GAUSSIAN PROCESSES FOR TIME SERIES
with application to air pollution forecasting

Raphaela Azar & Sbonelo Gumede

Minor dissertation presented in partial fullfilment of the requirements for the degree of Honours in Statistics & Data Science

Department of Statistical Sciences
University of Cape Town

Supervised by
Dr Birgit Erni

Abstract

In this paper we give a basic introduction into Gaussian Processes. We delve into how can you use a Gaussian process for time series analysis. Compare the peformance of Gaussian processes to other time series models. An area of application included is predicting the level of air pollution in the Table View station in Cape Town for the year of 2019. Another area of application is predicting the gold price for South Africa for the year of 2020. This is done in a Bayesian framework in order to make probabilistic statements and obtain a predictive distribution for future forecasts.

Literature Review

((Rasmussen and Williams, 2006)) We have a training set of $\mathcal{D}$ of n observations, $\mathcal{D} = \{(\mathbf{x}_i, y_i) \,|\, i = 1, \ldots, n\}$, where $\mathbf{x}$ denotes the input vector (covariates) of dimension $\mathcal{D}$ and y denotes a scalar output or target (dependent variable); the column vector of inputs for all n cases are arranged in the $\mathcal{D} \times n\, design\, matrix\, X$, and the targets are collected in the vector $\mathbf{y}$, so we can write $\mathcal{D} = (X, \mathbf{y})$.

Definition 2.1.1 ((Rasmussen and Williams, 2006)). *A Gaussian process is a collection of random variables, any finite number of which have a joint Gaussian distribution.*

A Gaussian process is completely specified by its mean function and co-variance function. We define mean mean function $m(\mathbf{x})$ and the covariance function $k(\mathbf{x}, \mathbf{x}')$ for a real process $f(\mathbf{x})$ as

$$m(\mathbf{x}) = \mathbb{E}[f(\mathbf{x})],$$

$$k(\mathbf{x}, \mathbf{x}') = \mathbb{E}[(f(\mathbf{x}) - m(\mathbf{x}))(f(\mathbf{x}') - m(\mathbf{x}'))],$$

and will write the Gaussian process as

$$f(\mathbf{x}) \sim \mathrm{GP}(m(\mathbf{x}),\, k(\mathbf{x}, \mathbf{x}')).$$

$$\mathbf{x} = \begin{bmatrix} X_1 \\ \vdots \\ X_p \end{bmatrix} \sim \mathcal{N}_p(\mathbf{m},\, \mathbf{K}),\ \text{where}\ \mathbf{m} = \begin{bmatrix} m(x_1) \\ m(x_2) \\ \vdots \\ m(x_p) \end{bmatrix}\ \text{and}\ \mathbf{K} = \begin{bmatrix} k(x_1,\, x_1) & k(x_1,\, x_2) & \cdots & k(x_1,\, x_p) \\ k(x_2,\, x_1) & k(x_2,\, x_2) & \cdots & k(x_2,\, x_p) \\ \vdots & \vdots & \ddots & \vdots \\ k(x_p,\, x_1) & k(x_p,\, x_2) & \cdots & k(x_p,\, x_p) \end{bmatrix}.$$

Consider a simple example shown in Figure 1, where $m(\mathbf{x}) = \mathbf{x}$ and $k(\mathbf{x}, \mathbf{x}') = \mathrm{I}_p$.

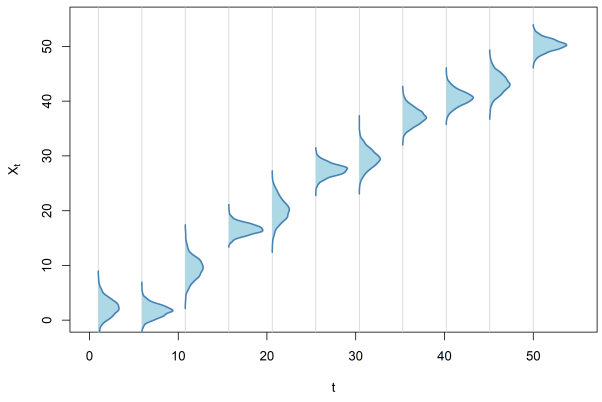


Figure 1: *An example of a Gaussian process on* $t \in \{0, 10, \ldots, 50\}$.

