Calibration and Downscaling of MERRA-2 Reanalysis Data for Improved Air Quality Assessments

1 Calibration Written

1.1 Data

As stated in the paper by Sayeed et al. (2022), the MERRA-2 aerosol variables are shown to be biased from the true measurements. Hence, in order to ensure the downscaling variables are accurate, we proposed to include a calibration phase where we used ground-level measurements of PM2.5 as a golden standard to calibrate the MERRA-2 variables for better downscaling quality.

Overall, the calibration phase of this project consists of mainly two datasets, as described below.

1.1.1 OpenAQ

OpenAQ is a non-profit organization providing universal access to air quality data, particularly PM2.5 data that we were interested in. In our study region, there are 11 measurement sites in total, all of which are U.S. embassies scattered in Southwest Asia. A list of the locations is included in the appendix. Overall, we gathered over 300,000 PM2.5 hourly measurements in the region from 2016 to 2024.

1.1.2 MERRA-2

The Modern-Era Retrospective Analysis for Research and Applications, Version 2, or MERRA-2 for short, is a worldwide reanalysis dataset provided by NASA for environmental studies. The MERRA-2 is a large, comprehensive dataset, and we only used a small subset of all the MERRA-2 variables. The full list of MERRA-2 variables used in this project along with their descriptions is included in the appendix. For this project, the data we selected comes in hourly format to match the OpenAQ PM2.5 measurements.

According to NASA, the mathematical relation between PM2.5 and the MERRA-2 aerosol variables can be summarized by the following equation:

$$\mathrm{PM}_{2.5} = \mathrm{DUST}_{2.5} + \mathrm{SS}_{2.5} + \mathrm{BC} + \mathrm{OC} + 1.375 \times \mathrm{SO}_4$$

As mentioned previously, this equation is empirically shown to be biased. Therefore, we aimed to find a better mathematical relation between PM2.5 and the MERRA-2 variables by fitting some machine learning models.

1.2 Methodology

Upon extracting the OpenAQ and MERRA-2 data, some preprocessing was completed. Specifically, since both the PM2.5 measurements and MERRA-2 come in hourly format, we joined the two datasets by matching the date and time at which the measurements were taken so that our models could use the MERRA-2 data to predict the PM2.5 value at the corresponding time step. Additionally, to exploit the temporal nature of environmental data, further data preprocessing included adding temporal lags of MERRA-2 variables in the previous 24 hours for the sake of building models that take advantages of this time series, such as LSTMs and transformers.

In the model fitting stage, we first referenced the methodology conducted by Sayeed et al. (2022) and attempted to replicate the results. We randomly selected 10% of all the PM2.5 measurements and fit various machine learning models to predict PM2.5 values based on the MERRA-2 data. The full list of machine learning models trained in this part is included in the appendix as well as in the next section. The XGBoost model was trained using the `XGBoost` library, the neural network using `PyTorch`, and other models using `scikit-learn` all with the default hyperparameters. Root mean squared error, or RMSE, was used for evaluation. This step aimed to ascertain the model that has the most potential in the calibration task, while keeping the validation process at a manageable level. The results will be discussed in the subsequent section.

After determining the best model, the entire dataset of all the PM2.5 measurements were split into a training set and a validation set with a 90:10 ratio. The best model type was then fit using the training set along with hyperparameter tuning based on the validation set. Root mean square error was again used as the evaluation metric. Feature engineering was applied, aiming to improve model performance.

In addition to hyperparameter tuning, different neural network architectures were considered during the model training process. Since the MERRA-2 data is essentially an hourly time series, LSTMs and transformers that are designed to handle sequential data were included as building blocks in the final neural network architecture. Furthermore, variational autoencoders or VAEs that aim to model the underlying distribution of the PM2.5 measurements were also part of the experiment during the calibration phase along with transfer learning, a powerful technique for training neural networks. Lastly, multiple loss functions such as mean squared error (MSE), mean absolute error (MAE), huber loss, and quantile loss for quantile regression were compared.

