APPLICATIONS OF MACHINE LEARNING WHEN DEVELOPING TEMPORALLY VARYING BACKGROUND AIR CONCENTRATION DATASETS USED IN AIR DISPERSION MODELS

Sophie Materia[1], Amy Schmidt[1] and Kirsty Robinson[2]

[1] Mott MacDonald, L17, Tower One, Collins Square, 727 Collins St, Docklands VIC 3008

[2] Mott MacDonald, Level 17, One Festival Tower Station Road Adelaide SA 5000

Abstract

This study investigates possible applications of machine learning when developing temporally varying background air concentration data for use in air dispersion models. Applications considered in this study include (1) using machine learning algorithms to replace high background concentrations which have occurred due to exceptional events (such as bushfires) and (2) to predict background concentrations accounting for climate change for use in future scenario air dispersion models. The results of this study indicate that machine learning models may be used to predict concentrations that are representative of typical conditions during exceptional events. Further, when a comparison against typical approaches was undertaken it was found that machine learning had the lowest error margin. The study found that climate change adjustments made to a 2023 dataset had little impact on predicted PM_{10} values compared to the unadjusted dataset. However, it highlighted that changes in temperature and wind speed influenced the predicted concentrations, and emphasised the potential of machine learning models to predict future background concentrations affected by climate change. It is the authors' opinion that further work is required to determine the implications of these findings with regards to the application of atmospheric dispersion modelling.

Keywords: machine learning, background air quality data, exceptional events, climate change projection.

1. Introduction

In recent years the exploration and adoption of machine learning algorithms has increased within the air quality research community (Mendez et al. 2023, p. 8). Machine learning algorithms are used in a variety of ways, including predicting air quality levels, using Geographic Information System (GIS) data to classify land use patterns and optimising environmental monitoring systems. This paper covers two potential applications of machine learning which are applied to temporally varying background air concentration data input into air dispersion models. It is noted that this study adopts an exploratory approach, intentionally avoiding an in-depth analysis of model parameters, with the primary aim of promoting the application of machine learning within the field of air quality consulting.

1.1. Machine learning algorithms

Although numerous machine learning approaches exist to solve different problems, it is common to use regression-based algorithms to predict air pollutant concentrations (including decision trees, Random Forest and K-nearest neighbours algorithms) (Mendez et al. 2023, p. 5). Regression-based algorithms work by estimating a function that maps input data to predicted output values. This allows for predictions to be made and an understanding of the underlying patterns in the data.

1.2. Background air concentration data used in air dispersion models

When undertaking air dispersion modelling, there are a number of methods available that can be used to form assumptions about background air concentrations. In particular, temporally varying data from ambient air monitoring stations are commonly input into dispersion models or added to incremental model predictions.

Often modifications are required to be made to the background data prior to use. Example modifications include removing high concentrations recorded during exceptional events (such as bushfires) and modifying background data to account for climate change for use in future scenario air dispersion models. These two modifications are the basis of this study and are herein after referred to as test cases.

2. Objectives

For the two identified test cases, this study investigates the following:

- **Test case 1: Exceptional events:** during exceptional events, can machine learning algorithms be used to predict values that are representative of typical conditions?
- **Test case 2: Climate change projection:** can machine learning algorithms be used to predict future scenario background concentrations which have been influenced by climate change?

3. Current approaches

3.1. Test case 1: Exceptional events

Current guidance in Australia and New Zealand was reviewed with regards to substituting high background concentrations which have occurred as a result of exceptional events. The recommended approach detailed in Victorian guidance was that "gaps [due to exceptional events removal] should be backfilled using the most recent non exceptional event day (EPA Victoria 2022 p.35). The remaining guidance reviewed did not provide details on substituting exceptional events.

In application, the following approaches are common:

- **Linear interpolation:** high values are filled in by estimating them based on neighbouring data points or the most recent non-exceptional event day.
- **Annual average / 70th percentile:** the annual average or commonly the 70th percentile is calculated and input into the missing data periods.

