Rainfall Runoff Modelling Using Gaussian Processes

C. Fischer and S. Kralisch

Institute for Geography, Department of Geoinformatics, Friedrich-Schiller-University, 07737 Jena, Germany

1. Objective

Runoff predictions with least effort and minimal data requirement.

Input Data:

Timeseries of

- Rainfall $I_t = \left(I_t^1, I_t^2, \ldots, I_t^r\right)$ and
- Runoff $Q_t = \left(Q_t^1, Q_t^2, \ldots, Q_t^s\right)$ where $t \in \{1, 2, \ldots, T\}$.

With $x_t := (I_t, Q_t)$ we denote all measurements at timestep t.

Assume there is an unknown function

$$f\left(x_t, x_{t-1}, \ldots, x_{t-k}\right) = Q_{t+1}^* + \epsilon,$$

which predicts future runoff Q_{t+1}^*, but is disturbed by noise ϵ.

Goal:

Find a function $\hat{f}$ which approximates f as good as possible.

2. Gaussian Processes

Think of a multivariate normal distribution with infinite many variables and you get a gaussian process.

Properties:

Figure 1: (a) Univariate normal distribution (b) Bivariate normal distribution

- Fully described by **mean function** $m(x)$ and **covariance function** $k(x)$
- Marginalization property:

$$p(x,y) \sim \mathcal{N}\left(\begin{bmatrix} a \\ b \end{bmatrix}, \begin{bmatrix} A & B \\ B^T & C \end{bmatrix}\right) \Rightarrow p(x) \sim \mathcal{N}(a, A)$$

- Conditional property:

$$p(x,y) \sim \mathcal{N}\left(\begin{bmatrix} a \\ b \end{bmatrix}, \begin{bmatrix} A & B \\ B^T & C \end{bmatrix}\right) \Rightarrow p(x|y) \sim \mathcal{N}\left(a + BC^{-1}(y-b), A - BC^{-1}B^T\right)$$

This gives information about x, if y is known.

Figure 2: Gaussian Process Regression Example - The shaded region represents the pointwise mean plus and minus two times the standard deviation.

3. Covariance Function

- Wide variety of covariance functions
- Incorporate a priori decisions

(a) Squared Exponential $\quad$ (b) Rational Quadratic $\quad$ (c) Matérn

$k_{SE}(r) = \exp\left(-\frac{r^2}{2l^2}\right)$ $\quad$ $k_{RQ}(r) = \left(1 + \frac{r^2}{2\alpha l^2}\right)^{-\alpha}$ $\quad$ $k_{Matern}(r) = \frac{2^{1-\nu}}{\Gamma(\nu)}\left(\frac{\sqrt{2\nu}r}{l}\right)^{\nu} K_\nu\left(\frac{\sqrt{2\nu}r}{l}\right)$
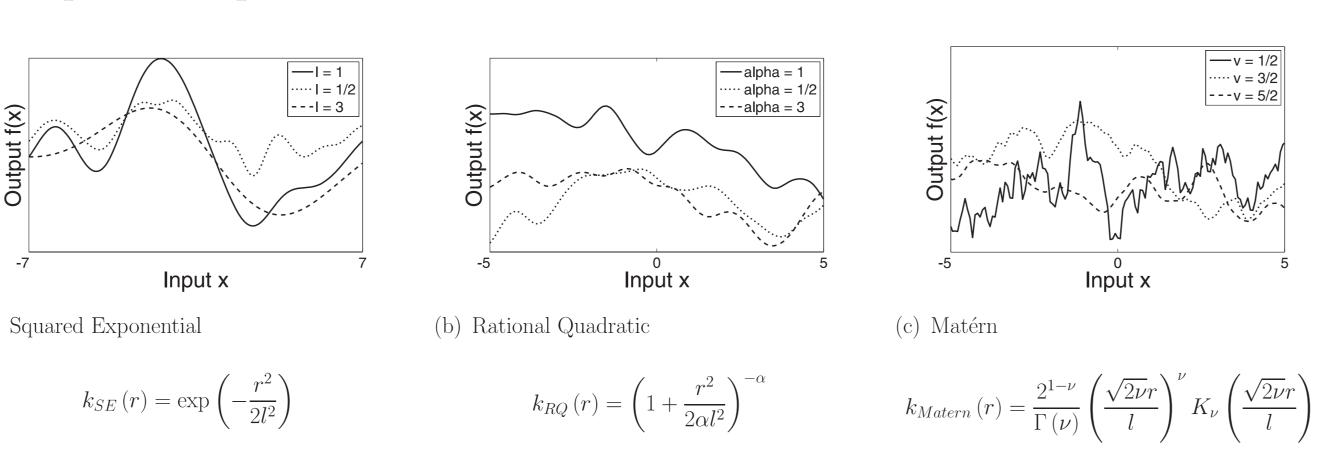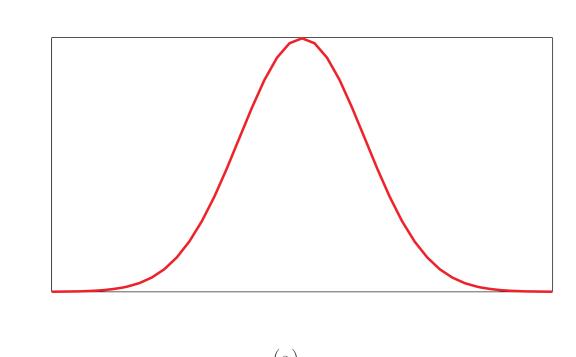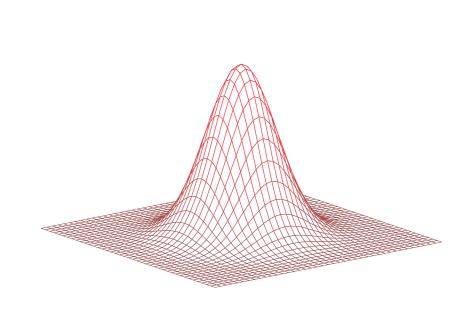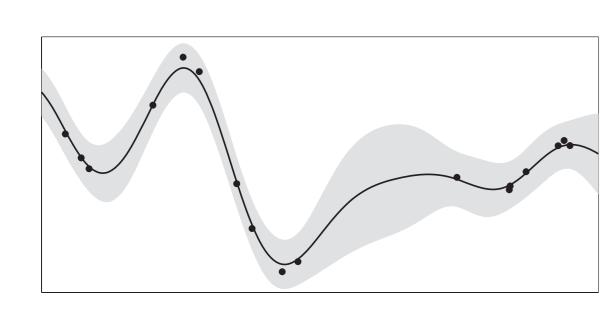
Figure 3: Random functions drawn from Gaussian Processes with different covariance functions and different parameters