Upon finishing the training phase, the final model was validated on a dataset of PM2.5 measured daily in Kuwait in the 2004-2005 period. This validation step serves to evaluate the generalizability of the model outside of the available time range of the training data and is crucial since the overall objective of the project is to downscale the MERRA-2 components in the 2002-2016 period.

1.3 Results & Discussion

1.3.1 Subset Validation

As mentioned in the previous section, various machine learning models have been fit and validated on a 10% subset of all the available PM2.5 measurements. RMSE was used as the evaluation metric for this part. The results of the subset validation are shown in *Table 1* below.

Model	**Validation RMSE**
Baseline	51.166
OLS	47.841
Ridge	49.483
Lasso	50.444
SGD	1.680×10^{18}
KNN	53.526
SVM	52.697
AdaBoost	128.360
GB	45.063
HistGB	43.048
DT	62.082
RF	**42.246**
XGB	**43.971**
NN	**44.113**

Table 1: Subset Evaluation

From *Table 1* above, we see that the models having the lowest validation RMSE on the 10% subset are random forest, XGBoost, and neural network. These results mostly agreed with the results obtained by Sayeed et al. (2022). Note that HistGB has an even lower validation RMSE than XGBoost and neural network. However, considering that

XGBoost is a more powerful model and a superset of HistGB, we decided to move forward with XGBoost instead of HistGB.

1.3.2 One-Step Prediction

As stated in the methodology section, two types of data preprocessing were done. One of them was matching the date and time of the PM2.5 measurements and MERRA-2 and using MERRA-2 to predict the PM2.5 value at the corresponding time step. For this part, we trained various fully-connected neural networks by experimenting different sets of hyperparameters. We also applied feature engineering by adding cyclical encodings for date and time, incorporating embedding layers for sites and seasons. Since our study region contains some countries in the Middle East which belong to a region that is more arid than other countries, embeddings for aridity were as well considered. As we added more features, the validation RMSE of the neural networks gradually decreased and is shown in *Table 2* below.

Additionally, as mentioned, random forest and XGBoost were the other two types of models that have shown potentials in this calibration task. Therefore, we made some efforts in fitting these models and tuning the hyperparameters for them. The results with comparison to the final neural network are shown in *Table 3* below.

Data	**Validation RMSE**
Only MERRA-2	38.184
Cyclical Encoding	35.665
Time & Site	36.810
Site Embeddings	29.343
All Embeddings	26.210
Aridity	25.652
Fine-tuned	**24.393**

Table 2: RMSE at Different Stages

Model	**Validation RMSE**
Neural Network	**24.393**
XGBoost	25.907
Random Forest	40.674

Table 3: Model Type Comparison

From the tables above, we observed that the fine-tuned neural network obtained the lowest validation RMSE of 24.393, outperforming the XGBoost model and the random forest. This is expected as neural networks are a comprehensive set of models that can learn arbitrary non-linear functions well and thus have great potentials. As shown in *Table 2*, we also see that more added features also led to lower validation RMSE. The final model thus had embeddings for sites, seasons, and aridity.

It is worth to note that in the paper by Sayeed et al. (2022), the former researchers successfully achieved a validation RMSE of 7.15 on their random forest. This shows that there is still a large gap in the performance between our random forest and the previous successful model. An observation, though, on our current study is that the region of interest is Southwest Asia instead of the U.S. in the literature. This factor poses a significant challenge because Southwest Asia, especially the Middle East region, is an area that is much more arid than the U.S, and thus originally has PM2.5 measurements that are highly variable. It is obvious that the prediction task should be more difficult in the current study.

