Multi-output Gaussian process for river water height estimation

Trung Kien Tran, Quoc Long Tran, Emmanuel Vazquez, Merlin Keller, Kaniav Kamary

MULTI-OUTPUT GAUSSIAN PROCESS FOR RIVER WATER HEIGHT ESTIMATION

TRAN Trung Kien
trungkien86@gmail.com
Master student Paris-Saclay University and VNU University of Engineering and Technology Vietnam

TRAN Quoc Long
tqlong@vnu.edu.vn
Doctor, VNU University of Engineering and Technolog France

Merlin KELLER
merlin.keller@edf.fr
Doctor, Électricité de France RD France

Emmanuel VAZQUEZ
emmanuel.vazquez@centralesupelec.fr
Professor, CentraleSupélec, Paris-Saclay University France

Kaniav KAMARY
kaniav.kamary@centralesupelec.fr
Doctor, CentraleSupélec, Paris-Saclay University France

ABSTRACT

This paper aims to improve the Garonne River water height estimation by employing a method that utilizes the Multi-output Gaussian Process with stationary kernels, in conjunction with a Coregionalization model. This approach effectively addresses the challenge by constructing a model based on the data related to Garonne River water height, enabling highly accurate predictions for multiple outputs simultaneously. This methodology proves to be more convenient and superior compared to using a Single-Output Gaussian Process (SOGP) and simplify the process of training, with acceptable time.

KEYWORDS

Flood forecasting, Single-output Gaussian Process, Multi-Output Gaussian Process, Kernel, River Water height estimation

1 INTRODUCTION

Every year, floods disrupt the lives of millions of people, cause significant financial losses. Flood forecasting improve preparedness and response capabilities, reduce economic losses, protect lives and property in the world. Methods used include hydrological models, which simulate the water cycle and river flow, and meteorological models, which predict weather patterns that lead to heavy rainfall and flooding. Other methods using Single-output Gaussian process, mix model, threshold or machine learning. The Garonne River is located in southwest France, originating in the Spanish Pyrenees and flowing into the Atlantic Ocean. EDF (Électricité de France) operates numerous electric power plants situated in close proximity to water sources, serving purposes such as cooling or as primary energy sources, especially in the case of water dams [4]. Protection against floods is one of the main concerns toward ensuring the safety and reliability of its industrial park. The Saint-Vernant model can be thought of as a mathematical formula for water height estimation:

$$\mathbf{h} = f\,(q, \mathbf{K_s}) \tag{1}$$

where $\mathbf{h}$ is the water heights throughout the mesh, over a certain amount of time and depends on :

- controlled variable q : the water discharge at the entry of the river segment.
- a vector $\mathbf{K_s}$ representing 5 uncertain parameters, quantifying the smoothness coefficient of the river bed across 5 subdivisions, assuming homogeneity in terms of regularity for each subdivision.

Figure 1: Garrone river area

In [4], Dr.Kaniav Kamary, Dr. Merlin Keller, et al, used mixtre model to predict water heigh at 2 position Marmande and Mas Agenais.The simulation also used Gaussian Process and runs independently of the two outputs.
In both case, the Matern 3/2 covariance kernel has the best prediction performance, with LOO (leave one out) prediction scores Q^2 at Mas Agenais is 98.8% Q^2 at Marmande is 99.9% .
There are 4 outputs at: Reole, Marmande, Mas Agenais, Tonneins. This prompted us to study a solution for predicting multiple (4) outputs simultaneously.

2 BACKGROUND AND RELATED WORKS

2.1 Gaussian Process

A Gaussian process (GP) is a collection of random variables, any sub set of which have a joint Gaussian distribution. It is represented as:

$$f(\mathbf{x}) \sim \mathcal{GP}(m(\mathbf{x}), \mathcal{K}(\mathbf{x}, \mathbf{x}')) \tag{2}$$

where $m(\mathbf{x})$ is the mean function that characterizes the expected value of $f(\mathbf{x})$, $k(\mathbf{x}, \mathbf{x}')$ is the covariance function that determines the covariance between $f(\mathbf{x})$ and $f(\mathbf{x}')$. The mean function $m(\mathbf{x})$ captures the prior knowledge or expectations about the underlying function, it is often assumed to be 0. The covariance function $k(\mathbf{x}, \mathbf{x}')$ or the kernel function present the correlation between function values. Gaussian processes are widely used in machine learning tasks as regression, classification, and optimization due to their ability to capture uncertainty and make probabilistic predictions.

