31 décembre 2013

Georges-Marie Saulnier, William Castaings, Anne Johannet, Gérard Dreyfus

Tâche 03
Approche hybride
De la correction des erreurs à la sélection de variables

Sommaire

1. INTRODUCTION

Cette tâche est la moins aboutie du projet. Compte tenu de l'éloignement des communautés scientifiques engagées dans cette tâche et de premiers résultats décevants, il n'a guère été possible d'aller plus loin que l'échange scientifique, fructueux d'un point de vue exploratoire mais peu productif en résultats quantitatifs. Si ce projet a permis dans cette tâche des perspectives intéressantes pour les partenaires et de dessiner des pistes de futures collaborations, les résultats ne sont pas à la hauteur des espérances initiales.

Rappelons toutefois la problématique. La question scientifique posée était celle-ci :

- Comment **améliorer la justesse** (au sens du réalisme physique des modélisations hydrologiques utilisées et améliorées), **la précision** (au sens de la proximité entre les prévisions hydrologiques et les observations de terrain) **et la robustesse** (au sens de la fiabilité du modèle hydrologique sur l'ensemble des crues rapides de la base de données) pour des horizons temporels permettant d'améliorer la sécurité des personnes et des biens
- En **combinant les apports respectifs** et complémentarités entre **l'apprentissage statistique et la modélisation physique**.

Quelles sont les complémentarités et les possibilités de synergie ? Comment les exploiter pour le contexte des prévisions hydrologiques des crues rapides ?

Deux étapes étaient envisagées :

- Améliorer les prévisions du modèle hydrologique physique aux moyens de modèles utilisant l'apprentissage statistique (correction des erreurs).
- Fusionner les deux approches de modélisations en substituant certains modules du modèle physique, les plus discutables, par des modèles fondés sur l'apprentissage statistique (modélisation hybride).

La première étape ayant des résultats finalement décevants, la deuxième étape n'a pas pu être abordée.

2. CORRECTION DES ERREURS

Les motivations fondant la proposition initiale de corriger les erreurs d'un modèle hydrologique physique par un modèle d'apprentissage statistique reposaient sur le constat que n'importe quelle modélisation numérique possède un socle incompressible de sources d'incertitudes et d'erreurs. Les modèles sont imparfaits par nature, ils sont de plus contraints par des données incertaines et enfin les interactions entre sources d'incertitudes ne sont pas toujours analytiquement identifiables et représentables.

Les erreurs d'un modèle hydrologique physique sont ainsi non indépendantes et identiquement distribuées sur les différentes phases d'une crue. L'auto-corrélation des résidus de la prévision hydrologique (en général les chroniques de débits à une section donnée d'une rivière surveillée) est ainsi principalement déterminée par la structure du modèle hydrologique physique.

Mais si les erreurs d'un modèle physique ne sont ainsi pas « aléatoires » mais influencées par les insuffisances et approximations des hypothèses physiques le fondant, dans quelle mesure un apprentissage statistique ne parviendrait il pas à interpoler les « régularités » expliquant en partie les erreurs du modèle physique ? Si la réponse est positive l'application de ce modèle à apprentissage statistique en post-traitement des simulations du modèle physique permet elle de diminuer les erreurs de prévision du modèle physique ?

Une démarche en trois étapes a donc été entreprise :

- Calibration du modèle physique sur l'ensemble des épisodes de crues disponibles sur le projet.
- Calculs des erreurs commises par le modèle physique à chaque pas de temps de la simulation
- Apprentissage statistique (par réseau de neurones) des erreurs commises par le modèle
- Ajouter l'erreur calculée par le réseau de neurones aux prévisions hydrologiques du modèle physique

2.1 MODELE HYDROLOGIQUE PHYSIQUE

L'approche de modélisation utilisée dans cette étude est fondée sur l'approche TOPMODEL (TOPography based MODEL, Beven et Kirkby 1979 et Beven et al. 1995). C'est un cadre de modélisation particulièrement populaire qui a connu un vif succès et un développement continu depuis sa conception dans les années 80.

Ses principaux avantages sont les suivants :

- il offre un cadre de modélisation souple qui permet d'adapter "facilement" des modèles existants à des contextes particuliers
- il est à base physique et permet d'intégrer la connaissance hydrologique au fur et à mesure qu'elle se développe
- il est spatialisé au sens où il permet de prédire les différents flux hydrologiques (ruissellement, infiltration, percolation, évapotranspiration) en tous points d'un bassin
- il propose la définition d'indice de similarité hydrologique qui permet une résolution élégante et numériquement très efficace des calculs.