1.3.3 Evaluation & Limitations

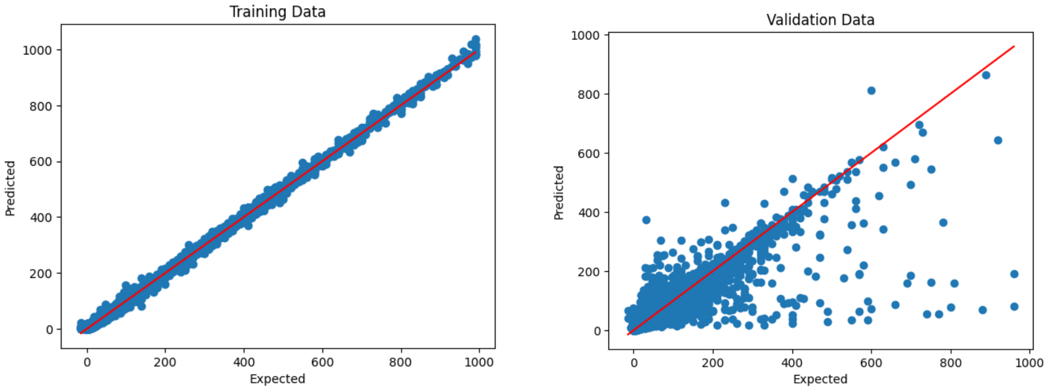
To better understand the unsatisfactory performance of the models, we made some efforts to evaluate and diagnose our fine-tuned neural network above. Since we had 11 sites in total, we first tried leave-one-out cross validation – by fitting the model on 10 of the 11 sites and validating it on the last one. We iteratively applied this technique on all 11 sites and obtained the following results as shown in *Table 4*.

Site	**Training RMSE**	**Validation RMSE**
Baghdad Airnow	8.8	45.0
Baghdad Embassy	**8.7**	**55.4**
Bahrain	8.2	44.2
Doha	8.5	33.0
Dubai	8.1	36.5
Kuwait	**7.5**	**51.3**
Kyrgyzstan	8.1	33.7
Tajikistan Dushanbe	8.4	32.9
Tajikistan Embassy	7.9	43.8
Turkmenistan	8.5	44.2
Uzbekistan	8.5	39.2

Table 4: Leave-One-Out Cross Validation

From *Table 4*, notice that the neural network is capable of fitting the PM2.5 values well, as shown by the low training RMSE. However, the model failed to generalize to unseen data, and is particularly seen when Baghdad Embassy or Kuwait was left out. This supported our earlier claim that arid areas pose are more problematic when it comes to predicting PM2.5 – these two sites are located in the Middle East, an area much more arid than others.

Besides leave-one-out cross validation, we also tried plotting the expected PM2.5 measurements vs. the predicted PM2.5 values and the plots are shown in *Figure 1* below.



(a) Training, $R^2 = 0.997$

(b) Validation, $R^2 = 0.769$

Figure 1: Expected vs. Predicted PM2.5 Values

From the figure, we see that the neural network was fitting the training data very well, proven by the training R^2 value and the straight line in the plot. However, the validation plot looks much more scattered. Particularly, the model seemed to constantly underestimate PM2.5 values that were expected to be greater than 500 $\mu g/m^3$, leading to the high validation RMSE. The following section thus describes how we attempted to alleviate this bottleneck by introducing more neural network architectures.

Since the model was constantly underestimating high PM2.5 values, we were curious of whether high PM2.5 values were common in the dataset and so we conducted a diagnostic for summary statistics. *Figure 2* below shows the boxplots of the PM2.5 ground measurements classified by measurement sites.

