Data Estimation & Inference

Leonard Berrada

October 24, 2015

1 Introduction

The aim of this report is to shed light on the code[1] and experiments I made for the laboratory session. As a quick reminder, the objective of the lab is to implement a Gaussian Process Regression to predict missing values in the Sotonmet dataset, which consists in weather data measured in the port of Southampton. I will explain in this report the choices I made for the implementation, and I will present the results I have obtained.

2 Gaussian Process Regression: My Implementation

NB We follow the notations of [2] and therefore do not introduce them again.

I present in this section a few insights that might be of interest about my implementation of the Gaussian Process Regression.

2.1 Using Choleski Decomposition Efficiently

When predicting, we need to estimate:

$$\begin{aligned} \bar{y_*} &= K_* K^{-1} \mathbf{y} \\ var(y_*) &= K_{**} - K_* K^{-1} K_*^T \end{aligned}$$

Therefore, the repeated term is $K_* K^{-1}$, which we want to compute only once with Choleski decomposition. K_* is a row vector $(1 \times n)$, and K^{-1} a square matrix $(n \times n)$. Using the `numpy.linalg.solve` function, we can only solve an equation of the form $Au = v$ with u and v column vectors, and not $uA = v$, with u and v row vectors. We derive here how to compute $K_* K^{-1}$ with the right formulation:

$$\begin{aligned} K &= LL^T \text{ , hence:} \\ K_* K^{-1} &= K_* \left(LL^T\right)^{-1} \\ &= K_* \left(L^T\right)^{-1} L^{-1} \\ &= K_* \left(L^{-1}\right)^T L^{-1} \\ &= \left(L^{-1} K_*^T\right)^T L^{-1} \\ &= \left(\left(L^{-1}\right)^T L^{-1} K_*^T\right)^T \end{aligned}$$

This new form allows us to use the `numpy.linalg.solve` as is.

2.2 Kernels Implemented

I have used [1], [2] and [3] to implement the following kernels:

- Exponential Quadratic : $k_{EQ}(x, x') = \sigma_f^2 \exp\left(-\frac{(x - x')^2}{2l^2}\right)$

  Hyperparameters : $\theta_{EQ} = (\sigma_f, l) \left(\mathbb{R}_+^*\right)^2$

[1] The code can be found at `https://github.com/leonardbj/AIMS`

- Periodic : $k_P(x, x') = \sigma_f^2 \exp\left(-2\sin^2\left(\nu\pi|x - x'|\right)\right)$
  Hyperparameters : $\theta_P = (\sigma_f, \nu) \left(\mathbb{R}_+^*\right)^2$

- Rational Quadratic : $k_{RQ}(x, , x') = \sigma_f^2 \left(1 + \frac{1}{2\nu}\left(\frac{|x - x'|}{l}\right)^2\right)^{-\nu}$
  Hyperparameters : $\theta_{RQ} = (\sigma_f, l, \nu) \in \left(\mathbb{R}_+^*\right)^3$

- Matérn : $k_M(x, x') = \sigma_f^2 \frac{1}{2^{\nu-1}\Gamma(\nu)} \left(\frac{2\sqrt{\nu}|x - x'|}{l}\right)^\nu \mathcal{H}_\nu \left(\frac{2\sqrt{\nu}|x - x'|}{l}\right)$
  Hyperparameters : $\theta_M = (\sigma_f, l, \nu) \in \left(\mathbb{R}_+^*\right)^3$
  NB : Γ is the Gamma function, and $\mathcal{H}_\nu$ the modified Bessel function of the second kind of real order ν
  In practice I encountered numerical problems when using this general form, so I reduced the cases to :
  - Matérn 1/2 : $k_{M1}(x, x') = \sigma_f^2 \exp(\frac{-|x - x'|}{l})$
    Hyperparameters : $\theta_{M1} = (\sigma_f, l) \left(\mathbb{R}_+^*\right)^2$
  - Matérn 3/2 : $k_{M3}(x, x') = \sigma_f^2(1 + \sqrt{3}\frac{|x - x'|}{l}) \exp(\frac{-\sqrt{3}|x - x'|}{l})$
    Hyperparameters : $\theta_{M3} = (\sigma_f, l) \left(\mathbb{R}_+^*\right)^2$

Independently to the choice of the kernel, we add a data-characteristic noise σ_n^2 on the diagonal of the kernel matrix, which counts as an additional hyperparameter, also taking value in $\mathbb{R}_+^*$.

2.2.1 Mean Functions

I have implemented the following functions:

- constant : $\mathbf{m}(\mathbf{x}) = \alpha$
  Hyperparameters : $\theta_C = (\alpha) \in \mathbb{R}$
- linear : $\mathbf{m}(\mathbf{x}) = \alpha\mathbf{x} + \beta$
  Hyperparameters : $\theta_L = (\alpha, \beta) \in (\mathbb{R})^2$
- periodic : $\mathbf{m}(\mathbf{x}) = \alpha \sin(\frac{2\pi\mathbf{x}}{\rho})$
  Hyperparameters : $\theta_{P'} = (\alpha, \rho) \in \left(\mathbb{R}_+^*\right)^2$

2.2.2 Combining Methods

My implementation allows to handle any combination of multiplication or addition of implemented kernels, and the same goes for the mean functions : I parse a string given in argument, recognize the operators ('+' or '*') and automatically collect the parameters, with their priors and range of possible values. This will allow us to explore a great range of combinations during the model selection phase.

NB Early experiments showed that the linear mean did not yield better results, even when combined with other mean functions. We therefore discarded it for the model selection (in order to reduce the number of possibilities to test).

