covers 220 sq.km in southern France. It is part of the *Cévennes* range which is known as a preferential location for the well-known meteorological phenomenon named *cevenols episodes* (Fig. 1). These episodes consist in short duration (less than 2 days) very heavy rainfall events. The elevation of Mialet basin ranges from 150 *m.a.s.l.* to 1170 *m.a.s.l.* and its mean slope is about 33 %. As for the most of basins of the *Cévennes*, these characteristics lead to limited infiltration or underground flow and thus to a high drainage density. Its response time is relatively short: between 2-4 hours [4]. The area is dominated by a metamorphic formation essentially with 95 % of mica-schist and gneiss, which lead to a poorly porous and impermeable rocky sub-soil. The land use is almost homogeneous while covered by natural vegetation (chestnut trees, conifers, mixed forest and bush) for 92 %. The rest is shared between rocks and urban areas.



Typically, in Mediterranean regions, heavy rainfalls sometimes exceed 500 mm in only 24 h, to be compared to the 600 mm that fall on Paris annually. They are mainly produced by convective events, triggered either by relief, by a wind convergence, or by both. For example, in September 2002, the Gard (France) department has registered 687 mm of rainfall in 24h with 137 mm in only one hour at *Anduze* (a few km distant from *Mialet*).

Fig. 1. The study area (by Artigue, 2012)

2.2 Database

2.2.1. Presentation.

The database used in this study is essentially compounded with hourly observations from 1992 to 2002 and 5 minutes time step observations from 2002 to 2008 on three rain gauges and one hydrometric station at the outlet at *Mialet* (Fig.). From upstream to downstream, these stations are: BDC (*Barre des Cévennes*), SRDT (*Saint-Roman de Tousque*) and *Mialet* which coincide with the discharge station. They are all managed by the local Flood Forecasting Service (*SPC Grand Delta*). 58 events were extracted at 30 min time-step (based on rainfall events having at least 100 mm accumulation in 48 h on any of the rain gauges). Data description is synthetized in Tables 1 & 2.

Table 1. Data description

	Rainfall (mm)			Discharge	
	BDC	SRDT	Mialet	(m^3s^{-1})	($m^3s^{-1}km^2$)
Maximum (30 min)	33.3	41.8	62.0	819.3	3.72
Median (30 min)	0.3	0.3	0.2	29.3	0.13
Moy	1.0	1.3	1.2	43.4	0.20
Min	0	0	0	2.13	0.010

Table 2. Test event description

Event	Date	Duration	Maximum of discharge (m^3s^{-1})	Mean discharge (m^3s^{-1})	Cumulative rainfall (mm)	Intensity ($mm.h^{-1}$)
13	Sept. 00	26 h	454.2	70	230	40

2.3 Artificial Neural network

2.3.1. General presentation.

A neural network is a combination of parametrized functions called neurons that calculate their parameters thanks to a database using a training process [10]. The most popular model is the multilayer perceptron (MLP), which generally contains one or more hidden layers of nonlinear neurons and one output linear neuron. Each

hidden neuron computes a non-linear function of a weighted sum of the input variables, then the output neuron computes the linear combination of the outputs of the hidden ones.

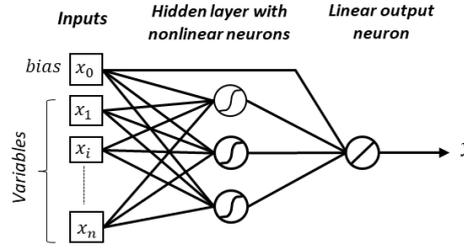


Fig. 2. Multilayer perceptron with a single hidden layer

The MLP is very popular due to its two main properties: universal approximation and parsimony. The first one states the capability to successfully approximate any differentiable function with an arbitrary level of accuracy [11]. The latter states how the multilayer perceptron needs fewer parameters to successfully fit a non-linear function, compared to others statistic model that linearly depend on their parameters [12]. The more general model of neuron calculates its output y as following:

$$y = f\left(\sum_{j=1}^n c_j \cdot x_j\right) = f(x_1, \dots, x_n; c_1, \dots, c_n) \quad (1)$$

with x_j , the input variable j ; c_j , the parameter linking the variable x_j to the neuron; $f(\cdot)$, the activation function (usually a sigmoid). The dynamic properties of the identified process can be considered thanks to three kinds of models [13].

- **Static model**

The static model is a digital filter with a finite impulse response. It calculates the following equation:

$$\hat{y}(k) = \varphi(\mathbf{x}(k), \dots, \mathbf{x}(k - n_r + 1), \mathbf{C}) \quad (2)$$

with $\hat{y}(k)$, the estimated output at the discrete time k ; φ_{rn} , the non-linear function implemented by the model; \mathbf{x} is the input vector; n , the sliding time-windows size defining the length of the necessary exogenous data; \mathbf{C} , the vector of the parameters. This model is known for having more parameters than the following models.

- **Recurrent model**

The recurrent model allows identification of dynamical processes (Infinite Impulse Response), it is implemented following the equation (3).

$$\hat{y}(k) = \varphi(\hat{\mathbf{y}}(k - 1), \dots, \hat{\mathbf{y}}(k - r); \mathbf{x}(k), \mathbf{x}(k - 1) \dots, \mathbf{x}(k - n_r + 1); \mathbf{C}) \quad (3)$$

With r , the order of the recurrent model; n_r , the depth of the sliding time-window used to consider the input variables. One must distinguish the recurrent variable (y) from the exogenous variables (\mathbf{x}). This model can deliver forecasts for an undetermined forecasting horizon providing the availability of the exogenous variables.

- **Feed-forward model**

In the feed-forward model, the recurrent input is substituted by the measurements of the process output at previous times step. This model is non recurrent; but it can identify dynamical processes. This model is the most used and generally provides the best results. Nevertheless, we have observed that it generally has difficulties to model the dynamics of the process (cited in Artigue et al 2012). It calculates:

$$\hat{y}(k) = \varphi(\mathbf{y}(k - 1), \dots, \mathbf{y}(k - r); \mathbf{x}(k), \mathbf{x}(k - 1) \dots, \mathbf{x}(k - n_r + 1); \mathbf{C}) \quad (4)$$

with $y(\cdot)$, the observed value of the modelled variable at the discrete time k .

These three categories of models will be compared in this study.

2.3.2. Training

As data-driven models, neural networks design is based on a database. Training consists in calculating the set of parameters of the model in order to minimize the least square cost function on the training set [10]. Because the model is non linear, this minimization is iteratively calculated.

Nevertheless, as the goal of the model is to be able to generalize the trained behavior to any set of data never seen, the quality of the model must be validated on another set, independent from the training set that is called

“test set”. The bias-variance dilemma [14] shows an important limitation: the training error is not representative of the test error, and the difference increases with the complexity of the model (*i.e.* the number of free parameters of the model). The bias-variance dilemma may be avoided using regularization methods.

2.3.2. Regularization methods

Early stopping

Early stopping was presented by [15] as a regularization method. It consists in stopping the training before the full convergence. To this end a supplementary subset, called stop set, is defined whose goal is to evaluate the ability of generalization of the model during the training. This subset is independent from the training set. Training is stopped when the error on the stop set begins to increase. The stop set is used to stop the training, the performances of this set are thus overestimated compared to those of the test set. Nevertheless, this set is usually (improperly) called “validation set” in the literature.

Cross validation

Proposed by [16], cross validation allows to select a model having the lower variance. To this end the training set is divided in K subset and the error is calculated on the remaining ($K-1$) subsets in the training set. After K trainings, the cross-validation score is calculated, for example by the mean of the previously obtained errors. Based on the cross validation score it is possible to select the model that has the lowest variance, minimizing by this way the bias on the training set and the variance on validation sets. This method allows to select input variables, the order (r), and the number of hidden neurons.

Ensemble model

Darras *et al.* [9] showed that, surprisingly, cross validation was not able to successfully select the best initialization of parameters. In order to diminish the sensitivity of the output to the parameter’s initialization, they propose to create an ensemble model of M members [17] and to calculate the output of the ensemble, at each time step, by the median of the M members.

2.3.3 Design of the model

In this study, regularization methods are applied by: (*i*) dividing the dataset in three subsets (training, stop and test sets), (*ii*) using cross correlation to select the architecture of the model in the following succession: inputs (n_r) except for rain gauges, order (r), number of hidden neurons (h), and (*iii*) using 20 members in the ensemble.

Three kinds of sliding window widths are tried based on the rainfall-runoff cross-correlogram.

2.3.4. Performance criteria

Several criteria are used to assess the performance of a model. The determination coefficient R^2 [18]; the *Synchronous percentage of the peak discharge* (S_{PPD}) and the *Peak delay* as two peak assessment criteria [4]. They are briefly detailed below:

- **R^2 criterion**

$$R^2 = 1 - \frac{\sum_{k=1}^n (y_k - \hat{y}_k)^2}{\sum_{k=1}^n (y_k - \bar{y}_o)^2} \quad (5)$$

with the same notations as previously.

The nearest than 1 the Nash-Sutcliff efficiency is, the best the results are. Nevertheless, this criterion can reach good values even if the model proposes bad forecasts.

- **Peak analysis**

The quality of the flood prediction is analyzed regarding the quality of the peak using two criteria defined by [4].

Synchronous percentage of the peak discharge: SPPD

The synchronous percentage of the peak discharge: SPPD [4] is a relevant criterion to assess flash flood modeling performance of a model on the peak discharge. It shows the simulation quality at the peak discharge through the ratio between the observed and simulated discharges at the observed peak discharge moment (k_o^{max}).

$$S_{PPD} = 100 \frac{\hat{y}_o^{k_o^{max}}}{y_o^{k_o^{max}}} \quad (6)$$

Peak delay (P_D)

The peak delay [4] indicates the duration between the maximum of simulated peak and measured peak. When the estimated peak is in advance, the peak delay is negative.

$$P_D = k_s^{max} - k_o^{max} \quad (7)$$

with k_s^{max} the instant of the peak of discharge (simulated or observed).

2.5. Extracting information: KnoX method

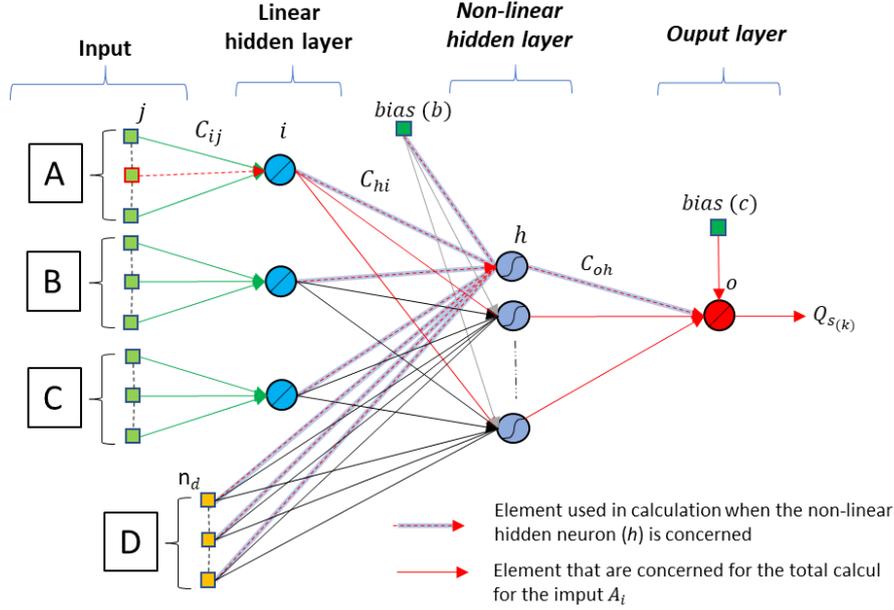


Fig. 3. Application of the KnoX method on the deep multilayer perceptron

$$P_{A(j)} = \frac{M|C_{ij}|}{\sum_{i=1}^{n_A} M|C_{ij}|} \sum_{h=1}^H \left(\frac{M|C_{hi}|}{\sum_{i=1}^{n_i} M|C_{hi}| + \sum_{d=1}^{n_d} M|C_{hd}| + b_h} \right) \left(\frac{M|C_{oh}|}{\sum_{h=1}^H M|C_{oh}| + c_o} \right) \quad (8)$$

and:

$$P_A = \sum_{j=1}^{n_A} (P_{A(j)}) \quad (9)$$

The KnoX method [8, 19] allows to calculate a simplified contribution of each input to the model output. This method is described for the general deep model (2 hidden layers) shown in Fig. 3. The principle of the method is that a contribution of an individual input variable can be quantified after training, by the product of the parameters linking this input to the output. The considered parameters are (i) “normalized” by the sum of the parameters linked to the same targeted neuron, and (ii) regularized by calculating the median of absolute values of their values for 20 different random initializations. This regularized value is noted as $M|C_{ij}|$ for the parameter C_{ij} linking the neuron (or input) j to the neuron i .

Regarding the model shown in Fig. 3, the contribution (P_A) of the input A (group of several delayed inputs) is the sum of the contributions of each individual delayed input of the group A. The equation calculating the contribution for just one element of the input A is provided in eq. (8). It is not possible to explain more comprehensively the method in the short present paper, so we suggest to the reader to refer to [8].

3 Results

Starting from previous works of [4] we chose the following exogenous variables: (i) *Barre des Cevennes* rain gauge, *Saint-Roman de Tousque* rain gauge and *Mialet* rain gauge, each one with a sliding window length $\{k, \dots, k-n_r+1\}$, (ii) the sum of the mean rain (over the three gauges) fallen from the beginning of the event. Of course, a bias input is used; several values were tried in order to evaluate the sensitivity of the KnoX method to its value.

3.1. Window widths selection thanks to correlation analysis

Widths of the rainfall windows applied to the model are selected thanks to cross correlation. Initially proposed by [20] Jenkins and Watts (1968), [1] generalizes the application of cross correlation in hydrology. The used equation in this study is presented in eq. (9).

$$C_{xy}(k) = \frac{Cov(x_i, y_{i+k})}{\sigma_x \sigma_y} = \frac{\frac{1}{n} \sum_{i=1}^{n-k} (x_i - \bar{x})(y_{i+k} - \bar{y})}{\sigma_x \sigma_y} \quad (10)$$

With $k = 0, 1, \dots$; where m is the truncation which is recommended to be $m=n/3$ (Mangin, 1984). [20] indicated that 2 hydrological variables can be considered as statistically independent if their cross-correlation is superior to 0.2. We thus select three possible lengths for the sliding windows of rain gauges inputs: (i) the number of time step between $C_{xy} = 0$ and $C_{xy} = 0.2$, that defines the memory effect; (ii) the window between $C_{xy} = 0.2$ (positive slope) and $C_{xy} = -0.2$ (negative slope) and (iii) all the m positive values of C_{xy} . Based on [20] the correlations between gauges as well as response times are indicated in Table 3.

Table 3. Correlation analysis of the data

<i>Rain gauge</i>	Mialet (h)	SRDT (h)	BDC (h)
Average response-time	2	3	4.5
Response-time range	1 – 3.5	2.5 – 4.5	4 – 5.5
Rainfall-discharge average cross-correlation	0.40	0.455	0.44
Rain gauge cross-correlation	Mialet	--	0.58
	SRDT	--	0.61

3.2. Model selection

A partial cross-validation score was operated on a subset of 17 most intense events in the database [3]. The number of hidden neurons was increased from 1 to 10. The best model was chosen according to the highest cross-validation score S_v estimated as following:

$$S_v = \frac{1}{K} \sqrt{\sum_{i=1}^K |E_i|^2} \quad (11)$$

Where E_i is the validation error of the subset i used in partial cross validation.

The output values are the result of the median of the outputs of an ensemble of 20 members differing only by their initialization before training.

Three bias values are considered (0.01; 0.1; 1), three depths of sliding windows (see section 3.1) and three kinds of models (see section 2.3), 27 different models have been designed following the procedure indicated in section 2.3.3. The best one in each kind of models has been chosen, regarding the test event, in order to have efficient models to analyze. Architectures presented in Table 4 were thus selected.

Table 4. Selected models

Input variables	Static	Recurrent	Feed-Forward
Rain-gauge window width (n_r) (BDC/SRDT/Mialet)	32/32/23	27/28/20	32/32/23
Rain cumul window width	3	3	3
Order (r)	x	3	3
Number of hidden nonlinear neurons	2	10	5
Bias value	1	0.01	1

3.3 Results

Obtained test set hydrographs are shown in the Fig 4 and their performances described in Table 5. It appears in **Fig.4** and **Table 5** that the best results are provided by the feed-forward model. This is usual because the feedforward model uses the previous observations of the modelled variable in input. The recurrent model is usually not as efficient but exhibits better dynamics, which is also frequently observed [4]. The static model presents an acceptable performance, being able to generate 63% of the peak discharge.

Table 5. The models performances on the test set

<i>Model</i>	R^2	<i>SPPD</i> %	$P_D(0.5h)$
Static	0.83	63,3	1
Recurrent.	0.89	78.5	0
Feed-Forward	0.99	99.3	1

After having verified that the models are convenient, it is possible to apply the KnoX method. The extracted contributions are presented in **Table 5**.

Regarding the rainfalls, one can note that in general, SRDT is the station with the highest contribution. The contributions do not change significantly for *Mialet* through all the models. BDC and *Mialet* are probably affected by their location close to the border of the basin whereas SRDT is close to the middle of the basin.

Regarding the balance between the state variables and the rainfalls, it appears that when the previous observed discharge is used as an input variable, it brings almost 50 % of the contribution to the output. This observation means that the model does not pay enough attention to rain inputs and this could be the reason of the sensitivity to parameters initialization. Beside this, it also appears that the state variables in the static model have lesser contribution than they do in the other two models. In general, from the static model to the feed-forward one, the total contributions of the state variables are respectively 45%, 61 % and 65 %, where the biggest parts are imputed to the previous observed discharge (feed-forward). These observations are fully consistent and the results seem highly interpretable.

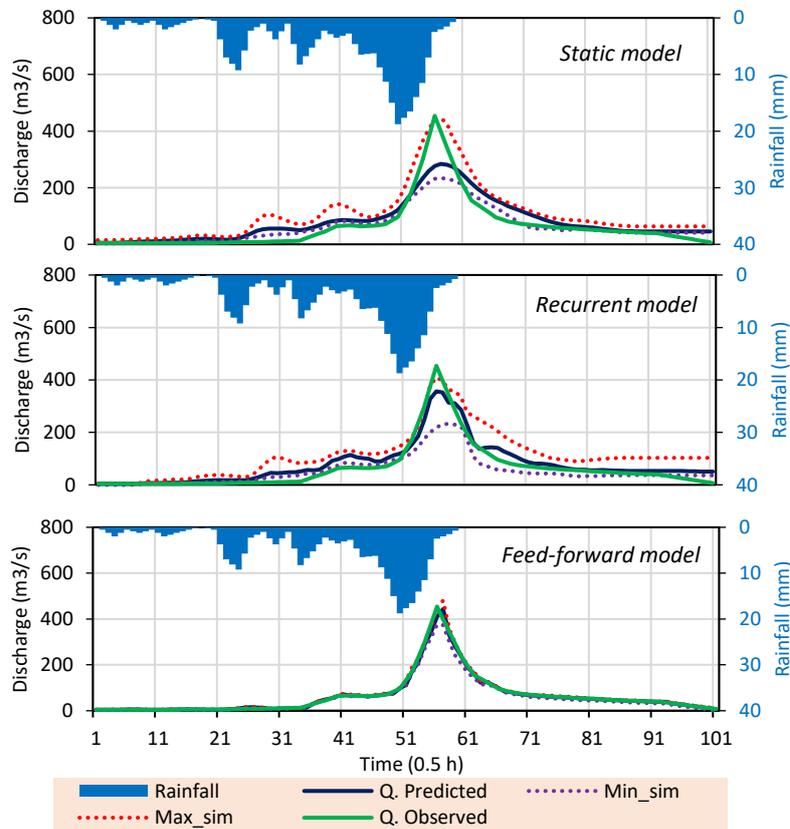


Fig. 4. Hydrographs for the test set. Min_sim and Max_sim correspond to the minimum and maximum values of the ensemble model. Q is the median of the 20 members of the ensemble.

Table 6. Contributions (P_A) for the variables, from each model, expressed in %.

<i>Name of variable</i>	<i>Static</i>	<i>Recurrent</i>	<i>Feed-forward</i>
<i>BDC</i>	13 %	12 %	5 %
<i>SRDT</i>	31 %	17 %	22 %
<i>Mialet</i>	11 %	11 %	9 %
<i>Cumulated rainfall</i>	31 %	20 %	12 %
<i>Previous Q. obs</i>	--	--	45 %
<i>Previous Q. calc</i>	--	25 %	--
<i>bias</i>	14 %	16 %	8 %

4 Interpretation

These results show how the kind of model can modify the contribution of explanatory variables on an observed phenomenon. Thus, some kind of models must be preferred when it comes to represent physical relations. It is also shown that the mean cumulative rainfall used here as a state variable plays a great role in models where the previous discharge is not used as input. This state variable seems to have a great interest in hydrologic modelling. The value of the bias, surprisingly, seems to have a role. It is usually interpreted as the base flow. Nevertheless, its behavior is consistent: it shows more involvement when the previous observed discharges are not used as input; then by complementarity with the humidity information, it guides the models to acceptably approximate the real discharge information.

5 Conclusion

Prediction of flash flood events is a very challenging task in the *Cévennes* range. It was previously realized using neural networks but sometimes appeared difficult to understand because of the specific behaviors of the models. In order to be able to improve these models, the present work takes steps to better understand the processes involved in such events. To this end, the KnoX method, developed to extract information from a neural network model was applied to the *Gardon de Mialet* Basin. The obtained results show that by using relevant variables properly combined on whatever the network used here, efficient model can be built out. Besides, the KnoX method allows to see how the variables are handled by the model to approximate the phenomenon. There has been evidence that the variables do not express themselves in the same way through the different models used. As it is understandable, sometimes, the choice for a model is commanded by the situations in presence. The information extracted from the network can probably be used to compare to some physical meaningful characteristics of watershed or events, such as the Thiessen polygons, the response time, the cross correlation etc. It provided also some guidelines to deal with the sensitivity of the model to the parameter's initialization.

6 Acknowledgement

The authors thank the METEO-France weather agency, the SPGD flood-forecasting agency for providing rainfall datasets. Our gratitude is extended to Bruno Janet for the stimulating collaboration shared with the SCHAPI Unit, and to Roger Moussa and Pierre Roussel-Ragot for the helpful discussions and support. The constant effort made by Dominique Bertin and the Geonosis Company to enhance and develop the neural network software RNF Pro are thereby acknowledged as well.

7 References

1. Rouzeau, M., Xavier M., and Pauc, J.C. 2010. "Retour d'expériences des inondations survenues dans le département du Var les 15 et 16 juin 2010." http://cgedd.documentation.developpement-durable.gouv.fr/documents/cgedd/007394-01_rapport.pdf.
2. J. Roberts, Stephen, and Will Penny. 1981. *Neural Networks: Friends or Foes?* Sensor Review. Vol. 17. London: MCB University Press.
3. Toukourou M., Johannet A., Dreyfus G.,Ayril P.A. 2011. Rainfall-runoff Modeling of Flash Floods in the Absence of Rainfall Forecasts: the Case of "Cévenol Flash Floods", *App. Intelligence*, 35 2,178-189.
4. Artigue, G,et al. 2012. "Flash Flood Forecasting in Poorly Gauged Basins Using Neural Networks: Case Study of the Gardon de Mialet Basin (Southern France)". *NHESS*, 12(11): 3307-24.
5. Oussar, Yacine, and Gérard Dreyfus. 2001. "How to Be a Gray Box: Dynamic Semi-Physical Modeling." *Neural Networks* 14 (9): 1161-72. [https://doi.org/10.1016/S0893-6080\(01\)00096-X](https://doi.org/10.1016/S0893-6080(01)00096-X)
6. Johannet, Anne, B Vayssade, and Dominique Bertin. 2007. "Neural Networks: From Black Box towards Transparent Box - Application to Evapotranspiration Modelling." *Int. Journal of Comp. Int.* 24 (1): 162.
7. Kong-A-Siou, L., et al, S.: KnoX method, or Knowledge eXtraction from neural network model. Case study on the Lez karst aquifer (southern France), *J. Hydrol.*, 507, 19–32.
8. Darras, T., *et al.* 2015. Identification of spatial and temporal contributions of rainfalls to flash floods using neural network modelling: case study on the Lez basin (southern France) *Hydrol. Earth Syst. Sci.*, 19, 4397–4410, 2015
9. Darras, T., Johannet, A., Vayssade, B., Kong-A-Siou, L. and Pistre, S. (2014). Influence of the Initialization of Multilayer Perceptron for Flash Floods Forecasting: How Designing a Robust Model, (ITISE 2014), Ruiz, IR, Garcia, GR Eds, 687-698.
10. Dreyfus, G. 2005. *Neural networks, methodology and applications*, Springer, Berlin.
11. Hornik, Kurt, Maxwell Stinchcombe, and Halbert White. 1989. "Multilayer Feedforward Networks Are Universal Approximators." *Neural Networks* 2 (5): 359-66.

12. Barron, A R. 1993. "Universal Approximation Bounds for Superpositions of a Sigmoidal Function." *IEEE Trans. Inf. Theor.* 39 (3): 930-45. <https://doi.org/10.1109/18.256500>.
13. Nerrand, O., P. Roussel-Ragot, L. Personnaz, G. Dreyfus, and S. Marcos. 1993. "Neural Networks and Nonlinear Adaptive Filtering: Unifying Concepts and New Algorithms." *Neural Comp* 5 (2): 165-99.
14. Geman, Stuart, Elie Bienenstock, and René Doursat. 1992. "Neural Networks and the Bias/Variance Dilemma." *Neural Computation* 4 (1): 1-58.
15. Sjöberg, J., *et al.* 1995. "Nonlinear Black-Box Modeling in System Identification: A Unified Overview." *Automatica* 31 (12): 1691-1724.
16. Stone, M. 1976. "Cross-Validatory Choice and Assessment of Statistical Predictions (With Discussion)." *Journal of the Royal Statistical Society: Series B (Methodological)* 38 (1): 102-102.
17. Dietterich T.G., 2000. Ensemble Methods in Machine Learning, in J. Kittler and F. Roli (Ed.), *First Int. Workshop on Multiple Classifier Systems, LNCS*, p. 1-15, New York: Springer Verlag.
18. Nash, J.E., and J.V. Sutcliffe. 1970. "River Flow Forecasting through Conceptual Models Part I - A Discussion of Principles." *Journal of Hydrology* 10 (3): 282-90.
19. Jenkins, G.-M., Watts, D.-G. 1969 "Spectral analysis and its applications", Holden-Day, - 525 pages
20. Mangin, Alain. 1984. "Pour Une Meilleure Connaissance Des Systèmes Hydrologiques à Partir Des Analyses Corrélatoire et Spectrale." *Journal of Hydrology* 67 (1-4): 25-43.

Points de valorisation de la publication

Les travaux décrits par cette publication montrent que les performances sont meilleures pour le modèle statique dirigé que pour le récurrent, lui-même plus performant que le modèle statique. Pour autant, ce dernier modèle, dépourvu d'information sur l'état du système, à l'exception du cumul depuis le début de l'événement, qui n'est qu'un proxy de cette information, est celui dont l'extraction des informations portées par les couches profondes semble faire le plus de sens en termes de répartition spatiale des précipitations. Dans une moindre mesure, on retrouve ces traits dans le modèle récurrent. Le modèle statique dirigé, quant à lui, propose une répartition des poids associés aux variables d'entrée sur la couche la plus profonde moins proche de ce que l'on pourrait décrire physiquement. Cela vient du fait que l'information des dernières observations de débit prend un poids très important qui tend à écraser celui des variables exogènes.

Il en résulte qu'un modèle laissé libre de sélectionner ses poids se rapproche d'un certain sens physique présent dans la nature et que le meilleur compromis entre performance et signification physique des paramètres est le modèle récurrent.

Ces modèles de simulation ne sont évidemment pas utiles en conditions opérationnelles, sauf à les alimenter avec des prévisions de pluie suffisamment exactes. Notons que dans le cas du modèle dirigé, la dépendance aux sorties précédemment observées est élevée, ce qui peut être un facteur limitant en conditions opérationnelles. Ceci étant, les conclusions restent entières et transposables à des modèles de prévision.

Chapitre VI : Deep Multilayer Perceptron for Knowledge Extraction: Understanding the Gardon de Mialet Flash Floods Modelling

Introduction de la publication

Ce chapitre est une extension de la publication présentée au niveau du *Chapitre V*. Il a fait l'objet d'une publication sous forme d'un « book chapter » dans Springer Nature book « Contributions to statistic », édition 2020. Dans ce chapitre, les informations extraites des couches profondes des modèles neuronaux sont comparées à la variabilité spatiale des pluies sur le bassin versant, aux temps de réponses en fonction du pluviomètre considéré et aux corrélogrammes pluies-débit. Une analyse est également réalisée sur l'influence de la valeur du biais qui peut être considéré comme analogue au débit de base.