Covariance functions

((Roberts et al., 2013)) In the following section, we briefly describe commonly used kernels. We start with a simple white noise, and then consider common *stationary* covariances, both uni- and multi-dimensional. We finish this section with periodic and quasi-periodic kernel functions. We note that sums (and products) of valid covariance kernels give valid covariance functions (i.e. the resultant covariance matrices are positive semi-definite) and so we may entertain with ease multiple explanatory hypothesis.

White noise with variance σ^2 is represented by

$$k(x_i, x_j) = \sigma^2 \delta_{ij}, \text{ where } \delta_{ij} = 1 \text{ for } i = j \text{ and } \delta_{ij} = 0 \text{ for } i \neq j.$$

This kernel allows us to entertain uncertainity in our observed data and is so typically found added to other kernels.

The SE kernel is given by

$$k(x_i, x_j) = h^2 \exp(-(\frac{x_i - x_j}{\lambda})^2),$$

where h is an output-scale amplitude and λ is an input (length, or time) scale. This gives rather smooth variations with typical time scale of λ and admits functions drawn from the GP that are infinitely differentiable.

The rational quadratic (RQ) kernel is given by

$$k(x_i, x_j) = h^2 (1 + \frac{(x_i - x_j)^2}{\alpha \lambda^2})^{-\alpha},$$

where α is known as the index. ((Rasmussen and Williams, 2006)) show that this is equivalent to a scale mixture of SE kernels with different length scales, the latter distributed according to a Beta distribution with parameters α and λ^{-2}. This gives variations with a range of time scales, the distribution peaking around λ but extending significantly longer period (but remaining rather smooth). When $\alpha \to \infty$, the RQ kernel reduces to the SE kernel with length scale λ.

The Matérn class of covariance functions is defined by

$$k(x_i, x_j) = h^2 \frac{1}{\Gamma(v) 2^{v-1}} (2\sqrt{v} \frac{|x_i - x_j|}{\lambda}) \mathbb{B}_v (2\sqrt{v} \frac{|x_i - x_j|}{\lambda}),$$

where h is the output scale, λ is the input scale, $\Gamma()$ is the standard Gamma function and $\mathbb{B}()$ is the modified Bessel function of second order. The additional hyperparameter v controls the degree of differentiability of the resultant functions modelled by a GP with a Matérn covariance function, such that there are only $(v + \frac{1}{2})$ times differentiable. As $v \to \infty$, so the functions become infinitely differentiable and the Matern kernel becomes the SE one. Taking $v = \frac{1}{2}$ gives the exponential kernel

$$k(x_i, x_j) = h^2 \exp(-(\frac{x_i - x_j}{\lambda})^2),$$

which results in functions that are only once differentiable, and corresponds to the Ornstein-Unlenbeck process.

Prediction

((Betancourt, 2020)) To simulate sampling a function from a Gaussian process and then evaluating the sampled function at the grid of covariate values we can just directly sample from a multivariate normal random number generator.

We can also consider *predictions* by taking advantage of the conditional structure of a multivariate normal density function. Consider a grid of observed covariate values

$$\{x_1^{\text{obs}}, \ldots, x_{N_{\text{obs}}}^{\text{obs}}\}$$

where we know the function values $f(x_n^{\text{obs}})$, and a grid of unobserved covariate values where we want to predict the functional values,

$$\{x_1^{\text{pred}}, \ldots, x_{N_{\text{pred}}}^{\text{pred}}\}.$$

The parameters of the multivariate normal density function of the combined covariate values decomposes into the parameters of the multivariate normal density function for each component grid plus mixed covariate function evaluations.