Les processus hydrologiques représentés dans cette modélisation à l'échelle du pixel MNT sont les suivants :

- ruissellement sur surfaces saturées: les écoulements préférentiels des premiers mètres de sol peuvent être localement et temporairement déjà saturés en eau. Les précipitations tombant sur ces zones ruissellent alors en forte proportion à la

surface du sol jusqu'à rejoindre le réseau hydrographique ("le sol pourrait absorber rapidement l'eau des précipitations mais il est est déjà saturé en eau").
- ruissellement hortonien : par dépassement des capacités d'infiltration des sols, les précipitations ruissellent dans une forte proportion à la surface du sol jusqu'à rejoindre le réseau hydrographique ("le sol n'est pas saturé en eau mais il ne peut absorber assez vite l'eau des précipitations").
- reprise évaporative : l'eau des premiers mètres du sol contribuent à la reprise évaporative en fonction du taux d'évapotranspiration potentielle et de l'eau contenu dans le sol.
- percolation profonde : une partie de l'eau du sol peut contribuer à l'alimentation de nappes profondes.
- exfiltration du sol : les écoulements latéraux dans les premiers mètres du sol peuvent exfiltrer dans le réseau hydrographique et contribuer également à la genèse des débits.
- transferts du ruissellement à la surface du sol et dans le réseau hydrographique.

Sans rentrer dans le détail, l'originalité des TOPMODELs réside dans la résolution de la dynamique spatio-temporelle des aires saturées basée sur la définition d'un indice de similarité hydrologique décrivant leur invariance spatiale.

Dans le cadre de ce projet et en préalable à cette tâche 3 (mais aussi de la tâche 5), ce modèle hydrologique a été analysé par analyse de sensibilité probabiliste. L'exploration stochastique de l'espace des paramètres a été effectuée par une approche de type Quasi-Monte Carlo (QMC) et la variabilité résultante a été analysée par analyse de la contribution à la moyenne et décomposition de la variance fonctionnelle. De manière similaire à l'analyse de sensibilité régionalisée très utilisée dans la communauté des hydrologues, un filtrage de Monte Carlo suivi d'un test statistique (i.e. test de Kolmogorov-Smirnov) permettant de déterminer les paramètres dont la valeur est influente sur la mesure de performance utilisée pour le calage a été effectué.

Par ailleurs, une version du modèle adjoint, élément central de la méthode d'assimilation variationelle de données, a été développée. Une procédure de calage des paramètres a été alors mise en œuvre à partir du gradient par rapport aux paramètres calculé par le modèle adjoint. Elle est basée sur un algorithme de descente ou l'optimisation est effectuée par une méthode de type quasi-newton avec contraintes inégalités (bornes sur les paramètres). Le calage sur 20 épisodes cévenols ayant touché le bassin versant des Gardons à Anduze a été ainsi effectué.

2.2 CORRECTION DES ERREURS DU MODELE PHYSIQUE PAR RESEAU DE NEURONES

L'apprentissage statistique des erreurs du modèle physique du réseau de neurones a été effectué par l'Ecole des Mines d'Alès après fourniture par le laboratoire EDYTEM des fichiers de simulations du modèle hydrologique physique. Des comparaisons ont été alors faites entre :

- les débits calculés par le modèle physique auxquels ont été ajoutés les erreurs simulées par le réseau de neurones
- et la correction triviale du modèle physique consistant à rajouter aux débits simulés par le modèle physique les erreurs observées aux pas de temps précédent.