Deep Multilayer Perceptron for Knowledge Extraction: Understanding the *Gardon de Mialet* Flash Floods Modelling

Bob E. Saint Fleur ^{1,2}, Guillaume Artigue ¹, Anne Johannet ¹, Séverin Pistre ²

¹ LGEI, IMT Mines Alès, Alès, France

² Hydrosociences Montpellier, Univ Montpellier, CNRS, IRD, 34090 Montpellier, France

guillaume.artigue@mines-ales.fr

Abstract. Flash floods frequently hit Southern France and cause heavy damages and fatalities. To enhance persons and goods safety, official flood forecasting services in France need accurate information and efficient models to optimize their decisions and policy in crisis management. Their forecasting is a serious challenge as heavy rainfalls that cause such floods are very heterogeneous in time and space. Such phenomena are typically nonlinear and more complex than classical flood events. This analysis had led to consider complementary alternatives to enhance the management of such situations. For decades, artificial neural networks have been proved very efficient to model nonlinear phenomena, particularly rainfall-discharge relations in various types of basins. They are applied in this study with two main goals: first modelling flash floods on the *Gardon de Mialet* basin (Southern France); second, extract internal information from the model by using the KnoX: knowledge extraction method to provide new ways to improve models. The first analysis shows that the kind of nonlinear predictor strongly influences the representation of information: e.g. the main influent variable (rainfall) is more important in the recurrent and static models than in the feed-forward one. For understanding "long-term" flash floods genesis, recurrent and static models appear thus as better candidates, despite their lower performance. Besides, the distribution of weights linking the exogenous variables to the first layer of neurons is consistent with the physical considerations about spatial distribution of rainfall and response time of the hydrological system.

Keywords: Neural networks, flash floods, knowledge extraction, deep learning.

1 Introduction

In the Mediterranean regions, flash floods due to heavy rainfalls frequently occur and cause numerous fatalities and costly damages. During the last few years, Southern France has been particularly exposed to these catastrophic events. In such cases, in only one event, there can be more than 20 fatalities, and damages that can reach more than one billion euros, in only one event [1]. Facing these issues, authorities need reliable forecasts for early warning purposes. Unfortunately, both the short-term rainfall forecasts and the processes leading to the discharge response remain poorly known at the space and time scales required. It is thus difficult to provide forecasts using the traditional coupling between a meteorological model and a physically based hydrological model.

Artificial neural networks therefore appear as an alternative paradigm as they are able to provide forecasts of an output (discharge) without making any other hypothesis on the system than the causality between rainfall and discharge. Artificial neural networks have been applied in a wide variety of domains, as they are essentially based on data and training [2]. They appear as particularly suitable for identifying the generating processes in hydrological time series because of their ability to model nonlinear dynamic systems [3,4]. However, due to their statistical origin, it is difficult to associate meaning to their internal parameters, and they are rightly considered as black-box models. For this reason and to enhance the understanding of the behavior of both the model and the physical processes, several works have been done to bring more transparency in the operating mode and introduced concepts of gray-box and transparent-box models [5, 6]. Some other works have been conducted to make neural networks models more hydrologically meaningful [6, 7, 8].

2 Material and methods

2.1 Study area: location and general description

The *Gardon de Mialet* basin covers 220 sq.km in Southern France. It is part of the *Cévennes* range, which is known as a preferential location for the well-known meteorological phenomenon named "cévenols episodes" (**Fig. 1**). These episodes consist in short duration (less than 2 days) very heavy rainfall events.

The elevation of the *Gardon de Mialet* basin ranges from 150 m.a.s.l. to 1170 m.a.s.l. and its mean slope is about 33 %. As for the most of basins of the *Cévennes*, these characteristics lead to limited deep infiltration or deep underground flow, and thus to a high drainage density. Its response time is relatively short: between 2-4 hours [4]. The area is dominated by a metamorphic formation with 95 % of mica-schist and gneiss, which leads to a poorly porous and impermeable rocky sub-soil. The land use is almost homogeneous while covered by natural

vegetation (chestnut trees, conifers, mixed forest and bush) for 92 %. The rest is shared between rocks and urban areas.

Typically, in Mediterranean regions, heavy rainfalls sometimes exceed 500 mm in only 24 h, to be compared to the 600 mm that fall on Paris annually. They are mainly produced by convective events, triggered either by relief, by a wind convergence, or by both. For example, in September 2002, the Gard (France) department has registered 687 mm of rainfall in 24h with 137 mm in only one hour at *Anduze* (a few kilometers distant from *Mialet*).

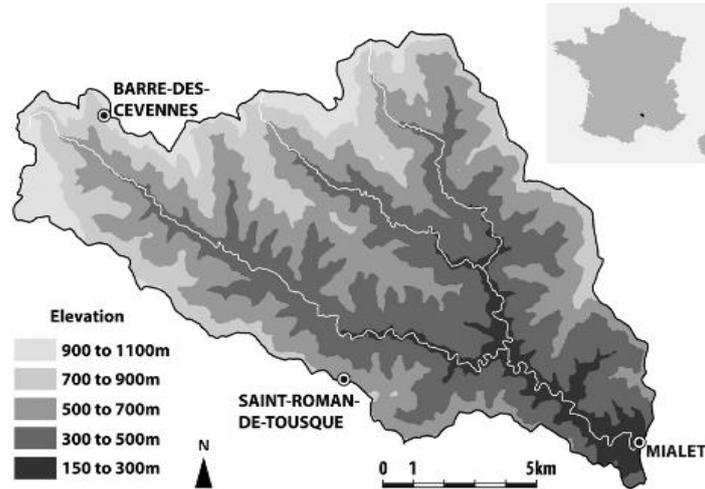


Fig. 1. The study area (Artigue, 2012)

2.2 Database

The database used in this study is essentially compounded with hourly observations from 1992 to 2002, and 5 minutes time-step observations from 2002 to 2008, on three rain gauges and one hydrometric station at the outlet at Mialet (**Fig. 1**). From upstream to downstream, these stations are: BDC (Barre des Cévennes), SRDT (Saint-Roman de Tousque) and Mialet which coincides with the discharge station. They are all managed by the local Flood Forecasting Service (SPC Grand Delta). 58 events were extracted at 30 minutes time-step (based on rainfall events having at least 100 mm accumulation in 48 h on any of the rain gauges). Data description is synthetized in **Tables 1 & 2**.

Table 4. Data description

	Rainfall (mm)			Discharge	
	<i>BDC</i>	<i>SRDT</i>	<i>Mialet</i>	(m^3s^{-1})	($m^3s^{-1}km^{-2}$)
<i>Maximum (30 min)</i>	33.3	41.8	62.0	819.3	3.72
<i>Median (30 min)</i>	0.3	0.3	0.2	29.3	0.13
<i>Moy</i>	1.0	1.3	1.2	43.4	0.20
<i>Min</i>	0	0	0	2.13	0.010

Table 5. Test event description

Event	Date	Duration	Maximum of discharge (m^3s^{-1})	Mean discharge (m^3s^{-1})	Cumulative rainfall (mm)	Intensity ($mm.h^{-1}$)
13	Sept. 00	26 h	454	70	230	40

2.3 Artificial neural networks

As widely explained in [9, 4], three kinds of neural networks models have been used in this study: a static model, a recurrent model and a feedforward model. The same references should provide the reader guidance about the implemented methods for the control of the bias-variance dilemma and of overtraining (early stopping, cross-validation, ensemble model) and about the performance criteria used (R^2 criterion and peak analysis). Only the part about knowledge extraction is reminded here, due to its important role in the study.

2.4 Extracting information: KnoX method

First, the KnoX method is applied to a specific architecture, based on multilayer perceptron, which represents the behavior of the physical process, in order to constrain the model to represent this physical behavior [7]. As the rain is essentially added in the first step of the rainfall-runoff transformation, we have introduced one layer of linear neurons implementing the addition of rains fallen at different time-steps (delayed rains). This supplementary layer is called "i" (linear hidden neurons) in Fig 2. The second hidden layer (non-linear hidden layer) calculates a non-linear combination of the "locally added" rains.

The KnoX method [7, 8, 9] allows calculating a "simplified" contribution of each input to the model output. This method is described for the general deep model (two hidden layers) shown in Fig. 2. The principle of the method is that a contribution of an individual input variable can be quantified, after training, by the product of the parameter's chain linking this input to the output. The considered parameters are (i) "normalized" by the sum of the parameters linked to the same targeted neuron, and (ii) made independent from the model initialization by calculating the median of absolute values of their values for 20 different random initializations. This regularized value is noted as $^M|C_{ij}|$ for the parameter C_{ij} linking the neuron (or input) j to the neuron i .

As the value of the sigmoid is not taken into account in eq. 2, this contribution can be seen as the contribution of the "linearized" model. Nevertheless the model is really a non-linear model.

Regarding the model shown in Fig. 2, it appears that inputs are applied in several groups, for example A, B, \dots . Each group corresponds to a variable, for example the rain gauge of *Mialet*, or the previous discharge (D). As the output depends dynamically on these inputs, following a complex and unknown multi-scale relation, these inputs are applied at several time-step in order to allow the model to estimate these multi-scale relations. Thus, the contribution (P_A) of the grouped inputs A (including several delayed inputs) is the sum of the contributions of each individual delayed input of the group A . The equation calculating the contribution for just one element (the value for the delay j) of the input A is provided in eq. (1). Unhopefully, it is not possible to explain more comprehensively the method in the short present paper, so we suggest to the reader to refer to [7, 8].

$$P_{A(j)} = \frac{^M|C_{ij}|}{\sum_{i=1}^{n_A} ^M|C_{ij}|} \sum_{h=1}^H \left(\frac{^M|C_{hi}|}{\sum_{i=1}^{n_i} ^M|C_{hi}| + \sum_{d=1}^{n_d} ^M|C_{hd}| + b_h} \right) \left(\frac{^M|C_{oh}|}{\sum_{h=1}^H ^M|C_{oh}| + c_o} \right) \quad (1)$$

and:

$$P_A = \sum_{j=1}^{n_A} (P_{A(j)}) \quad (2)$$

Where the categories of parameters C_{ij} , C_{hi} , C_{oh} , C_{hd} are shown on Fig. 2; n_A is the number of inputs in the group A ; H the number of hidden non-linear neurons; n_j the number of hidden linear neurons (first hidden layer); n_d is the number of delayed inputs of the group D ; b_h is the bias inputted to the non-linear hidden input and c_o is the bias parameter inputted to the output neuron.

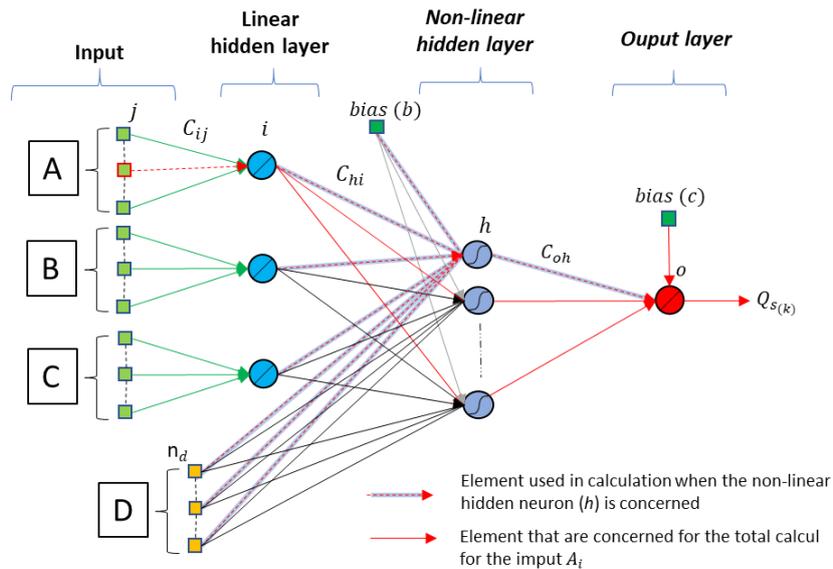


Fig. 2. Application of the KnoX method on the deep multilayer perceptron

3 Results

3.1 Choice of variables

Starting from previous works of [4], we chose the following exogenous variables: (i) *Barre des Cevennes* rain gauge, *Saint-Roman de Tousque* rain gauge and *Mialet* rain gauge, each one with a sliding window length $\{k, \dots, k-n_r+1\}$, (ii) the sum of the mean rain (mean calculated over the three gauges) fallen from the beginning of the event. Of course, a bias input is used; several values were tried in order to evaluate the sensitivity of the KnoX method to its value.

Depending on the kind of considered model, states variables can be added: previous observed discharges for the feed-forward model, and previous estimated discharges for the recurrent one. The static model only takes rains and mean rains into account [9].

3.2 Model selection

Model selection is a key issue of machine learning. The goal is to define accurately the architecture of the model managing the bias-variance tradeoff. This was done in this following work [10] using cross correlation, cross validation, and early stopping using the following rules.

- Hyper-parameters are adjusted for each one of the three kinds of model (static, feed-forward, recurrent): input sliding windows width (n_A, n_B, n_C, n_D), number of non-linear hidden neurons (h).
- Widths of the rainfall windows applied to the model, $\{n_A, n_B, n_C\}$, are selected thanks to cross correlation [11]. Initially proposed by [12], [13] generalizes the application of cross correlation in hydrology. The used equation in this study is presented in eq. (3).

$$C_{xy}(k) = \frac{Cov(x_i, y_{i+k})}{\sigma_x \sigma_y} = \frac{\frac{1}{n} \sum_{i=1}^{n-k} (x_i - \bar{x})(y_{i+k} - \bar{y})}{\sigma_x \sigma_y} \quad (3)$$

With $k \in \mathbb{N}^+$; the truncation m , which is the maximum value of k , is recommended to be $m=n/3$. [12] indicated that two hydrological variables can be considered as statistically independent if their cross-correlation is superior to 0.2. Starting from this work, we selected three possible lengths for the sliding windows of rain gauges inputs: (i) the number of time-step between $C_{xy}=0$ and $C_{xy}=0.2$, that defines the memory effect (called *memory window*); (ii) the sliding window between $C_{xy}=0.2$ (positive slope) and $C_{xy}=0.2$ (negative slope) (called *strong correlation window*), and (iii) all the m positive values of C_{xy} (called *full correlation window*). Based on [12], the correlations between gauges and response times are indicated in **Table 3**.

Table 3. Correlation analysis of the data

Rain gauge	Mialet (h)	SRDT (h)	BDC (h)
Average response-time	2	3	4.5
Response-time range	1 – 3.5	2.5 – 4.5	4 – 5.5
Rainfall-discharge average cross-correlation	0.40	0.455	0.44
Rain gauge cross-correlation	Mialet	--	0.58
	SRDT	--	--
			0.61

- The partial cross-validation score was operated on a subset of K events, the 17 most intense events in the database [3].

- The number of hidden neurons was increased from 1 to 10. The best model was chosen according to the highest cross-validation score S_v estimated as following:

$$S_v = \frac{1}{K} \sqrt{\sum_{i=1}^K |E_i|^2} \quad (4)$$

Where E_i is the validation error of the subset i used in partial cross validation.

- An ensemble model is used in order to regularize on the initialization of parameters; moreover, the output values are the result of the median of the outputs of an ensemble of 20 members differing only by their initialization before training [8].

- Three bias values were considered (0.01; 0.1; 1), three depths of sliding windows and three kinds of models (see section 2.3), *i.e.* 27 different models have been designed following the procedure indicated in [9]. The best one in each kind of models has been chosen, regarding the test event, in order to have the most efficient models to analyze. Architectures presented in **Table 3** were thus selected.

Table 3. Selected models

Input variables	Static	Recurrent	Feed-Forward
Rain-gauge window width (n_r) (BDC/SRDT/Mialet)	32/32/23	27/28/20	32/32/23
Cumulative rainfall window width	3	3	3
Order (r) / Previous observed outputs	x	3	3
Number of hidden nonlinear neurons	2	10	5
Bias value	1	0.01	0.01

3.3 Discharge estimation

As shown in [9, 4], the best results are provided by the feed-forward model. This is usual because the feedforward model uses the previous observed output as a state variable. The recurrent model is not as efficient but exhibits better dynamics, which is also frequently observed [4]. The static model presents an acceptable performance, being able to generate 63% of the peak discharge (**Table 4; Fig 3**).

Table 4. Models performances on the test set

Model	R^2	SPPD %	$P_D(0.5h)$
Static	0.83	63,3	1
Recurrent.	0.89	78.5	0
Feed-Forward	0.99	99.3	1

3.4 Contributions of input variables

After having verified that the models are convenient, it is possible to apply the KnoX method. The extracted contributions are presented in **Table 5** [9].

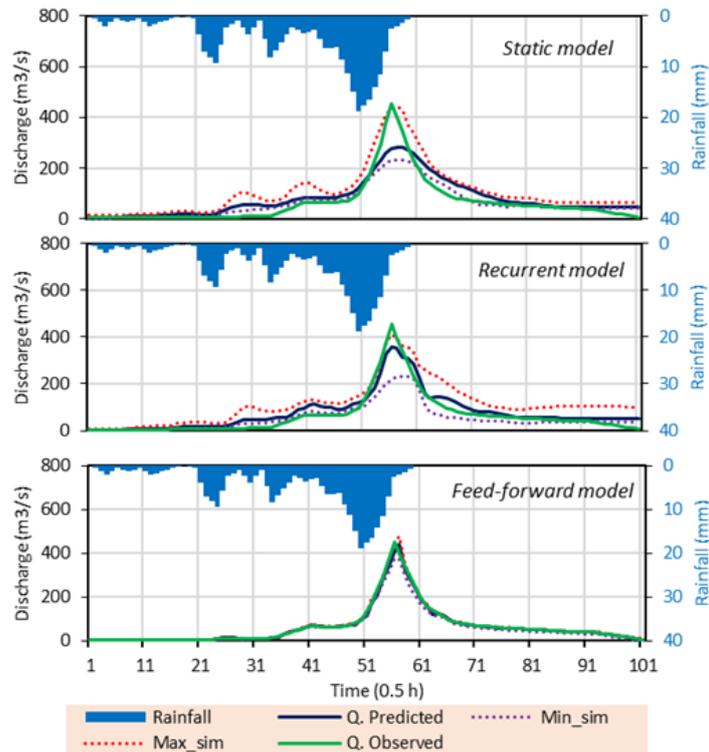


Fig. 3. Hydrographs for the test set. Min_sim and Max_sim correspond to the minimum and maximum values of the ensemble model. Q is the median of the 20 members of the ensemble.

It is interesting to compare the relative weights of the three rain gauges with a classic method dedicated to distribute rainfall on a watershed and widely used in hydrology: the Thiessen polygons method (or Voronoï polygons). This comparison is presented in **Fig. 4**.

Table 5. Contributions (P_x) for the variables, from each model, expressed in %.

<i>Name of variable X</i>	<i>Static</i>	<i>Recurrent</i>	<i>Feed-forward</i>
<i>BDC</i>	11 %	10 %	9 %
<i>SRDT</i>	31 %	17 %	22 %
<i>Mialet</i>	13 %	12 %	5 %
<i>Cumulated rainfall</i>	31 %	20 %	12 %
<i>Previous Q. obs</i>	--	--	45 %
<i>Previous Q. calc</i>	--	25 %	--
<i>bias</i>	14 %	16 %	7 %
<i>Total</i>	100 %	100 %	100 %

As *Mialet* (MIA) is at the outlet of the basin and *Barre-des-Cévennes* (BDC) at the top of the basin, they are both represented with less contribution than *Saint-Roman-de-Tousque* (SRDT, near the middle of the basin) by the Thiessen polygon method. It is more or less also the case for the neural network models, with a very similar distribution to Thiessen distribution for the static model, being a little more different for the recurrent model and even more for the feed-forward model (providing the best results).

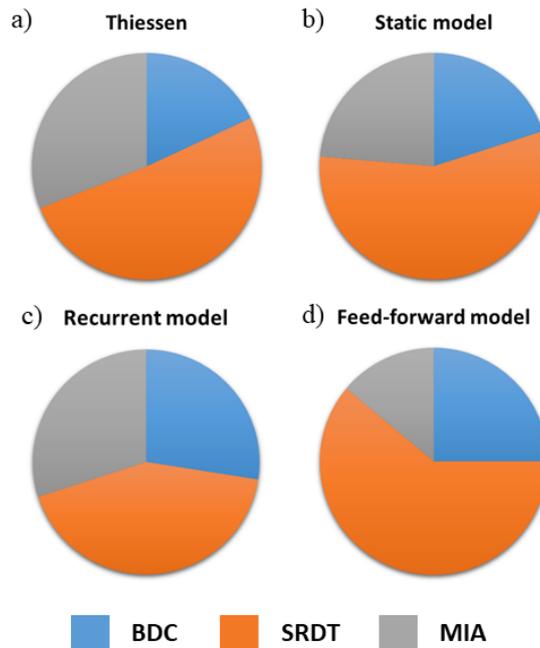


Fig. 4. Thiessen method weights (a) and relative weights from the models of the three rain gauges (b, c, d).

3.5 Results: contributions as a function of time windows

Here, we have considered the distribution of contributions amongst the time delay in the first layer of parameters (arriving to the linear neurons in Fig 2). We compare the sum of these contributions (for the three rain gauges) to the cross-correlogram of the average rainfall (average of the three rain gauges) and the discharge. This comparison involves the three selected models presented in section 3.1 (Fig. 5).

The static model shows the greatest similarity with the cross-correlogram, for the total contributions and for the relative contributions of each rain gauge. Regarding the response time (time corresponding to the peak of the cross-correlogram) the static model seems also to be the best. This result is logical because the variables taken into account by the static models are similar to those considered by the cross-correlogram: only rains.

For the three models, the SRDT rain gauge is the most represented in most of the time lags considered and not only in general, as shown in section 3.4.

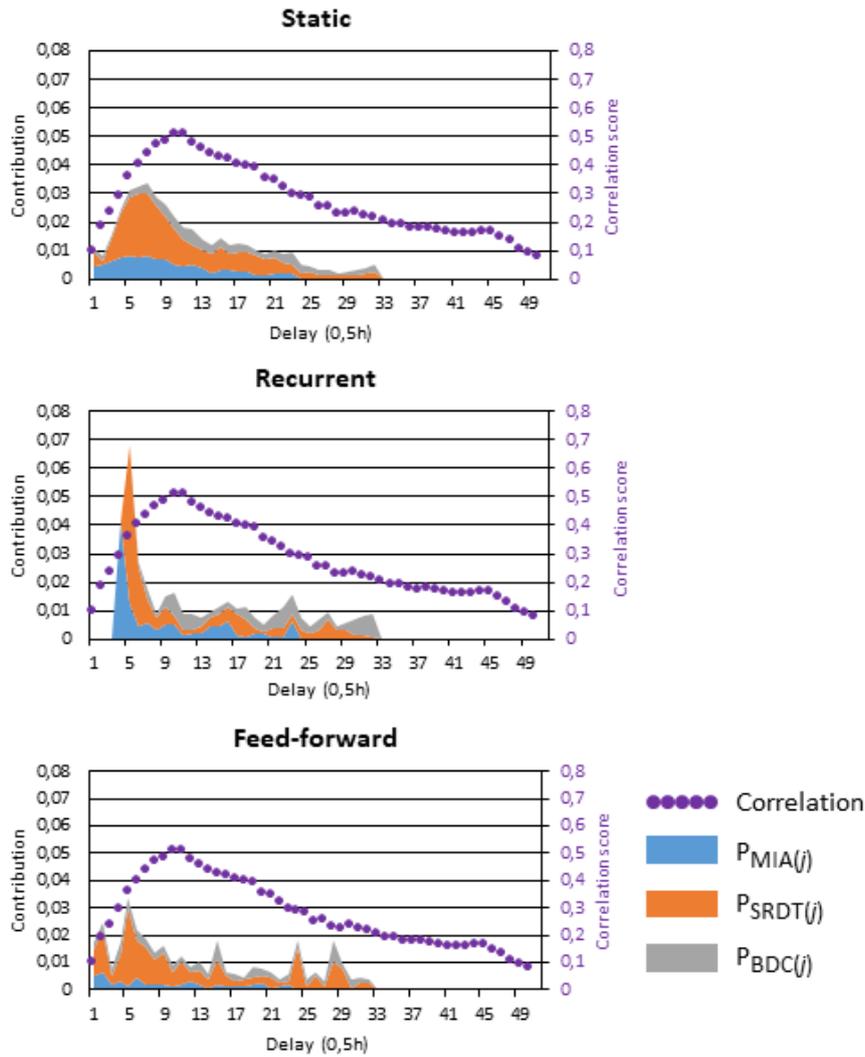


Fig. 5. Cross-correlogram (mean rainfalls-Mialet discharge) and distribution of contributions calculated as indicated in section 2.4.

3.6 Results: effects of the bias

Before obtaining the selected models, many different combinations were tried during optimization. Among these combinations, three values of bias have been experimented, each separated by an order of magnitude: 0.01, 0.1 and 1. The **Fig. 6** shows the contributions of *Saint-Roman-de-Tousque* amongst the delays of the input time window, for the three types of models (static, recurrent, feed-forward) and for the three time-windows defined in section 3.2 (memory, strong correlation and full correlation windows). The other rain gauges have not been presented here due to the large number of figures it would have produced; but the *Saint-Roman-de-Tousque* station is representative of the three rain gauges from this point of view.

It can be noticed that the bias value does not deeply impact the contributions of the input variables. In particular, it does not change the general shape of these contributions even if in some cases, moderate amplitude differences appear.

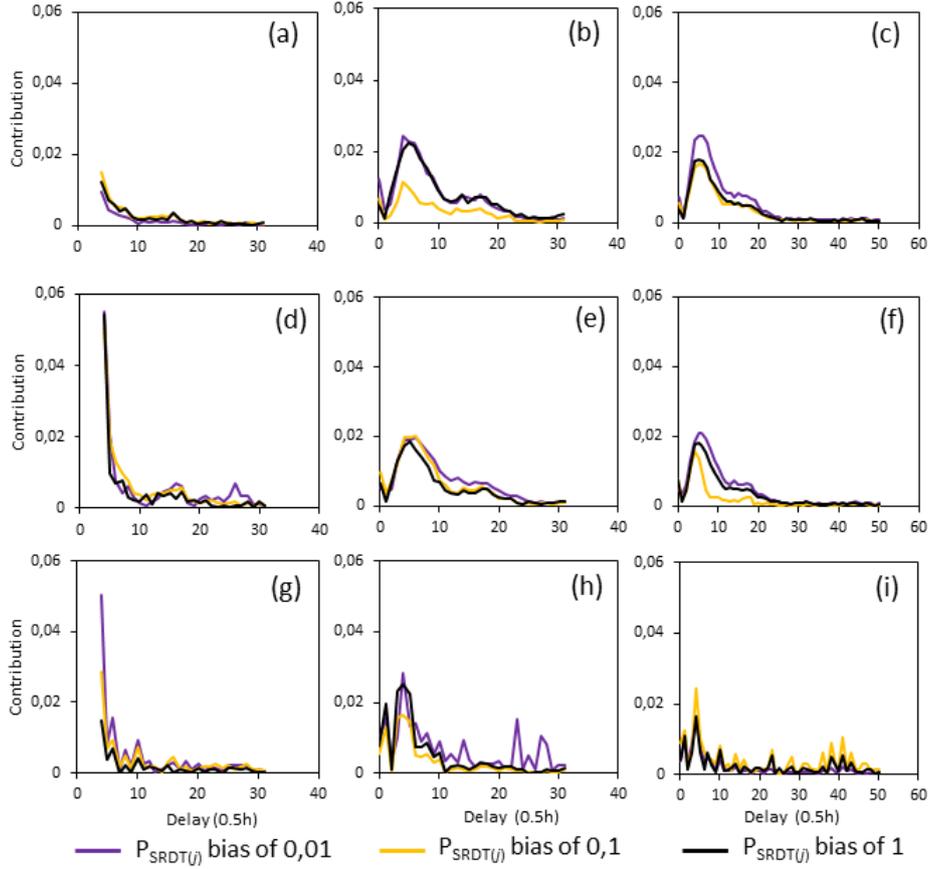


Fig. 6. *Saint-Roman-de-Tousque* contributions calculated as indicated in section 2.4 with different bias and in different modeling configurations: (a), (b) and (c) are for static models; (d), (e) and (f) are for recurrent models; (g), (h), and (i) are for feedforward models whereas (a), (d) and (g) are for memory windows; (b), (e) and (h) are for strong correlation windows and (c), (f) and (i) are for full correlation windows.

4 Discussion

These results show how the kind of model takes into account explanatory variables on an observed phenomenon. Even if they use the same exogenous variables in the same context, their performances and behaviors are different due to their configuration and architecture.