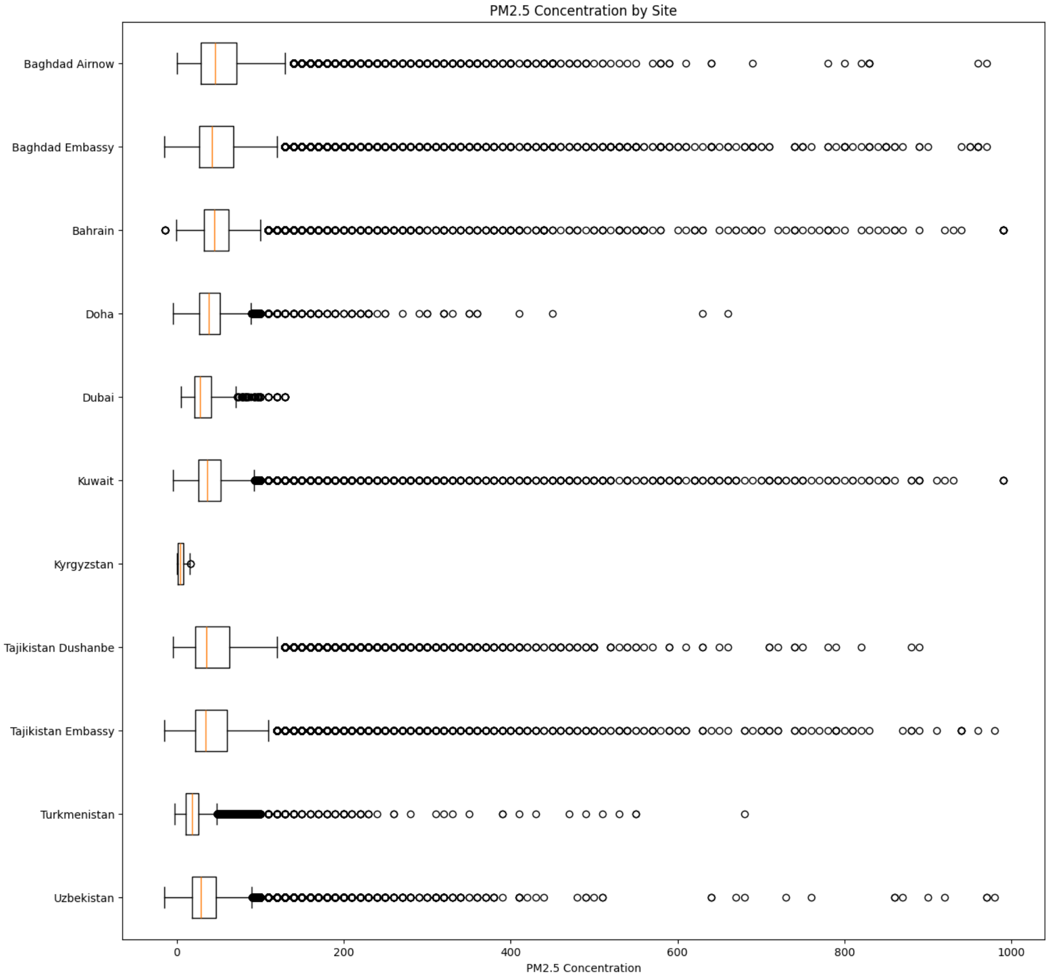


Figure 2: PM2.5 Concentration by Sites

Note that other than Kyrgyzstan, all other measurement sites have a large number of outliers that have PM2.5 values greater than 100 $\mu g/m^3$, while the majority of values fall below the 100 $\mu g/m^3$ mark. This supports that the model had difficulties making high PM2.5 predictions since there was not much data to train on. On the other hand, it was not surprising that Kyrgyzstan had the lowest validation RMSE in the leave-one-out practice because there weren't too many outliers associated with this measurement site, thus easier to generalize.

Besides the above observations, we also conducted some additional data diagnostics for the MERRA-2 data. Looking at the full list of the chosen MERRA-2 variables, note that there are some pairs of variables that look similar, such as SSSMASS and SSSMASS25. Therefore, we plotted the correlation plot of the MERRA-2 variables in *Figure 3* below.