3.2. Test case 2: Climate change projection

To understand the current approach for applying background concentration data to future scenario dispersion models, a review of industry best practice methods as documented in air quality assessments for major infrastructure projects across Australia was completed. The review found that the assessment of future conditions used historic background concentration data for the same historic period as the meteorological data used in the model (Melbourne Metro Rail Authority 2016; State Government of Victoria 2019). No adjustments were made to the data to account for climate change and background concentrations were assumed to remain unchanged for future years (WestConnex Delivery Authority 2015).

Whilst this test case explores the influence of climate change on background pollutant concentrations, it should be noted that there are numerous additional factors that could be reasonably expected to influence future background pollutant concentrations. These factors include future changes to industry, societal behaviour and vehicle number / fleet composition.

4. Methodology

4.1. Model selection

There are many machine learning models available, however this study has used the Random Forest machine learning model which is widely adopted for air quality prediction due to its ability to handle non-linear relationships between the input variables and the target variable (i.e., the variable being predicted) (Gaikar et al. 2023).

4.2. Data sources

Data measured between 2018 to 2023 from the South Australia Environment Protection Authority (EPA) monitoring station Le Fevre 1 was used in this study. The input variables were one hour averaged meteorology data including: temperature, wind speed, wind direction, relative humidity and temporal variables including the date and hour of day. The target variable was one hour averaged PM_{10}. In other words, the Random Forest algorithm used the relationship between the meteorological / temporal data and PM_{10} measured at the Le Fevre 1 air quality monitoring station to predict PM_{10} concentrations for the two test cases.

4.3. Train-test split

Train-test split is a technique used in machine learning to evaluate the performance of a model on unseen data. The process involves splitting a dataset into two parts: a training set and a testing set. The training set, which typically comprises of 80% of the data, is used to train the model so that it can generalise to new, unseen data. The testing set, comprising of the remaining 20% of the data, is used to evaluate the model's performance.

4.4. Model evaluation

The following metrics were used to evaluate the performance of the Random Forest models.

Root Mean Squared Error (RMSE): quantifies the difference between the predicted values and the actual values by calculating the square root of the average of the squared differences. Lower values of RMSE indicate better performance of the model.

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - x_i)^2} \quad (1)$$

R squared (R^2): measures how well a statistical model predicts an outcome. Specifically, it represents the proportion of variation in the dependent variable (the outcome / prediction) that is

explained by the model. R-squared values range from 0 to 1, with a value of 1 indicating a perfect fit.

$$R^2 = 1 - \frac{RSS}{TSS} = 1 - \frac{\sum_i (y_i - x_i)^2}{\sum_i (y_i - z_i)^2} \quad (2)$$

4.5. Test cases

Test case specific methodologies are outlined in the below sections.

4.5.1. Test case 1: Exceptional events

Hourly averaged meteorology (wind speed, wind direction, temperature and relative humidity), temporal data and PM_{10} concentrations measured at the Le Fevre 1 air quality monitoring station between 2018 and 2023 were randomly split into training and testing sets.

Three scenarios were undertaken in order to analyse this test case, namely:

- **Scenario 1:** Training set and testing set include exceptional events. Provides a baseline model with no data removed.
- **Scenario 2:** Training set and testing set exclude exceptional events. Analyses the difference in model predictions when exceptional events are excluded.
- **Scenario 3:** Training set excludes exceptional events and testing set includes exceptional events. Determines the model's effectiveness in predicting concentrations during exceptional events.

For the purpose of this study an exceptional event was considered to have occurred when the Le Fevre 1 24-hour average PM_{10} concentration exceeded the 50 µg/m³ criterion (Government of South Australia 2023 p.19) in addition to two other nearby stations having also exceeded the criterion for that day (i.e., a total of three or more stations having exceeded the 24-hour criterion for that day). The nearby monitoring stations included in the analysis were Christie Downs, Netley, Adelaide CBD, Le Fevre 2, Elizabeth and Kensington.

4.5.2. Test case 2: Climate change projection

A meteorology file adjusted to account for climate change was developed for 2090 which used the same input parameters as described for the exceptional events test case. This 2090 file was created by modifying 2023 meteorology data. The *Guide to Climate Projections for Risk Assessment and Planning in South Australia 2022* ('the Guide') was used to determine representative adjustments to temperature and wind speed, accounting for climate change projections (Department for Environment and Water 2022). All other parameters remained unchanged.