4.1 Selecting a model type for physical knowledge extraction

Analyzing the contributions assigned to each input variable (Table 5), it appears that:

- The static model strongly uses exogenous variables (total contribution of 55%) and uses an important contribution (31%) to the cumulated rainfall that is useful to represent the soil saturation and could thus be considered as a substitute to a state variable.
- The recurrent model uses mostly previous estimated discharge (25%), whereas the total contribution of cumulated rainfall (20%) and of exogenous variables (40%) is lower than for the static model.
- The feedforward model uses a smaller contribution for rains (12% for cumulated rainfall and 36% for exogenous variables) whereas previous observed values of discharge contribution is predominant (45%).

As foreseen by [14], the optimal type of model is strongly linked to the quality of explanatory information that is given to the model during the training phase. Here, we show that, despite its low performance, the static model is forced to represent, the physical relationship between exogenous variables and the output, whereas the recurrent model and the feedforward model are helped in this task resp. by the previous estimated or observed discharge. Consequently, the total use of the exogenous variables decreases when state variables information increases. If we compare the relative contributions of the three rain gauges with the Thiessen polygons, we observe a decrease of the similarity while state variables are added.

Finally, in this study, the best tradeoff between model performance and knowledge extraction capacities seems to be provided by the recurrent model. Nevertheless, this conclusion is based on one test set, it should be confirmed by further studies.

4.2 Response time and contributions

The cross-correlation provides a simple linear representation of the behavior of the modeled system and allows estimating the response time. Here again, while the contributions of state variables appear, the similarity with the correlogram decreases. This could be interpreted as a confirmation that recurrent and feed-forward models represent well the behavior that takes profit of the rich information provided by the previous discharge input (estimated or simulated): the accumulation of previously fallen rains. The less the model is helped by the previous discharge input, the more it is forced to represent well the role of recent and ancient rainfalls. This appears in Fig. 6.d and 6.g with great values of recent rains contributions; on the equivalent contributions in 6.e and the "noisy" contribution of the Fig 6.i (feed-forward with the maximum window-width).

4.3 Bias input importance

The bias input plays a role that is usually interpreted in hydrology as the base flow (remaining discharge when there is no rainfall). In this case, its contribution is consistent: it is significantly less involved in the calculation of the output when the previous observed discharges are used as input (the previous base flow is thus applied by the inputs). In the other cases, it seems to guide the models to acceptably approximate the discharge information when necessary.

If the bias input seems necessary to guide the model, its value does not deeply change the distribution of the contribution of the rain gauges as a function of the instant of the time window. One could suppose that changing an order of magnitude in the bias input value can easily be counterbalanced during the training step by applying a proportional modification to the weights applied to this input.

5 Conclusions and perspectives

Flash flood forecasting is a very challenging task, especially in the *Cévennes* range. Several examples of robust forecasts using neural networks have been published but the results did not always allow understanding how close the model was to the physical behavior of the basin, in addition of being close to the observed output. The obtained results prove again that when using relevant and properly combined variables on any of the networks used here, an efficient model can be implemented.

Nevertheless, enhancing these models and applying them to an increasing number of basins, in a context of climate change, and with various characteristics, requires a better understanding of the processes involved in their operation as well as in such flood events. For this purpose, the KnoX method, developed to extract information from a neural network model, was applied to the *Gardon de Mialet* basin. This method allows understanding how the variables are handled by the model to approximate the modeled phenomenon. First it appears that the bias input was consistently used to model the base flow. Then, interestingly, there has been evidence that the variables do not express themselves in the same way depending on the different models used. It was known that the choice of a model must be driven by the modeling goal (for example a recurrent model for a long-term prediction). Besides being driven by the modeling goal, it appears that the choice for a model might be guided by the situation: availability (real-time and historical) of data, quality and explanatory nature of the data. In this study, this results in three kinds of model: static, recurrent, feed-forward, showing increasing performances while taking into account more realistic state variables. On the other hand, if assess the performance of a kind of model by the ability to extract physical information from it, the ranking is reversed and the less the model considers state variables, the more the design of its estimator will adopt behaviors that mimic the physical processes.

Finally it appears that the KnoX method shows very interesting capabilities; the next steps will consist in generalizing this method to other sites and other rainfall events in the *Cévennes* range, with an increasing complexity in the physical processes to extract (dams and/or karst systems for example)..

6 Acknowledgements

The authors thank IMT Mines Alès that funded the present study; the SPGD flood-forecasting agency for providing datasets. Our gratitude is extended to B. Janet for the stimulating collaboration shared with the SCHAPI and to R.

Moussa and P. Roussel-Ragot for the helpful discussions and support. The constant effort made by D. Bertin and the Geonosis Company to enhance and develop the neural network software RNF Pro are thereby acknowledged as well.

7 References

- Rouzeau, M., Martin, X., Pauc, J.C. Retour d'expérience des inondations survenues dans le département du Var les 15 et 16 juin 2010. http://cgedd.documentation.developpement-durable.gouv.fr/documents/cgedd/007394-01_rapport.pdf (2010).
- Roberts, S. J., Penny, W. Neural networks: friends or foes? *Sensor Review*, 17(1), 64-70 (1997).
- Toukourou M., Johannet, A., Dreyfus, G., Ayrat, P.A. Rainfall-runoff Modeling of Flash Floods in the Absence of Rainfall Forecasts: the Case of "Cévenol Flash Floods". *App. Intelligence*, 35 2,178-189 (2011).
- Artigue, G., Johannet, A., Borrell, V., Pistre, S. Flash Flood Forecasting in Poorly Gauged Basins Using Neural Networks: Case Study of the Gardon de Mialet Basin (Southern France). *NHESS*, 12(11): 3307-24 (2012).
- Oussar, Y., Dreyfus G. How to Be a Gray Box: Dynamic Semi-Physical Modeling. *Neural Networks* 14 (9): 1161-72 (2001).
- Johannet, A., Vayssade, B., Bertin, D. Neural Networks: From Black Box towards Transparent Box - Application to ETP Modelling. *Int. J. of Comp. Int.* 24 (1): 162 (2007).
- Kong-A-Siou, L., Cros, K., Johannet, A., Borrel-Estupina, V., Pistre, S. KnoX method, or Knowledge eXtraction from neural network model. Case study on the Lez karst aquifer (southern France), *J. Hydrol.*, 507, 19–32 (2013).
- Darras, T., Borrel-Estupina, V., Kong-A-Siou, L., *et al.* Identification of spatial and temporal contributions of rainfalls to flash floods using neural network modelling: case study on the Lez basin (southern France). *Hydrol. Earth Syst. Sci.*, 19, 4397–4410, (2015).
- Saint-Fleur, B., Artigue, G., Johannet, A., Pistre, S. Knowledge Extraction (KnoX) in Deep Learning: Application to the Gardon de Mialet Flash Floods Modelling. In: *Proceedings ITISE-2019*, pp. 178-189. Granada, 25th-27th September (2019).
- Kong-A-Siou, L., Johannet, A., Borrell V., Pistre, S., 2012. Optimization of the generalization capability for rainfall–runoff modeling by neural networks: the case of the Lez aquifer (southern France). *Environ Earth Sci* 65, 2365–2375.
- Kong-A-Siou, L., Johannet, A., Borrell, V., 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, 367–380.
- Jenkins, G.M., Watts, D.G. *Spectral analysis and its applications*. Holden-Day (1969).
- Mangin, A. Pour Une Meilleure Connaissance Des Systèmes Hydrologiques à Partir Des Analyses Corrélatoire et Spectrale. *Journal of Hydrology* 67 (1-4): 25-43 (1984).
- Nerrand, O., Roussel-Ragot, P., Personnaz, L., Dreyfus, G., Marcos, S. Neural Networks and Nonlinear Adaptive Filtering: Unifying Concepts and New Algorithms. *Neural Comp* 5 (2): 165-99 (1993).

Points de valorisation de l'article

Sur le même bassin versant expérimental et la même base de données que dans le *Chapitre V*, les compléments apportées ici sont :

- L'utilisation de la corrélation croisée pour dimensionner les fenêtres temporelles d'entrée,
- L'influence de la valeur de l'entrée constante sur les modèles,
- La comparaison du poids des variables exogènes avec celle de la méthode des polygones de Thiessen,
- La comparaison de la distribution temporelle de la contribution des variables sur la longueur de la fenêtre temporelle avec celle de la corrélation croisée.

Premièrement, on n'a pas noté de différence majeure entre les résultats lorsque l'on a fait varier l'entrée constante de 0.01 à 1, en passant par 0.1. Tout juste note-t-on que cette valeur, traditionnellement considérée comme un proxy du débit de base, est moins utilisée pour le modèle dirigé, ce qui semble accréditer cette hypothèse.

Comme dans le *Chapitre V*, on a pu noter que le modèle statique présentait une proximité plus grande avec la méthode des polygones de Thiessen en termes de répartition des poids des trois pluviomètres. Vient ensuite le modèle récurrent et enfin le modèle dirigé. L'hypothèse exposée dans le *Chapitre V* selon laquelle le débit observé tend à écraser les contributions des variables exogènes semble se confirmer à la lumière de ces résultats. Si l'on considère que la méthode des polygones de Thiessen apporte une information pertinente quant à la répartition spatiale des précipitations, il semble que le modèle soit en mesure d'approcher cette répartition lorsqu'il n'est pas trop influencé par des variables d'état.

Enfin, ici encore, l'apparition de contributions liées aux variables d'état semble faire diminuer la ressemblance entre la distribution temporelle du poids des entrées et la corrélation croisée entre ces entrées et la sortie. Le modèle statique est donc le plus ressemblant, suivi du modèle récurrent et enfin du modèle dirigé, ces deux derniers montrant que sur une grande partie de leurs entrées anciennes, les poids affectés s'apparentent à du bruit.

Chapitre VII : Initialisation des couches profondes avec la corrélation croisée

1 Introduction

Nous avons vu dans la présentation du Chapitre II (section 2.3.4.1) que l'initialisation du modèle neuronal pouvait générer une forte variabilité des sorties. Pour pallier cette variabilité nous avons proposé d'utiliser un modèle d'ensemble comme méthode de régularisation. Deux hypothèses sont émises pour expliquer ce comportement :

- Comme les incertitudes sur les mesures des précipitations sont fortes pour la prévision des crues éclair, ce « bruit » contenu dans les données, très important, pourrait générer cette sensibilité à l'initialisation car le contenu informatif des données serait trop faible.
- Même avec une seule couche cachée, la couche la plus « profonde » a du mal à être apprise, peut-être du fait du problème du gradient évanescent.

Ce chapitre a pour objet de tester la seconde hypothèse. Nous envisageons ici de créer des modèles profonds pour réaliser la prévision des crues du Gardon de Mialet, et d'initialiser les couches profondes des modèles avec une information issue des signaux : la corrélation croisée entre la variable d'entrée considérée et la variable de sortie. Ceci permettrait *in fine* de faciliter l'apprentissage des couches superficielles.

2 Approche expérimentale

2.1 Architecture générique

L'architecture profonde présentée en Figure 33 a été conçue pour, d'une part, permettre un traitement non linéaire dans les couches profondes, variable par variable, et d'autre part pour avoir suffisamment de couches dans le modèle pour que le phénomène du gradient évanescent puisse se manifester.

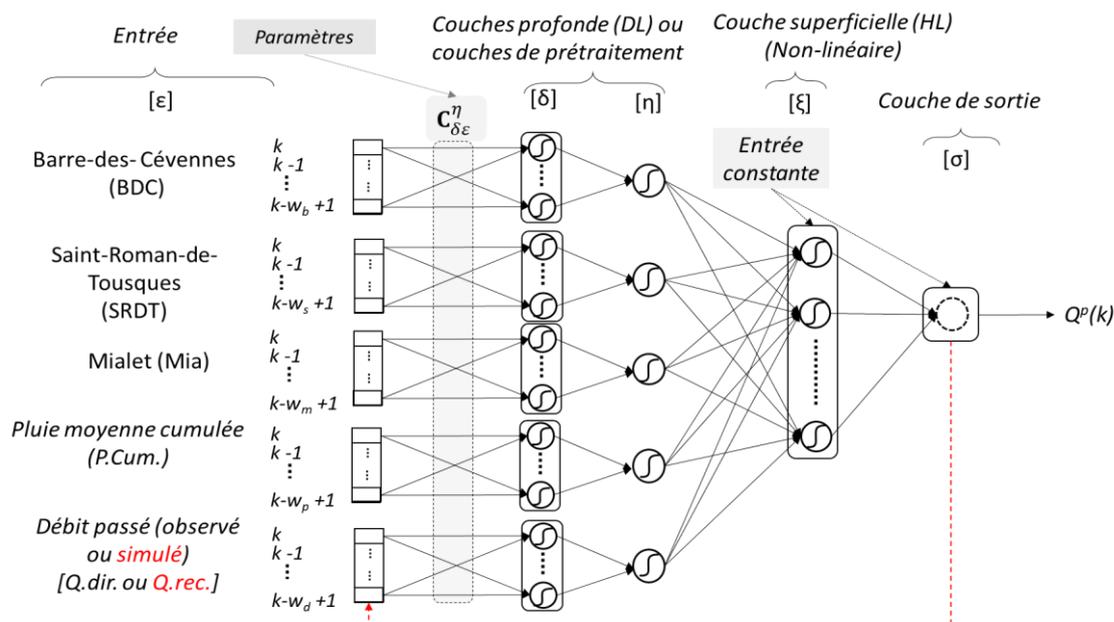


Figure 33 : Architecture générique du perceptron multicouche utilisé

Chaque variable, par exemple la pluviométrie à *Barre des Cévennes*, est appliquée à un « module » de deux couches comportant : la couche la plus profonde appartenant à la couche δ de la Figure 33, et connectée à son neurone de « sortie » (attention ce n'est pas une sortie du réseau, nous ne disposons pas de valeurs mesurées pour cette couche), délivrant la contribution de la variable considérée, appartenant à la couche η . La couche η comporte donc 5 neurones non linéaires qui effectuent une compression de l'information pour limiter le nombre de paramètres et donc la complexité. Ces neurones jouent le rôle d'entrées aux couches superficielles : la couche ξ non linéaire et la couche σ linéaire, qui effectuent une transformation non linéaire des informations de la couche η .

Les paramètres qui vont être initialisés avec les corrélations croisées sont les paramètres liant les entrées ε à la couche δ noté $C_{\delta\varepsilon}^{\eta}$ (Figure 33), autrement appelés « paramètres de la couche δ ».

2.2 Sélection des hyperparamètres de l'architecture

La sélection des hyperparamètres doit intervenir sur tous les éléments que nous devons ajuster, ou faire varier, pour définir complètement l'architecture :

- Fenêtres temporelles des variables exogènes et de la variable récurrente,
- Initialisation des paramètres de la couche la plus profonde (aléatoire/corrélation croisée)
- Nombre de neurones des couches ξ et δ ,
- Méthode de régularisation (arrêt précoce, modération des poids),
- Type d'architecture (récurrent/dirigé)

Dans ce qui suit chaque point sera évalué tour à tour, puis nous synthétiserons pour présenter les résultats des meilleurs modèles.

L'horizon de prévision est fixé à 2h sans prévision des pluies afin de pouvoir comparer les résultats obtenus avec ceux des travaux précédents (Artigue et al, Saint Fleur et al., 2021 in NHES).

2.2.1 Fenêtres temporelles des variables exogènes et de la variable récurrente

2.2.1.1 Variables exogènes

Les fenêtres temporelles des *variables de pluviométrie* ont été fixées en utilisant la corrélation croisée entre chaque variable et la sortie de débit. Figure 34 présente comment le temps de réponse de l'hydrosystème peut être défini par le retard correspondant au maximum du corrélogramme.

A ce temps de réponse, nous ajoutons 2 pas de temps afin de prendre en compte la variabilité de ce temps de réponse en fonction des événements de pluie (Artigue, 2012 ; Kong A Siou *et al.*, 2011) et cette durée définit la profondeur de fenêtre temporelle de la variable de pluviométrie considérée.

Tableau 3 : Dimensions des profondeurs des fenêtres temporelles des variables d'entrée

Variable	BDC	SRDT	Mialet	Pluie moyenne cumulée	Débit observé/simulé précédemment
<i>Temps de réponse (T_r)</i>	11	7	6	--	--
<i>Profondeur historique (T_r+2)</i>	13	9	8	3	3

Concernant les variables *observées de débit* et de *pluie moyenne cumulée*, nous choisissons arbitrairement, par expérience et en s'inspirant des travaux faits en théorie du contrôle, une fenêtre temporelle qui permette au modèle de disposer de : (i) la position dans la crue (la valeur du débit, et du cumul de pluie au temps discret k), la valeur de la vitesse (2 valeurs du débit, et du cumul de pluie

au temps discrets k et $k-1$), (ii) la valeur de l'accélération (3 valeurs du débit et du cumul de pluie au temps discrets k , $k-1$ et $k-2$).

On utilise donc trois valeurs pour appliquer la variable de débit observé en entrée : $Q^p(k)$, $Q^p(k-1)$, $Q^p(k-2)$, comme pour appliquer la variable de cumul des précipitations.

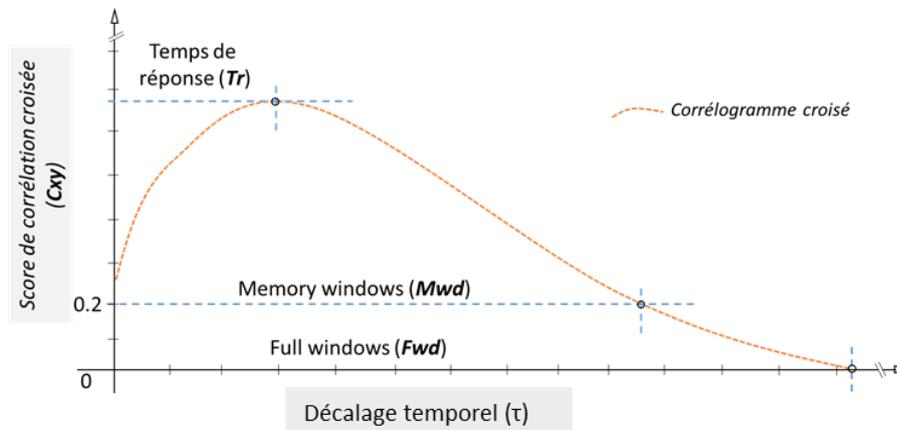


Figure 34: Repère de sélection de la profondeur historique par la corrélation croisée

2.2.1.2 Variable récurrente

Lorsque le débit estimé par le modèle est renvoyé en entrée pour constituer un modèle récurrent, trois valeurs de débits estimés : $Q(k)$, $Q(k-1)$, $Q(k-2)$ sont utilisées, par analogie avec ce qui a été présenté concernant les entrées de débit observé.

2.2.2 Initialisation de la couche la plus profonde (aléatoire/corrélation croisée)

Pour réaliser l'initialisation de la couche la plus profonde (couche δ de Figure 33) nous envisageons plusieurs possibilités qui sont décrites ci-après.

2.2.2.1 Paramètres initialisé ou fixés

Les valeurs des paramètres de la couche δ peuvent être initialisées avec les corrélations croisées au même moment qu'a lieu l'initialisation aléatoire des autres paramètres, ou l'initialisation peut être conservée, et ces paramètres ne sont alors pas appris : ils sont définitivement *fixés*.

On dira que les paramètres de la couche δ sont *initialisés*, ou *fixés*, à la valeur de la corrélation croisée.

2.2.2.2 Calcul de la corrélation croisée partielle ou totale

Nous avons vu en section 2.3.5.2 du *Chapitre II* que la validation croisée était réalisée dans ce travail grâce à la validation croisée partielle. La question se pose alors de déterminer quelle corrélation croisée nous utilisons pour initialiser la couche δ . En d'autres termes, la corrélation croisée peut-être calculée sur tous les événements de la base, ou seulement sur les événements qui sont utilisés pour la validation croisée. On va alors parler de corrélation croisée *partielle* ou *totale*.

En synthèse, cinq possibilités d'initialisation de la couche des paramètres de la couche d sont donc possibles ; elles sont respectivement notées :

- Paramètres aléatoires et libres (*Randomly initialized* notée iRNd)
- Paramètres initialisés par la corrélation croisée partielle et appris (initialisation notée iCP)
- Paramètres initialisés par la corrélation croisée totale et appris (initialisation notée iCT)
- Paramètres initialisés par la corrélation croisée partielle et fixés (initialisation notée fCP)
- Paramètres initialisés par la corrélation croisée totale et fixés (initialisation notée fCT)

Dans les deux derniers cas, les paramètres ne sont pas appris, seulement initialisés et laissés fixés.

2.2.3 Facteur d'échelle à appliquer aux corrélations croisées

L'initialisation des paramètres d'un réseau de neurones est une étape déterminante pour le bon fonctionnement du modèle. Lorsque l'apprentissage est réalisé avec la rétropropagation, en particulier pour calculer le gradient de la fonction de coût, les paramètres :

- ne peuvent pas avoir tous la même valeur, autrement les incréments aux paramètres sont tous identiques
- ne peuvent pas être trop grands sous peine de saturer les sigmoïdes de la couche qu'ils alimentent. En effet si le potentiel des neurones est trop grand, car les paramètres sont trop grands, alors la sortie des neurones se trouve dans la zone saturée de la sigmoïde et la dérivée de la sigmoïde est presque nulle : l'apprentissage ne démarre pas (Éq.28).

Pour éviter ces deux difficultés, on choisit en général d'initialiser les paramètres de manière aléatoire entre deux valeurs qui dépendent du nombre d'entrées : plus il y a d'entrées qui s'ajoutent dans le potentiel, plus les valeurs d'initialisation doivent être faibles. Dans ce travail les paramètres sont initialisés aléatoirement de la manière suivante :

1. C_{ij} est tiré avec une distribution aléatoire et uniforme dans l'intervalle : $[-1, 1[$
2. La valeur des paramètres est réduite : $C_{ij} = \frac{C_{ij}}{\sigma n_e}$,

Où σ est l'écart-type de la distribution aléatoire initiale et n_e est le nombre d'entrées arrivant sur le neurone considéré.

Ainsi lorsque les paramètres sont initialisés aléatoirement, l'amplitude des paramètres est réduite par l'écart type de la distribution aléatoire et par le nombre des entrées arrivant à chaque neurone.

Il est clair que si l'on initialise les paramètres avec la corrélation croisée, il faudra que les deux types d'initialisations conduisent aux mêmes ordres de grandeurs des paramètres initiaux, sous peine de fausser les comparaisons.

Pour cela nous avons introduit un facteur d'échelle au calcul des corrélations croisées : nous les multiplions par l'inverse du nombre d'entrée. La Figure 35.a illustre une comparaison entre les valeurs des paramètres à l'issue d'une initialisation aléatoire et d'une initialisation par la corrélation croisée avec le facteur d'échelle (courbe **bleue**), ou la corrélation croisée centrée (courbe **noire**). On peut y noter que les valeurs des paramètres sont assez faibles, mais du même ordre de grandeur (10^{-2}).

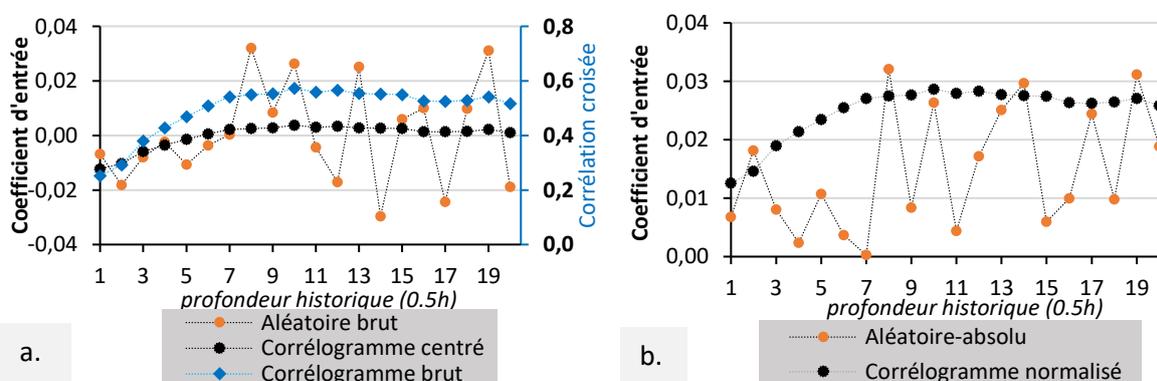


Figure 35 : Comparisons des valeurs des paramètres en fonction de la profondeur de la fenêtre temporelle d'entrée de 20 pas de temps. Graphe de gauche : valeurs brutes ; graphe de droite : valeurs absolues.

La Figure 35.b présente la comparaison entre les valeurs absolues, d'une part des paramètres initialisés aléatoirement, et d'autre part des paramètres initialisés avec la corrélation croisée et le facteur d'échelle.

L'application du coefficient de mise à l'échelle a ainsi permis de ramener les valeurs de corrélations croisées à un même ordre de grandeur que celles caractérisant les valeurs aléatoires.

Les variables ont été sélectionnées suivant la significativité de leur corrélation croisée au débit tout en tenant compte des premières expériences menées par Artigue (2012) sur ce bassin versant. Par contre, pour ces expériences-ci, du fait que nous nous limitons à la comparaison de l'effet du type d'initialisation sur les modèles, tout ce qui définit les modèles sera identique par ailleurs.

2.2.4 Sélection de la complexité

Les nombres de neurones cachés des couches δ et ξ ont été sélectionnés en se basant sur les scores de persistance en validation croisée, avec l'initialisation aléatoire. Les complexités ainsi obtenues ont ensuite été utilisées pour les autres cas de figure car la comparaison entre les différentes initialisations n'est possible que s'il n'y a qu'un élément qui change : l'initialisation. Les nombres de neurones sont donc identiques pour toutes les configurations d'un même type d'architecture. Savoir si l'initialisation de la couche δ par la corrélation croisée diminue ou augmente la complexité des modèles est une question que nous avons dû laisser en suspens dans ce travail, faute de temps.

En revanche, deux processus de sélection de la complexité ont été menés séparément pour les modèles récurrents et les modèles dirigés. La sélection dans les deux cas a été faite en deux étapes :

1. Complexité de la couche cachée ξ : sélection par la validation croisée partielle en considérant des complexités croissantes allant de deux à dix neurones cachés non-linéaires.
2. Complexité des couches profondes δ : sélection par la validation croisée partielle réalisée successivement en commençant par les couches reliées aux variables ayant les plus fortes corrélations croisées avec la sortie, et pour des nombres de neurones allant de deux à cinq neurones non-linéaires. Il n'était en effet pas possible de tester simultanément le dimensionnement de ces cinq couches cachées du fait de l'explosion du nombre de configurations ainsi engendrées.

2.3 Choix des méthodes de régularisation et de leurs hyperparamètres

Lors de sa thèse portant sur le Gardon d'Anduze, M. Toukourou a comparé l'utilisation des deux méthodes de régularisation : modération des poids et arrêt précoce (Toukourou, Johannet and Dreyfus, 2009). Ses conclusions étaient que la modération des poids n'apportait pas une meilleure généralisation ou même la dégradait par rapport à l'arrêt précoce. Cependant ces travaux avaient été faits en prenant en compte des contraintes qui ne nous paraissent plus utiles aujourd'hui : l'hyperparamètre γ contrôlait à la fois le terme intégrant l'erreur quadratique par un facteur (γ) et celui contrôlant la norme des paramètres ($1-\gamma$) (Éq.31). Il nous a paru plus cohérent, toujours pour ne faire varier qu'un élément à la fois, de n'utiliser l'hyperparamètre γ que pour pondérer le terme de norme des paramètres en le faisant varier plus finement. La règle utilisée est donc la suivante :

$$J^{mp}(\mathbf{C}) = J + \gamma \|\mathbf{C}\|^2 \quad (\text{Éq. 33})$$

Avec :

- γ , l'hyperparamètre gérant la pondération entre l'erreur quadratique et la norme des paramètres au carré $\|\mathbf{C}\|^2$,
- J , la fonction de coût quadratique, comme précédemment définie (Éq.19).

Comme ces deux méthodes de régularisation ont un objectif identique, à ce stade nous pouvons utiliser :

- L'arrêt précoce uniquement (notée ci-après APs) ;
- La modération des poids uniquement (notée ci-après MPs) ;
- Les deux méthodes de manière conjointe (notée ci-après ApMp).

2.3.1 Sélection de l'hyperparamètre γ .

Pour sélectionner la valeur optimale de l'hyperparamètre γ de la méthode de régularisation des poids, nous avons procédé à une exploration des valeurs assez large : de 10^{-1} à 10^{-4} par facteur de 10. L'évolution du score de validation croisée est représentée en Figure 36. Ces expérimentations ont été faites avec les configuration suivantes :

- 50 itérations d'apprentissage, 10 modèles d'ensemble,
- Deux architectures : dirigé, récurrent.

