GenHack2 - Hackathon for Generative modeling : Simulation of global warming Sea Surface Temperatures Part 2: Conditional Extension

December 4, 2022

1 Introduction

1.1 Context

In the second part of the competition, we extend the previous modelisation to a Conditional Generative model in order to perform simulations of SST at unseen stations.

1.2 Objective

Given a vector of input noise vector $Z \in \mathbb{R}^{d_z}$ and a location variable $Y \in \mathbb{R}^{d_6 \times 2}$, you have to build a conditional generative model G_θ, parametrized by θ, that can simulate realistic samples

$$\widetilde{X}(Y) := G_\theta(Z, Y),$$

similar to a multivariate real climate variable of interest $X(Y) \in \mathbb{R}^6$ conditioned by the location covariate $Y \in \mathbb{R}^{6\times 2}$. The objective is to simulate potential future temperature values (**drawn at random during the year**), conditioned by the station's position, with spatial dependence between stations.

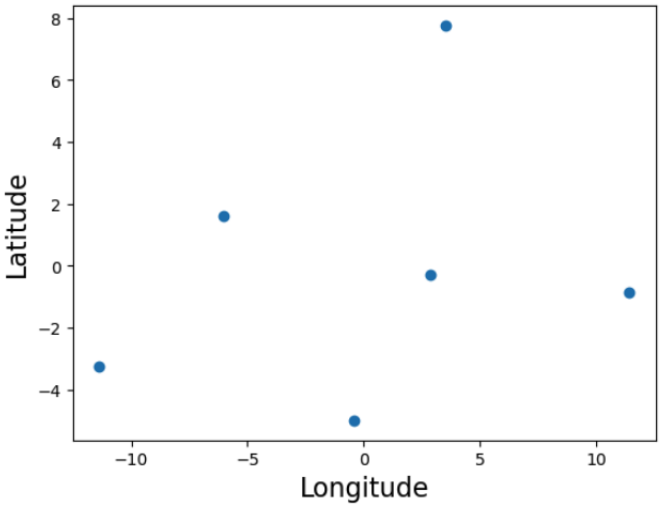

Figure 2: Position of the training stations

2 Data

2.1 Dataset description

We still consider the daily sea surface temperature (SST) in Kelvin from 1981-09-01 to 2016-12-31 (12,541 days) at 6 stations.

2.1.1 Train-test split

We consider the training period from 1981-09-01 to 2016-12-31 (n_{train} =12906 days) at the 6 known stations (see Figure 2) with (centered) coordinates [latitude, longitude] :

$$\begin{aligned} S_{\text{train}} = \Big\{ & S1 : [-3.242, -11.375], \\ & S2 : [-4.992, -0.425], \\ & S3 : [-0.292, 2.875], \\ & S4 : [7.758, 3.525], \\ & S5 : [1.608, -6.025], \\ & S6 : [-0.842, 11.425] \Big\}. \end{aligned}$$

The testing set will contain data from 2008-01-01 to 2016-12-31 (n_{test} =3288 days) at 6 **unknown stations** located at the neighborhood of the training ones. Both the training and the testing stations are located in Figure 3, but **you cannot know where they are.**

Figure 3: Map containing both the training and the testing stations

2.1.2 Data processing

The same **seasonality** at each station **in the training dataset** was removed.

During each evaluation, we will evaluate your model

$$(Z, Y) \in \mathbb{R}^{d_z + d_{6\times 2}} \mapsto G_\theta(Z, Y) \in \mathbb{R}^6,$$

which **must** have the following structure

$$G\left(\begin{bmatrix} Z_1 \\ \vdots \\ Z_{d_z} \\ Y_{\text{lat}_{\text{S1}}} \\ Y_{\text{lat}_{\text{S2}}} \\ \vdots \\ Y_{\text{long}_{\text{S5}}} \\ Y_{\text{long}_{\text{S6}}} \end{bmatrix}\right) = \begin{bmatrix} \widetilde{X}(Y)_{\text{S1}} \\ \vdots \\ \widetilde{X}(Y)_{\text{S6}} \end{bmatrix}.$$

The constraints on the latent dimension ($d_z \leq 50$) and the evaluation score metrics remain the same.

3 General rules

Exactly the same as the previous ones, except few updates:

- `model_cond.py`: new python file containing your conditional generative model and for loading the parameters. You can of course keep the same previous model, but keep in mind that this new model **must** take as input a position variable. **Do modify ✓**

- `data/data_test.csv`: testing data of Evaluations 1 and 2 (2008-01-01 to 2016-12-31 at the 6 known stations)
- `data/position.npy`: coordinates (latitude, longitude) of the 6 known stations.