Adjusted temperature and wind speed values aligned with the medium Representative Concentration Pathway (RCP) 4.5, for the year 2090, representing the mean projected change. RCP4.5 was chosen to reflect an optimistic scenario where climate change mitigation measures have been implemented. Adjustment values were seasonally dependent and are shown in Table 1.

Table 1. Climate change projection adjustment values under RCP4.5 for 2090

Season	Mean projected temperature (°C) change	Mean projected wind speed (%) change
Summer	+1.85	-1.8
Autumn	+1.65	-1.8
Winter	+1.85	-1.8
Spring	+2.1	-1.8

The approach to adjusting temperature and wind speed, outlined above, was simplified for the purpose of determining if a difference in PM_{10} concentration would be predicted for the climate change adjusted meteorological file (2090) compared with the unadjusted meteorological file (2023).

The same model, as described in Test case 1 Scenario 1 was used (i.e, the training set remained unchanged), however concentrations were instead predicted on the climate change 2090 meteorological file (i.e. the testing set was updated to the climate change 2090 meteorological file).

5. Results

5.1. Test case 1: Exceptional events

5.1.1. Model performance

This section provides an overview of model performance and differences observed between the scenarios. A summary of the model error parameters is provided in Table 2 and it can be seen that:

- Scenario 1 vs Scenario 2: A slight improvement to the model performance occurred when exceptional events were excluded from both the training and testing sets (i.e., when exceptional events were removed from the baseline). This is observed in the RMSE value decreasing from 11.8 to 8.9 between Scenario 1 and 2.
- Scenario 2 vs Scenario 3: An expected reduction in model accuracy occurred when the exceptional events were added back into the testing set (RMSE increased from 8.9 to 16.8 between Scenario 2 and 3). This is due to the model having not been exposed to exceptional events during training and therefore not predicting elevated concentrations during these events.

Table 2. Error parameters - exceptional event scenarios

Scenario	R^2	RMSE
1	0.47	11.8
2	0.50	8.9
3	0.31	16.8

A further difference between the models was observed when reviewing the variable importance, which refers to the significance of each feature (or predictor variable) in making predictions. In Scenario 1 temperature had the highest variable importance whereas in Scenario 2 and 3 wind direction had the highest variable importance.

It was found that across the three scenarios, typically the algorithm overpredicted lower measured concentrations (< 25 µg/m³) and underpredicted higher measured concentrations (> 25 µg/m³). Figure 1 shows a scatterplot for Scenario 2 (excluding exceptional events) which shows observed concentrations plotted against predicted concentrations.

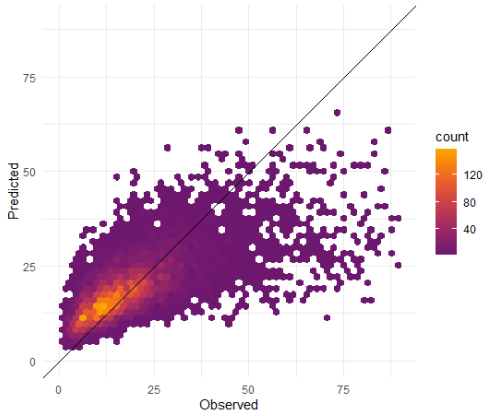


Figure 1. Scenario 2 – observed vs predicted one-hour average PM_{10} concentrations

5.1.2. Exceptional event analysis

This section provides an analysis of the model's effectiveness to predict conditions representative of typical conditions during exceptional events (Scenario 3). The exceptional events with the ten highest 24-hour average observed concentrations are analysed in this section.

Table 3 provides a comparison of the observed and predicted 24-hour average PM_{10} concentrations on these days. It can be seen that, with the exception of the two highest observed days, on average the predicted concentrations are ~36 µg/m³ lower than observed concentrations.

