Forecasting the Weather Parameters of Jena City using ML (RNN)

Team members:

Adabala Raja Venkata Sai Naresh

Govardhan Butta

Problem:

- Forecasting the weather data (Temperature, Air density, …) in urban cities is of vital importance for pollutant movement studies [1].
- Environmental Scientists who are studying about air pollutants do need predicted weather data for better modelling the pollutants movement in city and in determining the pollution concentration.
- Pollution is a huge problem in Cities and City residents will have to take necessary steps (like: to avoid jogging on bad air quality days or wearing mask) to protect themselves from harmful pollutants.
- [1] *https://scied.ucar.edu/learning-zone/air-quality/how-weather-affects-air-quality*

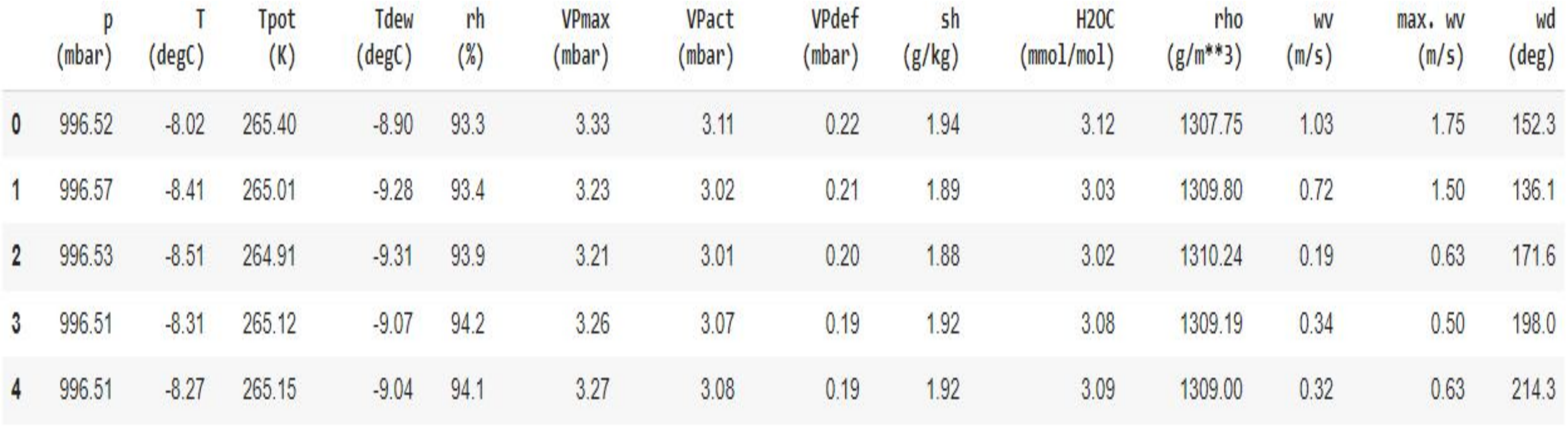
Dataset Available

- Jena Climate dataset is weather time series dataset recorded at the Weather Station of the Max Planck Institute for Biogeochemistry in Jena, Germany.

What is in data ?

- Jena Climate dataset is made up of 14 different quantities (such air temperature, atmospheric pressure, humidity, wind direction, and so on) were recorded every 10 minutes, over several years. This dataset covers data from January 1st 2009 to December 31st 2016.

Data



	p (mbar)	T (degC)	Tpot (K)	Tdew (degC)	rh (%)	VPmax (mbar)	VPact (mbar)	VPdef (mbar)	sh (g/kg)	H2OC (mmol/mol)	rho (g/m**3)	wv (m/s)	max. wv (m/s)	wd (deg)
0	996.52	-8.02	265.40	-8.90	93.3	3.33	3.11	0.22	1.94	3.12	1307.75	1.03	1.75	152.3
1	996.57	-8.41	265.01	-9.28	93.4	3.23	3.02	0.21	1.89	3.03	1309.80	0.72	1.50	136.1
2	996.53	-8.51	264.91	-9.31	93.9	3.21	3.01	0.20	1.88	3.02	1310.24	0.19	0.63	171.6
3	996.51	-8.31	265.12	-9.07	94.2	3.26	3.07	0.19	1.92	3.08	1309.19	0.34	0.50	198.0
4	996.51	-8.27	265.15	-9.04	94.1	3.27	3.08	0.19	1.92	3.09	1309.00	0.32	0.63	214.3

No of Data Points : 420551
No of Features : 14

All about data

0. Date Time: date-time reference

1. p (mbar): The pascal SI derived unit of pressure used to quantify internal pressure. Meteorological reports typically state atmospheric pressure in millibars.

2. T(degC): Temperature in Celsius

3. Tpot (K): Temperature in Kelvin

4. Tdew (degC): Temperature in Celsius relative to humidity. Dew Point is a measure of the absolute amount of water in the air, the DP is the temperature at which the air cannot hold all the moisture in it and water condenses.

All about data

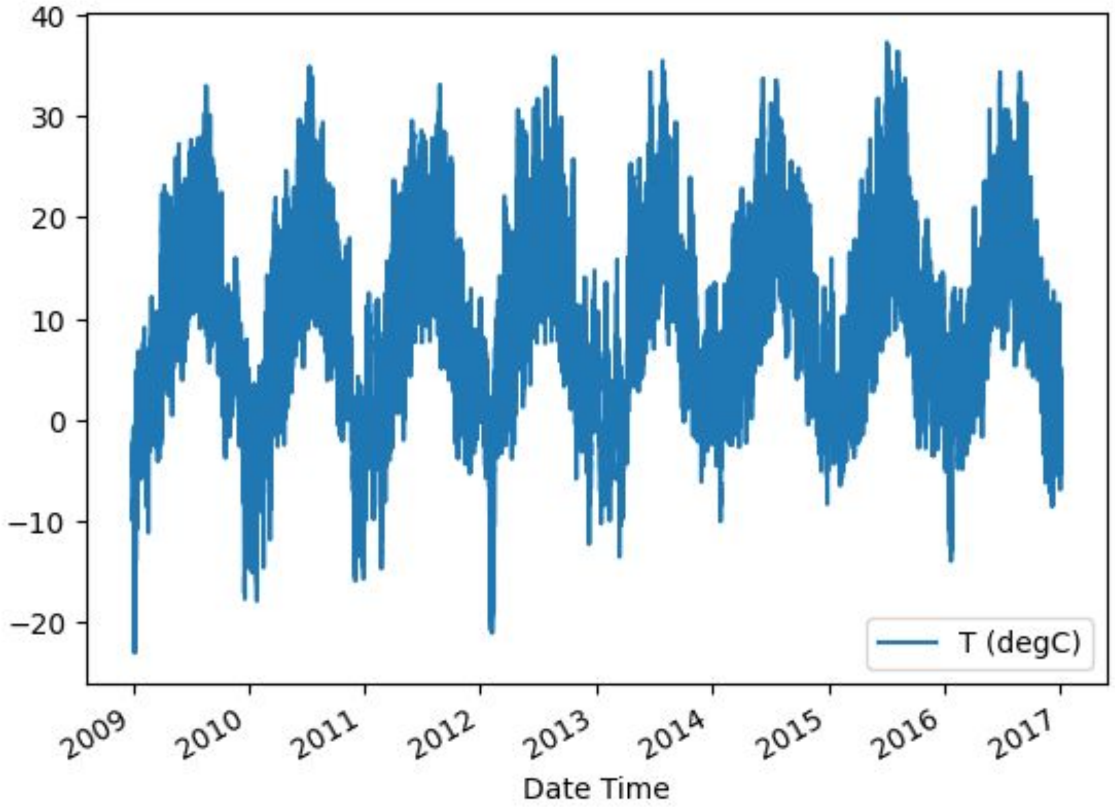
5. **rh (%):** Relative Humidity is a measure of how saturated the air is with water vapor, the %RH determines the amount of water contained within collection objects.
6. **VPmax (mbar):** Saturation vapor pressure
7. **VPact (mbar):** Vapor pressure
8. **VPdef (mbar):** Vapor pressure deficit
9. **sh (g/kg):** Specific humidity
10. **H2OC (mmol/mol):** Water vapor concentration