On note que pour les deux architectures (dirigée ou récurrente), les résultats sont sensiblement les mêmes et que la valeur de 10^{-4} semble la meilleure.

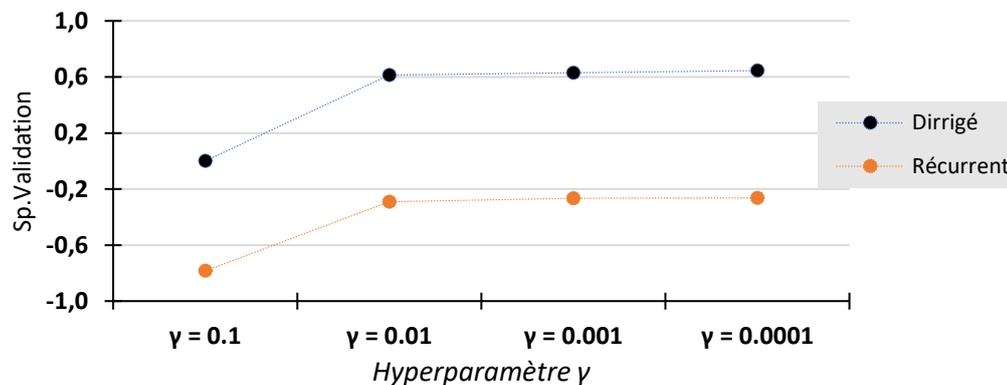


Figure 36 : Score de la validation croisée (persistance) en fonction de l'hyperparamètre de la modération des poids

3 Résultats et interprétation

3.1 Complexité des modèles

3.1.1 Neurones de la couche superficielle (couche cachée classique)

Le nombre de neurones de la couche cachée ξ des deux types d'architectures (dirigé et récurrent) a été sélectionnée en prenant la persistance comme critère de validation croisée. Les Figure 37.a&b présentent l'évolution de ce score en fonction du nombre de neurones cachés de la couche ξ pour, respectivement, les modèles dirigé et récurrent.

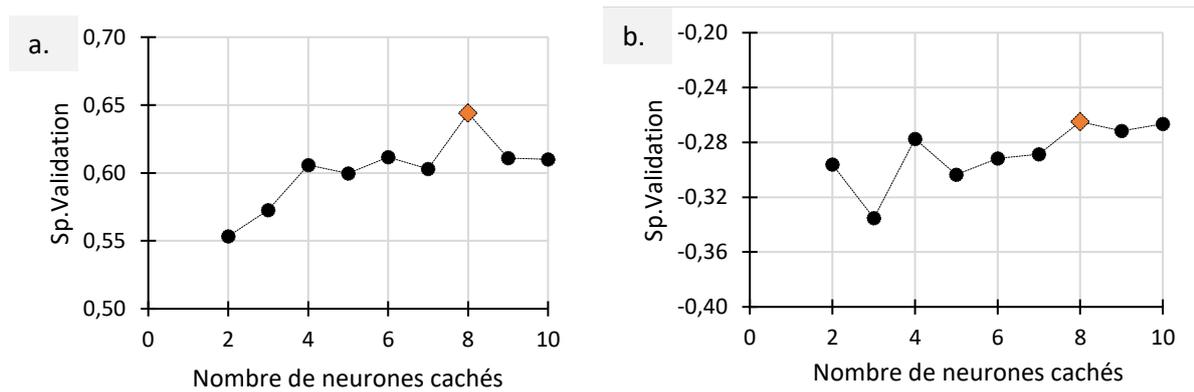


Figure 37 : Scores de validation croisée en fonction du nombre de neurones de la couche ξ pour le modèle dirigé (a), et le récurrent (b).

On note que le nombre optimal de neurones cachés dans la couche ξ sont de 8 pour le modèle dirigé, avec un score de 0,64. Tandis qu'il est de 10 pour le modèle récurrent avec un score négatif de -0,26 environ.

Ces complexités sont assez élevées soulignant que les couches profondes ne semblent pas avoir simplifié la fonction à réaliser par les couches superficielles.

Par ailleurs, le modèle récurrent est nettement moins bon que le modèle dirigé, comme cela est souvent le cas pour les faibles horizons de prévision.

3.1.2 Nombre de neurones des couches δ connectées aux variables d'entrée

Le Tableau 4 et le Tableau 5 synthétisent le dimensionnement des neurones des couches profondes connectées aux différentes variables d'entrée, respectivement pour les modèles dirigé et récurrent. La dimension de chacune de ces couches δ a été déterminée en considérant une complexité croissante de deux à cinq neurones non-linéaires. Pour limiter l'explosion des combinaisons possibles le nombre de neurones des couches δ reliées aux variables les plus explicatives (corrélacion croisée la plus forte) ont été déterminées les premières.

Pour les deux architectures, les complexités des couches δ ont donc été dimensionnées dans l'ordre suivant : Q_{obs} (respectivement Q_{rec}), P_{SRDT} , P_{BDC} , P_{MIA} , P_{Cum} .

Tableau 4 : dimensionnement des couches profondes au niveau des modèles dirigés

Nombre de neurones de la couche δ pour le modèle dirigé	Score de validation croisée (persistance)				
	Débit observé	SRDT	BDC	Mialet	P.Cum.
2	0,634	0,617	0,612	0,612	0,619
3	0,644	0,644	0,644	0,644	0,644
4	0,608	0,617	0,619	0,635	0,617
5	0,611	0,61	0,621	0,619	0,622
Décision	3	3	3	3	3

Ainsi les couches δ sont toutes dimensionnées avec trois neurones sur chacune des couches profondes. On voit que les scores de persistance maximale ne bougent quasiment pas au niveau de ce modèle, il reste à 0.644 qui est le même que celui obtenu à l'issue de la complexité optimale sélectionnée sur la couche ξ de l'entrée de débit observé aux pas de temps précédent.

Tableau 5 : dimensionnement des couches profondes au niveau des modèles récurrents

Nombre de neurones de la couche δ pour le modèle récurrent	Score de validation croisée (persistance)				
	Ordre	SRDT	BDC	Mialet	P.Cum.
2	-0,278	-0,288	-0,281	-0,248	-0,252
3	-0,253	-0,27	-0,253	-0,253	-0,248
4	-0,335	-0,277	-0,278	-0,271	-0,297
5	-0,273	-0,253	-0,278	-0,295	-0,275
Décision	3	5	3	2	3

Le modèle récurrent, exigeant plus d'informations venant des entrées exogènes, nécessite cinq neurones sur la couche δ reliée aux entrées du pluviomètre SRDT situé au milieu du bassin versant, deux neurones sur la couche δ reliée aux entrées du pluviomètre à l'exutoire, et trois neurones sur les couches δ reliées aux entrées des autres variables.

On peut noter que le score de la validation croisée augmente au fil de la sélection appliquée aux différentes variables. En effet, à l'issue de la sélection de l'ordre il vaut -0,253. La sélection des complexités successives l'amène à -0,248, qui est cependant très peu différent. Le score de validation croisée final est donc de -0,248 pour le modèle récurrent.

3.2 Type d'initialisation et méthodes de régularisation

Dans cette partie, nous étudions et comparons les scores de validation croisée en fonction :

- des cinq types d'initialisation de la couche profonde envisagés,
- des trois méthodes de régularisation envisagées : arrêt précoce seul, modération des poids seule, combinaison des deux.

Les performances associées à ces essais sont présentées en trois tableaux. Chaque tableau récapitule les critères de performance en validation croisée des deux architectures (dirigé et récurrent) pour toutes les initialisations envisagées.

Le Tableau 6 concerne l'utilisation conjointe des deux méthodes de régularisation (arrêt précoce et modération des poids). Le Tableau 7 est semblable, mais présente l'incrément des scores par rapport au Tableau 6 pour l'utilisation de l'arrêt précoce seul. La visualisation des incréments permet d'éviter la comparaison ligne à ligne et d'avoir rapidement une idée d'ensemble. De même le Tableau 8 suit le même mode de présentation mais concerne l'utilisation de la modération des poids seule.

A la lecture du Tableau 6, en prenant pour critère le score de persistance (il y a quelques différences avec le score de Nash), on constate que pour le modèle dirigé, l'initialisation aléatoire des paramètres contribue au meilleur score. Cette performance vaut 0,645, un peu supérieure aux configurations iCP (2,7 %) et iCT (3,1 %).

Concernant les architectures récurrentes, c'est l'initialisation par la corrélation croisée totale qui conduit à la plus grande performance. Sa persistance vaut -0,236, très légèrement supérieure à celle de la configuration iCP (1,2%) et 2,8% (iRnD).

On note également que le fait de fixer les paramètres profonds avec la corrélation croisée, qu'elle soit partielle ou totale diminue les performances très significativement pour le modèle dirigé, et notablement pour le modèle récurrent.

Tableau 6 : Performance en validation croisée des modèles avec l'utilisation conjointe des deux méthodes de régularisation (arrêt précoce et modération des poids)

Architecture	Gestion des coefficients des couches profondes	Performance en validation croisée		
		Ecart-type des scores durant la validation croisée (persistance)	Score de persistance	Score de Nash
Dirigée	Aléatoire (iRnD)	0,030	0,645	0,885
	Xcor. Partielle (iCP)	0,025	0,614	0,884
	Xcor. Totale (iCT)	0,029	0,618	0,885
	Xcor. partielle fixés (fCP)	0,048	0,144	0,791
	Xcor. totale fixés (fCT)	0,035	0,134	0,789
Modèle récurrent	Aléatoire (iRnD)	0,085	-0,264	0,731
	Xcor. Partielle (iCP)	0,115	-0,248	0,737
	Xcor. Totale (iCT)	0,091	-0,236	0,729
	Xcor. partielle fixés (fCP)	0,044	-0,410	0,725
	Xcor. totale fixés (fCT)	0,066	-0,417	0,720

En analysant les écart-types, pour les architectures dirigées, l'initialisation par la corrélation croisée ne diminue pas significativement la dispersion des scores de validation pendant la validation croisée : 0,025 et 0,029 contre 0,030 (iRnD). De manière contre intuitive, du moins dans le cadre de ces expériences, le fait de fixer les paramètres augmente encore la dispersion des scores. Pour les architectures récurrentes, c'est l'inverse, ce sont les modèles dont les couches profondes sont fixées qui ont une dispersion la plus faible. Dans ce dernier cas, cela paraît cohérent car les corrélations peuvent apporter une information utile au modèles récurrents.

En ce qui concerne les critères de Nash des modèles dirigés aucun n'est réellement mauvais et certains sont bons (0,88 pour iRnD, iCP et iCT). Pour ces modèles récurrents, les critères de Nash sont nettement plus faibles et restent entre 0,72 et 0,73 pour les différents types d'initialisation.

On peut tirer deux conclusions par rapport à ces résultats : d'une part l'initialisation par la corrélation croisée, totale ou partielle, n'apporte pas de plus-value notable que ce soit sur les critères ou sur la dispersion des scores de validation croisée.

D'autre part, le fait que les modèles simplement initialisés par rapport aux modèles dont les paramètres sont fixés soient toujours meilleurs laisse à penser que l'apprentissage se réalise correctement dans les couches profondes et ne subit donc pas d'effet de type « gradient évanescant ».

3.3.1 Comparaison des méthodes de régularisation : ApMp et APs

Le Tableau 7 présente les variations observées au tableau des critères obtenus lorsque la modération des poids est utilisée en conjonction avec l'arrêt précoce (ApMp), lorsque la méthode de régularisation est l'arrêt précoce seul (APs).

Les valeurs en rouge correspondent à une variation positive (APs meilleur), les valeurs en noir pour une variation négative (ApMp meilleure), et le tiret lorsque qu'il n'y a pas de variation.

On note tout d'abord que les variations sont toutes extrêmement faibles, de l'ordre de 10^{-3} pour la plupart. Les deux méthodes sont donc globalement équivalentes.

Tableau 7 : Variation de performance ApMp par rapport à l'arrêt précoce seul, APs

Architecture	Gestion des coefficients des couches profondes	Performance en validation croisée		
		Ecart-type des scores durant la validation croisée (persistance)	Score de persistance	Nash
Modèle dirigé	Aléatoire (iRnD)	0,003	0,001	--
	Xcor. Partielle (iCP)	0,002	0,003	--
	Xcor. Totale (iCT)	-0,001	0,001	0,001
	Xcor. partielle fixés (fCP)	0,002	0,001	--
	Xcor. totale fixés (fCT)	-0,005	-0,002	--
Modèle récurrent	Aléatoire (iRnD)	--	-0,016	-0,005
	Xcor. Partielle (iCP)	0,014	0,017	--
	Xcor. Totale (iCT)	0,006	0,006	-0,003
	Xcor. partielle fixés (fCP)	-0,001	0,003	0,001
	Xcor. totale fixés (fCT)	--	-0,001	--

3.3.2 Comparaison des méthodes de régularisation : ApMp et MPs

Intéressons-nous maintenant aux variations observées entre le tableau des critères obtenus lorsque la modération des poids est utilisée en conjonction avec l'arrêt précoce (ApMp), ou lorsque la méthode de régularisation est la modération des poids seule (MPs).

Les valeurs en rouge correspondent à une variation positive (MPs meilleure), les valeurs en noir pour une variation négative (ApMp meilleure), et le tiret lorsque qu'il n'y a pas de variation.

On note tout d'abord que les variations sont en général extrêmement faibles, de l'ordre de 10^{-3} pour la plupart. Néanmoins dans certains cas elles ne sont pas négligeables, par exemple pour l'initialisation avec la corrélation partielle du modèle récurrent (0,129). La régularisation MPs est alors moins bonne.

Ceci confirme les résultats obtenus par Toukourou et al. (2009). La modération des poids seule semble donc à éviter sur ces types de données et de bassin. Mais il est difficile de généraliser à partir de cette seule étude à tous types de bassins.

Tableau 8 : Gain ou déficit de performance de la combinaison à la modération des poids seule

Architecture	Gestion des coefficients des couches profondes	Performance en validation croisée		
		Ecart-type des scores durant la validation croisée (persistance)	Score de persistance	Nash
Modèle dirigé	Aléatoire (iRnD)	0.007	0.002	0.008
	Xcor. Partielle (iCP)	-0.015	-0.023	0.007
	Xcor. Totale (iCT)	-0.002	-0.021	0.009
	Xcor. partielle fixés (fCP)	0.032	0.070	0.002
	Xcor. totale fixés (fCT)	0.012	0.060	0.001
Modèle récurrent	Aléatoire (iRnD)	-0.010	0.061	0.002
	Xcor. Partielle (iCP)	0.001	0.129	0.009
	Xcor. Totale (iCT)	0.038	0.082	0.009
	Xcor. partielle fixés (fCP)	0.006	0.003	0.006
	Xcor. totale fixés (fCT)	0.005	0.003	0.002

3.4 Visualisation des prévisions sur l'évènement de test

3.4.1 Hydrogrammes sur l'épisode 13

Après l'analyse des tableaux précédents comparant les trois méthodes de régularisation mesurées sur les scores de validation croisée, nous nous intéressons dans cette partie aux scores sur l'évènement de test. Rappelons que ce dernier est le second évènement le plus intense de la base de données. Nous apprécierons la généralisation effectuée sur cet évènement au travers de deux méthodes très classiques : l'analyse visuelle des hydrogrammes de prévision et le tableau des scores sur ce seul évènement de test. La prévision est réalisée à 2h.

3.4.1.1 Prédiction avec combinaison arrêt précoce et Modération des poids

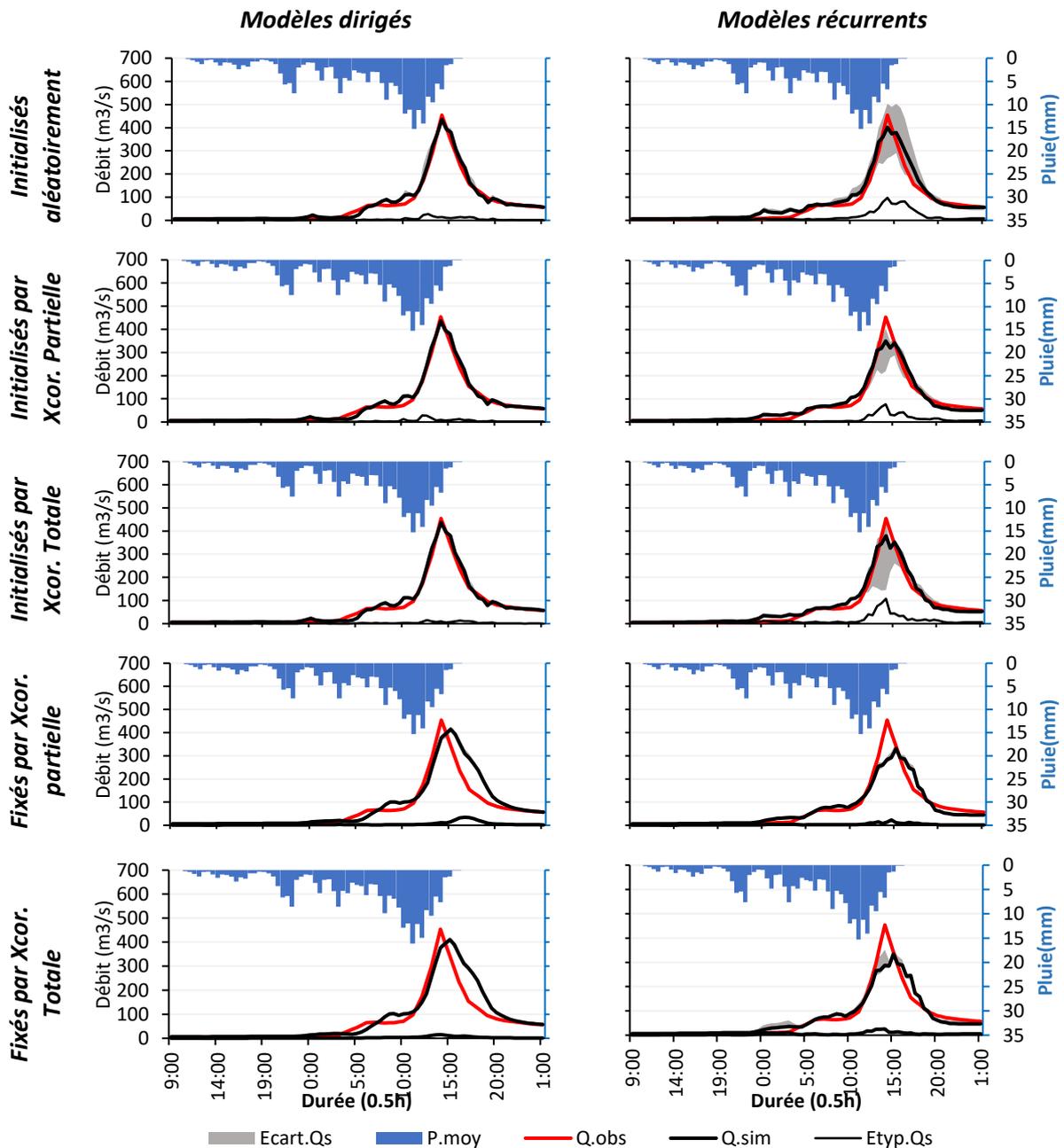


Figure 38 : Prédiction des crues en fonction des différents types d'initialisation avec la combinaison de l'arrêt précoce et la modération des poids

Les hydrogrammes prévus, issus de la médiane du modèle d'ensemble sont représentés en trait continu noir. La dispersion du modèle d'ensemble est représentée avec une zone grise qui représente la seconde valeur la plus haute et l'avant dernière valeur la plus basse. Nous n'avons pas visualisé les deux valeurs extrêmes pour ne pas être sujet à des effets dus aux cas anormaux (outliers).

On peut apprécier visuellement la performance de ces modèles en trois angles :

- Le premier concerne les modèles dirigés dont les coefficients ne sont pas fixés. Ceux-ci sont tous très performants avec pratiquement très peu de dispersion des prévisions. Ceci est confirmé par le tableau des critères (*Tableau 9 : Performance en test des modèles obtenus avec l'utilisation conjointe des deux méthodes de régularisation (arrêt précoce et modération des poids)*).
- Le second angle est identique au premier mais concerne les modèles récurrents. On peut y constater également une assez bonne performance sur ces prévisions, mais la dispersion sur les pics de crue est plus importante au niveau du modèle initialisé aléatoirement, cette dispersion est tout autant présente à la montée et à la descente des crues. Les critères ne prenant en compte que la médiane ne font pas apparaître la dispersion et le modèle le meilleur est iCT ($C_p= 0,89$). Encore une fois l'initialisation avec la corrélation totale aide le modèle récurrent. L'amélioration est sensible par rapport au modèle iRnD ($C_p=0,80$).
- Le troisième angle concerne les modèles dont les couches profondes sont fixées (pas d'apprentissage sur ces couches). La dispersion est très atténuée. Elle est pratiquement concentrée au niveau du pic.

Tableau 9 : Performance en test des modèles obtenus avec l'utilisation conjointe des deux méthodes de régularisation (arrêt précoce et modération des poids)

Architectures	Gestion des coefficients des couches profondes	Performances en test				
		C_p	P_d	SPPD	%Pic	Nash
Valeurs de référence		1	0	100%	100%	1
Dirigé	Aléatoire (iRnD)	0,969	0	96,0	96,0	0,988
	Xcor. Partielle (iCP)	0,971	0	96,1	96,1	0,987
	Xcor. Totale (iCT)	0,968	0	96,2	96,2	0,987
	Xcor. partielle fixés (fCP)	0,711	1	83,3	91,4	0,867
	Xcor. totale fixés (fCT)	0,698	1	83,0	90,5	0,865
Récurrent	Aléatoire (iRnD)	0,804	0	88,4	88,4	0,947
	Xcor. Partielle (iCP)	0,872	0	77,2	77,2	0,956
	Xcor. Totale (iCT)	0,890	0	83,5	83,5	0,959
	Xcor. partielle fixés (fCP)	0,807	1	62,5	72,9	0,912
	Xcor. totale fixés (fCT)	0,807	1	63,4	73,5	0,913

De manière plus générale, on note tout d'abord que les critères de persistance du modèle dirigé sur le test sont de 0,97 pour les trois types d'initialisation, le constat est le même pour les autres critères. Lorsque l'on a fixé les paramètres de la couche profonde, les performances sont significativement dégradées. Quant aux modèles récurrents, l'analyse est partagée, le critère de persistance est meilleur pour le modèle iCT, mais le SPPD est meilleur pour iRnD. Les critères sont cependant tous bons. En ce qui concerne les modèles dont les couches profondes sont fixées, le modèle récurrent est celui à qui cela profite le plus ($C_p= 0, 804$), contre ($C_p= 0,70$) pour le modèle dirigé, comme constaté auparavant.

3.4.1.2 Variation de performance ApMp par rapport à l'arrêt précoce seul, APs

Les hydrogrammes de prévision de la Figure 39 présentent les résultats en test lorsque la méthode de régularisation est l'arrêt précoce. Entre ces graphiques et leurs prédécesseurs obtenus avec l'utilisation conjointe de l'arrêt précoce et de la modération des poids, on ne note aucune différence.

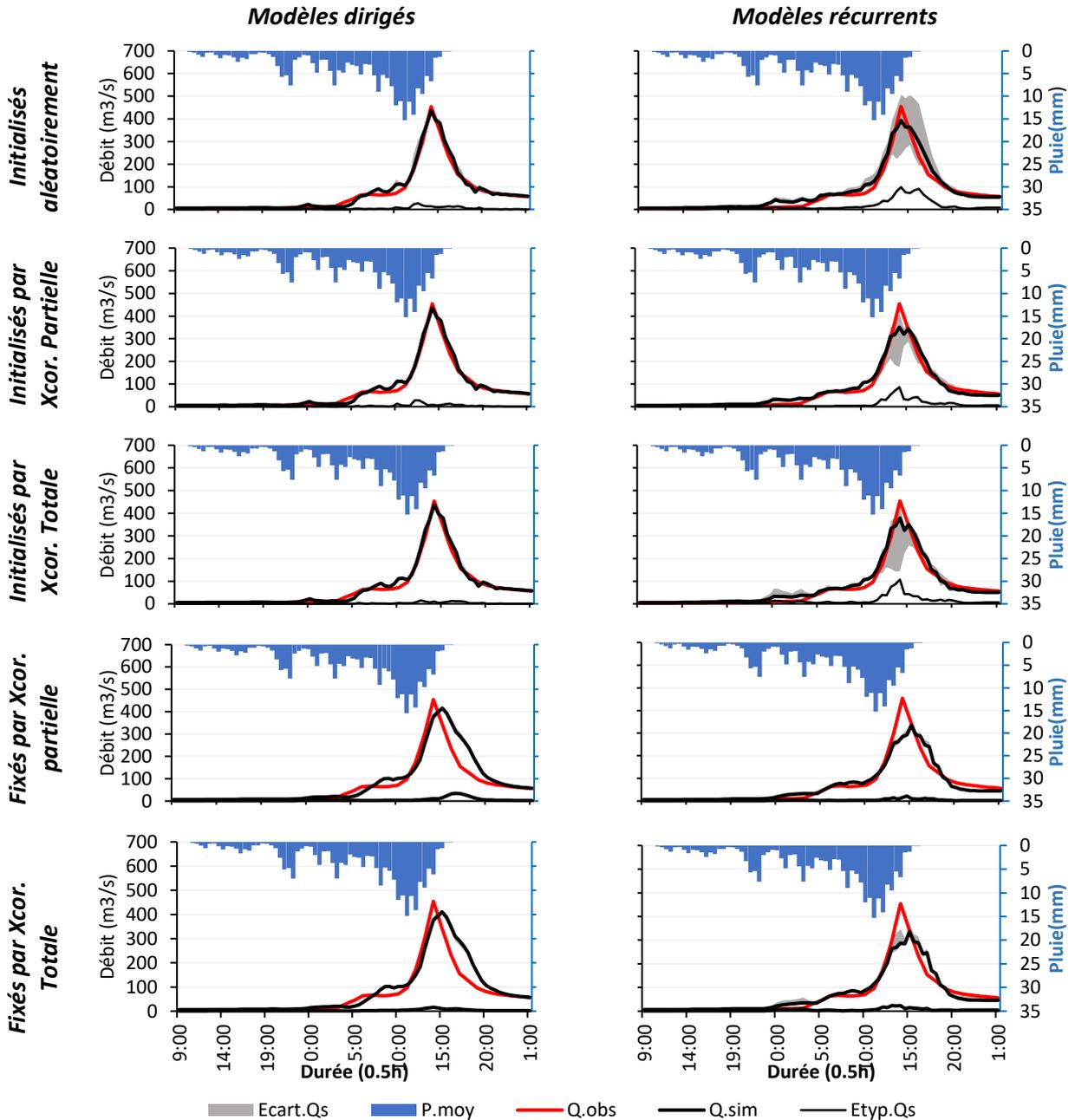


Figure 39 : Prédiction des crues en fonction des différents types d'initialisation régulée par l'arrêt précoce seul

Au niveau du Tableau 10 nous présentons les variations par rapport au Tableau 9 précédent. Les tirets (--) indiquent les cas où les variations sont nulles. Les variations négatives sont en noir (combinaison plus avantageuse) et en rouge elles indiquent les cas où l'arrêt précoce seul est plus avantageux.

Au niveau du modèle dirigé, on peut voir que globalement la combinaison et l'arrêt précoce seul procurent le même niveau de performance sur le test. Pour le modèle récurrent, c'est plus nuancé et l'avantage dû à l'arrêt précoce seul semble faible et très marginal.

Tableau 10 : Gain ou déficit de performance en test de la combinaison par rapport à l'arrêt précoce seul

Architecture	Gestion des coefficients des couches profondes	Performance en test				
		Cp	Pd	SPPD(%)	%Pic	Nash
Dirigé	Aléatoire (iRnD)	0,001	--	--	--	--
	Xcor. Partielle (iCP)	--	--	0.1	0.1	--
	Xcor. Totale (iCT)	--	--	--	--	--
	Xcor. partielle fixés (fCP)	0,001	--	-0.3	-0.1	--
	Xcor. totale fixés (fCT)	-0,003	--	--	--	-0,001
Récurent	Aléatoire (iRnD)	0,004	--	1.8	1.8	-0,004
	Xcor. Partielle (iCP)	--	--	0.1	0.1	0,001
	Xcor. Totale (iCT)	0,015	--	0.1	0.1	-0,001
	Xcor. partielle fixés (fCP)	--	--	-0.2	-0.7	-0,001
	Xcor. totale fixés (fCT)	--	--	--	-0.3	--

3.4.1.3 Variation de performance ApMp par rapport la modération des poids seule

Lorsqu'on utilise la modération des poids comme unique méthode de régularisation on peut noter des différences par rapport à la combinaison des deux méthodes : la combinaison donne des résultats meilleurs. Nous présentons ici le tableau correspondant pour faciliter la navigation du lecteur entre les deux tableaux dans sa logique de comparaison.

