# 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 28, 2024

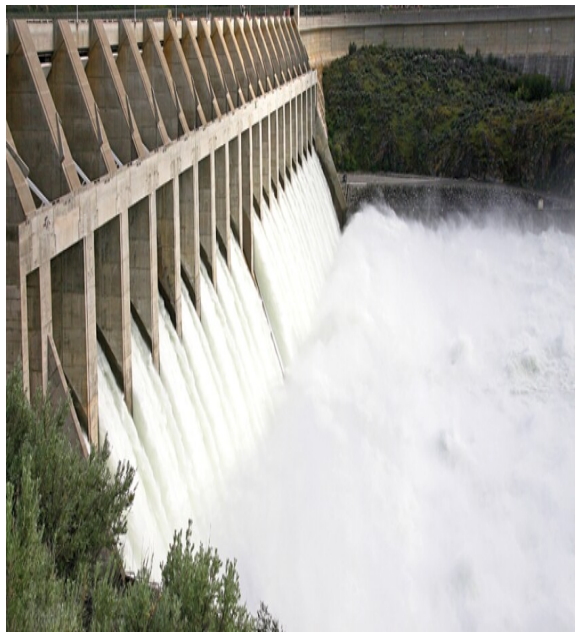

# Introduction

# 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

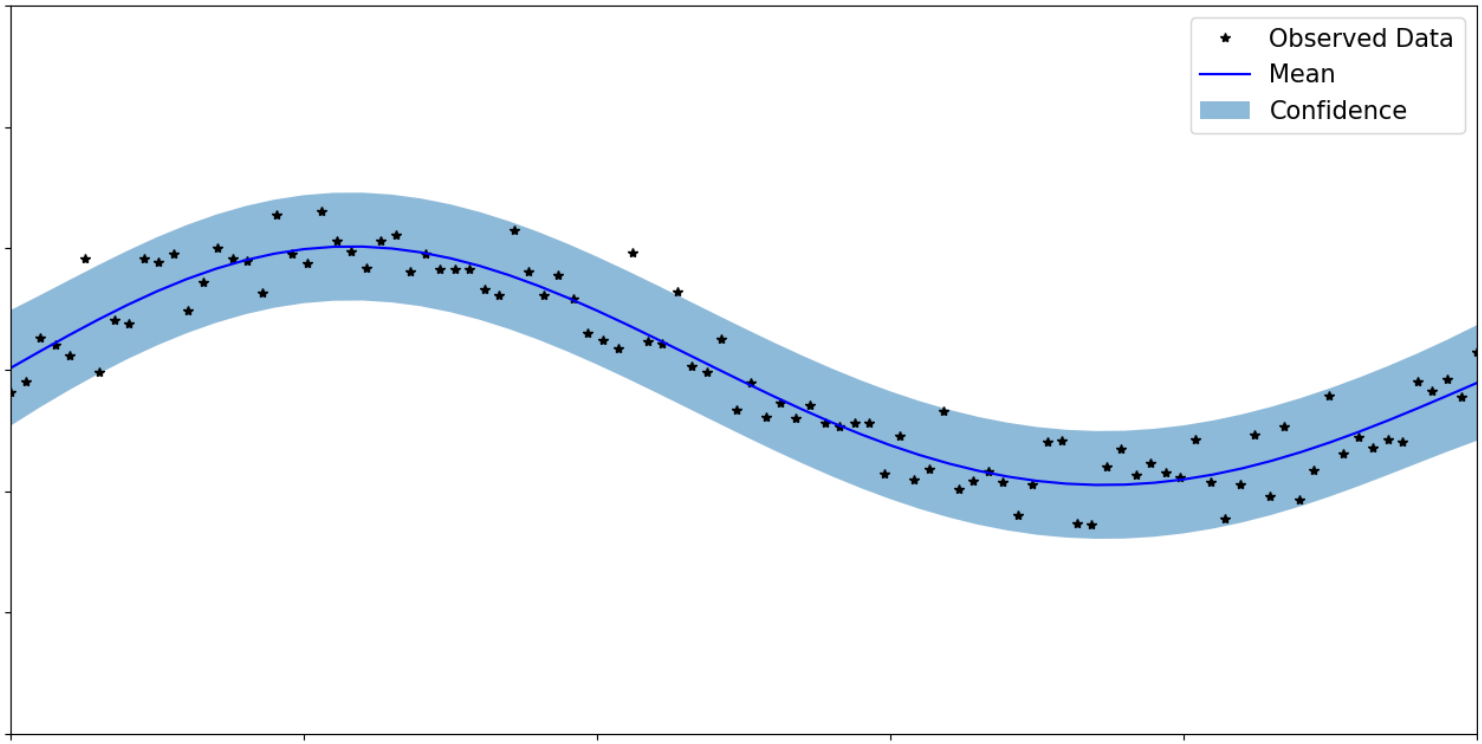
Understanding hydrological processes has become increasingly critical in the field of natural resource management, anticipation capacity of extreme hydrological events such as droughts and heavy rainfall.