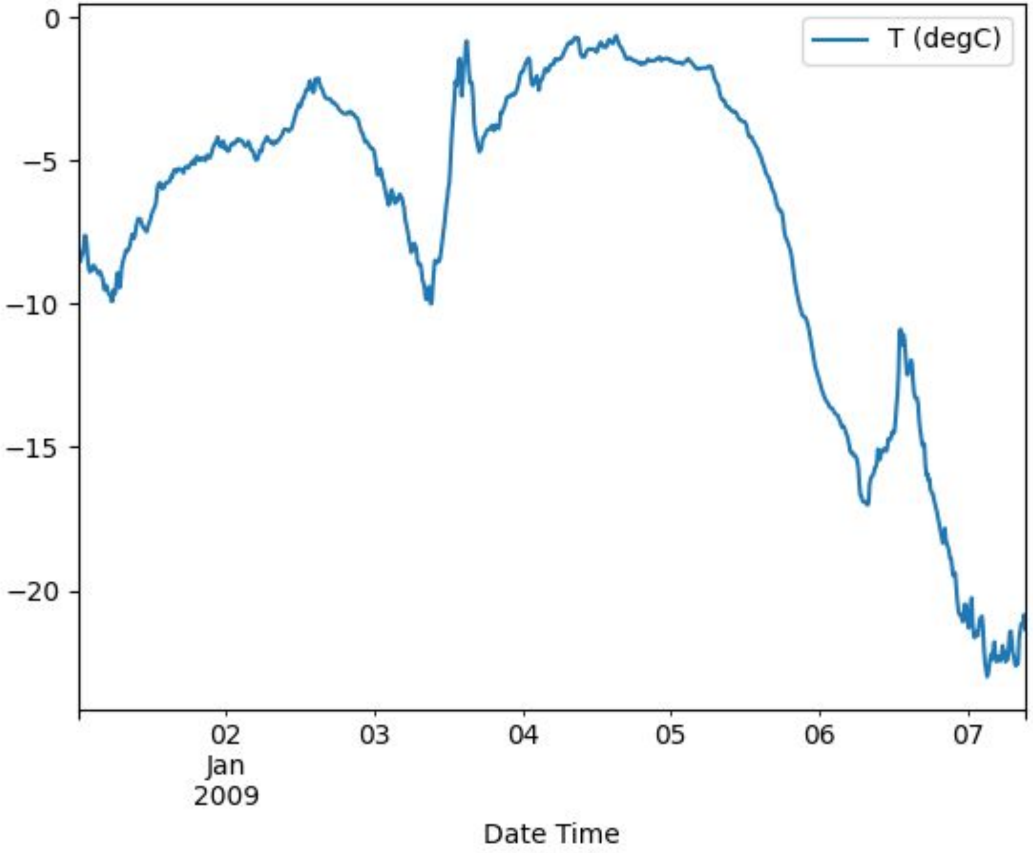
All about data

11. **rho (g/m ** 3):** Air Tight (Density)
12. **wv (m/s):** Wind speed
13. **max. wv (m/s):** Maximum wind speed
14. **wd (deg):** Wind direction in degrees

Data Visualization:

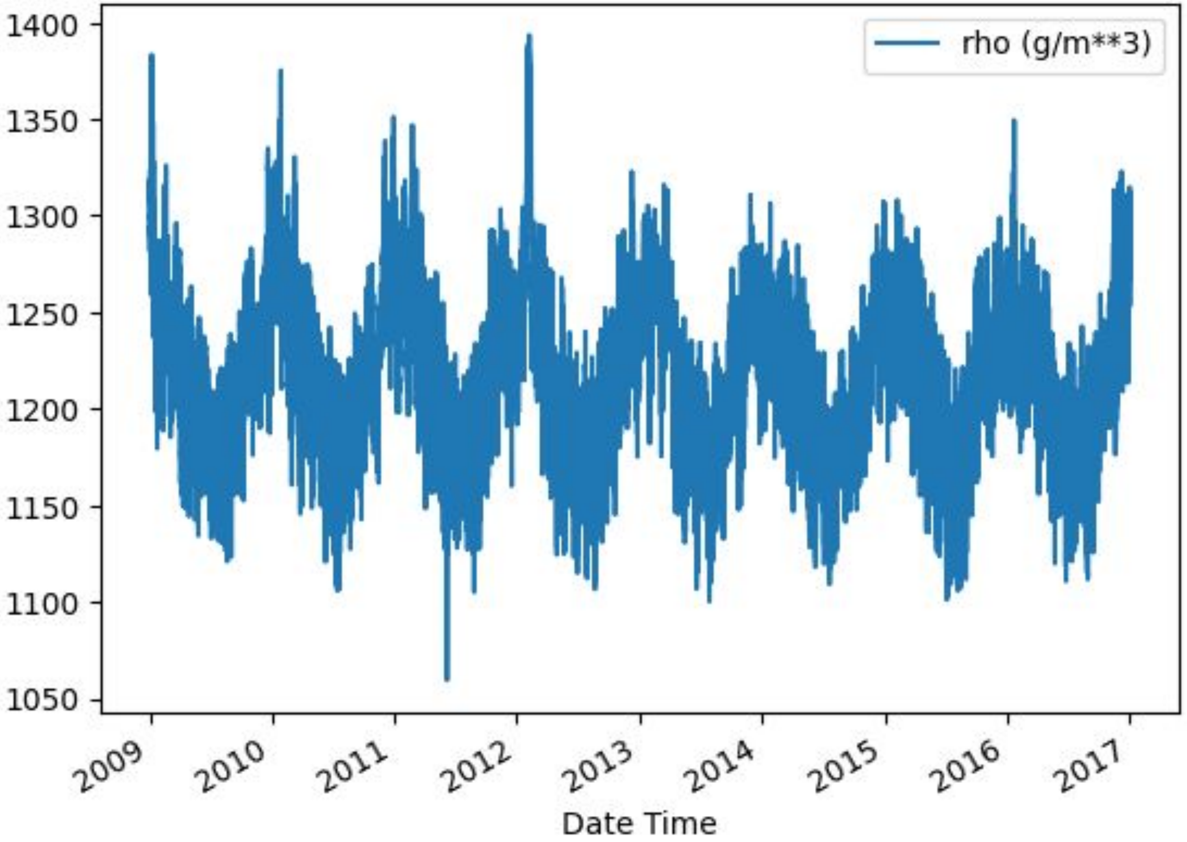
- **Temperature**

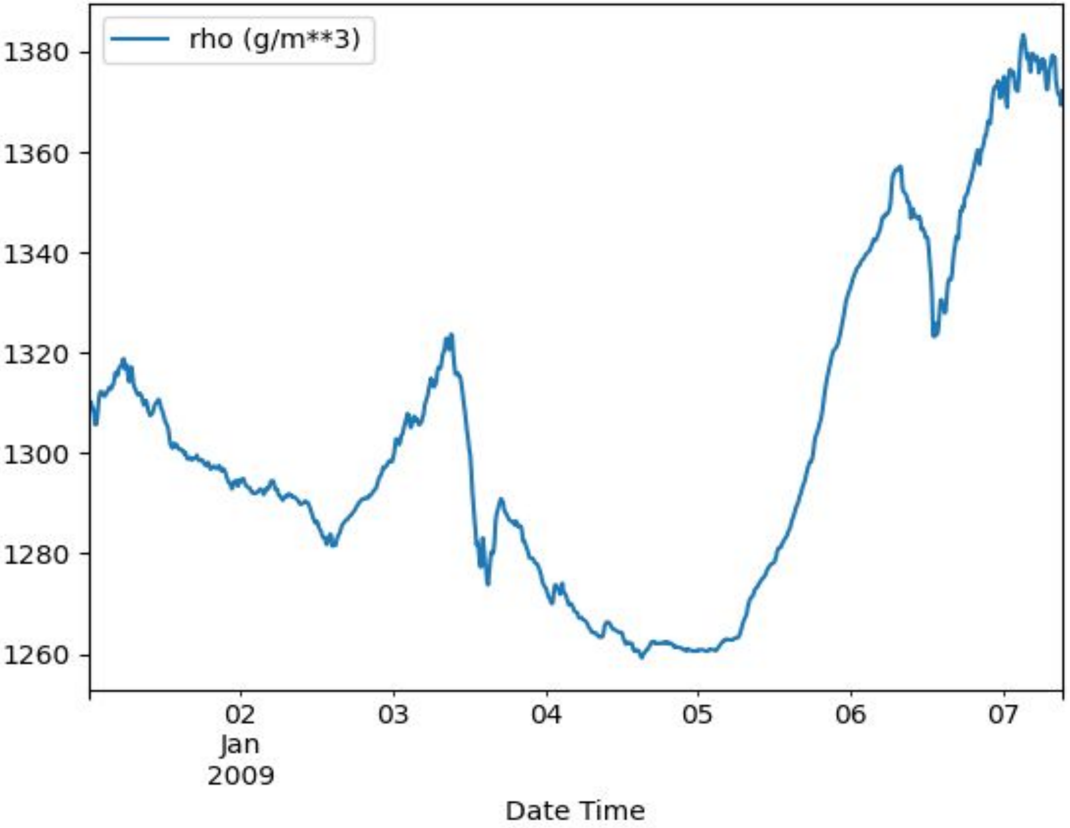




Data Visualization

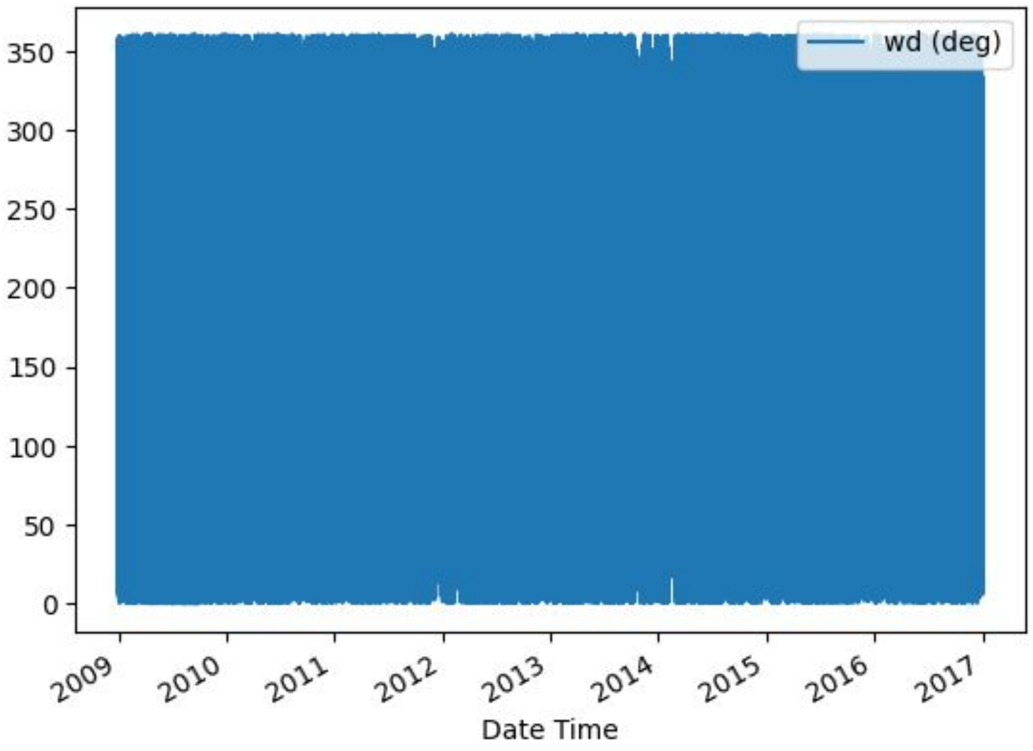
- **Air Density (Tight)**

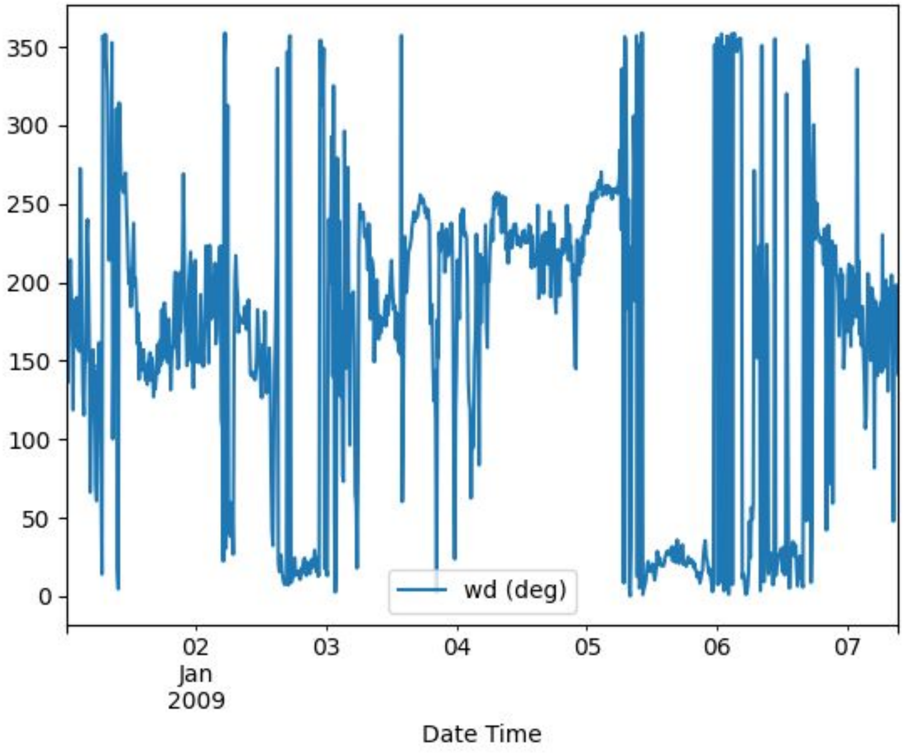




Data Visualization

- **Wind Direction**:



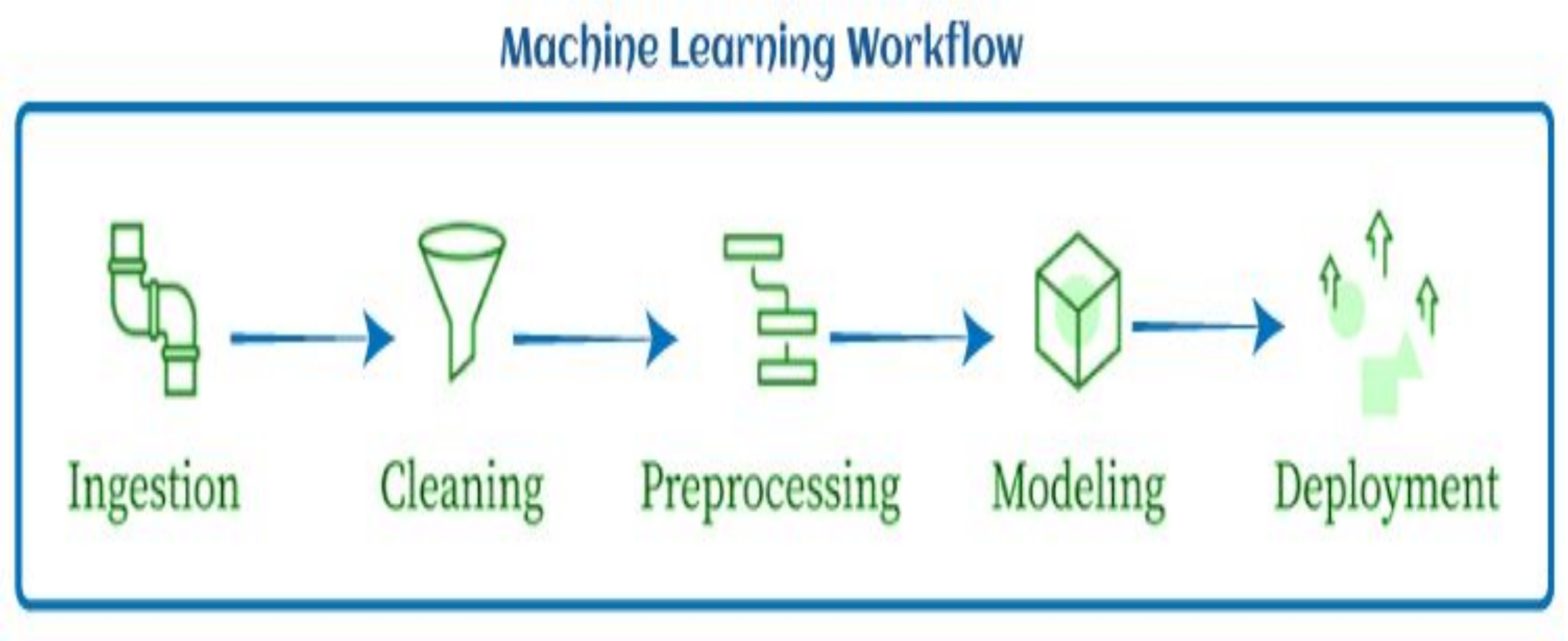


Goal of the project

- Given a time series of hourly measurement of various atmospheric parameters, predict the temperature, air density, and wind direction 24 hours in the future using **DL based Time Series model (RNN)**

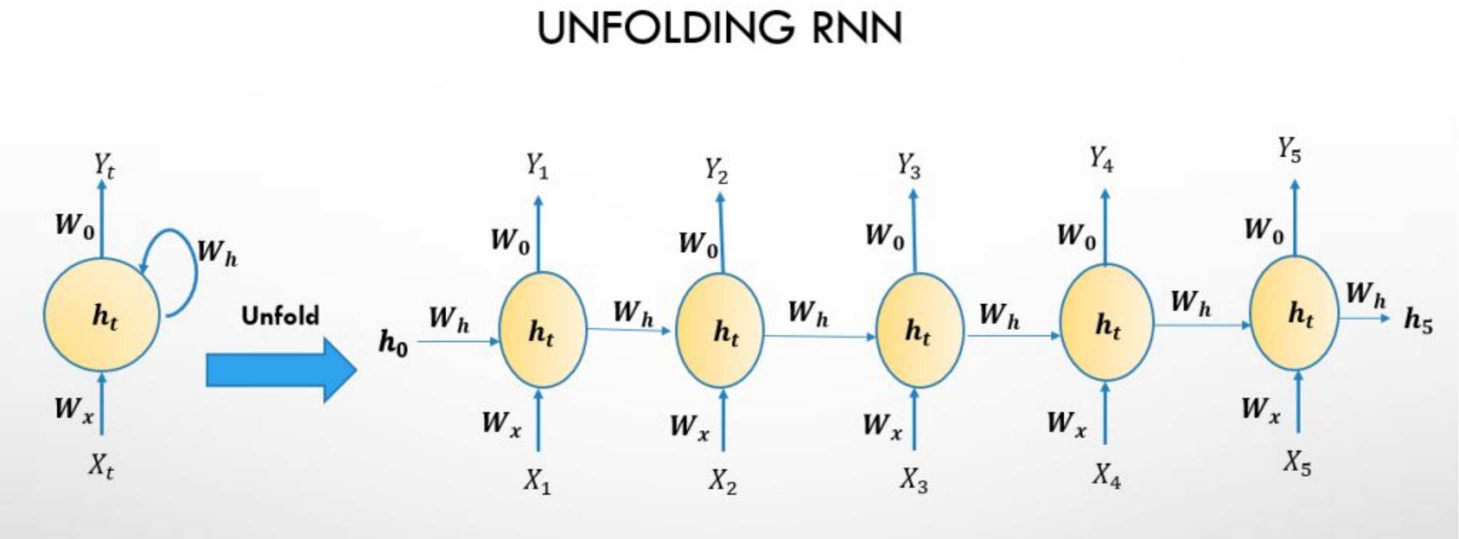
Different Kinds of Time Series

- Classification
- Event Detection
- Regression (Our Problem comes under this task)

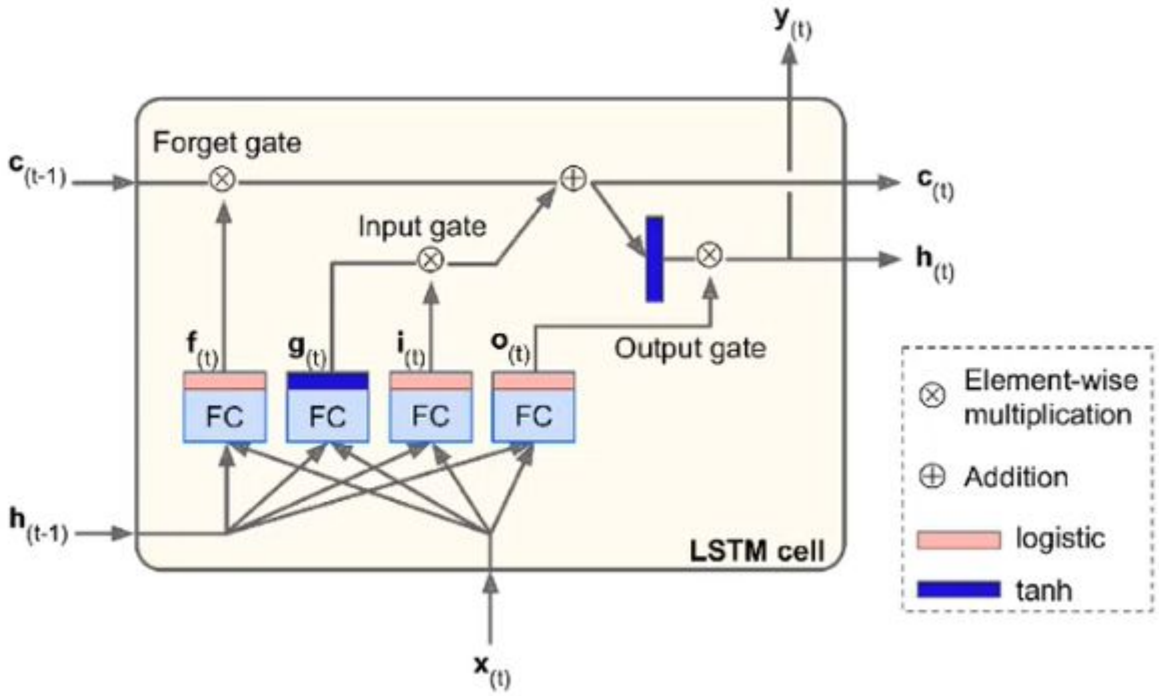

Machine Learning Workflow
Ingestion
Cleaning
Preprocessing
Modeling
Deployment

Recurrent Neural Network (RNN)

- RNNs perform better than Dense or Convolutional Network for time series data. [References]
- A recurrent neuron receives an input and produces an output and sends it back to itself.