Tableau 11 : Gain ou déficit de performance en test de la combinaison par rapport à la modération des poids seul

Architecture	Gestion des coefficients des couches profondes	Performance en test				
		Cp	Pd	SPPD (%)	%Pic	Nash
Dirigé	Aléatoire (iRnD)	0,021	--	-3,2	-3,2	0,006
	Xcor. Partielle (iCP)	0,034	--	0,8	0,8	0,012
	Xcor. Totale (iCT)	0,033	--	3,3	3,3	0,013
	Xcor. partielle fixés (fCP)	0,050	--	-17,0	-13,0	0,021
	Xcor. totale fixés (fCT)	0,049	--	-17,1	-13,1	0,026
Récurent	Aléatoire (iRnD)	-0,031	-1	17,5	16,6	0,004
	Xcor. Partielle (iCP)	0,030	-1	7,2	1,4	0,011
	Xcor. Totale (iCT)	-0,008	--	1,0	1,0	-0,003
	Xcor. partielle fixés (fCP)	-0,012	--	-0,9	-3,0	-0,007
	Xcor. totale fixés (fCT)	-0,002	--	-0,3	0,5	-0,002

Contrairement à la comparaison de la combinaison ApMp avec l'arrêt précoce seul, l'utilisation de la modération des poids comme unique méthode de régularisation peut dégrader ou améliorer le modèle dirigé quel que soit le type de gestion de la couche profonde. Concernant la fixation des couches profondes, le gain peut être notable : l'amélioration des pics atteint 17%. En revanche, pour le critère de persistance la modération des poids n'apporte aucune amélioration par rapport à la combinaison ApMp. Le support graphique de ces résultats est présenté au niveau de la Figure 39 à page suivante.

En effet, si on repart des trois angles d'analyses précédemment mentionnés, on peut constater que même si globalement les performances sont assez bonnes, il apparaît une dispersion assez remarquable sur la courbe de crues des modèles dirigés. Pour les modèles récurrents, la dispersion persiste au niveau des pics des crues indépendamment du type d'initialisation des paramètres de la couche profonde. Quant aux cas où les paramètres des couches profondes sont fixés, la dispersion a pratiquement disparu au niveau de la montée des crues indifféremment de l'architecture. Contrairement aux modèles récurrents, les pics sont assez bien anticipés au niveau des modèles dirigés. Les dispersions qui ont été pratiquement absentes au niveau des deux cas précédents sont plus visibles quand la modération des poids est utilisée seule.

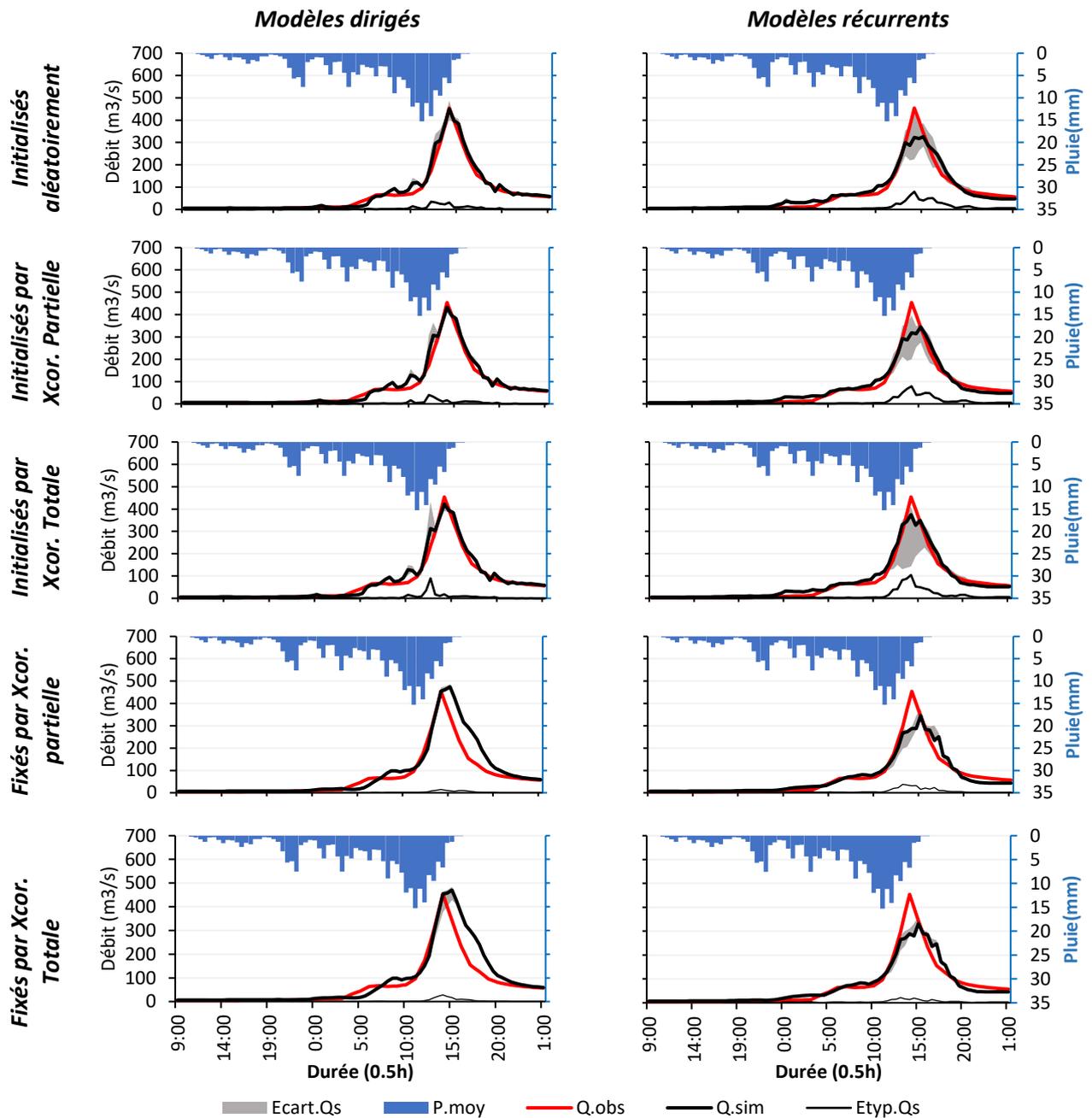


Figure 40 : Prédiction des crues en fonction des différents types d'initialisation régulée par la modération des poids seule

3.4.2 Comparaison des écart-types sur la prévision du débit entre les différentes approches d'initialisation

Les analyses précédentes ont montré que l'on pouvait noter un faible avantage à utiliser les régularisations par modération des poids et arrêt précoce conjointement. Nous nous intéressons maintenant à la dispersion des valeurs des membres de l'ensemble. La Figure 41.a& b présente les écart-types du débit prévu à deux heures sur 20 modèles d'ensembles où la combinaison des deux méthodes de régularisation est utilisée. Les graphiques c-d puis e-f représentent respectivement la différence entre la combinaison ApMp et l'utilisation seule de l'arrêt précoce puis la modération des poids.

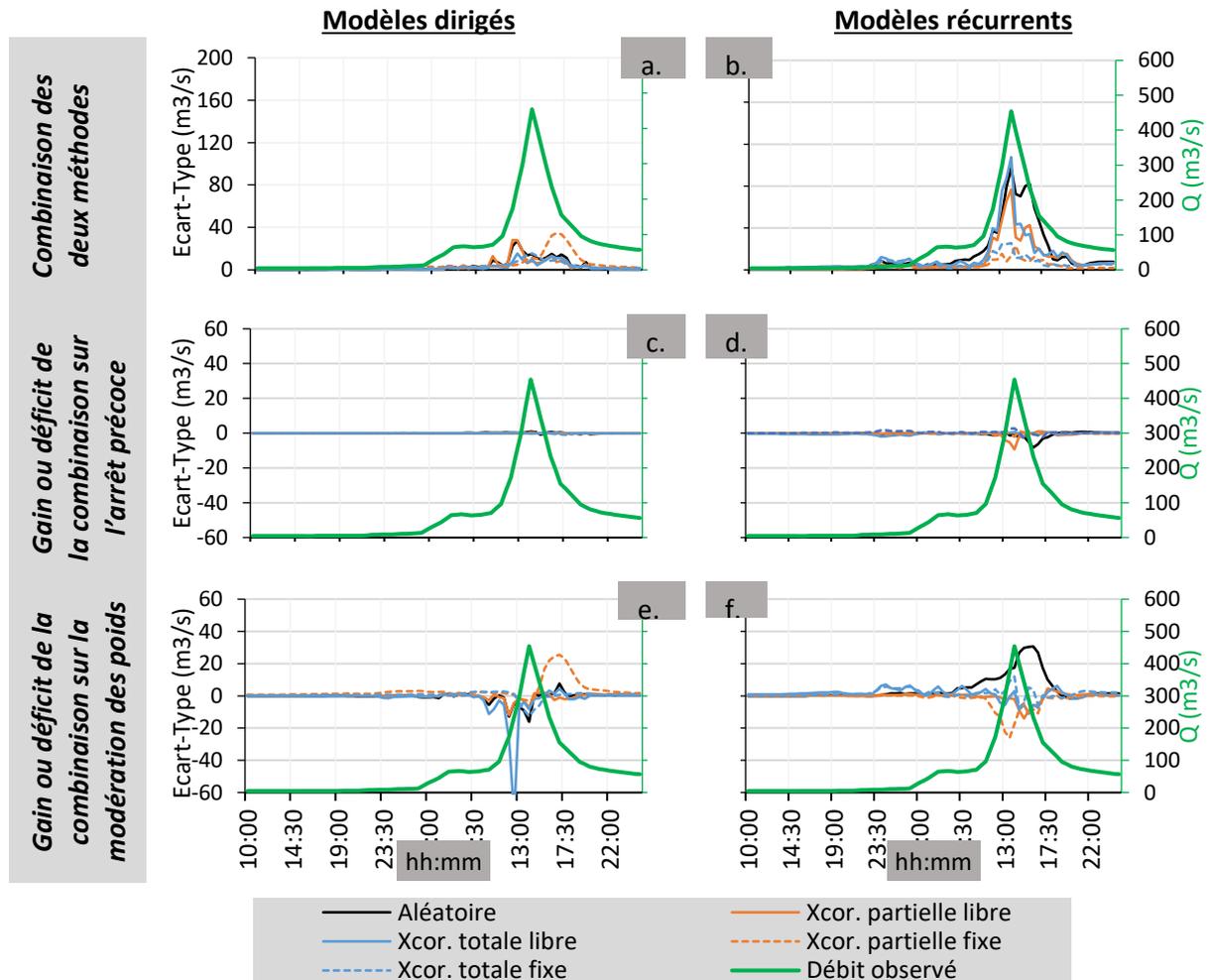


Figure 41 : Comparaison de la variabilité des prévisions des crues en fonction du type d'initialisation et de la méthode de régularisation. Le débit observé (vert, trait plein) ; écart-type : l'initialisation aléatoire (noir, trait plein) ; initialisation et fixation par corrélation croisée totale (bleu, trait plein et continu) ; initialisation et fixation par corrélation croisée partielle (orange, trait plein et continu)

Comme on a pu le constater au niveau des hydrogrammes de la section précédente, la dispersion des prévisions est plus importante pour les modèles récurrents qu'elle ne l'est sur les modèles dirigés. Ceci est clair en comparant les graphiques a & b. Quant aux types d'initialisation, on peut voir sur le modèle dirigé (a.) qu'il y a deux petites bosses de part et d'autre de l'instant du pic de la crue qui sont dû essentiellement à l'initialisation aléatoire et aux initialisations par la corrélation croisée partielle. Ces valeurs indiquent des écart-type d'à peu près $40 \text{ m}^3/\text{s}$. Quant au modèle récurrent, ces dispersions sont à peu près trois fois plus importantes pour les modèles dont les coefficients n'ont pas été fixés. Par ailleurs, les graphes c-d indiquent que l'approche combinée est vraiment équivalente à l'arrêt

précoce seul, tandis que les graphes e-f montrent des écarts importants entre la méthode combinée et la modération des poids seule. Sur cette dernière comparaison, il est difficile de pouvoir tirer des conclusions car il ne sort aucune tendance en fonction du type de gestion de la couche la plus profonde.

4 Discussion

Ces expériences avaient pour objectif de comparer différents types d'initialisation, ou de fixation, des paramètres des couches profondes des réseaux de neurones profonds. À notre connaissance ceci n'a jamais été étudié. Il convient de rappeler que le dimensionnement du modèle a été réalisé sur des modèles initialisés aléatoirement. Le dimensionnement n'a pas été refait pour chaque type d'initialisation différent, pour pouvoir comparer en ne changeant d'un item à chaque fois.

La première constatation est que les complexités de la couches ξ sont plus importantes, que ce soit pour les modèles récurrents ou dirigés. L'intégration de couches profondes a donc complexifié la tâche des couches superficielles.

Concernant la gestion des couches profondes, il apparaît que l'initialisation avec la corrélation croisée n'apporte que des bénéfices marginaux :

- Quasiment pas d'amélioration des scores ou des hydrogrammes, parfois une diminution de la variabilité des membres de l'ensemble de test.
- L'initialisation avec la corrélation croisée totale peut améliorer les modèles récurrents
- Concernant le fait de fixer les paramètres avec la corrélation croisée, elle dégrade en général les performances du modèle dirigé, mais peut apporter des bénéfices sur la dispersion des membres de l'ensemble du modèle récurrent.

Si l'on s'intéresse maintenant aux méthodes de régularisation, l'utilisation conjointe de l'arrêt précoce et de la modération des poids semble la meilleure des préconisations, même si ses performances en généralisation sont quasiment équivalentes à celle de l'arrêt précoce seul, avec quelques cas où elle présente un avantage. En revanche on note une plus faible performance de la modération des poids seule, ceci confirme l'étude de Toukourou et al. (2009).

Conclusion du chapitre et perspectives

Même si l'idée d'initialiser les couches profondes du modèle neuronal avec les valeurs des corrélations croisées avait suscité de grands espoirs il s'avère qu'elle ne permet pas de pallier significativement les questions soulevées en introduction de ce chapitre : expliquer, ou pallier, la forte variabilité des sorties des modèles en fonction de l'initialisation de leurs paramètres.

Des avantages sont cependant apparus marginalement qui pourraient permettre d'améliorer les modèles récurrents, par exemple en initialisant leurs couches profondes avec la corrélation croisée totale.

Peut-être plus intéressant, si l'on compare les critères obtenus avec les architectures très profondes de ce chapitre, on note qu'ils ont été améliorés par rapport aux critères du modèle profond de bassins emboîtés chapitre VI, pour l'horizon de 2h. Le critère de persistance sur l'évènement 13 est passé de 0.79 à 0.89 et le critère de Nash de 0.88 à 0.96 pour le modèle récurrent avec l'initialisation par corrélation croisée totale. Ces modèles très profonds ont donc vraisemblablement calculé des prétraitements efficaces sur les entrées exogènes ou les variables jouant le rôle de variables d'état (débit précédent, ou cumul de précipitation). L'augmentation de la complexité notée sur la couche ξ , que ce soit pour les modèles récurrents ou dirigés a donc une nécessité.

Ce dernier élément ouvre de nombreuses perspectives parmi lesquelles l'initialisation du modèle emboité avec la corrélation croisée totale, l'étude plus générale de la combinaison des méthodes de régularisation : modération des poids et arrêt précoce, ou encore l'investigation d'architectures encore plus profondes.

Pour finir, ces expériences ont été menées sur un seul bassin versant, et testées sur un seul événement. Il est donc évident que ces résultats, quoique prometteurs, ne peuvent pas être généralisés en l'état. De ce fait, leur application sur un bassin versant d'une plus grande complexité peut se révéler pertinente. Nous pensons ici à un bassin versant comportant une composante karstique forte, un bassin versant subissant l'influence des neiges ou un bassin versant comportant des modifications anthropiques fortes (barrage, urbain, ...).

Conclusion générale et perspectives

Comme nous l'avons souligné dans l'introduction, les crues éclairs sont un des risques naturels entraînant le plus de décès et de dégâts au monde. Face aux enjeux associés, les pouvoirs publics se sont organisés ; en France, le SCHAPI (Service Central d'Hydrométéorologie et d'Appui à la Prévision des Inondations) et les SPC (Services de Prévision de Crues) sont directement impliqués dans la prévision de ce genre de menace.

L'exercice de prévision des crues rapides est scientifiquement difficile pour plusieurs raisons : d'une part les bassins versants sont des systèmes dynamiques non-linéaires ; d'autre part leur observation est très difficile du fait des hétérogénéités spatiales et temporelles conduisant à des effets d'échelles difficiles à qualifier ; enfin, le couplage de ces phénomènes avec le substratum qui les reçoit conduit à des phénomènes mal connus, insuffisamment documentés.

Face à ces difficultés le Laboratoire HSM développe depuis près de trois décennies des modèles à apprentissage statistiques, ou *machine learning*, qui sont particulièrement adaptés à ce type de problématiques : ils peuvent représenter les phénomènes dynamiques et non linéaires inconnus grâce à seulement une base de données, composée des entrées et des sorties, du fait de leurs capacités d'apprentissage. Cependant la capacité d'apprentissage est limitée par le dilemme biais-variance qui transcrit le fait que plus il y a de « bruit » dans les données plus le modèle neuronal a du mal à généraliser ce qu'il a appris à une autre base de données que sa base d'apprentissage. Cette limite associée aux modèles neuronaux est d'autant plus critique que le modèle est complexe, c'est pourquoi le Laboratoire développe et perfectionne depuis plus de dix thèses la méthode de sélection de la complexité du modèle afin de préserver les qualités de « généralisation » du modèle, c'est-à-dire sa capacité à prévoir à des données inconnues ce qu'il a appris sur ses données d'apprentissage. Plus récemment les actualités scientifiques ont mis en lumière les modèles dits « profonds » capables de réaliser avec brio des fonctions difficiles comme la traduction automatique du langage, la conduite de véhicules autonomes ou autres applications de reconnaissance d'images.

Dans ce contexte notre travail visait à développer des modèles plus profonds que les modèles habituellement utilisés, pour effectuer la prévision des crues éclair en facilitant une meilleure prise en compte des comportements multi-échelles. Une architecture profonde aurait son intérêt, afin de calculer des prétraitements dans les couches dites « profondes » qui extraient les caractéristiques des signaux, avant que les couches dites « superficielles » ne les assemblent pour satisfaire à l'objectif. Pour ce faire nous avons choisi le bassin versant de Mialet, appartenant à la bordure cévenole, sur lequel se sont déjà focalisés deux travaux de thèse.

Cette étude a été menée en plusieurs étapes : la première partie visait à représenter par le modèle les « bassins emboîtés » qui peuvent représenter les différents sous bassins d'amont en aval, chacun prenant en compte le débit amont et une nouvelle source de pluviométrie. Pour ce faire un modèle profond récurrent a été utilisé comportant trois couches cachées de neurones non linéaires. Cette première partie a mis en évidence, une fois de plus, la qualité de la méthode de sélection de modèles utilisée au Laboratoire, qui a permis sur ce cas d'étude de sélectionner des historiques de précipitation pour alimenter chacune des couches, prenant en compte l'agrandissement du bassin et l'allongement associé du temps de réponse. De même la sélection du nombre de neurones cachés de chaque couche a conduit à une augmentation de la complexité d'amont en aval, les fonctions calculées à l'aval devant prendre en compte à la fois des dynamiques rapides (pluie proche) et des dynamiques plus lentes (débit amont). Le processus de sélection a donc été capable de sélectionner et représenter les différentes échelles utiles de temps et d'espace.

Une alternative à cette modélisation emboîtée a été de s'intéresser à l'importance du contenu informatif des trois pluviomètres utilisés dans ce bassin, et de déléguer à une seule couche cachée la prise en compte de la transformation des pluies en débit, sans emboîtement. La méthode KnoX développée dans la thèse de Line Kong A Siou a été appliquée pour extraire l'information que le modèle utilisait pour chaque pluviomètre et de comparer cette utilisation pour les différents types de modèles utilisés pour représenter le bassin versant : le modèle statique (uniquement pluie-débit estimé), le modèle dirigé (pluie+débites précédents observés – débit estimé), le modèle récurrent (pluie+débites précédents estimés – débit estimé). Le modèle comportait deux couches cachées dont la plus profonde était linéaire. L'analyse de l'information contenue dans la couche la plus profonde, pour chaque pluviomètre, a pu être mise en relation avec le type de modèle utilisé : le modèle statique qui ne reçoit que les précipitations a représenté cette information de manière très cohérente dans ses paramètres, suivi de près par le modèle récurrent qui reçoit en entrée l'estimation de son débit. Enfin le modèle dirigé qui reçoit en entrée la valeur de son débit observé au pas de temps précédent avait assez peu besoin de l'information ancienne des précipitations qu'il recevait déjà sous forme des débits. Il a donc estimé une répartition de l'information axée principalement sur cette information de débits précédents observés. Ce travail a également permis de comparer les performances relatives des trois types de modèles, qui correspondait aux attentes : le modèle purement statique étant le moins performant et le modèle dirigé le plus performant. Cette analyse a pu être présentée à la Conférence ITISE à Grenade en 2019 et publiée dans les actes de la conférence.

À partir de ce travail, le même modèle a été utilisé et son information a alors été analysée pour extraire les contributions des pluviomètres à la prévision du débit. Ces informations ont été comparées à la contribution des pluviomètres estimée par la technique des polygones de Thiessen. Cette contribution a aussi été comparée à celle délivrée par les corrélations croisées entre chaque pluviomètre et la sortie de débit. Ce travail a été publié sous forme d'un chapitre de livre dans la collection LNCS de Springer.

La dernière partie de ce travail, non encore publiée, a tenté, en quelques sortes, d'inverser le processus ; il s'agissait d'introduire des informations issues des signaux dans les couches profondes afin d'améliorer leurs performances et en particulier de diminuer l'impact de l'initialisation aléatoire des couches profondes. Si ces derniers travaux n'ont pas apporté de résultats très forts, en revanche ils ouvrent la voie à plusieurs questionnements : la corrélation croisée est-elle la meilleure information à introduire dans les couches profondes ? Le modèle récurrent est-il toujours le modèle qui bénéficie le plus de cet apport ? Il peut donc se révéler envisageable de trouver des approches méthodologiques permettant d'introduire des informations issues de l'occupation et de la couverture du sol, du taux d'imperméabilisation des versants, du taux et de l'influence de la karstification et ou des fractures, etc..., caractérisant le bassin versant en question.

Plus généralement ces travaux de thèse se sont intéressés à utiliser des modèles de réseaux de neurones profonds avec une complexité maîtrisée afin d'optimiser les capacités de généralisation des prédicteurs. Cette complexité maîtrisée a plusieurs avantages ; d'une part elle contribue à une meilleure généralisation, d'autre part, elle permet à la règle d'apprentissage d'être efficace et de ne pas saturer les sigmoïdes entraînant une atténuation du gradient et une baisse des capacités d'apprentissage, phénomène appelé le gradient évanescent.

Nous avons montré que contrairement à ce qui est souvent annoncé, les modèles à réseaux de neurones, même profonds peuvent être interprétés, et que ces interprétations sont cohérentes avec le fonctionnement connu du bassin. Ils apparaissent donc comme dignes de confiance pour effectuer des prévisions

On doit noter que ces expériences n'ont été réalisées que sur un seul bassin versant, elles ont également été testées pour la plupart sur un seul épisode pluvieux parmi les plus intenses de la base. Ceci nous amène à considérer qu'il serait important qu'elles soient étendues à d'autres contextes hydrologiques. Cette perspective devrait nous permettre de mettre à l'épreuve la généralisation de la démarche. Cependant, nous osons espérer que l'extension de ce travail sur un bassin versant comportant une composante karstique par exemple pourrait fournir des résultats beaucoup plus forts et intéressants. Dans ce sens, on peut orienter le modèle vers la prise en compte de l'aspect différé des écoulements imposés par la composante karstique vis à vis des écoulements de surfaces. Cela peut se faire en faisant ressortir, puis en les intégrant séparément, les différentes composantes de la réponse hydrologique globale du bassin versant. Il peut donc s'en découler diverses autres idées de ce type permettant de bien prendre en compte les comportements multi-échelles des hydrosystèmes à travers ces modèles neuronaux.

Au-delà de cela, dans le contexte de la crise climatique qui s'annonce et de la consommation grandissante du secteur du numérique, concevoir des modèles parcimonieux et peu consommateurs en énergie est une vertu.

Références bibliographique

- Ahmadalipour, A. and Moradkhani, H. (2019) 'A data-driven analysis of flash flood hazard, fatalities, and damages over the CONUS during 1996–2017', *Journal of Hydrology*, 578, p. 124106. doi: <https://doi.org/10.1016/j.jhydrol.2019.124106>.
- Akil, N., Artigue, G., Savary, M. Johannet, A. and Vinches, M. (2021) 'Uncertainty Estimation in Hydrogeological Forecasting with Neural Networks: Impact of Spatial Distribution of Rainfalls and Random Initialization of the Model', *Water*. doi: 10.3390/w13121690.
- Alfieri, L., Smith, P. J., Thielen-Del Pozo, P. J. and Beven, K. J. (2011) 'A staggered approach to flash flood forecasting - Case study in the Cévennes region', *Advances in Geosciences*, 29, pp. 13–20. doi: 10.5194/adgeo-29-13-2011.
- Alfieri, L., Bisselink, B., Dottori, F., Naumann, G., de Roo, A., Salamon, P., Wyser, K. and Feyen, L. (2017) 'Global projections of river flood risk in a warmer world', *Earth's Future*, 5(2), pp. 171–182. doi: 10.1002/2016EF000485.
- Ali, Z., Hussain, I., Faisal, M., Nazir, H. M., Hussain, T., Shad, M. Y., Mohamd Shoukry, A. and Hussain Gani, S. (2017) 'Forecasting drought using multilayer perceptron artificial neural network model', *Advances in Meteorology*, 2017.
- Araujo, P., Astray, G., Ferrerio-Lage, J. A., Mejuto, J. C., Rodriguez-Suarez, J. A. and Soto, B. (2011) 'Multilayer perceptron neural network for flow prediction', *Journal of Environmental Monitoring*, 13(1), pp. 35–41.
- Arjovsky, M., Shah, A. and Bengio, Y. (2015) 'Unitary Evolution Recurrent Neural Networks'. Edited by M. F. Balcan and K. Q. Weinberger. PMLR, pp. 1120–1128.
- Arplt, D., Jastrzębski S., Bails, N., Krueger, D., Bengio, E., Kanwal, M. S., Maharaj, T., Fischer, A., Courville, A., Bengio, Y. and Lacoste-Julien, S. (2017) 'A closer look at memorization in deep networks', in *34th International Conference on Machine Learning, ICML 2017*, pp. 350–359.
- 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 and Earth System Science*, 12(11), pp. 3307–3324. doi: 10.5194/nhess-12-3307-2012.
- Artigue, G. (2012) *Prévision des crues éclair par réseaux de neurones: généralisation aux bassins non jaugés*. Université de Montpellier II.
- Atmaja, B. T. and Akagi, M. (2020) *Deep Multilayer Perceptrons for Dimensional Speech Emotion Recognition, 2020 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference, APSIPA ASC 2020 - Proceedings*.
- Ayral, P.-A. (2005) 'Contribution à la spatialisation du modèle opérationnel de prévision des crues éclair ALHTAÏR', *Etudes de Géographie et Physiques*, XXXII(August), p. 315.
- Bai, S., Kolter, J. Z. and Koltun, V. (2018) 'Convolutional sequence modeling revisited'.
- Barron, A. R. (1993) 'Universal approximation bounds for superpositions of a sigmoidal function', *IEEE Transactions on Information Theory*, 39(3), pp. 930–945. doi: 10.1109/18.256500.
- Bartholmes, J. and Todini, E. (2005) 'Coupling meteorological and hydrological models for flood forecasting', *Hydrology and Earth System Sciences*, 9(4), pp. 333–346. doi: 10.5194/hess-9-333-2005.
- Bartlett, P. L. (1997) 'For valid generalization, the size of the weights is more important than the size

- of the network', *Advances in Neural Information Processing Systems*. Edited by eds. M.C. Mozer, M.I. Jordan, and T. Petsche, 9, pp. 134–140.
- Bell, T. L., Abdullah, A., Martin, R. L. and North, G. R. (1990) 'Sampling errors for satellite-derived tropical rainfall: Monte Carlo study using a space-time stochastic model', *Journal of Geophysical Research*, 95(D3), pp. 2195–2205. doi: 10.1029/JD095iD03p02195.
- Bellon, A. and Austin, G. (1984) 'The accuracy of short-term radar rainfall forecasts', *Journal of Hydrology*, 70, pp. 35–49.
- Bengio, Y., Simard, P. and Frasconi, P. (1994) 'Learning Long-Term Dependencies with Gradient Descent is Difficult', *IEEE Transactions on Neural Networks*, 5(2), pp. 157–166. doi: 10.1109/72.279181.
- Bergström, S. (1995) 'The HBV model.', *Computer models of watershed hydrology.*, pp. 443–476.
- Bertin, D. and Lambert, J. (1993) *Contribution à la modélisation du système karstique : application à la prédiction des débits d'eau souterrains" Rapport de stage d'initiation à la recherche de l'Ecole des Mines d'Alès*. Alès, France.
- Bessiere, H. (2008) 'Assimilation de données variationnelle pour la modélisation hydrologique distribuée des crues à cinétique rapide', *Sciences-New York*.
- Beven, K. (1989) 'Changing ideas in hydrology - The case of physically-based models', *Journal of Hydrology*, 105(1–2), pp. 157–172. doi: 10.1016/0022-1694(89)90101-7.
- Beven, K. (1993) 'Prophecy, reality and uncertainty in distributed hydrological modelling', *Advances in Water Resources*, 16(1), pp. 41–51. doi: 10.1016/0309-1708(93)90028-E.
- Beven, K. (1997) 'TOPMODEL: a critique', *Hydrological Processes*, 11(9), pp. 1069–1085. doi: 10.1002/(SICI)1099-1085(199707)11:9<1069::AID-HYP545>3.0.CO;2-O.
- Beven, K. and Binley, A. (1992) 'The future of distributed models: Model calibration and uncertainty prediction', *Hydrological Processes*, 6(3), pp. 279–298. doi: 10.1002/hyp.3360060305.
- Bishop, C. M. (1995) *Neural networks for pattern recognition*. Cambridge, UK: Oxford university press. doi: 10.1007/BF02811896.
- Blöschl, G. and Grayson, R. (2001) 'Spatial Observations and Interpolation', *Spatial Patterns in Catchment Hydrology: Observations and Modelling*, 113(10), pp. 13–16.
- Borga, M., Anagnostou, E., Blöschl, G. and Creutin, J. (2011) 'Flash flood forecasting, warning and risk management: The HYDRATE project', *Environmental Science & Policy - ENVIRON SCI POLICY*, 14, pp. 834–844. doi: 10.1016/j.envsci.2011.05.017.
- Bornancin-Plantier, A. (2013) *Conception de modèles de prévision des crues éclair par apprentissage artificiel*. UNIVERSITÉ PIERRE ET MARIE CURIE.
- Borovykh, A., Bohte, S. and Oosterlee, C. W. (2017) *Conditional time series forecasting with convolutional neural networks, Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*.
- Borrell, V. E. (2004) *Vers une modélisation hydrologique adaptée à la prévision opérationnelle des crues éclair. Application à de petits bassins versants du sud de la France*.
- Borrell, V. E., Chorda, J. and Dartus, D. (2005) 'Prévision des crues éclair', *Comptes Rendus - Geoscience*, 337(13), pp. 1109–1119. doi: 10.1016/j.crte.2005.06.009.
- Bosch, J. M. and Hewlett, J. D. (1982) 'A review of catchment experiments to determine the effect of vegetation changes on water yield and evapotranspiration', *Journal of Hydrology*, pp. 3–23. doi:

10.1016/0022-1694(82)90117-2.