$$\mathbf{m} = \begin{bmatrix} m(x_1^{\text{obs}}) \\ \vdots \\ m(x_{N_{\text{obs}}}^{\text{obs}}) \\ m(x_1^{\text{pred}}) \\ \vdots \\ m(x_{N_{\text{pred}}}^{\text{pred}}) \end{bmatrix} \text{ and } \mathbf{K} = \begin{bmatrix} k(x_1^{\text{obs}}, x_1^{\text{obs}}) & \cdots & k(x_1^{\text{obs}}, x_{N_{\text{obs}}}^{\text{obs}}) & k(x_1^{\text{obs}}, x_1^{\text{pred}}) & \cdots & k(x_1^{\text{obs}}, x_{N_{\text{pred}}}^{\text{pred}}) \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ k(x_{N_{\text{obs}}}^{\text{obs}}, x_1^{\text{obs}}) & \cdots & k(x_{N_{\text{obs}}}^{\text{obs}}, x_{N_{\text{obs}}}^{\text{obs}}) & k(x_{N_{\text{obs}}}^{\text{obs}}, x_1^{\text{pred}}) & \cdots & k(x_{N_{\text{obs}}}^{\text{obs}}, x_{N_{\text{pred}}}^{\text{pred}}) \\ k(x_1^{\text{pred}}, x_1^{\text{obs}}) & \cdots & k(x_1^{\text{pred}}, x_{N_{\text{obs}}}^{\text{obs}}) & k(x_1^{\text{pred}}, x_1^{\text{pred}}) & \cdots & k(x_1^{\text{pred}}, x_{N_{\text{pred}}}^{\text{pred}}) \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ k(x_{N_{\text{pred}}}^{\text{pred}}, x_1^{\text{obs}}) & \cdots & k(x_{N_{\text{pred}}}^{\text{pred}}, x_{N_{\text{obs}}}^{\text{obs}}) & k(x_{N_{\text{pred}}}^{\text{pred}}, x_1^{\text{pred}}) & \cdots & k(x_{N_{\text{pred}}}^{\text{pred}}, x_{N_{\text{pred}}}^{\text{pred}}) \end{bmatrix},$$

more compactly,

$$\mathbf{m} = \begin{bmatrix} \mathbf{m}_{\text{obs}} \\ \\ \mathbf{m}_{\text{pred}} \end{bmatrix} \text{ and } \mathbf{K} = \begin{bmatrix} \mathbf{K}_{\text{obs}} & \mathbf{K}_{\text{mix}} \\ \\ (\mathbf{K}_{\text{mix}})^{\text{T}} & \mathbf{K}_{\text{pred}} \end{bmatrix}$$

.

The conditional probability density function for the unobserved function values $(f_{\text{pred}})_n = f(x_n^{\text{pred}})$ given the observed function values $(f_{\text{obs}})_n = f(x_n^{\text{obs}})$ falls into the multivariate normal family,

$$\pi(\mathbf{f}_{\text{pred}}|\mathbf{f}_{\text{obs}}) \sim \mathcal{N}_n(\boldsymbol{\mu}, \boldsymbol{\Sigma})$$

with the location parameters

$$\boldsymbol{\mu} = \mathbf{m}_{\text{pred}} + (\mathbf{K}_{\text{mix}})^{\text{T}} \cdot (\mathbf{K}_{\text{obs}})^{-1} \cdot (\mathbf{f}_{\text{obs}} - \mathbf{m}_{\text{obs}})$$

and the covariance matrix parameters

$$\boldsymbol{\Sigma} = \mathbf{K}_{\text{pred}} - (\mathbf{K}_{\text{mix}})^{\text{T}} \cdot (\mathbf{K}_{\text{obs}})^{-1} \cdot \mathbf{K}_{\text{mix}}.$$

Inference with Gaussian process priors proceeds similarly. Once we've identified the covariate values at which we have observations or will want to make predictions we can specify the marginalized prior model with the corresponding multivariate normal density function. If the observational model is normal then we can generate predictions analytically with the conditioning operations for the posterior covariance function.

Air pollution example

Exploratory data analysis

Variable	Name	Description	Unit
NO_2	Nitrogen dioxide	A harmful gas from vehicles and industry.	$\mu g/m^3$
PM_{10}	Particulate matter 10	Small inhalable dust particles.	$\mu g/m^3$
SO_2	Sulphur dioxide	Mainly from burning fossil fuels.	$\mu g/m^3$
Speed	Wind speed	How fast the wind is moving.	m/s

Table 1: *Description of variables of the air pollution dataset.*

Variable	Min.	1st Qu.	Median	Mean	Std.	3rd Qu.	Max.	NA's
NO_2	0.0	5.0	9.0	12.73	10.72	17.0	113.0	734
PM_{10}	0.0	12.0	17.0	19.82	12.30	24.0	158.0	298
SO_2	0.0	2.0	3.0	6.20	11.30	5.0	142.0	638
Speed	0.5	2.3	3.6	3.74	1.71	4.9	11.2	918