2.2 Multi output Gaussian Process

The Multi-output Gaussian process (MOGP) is a probabilistic model that extends the concept of a Gaussian process (GP) to predict multiple output variables simultaneously. Let $\mathbf{X} \in \mathbb{R}^{n \times d}$ represent the input matrix, n denotes the number of sample and d represents the input dimensionality. The corresponding output matrix, $\mathbf{Y} \in \mathbb{R}^{n \times m}$, consists of m output variables, $\mathbf{f}(\mathbf{x}) = [f_1(\mathbf{x}), \cdots, f_D(\mathbf{x})]^\top \in \mathbf{R}^D$. $f_d(\mathbf{x})$ denotes the latent function of the d-th output evaluated at $\mathbf{x}$. Function $\mathbf{f}(\cdot)$ is presented as:

$$\mathbf{f}(\mathbf{x}) \sim \mathcal{GP}\left(\mathbf{m}(\mathbf{x}), \mathcal{K}\left(\mathbf{x}, \mathbf{x}'\right)\right) \tag{3}$$

The covariance matrix $\mathcal{K}$:

$$\mathcal{K} = \begin{bmatrix} k_1(\mathbf{X}, \mathbf{X}') & \cdots & k_1(\mathbf{X}, \mathbf{X}') \\ \vdots & \ddots & \vdots \\ k_m(\mathbf{X}, \mathbf{X}') & \cdots & k_m(\mathbf{X}, \mathbf{X}') \end{bmatrix} \tag{4}$$

Here, $\mathbf{m}(\cdot) = \{m_d(\cdot)\}_{d=1}^D$ is the mean function for the d-th output. $\mathcal{K}(\mathbf{x}, \mathbf{x}') \in \mathbf{R}^{D \times D}$ is a positive semi-definite matrix, and $(\mathcal{K}(\mathbf{x}, \mathbf{x}'))_{d,d'} \in \mathbf{R}$ is the covariance between $f_d(\mathbf{x})$ and $f_{d'}(\mathbf{x}')$ where $d, d' \in \{1, \cdots, D\}$.

2.3 Stationary kernel

Stationary kernels are defined as $\mathcal{K}(\boldsymbol{x}, \boldsymbol{x}') = \mathcal{K}(\boldsymbol{r})$, with $\boldsymbol{r} = \boldsymbol{x} - \boldsymbol{x}'$. In other words, the output solely depend on the relative difference between the input values.[6] Some examples of stationary kernels:

Squared exponential kernel (SE). also known as the radial basis kernel (RBF):

$$\mathcal{K}\left(\boldsymbol{x}, \boldsymbol{x}'; \ell\right) = \exp\left(-\frac{\|\boldsymbol{x} - \boldsymbol{x}'\|^2}{2\ell^2}\right) \tag{5}$$

where ℓ is the length-scale of the kernel.

Matérn kernels. [6] In numerous applications, the Matérn kernel is often preferred over the Squared Exponential (SE) kernel. Matérn kernel generates rougher functions that can effectively capture local fluctuations without requiring an excessively small overall length scale

$$\mathcal{K}(r; \nu, \ell) = \frac{2^{1-\nu}}{\Gamma(\nu)} \left(\frac{\sqrt{2\nu} r}{\ell}\right)^\nu K_\nu \left(\frac{\sqrt{2\nu} r}{\ell}\right) \tag{6}$$

K_ν represents a modified Bessel function, and ℓ denotes the length scale.

2.4 Related works

[2] Zvika Ben-Haim, et al have developed a framework for addressing the inertial variant of the Saint-Venant equations, representing the area of focus through a 2D grid where each cell spans approximately 10 m. A trouble with hydraulic modeling is the complex computation in many days. They explored ML algorithm have better in to traditional numerical models as hydraulic model. In [7], Inundation modeling in Google flood forecasting system implemented in Bangladesh 2021. Flood inundation was calculated using the Threshold, Hydraulic and Manifold methods. Threshold approach models the extent of inundation, Manifold models the extent and the height of water. Their experiment showed that the Threshold model and Manifold (a machine-learning method) have better performance than the hydraulic model. [5] A. F. Lopez-Lopera, et al developed a model that integrates inputs varying over time, providing insights into spatially varied inland flood conditions using MOGP. Many stationary kernels were tested: Squared Exponential (SE), Matern 5/2, Matern 3/2. Their experiments utilized datasets with 8 inputs and 20 outputs.

