Note that for this problem, we utilize data produced by a physics based process model. This is used to simulate stream temperature values as stream temperature data in the real world often contains many missing values. This data is provided in the dwallin\_stream\_preds.csv. We will also call this as the pretraining stream temperature throughout this document.

The purpose of using physics based simulated data for pretraining is to guide our model to learn from the stream temperature data with higher efficiency. We then finetune on real-world data to ensure that our model can learn any inaccuracies from the simulated data and rectify itself on the real-world.

The input data for pretraining and finetuning is contained in the following files:

* gridMET\_area\_weighted.csv
  + Minimum air temperature
  + Maximum air temperature
  + Shortwave radiation
* reservoir\_releases\_total.csv
  + Reservoir release values
* temperature\_observations\_forecast\_sites.csv
  + Ground truth observed stream temperature
* dwallin\_stream\_preds.csv
  + Simulated ground truth stream temperature

Forecasting data is contained in the following files:

* forecast\_release\_data.csv
  + Reservoir release values for forecasting
* forecast\_data\_E{0-30}.csv
  + Minimum air temperature
  + Maximum air temperature
  + Shortwave radiation
  + Stream temperature

The stream temperature preparatory Colab notebook will provide some steps to preprocess the files for pretraining and finetuning data. This includes ensuring all temperature variables are in Celsius, ensuring that pretraining and finetuning is within the appropriate time range, and so on.

Other preprocessing steps for infilling the data (i.e., filling in missing values in the finetuning stream temperature using the pretraining stream temperature) or normalizing must be done in the group project. You will find this significantly boosts your model’s performance and training efficiency.

Feel free to explore how to preprocess the forecast data, by taking the time to ensure that all variables are in the appropriate range and so.