(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.

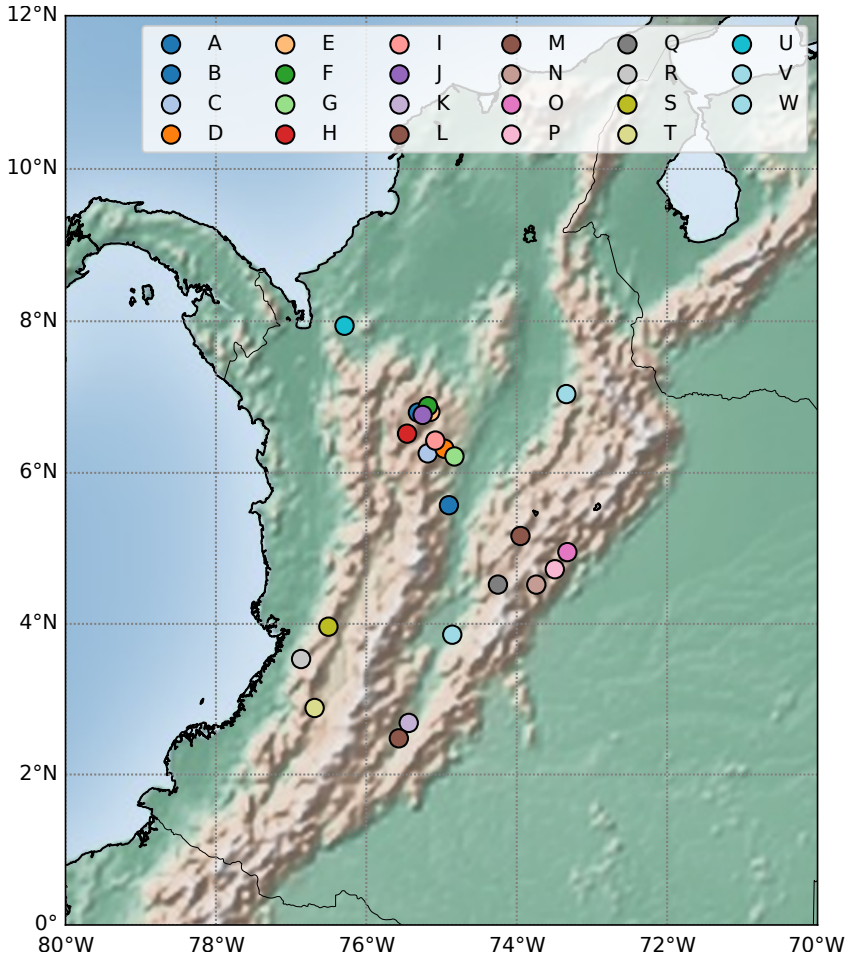



Figure: Reservoir locations in Colombia.

