Combinaison bayésienne multi-modèle et Cartographie de consensus

Marine Riffard

Combinaison bayésienne multi-modèle et Cartographie de consensus

Marine Riffard (Irstea – Antony).

La cartographie finale résulte d'une combinaison de trois modèles de prédétermination pour les débits d'étiage et les modules, et des estimations de la méthode développée à Antony pour les débits de crue.
Cette façon de faire permet à chaque modèle de s'exprimer, au prorata de son efficience. En stratifiant les bassins versants par catégories de surface (découpage qui est apparu pertinent), on s'aperçoit que les modèles ne sont pas toujours aussi performants selon les classes, mais que chacun peut apporter une information qui améliore les résultats d'un multi-modèle lorsqu'elle est prise en compte. Nous avons donc choisi d'opter pour une combinaison multi-modèle bayésienne plutôt que de faire un choix de modèle unique basé sur un critère de performance défini arbitrairement. Ce type de combinaison, outre le fait qu'il permet d'ajouter autant de nouvelles variables que l'on souhaite, donne également un accès direct à l'incertitude associée à l'estimation du multi-modèle.

1. Principe

- ***La combinaison bayésienne multi-modèle***

La statistique bayésienne est la statistique des probabilités conditionnelles. Ainsi, on exprime une variable d'intérêt particulière inconnue (ex. un débit d'étiage) conditionnellement à d'autres variables que l'on connait (ex. des estimations issues de différents modèles de prédétermination) et en utilisant le théorème de Bayes.

L'objectif est d'affiner la distribution de probabilités de la variable cible, au fur et à mesure que celle-ci est expliquée par d'autres variables. A chaque nouvelle mise à jour, la précision associée à la nouvelle estimation de la variable cible est augmentée et sa distribution modifiée.

Cette combinaison multi-modèle se fait en plusieurs étapes :

- Définition d'un *prior* : ajustement d'un modèle probabiliste sensé représenter la distribution de probabilité de la variable d'intérêt. Ce modèle est ajusté sur la base de la connaissance *a priori* qu'on a de la variable cible;

- Application du théorème de Bayes : on exprime la fonction de probabilité de la variable d'intérêt en fonction d'une variable explicative;

- Mise à jour de la nouvelle distribution *a posteriori* par le calcul des nouveaux paramètres de la loi de distribution de la variable d'intérêt.

Ces trois étapes sont répétées autant de fois que l'on a de variables explicatives disponibles. Dans notre cas, nous disposons pour le débit d'étiage et le module, de trois modèles de prédétermination ayant fourni des estimations.

- ***Les données***

  - ***Echantillon de référence***

L'échantillon de référence (632 bassins versants) n'est plus stratifié car on travaille sur les lames d'eau produite sur un maillage fin, et on cumule les productions au fur et à mesure que les surfaces

augmentent selon un plan de drainage. On évite ainsi les discontinuités dans les écoulements et on s'affranchie du biais lié à la surface.

