Document associé à la Charte Qualité de l'hydrométrie 2017 – Chapitre 6		
Liberté • Égalité • Fraternité RÉPUBLIQUE FRANÇAISE MINISTÈRE DE LA TRANSITION ÉCOLOGIQUE ET SOLIDAIRE	**Fiche 6.1 :** **Modèle linéaire et incertitude associée**	GROUPE DOPPLER HYDROMETRIE
Rédacteur : Perret C., famillecperret@orange.fr Vérificateurs : Belleville A. (EDF DTG), Lang M. (IRSTEA), Garçon R. (EDF DTG), Paquet E. (EDF DTG)		Mise à jour : Février 2018

1. Introduction

Dans son chapitre 6 Traitement des données, la charte qualité de l'hydrométrie suggère d'utiliser le modèle linéaire comme moyen de détection d'erreurs. La figure 6.5 de la Charte présente la comparaison d'une chronique de débits observés à un modèle élaboré par corrélation linéaire avec une station référence. Le modèle est muni des quantiles 10 et 90 qui représentent un intervalle de confiance à 80%. Cette notion n'est pas développée dans le corps de la charte et il parait nécessaire de la détailler à travers une fiche dédiée.
On propose tout d'abord de revenir plus en détails sur le modèle linéaire pour notamment mieux définir les conditions d'utilisation de celui-ci.
Les aspects théoriques sont illustrés à travers un exemple composé de deux séries de débits moyens mensuels de 30 années soit 360 occurrences. La station S_1 contrôle un bassin versant de 2170 km^2 et la station S_2 un bassin de 3580 km^2.

2. Modèle linéaire Y' = aX + b

Soit deux séries de données de débits X (station S_1) et Y (station S_2) qui comportent chacune N observations de débit :

$$X = \{X_1, X_2, \ldots\ldots\ldots\ldots, X_n\}$$
$$Y = \{Y_1, Y_2, \ldots\ldots\ldots\ldots, Y_n\}$$

On cherche à critiquer la série Y à partir de X. Un modèle linéaire peut être construit entre les deux séries de données. Il peut s'écrire de la manière suivante :

$$Y = a.X + b + \varepsilon \quad (1)$$

Si on pose :

$$Y' = a.X + b \quad (2)$$

On peut écrire :

$$\varepsilon = Y - Y' \quad (3)$$

ε représente l'ensemble des écarts entre les valeurs observées et les valeurs modélisées à partir de la série observée X. ε est couramment appelé ensemble des résidus du modèle linéaire.
Y' représente l'ensemble des valeurs modélisées de Y.

Détermination des coefficients a et b par la méthode des moindres carrés
La somme du carré des résidus S est une mesure qui permet d'évaluer l'efficacité de l'ajustement (Murray R. Spiegel, 1972) et « parmi toutes les courbes qui approchent un ensemble de données », la meilleure est celle qui minimise la somme du carré des résidus que l'on nommera S.

$$S = \Sigma\varepsilon_i^2 = (Y_1 - aX_1 - b)^2 + (Y_2 - aX_2 - b)^2 + \ldots\ldots\ldots\ldots\ldots + (Y_n - aX_n - b)^2 \quad (4)$$

S est minimale lorsque les dérivées partielles de S par rapport à a et b sont nulles. Elles s'expriment de la manière suivante :

$$\frac{\delta S}{\delta b} = 2[(Y_1 - b - aX_1) + (Y_2 - b - aX_2) + \ldots + (Y_n - b - aX_n)] = 0 \quad (5.1)$$

$$\frac{\delta S}{\delta a} = 2[(Y_1 - b - aX_1)X_1 + (Y_2 - b - aX_2)X_2 + \ldots + (Y_n - b - aX_n)X_n] = 0 \quad (5.2)$$

d'où :
(Dans les expressions qui suivent, on a supprimé l'indice i pour alléger la présentation)

$$\Sigma Y = a.\Sigma X + N.b \quad (6.1)$$

$$\Sigma XY = a.\Sigma X^2 + b.\Sigma X \quad (6.2)$$

Des équations (5.1) ou (6.1), on peut déduire une propriété importante de la régression linéaire : la somme des résidus est nulle et on peut conclure que le modèle linéaire est sans biais : $\Sigma\varepsilon_i = 0$.

Le système d'équation (6) se résout aisément et il vient :

$$a = \frac{N\Sigma XY - (\Sigma X)(\Sigma Y)}{N\Sigma X^2 - (\Sigma X)^2} \quad (7.1)$$

$$b = \frac{(\Sigma Y)(\Sigma X^2) - (\Sigma X)(\Sigma XY)}{N\Sigma X^2 - (\Sigma X)^2} \quad (7.2)$$

Coefficient de corrélation
Le coefficient de corrélation r mesure l'orthogonalité au sens géométrique des deux séries X et Y. Il constitue un critère pour évaluer le « lien » qui relie les deux variables. Il vaut :

$$r = \frac{\Sigma(X - \bar{X})(Y - \bar{Y})}{\sqrt{(\Sigma(X - \bar{X})^2)(\Sigma(Y - \bar{Y})^2)}} \quad (8)$$

La valeur est comprise entre -1 et 1. Si r vaut 0, les deux vecteurs X et Y sont orthogonaux donc il n'y a pas de lien entre les deux séries, donc pas de corrélation. On peut dire que les deux séries sont indépendantes. Si r vaut 1, les deux vecteurs X et Y sont colinéaires donc parallèles. Si r vaut –1, les deux vecteurs sont colinéaires de sens inverse.
On remarquera au passage que la notion de corrélation est indépendante de tout modèle, linéaire ou autre.