Table 3. Exceptional events – observed vs predicted 24-hour average PM_{10} concentrations

Date	24-hour average PM_{10} concentration (µg/m³)	
	Observed	Predicted
13/04/2021	124	31
20/12/2019	107	43
24/05/2021	75	39
23/12/2019	75	28
21/11/2019	72	24
19/09/2019	71	37
20/11/2019	70	41
22/03/2018	69	28
05/04/2019	66	36
01/03/2019	64	40

When analysing the hourly concentrations on these days (refer to Appendix A), it was found that the machine learning algorithm did not predict any extreme peak values - unlike those that were observed.

In some instances, similar predicted and observed concentrations were observed prior to or after the exceptional event occurring (i.e., in the morning hours if the exceptional event occurred in the afternoon). Figure 2 shows an example of this occurring on 21/11/2019 and it can also be seen on 19/09/2019 and 22/03/2018 in Appendix A.

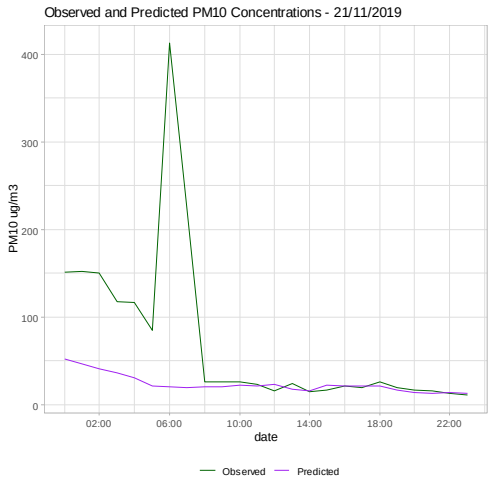


Figure 2. Observed vs predicted hourly PM_{10} concentrations – 21/11/2019

There were some instances where, prior to the exceptional event occurring, the predicted concentrations were incorrectly higher than the observed concentrations. An example of this occurred on 05/04/2019, as shown in Appendix A.

5.1.3. Comparison against current approaches

A comparison of the machine learning approach against two typical approaches: applying the 70th

percentile, and applying the average of the two valid 24-hour averages either side of the exceptional event was undertaken. This analysis was applied to the ten highest 24-hour average observed concentrations, resulting from exceptional events.

It was found that applying the 70th percentile typically resulted in the lowest background concentrations and machine learning typically resulted in the highest concentrations. The average of the valid day before and after typically fell between these two.

It is not possible to determine which of these approaches is most accurate as it is unknown what the concentrations would have been on these days if the exceptional events did not occur. Therefore, in order to identify the most accurate method, the analysis was extended to the entire dataset, excluding exceptional events.

A summary of the results of this analysis is provided in Table 4. It was found that when comparing against observed values, the 70th percentile approach resulted in the highest error (RMSE = 9.6) whilst use of machine learning resulted in the lowest error (RMSE = 5.1). The error associated with applying the average either side was between these two (RMSE = 7).

The corresponding scatterplots for the average either side and machine learning are provided in Appendix B.1.

Table 4. Error parameters – current approaches vs machine learning

Scenario	R^2	RMSE
70th percentile	0	9.6
Average either side	0.42	7.0
Machine learning	0.74	5.1

5.2. Test case 2: Climate change projection

This section provides an analysis of the ability of machine learning to predict appropriate future scenario background concentrations which have been influenced by climate change.

5.2.1. Comparison of observed (2023) and future (2090) predicted concentrations

Figure 3 shows the observed 24-hour PM_{10} concentrations for 2023 compared against the future (2090) predicted 24-hour PM_{10} concentrations for the climate change adjusted dataset. Future predicted values were generally higher than observed concentrations for values less than 18 µg/m³, and lower than observed concentrations for values greater than 18 µg/m³.

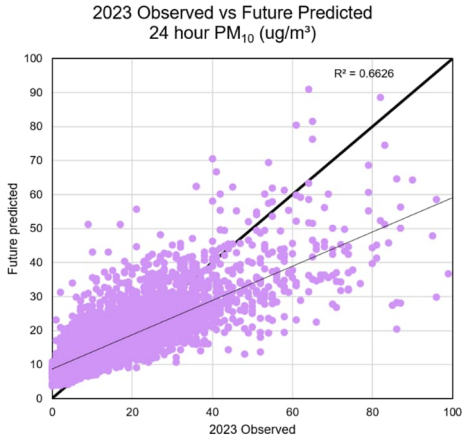


Figure 3: 24-hour average 2023 observed vs future (2090) predicted PM_{10} concentrations

The two datasets had an R^2 value of 0.7. Given the climate change adjustments, variability between the datasets was expected.