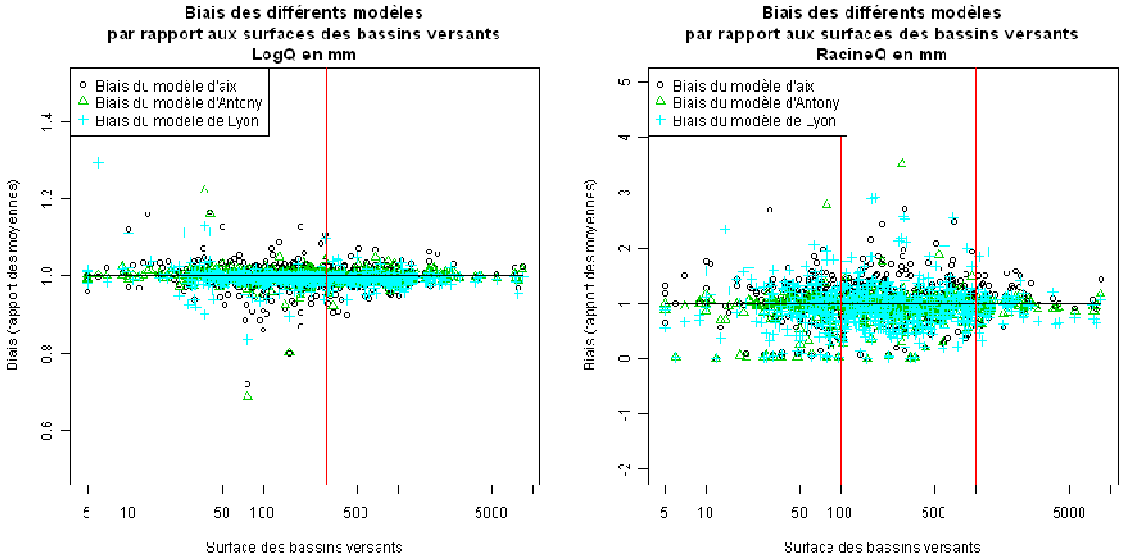


Figure 1 : Biais des modèles selon la variable d'intérêt (QA et QMNA5)

- ***Variables***

Nous utilisons comme variables les estimations des trois méthodes de prédétermination des débits développées par les trois équipes Cemagref/Irstea dans le cadre de la convention. Suite à une analyse réalisée au préalable (voir le rapport complet sur la combinaison multi-modèle), ce sont les racines des débits exprimés en mm/mois pour les QMNA5 et les Ln des débits en mm/an pour les QA qui seront les variables de calage du multi-modèle.

2. Application aux débits d'étiage et aux débits moyens

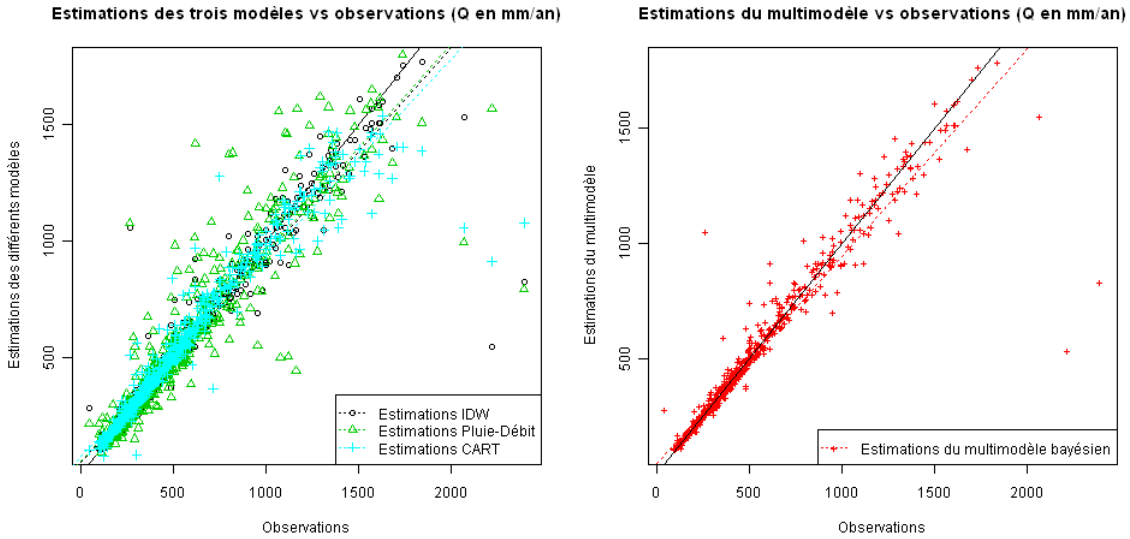
Nous ne présentons pas ici les détails des calculs et des différentes étapes. Nous présentons uniquement les résultats en termes de coefficients de détermination r^2 du multi-modèle sur l'échantillon de référence.

- ***Résultats sur le débit d'étiage (Q_{mna5})***

Tableau 1 : Résultats du multi-modèle sur le Q_{MNA5} – Valeurs des coefficients de détermination r^2

	Racine Q	Q
Antony	0.91	0.94
Aix	0.76	0.72
Lyon	0.62	0.53
Multi-Modèle	0.91	0.94

Dans tous les cas et sur l'échantillon stratifié ou global, le multi-modèle apparait au moins aussi performant que le meilleur des trois modèles de prédétermination (cf. Tableau 1). L'intérêt de ce type de combinaison qui laisse tous les modèles de prédétermination s'exprimer est assez visible. Et l'intérêt d'avoir plusieurs méthodes de prédétermination est également assez clair. Chaque modèle initial apporte une information nouvelle et permet d'affiner l'estimation finale.



