# Data Integration and Interpolation for Lake Quality Modeling

## 1. Introduction

This report describes the process of integrating multi-source datasets for weekly lake monitoring, followed by interpolation of missing values to support machine learning models such as LSTM. The goal is to build a reliable, complete, and temporally consistent dataset that includes climate drivers, lake physical characteristics, and quality indicators like chlorophyll-a (CHLA) and turbidity.

## 2. Dataset Merging

Weekly time series from the following sources were merged:  
- ERA5 climate data (2002–2022)  
- CHLA weekly means  
- Turbidity weekly means  
- LSWT weekly means (optional)  
  
Only lakes with available CHLA and turbidity data were retained. LSWT and other ERA5 variables were allowed to be missing and interpolated later. Merging was performed on a per-lake basis by matching on 'Lake\_ID', 'Year', and 'Week'.

## 3. Lake Parameters Integration

Each lake’s metadata (e.g., surface area, volume, depth, slope, watershed area) was integrated into the weekly files. This was done by joining static physical parameters to the corresponding time series using the 'Lake\_ID' key.

## 4. Interpolation of Missing Values

To prepare the dataset for time-series modeling:  
- Linear interpolation (limit=4 weeks) was used on all numeric time series columns.  
- Predictors such as 'lswt\_mean', 'temperature\_2m', and runoff metrics were additionally forward and backward filled.  
- If no values existed for a predictor (i.e., entire column NaN), remaining values were filled with 0.  
- Target variables 'chla\_mean' and 'turb\_mean' were NOT force-filled to avoid introducing artificial trends. Missing target values remain NaN and will be handled during training by excluding those sequences.

## 5. Output

Each lake is stored as a separate CSV file in the 'Interpolated\_Lake\_CSVs\_Strict' folder.  
These files include:  
- All merged weekly data from ERA5, CHLA, TURB, LSWT (if available)  
- Integrated static lake parameters  
- Cleaned and interpolated predictor variables  
- Preserved missing values for targets