LSTM Model for CHLA Prediction with Fire Predictors

# 1. Overview

This report documents the training and evaluation of an LSTM model for predicting Chlorophyll-a (CHLA) concentrations in lakes using both environmental and fire-related predictors. The model is trained on a balanced dataset that includes both fire-affected and no-fire lakes, enabling it to generalize across varying conditions.

# 2. Dataset Preparation

Weekly time series datasets were prepared for each lake by combining CHLA, LSWT, ERA5 climate predictors, lake-specific metadata, and top 5 fire-affected land cover indicators from FireCCI. The data was aggregated into a structured format per lake and split into two groups: fire-affected and no-fire lakes.

Each group was used to build LSTM sequences of 5 weeks in length. Seasonality features (week\_sin, week\_cos) were also added. The resulting datasets were normalized and split into training and testing subsets.

# 3. Model Architecture

The model architecture is based on an LSTM network enhanced with lake embeddings (Embedding Fusion). Each lake is assigned a unique embedding vector that is learned during training. These embeddings are fused with LSTM outputs to incorporate lake-specific behavior into predictions.

Key components:

* - 2-layer LSTM with 64 hidden units
* - Lake embedding layer (16 dimensions)
* - Dense layers to merge LSTM and embedding features
* - Final output layer for CHLA prediction

# 4. Training Procedure

The LSTM was trained using the Adam optimizer over 20 epochs. Mean Squared Error (MSE) was used as the loss function. Training was done on a balanced dataset (88 lakes from each group) to prevent model bias toward no-fire lakes.

# 5. Evaluation Results

Evaluation was done separately on fire and no-fire test sets using the model trained on combined data. The following metrics were used:

* - RMSE (Root Mean Squared Error)
* - R² (Coefficient of Determination)
* - rRMSE (Relative RMSE, % of mean CHLA)

The model showed significant improvement on fire-affected lakes compared to the no-fire-only model, validating the importance of including fire predictors in training.

# 6. Notes on Lake Embedding Fusion

Lake embeddings allow the model to learn unique characteristics of each lake (e.g., size, depth, elevation) without explicitly encoding all metadata in the input sequence. The concept is similar to entity embeddings in NLP and recommendation systems. Reference: Guo & Berkhahn (2016), 'Entity Embeddings of Categorical Variables'.