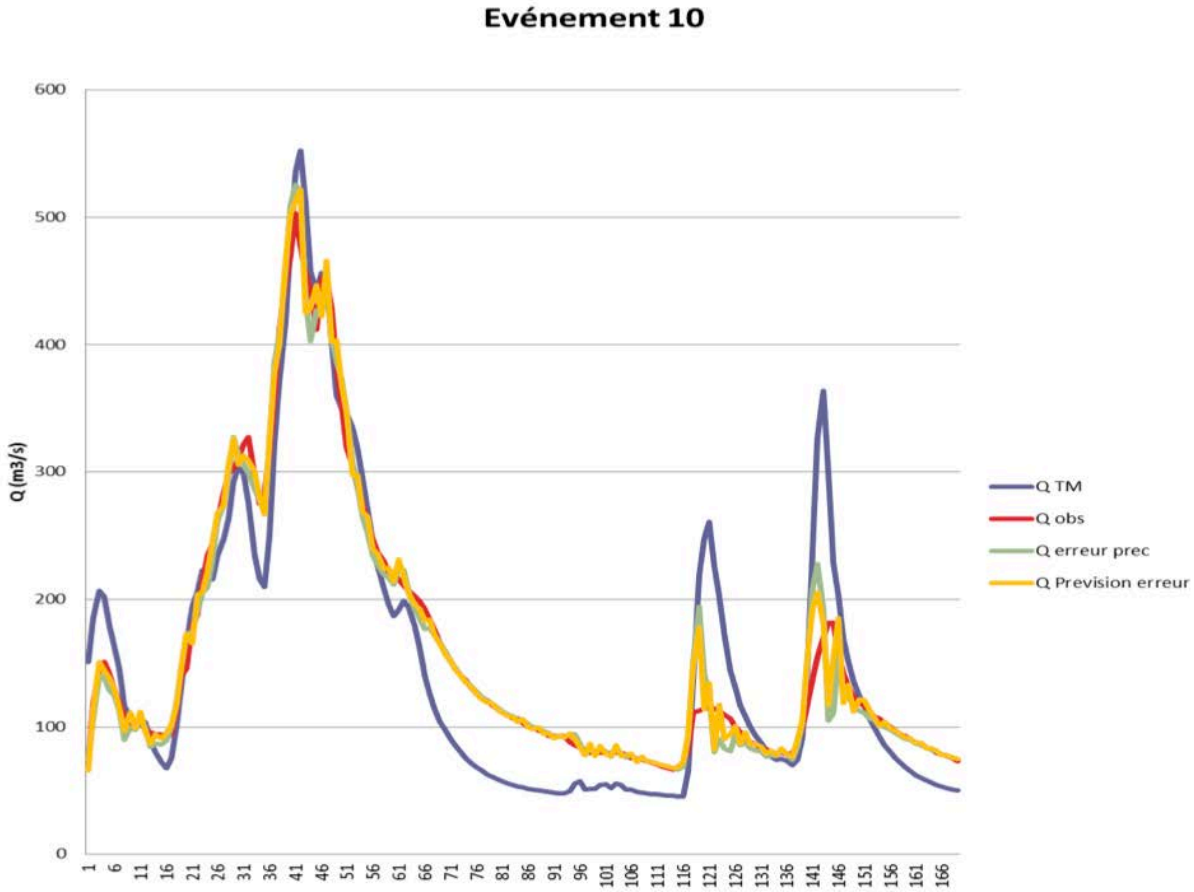

Figure 1 : comparaison des débits corrigés par post-application d'un réseau de neurones et débits corrigés trivialement (application de l'erreur observée précédemment).

Il apparaît que la correction triviale s'avère tout aussi efficace que celle utilisant le réseau de neurones. C'est encore plus vrai pour les échéances de prévisions courtes où la correction triviale est meilleure, et moins vrai pour les échéances de prévisions plus longues où la correction triviale devient abbérante comme cela est normal.

Des résultats comparables sont obtenus si le réseau est alimenté avec davantage de variables comme par exemple le contenu en eau moyen du bassin tel que simulé par le modèle physique. Autrement dit, l'ajout d'informations physiques sur le fonctionnement interne du bassin n'apporte que marginalement au réseau de neurones. Nous espérions ce faisant permettre au réseau de neurones de pouvoir tirer parti de l'information sur, par exemple, la phase de la crue correspondant au pas de temps de simulation.

Le résultat principal de cette étape est que si l'approche reste fondée scientifiquement (ceci sera expliqué par la suite), nous n'avons pu encore déterminer

la meilleure façon d'extraire les régularités physiques dans les erreurs du modèle physique.

3. MODELISATION SEMI-PHYSIQUE

3.1 CONCEPT D'INFORMATION MUTELLE MOYENNE

Suite à ce constat décevant, nous avons décidé d'approfondir l'amont de cette question de modélisation hybride. Notamment : quelles sont les variables du modèle physique les plus pertinentes pour réfléchir à une modélisation hybride.

Pour ce faire nous avons choisi de ne plus travailler classiquement sur l'auto-corrélation des erreurs pour identifier les variables explicatives les plus pertinentes mais d'utiliser le concept d'Information Mutuelle Moyenne (AMI), concept tiré de la théorie de l'information. L'AMI est une quantité mesurant la dépendance statistique de deux variables, même si cette dépendance est non linéaire. Ainsi la corrélation que nous utilisions au début du projet peut être vue comme un cas particulier de l'AMI quand la dépendance statistique est linéaire.

La justification de ce changement de concept pour l'étude préalable d'identification des variables les plus pertinentes pour initier une modélisation hybride est venue du constat d'une incohérence. En effet quand nous étudions l'auto-corrélation des erreurs du modèle physique (dans la figure suivante sur deux bassins cévenols : le bassin du Gardon à Anduze et le bassin de l'Ardèche à Vogue), nous constatons une auto-corrélation significative et attendue des erreurs du modèle.