3 Hyperparameter Estimation

3.1 Maximum Likelihood Estimation

As derived in [4], given a model $\mathcal{H}$ (i.e. a mean function $\mathbf{m}(\mathbf{x})$ and a covariance function $K(\mathbf{x})$), and θ a set of parameters over $\mathcal{H}$, we have the following marginal log-likelihood:

$$\log p(\mathbf{y}|\mathbf{x}, \theta, \mathcal{H}) = -\frac{1}{2}(\mathbf{y} - \mathbf{m}(\mathbf{x}))^T K(x)^{-1}(\mathbf{y} - \mathbf{m}(\mathbf{x})) - \frac{1}{2}\log|K| - \frac{n}{2}\log 2\pi$$

Removing the constant term, we obtain the following function to maximize:

$$\tilde{\mathcal{L}}(\theta) = -\frac{1}{2}(\mathbf{y} - \mathbf{m}(\mathbf{x}))^T K(x)^{-1}(\mathbf{y} - \mathbf{m}(\mathbf{x})) - \frac{1}{2}\log|K|$$

Then the Maximum Likelihood Estimator (MLE) is given by:

$$\hat{\theta}_{\mathrm{MLE}} = \arg\max_{\theta} \tilde{\mathcal{L}}(\theta)$$

3.2 Maximum A Posteriori Estimation

Given that $\tilde{\mathcal{L}}(\theta)$ is the log-likelihood (up to a constant), we have the following expression for the log posterior probability (up to a constant):

$$\begin{aligned}\tilde{\mathcal{P}}(\theta) &= \tilde{\mathcal{L}}(\theta) + \log p(\theta) \\ &= \tilde{\mathcal{L}}(\theta) + \sum_k \log p(\theta_k)\end{aligned}$$

We index with i the non-negative parameters. These have a normal prior $\mathcal{N}(\log(\mu_i), \sigma_i^2)$ on their logarithm. Other parameters, which take their value in all $\mathbb{R}$, are indexed by j and have a Gaussian prior $\mathcal{N}(\mu_j, \sigma_j^2)$. With the logarithmic change of variable, this yields (up to a constant):

$$\tilde{\mathcal{P}}(\theta) = \tilde{\mathcal{L}}(\theta) - \frac{1}{2}\sum_j \frac{1}{\sigma_j^2}(\theta_j - \mu_j)^2 - \frac{1}{2}\sum_i \frac{1}{\sigma_i^2}(\log(\theta_i) - \log(\mu_i))^2 - \sum_i \log(\theta_i)$$

And then:

$$\hat{\theta}_{\mathrm{MAP}} = \arg\max_{\theta} \tilde{\mathcal{P}}(\theta)$$

3.3 Practical Optimization

As we have seen, the θ_i have boundaries limits in the form of $]0, +\infty[$. We therefore the L-BFGS-B, a variant of the BFGS algorithm, to minimize $-\tilde{\mathcal{L}}(\theta)$ or $-\tilde{\mathcal{P}}(\theta)$ while respecting the constraints. We set boundaries to $(1\text{E}^{-6}, \infty)$[2] to deal with our constraints. This avoids any numerical error due to zero valued parameter. The use of the BFGS algorithm is justified by the regularity properties of either $\tilde{\mathcal{L}}(\theta)$ or $\tilde{\mathcal{P}}(\theta)$, which are both $\mathcal{C}^\infty$ functions of the hyperparameters for the mean functions and kernels presented here.

We try to give reasonable values for the means and variances of the Gaussian hyperpriors. Such parameters should probably depend on the predicted variable - for simplicity purposes, this is not the case in my code. The values can be found in in the source code, in `params.py`.

4 Model Selection

As described in [4][3], one may consider three levels of inference, one for each of the parameters, hyperparameters and model levels. The first level of inference is implicitly performed in the case of Gaussian Process Regression, during the computation of the mean and covariance matrix. The second one, for hyperparameters, has just been done in the previous section about hyperparameters tuning, and we address the third one here. More specifically:

- Level 2 : Hyperparameters θ (the use of $\mathcal{H}$ was implicit in the previous section)

$$\hat{\theta}_{\mathcal{H},\mathrm{MLE}} = \arg\max_{\theta} p(\mathbf{y}|\mathbf{x}, \theta, \mathcal{H}) = \arg\max_{\theta} \tilde{\mathcal{L}}_{\mathcal{H}}(\theta)$$

$$\hat{\theta}_{\mathcal{H},\mathrm{MAP}} = \arg\max_{\theta} p(\theta|\mathbf{x}, \mathbf{y}, \mathcal{H}) = \arg\max_{\theta} \tilde{\mathcal{P}}_{\mathcal{H}}(\theta)$$

- Level 3 : Model $\mathcal{H}$

$$\hat{\mathcal{H}}_{\mathrm{MLE}} = \arg\max_{\mathcal{H}} p(\mathbf{y}|\mathbf{x}, \mathcal{H})$$

$$\begin{aligned}\hat{\mathcal{H}}_{\mathrm{MAP}} &= \arg\max_{\mathcal{H}} p(\mathcal{H}|\mathbf{x}, \mathbf{y}) \\ &= \arg\max_{\mathcal{H}} p(\mathbf{y}|\mathbf{x}, \mathcal{H}) p(\mathcal{H})\end{aligned}$$

[2] $(1\text{E}^{-3}, \infty)$ for σ_n to ensure non-singularity during the optimization

[3] p. 109

We assume here that all models are equally likely, which yields in both case $\hat{\mathcal{H}} = \arg\max_{\mathcal{H}} p(\mathbf{y}|\mathbf{x}, \mathcal{H})$. The difficulty is to approximate $p(\mathbf{y}|\mathbf{x}, \mathcal{H}) = \int p(\mathbf{y}|\mathbf{x}, \theta, \mathcal{H})p(\theta|\mathcal{H})d\theta$. The authors of [4] recommend to approximate this integral by a Laplace approximation around $\hat{\theta}_{\mathrm{MLE}}$. In our case, we use a cruder approximation:

$$\hat{\mathcal{H}}_{\mathrm{MLE}} \approx \arg\max_{\mathcal{H}} p(\mathbf{y}|\mathbf{x}, \hat{\theta}_{\mathcal{H},\mathrm{MLE}}, \mathcal{H}) = \arg\max_{\mathcal{H}} \left(\max_{\theta} \tilde{\mathcal{L}}_{\mathcal{H}}(\theta) \right)$$

$$\hat{\mathcal{H}}_{\mathrm{MAP}} \approx \arg\max_{\mathcal{H}} p(\mathbf{y}|\mathbf{x}, \hat{\theta}_{\mathcal{H},\mathrm{MAP}}, \mathcal{H}) = \arg\max_{\mathcal{H}} \left(\max_{\theta} \tilde{\mathcal{P}}_{\mathcal{H}}(\theta) \right)$$

Put it differently, we simply choose the best optimized score to select a model, this score being (up to a constant) either the log-likelihood ($\tilde{\mathcal{P}}_{\mathcal{H}}(\theta)$) or the log-posterior ($\tilde{\mathcal{P}}_{\mathcal{H}}(\theta)$) according to the estimation chosen. This method has yielded the following results (see Appendix : Values Of Hyperparameters for all results):

Variable	Model Selected with MLE		Model Selected with MAP	
Tide	Mean: Constant + Periodic	Kernel: Matérn 3/2 × Periodic	Mean: Constant	Kernel: Matérn 1/2 × Periodic
Temperature	Mean: Constant	Kernel: Exp. Quadratic + Exp. Quadratic	Mean: Constant	Kernel: Matérn 1/2 × Periodic

Table 1: Model selected by campaign of results

Interestingly enough, the MAP estimation has chosen the same models for the two variables, although the optimized hyperparameters are obviously not the same. This might be because the priors happened to be closer to the optimal values found by the MLE, in which case it the result might be biased. One might be surprised by the selection of the sum of exponential for the ML estimation with the temperature. It is worth to say that the second exponential is initialized with a significantly larger length scale, in order to model longer-term correlations, which is what effectively happened during the optimization (second length scale ten times larger than the first one).