Boudevillain, B., Argence, S., Claud, C., Ducrocq, V., Joly, B., Joly, A., Lambert, D., Nuissier, O., Plu, M., Arbogast, P., Berne, A., Chaboureau, J.-P., Chapon, B., Crepin, F., Delrieu, G., Doerflinger, E., Funatsu, B. M., Kirstetter, P.-E., Masson, F., Maynard, K., Richard, E., Sanchez, E., Terray, L. et Walpersdorf, A. (2009) 'Projet Cyprim, partie I : Cyclogenèses et précipitations intenses en région méditerranéenne : origines et caractéristiques', *La Météorologie*, 8(66), p. 18. doi: 10.4267/2042/28828.

Boughton, W. and Droop, O. (2003) 'Continuous simulation for design flood estimation - A review', *Environmental Modelling and Software*, 18(4), pp. 309–318. doi: 10.1016/S1364-8152(03)00004-5.

Bowden, G. J., Dandy, G. C. and Maier, H. R. (2005) 'Input determination for neural network models in water resources applications. Part 1 - Background and methodology', *Journal of Hydrology*, 301(1–4), pp. 75–92. doi: 10.1016/j.jhydrol.2004.06.021.

Bowden, G., Maier, H. and Dandy, G. (2002) 'Optimal Division of Data for Neural Network Models in Water Resources Applications', *Water Resources Research - WATER RESOUR RES*, 38(2). doi: 10.1029/2001WR000266.

Box, G. E. P., Jenkins, G. M., Reinsel, G. C. and Ljung, G. M. (2015) *Time series analysis: forecasting and control*. John Wiley & Sons.

Box, G. E. P., Jenkins, G. M. and Reinsel, G. C. (2013) 'Time series analysis: Forecasting and control: Fourth edition', *Time Series Analysis: Forecasting and Control: Fourth Edition*, pp. 1–746. doi: 10.1002/9781118619193.

Braud, I., Roux, H., Anquetin, S., Maubourguet, M. M., Manus, C., Viallet, P. and Dartus, D. (2010) 'The use of distributed hydrological models for the Gard 2002 flash flood event: Analysis of associated hydrological processes', *Journal of Hydrology*, 394(1–2), pp. 162–181. doi: 10.1016/j.jhydrol.2010.03.033.

Campolo, M., Soldati, A. and Andreussi, P. (1999) 'Forecasting river flow rate during low-flow periods using neural networks', *Water resources research*, 35(11), pp. 3547–3552.

Cappus, P. (1960) 'Étude Des Lois De L'Écoulement - Application Au Calcul Et À La Prévision Des Débits', *Houille Blanche*, pp. 493–520. doi: 10.1051/lhb/1960007.

Chandar, S., Sankar, C., Vorontsov, E., Kahou, S. E. and Bengio, Y. (2019) 'Towards non-saturating recurrent units for modelling long-term dependencies', *33rd AAAI Conference on Artificial Intelligence, AAAI 2019, 31st Innovative Applications of Artificial Intelligence Conference, IAAI 2019 and the 9th AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2019*, 33(01), pp. 3280–3287. doi: 10.1609/aaai.v33i01.33013280.

Chaudhari, P., Choromanska, A., Soatto, S., LeCun, Y., Baldassi, C., Borgs, C., Chayes, J., Sagun, L. and Zecchina, R. (2017) 'Entropy-SGD: Biasing gradient descent into wide valleys', in *Machine Learning (cs.LG); Machine Learning (stat.ML)*. arXiv preprint arXiv:1611.01838, 2016.

Chen, J., Song, L., Wainwright, M. and Jordan, M. (2018) 'Learning to explain: An information-theoretic perspective on model interpretation', in *International Conference on Machine Learning*. PMLR, pp. 883–892.

Chiang, Y.-M., Chang, L.-C. and Chang, F.-J. (2004) 'Comparison of static-feedforward and dynamic-feedback neural networks for rainfall–runoff modeling', *Journal of hydrology*, 290(3–4), pp. 297–311.

Cho, K., Van Merriënboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H. and Bengio, Y. (2014) 'Learning phrase representations using RNN encoder-decoder for statistical machine translation', *EMNLP 2014 - 2014 Conference on Empirical Methods in Natural Language Processing*,

Proceedings of the Conference, pp. 1724–1734. doi: 10.3115/v1/d14-1179.

Cosandey, C. (2003) *Les eaux courantes: Géographie et environnement*, Belin Sup Géographie. Belin.

Costa, J. E., Cheng, R. T., Haeni, F. P., Melcher, N., Spicer, K. R., Hayes, E., Plant, W., Hayes, K., Teague, C. and Barrick, D. (2006) 'Use of radars to monitor stream discharge by noncontact methods', *Water Resources Research*, 42(7). doi: 10.1029/2005WR004430.

Coulibaly, P., Anctil, F., Aravena, R. and Bobée, B. (2001) 'Artificial neural network modeling of water table depth fluctuations', *Water resources research*, 37(4), pp. 885–896.

Coulibaly, P., Anctil, F. and Bobée, B. (2000) 'Daily reservoir inflow forecasting using artificial neural networks with stopped training approach', *Journal of Hydrology*, 230(3–4), pp. 244–257. doi: 10.1016/S0022-1694(00)00214-6.

Coustau, M. (2011) *Contribution à la prévision des crues sur le bassin du Lez : modélisation de la relation pluie-débit en zone karstique et impact de l'assimilation de débits*. Université de Montpellier II.

Le Coz, J., Hauet, A., Pierrefeu, G., Dramais, G. and Camenen, B. (2010) 'Performance of image-based velocimetry (LSPIV) applied to flash-flood discharge measurements in Mediterranean rivers', *Journal of Hydrology*, 394(1–2), pp. 42–52. doi: 10.1016/j.jhydrol.2010.05.049.

Creutin, J. D., Delrieu, G. and Lebel, T. (1985) 'Estimation des lames d'eau moyennes et exceptionnelles', in *Petite hydraulique : problématique de développement*, p. 11 multigr.

Creutin, J. D. and Obled, C. (1982) 'Objective analyses and mapping techniques for rainfall fields: An objective comparison', *Water Resources Research*, 18, pp. 413–431.

Cui, Z., Chen, W. and Chen, Y. (2016) 'Multi-Scale Convolutional Neural Networks for Time Series Classification'.

Cybenko, G. (1989) 'Approximation by superpositions of a sigmoidal function', *Mathematics of Control, Signals, and Systems*, 2(4), pp. 303–314. doi: 10.1007/BF02551274.

Darras, T., Johannet, A., Vayssade, B., Kong-A-Siou, L. and Pistre, S. (2014) 'Influence of the Initialization of Multilayer Perceptron for Flash Flood Forecasting: Design of a Robust Model', *International Work-Conference on Time Series (Itise 2014)*, pp. 687–698.

Darras, T. (2015) *Prévision des crues rapides par apprentissage statistique*. Université de Montpellier. doi: 10.13140/RG.2.1.4451.6241.

Darras, T., Johannet, A., Vayssade, B., Kong-A-Siou, L. and Pistre, S. (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), pp. 565–578. doi: 10.21701/bolgeomin.129.3.007.

Delhomme, J. P. (1978) 'Kriging in the hydrosociences', *Advances in Water Resources*, 1(5), pp. 251–266. doi: [https://doi.org/10.1016/0309-1708\(78\)90039-8](https://doi.org/10.1016/0309-1708(78)90039-8).

Delrieu, G., Bellon, A. and Creutin, J. D. (1988) 'Estimation de lames d'eau spatiales à l'aide de données de pluviomètres et de radar météorologique - Application au pas de temps journalier dans la région de Montréal', *Journal of Hydrology*, 98(3–4), pp. 315–344. doi: 10.1016/0022-1694(88)90020-0.

Deng, J., Dong, W., Socher, R., Li, L., Li, K. and Fei-Fei, L. (2009) 'ImageNet: A large-scale hierarchical image database', in *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 248–255. doi: 10.1109/CVPR.2009.5206848.

Destro, E., Amponsah, W., Nikolopoulos, E. I., Marchi, L., Marra, F., Zocatelli, D. and Borga, M. (2018) 'Coupled prediction of flash flood response and debris flow occurrence: Application on an alpine

extreme flood event', *Journal of Hydrology*, 558, pp. 225–237.

Diakakis, M., Priskos, G. and Skordoulis, M. (2018) 'Public perception of flood risk in flash flood prone areas of Eastern Mediterranean: The case of Attica Region in Greece', *International Journal of Disaster Risk Reduction*, 28, pp. 404–413. doi: 10.1016/j.ijdrr.2018.03.018.

Dinu, C., Drobot, R., Pricop, C. and Blidaru, T. V (2017) 'FLASH-FLOOD MODELLING WITH ARTIFICIAL NEURAL NETWORKS USING RADAR RAINFALL ESTIMATES.', *Mathematical Modeling in Civil Engineering*, 13(3).

Diskin, M. H. and Nazimov, N. (1995) 'Linear reservoir with feedback regulated inlet as a model for the infiltration process', *Journal of Hydrology*, 172(1), pp. 313–330. doi: [https://doi.org/10.1016/0022-1694\(95\)02709-X](https://doi.org/10.1016/0022-1694(95)02709-X).

Dolciné, L., Andrieu, H., Sempere-Torres, D. and Creutin, D. (2001) 'Flash Flood Forecasting with Coupled Precipitation Model in Mountainous Mediterranean Basin', *Journal of Hydrologic Engineering*, 6(1), pp. 1–10. doi: 10.1061/(asce)1084-0699(2001)6:1(1).

Doya, K. (1993) 'Bifurcations of Recurrent Neural Networks in Gradient Descent Learning', *IEEE Transactions on Neural Networks*, 1(75), pp. 1–11.

Dramais, G., Le Coz, J., Le Boursicaud, R., Hauet, A., Lagouy, M. (2014) 'Jaugeage par radar mobile, protocole et résultats', *Houille Blanche*, 100(3), pp. 23–29. doi: 10.1051/lhb/2014025.

Dreyfus, G. (2004) *Réseaux de neurones: méthodologie et applications*, Springer. Paris, France: Eyrolles.

Dreyfus, G., Martinez, J.-M., Samuelides, M., Gordon, M. B., Badran, F. and Thiria, S. (2008) *Apprentissage statistique*. 3rd edn. Edited by Eyrolles. Paris, France.

Duan, Q., Sorooshian, S. and Gupta, V. (1992) 'Effective and efficient global optimization for conceptual rainfall-runoff models', *Water Resources Research*, 28(4), pp. 1015–1031. doi: 10.1029/91WR02985.

Duchi, J., Hazan, E. and Singer, Y. (2011) 'Adaptive subgradient methods for online learning and stochastic optimization.', *Journal of machine learning research*, 12(7).

Dupasquier, B., Andrieu, H., Delrieu, G., Griffith, R. J. and Cluckie, I. (2000) 'Influence of the VPR on high frequency fluctuations between radar and raingage data', *Physics and Chemistry of the Earth, Part B: Hydrology, Oceans and Atmosphere*, 25(10–12), pp. 1021–1025. doi: 10.1016/S1464-1909(00)00146-5.

Edijatno N., Yang, X., Makhlof, Z. and Michel, C. (1999) 'GR3J: A daily watershed model with three free parameters', *Hydrological Sciences Journal*, 44(2), pp. 263–277. doi: 10.1080/02626669909492221.

Elman, J. L. (1990) 'Finding Structure in Time', *Cognitive Science*, 14(2), pp. 179–211. doi: https://doi.org/10.1207/s15516709cog1402_1.

Fabry, F., Bellon, A., Duncan, M. R. and Austin, G. L. (1994) 'High resolution rainfall measurements by radar for very small basins: the sampling problem reexamined', *Journal of Hydrology*, 161(1–4), pp. 415–428. doi: 10.1016/0022-1694(94)90138-4.

Fang, Z., Wang, Y., Peng, L. and Hong, H. (2021) 'Predicting flood susceptibility using LSTM neural networks', *Journal of Hydrology*, 594, p. 125734.

Fritsch, J. M. (1995) 'Land use and hydrological regimes: a case study of French Guianan watersheds', *Annales de Géographie*, 581(582), pp. 26–63. doi: 10.3406/geo.1995.13867.

Funahashi, K. I. (1989) 'On the approximate realization of continuous mappings by neural networks', *Neural Networks*, 2(3), pp. 183–192. doi: 10.1016/0893-6080(89)90003-8.

- Gallinari, P. and Cibas, T. (1999) 'Practical complexity control in multilayer perceptrons', *Signal Processing*, 74(1), pp. 29–46. doi: 10.1016/S0165-1684(98)00200-X.
- 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, pp. 14–24. doi: <https://doi.org/10.1016/j.atmosres.2013.09.016>.
- Gaume, E. (2002) *Éléments d'analyse sur les crues éclair*. Université du Québec.
- Gaume, E., Bain, V., Bernardara, P., Newinger, O., Barbuc, M., Bateman, A., Blaškovičová, 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), pp. 70–78. doi: 10.1016/j.jhydrol.2008.12.028.
- Gaume, É., Bain, V. and Borga, M. (2010) 'Les crues éclair en Europe le projet de recherches hydrate', *Bulletin des Laboratoires des Ponts et Chaussées*, (277), pp. 65–73.
- Gaume, E. and Borga, M. (2008) 'Post-Flood Field Investigations in Upland Catchments After Major Flash Floods: Proposal of a Methodology and Illustrations', *Journal of Flood Risk Management*, 1, pp. 175–189. doi: 10.1111/j.1753-318X.2008.00023.x.
- Geman, S., Bienenstock, E. and Doursat, R. (1992) 'Neural Networks and the Bias/Variance Dilemma', *Neural Computation*, 4(1), pp. 1–58. doi: 10.1162/neco.1992.4.1.1.
- Georgakakos, K. P. (1986) 'A generalized stochastic hydrometeorological model for flood and flash-flood forecasting: 2. Case studies', *Water Resources Research*, 22(13), pp. 2096–2106. doi: <https://doi.org/10.1029/WR022i013p02096>.
- Gers, F. A., Schmidhuber, J. and Cummins, F. (2000) 'Learning to forget: Continual prediction with LSTM', *Neural computation*, 12(10), pp. 2451–2471.
- Giuntoli, I., Renard, B. and Lang, M. (2012) 'Floods in France', in *IAHS-AISH Publication*, pp. 199–211. doi: 10.1201/b12348-13.
- Giustolisi, O. and Laucelli, D. (2005) 'Improving generalization of artificial neural networks in rainfall–runoff modelling/Amélioration de la généralisation de réseaux de neurones artificiels pour la modélisation pluie-débit', *Hydrological Sciences Journal*, 50(3).
- Glorot, X. and Bengio, Y. (2010) 'Understanding the difficulty of training deep feedforward neural networks', *Journal of Machine Learning Research*. Edited by Y. W. Teh and M. Titterington. PMLR, pp. 249–256.
- Golden, R. M. (1996) *Mathematical methods for neural network analysis and design*, *Computers & Mathematics with Applications*. Massachusetts Institute of Technology. doi: 10.1016/s0898-1221(97)90143-7.
- Gourley, J. J., Flamig, Z. L., Vergara, H., Kirstetter, P-E., Clark, R. A., Argyle, E., Arthur, A., Martinaitis, S., Terti, G., Erlingis, J. M., Hong, Y. and Howard, K. W. (2017) 'The FLASH Project: Improving the Tools for Flash Flood Monitoring and Prediction across the United States', *Bulletin of the American Meteorological Society*, 98(2), pp. 361–372. doi: 10.1175/BAMS-D-15-00247.1.
- Goyal, P., Misra, N., Panolan, F. and Zehavi, M. (2015) 'Deterministic algorithms for matching and packing problems based on representative sets', *SIAM Journal on Discrete Mathematics*, 29(4), pp. 1815–1836.
- Graves, A. (2012) 'Sequence transduction with recurrent neural networks', *arXiv preprint arXiv:1211.3711*.

- Green, H. W. and Ampt, G. A. (1911) 'Studies on Soil Physics.', *The Journal of Agricultural Science*, 4(1), pp. 1–24. doi: 10.1017/S0021859600001441.
- Haugh, L. D. and Box, G. E. P. (1977) 'Identification of dynamic regression (distributed lag) models connecting two time series', *Journal of the American Statistical Association*, 72(357), pp. 121–130.
- He, K., Zhang, X., Ren, S. and Sun, J. (2016) 'Deep residual learning for image recognition', in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*. IEEE, pp. 770–778. doi: 10.1109/CVPR.2016.90.
- Hewlett, J. D. and Bosch, J. M. (1984) 'The dependence of storm flows on rainfall intensity and vegetal cover in South Africa', *Journal of Hydrology*, 75(1–4), pp. 365–381. doi: 10.1016/0022-1694(84)90060-X.
- Hewlett, J. D., Fortson, J. C. and Cunningham, G. B. (1977) 'The effect of rainfall intensity on storm flow and peak discharge from forest land', *Water Resources Research*, 13(2), pp. 259–266. doi: 10.1029/WR013i002p00259.
- Hewlett, J. D. and Hibbert, A. R. (1963) 'Moisture and energy conditions within a sloping soil mass during drainage', *Journal of Geophysical Research (1896-1977)*, 68(4), pp. 1081–1087. doi: <https://doi.org/10.1029/JZ068i004p01081>.
- Hewlett J.D. and Hibbert A.R. (1967) 'Factors Affecting the Response of Small Watersheds to Precipitation in Humid Regions.', *Forest Hydrology*, 33(2), pp. 275–290.
- El Hihi, S. and Bengio, Y. (1996) 'Hierarchical recurrent neural networks for long-term dependencies', in *Advances in neural information processing systems*, pp. 493–499.
- Hinton, G. E., Srivastava, N., Krizhevsky, A., Sutskever, I. and Salakhutdinov, R. R. (2012) 'Improving neural networks by preventing co-adaptation of feature detectors', *arXiv preprint arXiv:1207.0580*.
- Hinton, G. E., Osindero, S. and Teh, Y. W. (2006) 'A fast learning algorithm for deep belief nets', *Neural Computation*, 18(7), pp. 1527–1554. doi: 10.1162/neco.2006.18.7.1527.
- Hochreiter, S. and Schmidhuber, J. (1997) 'Long Short-Term Memory', *Neural Computation*, 9(8), pp. 1735–1780. doi: 10.1162/neco.1997.9.8.1735.
- Hollis, G. E. (1975) 'The effect of urbanization on floods of different recurrence interval', *Water Resources Research*, 11(3), pp. 431–435. doi: 10.1029/WR011i003p00431.
- Hornik, K., Stinchcombe, M. and White, H. (1989) 'Multilayer feedforward networks are universal approximators', *Neural Networks*, 2(5), pp. 359–366. doi: 10.1016/0893-6080(89)90020-8.
- Horton, R. E. (1933) 'The Rôle of infiltration in the hydrologic cycle', *Eos, Transactions American Geophysical Union*, 14(1), pp. 446–460. doi: 10.1029/TR014i001p00446.
- Huet, P., Martin, X., Prime, J.L., Baumont, G., Foin, P., Laurain, C., Cannard, P., Duband, D., Brun, E., Neppel, L., Coeur, D., Benoit, G., Ledoux, B., Valerie, P., Gresillon, J. M. (2003) *Crues du Gard 2002 : retour d'expérience.*, REPONSES ENVIRONNEMENT. LA DOCUMENTATION FRANCAISE. PARIS CN - 4149-D.
- IAHS (1974) *Flash Floods Symposium (International Association of Hydrological Sciences)*. Paris.
- IPCC (2012) *Managing the risks of extreme events and disasters to advance climate change adaptation [Field, C.B., V. Barros, T.F. Stocker, D. Qin, D.J. Dokken, K.L. Ebi, M.D. Mastrandrea, K.J. Mach, G.-K. Plattner, S.K. Allen, M. Tignor, and P.M. Midgley (Eds.)]*. A s. Cambridge, UK, New York, USA, 582 pp.
- Jacq, V. (1996) *Inventaire des situations a precipitations diluviennes sur le Languedoc-Roussillon la Provence-Alpes Cote d'Azur et la Corse*. réédition, Report No.3 of Service central d'exploitation de la

meteorologie. réédition. Paris: Météo-France (Phénomènes remarquables).

Jaeger, H. (2012) *Long short-term memory in echo state networks: Details of a simulation study*. Jacobs University Bremen.

Jain, S. K., Das, A. and Srivastava, D. K. (1999) 'Application of ANN for reservoir inflow prediction and operation', *Journal of water resources planning and management*, 125(5), pp. 263–271.

Jamali, B., Bach, P. M. and Deletic, A. (2020) 'Rainwater harvesting for urban flood management – An integrated modelling framework', *Water Research*, 171, p. 115372. doi: <https://doi.org/10.1016/j.watres.2019.115372>.

Jeannin, P.Y.; Artigue, G.; Butscher, C.; Chang, Y.; Charlier, J.B.; Duran, L.; Gill, L.; Hartmann, A.; Johannet, A.; Jourde, H.; Kavousi, A.; Liesch, T.; Liu, Y.; Lüthi, M.; Malard, A.; Mazzilli, N.; Pardo-Igúzquiza, E.; Thiéry, D.; Reimann, T.; Schuler, P.; Wöhling, T.; Wunsch, A. Karst modelling challenge 1: Results of hydrological modelling, *Journal of Hydrology*, Volume 600, 2021, 126508, <https://doi.org/10.1016/j.jhydrol.2021.126508>.

Jing, L., Shen, Y., Dubcek, T., Peurifoy, J., Skirlo, S., LeCun, Y., Tegmark, M. and Soljačić, M. (2017) 'Tunable efficient unitary neural networks (EUNN) and their application to RNNs', in Precup, D. and Teh, Y. W. (eds) *34th International Conference on Machine Learning, ICML 2017*. PMLR (Proceedings of Machine Learning Research), pp. 2753–2761.

Johannet, A. (2011) *Modélisation par apprentissage statistique des systèmes naturels, ou en interaction avec un environnement naturel. Applications aux karsts, crues éclair et en robotique*. doi: 10.13140/RG.2.1.2950.6401 M4.

Johannet, A., Mangin, A. and Hulst, D. D. (1994) 'Subterranean Water Infiltration Modelling by Neural Networks : Use of Water Source Flow', in *International Conference on Artificial Neural Networks ICANN 94. Vol II*. Sorrento, Italy, pp. 1033–1036.

Johannet, A., Vayssade, B. and Bertin, D. (2008) 'Neural Networks: From Black Box towards Transparent Box Application to Evapotranspiration Modeling', *International Journal of Computational Intelligence*, 4(3), p. 162.

Jordan, P., Seed, A. and Austin, G. (2000) 'Sampling errors in radar estimates of rainfall', *Journal of Geophysical Research Atmospheres*, 105(D2), pp. 2247–2257. doi: 10.1029/1999JD900130.

Kanuparthi, B., Arpit, D., Kerg, G., Ke, N. R., Mitliagkas, I. and Bengio, Y. (2019) 'H-Detach: Modifying the LSTM gradient towards better optimization', *7th International Conference on Learning Representations, ICLR 2019*.

Kazuhiko, F., Takayuki, Y., Hitoki, I. and Yoichi, T. (2021) 'Current Status of River Discharge Observation Using Non-Contact Current Meter for Operational Use in Japan', *World Environmental and Water Resources Congress 2008*. (Proceedings), pp. 1–10. doi: doi:10.1061/40976(316)278.

Kingma, D. P. and Ba, J. (2014) 'Adam: A method for stochastic optimization', *arXiv preprint arXiv:1412.6980*.

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), pp. 1034–1044. doi: 10.1029/WR016i006p01034.

Kohonen, T. (1982) 'Self-organized formation of topologically correct feature maps', *Biological cybernetics*, 43(1), pp. 59–69.

Koistinen, J. and Pahukka, T. (1984) 'A comparison between different gauging techniques of daily rainfall', in *22nd Conf. on Radar Meteorology*. Zurich.

- Kong-A-Siou, L., Cros, K., Johannet, A., Borrell-Estupina, V. and Pistre, S. (2013) 'KnoX method, or Knowledge eXtraction from neural network model. Case study on the Lez karst aquifer (southern France)', *Journal of Hydrology*, 507, pp. 19–32. doi: 10.1016/j.jhydrol.2013.10.011.
- Kong A Siou, L., Johannet, A., Borrell-Estupina, 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), pp. 367–380. doi: 10.1016/j.jhydrol.2011.04.015.
- Kong A Siou, L. (2011) 'Modélisation des crues de bassins karstiques par réseaux de neurones. Cas du bassin du Lez (France)', <http://www.theses.fr>, p. 232.
- Kong A Siou, L., Johannet, A., Borrell-Estupina, V. 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), pp. 2365–2375. doi: 10.1007/s12665-011-1450-9.
- Krajewski, W. (1987) 'Co-Kriging of Radar-Rainfall and Rain Gauge Data', *Journal of Geophysical Research*, 92, pp. 9571–9580. doi: 10.1029/JD092iD08p09571.
- Kratzert, F., Klotz, D., Brenner, C., Schulz, K. and Herrnegger, M. (2018) 'Rainfall-runoff modelling using Long Short-Term Memory (LSTM) networks', *Hydrology and Earth System Sciences*, 22(11), pp. 6005–6022. doi: 10.5194/hess-22-6005-2018.
- Kratzert, F., Herrnegger, M., Klotz, D., Hochreiter, S. and Klambauer, G. (2019) 'NeuralHydrology – Interpreting LSTMs in Hydrology', *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 11700 LNCS(7), pp. 347–362. doi: 10.1007/978-3-030-28954-6_19.
- Kuczera, G. (1982) 'Robust flood frequency models', *Water Resources Research*, 18(2), pp. 315–324. doi: 10.1029/WR018i002p00315.
- Lavabre, J., Sempere-Torres, D. and Cernesson, F. (1991) 'Etude du comportement hydrologique d'un petit bassin versant méditerranéen après la destruction de l'écosystème forestier par un incendie', *Hydrologie Continentale*, 6, pp. 121–132.
- LeCun, Y., Bottou, L., Bengio, Y. and Haffner, P. (1998) 'Gradient-based learning applied to document recognition', *Proceedings of the IEEE*, 86(11), pp. 2278–2324.
- Lecun, Y., Bengio, Y. and Hinton, G. (2015) 'Deep learning', *Nature*. Nature Publishing Group, pp. 436–444. doi: 10.1038/nature14539.
- Lei, D., Chen, X. and Zhao, J. (2018) *Opening the black box of deep learning*.
- Levenberg, K. (1944) 'A method for the solution of certain non-linear problems in least squares', *Quarterly of Applied Mathematics*, 2(2), pp. 164–168. doi: 10.1090/qam/10666.
- Lezcano-Casado, M. and Martínez-Rubio, D. (2019) 'Cheap orthogonal constraints in neural networks: A simple parametrization of the orthogonal and unitary group', in Chaudhuri, K. and Salakhutdinov, R. (eds) *36th International Conference on Machine Learning, ICML 2019*. PMLR (Proceedings of Machine Learning Research), pp. 6734–6743.
- Li, W., Kiaghadi, A. and Dawson, C. N. (2021) 'High Temporal Resolution Rainfall Runoff Modelling Using Long-Short-Term-Memory (LSTM) Networks', *Neural Computing and Applications*, 33(4), pp. 1261–1278. doi: 10.1007/s00521-020-05010-6 M4.
- Liong, S.-Y., Lim, W.-H. and Paudyal, G. N. (2000) 'River stage forecasting in Bangladesh: neural network approach', *Journal of computing in civil engineering*, 14(1), pp. 1–8.

