Le modèle statistique de BaRatin

1 Modèles et hypothèses

1.1 Erreurs sur les jaugeages

Le modèle statistique de BaRatin suppose que les mesures de niveau et de débit des jaugeages $(\tilde{H}_i, \tilde{Q}_i)$ sont affectées par des erreurs gaussiennes de moyenne nulle (pas de biais) et d'écarts-types u_{H_i} et u_{Q_i}, qui sont les incertitudes-types des jaugeages. En première approche, il est recommandé d'ignorer l'incertitude sur les hauteurs jaugées ($u_{H_i} = 0$) : le cas où cette incertitude n'est pas nulle sera décrit dans la section 5. Mathématiquement, on utilise donc le modèle suivant :

$$\begin{aligned} \tilde{H}_i &= H_i \\ \tilde{Q}_i &= Q_i + \varepsilon_i^Q,\ \varepsilon_i^Q \sim \mathrm{N}(0, u_{Q_i}) \end{aligned} \quad (1)$$

Où H_i et Q_i représentent les valeurs réelles de hauteur et de débit, et ε_i^Q est l'erreur sur le débit jaugé.

1.2 Erreur restante (ou erreur structurelle)

La courbe de tarage est formalisée comme une fonction $f(h;\boldsymbol{\theta})$, où h est le niveau d'eau et $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_m)$ est le vecteur des m paramètres de la courbe de tarage (cf. fiche RCequation.pdf). On suppose que l'écart entre le débit réel et sa représentation mathématique f, forcément simplifiée, est une erreur gaussienne de moyenne nulle et d'écart-type $\sigma_f(h)$, qui peut varier en fonction de la hauteur :

$$Q_i = f(H_i;\boldsymbol{\theta}) + \varepsilon_i^f,\ \varepsilon_i^f \sim \mathrm{N}\left(0, \sigma_f(H_i)\right) \quad (2)$$

La façon dont l'écart-type de l'erreur restante peut varier en fonction de la hauteur est paramétrée comme une fonction du débit donné par la courbe de tarage. Deux options sont disponibles dans BaRatin :

$$\begin{aligned} &\text{Option 1 : écart-type constant, } \sigma_f(h;\boldsymbol{\gamma}) = \gamma_1 \\ &\text{Option 2 : } \sigma_f(h;\boldsymbol{\gamma}) = \gamma_1 + \gamma_2 f(h;\boldsymbol{\theta}) \end{aligned} \quad (3)$$

L'option 2 est recommandée par défaut, car on observe souvent que l'incertitude structurelle tend à augmenter avec le débit de la courbe de tarage.

1.3 Erreur totale

On suppose que l'erreur restante est indépendante de l'erreur de jaugeage. En combinant les équations (1) et (2), on aboutit au modèle d'erreur totale suivant :

$$\tilde{Q}_i = f(\tilde{H}_i;\boldsymbol{\theta}) + \varepsilon_i^f + \varepsilon_i^Q \quad \text{avec : } \varepsilon_i^f + \varepsilon_i^Q \sim \mathrm{N}\left(0, \sqrt{\sigma_f^2(\tilde{H}_i;\boldsymbol{\gamma}) + u_{Q_i}^2}\right) \quad (4)$$

Où $\tilde{H}_i$ et $\tilde{Q}_i$ sont les hauteurs et débits jaugés, et ε_i^f et ε_i^Q sont les erreurs gaussiennes sur la formulation mathématique de la courbe de tarage et sur les débits jaugés, respectivement. L'équation (4) stipule donc que le débit jaugé est égal au débit prédit par la courbe de tarage, plus une erreur liée à l'incertitude de jaugeage, plus une erreur liée à l'imperfection de la courbe de tarage.

L'équation (4) comporte plusieurs quantités inconnues: les paramètres $\boldsymbol{\theta}$ de la courbe de tarage et les paramètres $\boldsymbol{\gamma}$ gouvernant l'écart-type structurel. L'inférence sur ces quantités se fait en utilisant le formalisme bayésien (cf. fiche BayesianBasics.pdf). Il faut donc définir la vraisemblance et spécifier une distribution a priori comme décrit ci-après.

2 Information portée par les jaugeages : vraisemblance

D'après l'équation (4), un débit jaugé $\tilde{Q}_i$ suit une loi normale de moyenne $f(\tilde{H}_i;\boldsymbol{\theta})$ (i.e. le débit prédit par la courbe de tarage) et d'écart-type $\sqrt{\sigma_f^2(\tilde{H}_i;\boldsymbol{\gamma})+u_{Q_i}^2}$. En supposant que chaque débit jaugé est indépendant, on obtient la vraisemblance suivante :

$$p(\tilde{\boldsymbol{Q}}|\boldsymbol{\theta},\boldsymbol{\gamma},\tilde{\boldsymbol{H}}) = \prod_{i=1}^{N} p_{norm}\left(\tilde{Q}_i | f(\tilde{H}_i;\boldsymbol{\theta}), \sqrt{\sigma_f^2(\tilde{H}_i;\boldsymbol{\gamma})+u_{Q_i}^2}\right) \quad (5)$$

où $\tilde{\boldsymbol{Q}} = (\widetilde{Q_1},\dots,\widetilde{Q_N})$ sont les N débits jaugés et $p_{norm}(z|m,s)$ représente la densité de probabilité d'une loi normale de moyenne m et d'écart-type s, évaluée en une valeur z.

