Proposal

Problem: Predicting wave conditions, such as wave height, period, and direction, is crucial for surfers. Traditional numerical models are not always accessible. By leveraging historical buoy data, I aim to develop a more efficient Long Short-Term Memory (LSTM) model to forecast wave conditions for surf spots, using meteorological and spectral wave data from NOAA Station 46026 near San Francisco, CA.

Proposed Method: I will use a Long Short-Term Memory (LSTM) neural network, which is well-suited for time series forecasting due to its ability to capture temporal dependencies and trends. The LSTM model will be trained on historical data to predict future wave height, period, and direction. The dataset will include several years of historical data from NOAA, featuring wind speed, wind direction, water temperature, and wave-specific parameters like dominant wave period and mean wave direction. The approach will also account for wave transit time adjustments to tailor predictions specifically for nearby surf spots, ensuring that the model forecasts the conditions surfers would actually experience rather than just the buoy data.

Potential Obstacles:

1. **Wave Transit Time Adjustments**: Accurately calculating wave transit time to nearby surf spots can be challenging and may require fine-tuning.
2. **Model Overfitting**: Due to the high complexity of the LSTM model, ensuring generalizability across different seasons and weather patterns will require careful hyperparameter tuning and validation.
3. **Limited Real-Time Integration**: Incorporating real-time data in a timely manner could present challenges in building a robust pipeline.

Novelty: This project is unique in combining LSTM's ability to model temporal dependencies with wave transit time adjustments to tailor predictions specifically for surf spots near a given buoy, rather than just replicating existing buoy data.

Station Overview:

- Station ID: 46026 (San Francisco, CA)
- Data Type: Standard Meteorological Data which includes:
    - Year, Month, Day, Hour, Minute, Wind Direction (WDIR), Wind Speed (WSPD), Wind Gust Speed (GST), Significant Wave Height (WVHT), Dominant Wave Period (DPD), Average Wave Period (APD), Mean Wave Direction (MWD), Pressure (PRES), Air Temperature (ATMP), Water Temperature (WTMP), Dew Point (DEWP), Visibility (VIS), and Tidal Level (TIDE)

Plan:

Phase 1: Data Collection and Preprocessing:

- **Data Source**
  - NOAA Standard Meteorological Data:
  - https://www.ndbc.noaa.gov/historical_data.shtml
- **Download Data**: Collect standard meteorological data from the links provided for the years 2013 to 2023.
- **Combine and Clean**:
  - Extract the `.gz` files to obtain `.txt` files.
  - Load each file into a `pandas` DataFrame, and concatenate them into a single DataFrame.
- **Standardize the column names for consistency**.
- **Handle missing values**
- **Create a Timestamp**:
  - Combine `Year`, `Month`, `Day`, and `Hour` columns into a single `datetime` index.
- **Transit Time Adjusted Features**:
  - Calculate wave transit time using group velocity and adjust wave parameters (WaveHeight, Period, and Direction) for expected arrival time and wave data at the nearby surf spot.
  - By shifting each observation from the buoy by the estimated transit time, the dataset represents surf spot conditions The model will then use these adjusted values to make predictions, ensuring that it forecasts the surf spot's conditions rather than just the buoy's.
- **Feature Engineering:**
  - Create lagged features for wave height, period, and direction (e.g., `WaveHeight_1hr_lag`).
  - Calculate rolling averages (e.g., 3-hour rolling average of wave height).

Phase 2: Model Development Using LSTM

- Problem Setup:
  - Use the past `n` hours (e.g., last 48 hours) of wave height, period, and direction to predict the next 6, 12, and 24 hours at the surf spot.
- LSTM Model Architecture:
  - Input shape: `(number_of_time_steps, number_of_features)`, e.g., `(48, 3)`.
  - Use multiple LSTM layers followed by dense layers for final predictions.
- Model Training
  - Split the combined dataset into training and validation sets.

    - Use metrics like Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE) to evaluate model performance.

Phase 3: Model Evaluation

- Evaluate the model on a hold-out test set to ensure generalization.
- Compare the LSTM model with a baseline model, such as a simple linear regression or moving average model.
- Use MSE and R^2 to evaluate accuracy

Phase 5: Real-Time Data Integration (Might do)

- Once the model is validated on historical data, set up a pipeline to pull real-time data from NOAA Station 46026
- Implement a real-time forecasting system that automatically updates the rolling window with new observations and generates predictions.
- Develop a simple dashboard using `Streamlit` or `Dash` to display: Predicted wave height, period, and direction for the next few hours.
- Comparison with actual observations.
- Implement alert mechanisms for surfers based on ideal conditions (e.g., wave height > 2m).

Implementation Details

1. Tools & Libraries:
   - Python: For core programming and model development.
   - Pandas & NumPy: For data manipulation and cleaning.
   - Keras or Pytorch: For building and training the LSTM model.
   - Matplotlib & Seaborn: For visualizations.
   - Streamlit/Dash: For dashboard development.
   - NOAA API: For accessing real-time data.

Potential Extensions

1. Multi-Buoy Forecasting: Include data from nearby buoys to improve model accuracy and for use at different surf spots.