LSTM (Long Short Term Memory)

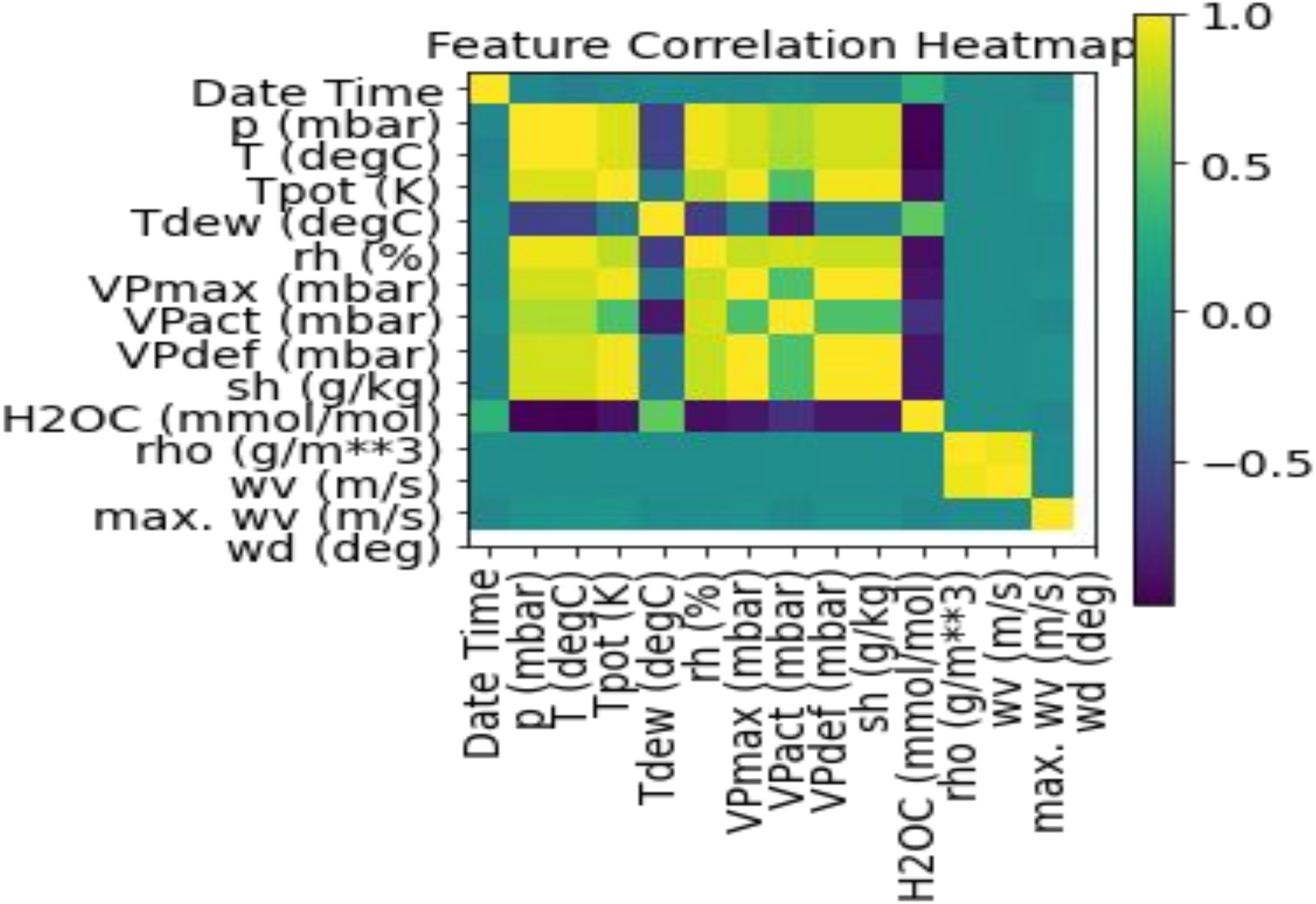


A common sense , Non Machine Learning Baseline

- Prediction based on the assumption that the temperature changes periodically.
- Hence, the temperature prediction of 24 hr later should be equal to the temperature right now.

Data Split:

Data	Percentage	Number of Samples
Train Data Set	55%	231303
Validation Data Set	25%	105137
Test Data Set	20%	84111

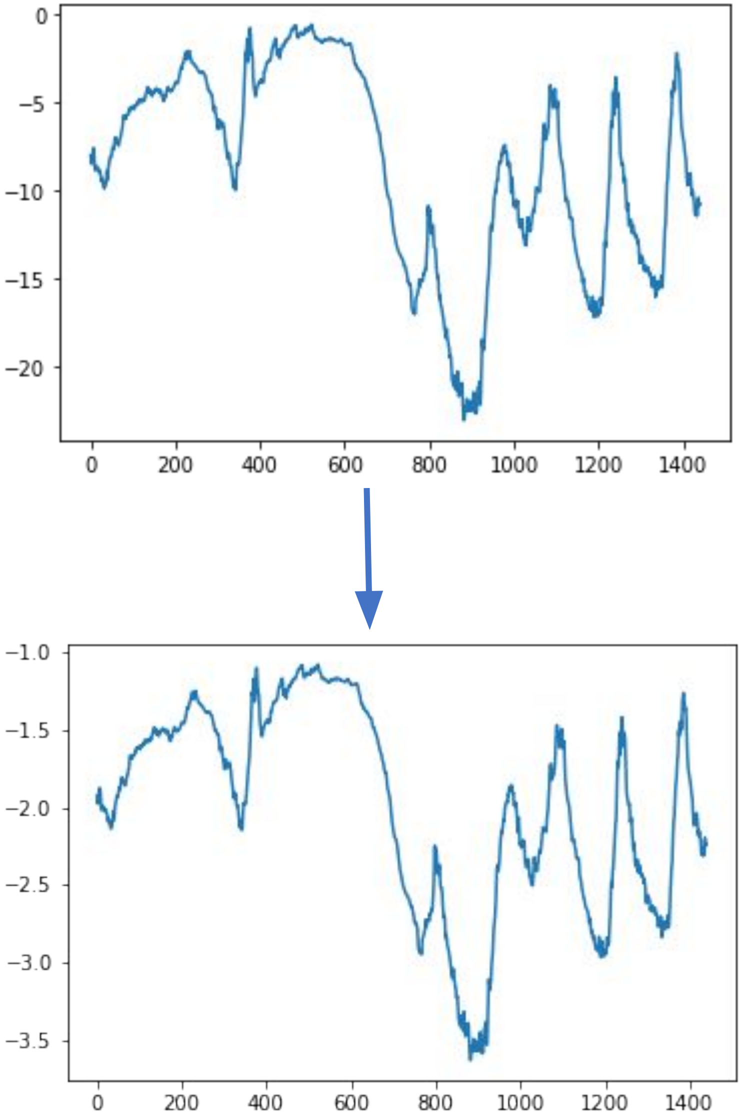

Feature Correlation Heatmap
Date Time
p (mbar)
T (degC)
Tpot (K)
Tdew (degC)
rh (%)
VPmax (mbar)
VPact (mbar)
VPdef (mbar)
sh (g/kg)
H2OC (mmol/mol)
rho (g/m**3)
wv (m/s)
max. wv (m/s)
wd (deg)
1.0
0.5
0.0
−0.5

