Stochastic Modeling of Multiple Streamflow Time Series in Colombian Based on Gaussian Processes

Author: Julián David Pastrana-Cortés

Director: Álvaro Angel Orozco-Gutiérrez

Co-director: David Augusto Cardenas-Peña

Automatic Research Group

August 26, 2024

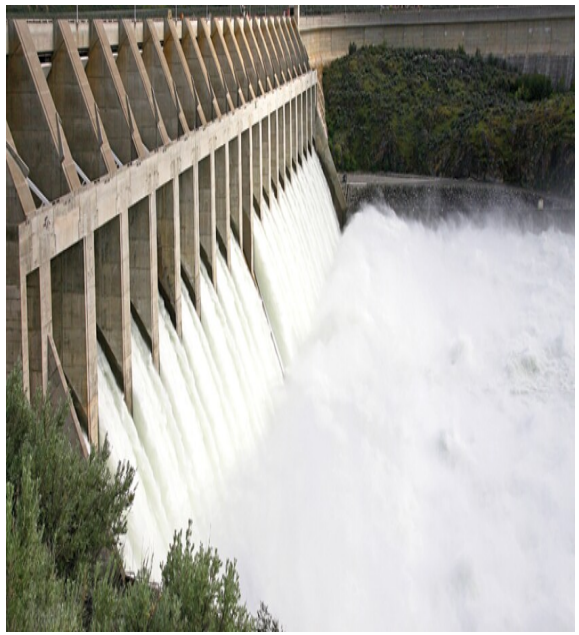
Introduction

Motivation

Understanding the implications of time series associated with hydrological variables, such as flow rates or reservoir levels, is essential for hydroelectric generation and the planning of other generation systems in Colombia

(a) Irrigation

(b) Flood control



(c) Hydropower generation

Challenges: non-linearities, high stochasticity, and complex water resource patterns.

The Importance of Hydrological Forecasting

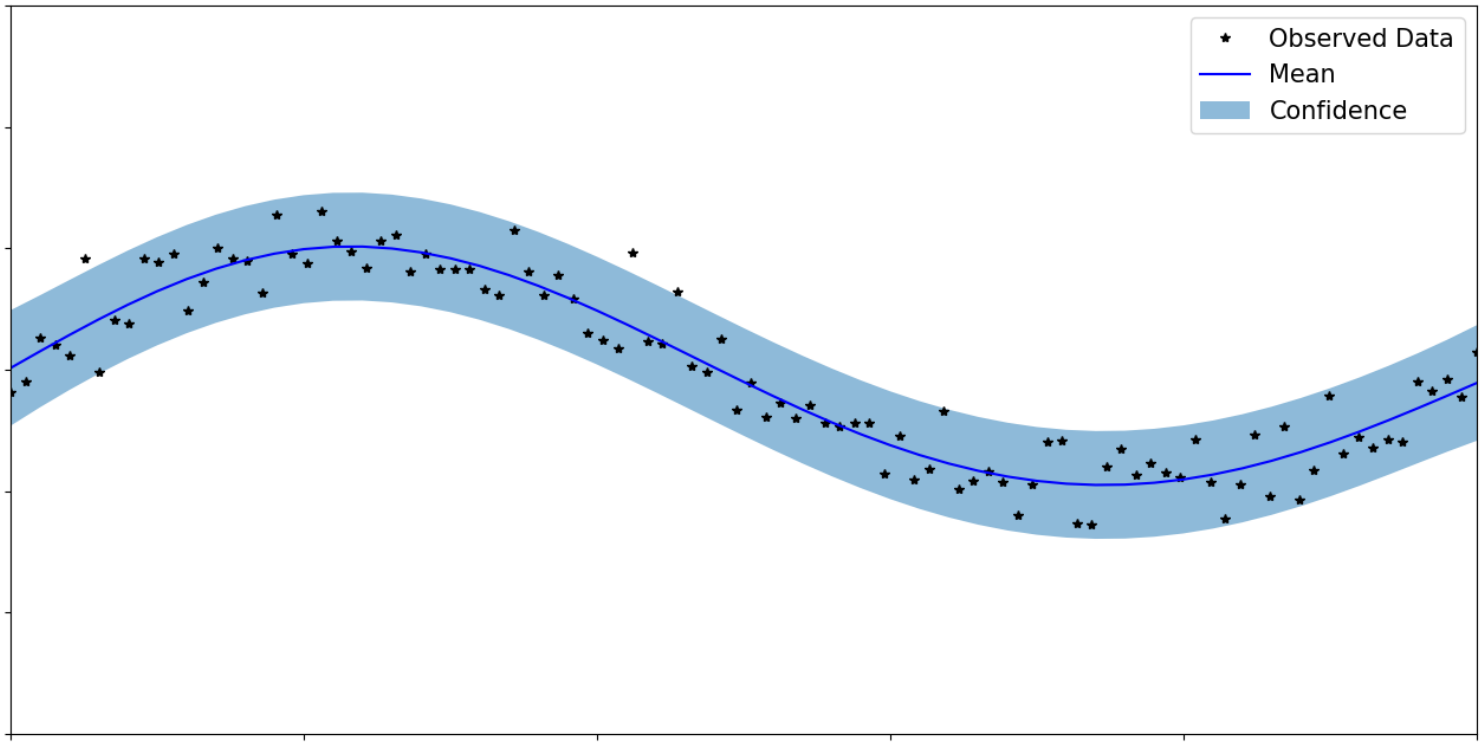
Understanding hydrological processes has become increasingly critical in the field of natural resource management, anticipation capacity of extreme hydrological events such as droughts and heavy rainfall.

