

Calibration of watershed models: How to deal with input uncertainty?

Peter Reichert

*Swiss Federal Institute for Environmental Science
and Technology (EAWAG)
8600 Dübendorf, Switzerland*

and

*Department of Environmental Sciences
Swiss Federal Institute of Technology (ETH)
8092 Zürich, Switzerland*

reichert@eawag.ch

Contents

- **Misunderstandings**
- **Concept of Talk**
- **Statistical Inference**
- **How to Consider Input Uncertainty**
- **Input Uncertainty Models**
- **Misunderstandings Revisited**
- **Key Questions**

Misunderstandings

Several misunderstandings pop up again and again.

1. If assumptions underlying least squares regression do not apply, we need replacement or extension of statistical inference theory.
2. As we will never have the 'true' model, we cannot apply statistical inference theory.
3. Statistical inference theory does not account for input uncertainty.

Concept of Talk

- Statistical inference theory is a very general framework of which many special cases are used in applications.
- Although a small number of such special cases dominate the literature (e.g. least squares regression), non-applicability of such a special case does not imply non-applicability of statistical inference theory.
- In this talk I would like to give an idea of how more innovative (and useful) applications of statistical inference theory could be designed, with special emphasis on input uncertainty.
- I hope that this stimulates the discussion of which techniques are best to apply how in hydrological modelling and if new techniques are required.

Statistical Inference

Notation:

- M : subscript characterizing the model;
- y : system variables predicted by the model;
- x : input variables of the model;
- θ : model parameters;
- \sim : variables with a \sim refer to measurements
(adding measurement error to variable uncertainty;
this includes aggregation and extrapolation error);
- dat: the subscript 'dat' indicates substitution of measured data for the corresponding variable;
- Y, X, Θ : capitalized variables represent random variables corresponding to the lower case variables;
- f : probability densities.

Statistical Inference

Model Equations:

$$\tilde{Y}_M(\Theta) \quad \text{or} \quad f_{M, \tilde{Y}|\Theta}(\tilde{y} | \theta) \quad , \quad f_{\Theta, \text{pri}}(\theta)$$

Frequentist Parameter Estimation (Maximum Likelihood):

$$\hat{\theta}_M(\tilde{y}_{\text{dat}}) = \operatorname{argmax}_{\theta} \left(f_{M, \tilde{Y}|\Theta}(\tilde{y}_{\text{dat}} | \theta) \right)$$

$$\hat{\Sigma}_M(\tilde{y}_{\text{dat}}) \text{ calculated from } \begin{cases} \hat{\theta}_M(\tilde{Y}_M(\theta)) \\ \hat{\theta}_M(\text{bootstrap sample of } \{\tilde{y}_{\text{dat}}\}) \end{cases}$$

Bayesian Parameter Estimation (Updating):

$$f_{\Theta, \text{post}}(\theta | \tilde{y}_{\text{dat}}) = c f_{M, \tilde{Y}|\Theta}(\tilde{y}_{\text{dat}} | \theta) f_{\Theta, \text{pri}}(\theta)$$