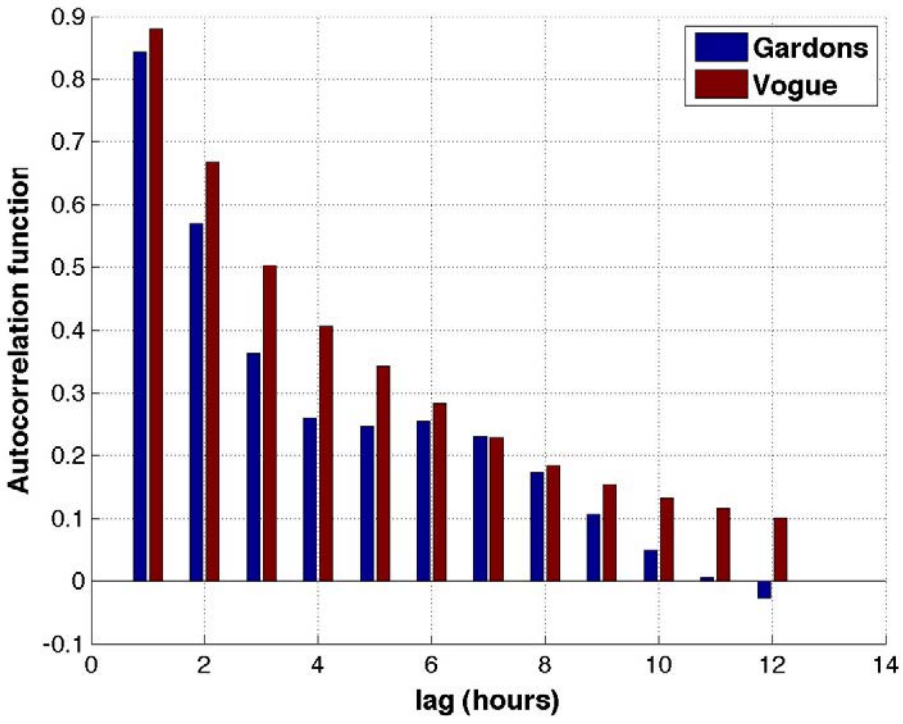

Figure 2 : auto-corrélation des erreurs du modèle hydrologique physique sur les bassins du Gardon à Anduze et de l'Ardèche à Vogue.

Mais quand on cherche à quantifier l'importance de la connaissance des précipitations dans l'auto-corrélation des erreurs du modèle hydrologique physique, nous observons bien (figure 3) dans le cas du bassin des Gardons à Anduze que l'auto corrélation des erreurs commises par le modèle hydrologique est supérieure à l'auto-corrélation des erreurs du modèle physique quand celui ci ne connaît pas les précipitations (ces erreurs sont estimées en comparant les prévisions du modèle hydrologique à l'horizon de 5h (temps de réponse du bassin) avec hypothèse de pluie nulle) :

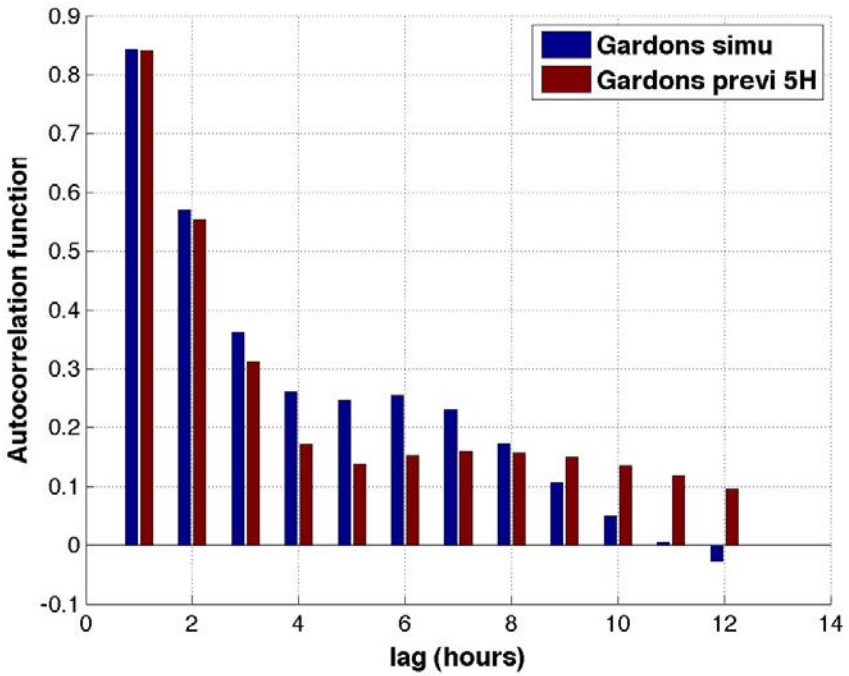

Figure 3 : Comparaison des fonctions d'autocorrélation moyennes de l'erreur modèle sur les épisodes disponibles en mode simulation et en mode prévision (5h). Bassin du Gardon à Anduze.

Mais dans le cas du bassin de l'Ardèche à Vogue (figure 4) nous constatons que l'auto-corrélation des erreurs est inférieure en connaissant la pluie qu'en ne la connaissant pas. Autrement dit, l'analyse de l'auto-corrélation semble indiquer que la dépendance des erreurs est renforcée par la méconnaissance des pluies à venir… C'est évidemment un résultat abérant.