Coefficient de détermination
Pour déterminer la qualité d'un modèle quelconque, Nash Sutcliffe (1970) ont proposé un critère (9) :

$$Nash = 1 - \frac{\Sigma(Y - Y')^2}{\Sigma(Y - \bar{Y})^2} \quad (9)$$

La valeur de Nash est comprise entre -∞ et 1. On considère que le modèle est correct lorsque la valeur de Nash est supérieure à 0,8.

Dans le cas de la régression linéaire classique, on a :

$$r^2 = 1 - \frac{\Sigma(\varepsilon_i)^2}{\Sigma(Y - \bar{Y})^2} \quad (10)$$

Elle est comprise entre 0 et 1 et prend le nom de coefficient de détermination.

On remarque tout d'abord que les expressions (9) et (10) sont les mêmes et on peut retenir que pour qualifier le modèle linéaire, r^2 et Nash sont équivalents ce qui revient à conclure que pour un modèle linéaire dont les coefficients sont ajustés par la méthode des moindres carrés, la valeur de Nash est toujours comprise entre 0 et 1.
En pratique, on pourra généralement considérer que le modèle n'explique pas du tout les valeurs observées si le coefficient de détermination tend vers 0 et qu'au contraire, il est de bonne qualité si la valeur tend vers 1. La méthode n'est toutefois pas sans défaut. Si l'un des couples (Xi, Yi) a une très forte valeur et que les autres points sont très concentrés dans un domaine restreint, on obtiendra un bon coefficient de détermination sans que celui-ci puisse être considéré comme significatif d'une bonne corrélation.

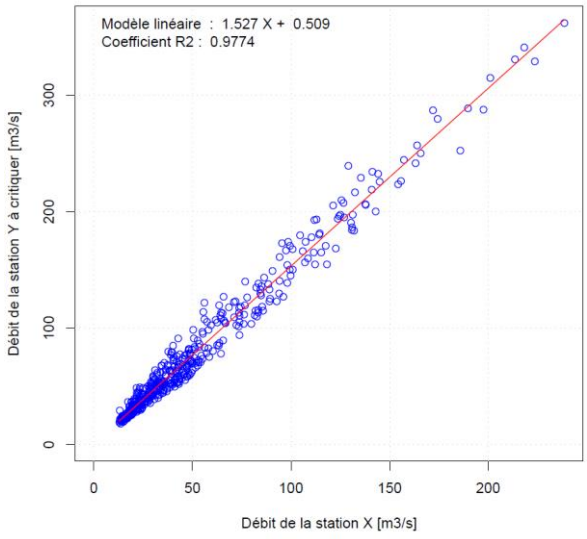

Figure 1 : Modèle linéaire entre la variable à expliquer (Données mensuelles de la station Y) et la variable d'explication (Données mensuelles de la station X)

On constate tout d'abord que le coefficient b (0,5 m^3/s) est faible si on le compare à la valeur des débits moyens observés à la station Y (Module égal à 78 m^3/s) soit moins de 1%. On remarque ensuite que le coefficient directeur a est comparable au rapport des bassins versants des deux stations puisqu'on obtient 1,527 pour a et 1,65 pour le rapport des bassins versants. Le coefficient de détermination est bon avec 0,977.

Conditions d'application

Il est toujours possible d'appliquer la méthode des moindres carrés à un nuage de points mais la probabilisation des résidus s'appuie sur trois conditions énoncées par Duband (1972) :

- Condition 1 : la variance des résidus doit être indépendante de la variable X.
- Condition 2 : les résidus successifs doivent être indépendants autrement dit, il ne doit pas y avoir d'auto corrélation des résidus.
- Condition 3 : La distribution des résidus doit suivre une loi Normale.

Duband juge la première condition « primordiale » et les deux autres « importantes mais pas essentielles ». On reviendra sur ces trois conditions un peu plus loin.

3. Modèle linéaire sans ordonnée à l'origine Y' = aX

De nombreux débats ont souvent eu lieu autour de la signification du coefficient b et notamment sur son sens physique. Aussi, certains hydrologues préfèrent-ils utiliser une relation linéaire sans coefficient b. Le modèle s'écrit de la manière suivante :

$$Y = a.X + \varepsilon \quad (11)$$

Si on pose :

$$Y' = a.X \quad (12)$$

On peut écrire :

$$\varepsilon = Y - Y' \quad (13)$$

ε représente l'écart entre la valeur observée et la valeur modélisée à partir de la série observée Q_1. Elle est couramment appelée résidu.