- Llasat, M. M C., Llasat-Botija, M., Prat, M A., Porcú, F., Price, C., Mugnai, A., Lagouvardos, K., Kotroni, V., Katsanos, D., Michaelides, S., Yair, Y., Savvidou, K. and Nicolaidis, K. (2010) 'High-impact floods and flash floods in Mediterranean countries: the FLASH preliminary database', *Advances in Geosciences*, 23, pp. 47–55. doi: 10.5194/adgeo-23-47-2010.
- Llasat, M C., Llasat-Botija, M., Petrucci, O., Pasqua, A. A., Rosselló, J., Vinet, F. and Boissier, L. (2013) 'Towards a database on societal impact of Mediterranean floods within the framework of the HYMEX project', *Natural Hazards and Earth System Sciences*, 13(5), pp. 1337–1350. doi: 10.5194/nhess-13-1337-2013.
- Luenberger, D. G. and Ye, Y. (1984) *Linear and nonlinear programming*. Springer.
- Lumbroso, D. and Gaume, E. (2012) 'Reducing the uncertainty in indirect estimates of extreme flash flood discharges', *Journal of Hydrology - J HYDROL*, 414. doi: 10.1016/j.jhydrol.2011.08.048.
- MacKay, D. J. C. (1992) 'A Practical Bayesian Framework for Backpropagation Networks', *Neural Computation*, 4(3), pp. 448–472. doi: 10.1162/neco.1992.4.3.448.
- Madsen, H. (2000) 'Automatic calibration of a conceptual rainfall-runoff model using multiple objectives', *Journal of Hydrology*, 235(3–4), pp. 276–288. doi: 10.1016/S0022-1694(00)00279-1.
- Maduranga, K. D. G., Helfrich, K. E. and Ye, Q. (2019) 'Complex unitary recurrent neural networks using scaled cayley transform', *33rd AAAI Conference on Artificial Intelligence, AAAI 2019, 31st Innovative Applications of Artificial Intelligence Conference, IAAI 2019 and the 9th AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2019*, 33(01), pp. 4528–4535. doi: 10.1609/aaai.v33i01.33014528.
- Maier, H. R., Jain, A., Dandy, G. C. and Sudheer, K. P. (2010) 'Methods used for the development of neural networks for the prediction of water resource variables in river systems: Current status and future directions', *Environmental Modelling and Software*, 25(8), pp. 891–909. doi: 10.1016/j.envsoft.2010.02.003.
- Maier, H. R. and Dandy, G. C. (1997) 'Determining inputs for neural network models of multivariate time series', *Computer-Aided Civil and Infrastructure Engineering*, 12(5), pp. 353–368.
- Maier, H. R. and Dandy, G. C. (2000) 'Neural networks for the prediction and forecasting of water resources variables: A review of modelling issues and applications', *Environmental Modelling and Software*, 15(1), pp. 101–124. doi: 10.1016/S1364-8152(99)00007-9.
- Maier, H. R., Dandy, G. C. and Burch, M. D. (1998) 'Use of artificial neural networks for modelling cyanobacteria *Anabaena* spp. in the River Murray, South Australia', *Ecological Modelling*, 105(2–3), pp. 257–272.
- Mangin, A. (1984) 'Pour une meilleure connaissance des systèmes hydrologiques à partir des analyses corrélatoire et spectrale', *Journal of Hydrology*, 67(1–4), pp. 25–43. doi: 10.1016/0022-1694(84)90230-0.
- 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 .', *Universite Montpellier li Sciences Et Techniques Du Languedoc*, pp. 1–214.
- Marcos, R., Llasat, M. C., Turco, M., Gilabert, J. and Llasat-Botija, M. (2016) 'Trends in flash flood events versus convective precipitation in the Mediterranean region: The case of Catalonia', *Journal of Hydrology*, 541, pp. 24–37. doi: 10.1016/j.jhydrol.2016.05.040.
- Marquardt, D. W. (1963) 'An Algorithm for Least-Squares Estimation of Nonlinear Parameters', *Journal*

of the Society for Industrial and Applied Mathematics, 11(2), pp. 431–441. doi: 10.1137/0111030.

Marr, B. (2020) *Tech Trends in Practice: The 25 Technologies that are Driving the 4th Industrial Revolution*. John Wiley & Sons.

Martens, J. and Sutskever, I. (2011) 'Learning recurrent neural networks with Hessian-free optimization', in *Proceedings of the 28th International Conference on Machine Learning, ICML 2011*, pp. 1033–1040.

Matalas, N. C. and Fiering, M. B. (1977) 'Water-Resource Systems Planning. Chapter 6 in Climate, Climatic Change, and Water Supply'. Washington, DC: National Academy of Sciences TS - EndNote Tagged Import Format M4 - Citavi.

Mathevet, T. (2005) *Quels Modèles Pluie-Debit Globaux au pas de temps horaire ?\rDéveloppements Empiriques et Comparaison de\rmodèles sur un large échantillon de Bassins Versants, Ecole Nationale du Génie Rural, des Eaux et Forêts*. ENGREF PARIS.

McCulloch, W. S. and Pitts, W. (1943) 'A logical calculus of the ideas immanent in nervous activity', *The Bulletin of Mathematical Biophysics*, 5(4), pp. 115–133. doi: 10.1007/BF02478259.

Meunier, M. (1996) 'Couvert forestier et crues sur les petits bassins versants de montagne', *Unasylva*, Influences(185), p. 62.

Mhammedi, Z., Hellicar, A., Rahman, A. and Bailey, J. (2017) 'Efficient orthogonal parametrisation of recurrent neural networks using householder reflections', in Precup, D. and Teh, Y. W. (eds) *34th International Conference on Machine Learning, ICML 2017*. PMLR (Proceedings of Machine Learning Research), pp. 3708–3720.

Mikolov, T., Karafiát, M., Burget, L., Cernocký, J. and Khudanpur, S. (2010) *Recurrent neural network based language model, Proceedings of the 11th Annual Conference of the International Speech Communication Association, INTERSPEECH 2010*.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S. and Dean, J. (2013) 'Distributed representations of words and phrases and their compositionality', in *Advances in neural information processing systems*, pp. 3111–3119.

Minsky, M., Papert, S. A. and Bottou, L. (2017) *Perceptrons, The MIT Press Ser.* Cambridge: MIT Press.

Mishkin, D. and Matas, J. (2016) 'All you need is a good init', *4th International Conference on Learning Representations, ICLR 2016 - Conference Track Proceedings*.

Montz, B. E. and Gruntfest, E. (2002) 'Flash flood mitigation: Recommendations for research and applications', *Environmental Hazards*, 4(1), pp. 15–22. doi: 10.3763/EHAZ.2002.0402.

Morel-Seytoux, H. J. (1978) 'Derivation of equations for variable rainfall infiltration', *Water Resources Research*, 14(4), pp. 561–568. doi: 10.1029/WR014i004p00561.

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), pp. 35–51. doi: 10.1016/j.jhydrol.2007.01.028.

Müller, J., Park, J., Sahu, R., Varadharajan, C., Arora, B., Faybishenko, B. and Agarwal, D. (2021) 'Surrogate optimization of deep neural networks for groundwater predictions', *Journal of Global Optimization*, 81(1), pp. 203–231. doi: 10.1007/s10898-020-00912-0.

Musy, A. (2009) *Hydrologie 1 : Une science de la nature Une gestion sociétale. 2ème édition revue et augmentée, Ingénierie de l'environnement*. PRESSES POLYTECHNIQUES ET UNIVERSITAIRES ROMANDES CN - EAU6823.

- Musy, A., Higy, C. and Reynard, E. (2014) *Hydrologie: Tome 1, Une science de la nature, une gestion sociétale*. 2nd edn. Edited by EPFL Press.
- Nash, J. E. and Sutcliffe, J. V. (1970) 'River flow forecasting through conceptual models part I - A discussion of principles', *Journal of Hydrology*, 10(3), pp. 282–290. doi: 10.1016/0022-1694(70)90255-6.
- Nérot, B. (2018) *'Réseaux de neurones : présentation de la méthode LSTM - Mise en œuvre sur un cas de prévision hydrologique'* Alès, France.
- Nerrand, O., Roussel-Ragot, P., Personnaz, L., Dreyfus, G. and Marcos, S. (1993) 'Neural Networks and Nonlinear Adaptive Filtering: Unifying Concepts and New Algorithms', *Neural Computation*, 5(2), pp. 165–199. doi: 10.1162/neco.1993.5.2.165.
- Ngo, P.-T. T., Hoang, N.-D., Pradhan, B., Nguyen, Q. K., Tran, X. T., Nguyen, Q. M., Nguyen, V. N., Samui, P. and Tien Bui, D. (2018) 'A Novel Hybrid Swarm Optimized Multilayer Neural Network for Spatial Prediction of Flash Floods in Tropical Areas Using Sentinel-1 SAR Imagery and Geospatial Data', *Sensors*. doi: 10.3390/s18113704.
- 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), pp. 165–178. doi: 10.1016/j.jhydrol.2010.12.020.
- Nystuen, J. A. (1998) 'Temporal sampling requirements for automatic rain gauges', *Journal of Atmospheric and Oceanic Technology*, 15(6), pp. 1253–1260. doi: 10.1175/1520-0426(1998)015<1253:TSRFAR>2.0.CO;2.
- OMM (1992) *Plan à long terme de l'OMM : orientations générales et stratégies 1992-2001 - Troisième Plan à long terme, partie I*. Genève: OMM.
- OMM (2011) *Manuel sur la prévision et l'annonce des crues*. 2011th edn, OMM-N° 1072. 2011th edn.
- Ouarda, T. B. M. J. and Ashkar, F. (1998) 'Effect of Trimming on LP III Flood Quantile Estimates', *Journal of Hydrologic Engineering*, 3(1), pp. 33–42. doi: 10.1061/(asce)1084-0699(1998)3:1(33).
- Oussar, Y. and Dreyfus, G. (2001) 'How to be a gray box: Dynamic semi-physical modeling', *Neural Networks*, 14(9), pp. 1161–1172. doi: 10.1016/S0893-6080(01)00096-X.
- Pascanu, R., Mikolov, T. and Bengio, Y. (2013) *On the difficulty of training recurrent neural networks, 30th International Conference on Machine Learning, ICML 2013*.
- Payraudeau, S. (2002) *Distributed modelling of nitrogen loads on small Mediterranean catchments T4 - Modélisation distribuée des flux d'azote sur des petits bassins versants méditerranéens*. ENGREF (AgroParisTech).
- Pellarin, T., Delrieu, G., Saulnier, G. M., Andrieu, H., Vignal, B. and Creutin, J. D. (2002) 'Hydrologic visibility of weather radar systems operating in mountainous regions: Case study for the Ardèche catchment (France)', *Journal of Hydrometeorology*, 3(5), pp. 539–555. doi: 10.1175/1525-7541(2002)003<0539:HVOWRS>2.0.CO;2.
- Perrin, C. (2000) *Vers une amélioration d'un modèle global pluie-débit au travers d'une approche comparative*. École doctorale terre, univers, environnement (Grenoble).
- Perrin, C., Michel, C. and Andréassian, V. (2007) 'Modèles hydrologiques du Génie Rural (GR)', *Cemagref, UR Hydrosystèmes et Bioprocédés*, pp. 1–16.
- Philip, J. R. (1957) 'The theory of infiltration: 1. The infiltration equation and its solution', *Soil Science*, 83(5), pp. 345–357. doi: 10.1097/00010694-195705000-00002.

- Picard, R. R. and Cook, R. D. (1984) 'Cross-validation of regression models', *Journal of the American Statistical Association*, 79(387), pp. 575–583. doi: 10.1080/01621459.1984.10478083.
- Piotrowski, A., Napiorkowski, J. J. and Rowiński, P. M. (2006) 'Flash-flood forecasting by means of neural networks and nearest neighbour approach—a comparative study', *Nonlinear Processes in Geophysics*, 13(4), pp. 443–448.
- Press, W. H., Teukolsky, S. A., Vetterling, W. T. and Flannery, B. P. (1992) *Numerical recipes in C: The art of scientific computing*. 2nd edn. Cambridge University Press.
- Price, C., Yair, Y., Mugnai, A., Lagouvardos, K., Llasat, M. C., Michaelides, S., Dayan, U., Dietrich, S., Di Paola, F. and Galanti, E. (2011) 'Using lightning data to better understand and predict flash floods in the Mediterranean', *Surveys in geophysics*, 32(6), p. 733.
- Qian, K., Mohamed, A. and Claudel, C. (2019) *Physics Informed Data Driven model for Flood Prediction: Application of Deep Learning in prediction of urban flood development*.
- Rasmussen, T. C., Baldwin, R. H., Dowd, J. F. and Williams, A. G. (2000) 'Tracer vs. Pressure Wave Velocities through Unsaturated Saprolite', *Soil Science Society of America Journal*, 64(1), pp. 75–85. doi: 10.2136/sssaj2000.64175x.
- Rojas, R. (2017) *Deepest Neural Networks*.
- Ruin, I., Lutoff, C. and Shabou, S. (2017) 'Anticipating or Coping: Behaviors in the Face of Flash Floods', in *Floods*, pp. 259–275. doi: 10.1016/B978-1-78548-269-4.50018-4.
- Rumelhart, D. E., Hinton, G. E. and Williams, R. J. (1986) 'Learning representations by back-propagating errors', *Nature*, 323(6088), pp. 533–536. doi: 10.1038/323533a0.
- Rumelhart, D. E. and McClelland, J. L. (1968) *Parallel distributed processing: explorations in the microstructure of cognition, vol. 1: foundations*. Cambridge, MA, USA: MIT Press.
- Sahoo, G. B., Ray, C. and De Carlo, E. H. (2006) 'Use of neural network to predict flash flood and attendant water qualities of a mountainous stream on Oahu, Hawaii', *Journal of Hydrology*, 327(3–4), pp. 525–538.
- Sahu, R. K., Müller, J., Park, J., Varadharajan, C., Arora, B., Faybishenko, B. and Agarwal, D. (2020) 'Impact of Input Feature Selection on Groundwater Level Prediction From a Multi-Layer Perceptron Neural Network', *Frontiers in Water*, 2, p. 46. doi: 10.3389/frwa.2020.573034.
- Saint Fleur, B. E., Artigue, G., Johannet, A. and Pistre, S. (2020) 'Deep Multilayer Perceptron for Knowledge Extraction: Understanding the Gardon de Mialet Flash Floods Modeling', in Valenzuela, O. et al. (eds) *Theory and Applications of Time Series Analysis*. Theory and. Cham: Springer International Publishing, pp. 333–348. doi: 10.1007/978-3-030-56219-9_22.
- Saint Fleur, B., Artigue, G., Johannet, A. and Pistre, S. (2019) 'Knowledge Extraction (KnoX) in Deep Learning: Application to the Gardon de Mialet Flash Floods Modelling', in O. Valenzuela, F. Rojas, L. J. Herrera, H. Pomares, & I. R. (Eds. . (ed.) *Itise 2019*. Granada, Spain: Web of sciences, pp. 178–189.
- Schleiter, I. M., Borchardt, D., Wagner, R., Dapper, T., Schmidt, K.-D., Schmidt, H.-H. and Werner, H. (1999) 'Modelling water quality, bioindication and population dynamics in lotic ecosystems using neural networks', *Ecological Modelling*, 120(2–3), pp. 271–286.
- Schmidhuber, J. (1992) 'Learning to Control Fast-Weight Memories: An Alternative to Dynamic Recurrent Networks', *Neural Computation*, 4(1), pp. 131–139. doi: 10.1162/neco.1992.4.1.131.
- Schofield, N. J. and Ruprecht, J. K. (1989) 'Regional analysis of stream salinisation in southwest Western Australia', *Journal of Hydrology*, 112(1), pp. 19–39. doi: [https://doi.org/10.1016/0022-1694\(89\)90178-](https://doi.org/10.1016/0022-1694(89)90178-)

9.

Sejnowski, T. J. and Rosenberg, C. R. (1987) 'Parallel networks that learn to pronounce English text', *Complex systems*, 1(1), pp. 145–168.

Shannon, C. E. (1998) 'Communication In The Presence Of Noise', *Proceedings of the IEEE*, 86(2), pp. 447–457. doi: 10.1109/JPROC.1998.659497.

Sharma, A. (2000) 'Seasonal to interannual rainfall probabilistic forecasts for improved water supply management: Part 3—A nonparametric probabilistic forecast model', *Journal of Hydrology*, 239(1–4), pp. 249–258.

Shen, C. (2018a) 'A Transdisciplinary Review of Deep Learning Research and Its Relevance for Water Resources Scientists', *Water Resources Research*, 54(11), pp. 8558–8593. doi: 10.1029/2018WR022643.

Shen, C. and Lawson, K. (2021) 'Applications of Deep Learning in Hydrology', *Deep Learning for the Earth Sciences*. (Wiley Online Books), pp. 283–297. doi: <https://doi.org/10.1002/9781119646181.ch19>.

Singh, V P (1995) 'What is environmental hydrology?', in Singh, Vijay P (ed.) *Environmental Hydrology*. Dordrecht: Springer Netherlands, pp. 1–12. doi: 10.1007/978-94-017-1439-6_1.

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), pp. 1691–1724. doi: 10.1016/0005-1098(95)00120-8.

Sood, A. and Smakhtin, V. (2015) 'Revue des modèles hydrologiques globaux', *Hydrological Sciences Journal*, 60(4), pp. 549–565. doi: 10.1080/02626667.2014.950580.

Steiner, M. (1996) 'Uncertainty of estimates of monthly areal rainfall for temporally sparse remote observations', *Water Resources Research*, 32(2), pp. 373–388. doi: 10.1029/95WR03396.

Steiner, M., Bell, T. L., Zhang, Y. and Wood, E. F. (2003) 'Comparison of two methods for estimating the sampling-related uncertainty of satellite rainfall averages based on a large radar dataset', *Journal of Climate*, 16(22), pp. 3759–3778. doi: 10.1175/1520-0442(2003)016<3759:COTMFE>2.0.CO;2.

Stone, M. (1976) 'Cross-Validatory Choice and Assessment of Statistical Predictions (With Discussion)', *Journal of the Royal Statistical Society: Series B (Methodological)*, 38(1), pp. 102–102. doi: 10.1111/j.2517-6161.1976.tb01573.x.

Sundermeyer, M., Schlüter, R. and Ney, H. (2012) 'LSTM neural networks for language modeling', in *Thirteenth annual conference of the international speech communication association*.

Sutton, R. S. and Barto, A. G. (2018) *Reinforcement learning: An introduction*. MIT press.

Swann, W. H. (1969) 'A survey of non-linear optimization techniques', *FEBS Letters*, 2(SUPPL. 1), pp. S39–S55. doi: 10.1016/0014-5793(69)80075-x.

Tabari, H. and Talaee, P. H. (2013) 'Moisture index for Iran: spatial and temporal analyses', *Global and Planetary Change*, 100, pp. 11–19.

Tabary, P., Scialom, G. and Protat, A. (2002) 'Un radar météorologique Doppler : pour quoi faire ?', *La Météorologie*, 8(38), p. 30. doi: 10.4267/2042/36232.

Taver, V., Johannet, A., Borrell-Estupina, V. and Pistre, S. (2015) 'Modèles de réseaux de neurones récurrents vs non-récurrents pour la modélisation non-stationnaire utilisant l'assimilation des données et l'adaptabilité', *Hydrological Sciences Journal*, 60(7–8), pp. 1242–1265. doi: 10.1080/02626667.2014.967696.

- Testud, J., Erwan, L. B., Estelle, O. and Mustapha, A.-M. (2000) 'The Rain Profiling Algorithm Applied to Polarimetric Weather Radar', *Journal of Atmospheric and Oceanic Technology*, 17, pp. 332–356.
- Tieleman, T. and Hinton, G. (2012) 'Lecture 6.5-rmsprop: Divide the gradient by a running average of its recent magnitude', *COURSERA: Neural networks for machine learning*, 4(2), pp. 26–31.
- Tien Bui, D., Pradhan, B., Lofman, O. and Revhaug, I. (2012) 'Landslide Susceptibility Assessment in Vietnam Using Support Vector Machines, Decision Tree, and Naïve Bayes Models', *Mathematical Problems in Engineering*. Edited by W.-C. Hong, 2012, p. 974638. doi: 10.1155/2012/974638.
- Tokar, A. S. and Johnson, P. A. (1999) 'Rainfall-runoff modeling using artificial neural networks', *Journal of Hydrologic Engineering*, 4(3), pp. 232–239.
- Toukourou, M., Johannet, A., Dreyfus, G. and Ayrat, P.-A. (2011) 'Rainfall-runoff modeling of flash floods in the absence of rainfall forecasts: The case of "cévenol flash floods"', *Appl. Intell.*, 35, pp. 178–189. doi: 10.1007/s10489-010-0210-y.
- Toukourou, M., Johannet, A., Dreyfus, G. (2009). 'Flash Flood Forecasting by Statistical Learning in the Absence of Rainfall Forecast: a Case Study' in Engineering Applications of Neural Networks, 11th International Conference, EANN 2009, London, UK, August 27-29. Proceedings Editors, Dominic Palmer-Brown, Chrisina Draganova, Elias Pimenidis, Haris Mouratidis, CCIS series, SpringerToukourou, M. S. (2009) *Application de l'apprentissage artificiel à la prévision des crues éclair*. Ecole Nationale des Mines de Paris.
- Trenn, S. (2008) 'Multilayer perceptrons: Approximation order and necessary number of hidden units', *IEEE Transactions on Neural Networks*, 19(5), pp. 836–844. doi: 10.1109/TNN.2007.912306.
- Trigo, R. M., Ramos, C., Pereira, S. S., Ramos, A. M., Zêzere, J. L. and Liberato, M. L.R. (2016) 'The deadliest storm of the 20th century striking Portugal: Flood impacts and atmospheric circulation', *Journal of Hydrology*, 541, pp. 597–610. doi: 10.1016/j.jhydrol.2015.10.036.
- Tukey, J. W. (1961) 'Discussion, Emphasizing the Connection Between Analysis of Variance and Spectrum Analysis', *Technometrics*, 3(2), pp. 191–219. doi: 10.1080/00401706.1961.10489940.
- Vannier, O., Anquetin, S. and Braud, I. (2016) 'Investigating the role of geology in the hydrological response of Mediterranean catchments prone to flash-floods: Regional modelling study and process understanding', *Journal of Hydrology*, 541, pp. 158–172. doi: 10.1016/j.jhydrol.2016.04.001.
- Villarini, G., Mandapaka, G., Mandapaka, P. V., Krajewski, W. F. and Moore, R. J. (2008) 'Rainfall and sampling uncertainties: A rain gauge perspective', *Journal of Geophysical Research: Atmospheres*, 113(D11). doi: <https://doi.org/10.1029/2007JD009214>.
- Vincent, P., Larochelle, H., Bengio, Y. and Manzagol, P. A. (2008) 'Extracting and composing robust features with denoising autoencoders', in *Proceedings of the 25th International Conference on Machine Learning*, pp. 1096–1103. doi: 10.1145/1390156.1390294.
- Wasko, C. and Sharma, A. (2017) 'Global assessment of flood and storm extremes with increased temperatures', *Scientific Reports*, 7(1), p. 7945. doi: 10.1038/s41598-017-08481-1.
- Werbos, P. J. (1988) 'Generalization of backpropagation with application to a recurrent gas market model', *Neural Networks*, 1(4), pp. 339–356. doi: [https://doi.org/10.1016/0893-6080\(88\)90007-X](https://doi.org/10.1016/0893-6080(88)90007-X).
- Wilson, J. W. and Brandes, E. A. (1979) 'Radar Measurement of Rainfall—A Summary', *Bulletin of the American Meteorological Society*, 60(9), pp. 1048–1060. doi: 10.1175/1520-0477(1979)060<1048:RMORS>2.0.CO;2.
- Xie, C., Zhang, Z., Yuille, A. L., Wang, J. and Ren, Z. (2018) 'Mitigating adversarial effects through randomization', *6th International Conference on Learning Representations, ICLR 2018 - Conference*

Track Proceedings.

Yang, X. and Parent, E. (1996) 'Analyse de fiabilité en modélisation hydrologique: Concepts et applications au modèle pluies-débits GR3', *Revue des sciences de l'eau*, 9(1), pp. 31–49. doi: 10.7202/705241ar

Yuan, Y., Xu, Y.-S. and Arulrajah, A. (2017) 'Sustainable Measures for Mitigation of Flooding Hazards: A Case Study in Shanghai, China', *Water*. doi: 10.3390/w9050310.

Zhang, J., Zhu, Y., Zhang, X., Ye, M. and Yang, J. (2018) 'Developing a Long Short-Term Memory (LSTM) based model for predicting water table depth in agricultural areas', *Journal of hydrology*, 561, pp. 918–929.

Zhang, S., Choromanska, A. and LeCun, Y. (2015) 'Deep learning with elastic averaging SGD: 3rd International Conference on Learning Representations, ICLR 2015', in *3rd International Conference on Learning Representations, ICLR 2015*.

Zhang, Z. and Brand, M. (2017) 'On the convergence of block coordinate descent in training DNNs with Tikhonov regularization', in *Advances in Neural Information Processing Systems*, pp. 1719–1728.

Zhang, Z., Wu, Y. and Wang, G. (2018) *BPGrad: Towards Global Optimality in Deep Learning via Branch and Pruning*, *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*. doi: 10.1109/CVPR.2018.00348.

Zolezzi, G. Zamlar, D., Laronne, J.-B., Salvaro, M., Piazza, F., Le Coz, J., Welber, M. and Dramais, G. (2011) 'A systematic test of surface velocity radar (SVR) to improve flood discharge prediction', in *AGU Fall Meeting Abstracts*.