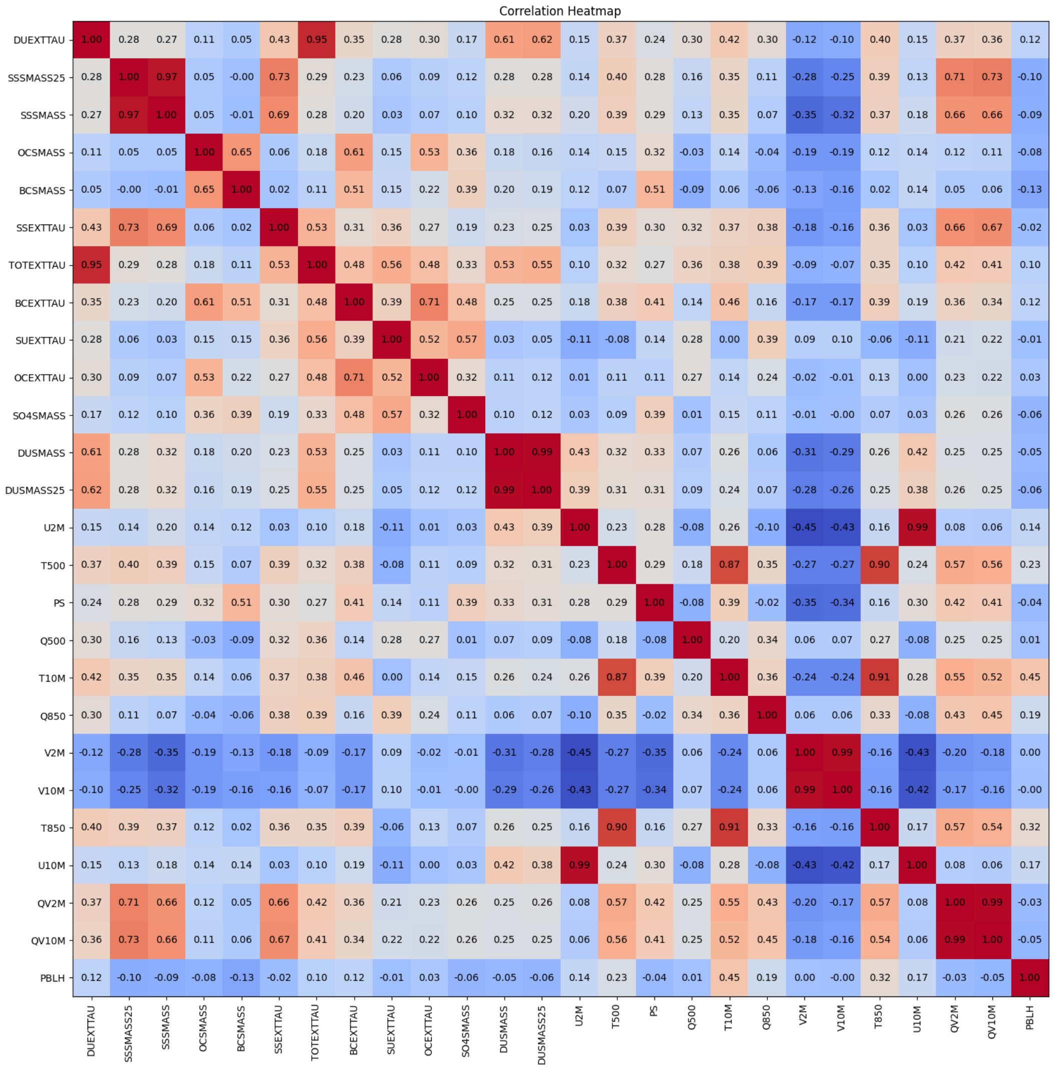


Figure 3: MERRA-2 Variable Correlations

Figure 3 shows that there are indeed pairs of variables that are highly correlated. In linear regression, these high correlations would pose severe problems of high variance inflation factors and render incorrect models, and might be a problem here as well. Further studies should omit the highly correlated variables and check whether the results might be better.

1.3.4 Time Series Data

For the sake of improving model performance of the predictions, we preprocessed the data to incorporate a temporal lag of 24 for the MERRA-2 variables for every time step so that our models not only have the current time but also past information to predict PM2.5. As such, we introduced LSTMs into our models, a recurrent neural network architecture that reads sequential data to make predictions. As the trending architecture, transformers were also built and compared with LSTMs on the PM2.5 prediction task.