Détermination du coefficient a par la méthode des moindres carrés

La méthode des moindres carrés peut aussi être appliquée pour trouver le coefficient a. La somme du carré des résidus devient :

$$S = (Y_1 - aX_1)^2 + (Y_2 - aX_2)^2 + \ldots\ldots\ldots\ldots\ldots + (Y_n - aX_n)^2 \quad (14)$$

S est minimale lorsque la dérivée de S par rapport à a est nulle :

$$\frac{\delta S}{\delta a} = 2[(X_1 Y_1 - aX_1^2) + (X_2 Y_2 - aX_2^2) + \ldots + (X_n Y_n - aX_n^2)] = 0 \quad (15)$$

d'où :

$$a = \frac{(\Sigma XY)}{(\Sigma X)^2} \quad (16)$$

On doit remarquer que le modèle ainsi constitué n'est pas sans biais car la somme des résidus n'est pas nulle, ce qui ne va pas sans inconvénients si on doit modéliser le comportement de ces derniers.

Autre méthode pour déterminer a

Pour obtenir un modèle sans biais, il suffit que le vecteur Y' passe par le centre de gravité du nuage. Dans ce cas :

$$a = \frac{(\Sigma Y)}{(\Sigma X)} \quad (17)$$

Dans ce cas $\Sigma(\varepsilon_i)^2$ n'est plus minimum.

Coefficient de détermination

On a vu plus haut que le coefficient de détermination ne pouvait être déduit du coefficient de corrélation que si le modèle était linéaire (y' = ax+b) et calé par la méthode des moindres carrés. Dans le cas du modèle (y' = ax), la méthode n'est pas applicable mais le Nash peut être calculé.

Lorsque vous calez un modèle y = ax avec le logiciel Excel, ce dernier vous propose pourtant un coefficient de détermination. Cette présentation est trompeuse car si vous vous amusez à saisir un nuage de point manifestement non corrélé et que vous l'ajustez par une régression y = ax, Excel vous propose alors un coefficient de détermination négatif ! « Après enquête », il s'avère que le coefficient proposé par Excel se révèle être en fait un critère de Nash. On rappelle que celui-ci est équivalent au coefficient de détermination pour une régression ax + b calée par la méthode des moindres carrés.

Qualification du modèle Y' = aX

On retiendra que la qualité du modèle Y' = ax devra être évaluée à partir d'un critère de Nash (équation 9).

Conclusion sur ce point

La Charte Qualité de l'Hydrométrie a recommandé non sans raison, de veiller à la signification physique du coefficient b.

Si le débit de la station explicative X située en amont de la station expliquée Y, tend vers 0, le débit de la station Y tend vers b. On en déduit que si la contribution du bassin versant contrôlé par X devient nulle, celle du bassin versant comprise entre X et Y ne s'annule pas. Physiquement, cela signifie que les lois d'évolution des écoulements spécifiques (cf. paragraphe 6.2.5 de la Charte) des deux bassins versants sont différentes, ce qui est tout à fait possible. Il est vrai alors que dans ce cas, il est souhaitable d'utiliser une station témoin situé sur l'affluent principal du tronçon intermédiaire et de construire un modèle avec deux variables explicatives, surtout si le coefficient b représente une valeur importante relativement aux débits observés à la station Y.

4. Incertitudes du modèle - Intervalle de prédiction ou intervalle de confiance ?

Les coefficients a et b de la régression linéaire (7.1 et 7.2) permettent de déterminer la meilleure représentation de la régression linéaire pour les populations X et Y. Ces dernières ne représentent toutefois que des échantillons des débits des stations S_1 et S_2. Deux échantillons différents représentant une période différente auraient donné un autre résultat pour a et b. Il faut donc admettre que la droite de régression caractérisée par ses coefficients a et b, est entachée d'une incertitude liée aussi aux incertitudes sur a et sur b. Sur le plan pratique, on peut envisager l'incertitude du modèle de deux manières. On se contente ici de donner les formulations usuelles et pour plus détails, on se référera à Obled, Bois, Zin (2007).

L'intervalle de confiance correspond à l'intervalle dans lequel la droite de régression peut se trouver avec une probabilité donnée. Autrement dit, l'intervalle de confiance est une mesure de l'incertitude du modèle moyen.

L'intervalle de confiance correspond à la combinaison de l'incertitude sur la valeur moyenne des résidus et celle correspondant à la valeur de a. En retenant l'hypothèse de normalité des résidus, il peut être élargi au seuil de confiance de 95% en étant multiplié par 1,96.

$$IC = \pm 1{,}96 \frac{\sigma_\varepsilon}{\sqrt{N}} \sqrt{1 + \frac{(X_i - \bar{X})^2}{\sigma_x^2}} \qquad (18)$$

σ_ε est l'écart type des résidus

σ_x est l'écart type de la population X

L'intervalle de confiance diminue avec le nombre d'observations.

L'intervalle de prédiction correspond à l'intervalle dans lequel un nouveau point d'observation peut se situer avec une probabilité donnée. Autrement dit : quelle est la probabilité qu'un débit observé se trouve dans l'intervalle calculé ?
L'intervalle de prédiction est égal à l'écart type des résidus $\sigma\varepsilon$ du modèle et en retenant l'hypothèse de normalité de ces derniers, il peut être élargi au seuil de confiance de 95% en étant multiplié par 1,96 si l'on retient l'hypothèse de normalité des résidus.

$$IP = \pm 1{,}96\,\sigma_\varepsilon \sqrt{1 + \frac{1}{N} + \frac{(X_i - \bar{X})^2}{N\,\sigma_x^2}} \qquad (19)$$

σ_ε est l'écart type des résidus
σ_x est l'écart type de la population X

Les notions d'intervalles de prédiction et d'intervalle de confiance sont donc deux manières d'aborder l'incertitude du modèle linéaire. Avant d'utiliser l'une ou l'autre, il est donc important de bien cerner la grandeur que l'on souhaite qualifier.
Intervalle de prédiction et intervalle de confiance s'évaluent à partir d'une analyse du comportement des résidus du modèle.
IP est finalement peu différent de $1.96\sigma_\varepsilon$ comme on le montre figure 2.