4. How to find a good covariance function?

- Measure of 'goodness': Log Validation Density Loss $\mathcal{L}(\theta)$

$$\mathcal{L}(\theta) = \sum_{i=1}^n \log(y_i | X, y_{-i}, \theta) = -\sum_{i=1}^n \frac{1}{2}\log\sigma_i^2 - \frac{(y_i - \mu_i)^2}{2\sigma_i^2} - \frac{1}{2}\log 2\pi$$

over Leave - One - Out Crossvalidation ($LOO - CV$) and parameters θ

- Time - Efficient Calculation is possible
- Maximize $\mathcal{L}(\theta)$ over θ via standard optimization method (e.g. Gradient Descent)
- Parameters are interpretable

5. Automatic relevance determination(ARD)

- For each input variable x_i introduce scaling parameter l_i
- Optimize l_i in respect to $\mathcal{L}(\theta)$
- A small value of l_i means that variable x_i has no influence and can be eliminated
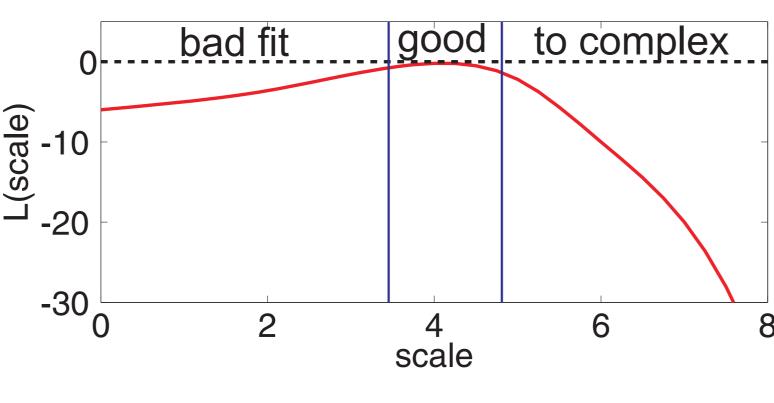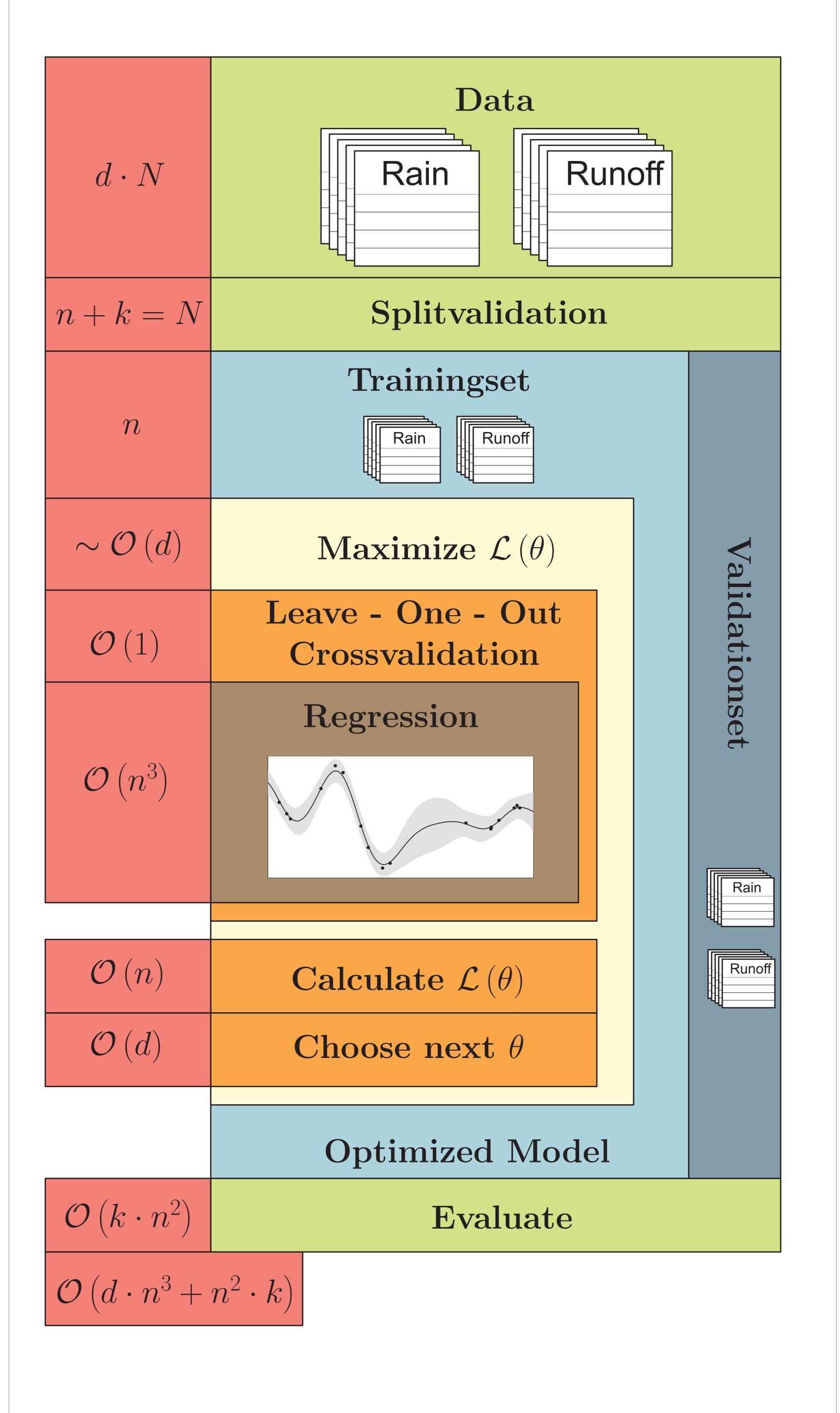- Parameters can easily used to get a sensitivity analysis

Figure 4: Model complexity and probability in comparison: Small values for l_i produce simple models, which can show bad fitting properties. On the other side, models with high length scales are more complex and tend to overfit the data.