Table 2: *Summary statistics of the air pollution dataset.*

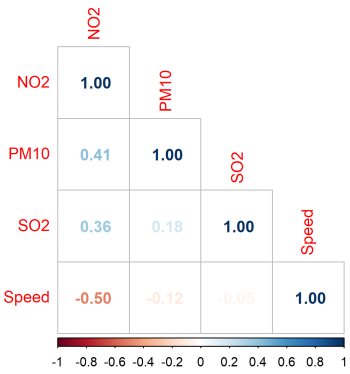
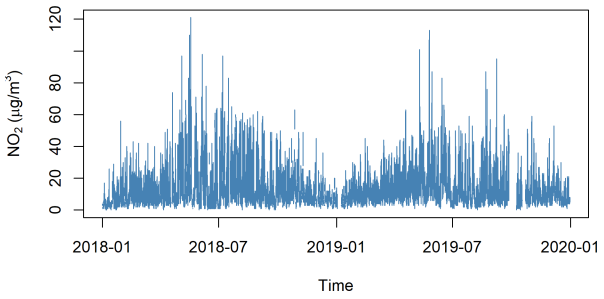
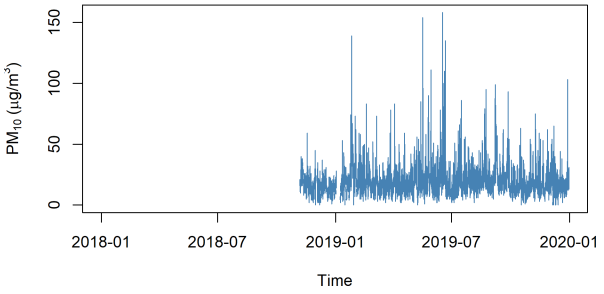


Figure 2: *Correlation plot of the air pollution dataset.*

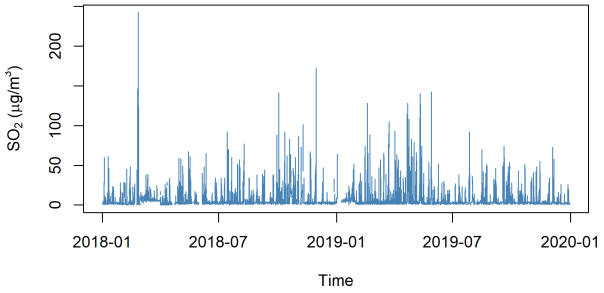
Our response variable NO_2 appears to be moderately positively correlated with PM_{10} and SO_2, and moderately negatively correlated with Speed. These are not ideal explanatory variables since we typically would like them to be strongly correlated with the response variable. The explanatory variables are weekly correlated with one another, whether it be positive or negative correlation. This is ideal since some models do not work well with correlated explanatory variables, often leading to unstable point estimates and inflated standard errors.