(a) Drought Condition

(b) Full Dam

The Proposed Model

A Gaussian Process (GP) is a Bayesian non-parametric model that provides not just point predictions but a full probability distribution, capturing uncertainty and enabling confidence intervals in time series forecasting.



Objectives

General Objective

Develop a stochastic forecasting model for making multiple simultaneous predictions of hydrological time series, taking advantage of cross-correlations among the tasks to improve the performance keeping the scalability in its implementation for the short-term horizon.

Specific Objectives

- Develop a model that allows the forecasting of hydrological time series, properly quantifying the uncertainty associated with each value within the prediction horizons.
- Design a multi-task forecasting methodology that captures and models cross-correlations between hydrological time series, to improve forecast accuracy within forecast horizons.
- Develop a multi-task prediction methodology that handles data constraints across reservoirs while maintaining high forecasting performance as measured by probabilistic metrics.

The Dataset

Problem Setting - Part 1

Consider a time-series vector of hydrological resources observed across all D outputs at the n-th time instant, denoted as $\mathbf{v}_n \in \mathbb{R}^D$. Our model employs the entire sequence of resource vectors from time n back to $n - T + 1$ as input to predict the resource vector at the future time step $n + H$. Here, T represents the model order, and H denotes the prediction horizon.
Consequently, we define the input vector $\mathbf{x}_n$ as follows:

$$\mathbf{x}_n = \begin{bmatrix} \mathbf{v}_n^\top \\ \mathbf{v}_{n-1}^\top \\ \vdots \\ \mathbf{v}_{n-T+1}^\top \end{bmatrix} \in \mathbb{R}^{DT}$$

Problem Setting - Part 2

The target output vector $\mathbf{y}_n$ is defined as follows:

$$\mathbf{y}_n = \mathbf{v}_{n+H}.$$

This formulation enables the model to leverage historical hydrological data for accurate future predictions. Accordingly, we build a dataset $\mathcal{D} = \{\boldsymbol{x}_n, \boldsymbol{y}_n\}_{n=1}^{N} = \{\mathbf{X}, \mathbf{y}\}$ comprising N input–output pairs, where $\boldsymbol{x}_n \in \mathcal{X} \subset \mathbb{R}^{DT}$ represents the dimensional input space. The target function is a vector-valued function where $\boldsymbol{y}_n \in \mathbb{R}^{D}$ comprises the observations of all outputs (tasks) at the same input $\boldsymbol{x}_n$.

The hydrological forecasting task for validating the proposed model regressors considers time-series streamflow contributions data from 23 Colombian reservoirs.