3 METHODOLOGY

3.1 Coregionalization Models

3.1.1 Linear Model of Coregionalization. In LMC [1], the outputs are represented as linear combinations of independent random functions. Given a set of outputs $\{f_d(\mathbf{x})\}_{d=1}^D$ where $\mathbf{x} \in \mathbf{R}^p$, each component f_d takes the form:

$$f_d(\mathbf{x}) = \sum_{q=1}^{Q} a_{d,q} u_q(\mathbf{x}) \tag{7}$$

where $u_q(\mathbf{x})$ is assumed to have zero mean and $\operatorname{cov}\left[u_q(\mathbf{x}), u_{q'}(\mathbf{x}')\right] = k_q(\mathbf{x}, \mathbf{x}')$ if $q = q'$, and $a_{d,q}$ are scalar coefficients, $\left\{u_q(\mathbf{x})\right\}_{q=1}^Q$ are independent for every $q \neq q'$.

$\{f_d(\mathbf{x})\}_{d=1}^D$ can be represented as:

$$f_d(\mathbf{x}) = \sum_{q=1}^{Q} \sum_{i=1}^{R_q} a_{d,q}^i u_q^i(\mathbf{x}) \tag{8}$$

where $u_q^i(\mathbf{x})$, with $q = 1, \ldots, Q$ and $i = 1, \ldots, R_q$, have zero mean and $\operatorname{cov}\left[u_q^i(\mathbf{x}), u_{q'}^{i'}(\mathbf{x}')\right] = k_q(\mathbf{x}, \mathbf{x}')$ if $i = i'$ and $q = q'$.

$$\operatorname{cov}\left[f_d(\mathbf{x}), f_{d'}(\mathbf{x}')\right] = \sum_{q=1}^{Q} \sum_{q'=1}^{Q} \sum_{i=1}^{R_q} \sum_{i'=1}^{R_q} a_{d,q}^i a_{d',q'}^{i'} \operatorname{cov}\left[u_q^i(\mathbf{x}), u_{q'}^{i'}(\mathbf{x}')\right] \tag{9}$$

$\operatorname{cov}[f_d(\mathbf{x}), f_{d'}(\mathbf{x}')]$ is expressed by $(\mathbf{K}(\mathbf{x}, \mathbf{x}'))_{d,d'}$. Because $u_q^i(\mathbf{x})$ are independent functions, the equation formulation can be represented as:

$$\left(\mathbf{K}\left(\mathbf{x},\mathbf{x}'\right)\right)_{d,d'} = \sum_{q=1}^{Q}\sum_{i=1}^{R_q} a_{d,q}^{i} a_{d',q}^{i} k_q\left(\mathbf{x},\mathbf{x}'\right) = \sum_{q=1}^{Q} b_{d,d'}^{q} k_q\left(\mathbf{x},\mathbf{x}'\right) \tag{10}$$

with $b_{d,d'}^{q} = \sum_{i=1}^{R_q} a_{d,q}^{i} a_{d',q}^{i}$. $\mathbf{K}\left(\mathbf{x},\mathbf{x}'\right)$ is represented as:

$$\mathbf{K}\left(\mathbf{x},\mathbf{x}'\right) = \sum_{q=1}^{Q} \mathbf{B}_q k_q\left(\mathbf{x},\mathbf{x}'\right) \tag{11}$$

where each $\mathbf{B}_q \in \mathbf{R}^{D\times D}$ is a coregionalization matrix. The rank matrix $\mathbf{B}_q$ is R_q.

3.1.2 Intrinsic Coregionalization Model. Intrinsic coregionalization model (ICM) is a specific case of the LMC [1], assumes that element $b_{d,d'}^{q}$ of the coregionalization matrix $\mathbf{B}_q$ can be represented as $b_{d,d'}^{q} = v_{d,d'} b_q$.

$$\begin{aligned}\operatorname{cov}\left[f_d(\mathbf{x}), f_{d'}\left(\mathbf{x}'\right)\right] &= \sum_{q=1}^{Q} v_{d,d'} b_q k_q\left(\mathbf{x},\mathbf{x}'\right), = v_{d,d'} \sum_{q=1}^{Q} b_q k_q\left(\mathbf{x},\mathbf{x}'\right) \\ &= v_{d,d'} k\left(\mathbf{x},\mathbf{x}'\right)\end{aligned} \tag{12}$$

where $k\left(\mathbf{x},\mathbf{x}'\right) = \sum_{q=1}^{Q} b_q k_q\left(\mathbf{x},\mathbf{x}'\right)$. This kernel is LMC kernel when $Q = 1$.

