Assignment 2

Forecasting Models with LamaH dataset

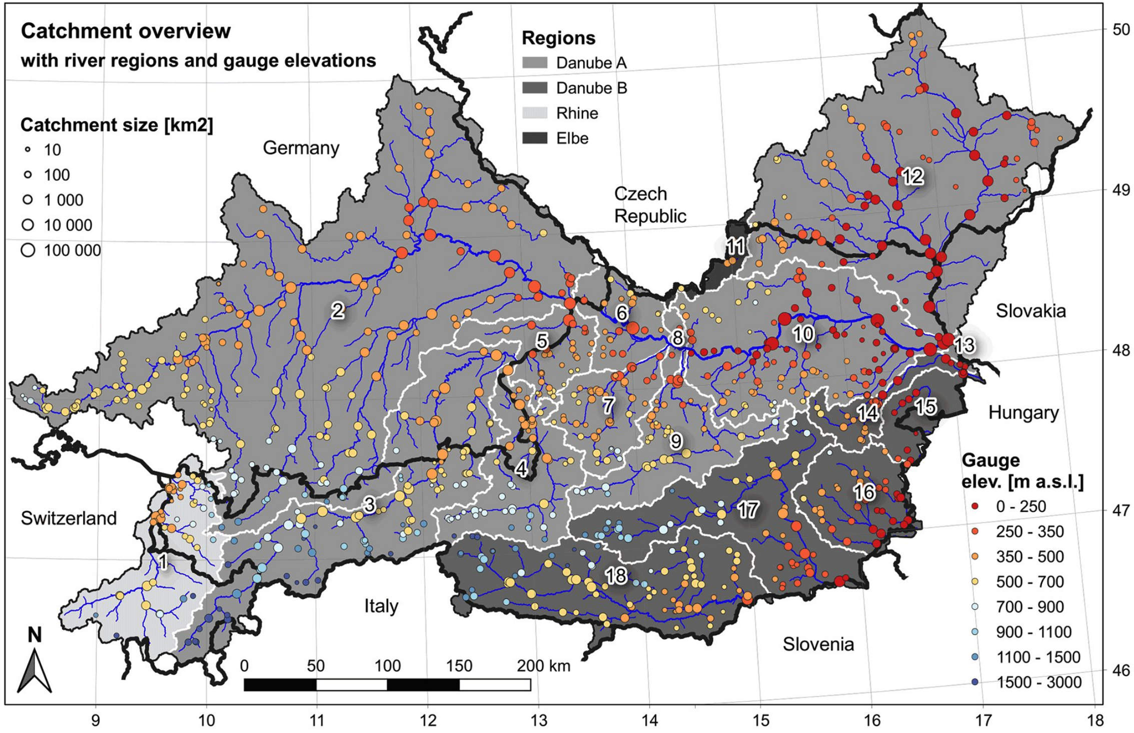
The project 2 aims to analyse the LamaH dataset and predicting the next day’s precipitation

Tasks:

- Preprocessing: analysis & imputation
- Developing ML models to forecast precipitation
- Analyzing the tradeoffs among the models
- Report
- Presentation

We expect all of you to work with the ***same group*** *members from Assignment 1. In case of any changes, please contact us.*

The LamaH Dataset (1)



- Overview of the area covered in LamaH (grey tones), and the runoff gauges with gauge elevation (circle color) and catchment area (circle size)
- LamaH is divided into different river regions, which are bordered by the white lines

The LamaH Dataset (2)

- **Hydrology** and **hydrological processes** are **characterized by high spatiotemporal variability**
- LamaH-CE (LArge-SaMple DAta for Hydrology and Environmental Sciences for Central Europe) → LamaH
- It **covers** the **entire upper Danube** to the state border of Austria–Slovakia, as well as all **other Austrian catchments including** their **foreign upstream areas**
- LamaH **covers** an **area** of **about 170 .000 km^2** in **9 countries**, ranging from lowland regions characterized by a continental climate to high alpine zones dominated by snow and ice

The LamaH Dataset (3)

