BaRatin statistical model

1 Models and hypotheses

1.1 Gauging errors

BaRatin assumes that stage/discharge measurements $(\tilde{H}_i, \tilde{Q}_i)$ are affected by Gaussian errors with zero mean (no bias) and known standard deviations u_{H_i} and u_{Q_i}. In general, we recommend ignoring stage uncertainty for gaugings ($u_{H_i} = 0$), at least as a first approximation. The case of non-negligible stage uncertainty will be described in section 5. The following statistical model is therefore used:

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

where H_i and Q_i denote the true stage/discharge values and ε_i^Q denotes the discharge error.

1.2 Remnant error (or structural error)

The rating curve is formalized as a function $f(h;\boldsymbol{\theta})$, where h is the stage and $\boldsymbol{\theta} = (\theta_1, \dots, \theta_m)$ is the vector containing the m parameters of the rating curve (see note RCequation.pdf). We assume that the difference between the true discharge and its simplified mathematical representation f is a realisation from a Gaussian distribution with zero mean and standard deviation $\sigma_f(h)$. The latter may vary as a function of stage:

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

The variation of the standard deviation $\sigma_f(h)$ with stage is parameterised as a function of the discharge computed from the rating curve. Two options are available in BaRatin:

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

Option 2 is recommended by default because it is frequently observed that structural uncertainty increases with the rating curve discharge.

1.3 Total error

Assuming that the remnant error and the gauging error are independent, the total error can be written as follows by combining equations (1) and (2):

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

where $\tilde{H}_i$ and $\tilde{Q}_i$ are the gauged stage/discharge, and ε_i^f and ε_i^Q are the Gaussian remnant and gauging errors, respectively. Equation (4) therefore states that the gauged discharge is equal to the discharge computed with the rating curve, plus an error due to gauging uncertainty, plus an error due to the imperfect formulation of the rating curve.

Several unknown quantities appear in equation (4): the rating curve parameters $\boldsymbol{\theta}$ and the parameters $\boldsymbol{\gamma}$ describing the standard deviation of remnant errors. Inference for these unknown quantities is based on a Bayesian approach (cf. note BayesianBasics.pdf). This requires deriving the likelihood function and specifying a prior distribution, as described next.

2 Information brought by the gaugings: likelihood

From equation (4), the gauged discharge $\tilde{Q}_i$ is a realisation from a Gaussian distribution with mean $f(\tilde{H}_i;\boldsymbol{\theta})$ (i.e. the rating curve discharge) and standard deviation $\sqrt{\sigma_f^2(\tilde{H}_i;\boldsymbol{\gamma})+u_{Q_i}^2}$. Assuming that all gauged discharges are mutually independent, the likelihood can be written as:

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

where $\tilde{\boldsymbol{Q}}=(\widetilde{Q_1},\dots,\widetilde{Q_N})$ denote the N gauged discharge and $p_{norm}(z|m,s)$ denote the probability density function (pdf) of a Gaussian distribution with mean m and standard deviation s, evaluated at some value z.

3 Information brought by hydraulics: prior distribution

The prior distribution offers the opportunity to include hydraulic knowledge, as discussed in the note HydraulicControls.pdf. BaRatin uses independent prior distributions for each inferred parameter, leading to the joint prior distribution:

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

4 Bayes theorem and posterior distribution

As explained in the note BayesianBasics.pdf, Bayes theorem is used to compute the pdf of the posterior distribution (up to a constant of proportionality):

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

As explained in the note MCMC.pdf, a MCMC sampler is then used to explore the posterior distribution. This results in a large number of realisations $(\boldsymbol{\theta}^{(j)},\boldsymbol{\gamma}^{(j)})_{j=1:M}$ from the posterior distribution. Each realisation can be associated with a rating curve (with parameters $\boldsymbol{\theta}^{(j)}$), which yields an ensemble of rating curves that are plausible given the gaugings and the hydraulic knowledge.

5 [Advanced] Handling stage uncertainty in gaugings

Abandoning the hypothesis that gauged stages are perfectly measured, equation (1) becomes:

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

Equation (4) describing the total error and used to compute the likelihood therefore has to be modified to include this non-zero stage error:

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

Unfortunately, equation (9) does not yield a closed-form expression for the likelihood. This is because the stage error ε_i^H propagates through the nonlinear rating curve function f. Consequently, the discharge error resulting from the stage error is not Gaussian. In order to circumvent this difficulty, the stage error ε_i^H is considered as a new unknown quantity that needs to be estimated (equivalently, this can be viewed as estimating the true stage or correcting the gauged stage). This estimation is constrained by the specified stage uncertainty u_{H_i}, that is used here as a prior information.

Adding the stage errors $\boldsymbol{\varepsilon} = (\varepsilon_1^H, \dots, \varepsilon_N^H)$ into the list of unknown parameters, the likelihood becomes:

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

The prior distribution becomes:

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

The posterior distribution is then derived (up to a constant of proportionality) using Bayes theorem:

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

Explicitly accounting for stage uncertainty in gaugings therefore has an important computational cost, because each stage value considered as uncertain correspond to an additional unknown parameter. Note however that these additional parameters are strongly constrained by the specified stage uncertainty (as long as u_{H_i} is not too large), which makes this estimation feasible in general.