Choosing these selected models with their optimal hyperparameters, we can show the plot of their prediction. In black is displayed the ground truth, in red the prediction, with dark red shade on the region $\pm\sigma_*$, and light red shade the 95% confidence interval ($\pm 1.96\sigma_*$) :

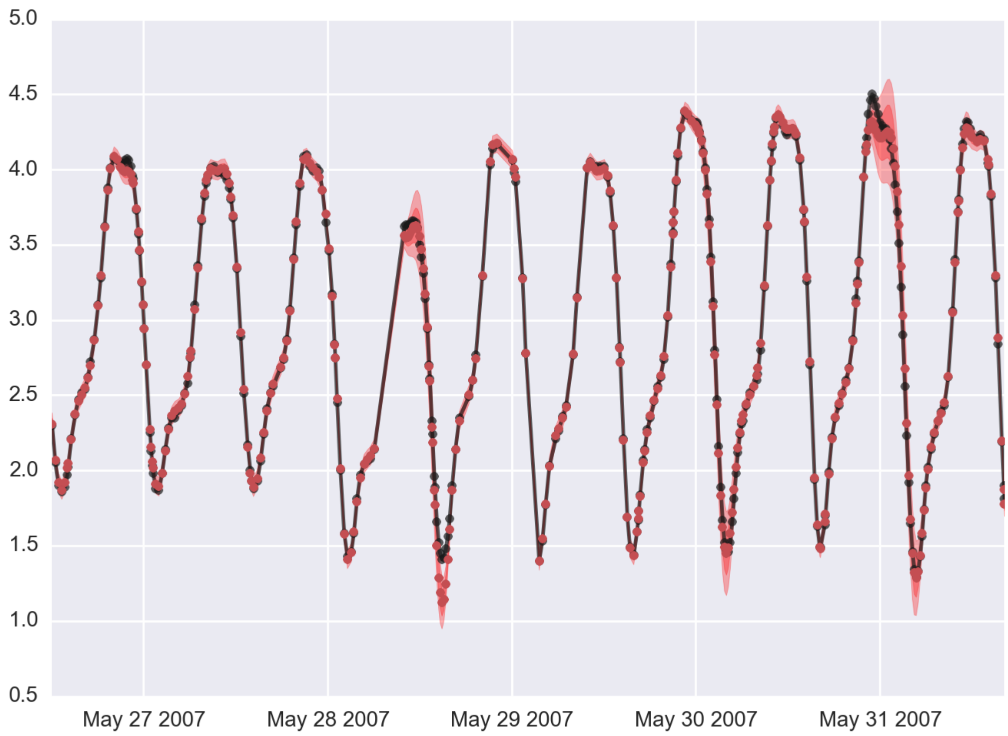

Figure 1: Prediction of the tide with model selected by ML estimation

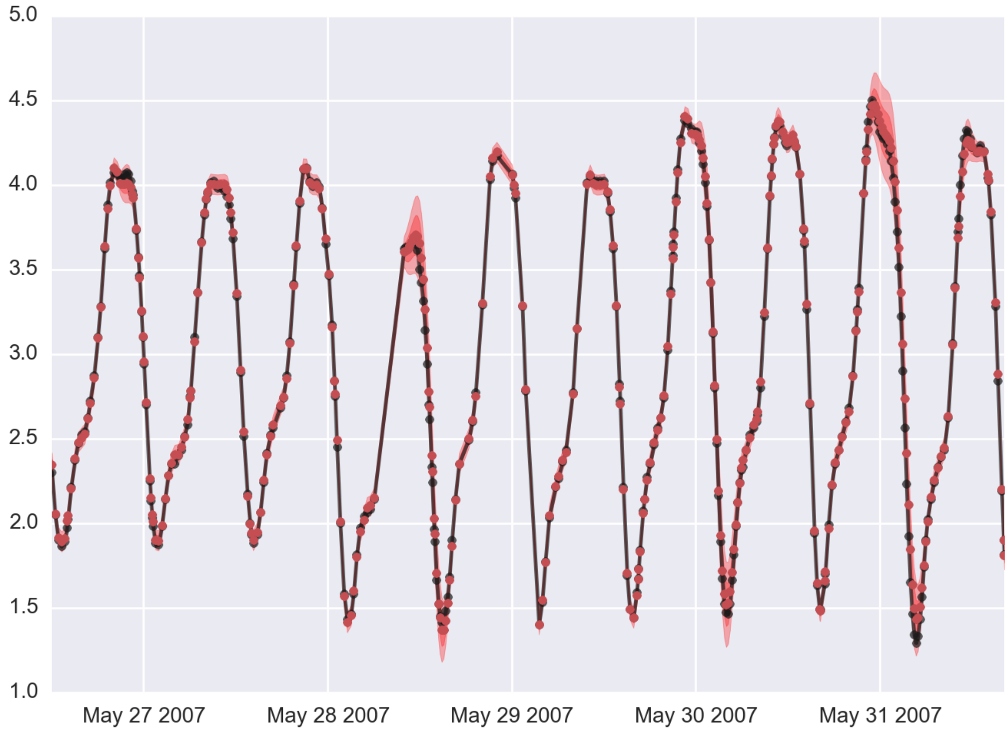

Figure 2: Prediction of the tide with model selected by MAP estimation

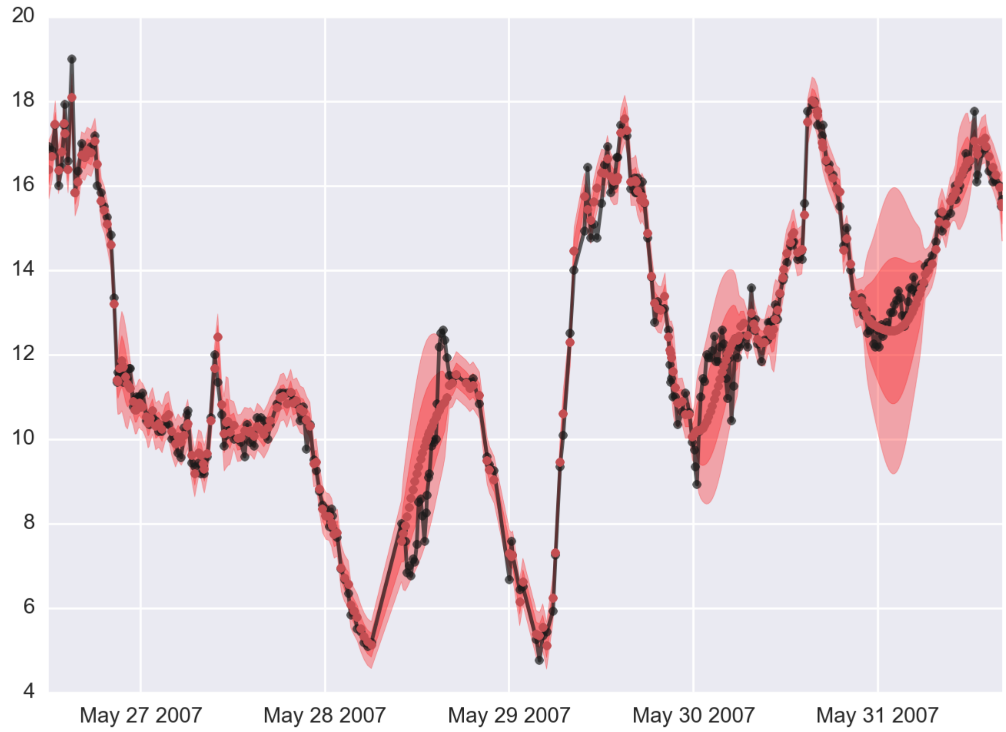

Figure 3: Prediction of the temperature with model selected by ML estimation

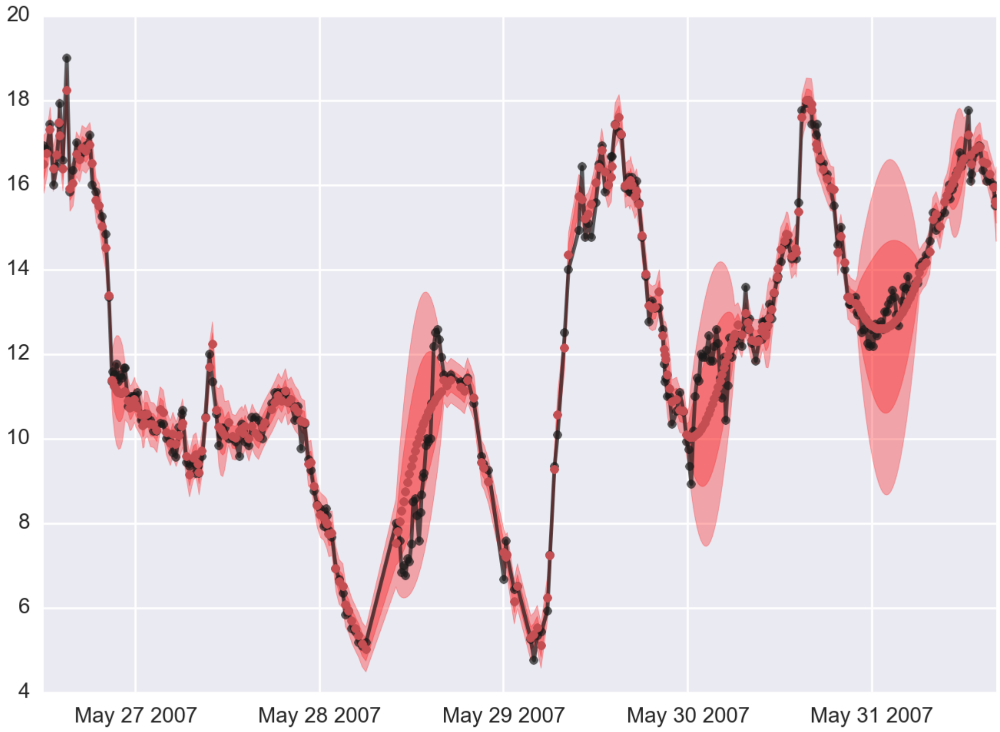

Figure 4: Prediction of the temperature with model selected by MAP estimation

5 Model Assessment

We have the opportunity to perform a real model assessment in our case, as we have access to the ground truth. This will enable us to see how our selected model really perform, as we can compare their predictions against the ground truth.

5.1 R^2 Score

We use the R^2 coefficient of determination:

$$R^2 = 1 - \frac{\sum_{i=1}^{n} (Y_*^i - Y_{\text{truth}}^i)^2}{(Y_*^i - \mu(\mathbf{Y}_*))^2}$$

With:

- $\mathbf{Y}_*$ the vector of predicted values : $\mathbf{Y}_* = (Y_*^1, ..., Y_*^n)$
- $\mathbf{Y}_{\text{truth}}$ the vector of true values : $\mathbf{Y}_{\text{truth}} = (Y_{\text{truth}}^1, ..., Y_{\text{truth}}^n)$
- $\mu(\mathbf{Y}_*)$ the empirical mean of the predicted values

Here are the results:

Variable	R^2 Score with MLE	R^2 Score with MAP
Tide	0.9972	**0.9979**
Temperature	**0.9686**	0.9659

Table 2: R^2 scores obtained with model selected - Best result in **bold** for each variable

One can notice that for both variables, the difference of performance between the ML and the MAP estimators is very low. This is probably because we do not have brought much external knowledge to the problem - an expert in weather or tide modeling might for instance provide with insights that could be fed to the model, whether on the choice of the model (implicit prior knowledge) or directly on the range or distribution of the hyperparameters (explicit prior).

5.2 Sequential Mode

Let us see some results in sequential mode, when using the optimal hyperparameters tuned on all the data[4]. We can notice that the tuning has been very effective : even in regions with low training data, the GP has learnt the trends to follow.