The coefficients $v_{d,d'} = \sum_{i=1}^{R_1} a_{d,1}^{i} a_{d',1}^{i} = b_{d,d'}^{1}$, the kernel matrix $\mathbf{K}\left(\mathbf{x},\mathbf{x}'\right) = k\left(\mathbf{x},\mathbf{x}'\right)\mathbf{B}$.

We have the formulation:

$$\mathbf{K}(\mathbf{X},\mathbf{X}) = \mathbf{B} \otimes k(\mathbf{X},\mathbf{X}) \tag{13}$$

3.2 Performance indicator

we use Q-squared (Q^2):

$$Q^2 = 1 - \frac{\sum_{i=1}^{N_{\text{test}}} \left(y_i - \widehat{y}_i\right)^2}{\sum_{i=1}^{N_{\text{test}}} \left(y_i - \bar{y}\right)^2} \tag{14}$$

where $\widehat{y}_1, \ldots, \widehat{y}_{N_{\text{test}}}$ represent the predictions, and $\bar{y}$ is the average of the test data $y_1, \ldots, y_{N_{\text{test}}}$. In the case of noise-free observations, Q^2 equals one when predictions exactly match the test data, zero when they match $\bar{y}$, and it becomes negative when they perform worse than $\bar{y}$.

3.3 Data set

The DOE dataset provided by EDF R&D contains 767 records with 5 input fields: Q,KS1,KS2,KS3,KS4,KS5 and 6 output field are water height at Reole, Marmande, Mas Agenais, Tonneis at stationary state.

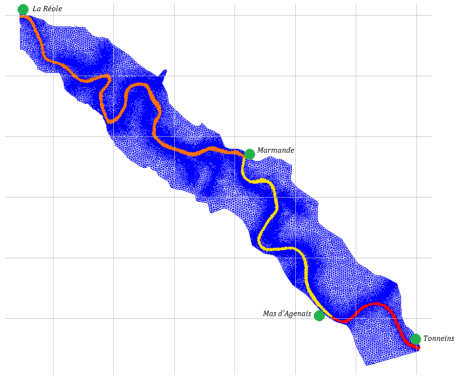

Figure 2: The Garrone river output positions

Preprocess. We sorted data by value in Q column, from smallest to largest value, and using sorted data for the input of program.

the value range of Q is much larger than other input data fields, so we regularize $X_1 = Q/100$ for a range from from 5.07 to 69.9.

- input X: (X_1 = Q/100,$X_2 = K_{S2}$,X_3=K_{S3},$X4$=K_{S4},X_5=K_{S5})
- output Y: (Y_1 = REOLE, Y_2 = MARMANDE, Y_3 = MAS AGENAIS,Y_4 = TONNEINS)

We add an extra column to our training dataset that contains an index that specifies which output is observed.

$$\mathrm{X}_{augmented} = \begin{bmatrix} X_1 & X_2 & X_3 & X_4 & X_5 & X_6 & 0 \\ X_1 & X_2 & X_3 & X_4 & X_5 & X_6 & 1 \\ X_1 & X_2 & X_3 & X_4 & X_5 & X_6 & 2 \\ X_1 & X_2 & X_3 & X_4 & X_5 & X_6 & 3 \end{bmatrix}$$

$$\mathrm{Y}_{augmented} = \begin{bmatrix} Y_1 & 0 \\ Y_2 & 1 \\ Y_3 & 2 \\ Y_4 & 3 \end{bmatrix}$$

X_i is the column vector of input field ith, Y_i is the column vector of output field number i.

3.4 Build the Intrinsic Coregionalization Model (ICM)

Because the ICM model is better than LMC in calculation resource saving and time consuming. We prefer to try the IMC [3] [1] first for Garonne river water height estimation at 4 positions: Reole, Marmande, Mas Agenais, Tonneins.

