OR 610 – Deep Learning

Student name: Tuan Nguyen

Proposal: Long-term Rainfall Prediction Using LSTM Deep Learning Model.

# I. Introduction

## 1.1 Rainfall prediction and challenges

Rainfall, or precipitations, are important in many different aspects of life and can have impacts in many industries such as agriculture, environment, industrial productions, constructions, transportation commercial airlines, ...etc. Excess or lack of rainfall can cause short and long term problems to affected areas in forms of floods or draughts. Therefore, accurately predicting in advance when rainfall events would occur can help mitigates economic and infrastructure damage as well as avoid human life lost.

Numerous existing research and literature have performed short term rainfall predictions (up to 7 days ahead, definition varies depending on agency) to great success. However, the performance of models that predict rainfall for longer periods (more than 14 days) is largely dependent on the quality of recorded data, the location selected to perform prediction as well as its climate characteristics. The main difficulty in long-term rainfall predictions is due to the butterfly effect that exist in weather conditions, where a small changes in initial conditions can lead to significant changes in future conditions, which in turn makes long term forecast volatile and less predictable.

## 1.2 Available prediction methods

The most common method of rainfall prediction currently employed globally is based on physical simulation models (more commonly known as the field of meteorology), where other weather information (cloud radar images, water vapor satellite images, ocean currents, temperature, windspeed, …etc.) is used to construct a simulation of the cloud movement and atmospheric conditions, which will then use to predict the probability of rainfall occurring at a given time interval. However, this method relies on iterative predictions, where one weather state would be used to predict the next state. This method is heavily influenced by the butterfly effect, when minor errors in initial iteration adds up overtime to product completely different results compare to actual future weather states.

A new emerging trend is to use statistical models to directly predict rainfall without relying on simulation output. These types of models use historical recorded weather data as input to predict future weather conditions. In the case of rainfall prediction, different ML/DL models such as GLM, SVM, RNN, LSTM have been used to predict long-term rainfall in many different areas in the world, using many different indicators such as windspeed, humidity, dew point, temperature as input, to different degree of success. Existing literature in the past 5 years shown that LSTM have proven to be effective in modeling relationships between current hydrological measurements and future rainfall events.

However, compare to physical based models, statistical models are not without drawbacks. This type of model can produce result faster at a cost of higher bias, as physical based model prediction and quickly adapts to new changes in weather states, whereas statistical model prediction would not be shifted significantly due to the small effects of new data compare to decades or centuries of recorded historical data.

# II. Model proposal

## Proposed approach.

Unlike most existing models which use other weather condition indicators such as humidity, temperatures, …etc., or simplify the long-term prediction problem by only calculating monthly average, the proposed approach for this project is to only use past recorded rainfall data to predict future rainfall, using a LSTM network or its variants. A 30-day future prediction will be calculated based on 60 days prior to the interested period. Ex: to make predictions for period from 1st Jan to 30th Jan, data from 2nd Nov to 31st Dec of previous year will be used, and for the period from 2nd Jan to 31st Jan, input data will be shifted 1 day to 3rd Nov -1st Jan.

The reason for only selecting rainfall as input is that this project could potentially serve as a proof-of-concept model that it is possible to predict long term future rainfall day by day without using physical model or relying on other predictors. Such model could be implemented in developing countries where state of the art weather monitoring stations and equipment are not always available.

The location chosen for this project is New Orleans, with the data collected from Louis Armstrong Airport weather station. This data is collected by NOAA and made publicly available from their website at [URL](https://www.ncdc.noaa.gov/cdo-web/datasets/GHCND/stations/GHCND:USW00012916/detail). The period of data used is from 1979 to 2022, with data from 1979-2015 will be used for training, while data from 2015 to 2022 will be used for testing and validation.

## Proposed methodology

Below is the proposed task list required to complete the project:

* Gather data from 1979 – 2022.
* Clean the data for anomalies.
* Select input – output sizes.
* Reformat data to correct dimensions
* Split data into train – test – validation sets.
* Select loss function.
* Create LSTM model architecture.
* Grid search for suitable hyperparameters
* Finalize the model and measure performance.

The most difficult step will be to create model architectural, if this doesn’t work out, an alternative would be to change the model architecture to a deep learning feed forward NN.

The 2nd most difficult would be to select input – output size, current proposal suggests input of 60 days prior, and output is 30 days ahead. This would mean that most of the output days would be used as inputs, which might lead to some unforeseen complications during prediction/validation.

Additionally, LSTM model will output a hidden state which would need to be transformed into scaler by a linear layer. This means that if the goal is 30 days ahead prediction, it will require running the model iteratively 30 times to get the desired output, which can potentially fall into the same problem of experiencing butterfly effect similarly to physical based models.

The metrics that will be used to measure model performance are Root-mean-square Error (RMSE) and Mean Absolute Percentage Error (MAPE), as these are common metric used in statistical rainfall prediction models.

## Possible experience

Below is the list of potential experience that I would be able to obtain during the course of this project:

* Gather data from real life source.
* Data preprocessing and visualization in Python.
* Design and implement deep learning techniques on a real-life problem.
* Learn how to implement and work with LSTM models.

Notes:

* Judge model performance at 7 – 14 – 21 - 28 days marks instead of just 30 days
* Define input, sequences of input