Statistical Inference

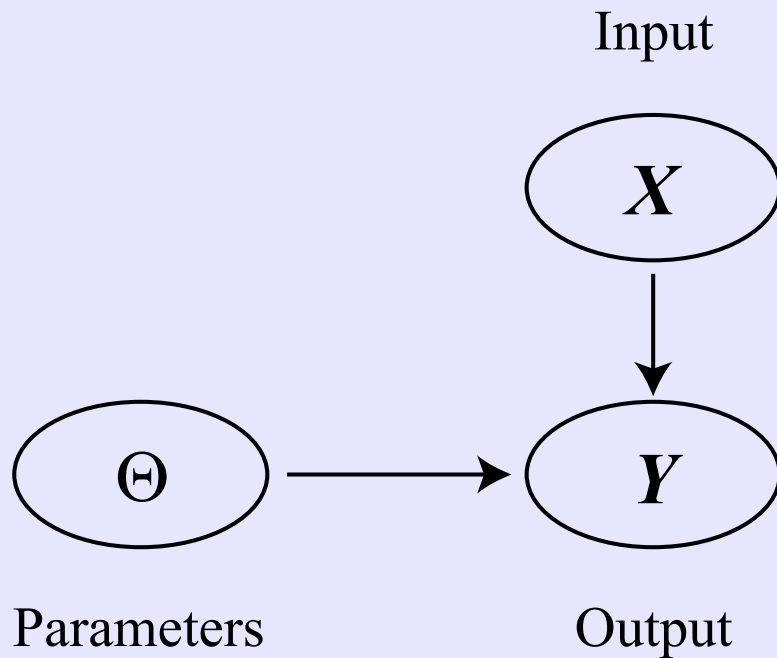
Different assumptions on

- which parameters are explicitly included,
- which variables are included in the model,
- which mechanistic description is used in the model,
- which distributional assumptions are made in the model,
- how input uncertainty is considered in the model.

lead to a huge class of techniques all based on the same basic methodology.

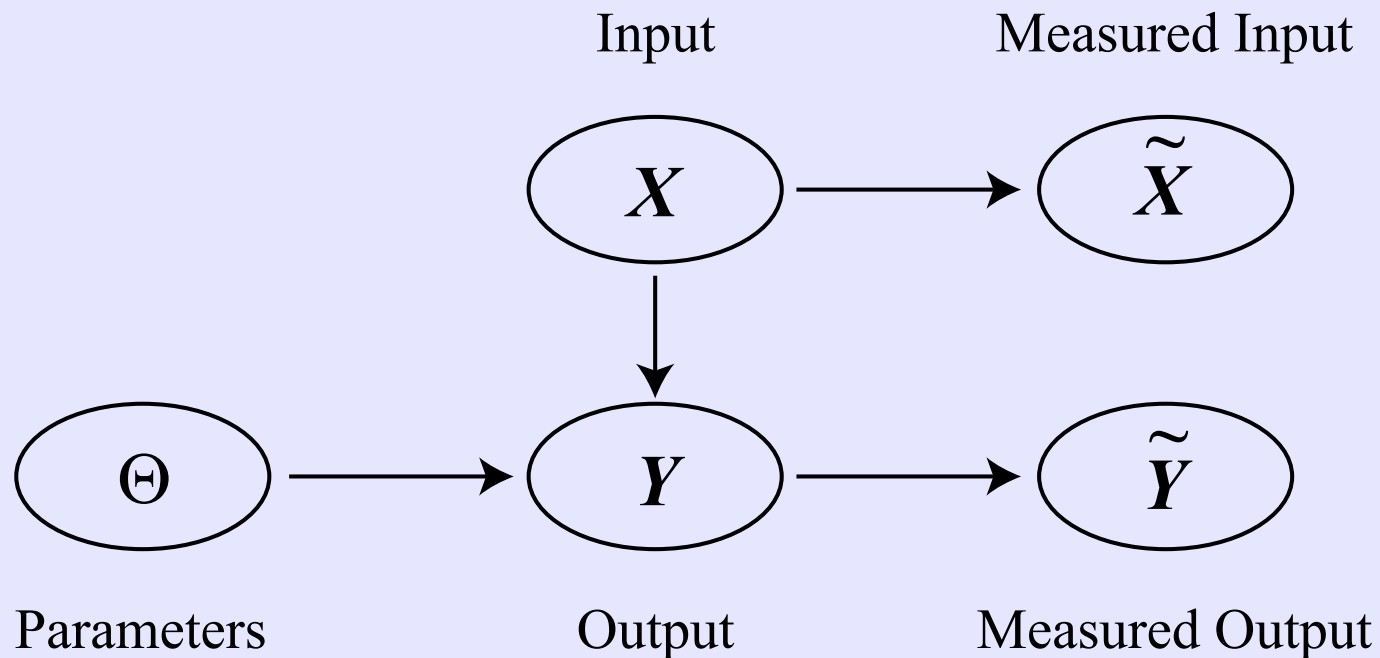
How to Consider Input Uncertainty

Dependence Structure of Hydrological Model:



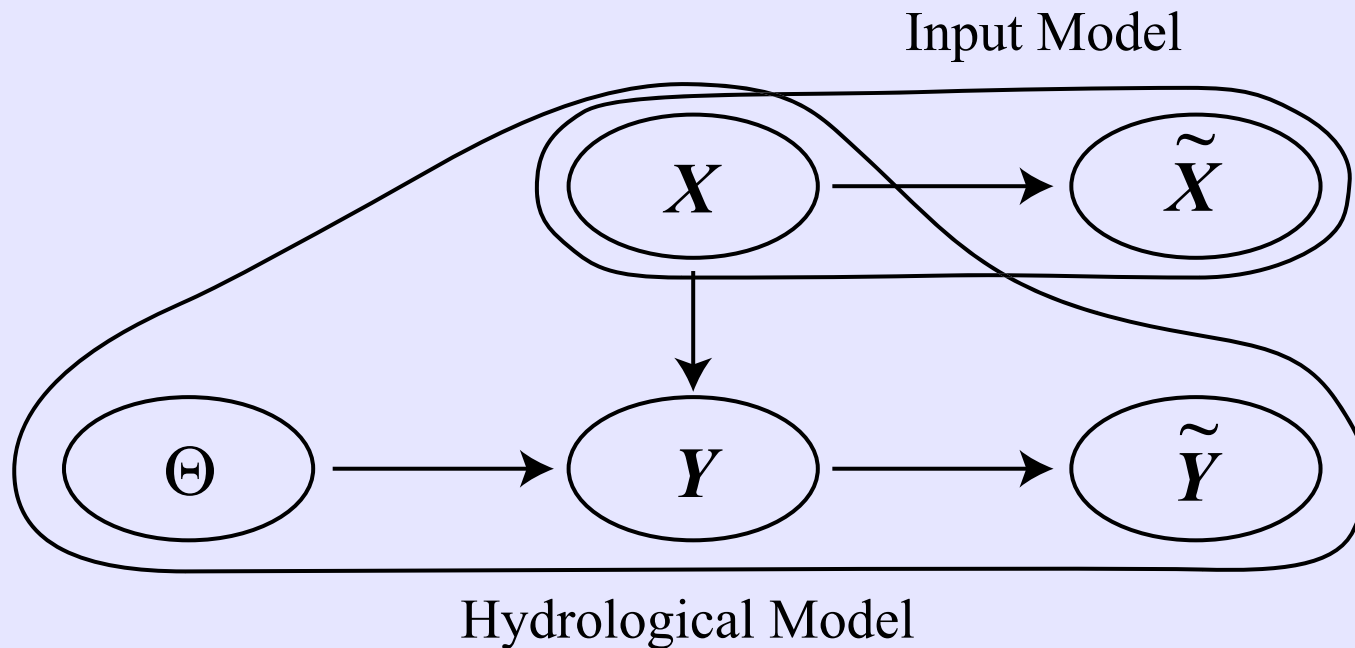
How to Consider Input Uncertainty

Dependence Structure of all Variables:



How to Consider Input Uncertainty

Division into Submodels:



How to Consider Input Uncertainty

Two Apparent Alternatives:

1. **Inference of parameters considering input uncertainty:**
explicitly formulating the model equations with inputs as random variables.
2. **Joint inference of parameters and input:**
add input variables to the parameters to be inferred and treat input measurements similarly to output measurements for joint inference.

In the following, it is shown that both alternatives lead to essentially the same results.

How to Consider Input Uncertainty

Required model formulations:

(we assume dependence structure according to the diagram shown earlier)

Model for measured output with making input uncertainty explicit:

$$f_{M, \tilde{Y} | \Theta}(\tilde{y} | \theta) = f_{M, \tilde{Y} | \Theta}(\tilde{y} | \theta, \mathbf{X})$$