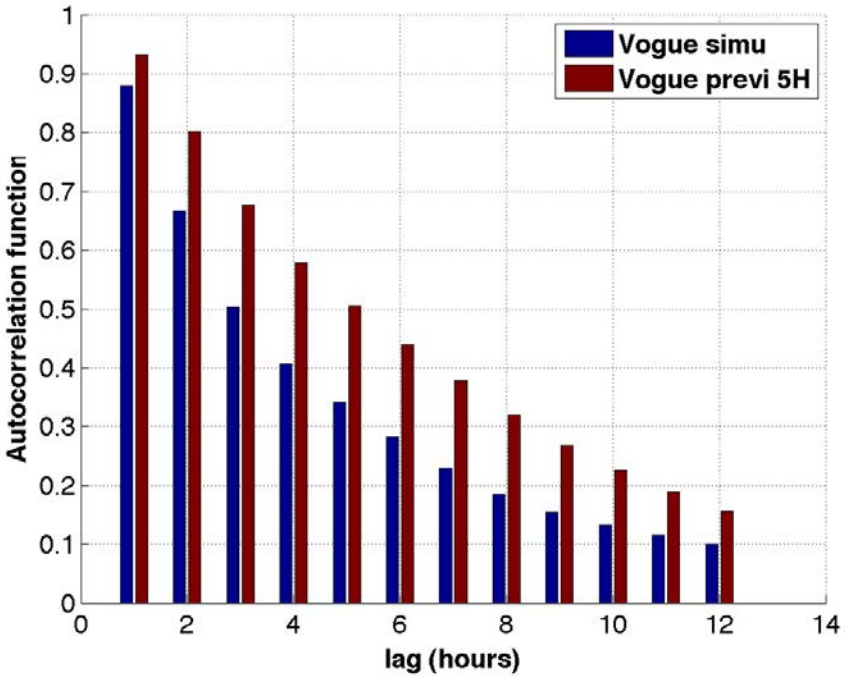

Figure 4 : Comparaison des fonctions d'autocorrélation moyennes de l'erreur modèle sur les épisodes disponibles en mode simulation et en mode prévision (5h). Bassin de l'Ardèche à Vogue.

Au contraire, **les mêmes analyses effectuées au moyen du concept d'AMI fournissent les deux figures suivantes, absolument cohérentes et interprétables physiquement**.

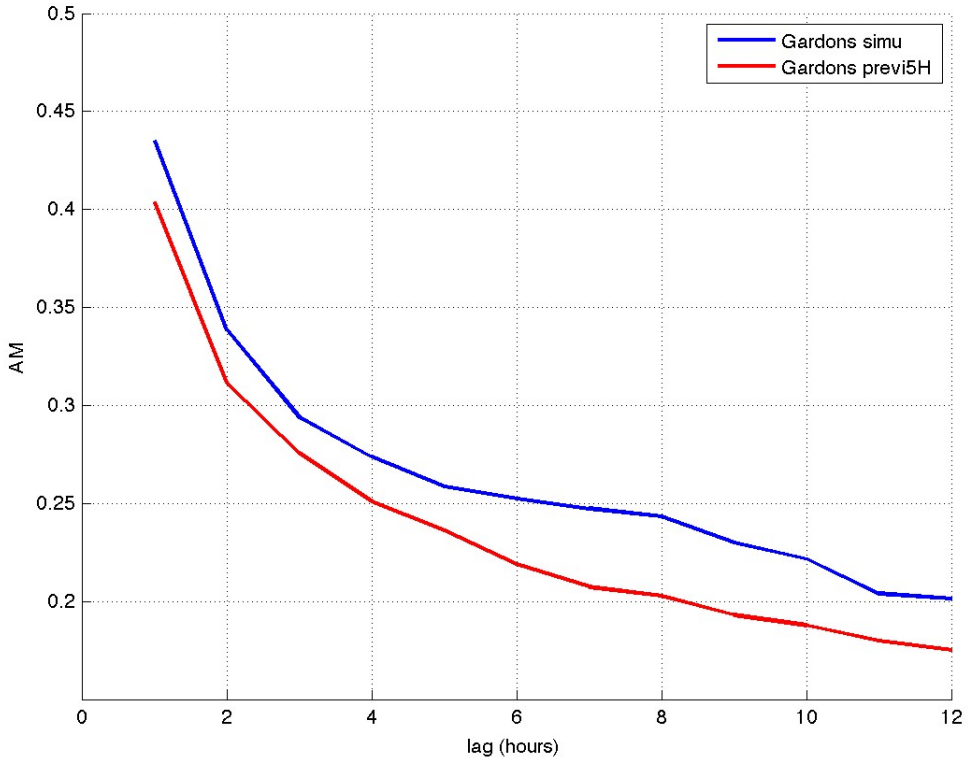

Figure 5 : Comparaison des fonctions d'autocorrélation AMI de l'erreur modèle sur les épisodes disponibles en mode simulation et en mode prévision (5h). Bassin du Gardon à Anduze