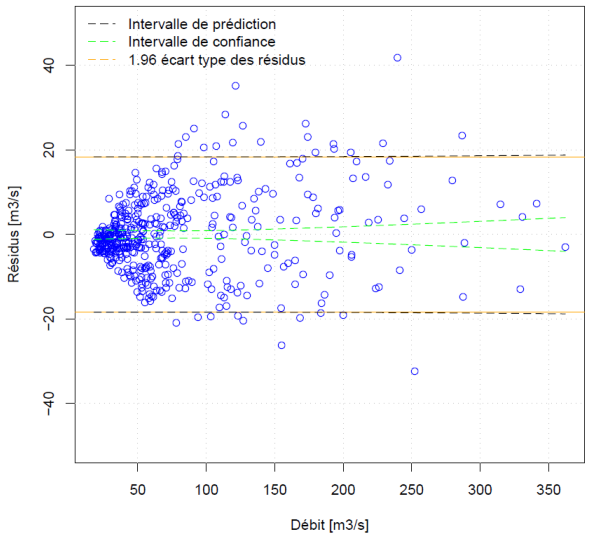

Figure 2 : Représentation des résidus du modèle linéaire en fonction du débit – Intervalle de confiance et Intervalle de prédiction

L'incertitude au seuil de confiance de 95% sur la valeur moyenne donnée par le modèle linéaire est faible comprise entre 0,84 m^3/s et 2 m^3/s en fonction du débit considéré.
Une valeur quelconque de débit observé a 95% de chance de présenter un écart de plus ou moins 18,3 m^3/s par rapport à la valeur moyenne du modèle linéaire.

La finalité de l'exercice de critique des données consiste au final à prendre la décision de ré-examiner ou non les paramètres de base qui ont servi à construire la série de données. Le modèle linéaire est un outil d'aide à la décision qui permet de répondre, en première approche, à la question : quelle est la probabilité qu'un débit observé se trouve dans l'intervalle proposé par le modèle ? On mesure à travers l'examen de la figure 2 que c'est l'utilisation de l'intervalle de prédiction qui permet d'y répondre.

Hétéroscédasticité des résidus

Une étude plus attentive laisse cependant planer un doute, car on peut suspecter qu'une des conditions énoncées plus haut n'est pas respectée : la valeur du résidu n'est pas indépendante du débit observé. On parle d'hétéroscédasticité des résidus. Pour s'en convaincre, on peut réaliser une analyse qui consiste à partager en deux la population des résidus après l'avoir classée en fonction des débits croissants. On calcule ensuite l'écart type de chacune des deux sous populations, ce qui permet de tester la réalité de l'écart constaté entre les deux. La figure 3 présente l'analyse graphiquement.

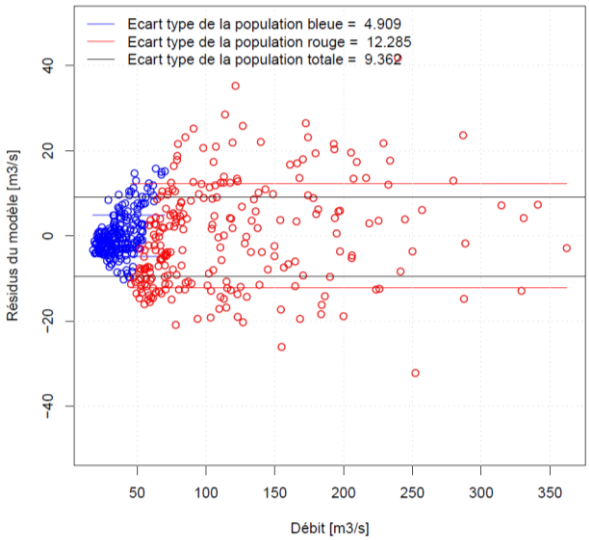

Figure 3 : Evolution des deux sous populations des résidus en fonction du débit – Valeurs des écarts types respectifs.

L'analyse montre clairement que si on attribue le même écart type à tous les débits pour évaluer l'incertitude du modèle, on commet une erreur : surestimation pour les débits faibles et sous estimation pour les débits forts. De plus, pour les faibles débits, on prend le risque de donner une valeur négative à la borne inférieure de l'intervalle de confiance.

Auto corrélation des résidus

Etablir s'il y a auto corrélation des résidus, consiste à calculer le coefficient de corrélation entre la série des résidus et elle-même décalée d'un pas de temps : 1 jour pour le pas de temps journalier, 1 mois pour le pas de temps mensuel.