La

Figure 3 présente les résultats sous forme graphique sur la variable d'intérêt finale (les Q_{mna5} exprimés en mm/mois). Le multi-modèle n'introduit pas de biais sur les estimations par rapport aux modèles initiaux et propose un biais équivalent à celui du meilleur modèle initial. Elle permet en outre de recentrer les estimations autour de la droite $Q_{observé}=Q_{estimé}$.

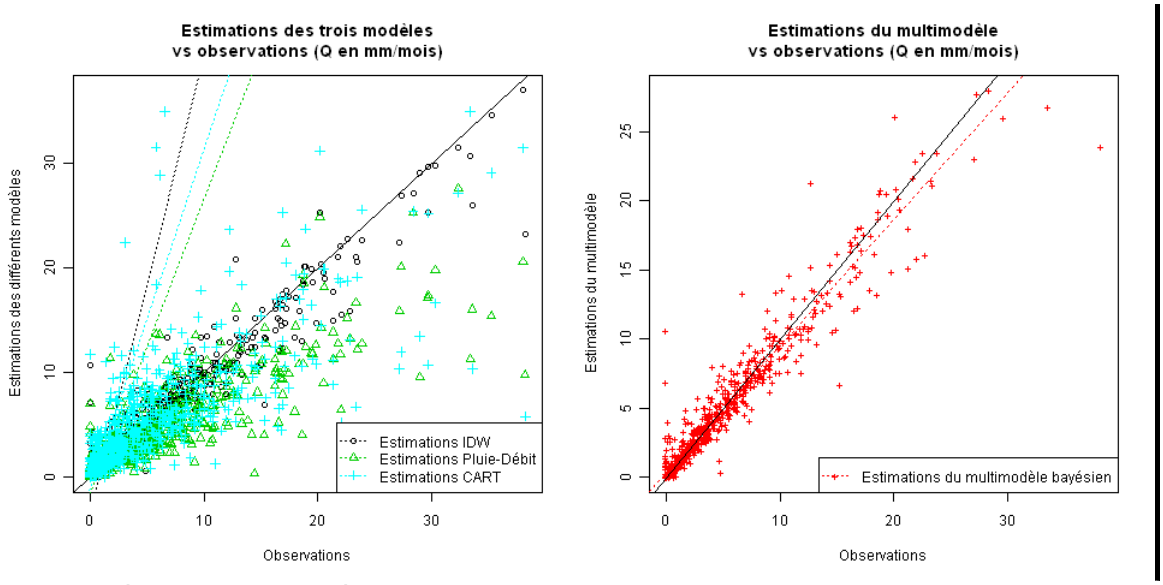


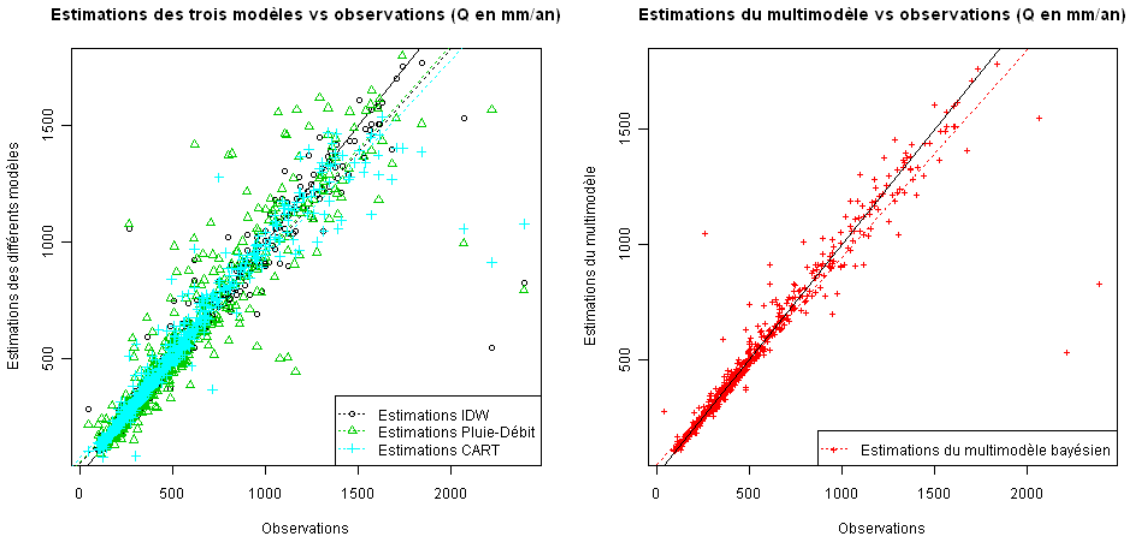
Figure 2 : Résultats du multi-modèle sur le Q_{mna5}