4 EXPERIMENT

We use Google Colab Pro+ with a GPU A100.

Train set and test set are divided in 3 ways, corresponding to 3 experiments:

- Experiment 1: the first 750/767 records for train set, the last 17/767 records for test set
- Experiment 2: randomly split 750/767 records for train set, 17/767 records for test set.
- Experiment 3: randomly split 80% records for train set, 20% record for test set.

Kernels and Model training: We tried kernels: Martern 5/2, Martern3/2, Martern1/2, Squared Exponential as base kernel and Coregionalization Model (ICM) for modelling. The models was trained in 15000 iterations.

Experiment result: Experimental results show that using the Intrinsic Coregionalization Model (ICM) is appropriate and gives good results for Garonne river flood water height estimation. Training time is acceptable, and the trained model's checkpoint can be loaded and used conveniently for prediction. All of 3 experiments, 95% prediction interval of the data points intersects with diagonal lines, so all of the predictive results in test data points is good. In the experiment 3, Q^2 with the Coregionalization model based on Matern 5/2 and the Coregionalization model based on Matern 5/2 are 99.9%Ẇith the Coregionalization model based on Matern 1/2 and the Coregionalization model based on Squared Exponential are 99.89 %Ċoregionalization model (ICM) based on Matern 5/2 and Matern 3/2 are quite better than the remaining 2 models.

No.	Matern 5/2 (h)	Matern 3/2 (h)	Matern 1/2 (h)	SE (h)
1	2.7	2.75	2.56	2.33
2	2.017	2.083	2.383	2.483
3	1.167	1.35	1.3	1.2

Table 1: Tranining time(h) for each base kernel in experiments

No.	Matern 5/2	Matern 3/2	Matern 1/2	SE
1	99.990053%	99.9911378%	99.950015%	99.92817%
2	99.88883%	99.89028%	99.89877%	99.86038%
3	99.9009%	99.9056%	99.89674%	99.88666%

Table 2: Q^2 for each base kernel in experiments

5 CONCLUSION

This paper has applied Multi-output Gaussian process (MOGP) for the flooding forecast problem in the Garonne river. It has addressed the challenge of estimating water heights in the Garonne river, specifically at Reole, Marmande, Mas Agenais, and Tonneins simultaneously. This was achieved through a novel approach using MOGP and a Coregionalization kernel model with stationary base kernels, resulting in strong prediction performance for multi-output simultaneously. In the experiments, our models performed exceptionally well, achieving Q2 = 99.9% on the test set. These experiments demonstrated that our method offers a comprehensive solution and represents an improvement over the baseline solution in [4]. It offers an efficient method for river water height estimation and flood forecasting problem.

REFERENCES

[1] Mauricio A. Alvarez, Lorenzo Rosasco, and Neil D. Lawrence. *Kernels for Vector-Valued Functions: a Review*. 2012. arXiv: 1106.6251 [stat.ML].

[2] Zvika Ben-Haim et al. *Inundation Modeling in Data Scarce Regions*. 2019. arXiv: 1910.05006 [cs.LG].

[3] Edwin V Bonilla, Kian Chai, and Christopher Williams. "Multi-task Gaussian Process Prediction". In: *Advances in Neural Information Processing Systems*. Ed. by J. Platt et al. Vol. 20. Curran Associates, Inc., 2007. URL: https://proceedings.neurips.cc/paper_files/paper/2007/file/66368270ffd51418ec58bd793f2d9b1b-Paper.pdf.

[4] Kaniav Kamary et al. *Computer code validation via mixture model estimation*. 2019. arXiv: 1903.03387 [stat.ME].

[5] Andrés F. López-Lopera et al. "Multioutput Gaussian processes with functional data: A study on coastal flood hazard assessment". In: *Reliability Engineering amp; System Safety* 218 (Feb. 2022), p. 108139. ISSN: 0951-8320. DOI: 10.1016/j.ress.2021.108139. URL: http://dx.doi.org/10.1016/j.ress.2021.108139.

[6] Kevin P. Murphy. *Probabilistic Machine Learning: Advanced Topics*. MIT Press, 2023. URL: http://probml.github.io/book2.

[7] Sella Nevo et al. *Flood forecasting with machine learning models in an operational framework*. 2021. arXiv: 2111.02780 [cs.LG].