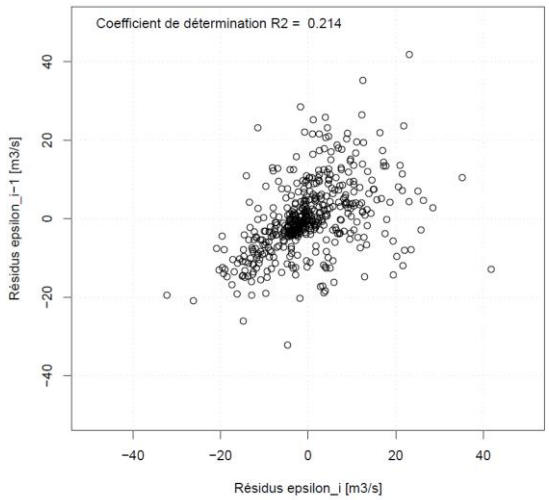

Figure 4 : Représentation des résidus du modèle en fonction des résidus du modèle décalés d'une occurrence

La figure 4 montre que la série des résidus du modèle présente une auto corrélation assez faible.

Normalité des résidus

Pour vérifier la normalité des résidus, on graphe l'histogramme de leur distribution et on la compare à la densité d'une distribution d'une loi Normale de moyenne égale à 0 et d'écart type égal à l'écart type des résidus $N(0, \sigma_\varepsilon)$.

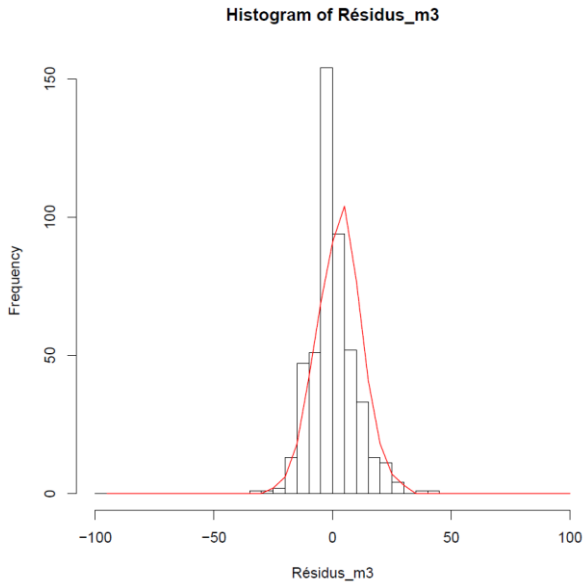

Figure 5 : Distribution des résidus du modèle linéaire et comparaison à la densité d'une distribution d'une loi Normale de moyenne égale à 0 et d'écart type égal à l'écart type des résidus.

La figure 5 montre que la distribution empirique des résidus du modèle est peu différente de celle d'une loi Normale.

Conclusions sur ce point

En première approche, on peut utiliser la formulation (19) pour calculer l'intervalle de prédiction au seuil de 70 % du modèle. Cette méthode reste cependant très imparfaite car en pratique, l'homoscédasticité des résidus du modèle n'est pas vérifiée car ceux ci évoluent avec le débit.

Il convient donc de rechercher une méthode d'évaluation de l'incertitude du modèle linéaire qui tienne compte de ce constat.

5. Formulation de l'incertitude du modèle linéaire

La méthode a été développée à partir de 2003 à EDF DTG par M. Hervé sous la houlette de R. Garçon et E. Paquet. Pour éviter l'écueil d'obtenir des débits négatifs lors du calcul de la borne basse de l'intervalle de prédiction, l'erreur du modèle est exprimée sous la forme d'un rapport de logarithmes :

$$\ln(Y) = \varepsilon + \ln(aX + B) \tag{20}$$

que l'on peut écrire également :

$$Y = Y' e^{\varepsilon} \tag{21}$$

L'expression (2) reste valable et on a donc :

$$\varepsilon = \ln\left(\frac{Y}{Y'}\right) = \ln(Y) - \ln(Y') \tag{22}$$

On retient comme hypothèses :

- le carré des erreurs du modèle linéaire est un bon estimateur de la variance des erreurs du modèle,
- et il peut être modéliser par une loi de type « puissance » :

$$\varepsilon^2 = b.Y^a \tag{23}$$

Pour trouver a et b, il suffit de linéariser l'expression de la manière suivante :

$$2\ln(\varepsilon) = \ln(b) + a.ln(Y) \tag{24}$$

a est calculé selon l'expression (7.1) en remplaçant X par ln(Y) et Y' par 2ln(ε).

Pour que le modèle soit exempt de biais, b vaut :

$$b = \frac{\overline{\varepsilon^2}}{\overline{Y^a}} \tag{25}$$

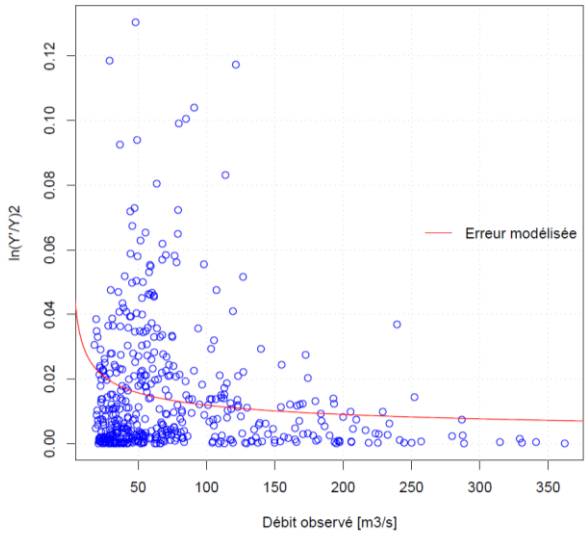

Figure 6 : Modélisation de ε^2 en fonction des données observées

Le calcul des bornes haute et basse avec intervalle de prédiction au seuil de confiance de 95%, se fait à l'aide des expressions suivantes qui sont déduites l'expression (21) :

$$Y'_{plus} = Y'.e^{1,96.\varepsilon'} \tag{26.1}$$

$$Y'_{moins} = Y'.e^{-1,96.\varepsilon'} \tag{26.2}$$

Avec :

$$\varepsilon' = \sqrt{b}.Y^{a/2} \tag{27}$$

6. Résultats

La figure 7 propose une représentation des écarts relatifs entre les débits modélisés et les débits observés. On a représenté :