Model for measured output, conditional on inputs:

$$f_{M, \tilde{Y} | \Theta, \mathbf{X}}(\tilde{y} | \theta, \mathbf{x})$$

Input measurement model:

$$f_I(\tilde{\mathbf{x}} | \mathbf{x})$$

How to Consider Input Uncertainty

1. Inference considering input uncertainty:

Likelihood function with explicit consideration of inputs:

$$f_{M, \tilde{Y} | \Theta}(\tilde{y} | \theta) = f_{M, \tilde{Y} | \Theta, \mathbf{X}}(\tilde{y} | \theta, \mathbf{X}) = \int f_{M, \tilde{Y} | \Theta, \mathbf{X}}(\tilde{y} | \theta, \mathbf{x}) f_{\mathbf{X}, \text{pri}}(\mathbf{x}) d\mathbf{x}$$

In the presence of input data, $f_{\mathbf{X}, \text{pri}}(\mathbf{x})$ should be updated:

$$f_{\mathbf{X}, \text{post}}(\mathbf{x} | \tilde{\mathbf{x}}_{\text{dat}}) \propto f_I(\tilde{\mathbf{x}}_{\text{dat}} | \mathbf{x}) f_{\mathbf{X}, \text{pri}}(\mathbf{x})$$

We then get the final form of the likelihood function:

$$f_{M, \tilde{Y} | \Theta}(\tilde{y} | \theta) \propto \int f_{M, \tilde{Y} | \Theta, \mathbf{X}}(\tilde{y} | \theta, \mathbf{x}) f_I(\tilde{\mathbf{x}}_{\text{dat}} | \mathbf{x}) f_{\mathbf{X}, \text{pri}}(\mathbf{x}) d\mathbf{x}$$

How to Consider Input Uncertainty

This leads to the posterior distribution of parameters:

$$f_{\Theta, \text{post}}(\boldsymbol{\theta} \mid \tilde{\mathbf{y}}_{\text{dat}}, \tilde{\mathbf{x}}_{\text{dat}}) \\ \propto \int f_{M, \tilde{\mathbf{Y}} \mid \Theta, \mathbf{X}}(\tilde{\mathbf{y}}_{\text{dat}} \mid \boldsymbol{\theta}, \mathbf{x}) f_I(\tilde{\mathbf{x}}_{\text{dat}} \mid \mathbf{x}) f_{\mathbf{X}, \text{pri}}(\mathbf{x}) \, d\mathbf{x} \cdot f_{\Theta, \text{pri}}(\boldsymbol{\theta})$$

How to Consider Input Uncertainty

2. Joint inference of parameters and inputs:

Likelihood function:

$$f_M(\tilde{y}, \tilde{\mathbf{x}} \mid \boldsymbol{\theta}, \mathbf{x}) = f_{M, \tilde{Y} \mid \Theta, \mathbf{X}}(\tilde{y} \mid \boldsymbol{\theta}, \mathbf{x}) f_I(\tilde{\mathbf{x}} \mid \mathbf{x})$$

Posterior distribution of parameters and inputs:

$$\begin{aligned} f_{\Theta, \text{post}}(\boldsymbol{\theta}, \mathbf{x} \mid \tilde{y}_{\text{dat}}, \tilde{\mathbf{x}}_{\text{dat}}) \\ \propto f_{M, \tilde{Y} \mid \Theta, \mathbf{X}}(\tilde{y}_{\text{dat}} \mid \boldsymbol{\theta}, \mathbf{x}) f_I(\tilde{\mathbf{x}}_{\text{dat}} \mid \mathbf{x}) f_{\mathbf{X}, \text{pri}}(\mathbf{x}) f_{\Theta, \text{pri}}(\boldsymbol{\theta}) \end{aligned}$$

Marginal posterior distribution of parameters:

$$\begin{aligned} f_{\Theta, \text{post}}(\boldsymbol{\theta} \mid \tilde{y}_{\text{dat}}, \tilde{\mathbf{x}}_{\text{dat}}) \\ \propto \int f_{M, \tilde{Y} \mid \Theta, \mathbf{X}}(\tilde{y}_{\text{dat}} \mid \boldsymbol{\theta}, \mathbf{x}) f_I(\tilde{\mathbf{x}}_{\text{dat}} \mid \mathbf{x}) f_{\mathbf{X}, \text{pri}}(\mathbf{x}) f_{\Theta, \text{pri}}(\boldsymbol{\theta}) d\mathbf{x} \end{aligned}$$

How to Consider Input Uncertainty

It is obviously the same to explicitly consider input uncertainty or to jointly infer parameters and input and marginalize over inputs to get the parameter distribution.