3 Information liée aux connaissances hydraulique : distribution a priori

La distribution a priori permet d'intégrer les connaissances hydrauliques discutées dans la fiche HydraulicControls.pdf. Dans BaRatin, on utilise des distributions a priori indépendantes sur chaque paramètre à estimer, conduisant à :

$$p(\boldsymbol{\theta},\boldsymbol{\gamma}) = p(\gamma_1)p(\gamma_2)\prod_{i=1}^{m} p(\theta_i) \quad (6)$$

4 Théorème de Bayes et distribution a posteriori

Comme détaillé dans la fiche BayesianBasics.pdf, le théorème de Bayes est ensuite utilisé pour calculer densité de la distribution a posteriori (à une constante de proportionnalité près):

$$p(\boldsymbol{\theta},\boldsymbol{\gamma}|\tilde{\boldsymbol{Q}},\tilde{\boldsymbol{H}}) \propto p(\tilde{\boldsymbol{Q}}|\boldsymbol{\theta},\boldsymbol{\gamma},\tilde{\boldsymbol{H}})p(\boldsymbol{\theta},\boldsymbol{\gamma}) \quad (7)$$

Comme détaillé dans la fiche MCMC.pdf, on utilise un algorithme MCMC pour explorer cette distribution a posteriori. On obtient ainsi un grand nombre de réalisations $(\boldsymbol{\theta}^{(j)},\boldsymbol{\gamma}^{(j)})_{j=1:M}$ issues de la distribution a posteriori. A chacune de ces réalisations correspond une courbe de tarage (de paramètres $\boldsymbol{\theta}^{(j)}$), ce qui conduit à générer un ensemble de courbes de tarage plausibles au vu des jaugeages et des connaissance hydrauliques a priori.

5 [Avancé] Cas des hauteurs jaugées incertaines

Si on s'affranchit de l'hypothèse que les hauteurs jaugées sont parfaites, l'équation (1) devient :

$$\tilde{H}_i = H_i + \varepsilon_i^H,\ \varepsilon_i^H \sim \mathrm{N}(0, u_{H_i}) \quad (8)$$

L'équation (4) représentant l'erreur totale et utilisée pour calculer la vraisemblance doit donc être modifiée pour incorporer l'erreur sur la hauteur jaugée :

$$\tilde{Q}_i = f(\tilde{H}_i - \varepsilon_i^H;\boldsymbol{\theta}) + \varepsilon_i^f + \varepsilon_i^Q \quad \text{avec : } \varepsilon_i^f + \varepsilon_i^Q \sim \mathrm{N}\left(0, \sqrt{\sigma_f^2(\tilde{H}_i;\boldsymbol{\gamma})+u_{Q_i}^2}\right) \quad (9)$$

Malheureusement, l'équation (9) ne permet pas directement d'écrire la vraisemblance. En effet, l'erreur sur la hauteur transite à travers le modèle non-linéaire de courbe de tarage. En conséquence, l'erreur de débit résultant de cette erreur de hauteur n'est pas gaussienne. Pour s'affranchir de ce problème, on considère dans BaRatin que l'erreur ε_i^H est un paramètre inconnu que l'on va chercher

à estimer (ou de manière équivalente, on va chercher à estimer la vraie hauteur, ou à corriger la hauteur jaugée). Cette estimation sera contrainte par l'incertitude spécifiée pour la hauteur jaugée, que l'on utilisera ici comme une distribution a priori.

Si l'on incorpore les erreurs de hauteur $\boldsymbol{\varepsilon} = (\varepsilon_1^H, \dots, \varepsilon_N^H)$ dans la liste des paramètres à estimer, la vraisemblance s'écrit:

$$p(\widetilde{\boldsymbol{Q}}|\boldsymbol{\theta}, \boldsymbol{\gamma}, \boldsymbol{\varepsilon}, \widetilde{\boldsymbol{H}}) = \prod_{i=1}^{N} p_{norm}\left(\tilde{Q}_i | f(\widetilde{H}_i - \varepsilon_i^H; \boldsymbol{\theta}), \sqrt{\sigma_f^2(\widetilde{H}_i; \boldsymbol{\gamma}) + u_{Q_i}^2}\right) \quad (10)$$

La distribution a priori devient :

$$p(\boldsymbol{\theta}, \boldsymbol{\gamma}, \boldsymbol{\varepsilon}) = p(\gamma_1)p(\gamma_2)\prod_{i=1}^{m} p(\theta_i)\prod_{j=1}^{N} p_{norm}(\varepsilon_i^H | 0, u_{H_i}) \quad (11)$$

La distribution a posteriori s'obtient classiquement avec le théorème de Bayes :

$$p(\boldsymbol{\theta}, \boldsymbol{\gamma}, \boldsymbol{\varepsilon} | \widetilde{\boldsymbol{Q}}, \widetilde{\boldsymbol{H}}) \propto p(\widetilde{\boldsymbol{Q}} | \boldsymbol{\theta}, \boldsymbol{\gamma}, \widetilde{\boldsymbol{\varepsilon}, \boldsymbol{H}}) p(\boldsymbol{\theta}, \boldsymbol{\gamma}, \boldsymbol{\varepsilon}) \quad (12)$$

La prise en compte explicite de l'incertitude sur les hauteurs jaugées a donc un coût important en termes de complexité et de temps de calcul, puisque chaque hauteur considérée comme incertaine ajoute un paramètre à estimer. Précisons néanmoins que ces paramètres sont en général très fortement contraints par l'incertitude de hauteur (tant que u_{H_i} n'est pas trop grand), ce qui rend cette estimation possible en général.