- ***Résultats sur le module (Q_a)***

Tableau 2 : Résultats du multi-modèle sur le Module – Valeurs des coefficients de détermination r^2

	Log des Q	Q
Antony	0.958	0.912
Aix	0.872	0.851
Lyon	0.405	0.907
Multi-Modèle	0.958	0.913

Dans tous les cas le multi-modèle apparait au moins aussi performant que le meilleur des trois modèles de prédétermination (cf. Tableau 2).



La

Figure 3 présente les résultats sous forme graphique sur la variable d'intérêt finale (les modules exprimés en mm/an). Le multi-modèle n'introduit pas de biais sur les estimations par rapport aux modèles initiaux et permet de recentrer les estimations autour de la droite $Q_{observé}=Q_{estimé}$.

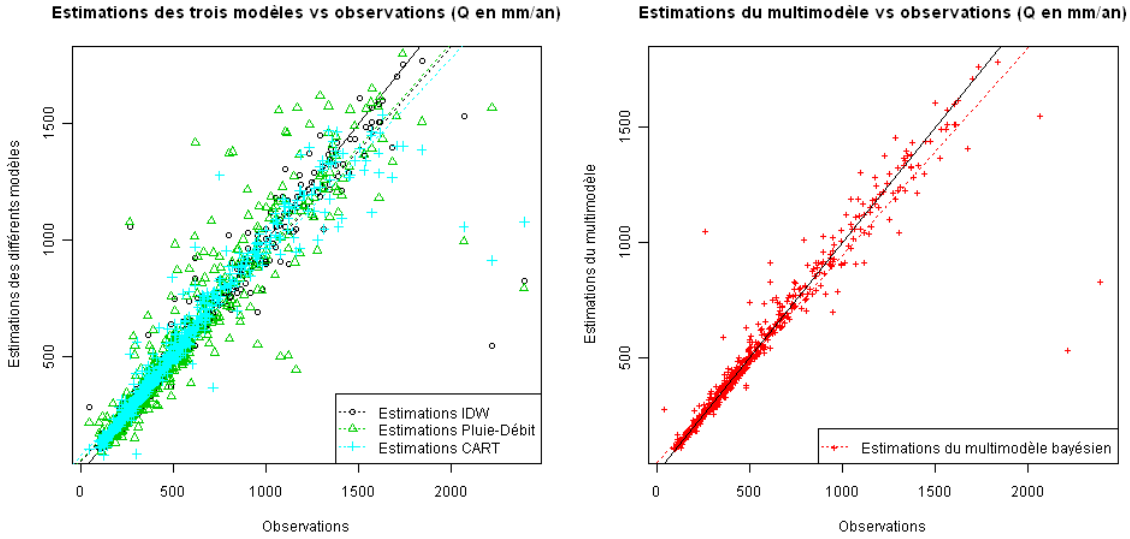


Figure 3 : Résultats du multi-modèle sur le Module

3. Cartographie de consensus et incertitude

- ***Quantification de l'incertitude***

Les distributions a postériori de nos variables cibles étant des distributions normales, elles sont caractérisées par une moyenne et une variance (écart type au carré). C'est cette variance qui nous donne accès à l'incertitude sur notre estimation finale.