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

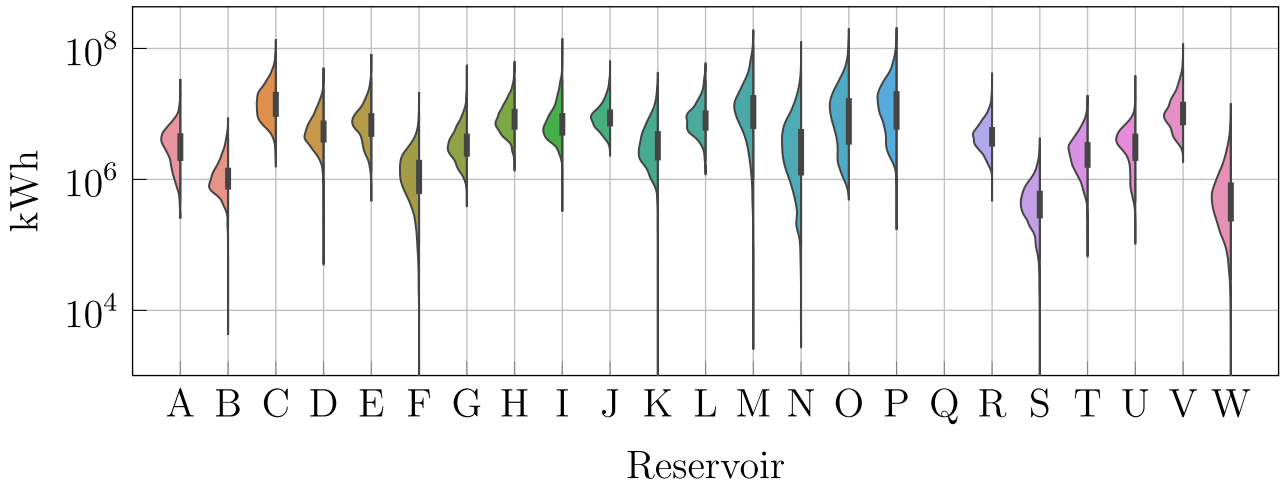



Figure: Time-series Violin plot depicting the streamflow contribution for each reservoir within the dataset.

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 GP framework, the dataset $\mathcal{D}$ is used to learn a random mapping function $\boldsymbol{f}(\cdot)$, capturing the relationship between $\boldsymbol{x}_n$ and $\boldsymbol{y}_n$. We define a GP distribution over $\boldsymbol{f}(\cdot)$ as:

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

where $\boldsymbol{f} : \mathcal{X} \to \mathbb{R}^D$ maps the input space, and $\boldsymbol{k} : \mathcal{X} \times \mathcal{X} \to \mathbb{R}^{D \times D}$ is the cross-covariance matrix function, parameterized by $\boldsymbol{\theta}$. Adding i.i.d Gaussian noise $\boldsymbol{\epsilon}$ with covariance $\Sigma_\epsilon = \text{diag}\{\sigma_{Nd}^2\}_{d=1}^D$, the model prediction becomes $\boldsymbol{y}_n = \boldsymbol{f}(\boldsymbol{x}_n) + \boldsymbol{\epsilon}$.

# Joint Distribution

Let $\mathbf{X}_* = \{\boldsymbol{x}_{n*}\}_{n=1}^{N_*}$ be a set of test points $\boldsymbol{x}_{n*} \in \mathcal{X}$ with test output vector $\mathbf{f}_* = [\boldsymbol{f}(\boldsymbol{x}_1)^\top, \boldsymbol{f}(\boldsymbol{x}_2)^\top, \cdots, \boldsymbol{f}(\boldsymbol{x}_{N_*})^\top]^\top \in \mathbb{R}^{N_* D}$. The joint Gaussian distribution over $\mathbf{y}$ and $\mathbf{f}_*$ is given by:

$$\begin{bmatrix} \mathbf{y} \\ \mathbf{f}_* \end{bmatrix} \sim \mathcal{N}\left(\mathbf{0}, \begin{bmatrix} \mathbf{K}_y & \mathbf{K}_* \\ \mathbf{K}_*^\top & \mathbf{K}_{**} \end{bmatrix}\right) \tag{2}$$

Here, $\mathbf{K}_y = \mathbf{K} + \Sigma_\epsilon$ with $\mathbf{K} \in \mathbb{R}^{ND \times ND}$, formed by evaluating the covariance function at all pairs in $\mathbf{X}$. The matrices $\mathbf{K}_{**} \in \mathbb{R}^{N_* D \times N_* D}$ and $\mathbf{K}_* \in \mathbb{R}^{ND \times N_* D}$ are formed by evaluating the covariance function across test inputs in $\mathbf{X}_*$ and test-train inputs, respectively.