5.2.2. Comparison of predicted (2023) and future (2090) predicted concentrations

To gain further insight into the influence of the climate change adjustments on the predictions, a comparison of 2023 predicted PM_{10} concentrations (with no climate change adjustment) was undertaken against future (2090) predicted PM_{10} concentrations (with climate change adjustment), as shown in Figure 4.

It can be seen that there was less variability between these two data sets (R^2 = 0.9). Similarly, future 2090 predicted values were generally higher than 2023 predicted concentrations for values less than 20 µg/m³ and lower than predicted concentrations for values greater than 20 µg/m³, although this trend was less distinct than comparisons between future predicted and observed data.

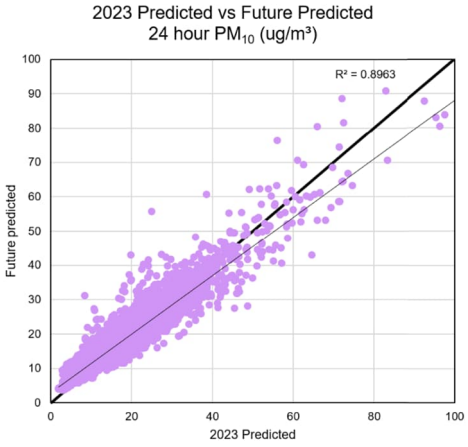


Figure 4: 24-hour 2023 predicted vs future (2090) predicted PM_{10} concentrations

5.2.3. Analysis of the influence of climate change adjustments on predictions

The influence of the adjusted temperature and wind speed values were analysed to further understand the cause of variation between future (2090) predicted and 2023 observed concentrations. Appendix B.2 shows how future (2090) predictions vary from 2023 predictions when temperature increases and wind speed decreases with values split into four quantiles. The observed trends match those between future (2090) and 2023 predicted concentrations, with R^2 values of approximately 0.9 for all quantiles. This supports that temperature and wind speed adjustments in the climate change dataset caused the differences in concentrations between future (2090) and 2023 predicted concentrations.

5.2.4. Summary

The analysis undertaken shows that climate change adjustments made to temperature and wind speed values in a 2023 dataset to reflect 2090 led to differences in predicted PM_{10} concentrations compared to an unadjusted 2023 dataset. However, the absolute change in concentration was not significant.

This study shows that machine learning models may help to predict future background concentrations affected by climate change. However, without knowing future PM_{10} values, the accuracy of these predictions cannot be determined. Further investigation is needed to understand when this approach is most effective for industry applications.

6. Conclusions

This study explored two specific applications of using machine learning to predict air concentrations input as background data into air dispersion models.

The results of this study indicate that machine learning models may be used to predict concentrations that are representative of typical conditions during exceptional events. Further, when a comparison against approaches typically used was undertaken it was found that using machine learning had the lowest error margin.

The study found that climate change adjustments made to a 2023 dataset had little impact on predicted PM_{10} values compared to the unadjusted dataset. However, it highlighted that changes in temperature and wind speed influenced the predicted concentrations, and emphasised the potential of machine learning models to predict future background concentrations affected by climate change.

It is the authors' opinion that further work is required to determine the implications of these findings with regards to the application of atmospheric dispersion modelling.

Acknowledgments

We extend our thanks to South Australia EPA for providing publicly available air quality and meteorology data, which formed the foundation of our research. We thank the Department for Environment and Water, South Australia for their development of 'the Guide' which was used to apply climate change projection adjustments to data in this study. Additionally, we thank David Carslaw, Stuart Grange and their teams, creators of the openair and rmweather R libraries. Their tools facilitated the analysis of air quality data, enabling us to derive meaningful insights.