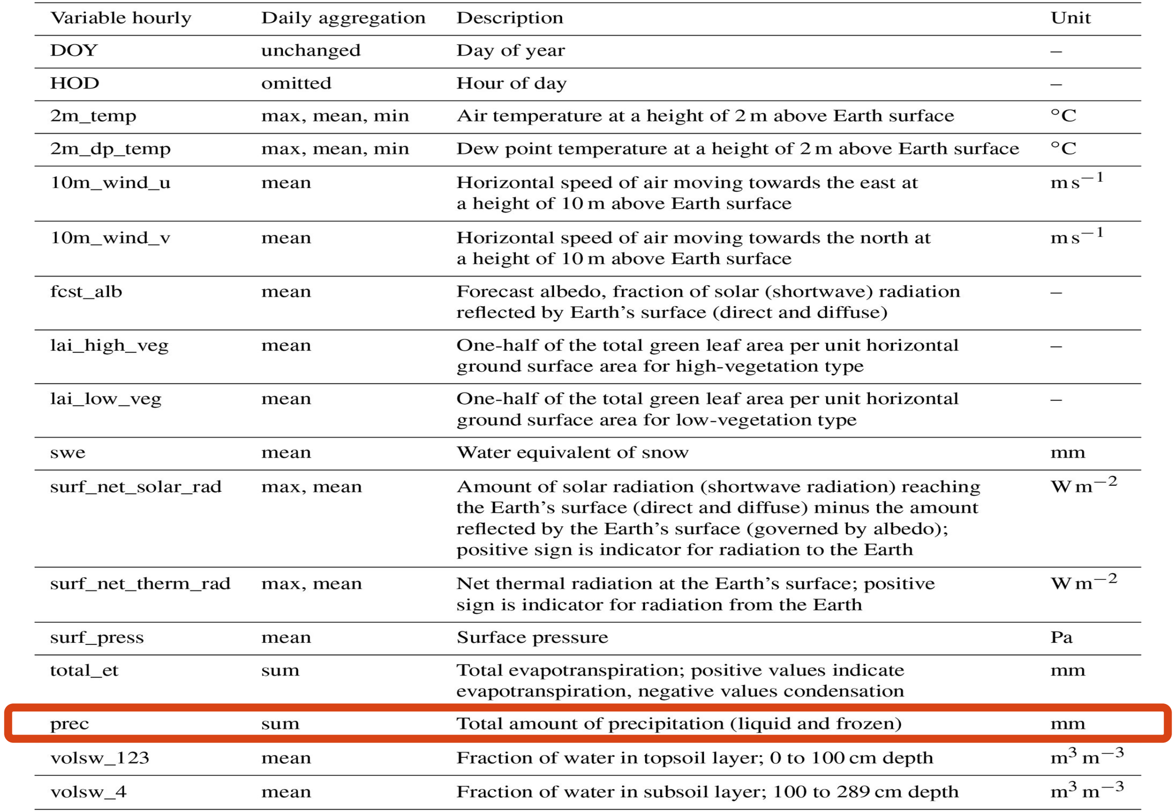
- It represent this variability in **859 gauged catchments** with **over 60 catchment attributes**, covering **topography**, **climatology**, **hydrology**, **land cover**, **vegetation**, **soil** and **geological properties**

- Also **contains** a **collection of runoff time-series** as well as **meteorological time-series-**

- These time-series are provided with a **daily** and **hourly** resolution

- All meteorological and the **majority** of **runoff time-series covers** a span of over **35 years**, which enables long-term analyses.