- L'intervalle de prédiction issu du calcul de l'écart type des résidus du modèle linéaire au seuil de confiance de 95%,
- L'intervalle de prédiction issu du modèle d'incertitude proposé chapitre 5 (équations 26.1 et 26.2), seuil de confiance de 95%.

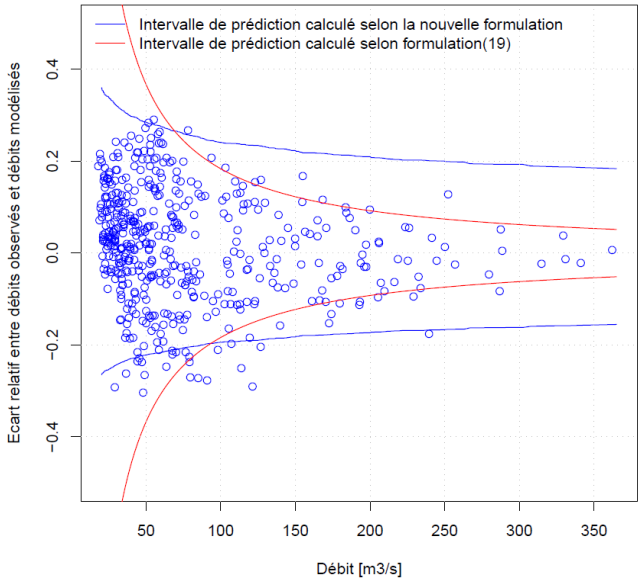

Figure 7 : Ecarts relatifs entre les débits modélisés et les débits observés munis des intervalles de prédiction au seuil de confiance de 95% calculés de deux façons différentes

On constate que le modèle proposé chapitre 5 permet de corriger l'effet d'hétéroscédasticité des résidus.

A titre d'illustration, on donne une représentation finale des débits observés pour l'année 1993 (signal noir) comparés aux intervalles de prédiction calculés selon les deux manières (Figure 8).

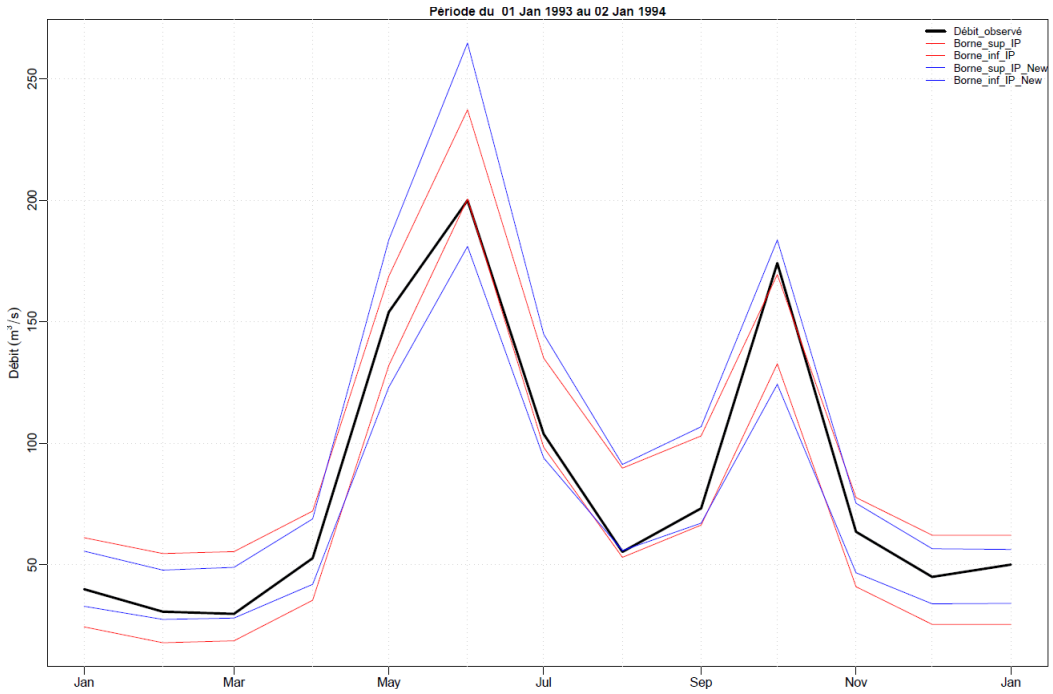

Figure 8 : Hydrogramme des débits observés comparés aux intervalles de prédiction calculés

Lorsqu'une valeur Qi (débit observé) ne rentre pas dans l'intervalle de prédiction, on ne doit pas nécessairement tirer la conclusion que cette valeur est erronée mais cela doit inciter le gestionnaire à ré-examiner les paramètres de base qui ont permis d'élaborer cette donnée : hauteur mesurée et courbe de tarage.