Normalizing the data:

$$\boxed{x' = (x - \mu) / \sigma}$$

Training set:

- `train_mean = raw_data[:num_train_samples].mean(axis=0)`
- `train_std =  raw_data[:num_train_samples].std(axis=0)`
- `raw_data -= train_mean`
- `raw_data /= train_std`



Normalized data of Temperature

Instantiating datasets for training, validation, and testing

- Sampling rate: 6 (only for data & not targets)
- Sequence Length: 144

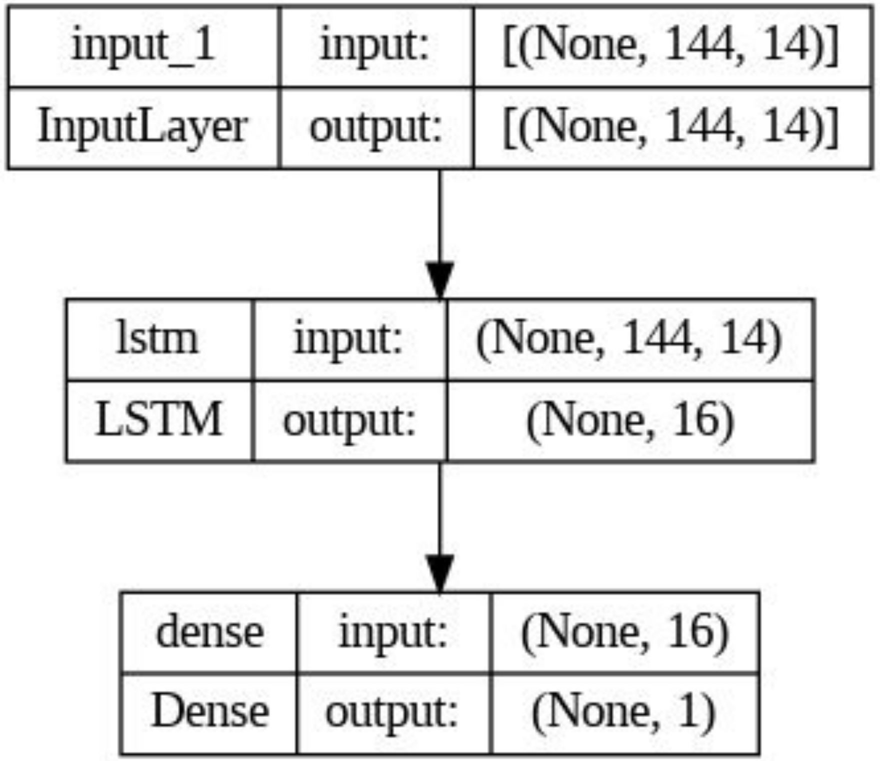
(data is recorded every 10 minutes, 24X6 = 144 data points per day)

- batch size : 256
- Samples Shape: (256, 144, 14)
- Targets Shape: (256,)

There are three built-in RNN layers in Keras:

- **keras.layers.SimpleRNN** : a fully-connected RNN where the output from previous time step is to be fed to next timestep.
- **keras.layers.GRU** : first proposed in Cho et al., 2014.
- **keras.layers.LSTMs** : first proposed in Hochreiter & Schmidhuber, 1997.

Our RNN (LSTM) Model



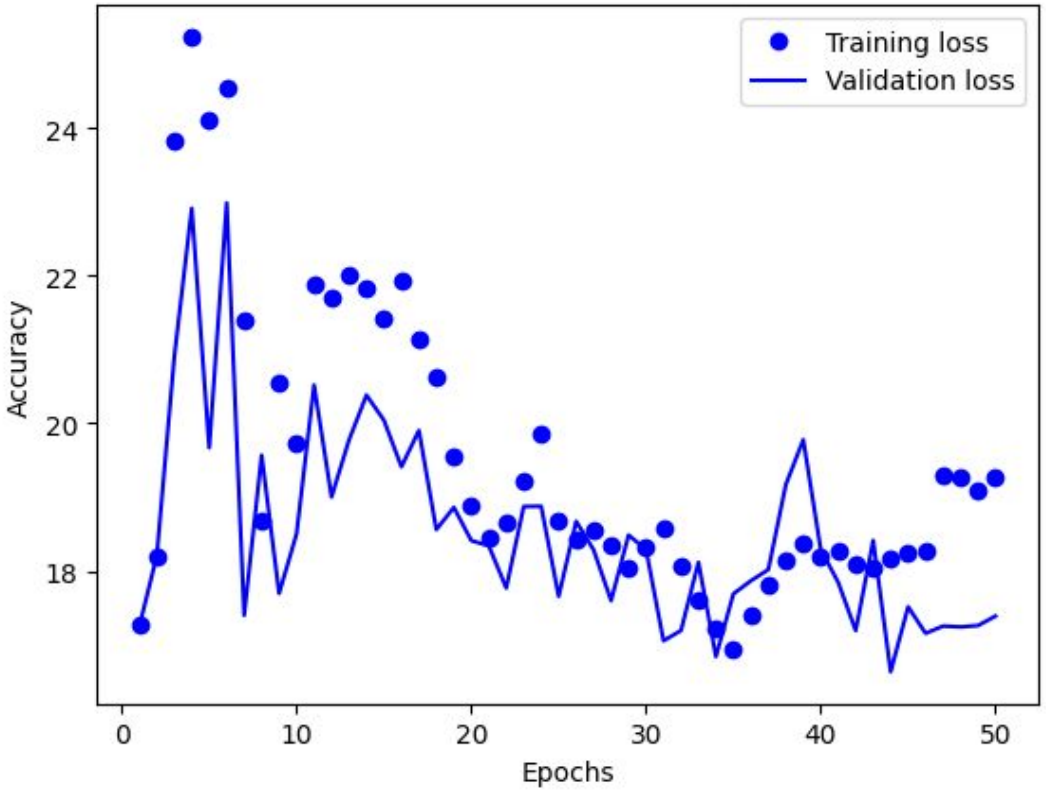
input_1	input:	[(None, 144, 14)]
InputLayer	output:	[(None, 144, 14)]