6. Procedure

$d \cdot N$	**Data** (Rain, Runoff)
$n + k = N$	**Splitvalidation**
n	**Trainingset** (Rain, Runoff)
$\sim \mathcal{O}(d)$	**Maximize $\mathcal{L}(\theta)$**
$\mathcal{O}(1)$	**Leave - One - Out Crossvalidation**
$\mathcal{O}(n^3)$	**Regression**
$\mathcal{O}(n)$	**Calculate $\mathcal{L}(\theta)$**
$\mathcal{O}(d)$	**Choose next θ**
	Optimized Model
$\mathcal{O}(k \cdot n^2)$	**Evaluate**
$\mathcal{O}(d \cdot n^3 + n^2 \cdot k)$	

Validationset (Rain, Runoff)

7. Results

Catchment Area - Ouse River

Forecast Horizon: 6h $\quad$ Forecast Horizon: 24h

Nash Sutcliffe:

$\quad$ 0.99 $\qquad$ 0.88

Other Catchment Areas

This method was tested in other meso scale catchment areas in south eastern germany with a forecast horizon of 24h. It shows good results in all but one case.

Forecast Horizon: 24h

Catchment Area - Gera $\quad$ Catchment Area - Roda $\quad$ Catchment Area - Wipper

Nash Sutcliffe:

$\quad$ 0.95 $\qquad$ $e_2 : 0.54$ $\qquad$ $e_2 : 0.95$