7. Applicabilité aux données journalières

Il est parfois recommandé, c'est le cas dans la Charte qualité de l'hydrométrie, de caler les coefficients a et b du modèle linéaire avec des données mensuelles et de les appliquer à des données journalières pour que les hypothèses de normalité des résidus et d'absence d'auto corrélation des résidus soient respectées. On propose ci-dessous l'analyse des résidus d'un modèle linéaire des séries journalières des mêmes stations dont les coefficients a et b ont été calés par la méthode des moindres carrés.

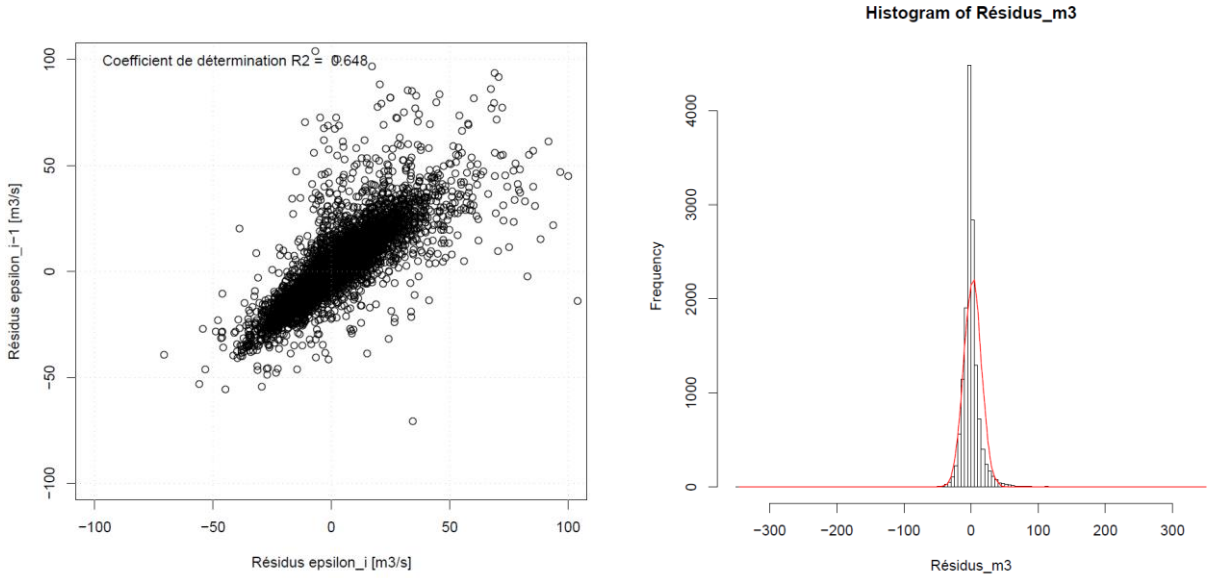

Figure 9 : Analyse de l'auto corrélation et de la normalité des résidus du modèle journalier

Force est de constater que sur cet exemple, les résidus sont mieux auto-corrélés en données journalières qu'en données mensuelles et que la normalité de leur distribution est correcte. Il arrive que ce soit l'inverse sur certaines séries. Lorsque l'auto corrélation des résidus est forte sur les données journalières, il est préférable d'utiliser les coefficients a et b calés sur les données mensuelles. On ne tirera pas de conclusions générales à partir de cet exemple mais on recommandera une analyse au cas par cas aux utilisateurs potentiels.

8. La corrélation double

On a vu (chapitre 3) qu'il était parfois très utile de construire un modèle avec plusieurs variables explicatives. Pour une démonstration exhaustive de la méthode, on se reportera aux documents cités comme références (Duband, 1972) (Obled, Bois, Zin, 2007). On propose ici de donner les éléments utiles pour une application avec deux variables explicatives, ce qui suffit dans la plupart des cas. On dispose d'une série de donnée Z issue d'une station S_3 qui permettrait de mieux expliquer les apports du bassin versant intermédiaire entre S_1 et S_2. La S_3 contrôle un bassin versant de 946 km^2. Le modèle s'écrit :

$$Y = a.X + bZ + c + \varepsilon \quad (28)$$

Si on pose :

$$Y' = a.X + bZ + c \quad (29)$$

On peut écrire :

$$\varepsilon = Y - Y' \quad (30)$$

Les coefficients a, b et c sont déterminés par la méthode des moindres carrés qui minimise la somme du carré des résidus. On ne reprendra pas ici la démonstration pour se contenter de présenter les expressions finales :

$$a = \frac{(r_1 - r_2 r_3)}{(1 - r_3^2)} \frac{\sigma_y}{\sigma_x} \quad (31)$$

$$b = \frac{(r_2 - r_1 r_3)}{(1 - r_3^2)} \frac{\sigma_y}{\sigma_z} \quad (32)$$

$$c = \bar{Y} - a\,\bar{X} - b\,\bar{Z} \quad (33)$$

où :

r_1, r_2 et r_3 représente respectivement les coefficients de corrélation (cf. expression 8) entre Y et X, Y et Z et X et Z.