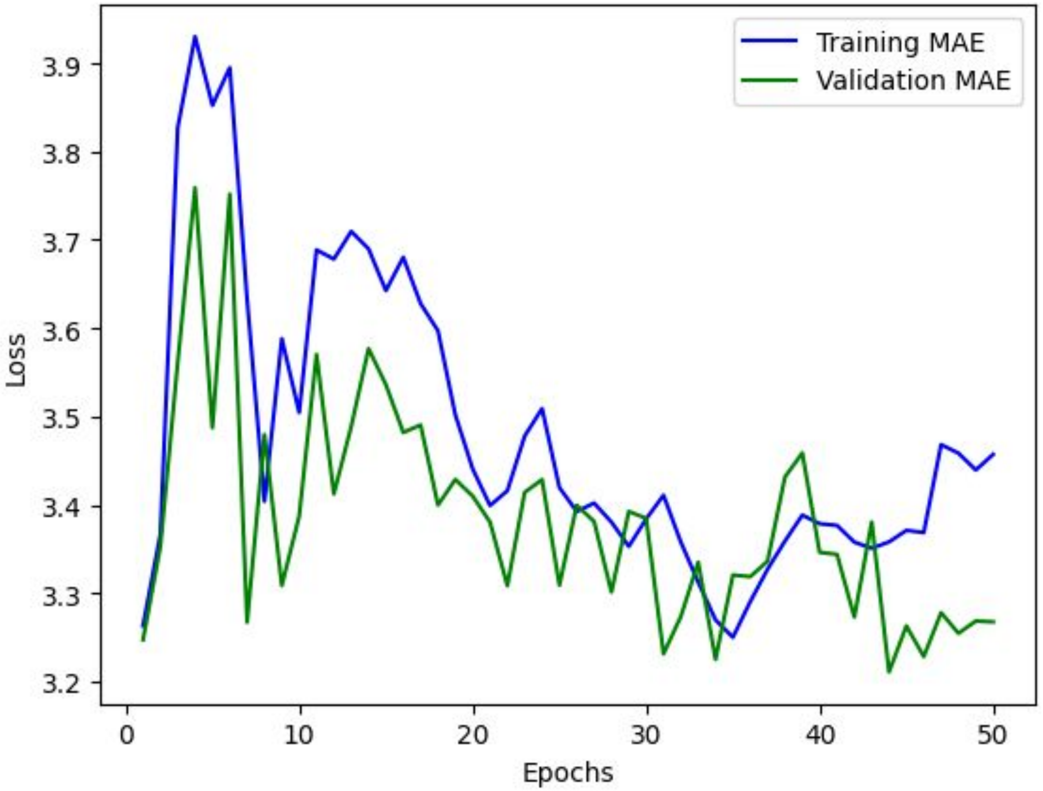
lstm	input:	(None, 144, 14)
LSTM	output:	(None, 16)

dense	input:	(None, 16)
Dense	output:	(None, 1)

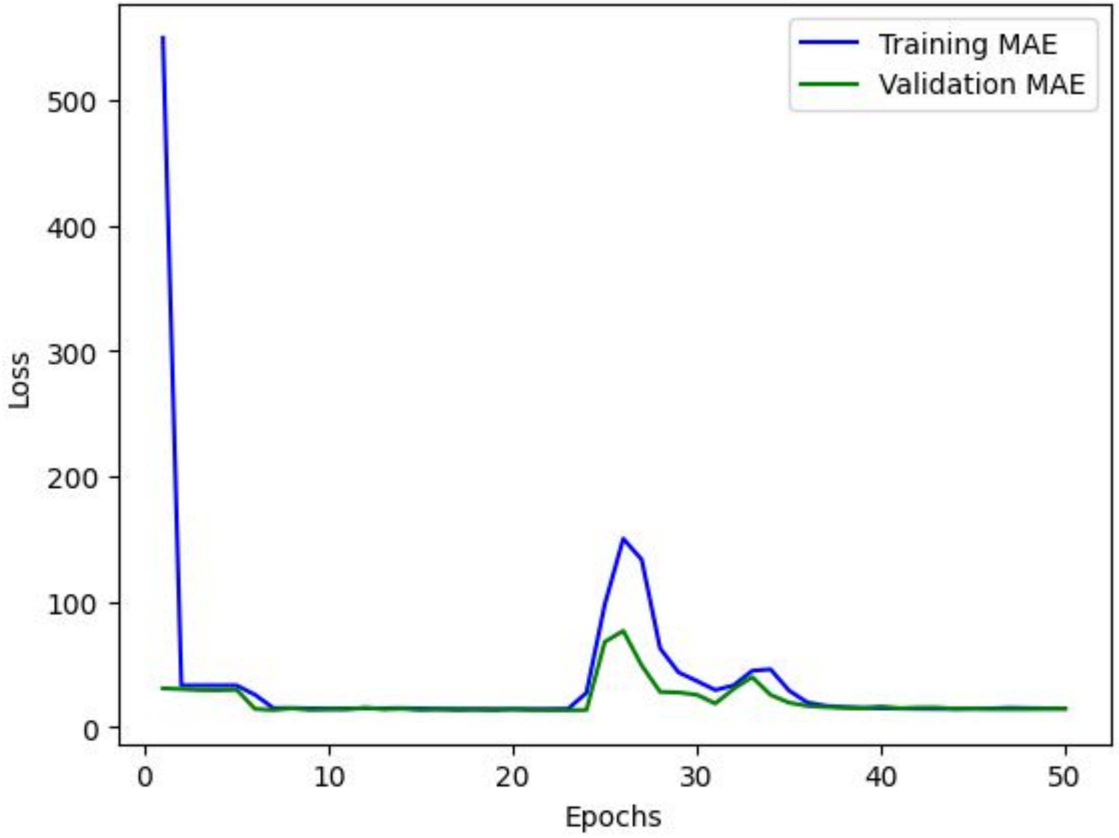
Optimizer	=	RMSprop
Loss	=	Mean Squared Error
Metrics	=	Mean Absolute Error
Epochs	=	50