Not initially marginalizing over the inputs has the advantage to give updated information on inputs which may be useful on its own but also for diagnosing potential problems of the modelling approach.

The integral is no problem in the second approach as we will construct a sample of this distribution; just ignoring sample information related to \mathbf{x} considers the integral.

Input Uncertainty Models

What we have to provide:

$f_{M, \tilde{Y} | \Theta, \mathbf{X}}(\tilde{y} | \theta, \mathbf{x})$: distribution of measured model results conditional on parameters and inputs (model and measurement error of outputs);

$f_I(\tilde{\mathbf{x}} | \mathbf{x})$: distribution of measured model input conditional on true model inputs (measurement error of inputs);

$f_{\mathbf{X}, \text{pri}}(\mathbf{x})$: prior distribution of model inputs;

$f_{\Theta, \text{pri}}(\theta)$: prior distribution of model parameters.

Different inference techniques differ in assumptions and formulations of these functions.

Input Uncertainty Models

Example 1

Deterministic model, $y_M(\boldsymbol{\theta}, \mathbf{x})$, with normally distributed output errors without input errors.

Model equations, conditional on inputs:

$$f_{M, \tilde{\mathbf{Y}} | \boldsymbol{\Theta}, \mathbf{X}}(\tilde{\mathbf{y}} | \boldsymbol{\theta}, \mathbf{x}) = \prod_{i=1}^{n_{\boldsymbol{\Theta}}} \frac{1}{\sqrt{2\pi}} \frac{1}{\sigma_{\tilde{y}_i}} \exp \left(-\frac{\left(\tilde{y}_i - y_{M,i}(\boldsymbol{\theta}, \mathbf{x}) \right)^2}{2\sigma_{\tilde{y}_i}^2} \right)$$

Input model:

$$f_I(\tilde{\mathbf{x}} | \mathbf{x}) = \delta(\tilde{\mathbf{x}} - \mathbf{x})$$

Input Uncertainty Models

Example 1 (continued)

Frequentist Parameter Estimation (Maximum Likelihood):

$$\hat{\theta}_M(\tilde{y}_{\text{dat}}, \tilde{\mathbf{x}}_{\text{dat}}) = \operatorname{argmin}_{\theta} \left(\sum_{i=1}^{n_{\Theta}} \frac{\left(\tilde{y}_{\text{dat},i} - y_{M,i}(\theta, \tilde{\mathbf{x}}_{\text{dat}}) \right)^2}{\sigma_{\tilde{y}_i}^2} \right)$$

Bayesian Parameter Estimation (Updating):

$$f_{\Theta, \text{post}}(\theta \mid \tilde{y}_{\text{dat}}, \tilde{\mathbf{x}}_{\text{dat}}) \propto \exp \left(- \sum_{i=1}^{n_{\Theta}} \frac{\left(\tilde{y}_{\text{dat},i} - y_{M,i}(\theta, \tilde{\mathbf{x}}_{\text{dat}}) \right)^2}{2\sigma_{\tilde{y}_i}^2} \right) \cdot f_{\Theta, \text{pri}}(\theta)$$

Input Uncertainty Models

Example 2

Deterministic model, $y_M(\boldsymbol{\theta}, \mathbf{x})$, with normally distributed output and input errors.

Model equations, conditional on inputs:

$$f_{M, \tilde{Y} | \Theta, \mathbf{X}}(\tilde{y} | \boldsymbol{\theta}, \mathbf{x}) = \prod_{i=1}^{n_{\Theta}} \frac{1}{\sqrt{2\pi}} \frac{1}{\sigma_{\tilde{y}_i}} \exp \left(-\frac{\left(\tilde{y}_i - y_{M,i}(\boldsymbol{\theta}, \mathbf{x}) \right)^2}{2\sigma_{\tilde{y}_i}^2} \right)$$