During the training process, we discovered that the loss function for the models largely determined model performance. Throughout the study, we compared multiple loss functions and logged the validation RMSE of the LSTM models in *Table 5* below.

Loss Function	**Validation RMSE**
Baseline (One-Step Model)	*24.393*
MSE Loss	22.91
Weighted MSE	28.62
Huber Loss	22.72
Pinball Loss	**21.40**

Table 5: Validation RMSE on Different Loss Functions

Note that pinball loss has the lowest validation among the four loss functions used. To use the pinball loss function, the LSTM models have been modified to include four heads in the output layer, each predicting the 0.5 quantile, the 0.9 quantile, the 0.95 quantile, and the 0.99 quantile, respectively. Upon finishing training the quantile LSTM models, the final predictions were then taken to be the 0.5 quantile. As mentioned previously, a major problem of the one-step prediction models was that the models underestimated many of the values that were expected to be greater than 500 $\mu g/m^3$. Adding higher quantiles during training thus helped the models to capture the greater PM2.5 values so that performance improved.

The mathematical equations for the loss functions are shown below:

$$\begin{aligned}
\text{MSE}(y,p) &= \frac{1}{n}\sum_{i=1}^{n}(y_i - p_i)^2 \\
\text{WeightedMSE}(y,p) &= \frac{1}{n}\sum_{i=1}^{n}\left(1+\left(\frac{y_i}{500}\right)^2\right)\times(y_i - p_i)^2 \\
\text{HuberLoss}(y,p,\delta) &= \begin{cases} 0.5(y_i-p_i)^2 & \text{if } |y_i-p_i| < \delta \\ \delta\cdot|y_i-p_i| - 0.5\delta^2 & \text{otherwise} \end{cases} \\
\text{PinballLoss}_\tau(y,p) &= \begin{cases} \tau(y_i-p_i) & \text{if } y_i \ge p_i \\ (1-\tau)(p_i-y_i) & \text{otherwise} \end{cases}
\end{aligned}$$

In addition to LSTM models, we also built transformers and variational autoencoders that were based on transformers and LSTMs. However, these models did not work as good as expected (worse than the LSTM model with pinball loss) and unfortunately the training logs were lost. Hence the results were not present here.

Transfer learning is a powerful technique in training neural networks. In the current study, we referenced the project conducted by Wang et al. (2022), in which the researchers utilized transfer learning for the MERRA-2 downscaling task. In the paper, the researchers developed an LSTM neural network architecture that had two process blocks consisted of solely fully-connected layers, two temporal blocks consisted of LSTM layers, and a transfer block where there was a transfer model trained on predicting MERRA-2 variables. The rationale was that by pre-training a model to understand the MERRA-2 data, the final downscaling model would have better information in the distribution of the MERRA-2 data in order to make better predictions. In the current study, we replicated the same model architecture and pre-trained a transfer model on the MERRA-2 data to predict itself. Nevertheless, as we applied the final model with the transfer block on the PM2.5 prediction task, we did not obtain great results and so the idea was later abandoned.

In summary, in this section of the project, we developed an LSTM model using the pinball loss and achieved a final validation RMSE of 21.40, which was the lowest among all the models. The R^2 value was also around 0.80, which was slightly higher than the one-step prediction model.

1.3.5 Out-of-Sample Validation

Upon obtaining the final LSTM model, we deployed the model on a daily PM2.5 dataset in Kuwait in the period of 2004-2005 to assess the generalizability of the model. Since the overall objective of the current study is to downscale MERRA-2 from 2002 to 2016, the accuracy of the calibration model in this period is necessary for the accuracy of the downscaling.

The bad news is that the model is not very generalizable. On the 2004-2005 Kuwait PM2.5 dataset, the model achieved an RMSE of only blank and an R^2 of blank, which show the poor performance.