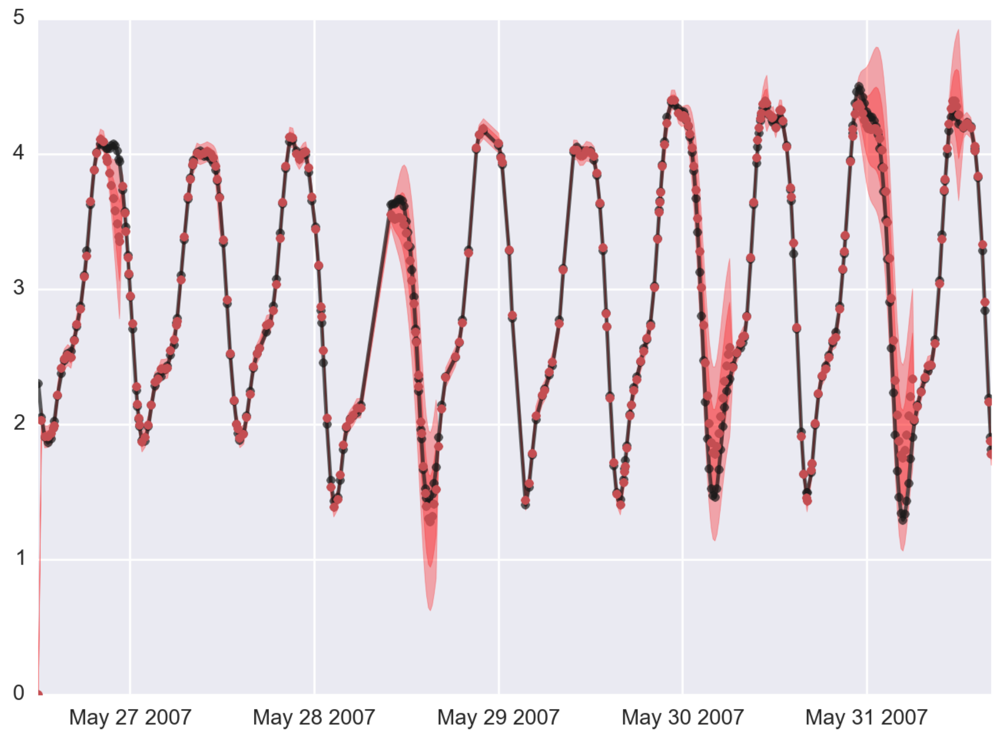

Figure 5: Sequential prediction of the tide with model selected by MLE

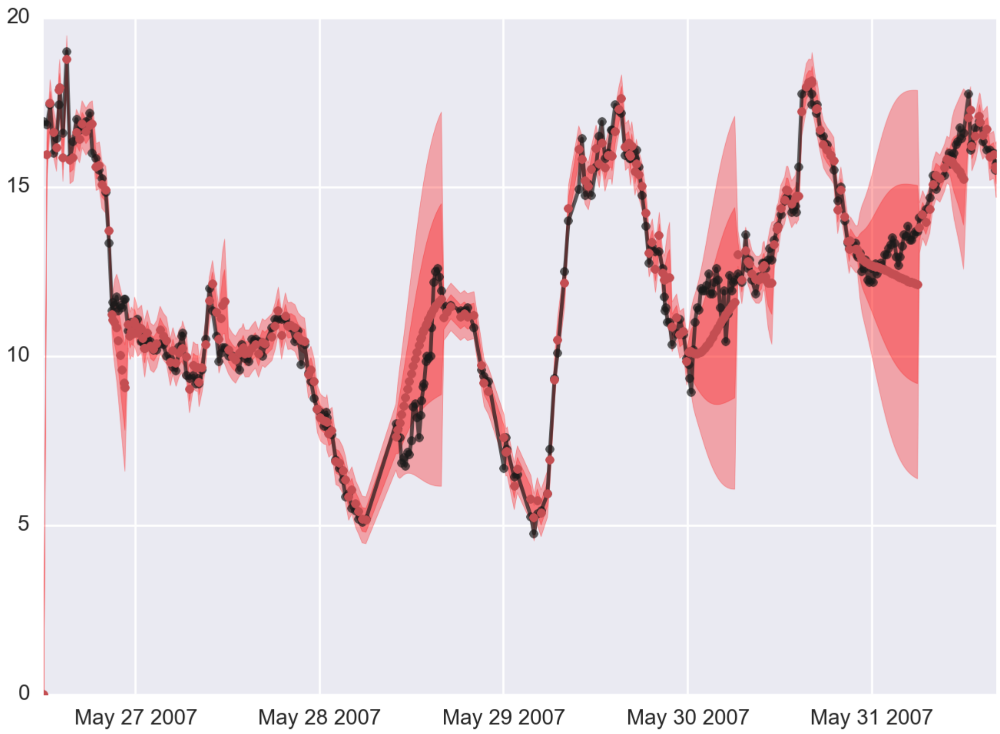

Figure 6: Sequential prediction of the temperature with model selected by MLE

[4]This is ”cheating”, as we are using information from the future when tuning on all the training data. However it was too expensive to perform model selection and tune the hyperparameters online.

Note that the first point set to zero is an artifact because no prediction can be made without the model having learned anything.

6 Conclusion

We have showed in this report how we have modeled, implemented and selected our model on the Sotonmet dataset. Gaussian Process Regression has proven to be an excellent solution to model the trends. We have followed what could be qualified as a Machine Learning approach, in the sense that the model selection has been blindly performed on a variety of automatically generated models. An alternative approach would be to focus on the modeling, and to have a closer look at the underlying physical phenomena, in order to provide with modeling choices to fit this reality.

Further steps could include more work on prior modeling or fusion of prediction from different variables, for the output as well as the input.

Appendix : Values Of Hyperparameters

All parameters for the hyperpriors (i.e. mean and standard deviation of the Gaussian models) can be found in the source code, in `params.py`. We give in this appendix the optimal estimated values obtained during a campaign of experiments:

ML Estimation On Tide

Kernel	Mean	Hyperparameters									Score
EQ+EQ	C	σ_n 0.029	$\sigma_{f,EQ}$ 0.81	l_{EQ} 18	$\sigma_{f,EQ}$ 0	l_{EQ} 1E+02	α 3				-2.42E+03
EQ+EQ	C+P	σ_n 0.029	$\sigma_{f,EQ}$ 0.81	l_{EQ} 18	$\sigma_{f,EQ}$ 0.0017	l_{EQ} 1E+02	α 3	σ_p 1	ρ 1.5E+02		-2.42E+03
EQ+P	C	σ_n 0.029	$\sigma_{f,EQ}$ 0.81	l_{EQ} 18	$\sigma_{f,P}$ 0	ρ 1.5E+02	α 3				-2.42E+03
EQ+P	C+P	σ_n 0.029	$\sigma_{f,EQ}$ 0.81	l_{EQ} 18	$\sigma_{f,P}$ 0	ρ 1.5E+02	α 3	σ_p 1	ρ 1.5E+02		-2.42E+03
RQ+P	C	σ_n 0.029	$\sigma_{f,RQ}$ 0.87	l_{RQ} 21	ν_{RQ} 3.8	$\sigma_{f,P}$ 0	ρ 1.5E+02	α 3			-2.42E+03
RQ+P	C+P	σ_n 0.029	$\sigma_{f,RQ}$ 0.95	l_{RQ} 23	ν_{RQ} 2.3	$\sigma_{f,P}$ 0	ρ 1.5E+02	α 3	σ_p 1	ρ 1.5E+02	-2.43E+03
M1+P	C	σ_n 0.001	$\sigma_{f,EQ}$ 0.28	l_{EQ} 27	$\sigma_{f,P}$ 12	ρ 1.5E+02	α 7.1				-1.81E+03
M1+P	C+P	σ_n 0.001	$\sigma_{f,EQ}$ 0.73	l_{EQ} 2.7E+02	$\sigma_{f,P}$ 0.061	ρ 1.1E+02	α 2.9	σ_p 1	ρ 1.5E+02		-1.97E+03
M3+P	C	σ_n 0.027	$\sigma_{f,EQ}$ 0.68	l_{EQ} 38	$\sigma_{f,P}$ 6.3	ρ 1.5E+02	α 22				-2.29E+03
M3+P	C+P	σ_n 0.027	$\sigma_{f,EQ}$ 0.71	l_{EQ} 40	$\sigma_{f,P}$ 8.8	ρ 1.5E+02	α 25	σ_p 1	ρ 1.5E+02		-2.3E+03
EQ×EQ	C	σ_n 0.029	$\sigma_{f,EQ}$ 0.98	l_{EQ} 18	$\sigma_{f,EQ}$ 0.82	l_{EQ} 1E+02	α 3				-2.42E+03
EQ×EQ	C+P	σ_n 0.029	$\sigma_{f,EQ}$ 1	l_{EQ} 18	$\sigma_{f,EQ}$ 0.8	l_{EQ} 1E+02	α 3	σ_p 1	ρ 1.5E+02		-2.42E+03
EQ×P	C	σ_n 0.029	$\sigma_{f,EQ}$ 1	l_{EQ} 1.7E+02	$\sigma_{f,P}$ 0.96	ρ 2.9E+02	α 2.9				-2.46E+03
EQ×P	C+P	σ_n 0.029	$\sigma_{f,EQ}$ 0.95	l_{EQ} 40	$\sigma_{f,P}$ 0.92	ρ 3E+02	α 3	σ_p 1	ρ 1.5E+02		-2.42E+03
RQ×P	C	σ_n 0.028	$\sigma_{f,RQ}$ 0.78	l_{RQ} 40	ν_{RQ} 54	$\sigma_{f,P}$ 0.78	ρ 1.6E+02	α 3			-2.32E+03
RQ×P	C+P	σ_n 0.029	$\sigma_{f,RQ}$ 0.99	l_{RQ} 1.7E+02	ν_{RQ} 54	$\sigma_{f,P}$ 0.97	ρ 2.9E+02	α 2.9	σ_p 1	ρ 1.5E+02	-2.46E+03
M1×P	C	σ_n 0.001	$\sigma_{f,EQ}$ 0.28	l_{EQ} 1.8E+03	$\sigma_{f,P}$ 4.5	ρ 3E+02	α 6.4				-2.37E+03
M1×P	C+P	σ_n 0.001	$\sigma_{f,EQ}$ 0.32	l_{EQ} 1.8E+03	$\sigma_{f,P}$ 3.4	ρ 3E+02	α 3	σ_p 1	ρ 1.5E+02		-2.4E+03
M3×P	C	σ_n 0.032	$\sigma_{f,EQ}$ 11	l_{EQ} 3.8E+02	$\sigma_{f,P}$ 0.17	ρ 2.9E+02	α 0.13				-2.45E+03
M3×P	**C+P**	σ_n 0.03	$\sigma_{f,EQ}$ 0.1	l_{EQ} 5E+02	$\sigma_{f,P}$ 9.4	ρ 3E+02	α 2.9	σ_p 1	ρ 1.5E+02		**-2.49E+03**