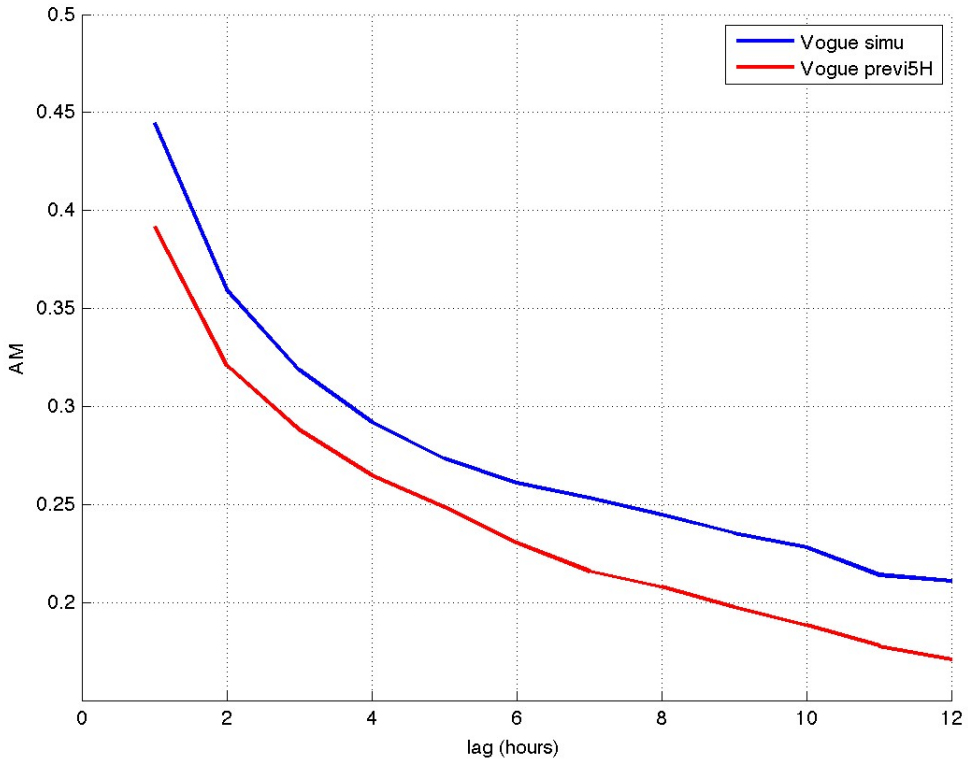

Figure 6 : Comparaison des fonctions d'autocorrélation AMI de l'erreur modèle sur les épisodes disponibles en mode simulation et en mode prévision (5h). Bassin de l'Ardèche à Vogue

Ayant introduit une nouvelle façon plus pertinente d'étudier la structure des erreurs du modèle physique, nous avons recherché les variables explicatives prépondérantes dans la genèse des erreurs du modèle hydrologique physique.

Les deux figures 7 et 8 suivantes illustrent la fonction AMI pour plusieurs pas de temps, ie la dépendance statistique entre d'une part les erreurs du modèle physique et les variables explicatives suivantes :

- l'erreur elle même (auto-corrélation non linéaire)
- la pluie (« rain » en vert dans les figures)
- le débit simulé (« qsim » en bleu dans les figures)
- le contenu en eau moyen du bassin (« dt » en rouge dans les figures)
- le pourcentage de ruissellement (« sat » en noir dans les figures)

On constate sur les deux figures que les variables ayant le plus de dépendance statistique avec les erreurs du modèle sont le contenu en eau moyen du bassin et le pourcentage de ruissellement.

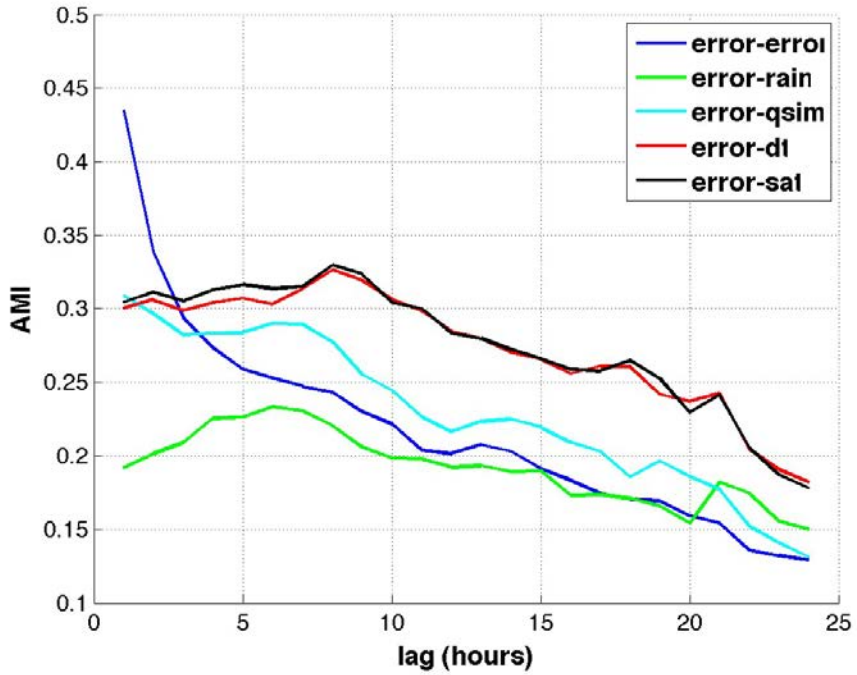

Figure 7 : Fonctions d'auto-corrélation non linaire (AMI) des erreurs et différentes variables explicatives. Bassin du Gardon à Anduze.