Input model:

$$f_I(\tilde{\mathbf{x}} | \mathbf{x}) = \prod_{i=1}^{n_{\mathbf{X}}} \frac{1}{\sqrt{2\pi}} \frac{1}{\sigma_{\tilde{x}_i}} \exp \left(-\frac{(\tilde{x}_i - x_i)^2}{2\sigma_{\tilde{x}_i}^2} \right)$$

Input Uncertainty Models

Example 2 (continued)

Frequentist Parameter Estimation (Maximum Likelihood):

$$\begin{aligned} & (\hat{\boldsymbol{\theta}}, \hat{\mathbf{x}})_{M}(\tilde{\mathbf{y}}_{\text{dat}}, \tilde{\mathbf{x}}_{\text{dat}}) \\ &= \operatorname{argmin}_{\boldsymbol{\theta}, \mathbf{x}} \left(\sum_{i=1}^{n_{\Theta}} \frac{(\tilde{y}_{\text{dat},i} - y_{M,i}(\boldsymbol{\theta}, \mathbf{x}))^2}{\sigma_{\tilde{y}_i}^2} + \sum_{i=1}^{n_{\mathbf{X}}} \frac{(\tilde{x}_{\text{dat},i} - x_i)^2}{\sigma_{\tilde{x}_i}^2} \right) \end{aligned}$$

Bayesian Parameter Estimation (Updating):

$$f_{\Theta, \text{post}}(\boldsymbol{\theta}, \mathbf{x} \mid \tilde{\mathbf{y}}_{\text{dat}}, \tilde{\mathbf{x}}_{\text{dat}}) \\ \propto \exp \left(- \sum_{i=1}^{n_{\Theta}} \frac{\left(\tilde{y}_{\text{dat}, i} - y_{M, i}(\boldsymbol{\theta}, \mathbf{x}) \right)^2}{2\sigma_{\tilde{y}_i}^2} \right) \\ \cdot \exp \left(- \sum_{i=1}^{n_{\mathbf{X}}} \frac{(\tilde{x}_{\text{dat}, i} - x_i)^2}{2\sigma_{\tilde{x}_i}^2} \right) \cdot f_{\Theta, \text{pri}}(\boldsymbol{\theta}) \cdot f_{\mathbf{X}, \text{pri}}(\mathbf{x})$$

Misunderstandings Revisited

1. If assumptions underlying least squares regression do not apply, we need replacement or extension of statistical inference theory.

Least squares regression is only a special case of statistical inference theory. More innovative assumptions have to be taken into account for hydrological model calibration. Those can easily be made within the framework of statistical inference theory.

Misunderstandings Revisited

2. As we will never have the 'true' model, we cannot apply statistical inference theory.

All methodologies have the problem that the results are not valid if the assumptions are violated. This is not only true for statistical inference theory. By analyzing distributions of residuals, the validity of statistical assumptions can be tested. Models have to be designed to minimize deviations from these assumptions and assumptions should be tested for robustness against small deviations.

Misunderstandings Revisited

3. Statistical inference theory does not account for input uncertainty.

Consideration of input uncertainty is crucial in hydrological modelling (independent of the applied inference methodology). This can easily be done within the framework of statistical inference theory. It should be a major issue at this workshop to discuss how this can best be done.

Key Questions

Questions, important to address in the workshop:

1. Which input models, $f_I(\tilde{\mathbf{x}} | \mathbf{x})$, do adequately describe input uncertainty in catchment-scale hydrologic models?
2. Which model is most adequate to describe hydrological output and its measurement in $f_{M, \tilde{Y} | \Theta, \mathbf{X}}(\tilde{y} | \boldsymbol{\theta}, \mathbf{x})$?
3. How to combine multiple objectives with the statistical approach?
4. Is there only a need for a better formulating statistical models or is there a need for conceptually different approaches? If yes, with respect to which aspects?

References/Acknowledgements

- The paper on input uncertainty by Kavetski, Franks and Kuczera (2003) provided the most important insights worked out in this talk.
- The book “An Introduction to Bayesian Inference in Econometrics” from 1971 gives an excellent treatment of the problem.
- Stimulating discussions with Mark Borsuk were very supportive for developing the formalism.