Table 3: ML estimation on the tide - Best result in **bold**

MAP Estimation On Tide

Kernel	Mean	Hyperparameters									Score
EQ+EQ	C	σ_n 0.029	$\sigma_{f,EQ}$ 0.8	l_{EQ} 18	$\sigma_{f,EQ}$ 0.11	l_{EQ} 1E+02	α 3				-2.4E+03
EQ+EQ	C+P	σ_n 0.029	$\sigma_{f,EQ}$ 0.8	l_{EQ} 18	$\sigma_{f,EQ}$ 0.11	l_{EQ} 1E+02	α 3	σ_p 0.0001	ρ 1.5E+02		-2.39E+03
EQ+P	C	σ_n 0.029	$\sigma_{f,EQ}$ 0.8	l_{EQ} 18	$\sigma_{f,P}$ 0.0001	ρ 1.5E+02	α 3				-2.4E+03
EQ+P	C+P	σ_n 0.029	$\sigma_{f,EQ}$ 0.8	l_{EQ} 18	$\sigma_{f,P}$ 0.0001	ρ 1.5E+02	α 3	σ_p 0.0001	ρ 1.5E+02		-2.39E+03
RQ+P	C	σ_n 0.029	$\sigma_{f,RQ}$ 0.96	l_{RQ} 24	ν_{RQ} 2	$\sigma_{f,P}$ 0.0001	ρ 1.5E+02	α 3			-2.4E+03
RQ+P	C+P	σ_n 0.029	$\sigma_{f,RQ}$ 0.88	l_{RQ} 22	ν_{RQ} 3.1	$\sigma_{f,P}$ 0.0001	ρ 1.5E+02	α 3	σ_p 0.0001	ρ 1.5E+02	-2.4E+03
M1+P	C	σ_n 0.0039	$\sigma_{f,EQ}$ 0.33	l_{EQ} 49	$\sigma_{f,P}$ 38	ρ 1.5E+02	α 27				-1.82E+03
M1+P	C+P	σ_n 0.0039	$\sigma_{f,EQ}$ 0.3	l_{EQ} 38	$\sigma_{f,P}$ 25	ρ 1.5E+02	α 17	σ_p 0.21	ρ 1.5E+02		-1.8E+03
M3+P	C	σ_n 0.027	$\sigma_{f,EQ}$ 1.4	l_{EQ} 72	$\sigma_{f,P}$ 0.0001	ρ 1.4E+02	α 2.9				-2.35E+03
M3+P	C+P	σ_n 0.026	$\sigma_{f,EQ}$ 0.48	l_{EQ} 26	$\sigma_{f,P}$ 9.3	ρ 1.5E+02	α 7.1	σ_p 0.0009	ρ 1.5E+02		-2.22E+03
EQ×EQ	C	σ_n 0.029	$\sigma_{f,EQ}$ 1.5	l_{EQ} 18	$\sigma_{f,EQ}$ 0.53	l_{EQ} 1E+02	α 3				-2.39E+03
EQ×EQ	C+P	σ_n 0.029	$\sigma_{f,EQ}$ 1.1	l_{EQ} 18	$\sigma_{f,EQ}$ 0.71	l_{EQ} 1E+02	α 3	σ_p 0.0001	ρ 1.5E+02		-2.39E+03
EQ×P	C	σ_n 0.029	$\sigma_{f,EQ}$ 0.96	l_{EQ} 1.7E+02	$\sigma_{f,P}$ 1	ρ 2.9E+02	α 2.9				-2.43E+03
EQ×P	C+P	σ_n 0.029	$\sigma_{f,EQ}$ 0.95	l_{EQ} 1.7E+02	$\sigma_{f,P}$ 1	ρ 2.9E+02	α 2.9	σ_p 0.0001	ρ 1.5E+02		-2.43E+03
RQ×P	C	σ_n 0.029	$\sigma_{f,RQ}$ 1	l_{RQ} 59	ν_{RQ} 2E+02	$\sigma_{f,P}$ 0.93	ρ 2.9E+02	α 3			-2.39E+03
RQ×P	C+P	σ_n 0.028	$\sigma_{f,RQ}$ 0.77	l_{RQ} 39	ν_{RQ} 51	$\sigma_{f,P}$ 0.78	ρ 1.6E+02	α 3	σ_p 0.0001	ρ 1.5E+02	-2.28E+03
M1×P	**C**	σ_n 0.025	$\sigma_{f,EQ}$ 0.71	l_{EQ} 1.8E+03	$\sigma_{f,P}$ 0.71	ρ 3E+02	α 3				**-2.48E+03**
M1×P	C+P	σ_n 0.025	$\sigma_{f,EQ}$ 1.4	l_{EQ} 1.8E+03	$\sigma_{f,P}$ 0.35	ρ 3E+02	α 2.9	σ_p 0.0001	ρ 1.5E+02		-2.48E+03
M3×P	C	σ_n 0.03	$\sigma_{f,EQ}$ 0.99	l_{EQ} 5.5E+02	$\sigma_{f,P}$ 1	ρ 3E+02	α 2.9				-2.47E+03
M3×P	C+P	σ_n 0.028	$\sigma_{f,EQ}$ 1.4	l_{EQ} 1.3E+02	$\sigma_{f,P}$ 1.1	ρ 3.9E+02	α 2.9	σ_p 0.0001	ρ 1.5E+02		-2.39E+03

Table 4: MAP estimation on the tide - Best result in **bold**

ML Estimation On Temperature

Kernel	Mean	Hyperparameters									Score
EQ+EQ	**C**	σ_n	$\sigma_{f,EQ}$	l_{EQ}	$\sigma_{f,EQ}$	l_{EQ}	α				**-416**
		0.25	0.47	3.8	2.9	41	12				
EQ+EQ	C+P	σ_n	$\sigma_{f,EQ}$	l_{EQ}	$\sigma_{f,EQ}$	l_{EQ}	α	σ_p	ρ		-416
		0.25	0.47	3.8	2.9	41	12	1	1.5E+02		
EQ+P	C	σ_n	$\sigma_{f,EQ}$	l_{EQ}	$\sigma_{f,P}$	ρ	α				-284
		0.32	2.7	13	0.55	1.5E+02	12				
EQ+P	C+P	σ_n	$\sigma_{f,EQ}$	l_{EQ}	$\sigma_{f,P}$	ρ	α	σ_p	ρ		-284
		0.32	2.7	13	0.55	1.5E+02	12	1	1.5E+02		
RQ+P	C	σ_n	$\sigma_{f,RQ}$	l_{RQ}	ν_{RQ}	$\sigma_{f,P}$	ρ	α			-390
		0.25	5.2	34	0.021	0	1.5E+02	15			
RQ+P	C+P	σ_n	$\sigma_{f,RQ}$	l_{RQ}	ν_{RQ}	$\sigma_{f,P}$	ρ	α	σ_p	ρ	-390
		0.25	7.2	47	0.011	0	1.5E+02	13	1	1.5E+02	
M1+P	C	σ_n	$\sigma_{f,EQ}$	l_{EQ}	$\sigma_{f,P}$	ρ	α				-367
		0.001	1.7	43	15	1.4E+02	4.4				
M1+P	C+P	σ_n	$\sigma_{f,EQ}$	l_{EQ}	$\sigma_{f,P}$	ρ	α	σ_p	ρ		-362
		0.001	1.5	36	9.5	1.5E+02	3.2	1	1.5E+02		
M3+P	C	σ_n	$\sigma_{f,EQ}$	l_{EQ}	$\sigma_{f,P}$	ρ	α				-385
		0.26	2.9	29	0	1.2E+02	12				
M3+P	C+P	σ_n	$\sigma_{f,EQ}$	l_{EQ}	$\sigma_{f,P}$	ρ	α	σ_p	ρ		-385
		0.26	2.9	29	0	1.5E+02	12	1	1.5E+02		
EQ×EQ	C	σ_n	$\sigma_{f,EQ}$	l_{EQ}	$\sigma_{f,EQ}$	l_{EQ}	α				-282
		0.32	1.8	13	1.5	1E+02	12				
EQ×EQ	C+P	σ_n	$\sigma_{f,EQ}$	l_{EQ}	$\sigma_{f,EQ}$	l_{EQ}	α	σ_p	ρ		-282
		0.32	1.8	13	1.5	1E+02	12	1	1.5E+02		
EQ×P	C	σ_n	$\sigma_{f,EQ}$	l_{EQ}	$\sigma_{f,P}$	ρ	α				-306
		0.3	1.6	56	1.6	1.5E+02	12				
EQ×P	C+P	σ_n	$\sigma_{f,EQ}$	l_{EQ}	$\sigma_{f,P}$	ρ	α	σ_p	ρ		-306
		0.3	1.6	56	1.6	1.5E+02	12	1	1.5E+02		
RQ×P	C	σ_n	$\sigma_{f,RQ}$	l_{RQ}	ν_{RQ}	$\sigma_{f,P}$	ρ	α			-324
		0.26	1.7	19	0.11	1.7	1.5E+02	13			
RQ×P	C+P	σ_n	$\sigma_{f,RQ}$	l_{RQ}	ν_{RQ}	$\sigma_{f,P}$	ρ	α	σ_p	ρ	-324
		0.26	1.7	19	0.11	1.7	1.5E+02	13	1	1.5E+02	
M1×P	C	σ_n	$\sigma_{f,EQ}$	l_{EQ}	$\sigma_{f,P}$	ρ	α				-354
		0.001	1.1	96	2.1	1.6E+02	12				
M1×P	C+P	σ_n	$\sigma_{f,EQ}$	l_{EQ}	$\sigma_{f,P}$	ρ	α	σ_p	ρ		-354
		0.001	0.89	89	2.6	1.6E+02	12	1	1.5E+02		
M3×P	C	σ_n	$\sigma_{f,EQ}$	l_{EQ}	$\sigma_{f,P}$	ρ	α				-384
		0.26	1.9	29	1.5	2.1E+03	12				
M3×P	C+P	σ_n	$\sigma_{f,EQ}$	l_{EQ}	$\sigma_{f,P}$	ρ	α	σ_p	ρ		-384
		0.26	1.8	29	1.6	2.7E+03	12	1	1.5E+02		