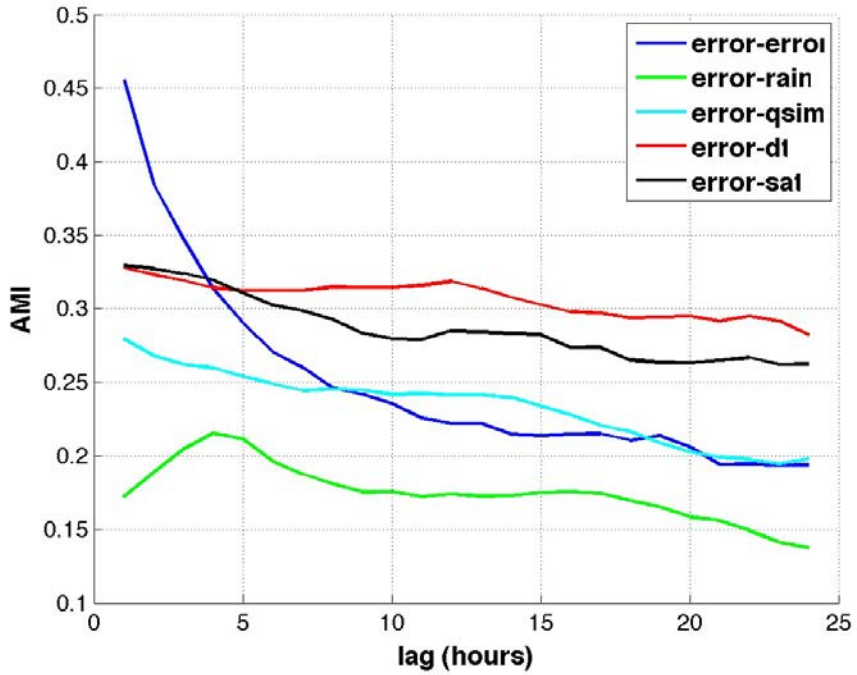

Figure 8 : Fonctions d'auto-corrélation non linaire (AMI) des erreurs et différentes variables explicatives. Bassin de l'Ardèche à Vogue

Ces analyses corroborent le sens physique : les variables les plus influantes sur les erreurs du modèle hydrologique sont les variables d'état du modèle physique.

Par voie de conséquence, on peut se demander quels modules du modèle hydrologique physique il conviendrait de questionner, voire d'améliorer au moyen d'une approche hybride dans la mesure où peu d'évidence permette d'améliorer ces modules représentant déjà l'état de l'Art. Ceci a été réalisé en effectuant le même type d'analyses sur les différentes composantes au débit que simulent le modèle physique :

- le débit d'exfiltration du sol (« exfil » en noir dans les figures)
- le débit de ruissellement de surface (« runoff » en rouge dans les figures)
- le débit total pour comparaison (« allflow » en vert dans les figures)
- la pluie pour comparaison (« rain » en bleu dans les figures).

Les figures 9 et 10 illustrent ces analyses.