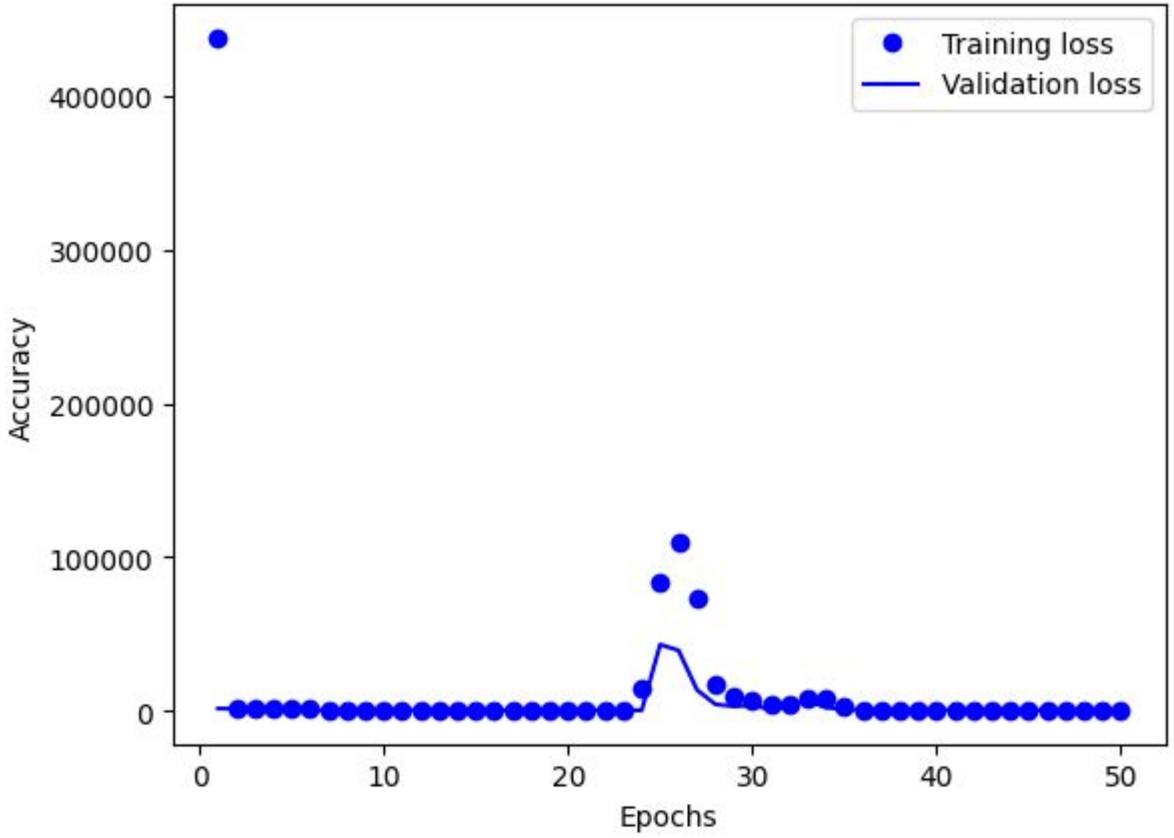
Evaluation : Temperature



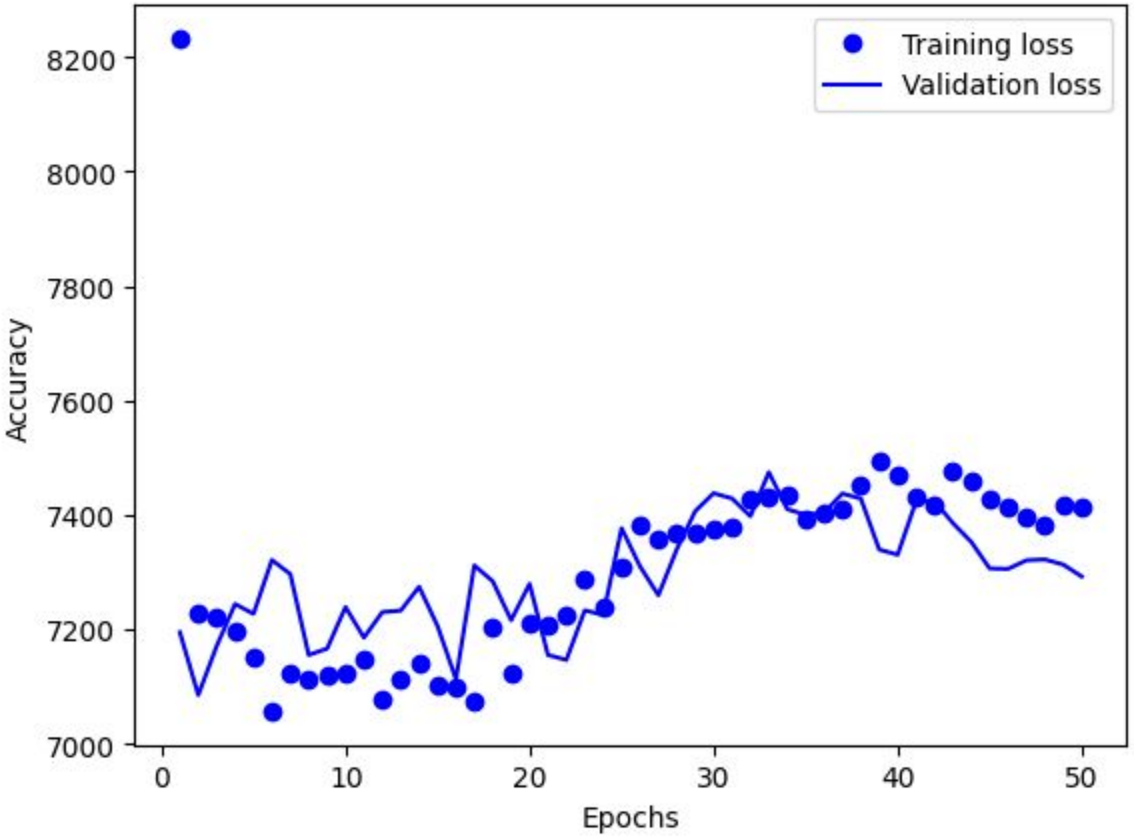


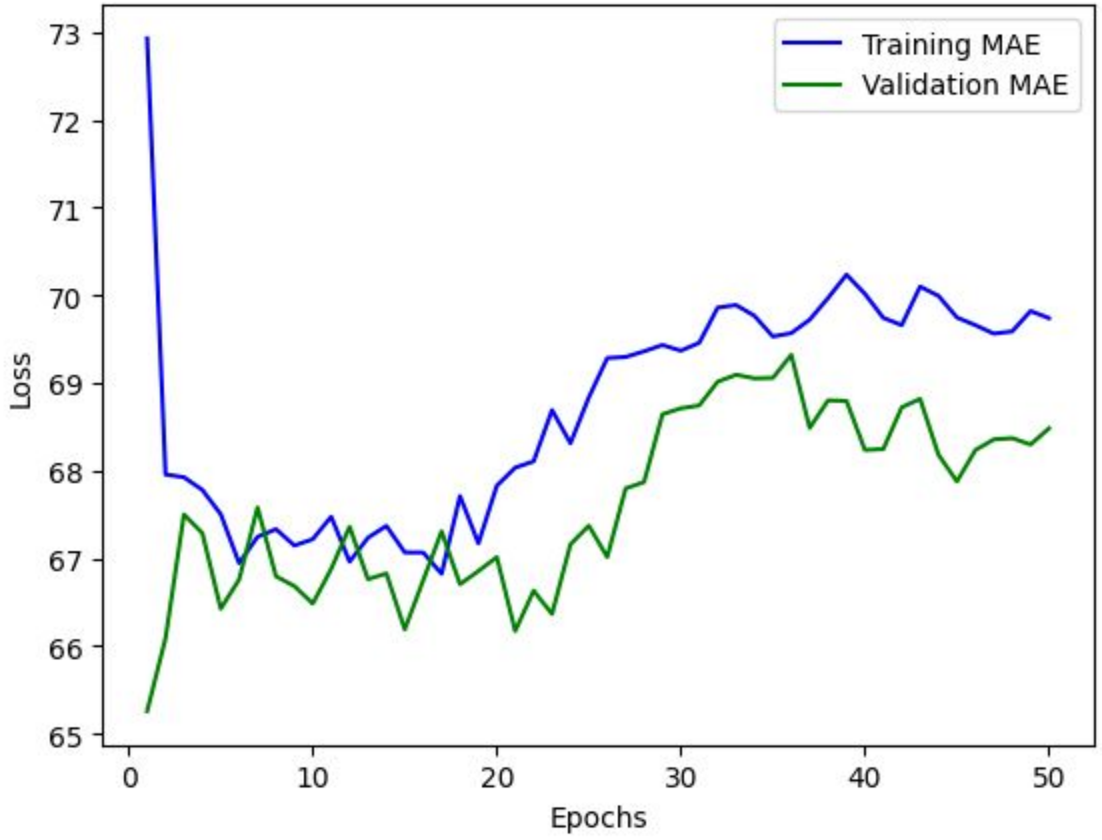
Evaluation: Air Density





Evaluation: Wind Direction





Evaluation Scores of LSTM and Common Baseline Model

Temperature Data	LSTM	Common Baseline
Validation Data MAE	2.68	2.53
Test Data MAE	2.62	2.59

Air Density Data	LSTM	Common Baseline
Validation Data MAE	14.30	18.29
Test Data MAE	14.26	19.31

Wind Direction Data	LSTM	Common Baseline
Validation Data MAE	66.17	80.81
Test Data MAE	66.72	81.87

Challenges

- Geoscience data availability
- Deep learning based time series (RNN)
- Adapting the baseline model for bench mark for final results.

Any Questions ?

Thank you!