Table 5: ML estimation on the temperature - Best result in **bold**

MAP Estimation On Temperature

Kernel	Mean	Hyperparameters									Score
EQ+EQ	C	σ_n	$\sigma_{f,EQ}$	l_{EQ}	$\sigma_{f,EQ}$	l_{EQ}	α				-393
		0.25	0.47	3.7	2.8	40	12				
EQ+P	C	σ_n	$\sigma_{f,EQ}$	l_{EQ}	$\sigma_{f,P}$	ρ	α				-251
		0.32	2.3	12	1.5	1.5E+02	10				
EQ+P	C+P	σ_n	$\sigma_{f,EQ}$	l_{EQ}	$\sigma_{f,P}$	ρ	α	σ_p	ρ		-252
		0.32	2.7	13	0.38	1.5E+02	12	0.0001	1.5E+02		
RQ+P	C	σ_n	$\sigma_{f,RQ}$	l_{RQ}	ν_{RQ}	$\sigma_{f,P}$	ρ	α			-367
		0.25	3.7	24	0.043	0.0001	1.1E+02	10			
RQ+P	C+P	σ_n	$\sigma_{f,RQ}$	l_{RQ}	ν_{RQ}	$\sigma_{f,P}$	ρ	α	σ_p	ρ	-361
		0.25	4.1	26	0.035	0.0001	1.5E+02	9.5	0.0001	1.5E+02	
M1+P	C	σ_n	$\sigma_{f,EQ}$	l_{EQ}	$\sigma_{f,P}$	ρ	α				-338
		0.13	1.7	58	46	1.5E+02	7.2				
M1+P	C+P	σ_n	$\sigma_{f,EQ}$	l_{EQ}	$\sigma_{f,P}$	ρ	α	σ_p	ρ		-326
		0.1	1.7	51	46	1.5E+02	7.3	0.0001	1.5E+02		
M3+P	C	σ_n	$\sigma_{f,EQ}$	l_{EQ}	$\sigma_{f,P}$	ρ	α				-363
		0.26	2.9	29	0.0001	1.4E+02	12				
M3+P	C+P	σ_n	$\sigma_{f,EQ}$	l_{EQ}	$\sigma_{f,P}$	ρ	α	σ_p	ρ		-375
		0.26	2.9	29	0.0001	0	12	0.0001	1.5E+02		
EQ×EQ	C	σ_n	$\sigma_{f,EQ}$	l_{EQ}	$\sigma_{f,EQ}$	l_{EQ}	α				-258
		0.32	4.5	13	0.59	1E+02	12				
EQ×EQ	C+P	σ_n	$\sigma_{f,EQ}$	l_{EQ}	$\sigma_{f,EQ}$	l_{EQ}	α	σ_p	ρ		-251
		0.32	4.5	13	0.59	1E+02	12	0.0001	1.5E+02		
EQ×P	C	σ_n	$\sigma_{f,EQ}$	l_{EQ}	$\sigma_{f,P}$	ρ	α				-279
		0.3	1.6	55	1.6	1.5E+02	12				
EQ×P	C+P	σ_n	$\sigma_{f,EQ}$	l_{EQ}	$\sigma_{f,P}$	ρ	α	σ_p	ρ		-271
		0.3	1.6	54	1.5	1.5E+02	12	0.0001	1.5E+02		
RQ×P	C	σ_n	$\sigma_{f,RQ}$	l_{RQ}	ν_{RQ}	$\sigma_{f,P}$	ρ	α			-297
		0.26	1.7	19	0.11	1.7	1.5E+02	12			
RQ×P	C+P	σ_n	$\sigma_{f,RQ}$	l_{RQ}	ν_{RQ}	$\sigma_{f,P}$	ρ	α	σ_p	ρ	-288
		0.26	1.8	23	0.071	1.8	1.5E+02	11	0.0001	1.5E+02	
M1×P	**C**	σ_n	$\sigma_{f,EQ}$	l_{EQ}	$\sigma_{f,P}$	ρ	α				**-408**
		0.16	1.5	2.2E+02	1.9	5.9E+02	12				
M1×P	C+P	σ_n	$\sigma_{f,EQ}$	l_{EQ}	$\sigma_{f,P}$	ρ	α	σ_p	ρ		-325
		0.14	1.5	1.3E+02	1.5	1.6E+02	12	0.0001	1.5E+02		
M3×P	C	σ_n	$\sigma_{f,EQ}$	l_{EQ}	$\sigma_{f,P}$	ρ	α				-355
		0.26	1.8	28	1.6	2E+03	12				

Table 6: MAP estimation on the temperature - Best result in **bold**

References

[1] David Duvenaud. *Automatic model construction with Gaussian processes.* PhD thesis, University of Cambridge, 2014.

[2] Mark Ebden. Gaussian processes for regression: A quick introduction. *The Website of Robotics Research Group in Department on Engineering Science, University of Oxford*, 2008.

[3] Marc G Genton. Classes of kernels for machine learning: a statistics perspective. *The Journal of Machine Learning Research*, 2:299–312, 2002.

[4] Christopher KI Williams and Carl Edward Rasmussen. Gaussian processes for machine learning. *the MIT Press*, 2(3):4, 2006.