Ainsi, pour chaque catégorie de surface, nous avons un écart type associé à l'estimation et nous pouvons encadrer nos valeurs dans un intervalle à 80% par exemple (intervalle défini pour l'étude) :

$Q_{référence}$ = ***m (moyenne de l'estimation) ± 1.28.σ***

Tableau 3 : Résultats du multi-modèle sur les deux variables– Valeurs des écart-types d'estimation

Erreur d'estimation	Multi-modèle de QMNA5	Multi-modèle de QA
Prior initial	1.22	0.66
Première mise à jour	0.75	0.51
Deuxième mise à jour	0.58	0.23
Troisième mise à jour	0.36	0.13

Ces valeurs sont applicables à des débits transformés. Si la loi qui caractérise la distribution de telles variables est une loi normale et permet donc d'avoir des intervalles d'erreur symétriques, lorsque l'on va repasser en débits non transformés par la racine, ces intervalles ne seront plus symétriques.
Par ailleurs, une valeur comprise entre 1 et 3 donne une information sur la convergence des modèles sur le tronçon sur lequel la valeur du débit est estimée. Cette note est appelée "fiabilité" de l'estimation, 3 étant appliqué à l'estimation la plus fiable, 1 à celle la moins fiable et N/A est utilisé pour une estimation pour laquelle un seul des modèles de prédétermination a fourni une estimation.

- ***Cartographie de consensus***

Les valeurs des débits estimés en l/s par le multi-modèle sont restituées sur chaque tronçon hydrographique issu du chainage de la base de données Carthage, version 2010. Sous forme de tableau, lorsque l'on connait ainsi le code du tronçon hydrographique, il est facile d'extraire la valeur qui intéresser l'utilisateur (on a ainsi accès à la valeur moyenne d'estimation, et aux limites inférieure et supérieure de cette valeur dans un intervalle à 80%).