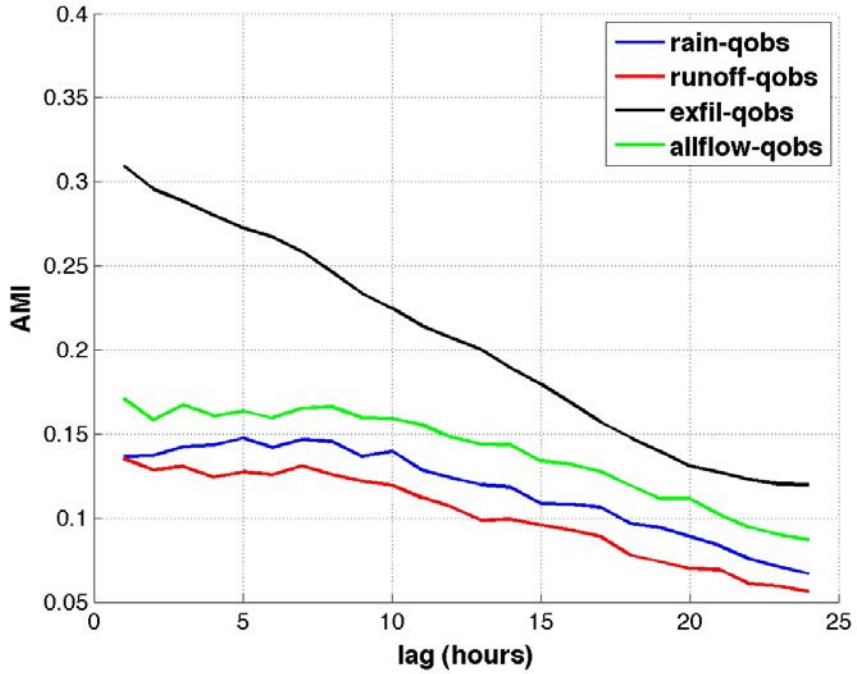

Figure 9 : Fonctions d'auto-corrélation non linaire (AMI) des erreurs et les différentes contributions au débit simulé. Bassin de l'Ardèche à Vogue

On constate sur cette figure 9 que le débit d'exfiltration du sol est la variable explicative expliquant le plus les erreurs du modèle physique. **Il est remarquable de constater que cette analyse corrobore une connaissance qualitative de la communauté des hydrologues** pratiquant ce type de modélisation physique, à savoir que la composante d'exfiltration du sol est la moins bien contrainte physiquement compte tenu du manque de données sur le comportement hydrodynamique des sols à l'échelle des versants et le moyens instrumentaux pour en mesurer les paramètres les plus caractéristiques.

Ce résultat est également constaté dans la figure 10 suivante illustrant la même analyse sur le bassin de l'Ardèche à Vogue.

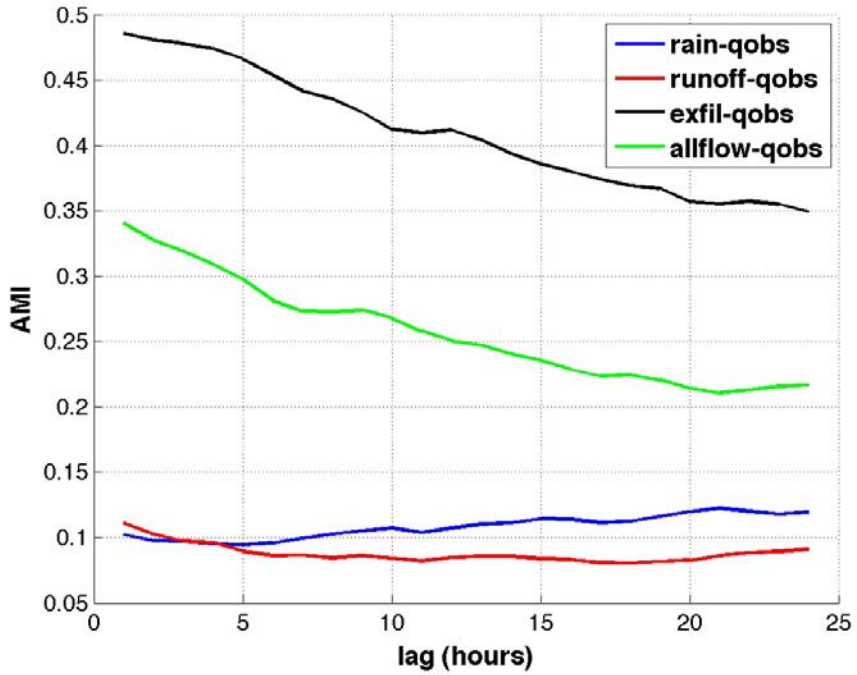

Figure 10 : Fonctions d'auto-corrélation non linaire (AMI) des erreurs et les différentes contributions au débit simulé. Bassin de l'Ardèche à Vogue

Ces analyses tendent à dessiner en perspective une étude hybride sur les modules du modèle physique représentant le fonctionnement hydrodynamique des sols à l'échelle des versants, davantage que sur les modules représentant les processus de ruissellement et de transferts hydraulique sur versant et en réseau.

4. CONCLUSIONS - PERSPECTIVES

Il semble donc à l'issue de cette tâche que nous avions sur-estimé la puissance des méthodologies pratiquées par les différents partenaires de ce projet (et de leur communauté scientifique respective) pour aborder la question de la modélisation hybride modèle physique – apprentissage statistique.

L'application d'une méthodologie somme toute raisonnable et logique a été décevante. Ceci nous a conduit à revisiter plus en amont la question de la dépendance statistique entre les sources d'erreurs du modèle physique et les différentes variables explicatives possibles.

Cette méthodologie nous a permis de corroborer de façon plus objective les modules du modèle physique qu'il conviendrait de revisiter. Il est donc un peu frustrant de constater à l'issue de cette tâche que nous pouvons prouver la pertinence d'une approche hybride sans avoir eu le temps d'aller plus loin. Mais le temps passé à définir et utiliser une méthodologie de définition plus objective et hiérarchisée des étapes d'une modélisation hybride est un résultat ouvrant des perspectives et transférables à d'autres modèles physiques.

http://blog.espci.fr/flash/