Meteorological Time-series



Variable hourly	Daily aggregation	Description	Unit
DOY	unchanged	Day of year	–
HOD	omitted	Hour of day	–
2m_temp	max, mean, min	Air temperature at a height of 2 m above Earth surface	°C
2m_dp_temp	max, mean, min	Dew point temperature at a height of 2 m above Earth surface	°C
10m_wind_u	mean	Horizontal speed of air moving towards the east at a height of 10 m above Earth surface	$\mathrm{m\,s^{-1}}$
10m_wind_v	mean	Horizontal speed of air moving towards the north at a height of 10 m above Earth surface	$\mathrm{m\,s^{-1}}$
fcst_alb	mean	Forecast albedo, fraction of solar (shortwave) radiation reflected by Earth's surface (direct and diffuse)	–
lai_high_veg	mean	One-half of the total green leaf area per unit horizontal ground surface area for high-vegetation type	–
lai_low_veg	mean	One-half of the total green leaf area per unit horizontal ground surface area for low-vegetation type	–
swe	mean	Water equivalent of snow	mm
surf_net_solar_rad	max, mean	Amount of solar radiation (shortwave radiation) reaching the Earth's surface (direct and diffuse) minus the amount reflected by the Earth's surface (governed by albedo); positive sign is indicator for radiation to the Earth	$\mathrm{W\,m^{-2}}$
surf_net_therm_rad	max, mean	Net thermal radiation at the Earth's surface; positive sign is indicator for radiation from the Earth	$\mathrm{W\,m^{-2}}$
surf_press	mean	Surface pressure	Pa
total_et	sum	Total evapotranspiration; positive values indicate evapotranspiration, negative values condensation	mm
prec	sum	Total amount of precipitation (liquid and frozen)	mm
volsw_123	mean	Fraction of water in topsoil layer; 0 to 100 cm depth	$\mathrm{m^3\,m^{-3}}$
volsw_4	mean	Fraction of water in subsoil layer; 100 to 289 cm depth	$\mathrm{m^3\,m^{-3}}$

Assignment Tools and Dataset

- Python, pandas, PyTorch, tensorflow, scikit-learn
- Matplotlib / R / matlab (plots & data analysis)
- LaTeX, Word (report)

- Dataset: LamaH-CE_daily.tar.gz
  - A_basins_total_upstrm-> 2_timeseries-> daily-> *.csv

*These are **suggestive** tools and programming environment, you can use any tools that suits your requirements & skill set!*

Assignment 2.1: Data Analysis & Preprocessing

- *Step 1 - **Data Loading***: load 100 random location files from the folder
- *Step 2 - **Null Value Analysis***: analyse and discuss if and which values are missing in the data
- *Step 3 - **Statistical Analysis***: basic data analysis like distributions, correlations…
  - A simple statistical graph or summary table can be depicted here
- Step 4 - **Data Preprocessing**: impute missing values and transform/normalize any data, if needed
- *Step 5 - **Discussing Data Quality & Engineering***: emphasize the importance of understanding data quality and the need for data engineering before model training

Assignment 2.1: Forecasting Models (1)

- *Step 1*:
  - **Model**: at least four types of models – time series / neural networks (NNs) Deep NNs (DNNs) / regression models and any other variants
  - **Dataset**: LamaH daily usage meteorological data
  - **Develop** the models, **measure** the **accuracy for one day ahead**
  - Compare your models with a **naive baseline** of your choise (e.g. MA, t-1, ARIMA, …)

- *Step 2*: perform a **feature importance study**, and report which features are significant in predicting the target variable

Assignment 2.1: Forecasting Models (2)

- Different forecasting models have their own way for presenting feature importance: follow the model specific method and present the table or plots

- Follow the standard ML practices such as 70% - 15% - 15% train, test and validation dataset

- Each model **needs not be** optimized for hyperparameters for initial experiments

- Select the **best performing model** and then optimize hyperparameters of that model to study the best achievable results

- Measure the performance metrics with **different time horizons** for your selected model: 1,3,7 day(s) ahead

Assignment 2: Report

- The report should have at least five sections:
  - *(i) Introduction*
  - *(ii) Background*
  - *(iii) Data Analysis*
  - *(iv) Experiments and Results*
  - *(v) Conclusions*
- In the experiments section, you are expected to provide **at least** three graphs (e.g, scatter plot, bar plot, …) for the following three measurements:
  - Model type vs. Accuracy (e.g. Root Mean Square Error)
  - Feature importance
  - Accuracy of the model before and after hyperparameter tuning
- The report should be between 6-8 pages including the references

You are encouraged to measure any other relevant metrics and add more results and insights!

Assignment 2: Submission and Timeline

Submission

- Report (PDF)
- Presentation file (PDF)
- Source code and artifacts created during the assignment (ZIP archive)

Timeline

- Submission deadline: 16th December 2025
- Presentation: 17th December 2025