References

Department for Environment and Water 2022, *Guide to climate projections for risk assessment and planning in South Australia*, Government of South Australia, through the Department for Environment and Water, Adelaide.

Environment Protection Authority Victoria 2022, *Guideline for assessing and minimising air pollution*, Publication 1961.

Gaikar D, Patel U, Vispute O, Singh S, Sanghvi T 2023, *'International Research Journal of Engineering and Technology (IRJET)'* **10**:04

Government of South Australia 2023, *'Environment Protection (Air Quality) Policy 2016'*, Version: 18.5.2023

Melbourne Metro Rail Authority 2016, *'Melbourne Metro Rail Project, Environment Effects Statement – Air Quality Impact Assessment'*, Melbourne Metro Rail Project.

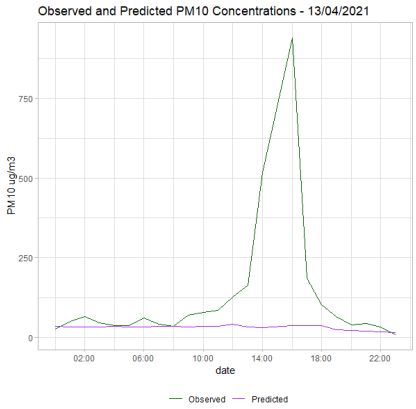
Mendez M, Merayo MG, Nunez M 2023, 'Machine learning algorithms to forecast air quality: a survey', *Artificial Intelligence Review (2023)* **56**:10031–10066.

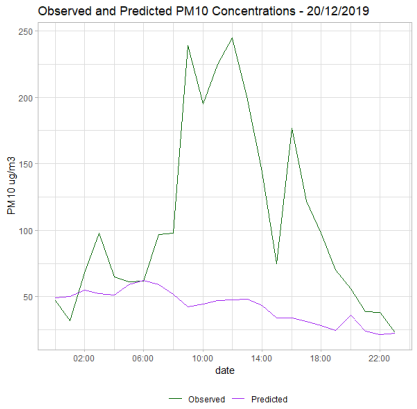
National Environment Protection Council, 2021, *National Environment Protection (Ambient Air Quality) Measure*, Office of Parliamentary Counsel, Canberra.

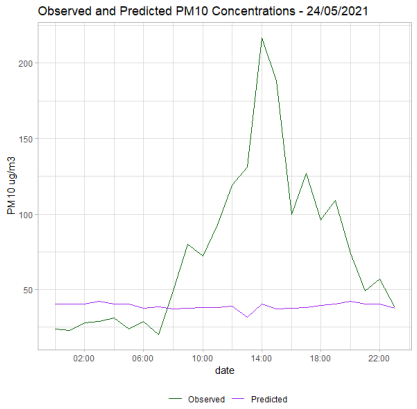
State Government of Victoria 2019, *'North East Link, Environment Effects Statement, Chapter 10 – Air Quality',* North East Link Project.

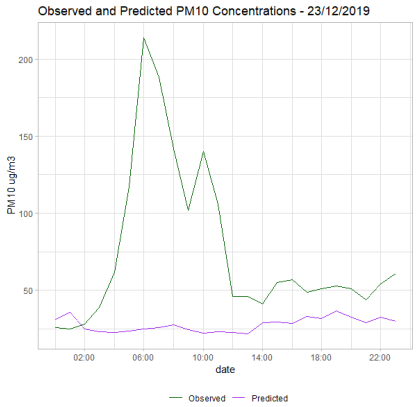
WestConnex Delivery Authority 2015, *'WestConnex M4 East Environmental Impact Assessment: Appendix H (Air Quality Impact Assessment)',* New South Wales Government.