1.4 Conclusion

In conclusion, on the OpenAQ PM2.5 dataset, our developed model was capable of predicting PM2.5 quite accurately using the MERRA-2 data. The final RMSE of 21.40 and R^2 of around 0.8 proved the capability of the LSTM model. However, the model is not ready to use for the downscaling task that will be carried out later, as the model was not generalizable to the 2004-2005 Kuwait dataset by only having an RMSE of blank and an R^2 of blank.

1.4.1 Limitations

There were several limitations associated with the study. The most important one must be the problem with the Kuwait dataset that we used to validate our LSTM model. The data that we trained our models on was hourly in both the PM2.5 and MERRA-2, while the Kuwait validation data was daily. To make up for this incompatibility, we had to average the MERRA-2 data used in the validation to a daily level so that we could hopefully predict the daily PM2.5 values accurately. However, this method is obviously not justified, but we unfortunately did not have applicable hourly PM2.5 data to validate our model on. Future researchers preferably should have more PM2.5 data in the period of 2002 to 2016 to validate the calibration model so that the downscaling portion can be more accurate.

2 Appendices

2.1 Data

2.1.1 List of MERRA-2 Aerosols Variables

- **BCEXTTAU** – black carbon
- **DUEXTTAU** – dust
- **SUEXTTAU** – SO_4
- **OCEXTTAU** – organic carbon
- **SSEXTTAU** – sea salt
- **TOTEXTTAU** – total aerosol
- **DUSMASS25** – dust PM2.5
- **SSSMASS25** – sea salt PM2.5
- **DUSMASS** – dust
- **BCSMASS** – black carbon
- **SO4SMASS** – SO_4
- **SSSMASS** – sea salt
- **OCSMASS** – organic carbon

2.1.2 List of MERRA-2 Meteorology Variables

- **PS** – surface pressure
- **Q500** – specific humidity at 500 hPa
- **Q850** – specific humidity at 850 hPa
- **T850** – air temperature at 850 hPa
- **T500** – air temperature at 500 hPa
- **T10M** – 10-meter air temperature
- **QV10M** – 10-meter specific humidity
- **QV2M** – 2-meter specific humidity
- **U10M** – 10-meter eastward wind
- **U2M** – 2-meter eastward wind
- **V10M** – 10-meter northward wind
- **V2M** – 2-meter northward wind
- **PBLH** – planetary boundary layer height

2.1.3 Location

- Baghdad, Iran (2 sites)
- Manama, Bahrain
- Doha, Qatar
- Dubai, UAE
- Kuwait City, Kuwait
- Osh, Kyrgyzstan
- Dushanbe, Tajikistan (2 sites)
- Ashgabat, Turkmenistan
- Tashkent, Uzbekistan

2.2 Methodology

1. Pulled ground-level measurements from OpenAQ
2. Pulled MERRA-2 data
3. Matched OpenAQ and MERRA-2 data by location and time

4. Replicated Sayeed et al. (2022) paper by evaluating OLS, Ridge, Lasso, Stochastic Gradient Descent, KNN, SVM, Ada Boost, Gradient Boosting, Decision Tree, Random Forest, NN, and XGBoost on a 10% subset of data
5. Tried different neural networks, training hyperparameters, and feature engineering
6. **Feature Engineering Steps:**

   (a) Only MERRA-2 variables for prediction

   (b) Added cyclical encoding for month, day, and hour

   $$\text{value}_{\text{sin}}(x) = \sin\left(\frac{2\pi x}{\text{max value}}\right)$$

   $$\text{value}_{\text{cos}}(x) = \cos\left(\frac{2\pi x}{\text{max value}}\right)$$

   For example, the max value for month is 12.

   (c) Added one-hot vector for sites

   (d) Changed site information into embeddings using an embedding layer in PyTorch, dim = 256

   (e) Added season information, also as embeddings, dim = 32

   (f) Used embeddings for all of site, season, hour, day, month, removed cyclical embeddings

   (g) Added aridity information, also as embeddings dim = 200