Project 2: Predicting Lobith Discharge from Atmospheric Data.
Deadline: Friday June 28, 23:59

Starting Notebook: Project_2_pre.ipynb

In this project, you will use atmospheric reanalysis (ERA5) data to predict the discharge of the Rhine river at Lobith (see Fig. 1), using a CNN approach. The data (years 2000-2020) are provided as well as the functions necessary to

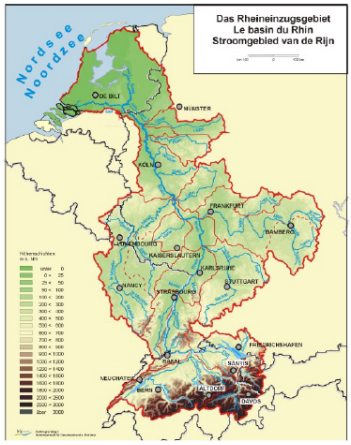

Figure 1: *The catchment area of the Rhine river.*

read it (see notebook "Project2_pre.ipynb").

a.
Make plots of the monthly mean precipitation P, temperature T and volumetric soil water VSW for January 2020.

b.
Determine the time delay (in days) between the discharge Q and the region-averaged precipitation P.

c.
Split the ERA5 data between training (2000-2016) and test (2016-2020). Normalize each grid point of the training and test data in the time dimension, using the mean and std of the training data. The 3 variables (P, T, VSW) must be normalized separately. Normalize the training labels (river discharge) in the time dimension, using their mean and std. Redefine your data samples using a running time window of dimension t. Each sample must have dimension $[t, lat, long, v]$ for TensorFlow or $[t, v, lat, long]$ for Pytorch. t is the dimension of the time window and v is the number of variables. Determine the dimension t based on the results of b.

d. Define a CNN by two Conv3D layers with RELU activation and Maxpooling 3D. Next, a flattening layer followed by a fully connected network with 3 hidden layers with dimensions 256, 128, and 64, respectively. Finally, an output layer with one neuron and a linear activation function. Train the CNN on T and P data with learning rate=10^{-4}, batch size=10, n_epochs=20, loss=mse (mean squared error) and optimizer=Adam. For training, use the normalized training labels. Plot the MSE with respect to the normalized training labels and the MAE (mean absolute error) with respect to the original river discharge values versus the number of epochs.

e. Plot the discharge Q for the test data (2016-2020) against the predicted values and determine the MAE.

f. Repeat c-e. but now including the VSW variable in the feature set.

g. Which variable T, P or VSM gives the dominant contribution to the performance of the CNN in (i) Winter and (ii) Summer?