$\bar{Y}, \bar{X}, \bar{Z}$ sont les moyennes des variables X, Y et Z.

L'application numérique à notre exemple de trois stations nous donne :

$$Y' = 0{,}864X + 1{,}52Z + 3{,}33$$

Le coefficient de détermination r^2 calculé selon l'expression (10) vaut : 0,9942.
On constate que l'apport de Z comme variable explicative permet d'améliorer la qualité du modèle.
L'évaluation de l'incertitude du modèle peut être conduite comme pour la régression simple (chapitres 4 et 5). On présente figure 10 le résultat de la méthode appliquée aux stations Y, X et Z.

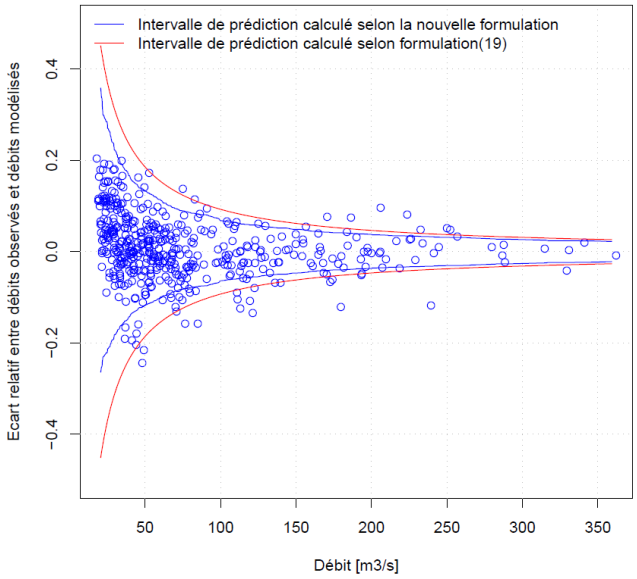

Figure 10 : Ecarts relatifs entre les débits modélisés et les débits observés munis des intervalles de prédiction au seuil de confiance de 95% calculés de deux façons différentes pour le modèle linéaire double

Si on compare les figures 7 et 10, on constate que le modèle linéaire double (Figure 10) permet de réduire l'intervalle de prédiction par rapport au modèle linéaire simple (Figure 7), ce qui augmente la pertinence de l'analyse lors de la critique de données. Inversement, il est difficile de trouver une pertinence physique aux coefficients a et b calculés, ce qui doit nous inciter à trouver une autre solution pour la combinaison des variables.

9. Corrélation sur la somme des variables explicatives

On reprend les variables S1, S2 et S3 du paragraphe 8 et on va considérer que la variable explicative du modèle est la somme X+Y. On revient donc à un modèle de corrélation simple et on déroule la méthodologie exposée au paragraphe 2. On obtient :

$$Y' = 1{,}067*(X + Z) + 2{,}22 \qquad (34)$$

avec un coefficient de détermination de 0,9927. La qualité de la corrélation est un peu moins bonne que pour la corrélation double mais reste supérieure à la corrélation calculée avec une seule variable explicative.
Le coefficient directeur du modèle linéaire a une signification physique acceptable compte tenu de la taille des bassins versants des trois stations. L'évaluation de l'incertitude du modèle est conduite de la même manière selon la méthodologie exposée au paragraphe 6. On obtient la figure 11 :

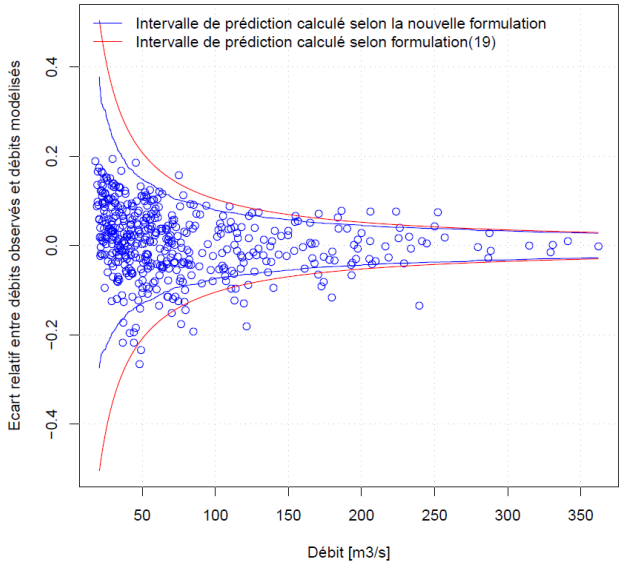

Figure 11 : Ecarts relatifs entre les débits modélisés et les débits observés munis des intervalles de prédiction au deuil de confiance de 95% calculés de deux façons différentes pour le modèle linéaire simple avec comme variable explicative (X + Z)

En examinant les figures 10 et 11, on peut voir que les intervalles de prédiction sont équivalents. Dans ce cas, on devrait préférer cette construction simple qui somme les variables d'entrée à la corrélation double.

10. Références

Duband D. (1973) Hydrologie approfondie Statistique appliquée – ENSHG INPG

Obled C. Bois P. Zin (2007) I. Introduction au traitement de données en hydrologie – ENSHMG INPG https://hydrologie.org/BIB/manuels/Bois-Obled-Zin.pdf

Murray R. Spiegel (1972) Théorie et application de la statistique –Série Schaum

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), 282–290