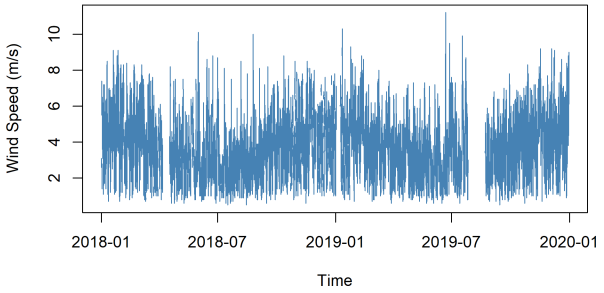
(a) Nitrogen dioxide



(b) Particulate matter 10

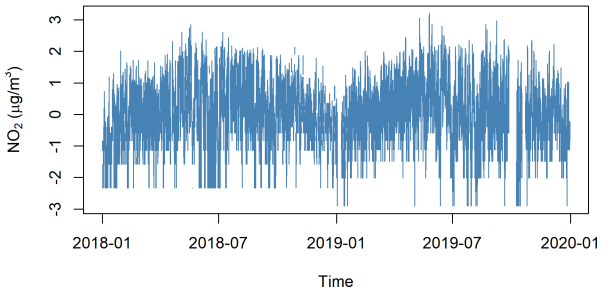


(c) Sulphur dioxide

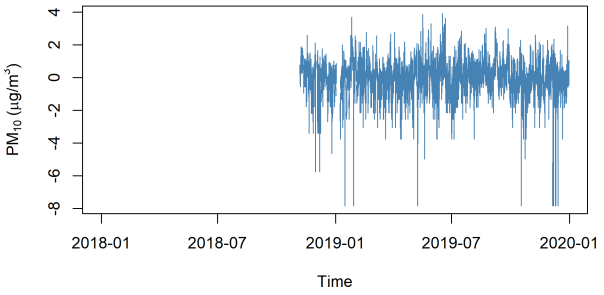


(d) Wind speed

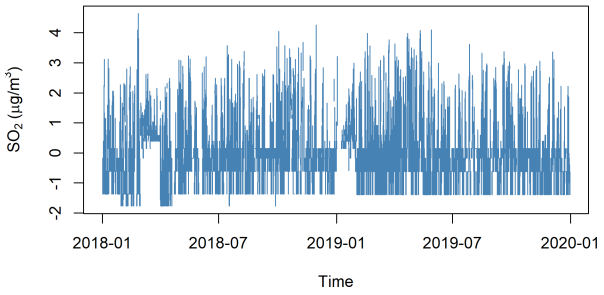
Figure 3: *Time-series plots of nitrogen dioxide and mereorological processes from 01/01/2018 to 31/12/2019 measured hourly. Similar cyclic patterns can be observed in the meteorological processes, and a weaker seasonality component can be noted in the yearly nitrogen dioxide processes.*



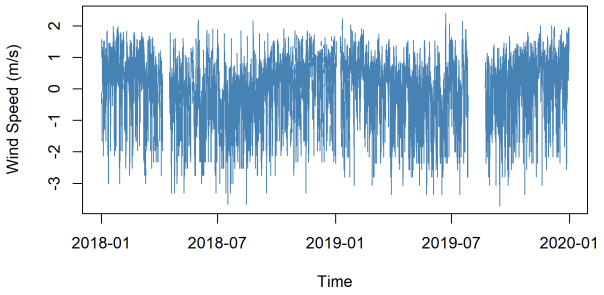
(a) Nitrogen dioxide



(b) Particulate matter 10



(c) Sulphur dioxide

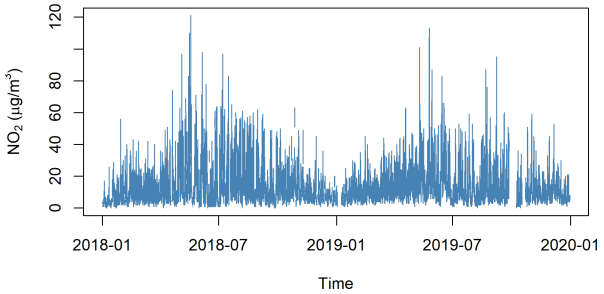


(d) Wind speed

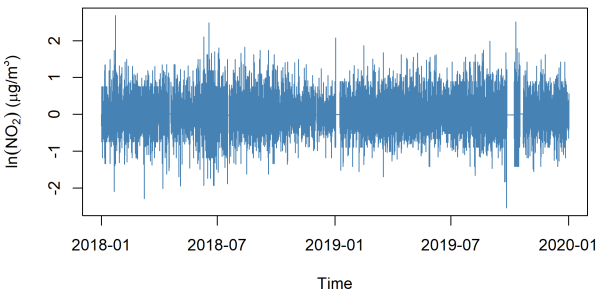
Figure 4: *Post-processed daily curves of the nitrogen dioxide and meteorological data, both standardized by their overall mean and standard deviation.*

Components of the time series

The response variable time-series x_t is not statationary because the mean and variance are changing with time. To achieve stationarity there is a need for detrending and a variance stabilizing transformation. In order to stabilize the variance we use Box-Cox transformations in the training set, $y_t = (x_t^\lambda - 1)/\lambda$. The Box-Cox transformations allow us to experiment with a wide variety of λ values. A good value of λ is one the makes the variation in the data constant through time (Neil Watson, 2024). The `R` package **`forecast`** was used to perform the Box-Cox transformation which yielded an optimal value of $\hat{\lambda} \approx 0$, suggesting a logarithmic transformation. Then, conducting first-order differencing at lag one to remove the trend $z_t = y_t - y_{t-1}$ for $t \in \{2, 3, \ldots, p\}$. The time series z_t is now stationary. Thus, we only need to account for the mean and variance in our models.



(a) Before transformations.



(b) After transformations.