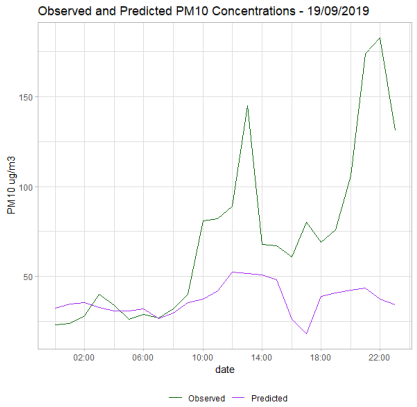
Appendix A – Exceptional Events Graphs

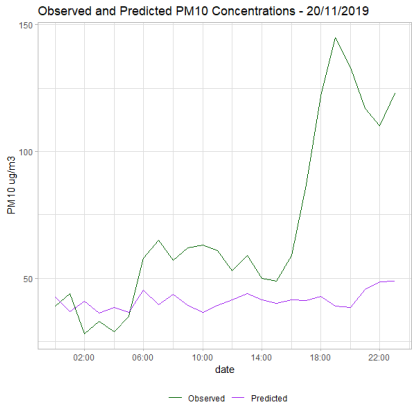


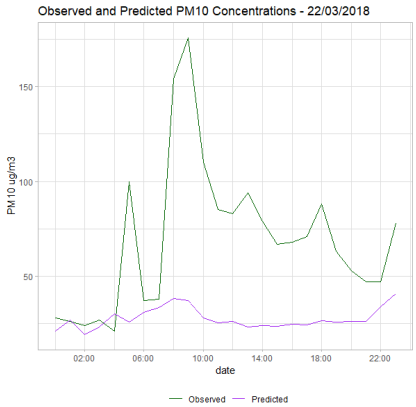


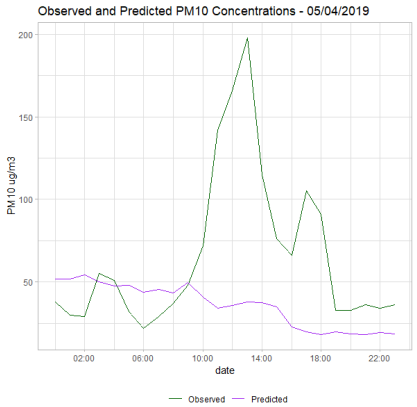


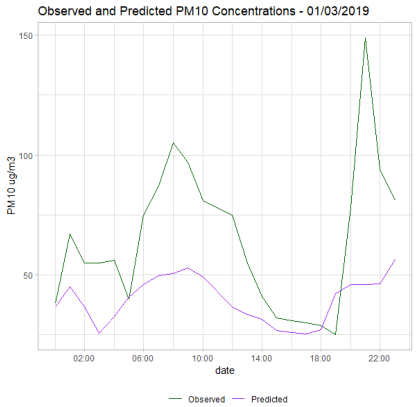






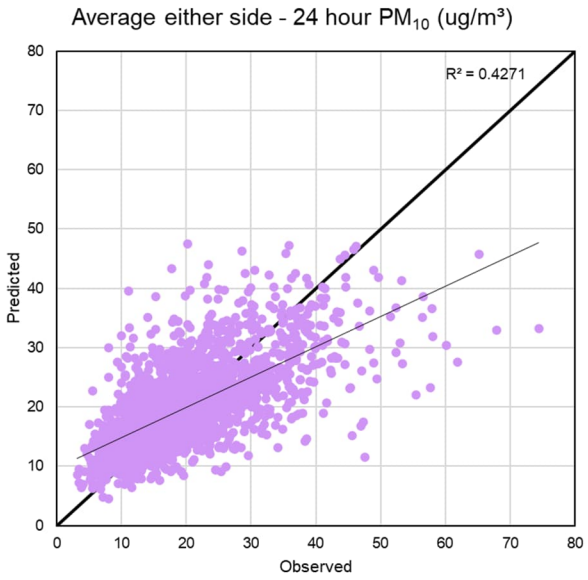




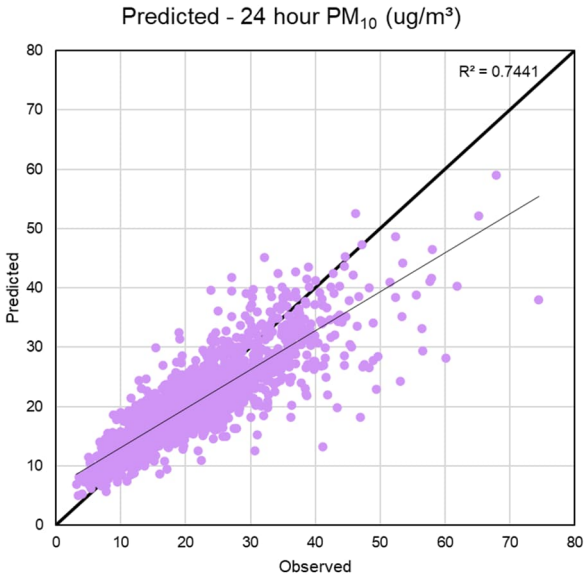


Appendix B – Supplementary Graphs

B.1 – Test case 1: Exceptional events – comparison against current approaches

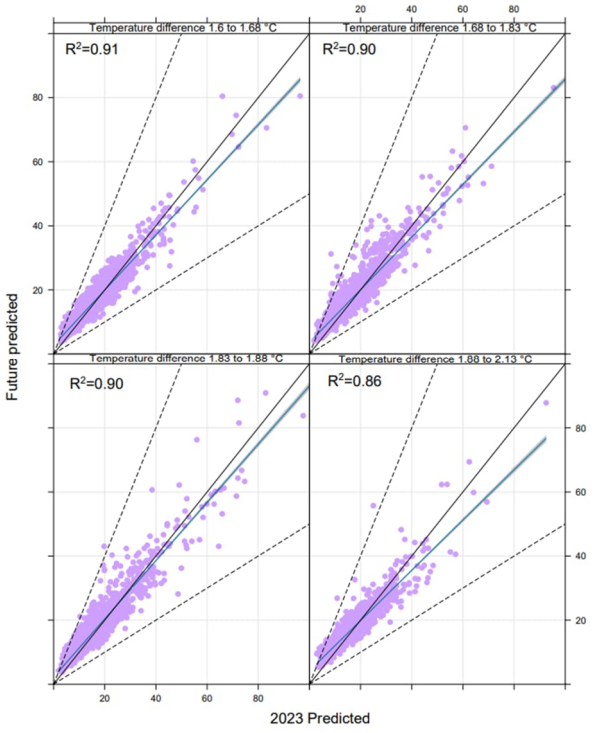