# Posterior Distribution

This notation allows for deriving the conditional distribution named posterior as follows:

$$\mathbf{f}_*|\mathbf{X}_*, \mathcal{D} \sim \mathcal{N}(\bar{\mathbf{f}_*}, \text{cov}(\mathbf{f}_*)) \tag{3}$$

with

$$\bar{\mathbf{f}_*} = \mathbf{K}_*^\top \mathbf{K}_y^{-1} \mathbf{y} \tag{4}$$

$$\text{cov}(\mathbf{f}_*) = \mathbf{K}_{**} - \mathbf{K}_*^\top \mathbf{K}_y^{-1} \mathbf{K}_* \tag{5}$$

# The Marginal Log-likelihood

Furthermore, the prediction performance archived by the conditional distribution depends on the selected parameter set $\boldsymbol{\theta}$ and observation noise matrix $\Sigma_\epsilon$. Both parameters are calculated by maximizing the marginal log-likelihood from Equation (2) where $p(\mathbf{y}) = \mathcal{N}(\mathbf{y} \mid \mathbf{0}, \mathbf{K}_y)$ as follows:

$$\begin{aligned} \{\boldsymbol{\theta}_{\text{opt}}, \Sigma_{\epsilon\text{opt}}\} &= \arg\max_{\boldsymbol{\theta}, \Sigma_\epsilon} \quad \ln p(\mathbf{y}) \\ &= \arg\min_{\boldsymbol{\theta}, \Sigma_\epsilon} \quad \frac{1}{2}\mathbf{y}^\top \mathbf{K}_y^{-1}\mathbf{y} + \frac{1}{2}\ln \mid \mathbf{K}_y \mid + \frac{ND}{2}\ln 2\pi \end{aligned} \tag{6}$$

The main challenge is their computational complexity of $\mathcal{O}(N^3D^3)$ and storage demand of $\mathcal{O}(N^2D^2)$ due to inverting the matrix $\mathbf{K}_y$.

# Variational Inference

A common approach to reduce GP complexity, consists of the introduction of a reduced set of $M \ll N$ inducing point locations, denoted by $Z$, containing vectors $\mathbf{z} \in \mathcal{X}$, and an inducing variable vector $\mathbf{u} = [\boldsymbol{f}(\boldsymbol{z}_1)^\top, \boldsymbol{f}(\boldsymbol{z}_2)^\top, \cdots, \boldsymbol{f}(\boldsymbol{z}_M)^\top]^\top \in \mathbb{R}^{MD}$. This approach leads to the following extended joint distribution:

$$\begin{bmatrix} \mathbf{u} \\ \mathbf{f} \end{bmatrix} \sim \mathcal{N}\left( \mathbf{0} \ , \begin{bmatrix} \mathbf{K}_{uu} & \mathbf{K}_{uf} \\ \mathbf{K}_{uf}^\top & \mathbf{K} \end{bmatrix} \right) \tag{7}$$

Here, $\mathbf{K}_{uu} \in \mathbb{R}^{MD \times MD}$ and $\mathbf{K}_{uf} \in \mathbb{R}^{MD \times ND}$ represent the block covariance matrices evaluated at the inducing point pairs and between the inducing points and training inputs, respectively. We call this model the Sparse Variational Gaussian Process (SVGP).