Autres références et sites internet

Fjodor Van Veen, 2016 : <https://www.asimovinstitute.org/neural-network-zoo/>, consulté le 17/09/2021)

INRAE, 2021 : « https://webgr.inrae.fr/recherche/modelisation_hydrologique/ », consulté le 15 juin 2021

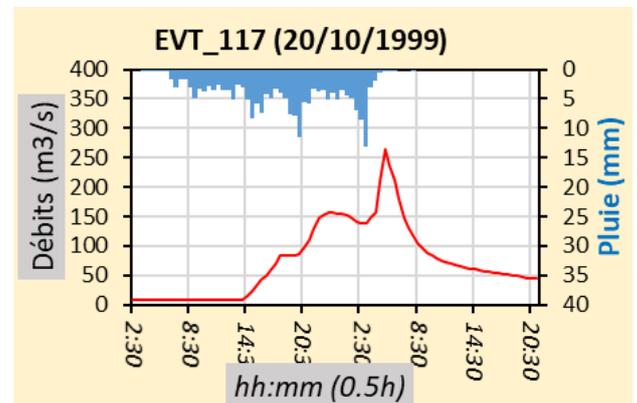
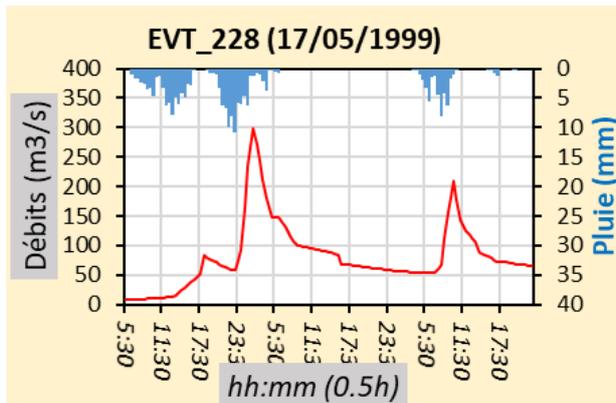
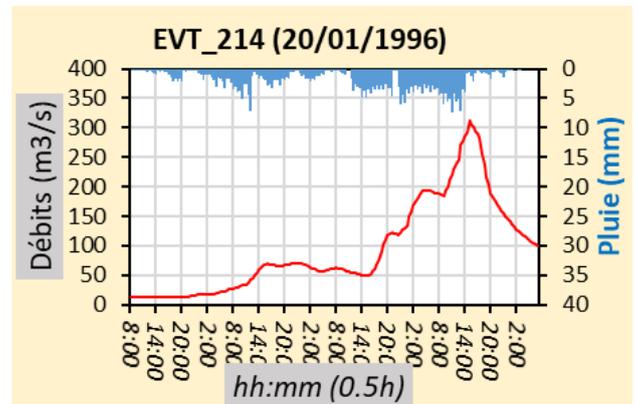
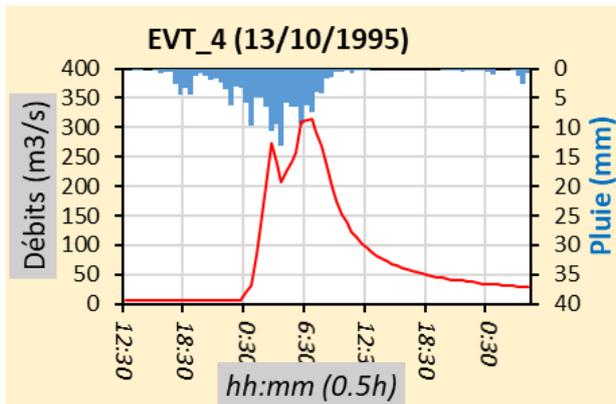
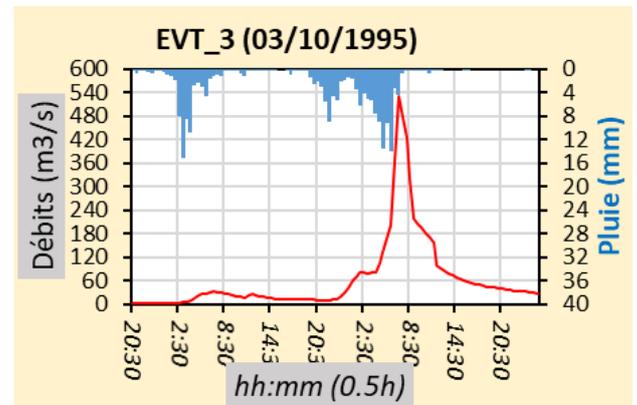
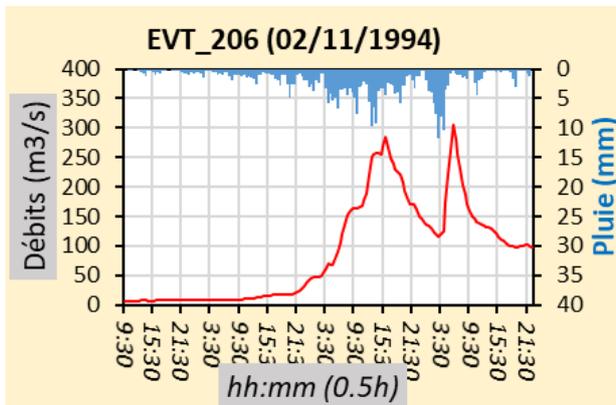
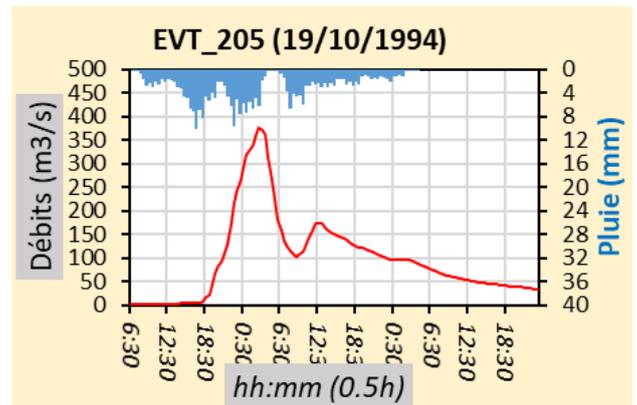
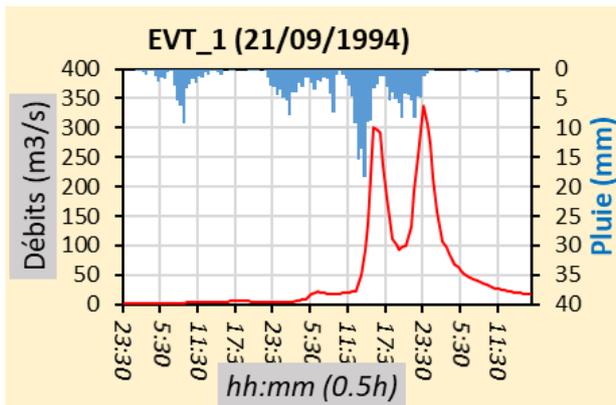
Ministry of Water Resources, 2018 : 2017 Water resources management yearbook of China, Dept. of Water Resources

Annexes

Annexe 1: Résumé des 58 évènements

Date	N°	Cumul (mm)			Max (Intensité ou débit) [30 min]			Durée (h)	Qs > 1 m³/s/km²
		BDC	SRDT	Mia	SRDT	Q (m³/s)	Qs (m³/s/km²)		
22/09/1993 04:00	200	16,8	169,3	103,4	39,3	118	0,54	39,5	
31/10/1993 06:00	202	160,7	190,2	210,1	7,3	79	0,36	119,5	
22/09/1994 00:30	1	315	293	261,5	16,6	337	1,53	67,5	V
19/10/1994 07:00	205	199,6	286	385,8	13,6	377	1,71	67,5	V
02/11/1994 10:00	206	411,5	331	234	10,2	304	1,38	92	V
10/09/1995 21:30	208	78	190,5	214	29,1	31	0,14	67,5	
03/10/1995 20:30	3	144	209,5	358,5	22,6	531	2,41	53	V
13/10/1995 13:00	4	70	203,5	267,5	18,7	313	1,42	42,5	V
24/11/1995 16:00	209	156,3	160,8	132,8	9,2	132	0,60	92	
11/01/1996 21:00	213	138,8	188,8	199,3	8	193	0,88	55	
20/01/1996 08:30	214	319,8	439,6	268,8	9,2	312	1,42	105	V
13/03/1996 17:30	216	94,5	145	130,5	3,5	37	0,17	92,5	
18/09/1996 19:00	218	185	254	155,5	21,1	95	0,43	42,5	
13/10/1996 04:30	219	158,5	163,5	107	7,6	43	0,20	86	
10/11/1996 13:00	8	215	225,5	225,5	11,5	269	1,22	80	
04/12/1996 00:30	220	236,5	230	171,6	5,5	66	0,30	117,5	
12/01/1998 13:30	222	124	155,5	71,6	7,7	200	0,91	67,5	S
29/04/1998 13:00	224	91	136,5	131	13,7	120	0,55	54,5	
25/09/1998 10:30	226	162,5	135,5	88,1	12,9	61	0,28	55	
30/12/1998 13:00	227	277,3	315,8	242,5	13,1	143	0,65	76,5	
17/05/1999 06:00	228	150,9	234,8	175	20,6	299	1,36	67	V
20/10/1999 03:00	117	207,5	305	158,9	19,6	263	1,20	47,5	V
11/11/1999 17:00	229	183,5	168,6	112	5,4	72	0,33	116,5	
28/09/2000 10:00	13	205,5	261	193	22,3	454	2,06	55	V
08/09/2002 12:30	19	70	303,5	346,5	101	833	3,79	33,5	V
09/10/2002 02:00	315	238	225	145	18	120	0,54	68,5	
09/12/2002 16:30	317	149,5	216,5	156	10	175	0,80	82,5	
30/09/2003 11:00	301	155	120,5	155	20,5	145	0,66	42,5	
15/11/2003 13:30	302	134,5	166	134,5	15	205	0,93	67	
22/11/2003 07:00	318	369,5	203,5	369,5	29,5	224	1,02	63,5	V
29/11/2003 18:00	303	344	329,5	344	26	340	1,55	117,5	V
30/03/2004 05:00	305	205	260	157	13,5	169	0,77	116,5	
28/04/2004 23:30	306	178,5	179	125	14,5	205	0,93	45	
26/10/2004 01:00	307	129,5	202,5	179,5	22	82	0,37	113,5	
05/09/2005 21:00	308	186,5	216,5	258	19	74	0,33	105	
28/01/2006 04:30	309	133	205	154	19	141	0,64	47,5	
13/09/2006 10:00	310	155,5	149	178,5	37	31	0,14	55	
18/10/2006 11:30	23	322,5	256	198	18,5	102	0,46	39,5	
20/11/2007 21:30	25	263	279	323,5	11	53	0,24	92,5	
03/01/2008 11:00	311	120	208,5	166	12,5	40	0,18	40	
31/10/2008 10:30	27	318,5	308,5	227	35	498	2,26	69	V
31/01/2009 05:00	324	173	200,5	266,5	9	160	0,73	116	
20/10/2009 07:00	327	161	171	151	16,5	99	0,45	47,5	
12/03/2011 11:30	331	263,5	343,5	257	10	152	0,69	117,5	
23/10/2011 16:00	387	163,5	88,5	130,5	11	24	0,11	58	
02/11/2011 12:00	332	508,5	606	429,5	37,5	430	1,95	95	V
16/03/2013 15:00	340	101,5	141,5	203,5	8,5	152	0,69	57,5	
17/09/2014 01:00	349	172,5	178,5	362	28	87	0,40	92,5	
09/10/2014 19:30	350	181	242,5	255	36,5	284	1,29	92,5	V
27/11/2014 18:00	353	247,5	202,5	116	21	183	0,83	75,5	
11/09/2015 23:30	355	179,5	172,5	249	22	200	0,91	56,5	
27/10/2015 00:00	356	110,5	162,5	234	30	494	2,24	57,5	V
14/09/2016 02:30	362	146,5	142,5	111,5	30,5	89	0,41	27	
12/10/2016 18:00	363	166	154	143,5	15	78	0,36	68,5	
20/11/2016 19:00	365	244,5	205	155,5	18,5	317	1,44	67,5	V
03/11/2017 21:30	372	0	195,5	138,5	38	182	0,83	42	
30/10/2018 19:30	385	126,5	146	56	19,5	99	0,45	62,5	
05/11/2018 07:30	386	229,5	235,5	70	21	275	1,25	132,5	V
Légende		1992-2002	2002-2008		2009-2019		V : Débit spécifique de plus de 1 m³/s/km²		

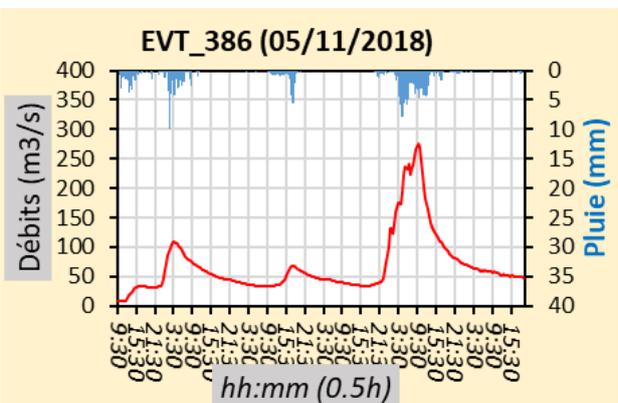
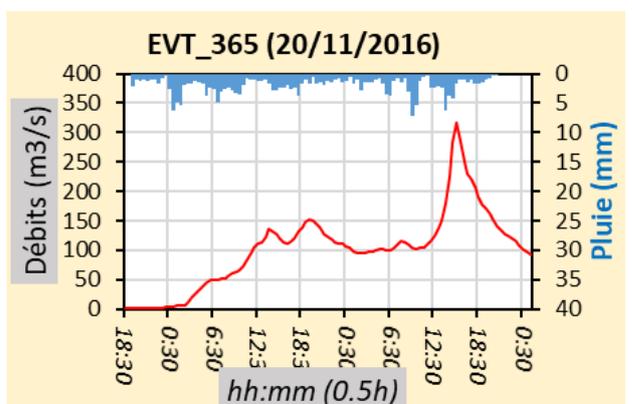
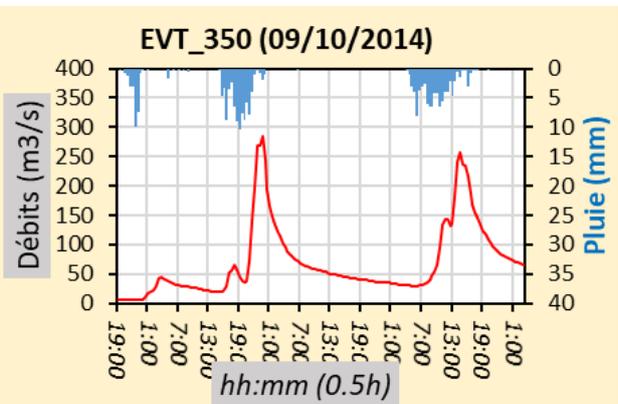
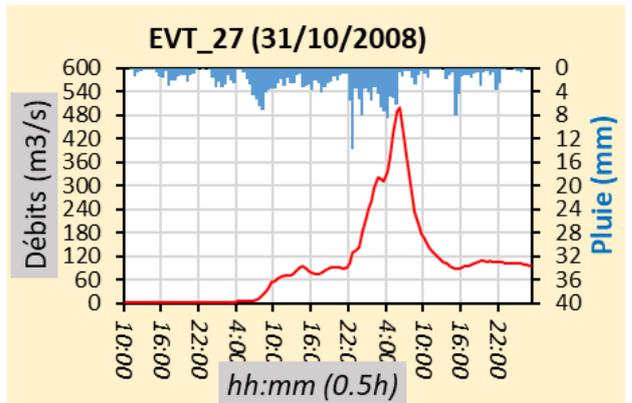
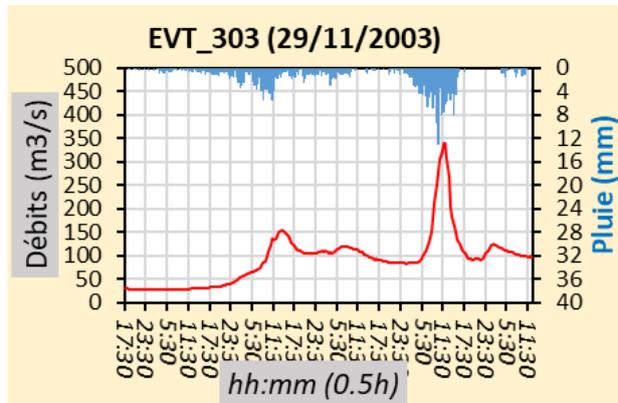
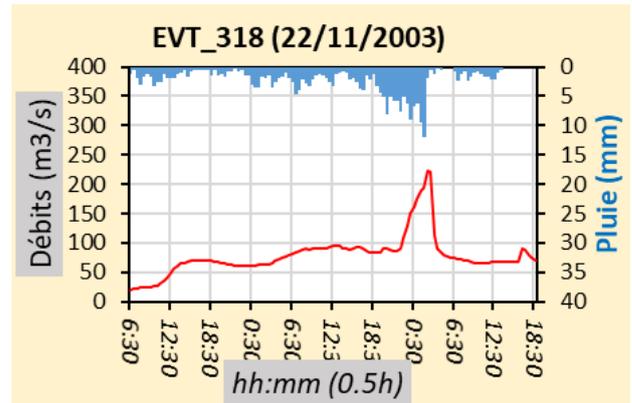
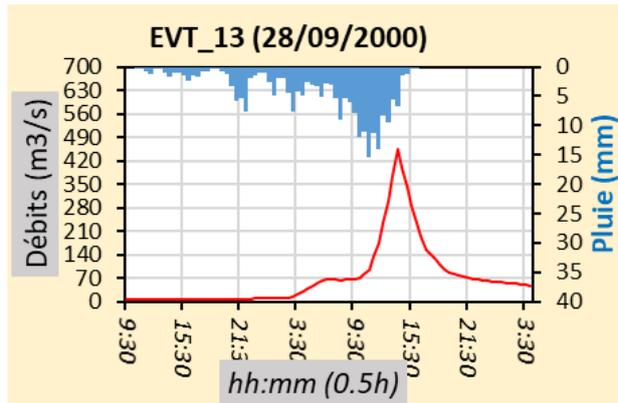
Annexe 2: Hydrogrammes des épisodes de la validation croisée de la période 1992-1999



■ Pluie

— Débit

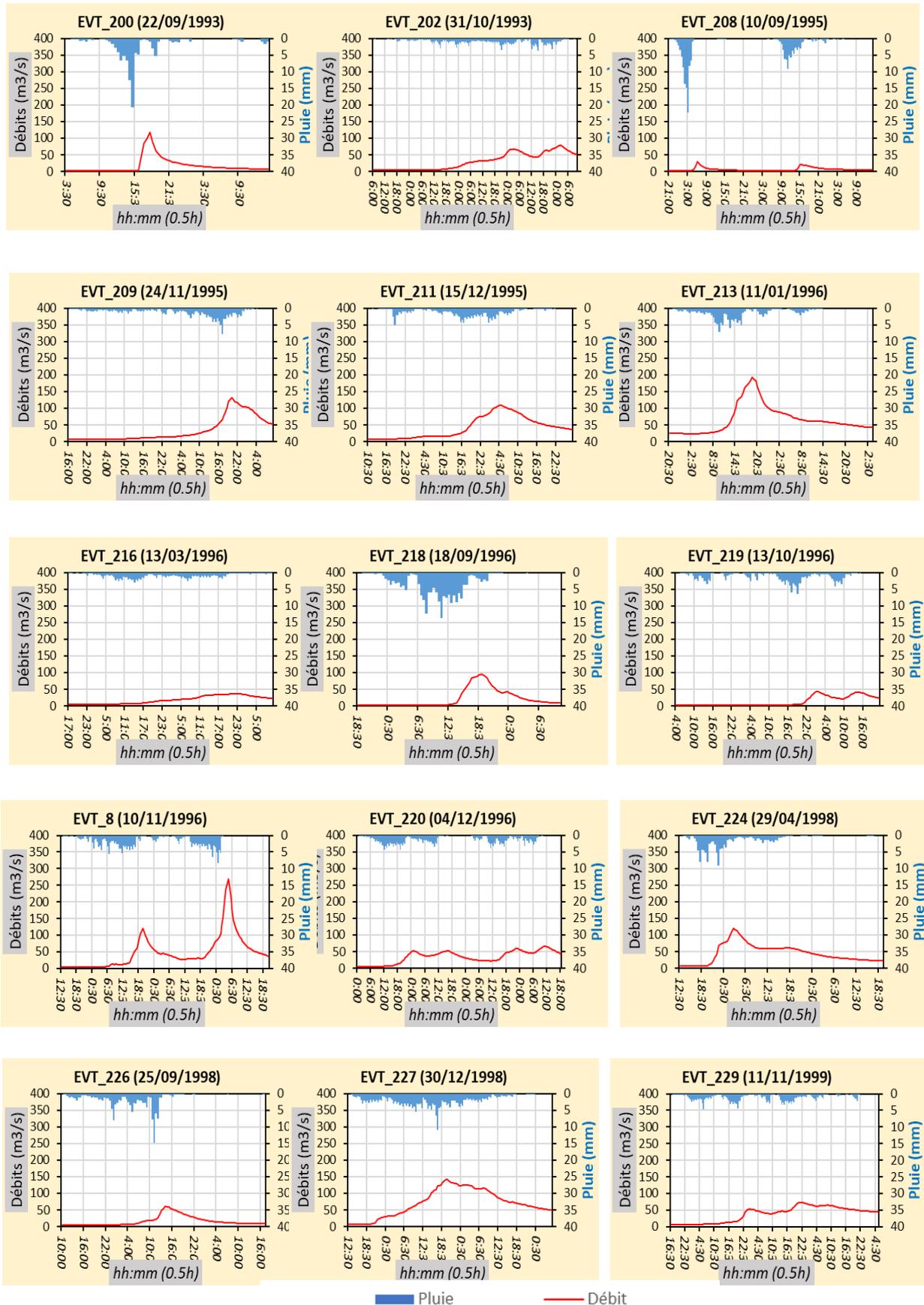
Annexe 3: Hydrogrammes des épisodes de la validation croisée de la période 2000-2019



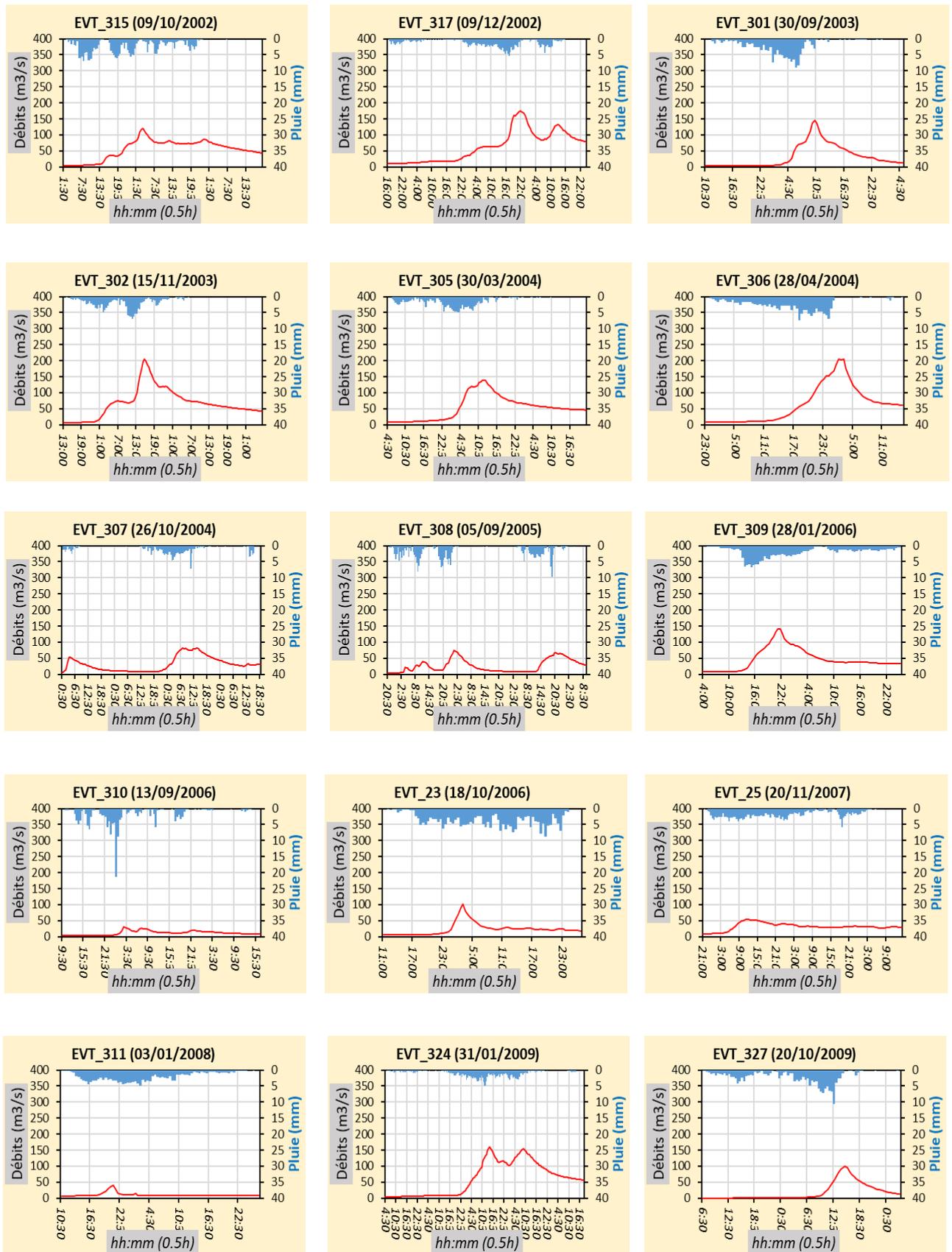
■ Pluie

— Débit

Annexe 4: Hydrogrammes des épisodes hors validation croisée de la période 1992-1999

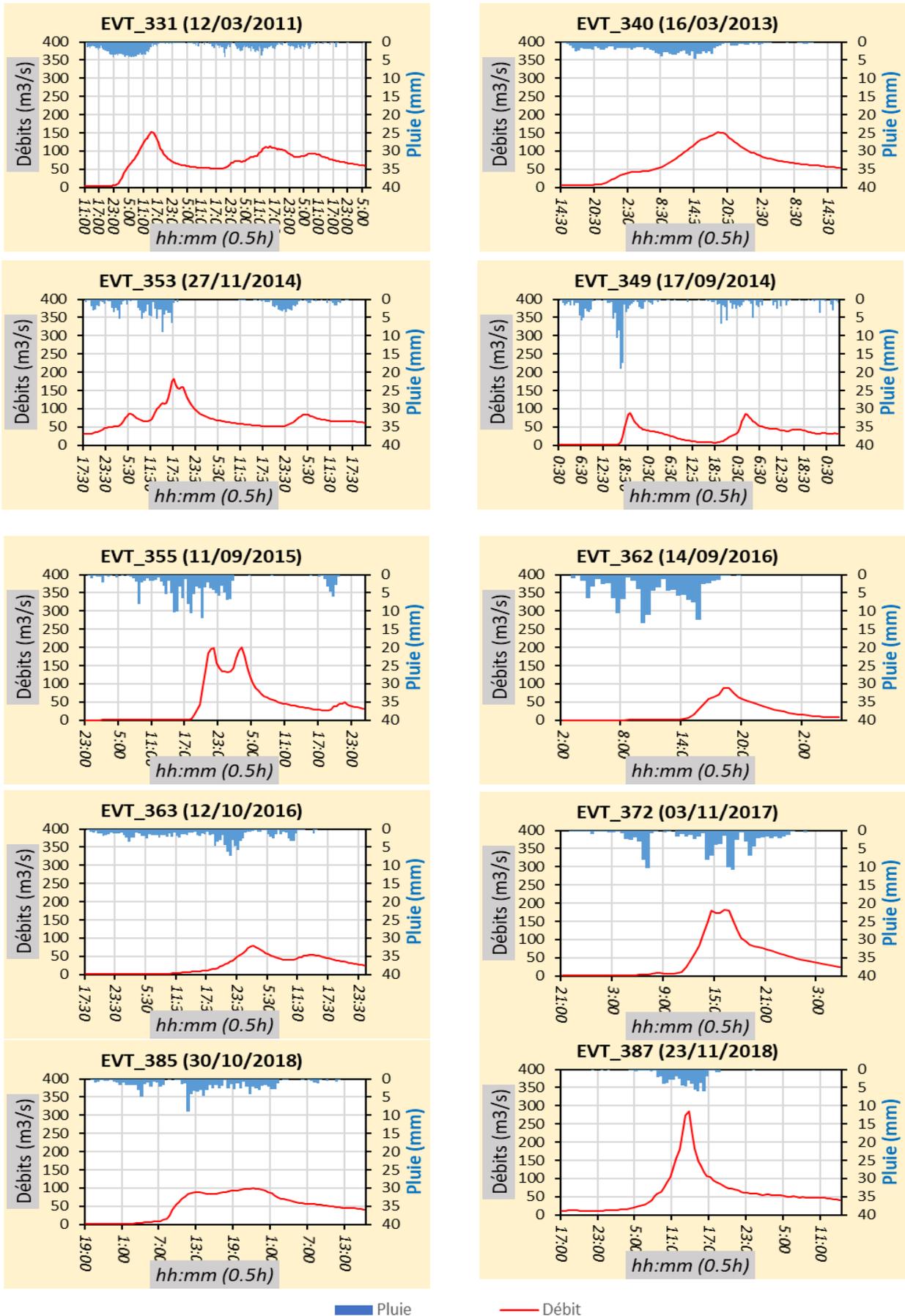


Annexe 5: Hydrogrammes des épisodes hors validation croisée de la période 2000-2009



■ Pluie — Débit

Annexe 6: Hydrogrammes des épisodes hors validation croisée de la période 2010-2019



Annexe 7: Les épisodes de Test (a) et d'Arrêt (b)