24-hour average observed vs average either side PM_{10} concentrations

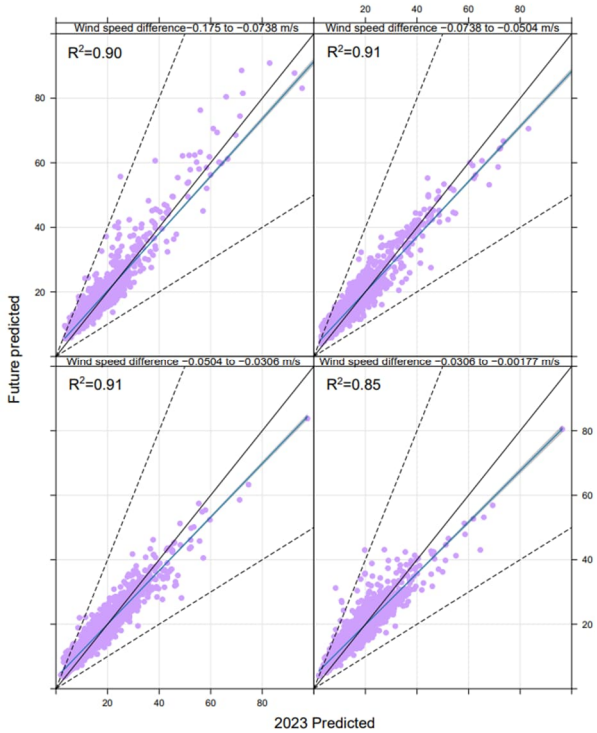


24-hour average observed vs machine learning PM_{10} concentrations

B.2 – Test case 2: Climate change projection – influence of climate change adjustments on predictions



Influence of temperature change on 2023 predicted vs. future predicted PM_{10} values ($\mu g/m^3$)



Influence of wind speed change on 2023 predicted vs. future predicted PM_{10} values (m/s)