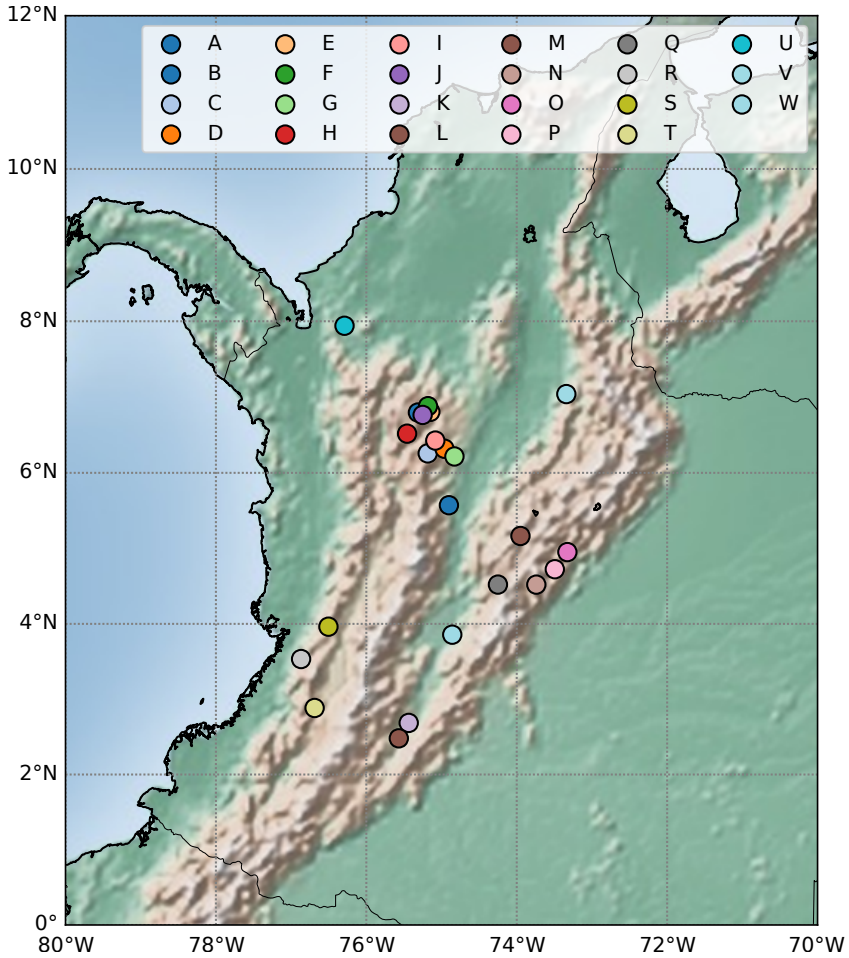


Figure: Reservoir locations in Colombia.

The dataset was selected due to the strong dependence of time series on weather patterns, which are closely tied to hydropower dispatch operations.

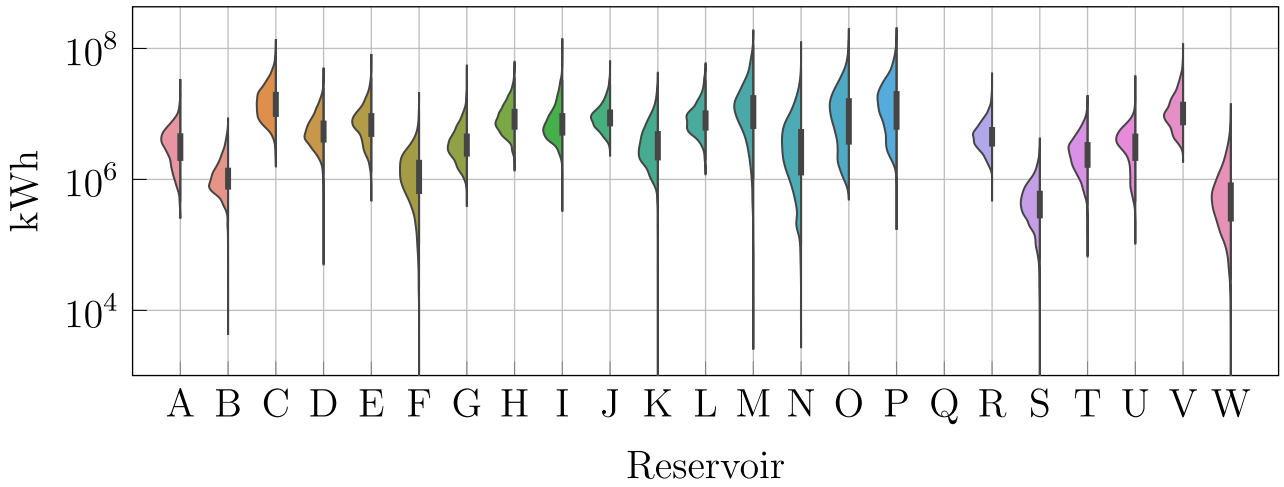


Figure: Time-series Violin plot depicting the streamflow contribution for each reservoir within the dataset.

These streamflow contributions were recorded daily from 1 January 2010 to 28 February 2022. While these contributions represent volumetric values, they are reported in kilowatt-hours (kWh) by the hydroelectric power plants.

Gaussian Process Regression

In the context of the GP framework, the dataset $\mathcal{D}$ supports the learning of a random mapping function $\boldsymbol{f}(\cdot)$ that captures the relationship between $\boldsymbol{x}_n$ and $\boldsymbol{y}_n$. Thus, we pose a GP distribution over $\boldsymbol{f}(\cdot)$ as follows:

$$\boldsymbol{f}(\boldsymbol{x}) \sim \mathcal{GP}\left(\mathbf{0},\, \boldsymbol{k}(\boldsymbol{x}, \boldsymbol{x}' \mid \boldsymbol{\theta})\right) \tag{1}$$

where $\boldsymbol{f} : \mathcal{X} \to \mathbb{R}^D$ represents the vector-valued function responsible for mapping the input space, and $\boldsymbol{k} : \mathcal{X} \times \mathcal{X} \to \mathbb{R}^{D \times D}$ symbolizes the cross-covariance matrix function, parameterized by vector $\boldsymbol{\theta}$. Adding i.i.d zero mean Gaussian noise to each task $\boldsymbol{\epsilon}$ with diagonal covariance matrix $\Sigma_\epsilon = \text{diag}\{\sigma_{Nd}^2\}_{d=1}^D \in \mathbb{R}^{D \times D}$ as a regularization term, the model prediction given by Equation (1) turns into $\boldsymbol{y}_n = \boldsymbol{f}(\boldsymbol{x}_n) + \boldsymbol{\epsilon}$, corresponding to a noise observation of the latent variable $\boldsymbol{f}$.