# The Variational Distribution

From the above, the conditional distribution is given by:

$$p(\mathbf{f} \mid \mathbf{u}) = \mathcal{N}\left(\mathbf{f} \mid \mathbf{K}_{uf}^{\top}\mathbf{K}_{uu}^{-1}\mathbf{u}, \mathbf{K} - \mathbf{K}_{uf}^{\top}\mathbf{K}_{uu}^{-1}\mathbf{K}_{uf}\right), \tag{8}$$

and the prior for the inducing variables is:

$$p(\mathbf{u}) = \mathcal{N}\left(\mathbf{u} \mid \mathbf{0}, \mathbf{K}_{uu}\right). \tag{9}$$

The joint posterior distribution over the latent variable $\mathbf{f}$ and inducing variable $\mathbf{u}$ is:

$$p(\mathbf{f}, \mathbf{u} \mid \mathbf{y}) = p(\mathbf{f} \mid \mathbf{u})p(\mathbf{u} \mid \mathbf{y}). \tag{10}$$

However, computing $p(\mathbf{u} \mid \mathbf{y})$ is typically intractable, so we use variational inference to approximate the posterior with a variational distribution $q(\mathbf{u}) = \mathcal{N}\left(\mathbf{u} \mid \boldsymbol{\mu}, \boldsymbol{S}\right)$, where $\boldsymbol{\mu} \in \mathbb{R}^{MD}$ and $\boldsymbol{S} \in \mathbb{R}^{MD \times MD}$. The approximate joint posterior then becomes:

$$p(\mathbf{f}, \mathbf{u} \mid \mathbf{y}) \approx q(\mathbf{f}, \mathbf{u}) = p(\mathbf{f} \mid \mathbf{u})q(\mathbf{u}). \tag{11}$$

# ELBO

Accordingly, the approximation of the posterior distribution consists of estimating the variational parameters $\boldsymbol{\mu}$ and $\boldsymbol{S}$, performed by maximizing an evidence lower bound (ELBO). Such ELBO is obtained from the log marginal likelihood:

$$\ln p(\mathbf{y}) \geq \int \int q(\mathbf{f}, \mathbf{u}) \ln \frac{p(\mathbf{y} \mid \mathbf{f}) p(\mathbf{f} \mid \mathbf{u}) p(\mathbf{u})}{q(\mathbf{f}, \mathbf{u})} d\mathbf{f} d\mathbf{u} = \mathcal{L} \tag{12}$$

resulting from the Jensen's inequality. We can rewrite the previous expression to derive the following:

$$\mathcal{L} = \mathbb{E}_{q(\mathbf{f})}\{\ln p(\mathbf{y} \mid \mathbf{f})\} - \mathsf{KL}\{q(\mathbf{u}) \parallel p(\mathbf{u})\} \tag{13}$$

Here, KL represents the Kullback–Leibler divergence between the Gaussian-shaped distributions.

# ELBO

With the i.i.d assumption, the likelihood function $p(\mathbf{y} \mid \mathbf{f})$ can be factorized over observations and outputs:

$$p(\mathbf{y} \mid \mathbf{f}) = \prod_{d=1}^{D} \prod_{n=1}^{N} p(y_{dn} \mid f_d(\mathbf{x}_n)), \tag{14}$$

where $f_d(\mathbf{x}_n)$ represents the latent function value for the $d$-th output at the $n$-th input, and $y_{dn}$ is the corresponding label. The lower bound can then be expressed as:

$$\mathcal{L} = \sum_{d=1}^{D} \sum_{n=1}^{N} \mathbb{E}_{q(f_d(\mathbf{x}_n))}\{\ln p(y_{dn} \mid f_d(\mathbf{x}_n))\} - \sum_{d=1}^{D} \mathrm{KL}\{q(\mathbf{u}_d) \parallel p(\mathbf{u}_d)\}. \tag{15}$$

The summation over $D$ and $N$ facilitates training through a mini-batch fashion.

# Posterior and Predictive Distribution

We define the variational distribution for $\mathbf{f}$ as:

$$q(\mathbf{f}) := \int p(\mathbf{f} \mid \mathbf{u}) q(\mathbf{u}) d\mathbf{u}$$
$$= \mathcal{N}\left(\mathbf{f} \mid \mathbf{K}_{uf}^{\top} \mathbf{K}_{uu}^{-1} \boldsymbol{\mu}, \mathbf{K} + \mathbf{K}_{uf}^{\top} \mathbf{K}_{uu}^{-1} (\mathbf{S} - \mathbf{K}_{uu}) \mathbf{K}_{uu}^{-1} \mathbf{K}_{uf}\right). \tag{16}$$