Figure 5: *Comparison of NO_2 time series before and after transformations.*

Model formulation

Priors

$\alpha \sim \text{half-normal}(0,\, \tau^2)$

$$\begin{aligned}\pi(\alpha) &= \frac{2}{\sqrt{2\pi\tau^2}}\exp(-\frac{\alpha^2}{2\tau^2}) \\ &\propto \exp(-\frac{\alpha^2}{2\tau^2}) \\ &\text{for } \alpha \geq 0.\end{aligned}$$

$\rho \sim \text{Inverse-Gamma}(\lambda,\, \beta)$

$$\begin{aligned}\pi(\rho) &= \frac{\beta^\lambda}{\Gamma(\lambda)}\rho^{-\lambda-1}\exp(-\frac{\beta}{\rho}) \\ &\propto \rho^{-\lambda-1}\exp(-\frac{\beta}{\rho}) \\ &\text{for } \rho > 0.\end{aligned}$$

$\sigma \sim \text{half-normal}(0,\, \phi^2)$

$$\begin{aligned}\pi(\sigma) &= \frac{2}{\sqrt{2\pi\phi^2}}\exp(-\frac{\sigma^2}{2\phi^2}) \\ &\propto \exp(-\frac{\sigma^2}{2\phi^2}) \\ &\text{for } \sigma \geq 0.\end{aligned}$$

Likelihood

$f(\mathbf{x}) \sim \text{GP}(m(\mathbf{x}),\, k(\mathbf{x}, \mathbf{x}'))$

$\mathbf{x} \sim \mathcal{N}_p(\mathbf{m},\, \mathbf{K})$

$$\begin{aligned}\mathcal{L}(\mathbf{m},\, \mathbf{K},\, \mathbf{x} |\, \alpha,\, \rho,\, \sigma) &= (2\pi)^{-\frac{p}{2}}\det(\mathbf{K})^{-\frac{1}{2}}\exp(-\frac{1}{2}(\mathbf{x}-\mathbf{m})^{\mathrm{T}}\mathbf{K}^{-1}(\mathbf{x}-\mathbf{m})) \\ &\propto \det(\mathbf{K})^{-\frac{1}{2}}\exp(-\frac{1}{2}(\mathbf{x}-\mathbf{m})^{\mathrm{T}}\mathbf{K}^{-1}(\mathbf{x}-\mathbf{m})) \\ &\text{for } \mathbf{x} \in \mathbb{R}^p.\end{aligned}$$

Posterior

$$\begin{aligned}\pi(\alpha,\, \rho,\, \sigma |\, \mathbf{x}) &\propto \mathcal{L}(\mathbf{m},\, \mathbf{K},\, \mathbf{x} |\, \alpha,\, \rho,\, \sigma) \cdot \pi(\alpha) \cdot \pi(\rho) \cdot \pi(\sigma) \\ &\propto \det(\mathbf{K})^{-\frac{1}{2}}\exp(-\frac{1}{2}(\mathbf{x}-\mathbf{m})^{\mathrm{T}}\mathbf{K}^{-1}(\mathbf{x}-\mathbf{m})) \cdot \exp(-\frac{\alpha^2}{2\tau^2}) \cdot \rho^{-\lambda-1}\exp(-\frac{\beta}{\rho}) \cdot \exp(-\frac{\sigma^2}{2\phi^2}) \\ &\text{for } \alpha \geq 0,\ \rho > 0,\ \sigma \geq 0.\end{aligned}$$

Results

Model	RMSE			MAE		
	Forecasts (h day time horizon)			Forecasts (h day time horizon)		
	24	168	744	24	168	744
Average	7.731	**5.397**	7.439	**6.042**	**4.254**	5.385
Naive	10.553	7.864	10.040	7.708	6.071	7.308
Drift	10.606	8.128	11.358	7.764	6.391	8.809
AR(1)	8.662	5.912	8.044	6.029	4.422	5.616
GP-0-WN	13.756	11.141	13.169	11.719	9.786	11.081
GP-0-SE						
GP-MLR-WN	**7.692**	6.134	**7.225**	6.119	4.490	**5.136**
GP-MLR-SE						

Table 3: *RMSE and MAE of forecasting models across horizons.*

Bibliography

1. Betancourt (2020). Robust Gaussian process modeling.
2. Roberts et al. (2013). Gaussian processes for time-series modelling. *Philosophical Transactions of the Royal Society.*
3. Rasmussen and Williams (2006). Gaussian Processes in Machine Learning. *The MIT Press.*