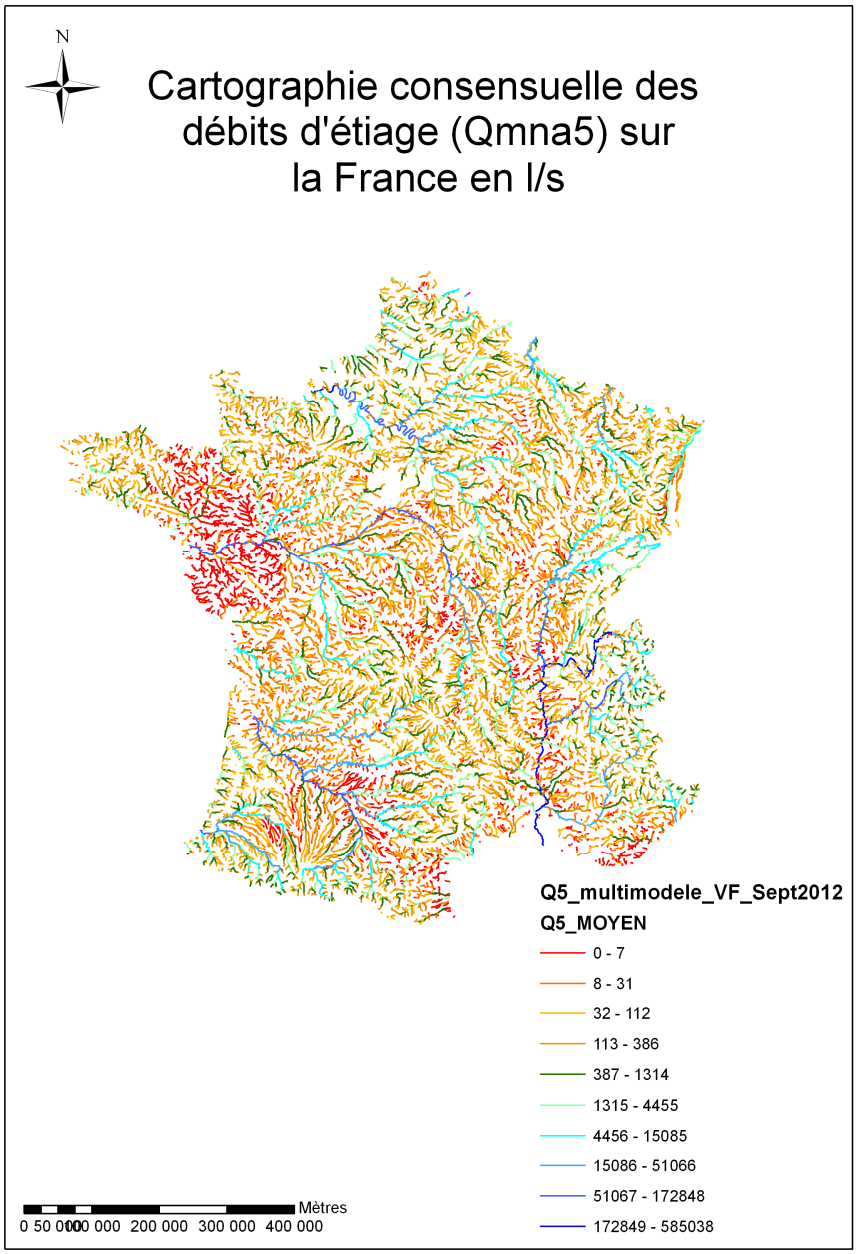


Figure 4 : Cartographie de consensus du QMNA5 sur la france entière en l/s

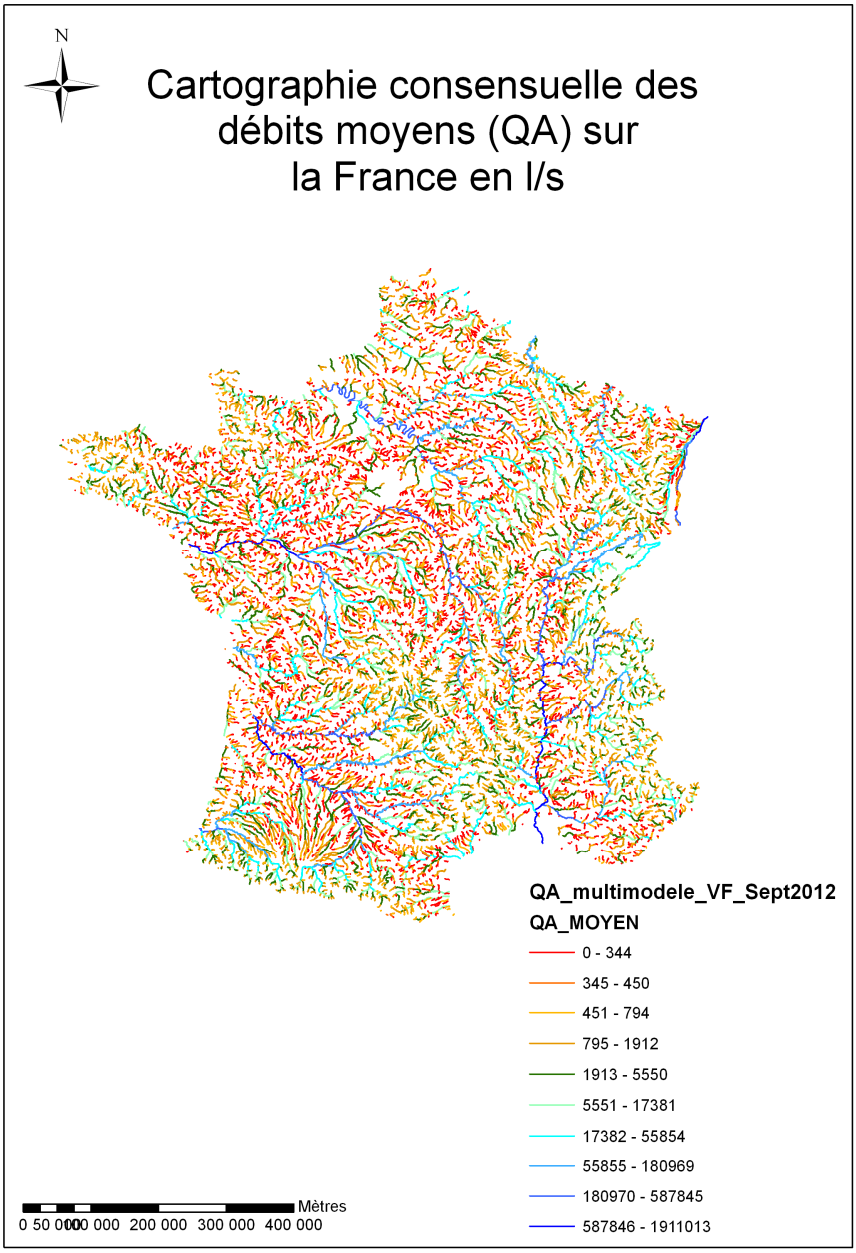


Figure 5 : Cartographie de consensus du module sur la France entière el l/s