This approach reduces complexity to $\mathcal{O}(NM^2D^3)$, as $\mathbf{K}_{uu}$ is smaller than $\mathbf{K}_y$. For predictions at new points $\mathbf{x}_*$, we compute:

$$p(\mathbf{f}_* \mid \mathbf{y}) \approx q(\mathbf{f}_*) = \int p(\mathbf{f}_* \mid \mathbf{u}) q(\mathbf{u}) d\mathbf{u}, \tag{17}$$

and adding Gaussian noise $\Sigma_\epsilon$ to the latent posterior $\mathbf{f}_* \mid \mathbf{y}$.

# Model Setup

The proposed approach factorizes the GP scalar covariance function into two kernels: $k_{\mathcal{X}}$, which models input correlations via relative distances, and $k_D$, which models task pair-wise correlations as follows:

$$k\left((\boldsymbol{x}, d), (\boldsymbol{x}', d') \mid \boldsymbol{\theta}\right) = k_{\mathcal{X}}\left(\boldsymbol{x}, \boldsymbol{x}' \mid \boldsymbol{\Theta}_d\right) k_D\left(d, d' \mid \sigma_d\right), \tag{18}$$

$$k_{\mathcal{X}}\left(\boldsymbol{x}, \boldsymbol{x}' \mid \boldsymbol{\Theta}_d\right) = \exp\left(-\frac{1}{2}(\boldsymbol{x} - \boldsymbol{x}')^{\top} \boldsymbol{\Theta}_d^{-2} (\boldsymbol{x} - \boldsymbol{x}')\right), \tag{19}$$

$$k_D\left(d, d' \mid \sigma_d^2\right) = \sigma_d^2 \, \delta_{d,d'}, \tag{20}$$

where $\delta_{d,d'}$ is the Kronecker delta, $\sigma_d^2$ is the output scale for task $d$, and $\boldsymbol{\Theta}_d$ is a diagonal matrix of lengthscale factors. The input kernel $k_{\mathcal{X}}$ uses a squared-exponential function, ensuring smooth data mapping. The task kernel $k_D$ avoids modeling task correlations directly, reducing complexity to $\mathcal{O}(NM^2D)$. Despite not modeling task correlations explicitly, the autoregressive inputs and shared inducing points still enable exploration of task dependencies.

To evaluate our model's performance, we use three metrics: Mean Squared Error (MSE), Mean Standardized Log Loss (MSLL), and Continuous Ranked Probability Score (CRPS).
MSE measures the average squared difference between the observed outcome $y_{dn}$ and the predicted expectation $\mathbb{E}\{y_{dn*}\}$:

$$\text{MSE} = \frac{1}{DN_*} \sum_{d=1}^{D} \sum_{n=1}^{N_*} (y_{dn} - \mathbb{E}\{y_{dn*}\})^2 \tag{21}$$

MSLL assesses the quality of probabilistic predictions by considering the log-likelihood of observed values:

$$\text{MSLL} = \frac{1}{2DN_*} \sum_{d=1}^{D} \sum_{n=1}^{N_*} \left( \frac{(y_{dn} - \mathbb{E}\{y_{dn*}\})^2}{\text{var}\{y_{dn*}\}} - \frac{(y_{dn} - \mu_d)^2}{\sigma_d^2} + \ln \left( \frac{\text{var}\{y_{dn*}\}}{\sigma_d^2} \right) \right) \tag{22}$$

CRPS evaluates probabilistic predictions by comparing the predicted CDF $\Phi(y_{dn})$ with the empirical CDF:

$$\text{CRPS} = \frac{1}{DN_*} \sum_{d=1}^{D} \sum_{n=1}^{N_*} \sqrt{\text{var}\{y_{dn*}\}} \left( \beta_{dn} \left( 2\Phi\left(\beta_{dn}\right) - 1 \right) + 2\phi\left(\beta_{dn}\right) - \frac{1}{\sqrt{\pi}} \right) \tag{23}$$

where $\beta_{dn} = \frac{y_{dn} - \mathbb{E}\{y_{dn*}\}}{}$