Authors:

Group 18

Thomas Koh 329888

Vishall Krishnan XXX

Shaohua Liu XXX

Yujing Yang XXX

Abstract

The planetary boundary layer (PBL) height is an important factor for measuring the pollution level of a given area. The DPIE have asked us to conduct an analysis comparing the PBL height values based on two different sources, the CL51 ceilometer and the CT model output. This report shows our preliminary exploratory data analysis conducted on the data and our proposal going forward

Comparison of pbl height based on cl51 ceilometer and ct model outputs in merriwa and lidcombe

An EDA analysis and scoped proposal

Contents

[A. Introduction 2](#_Toc71304833)

[Background: 2](#_Toc71304834)

[Problem statement: 2](#_Toc71304835)

[Intended outcomes of analysis: 2](#_Toc71304836)

[B. Related work (Yujing) 3](#_Toc71304837)

[C. Data analysis and preliminary modelling 4](#_Toc71304838)

[Data provided: 4](#_Toc71304839)

[Data pre-processing steps. 4](#_Toc71304840)

[D. Analysis 1 - Comparing the PBL height of both datasets. 8](#_Toc71304841)

[Visualisation of the CL51 and CTM PBL height data 8](#_Toc71304842)

[Statistical metrics 11](#_Toc71304843)

[Discussion 12](#_Toc71304844)

[E. Analysis 2 - Analysing factors that impact the PBLH 14](#_Toc71304845)

[F. Proposal for Semester 2 15](#_Toc71304846)

[G. Proposed timeline for Semester 2 15](#_Toc71304847)

## Introduction

### Background:

There has traditionally been a lot of pollution in New South Wales (NSW) due to mining in particular from mining dust. As such, there has been concerted effort to better monitor these pollution sites. As part of this effort The Department of Planning, Industry and Environment of the NSW Government (DPIE) have obtained new instruments as part of an industry programme to monitor pollution levels.

One such instrument is the CL51 ceilometer by Vaisala which is an air quality measurement tool. This is rolled out in two locations: (1) Merriwa and (2) Lidcombe. We have therefore been asked to investigate the efficacy of a new air quality measuring tool in measuring pollution levels in these two areas.

### Problem statement:

We have been notified by DPIE that the Planetary Boundary Layer Height (PBL height) is an important factor for measuring the pollution level of a given area.

We have subsequently been tasked to compare the quality of the PBL height as measured by the new CL51 ceilometer instrument. In doing so we aim to answer the following questions:

* How accurate and reliable is the PBL height data as obtained from the CL51 ceilometer;
* How do the readings from the CL51 ceilometer compare to the predicted PBL height as obtained from the CT model; and
* What are the potential improvements that can be made to the prediction of PBL height based on information obtained from the CL51 ceilometer.

For each of the two locations, we are presented with the following information:

* CL51 data for the period from 12/2/2021 to 19/2/2021; and
* Predicted PBL height values;

### Intended outcomes of analysis:

In Semester 1, our focus was on performing detailed exploratory data analysis on the PBL height data provided. The intended outcome of the analysis is as follows:

* To develop metrics for comparing the PBL height from the CL51 ceilometer against the predicted PBL height data from the CT model output; and
* To understand the factors that drive the change in PBL height.

## Related work (Yujing)

## Data analysis and preliminary modelling

In this section we outline the results of our preliminary analysis of the PBL height data provided.

### Data provided:

DPIE has provided the following PBL height data for the period between 12/2/2021 to 19/2/2021 at Merriwa and Lidcombe:

* Data from the Viasala CL51 ceilometer; and
* Data from the CT model.

Both dataset are provided in the form of CSV files.

The following table sets out the fields in each datafile provided as well as our understanding of what they represent.

In respect of the CL51 dataset:

|  |  |
| --- | --- |
| **Data fields** | **Description** |
| # Time | Datetime stamp of the observation in AEST |
| Layer\_QualityIndex | Quality of information read |
| bl\_height | PBL height in metres |
| Mean\_Layer\_Height | Mean of the PBL height in metres |
| n\_BL | Number of PBL observed |
| cloud\_stat | Number of cloud layers observed |
| cloud\_dat | Height of cloud data in metres |

In respect of the CT model dataset:

|  |  |
| --- | --- |
| **Data fields** | **Description** |
| lat | Latitude of observed location |
| lon | Longitude of observed location |
| date | Date of observation |
| time | Time of observation in UTC |
| temperature | Predicted temperature in Celsius |
| mixing\_height | Predicted PBL height in metres |

### Data pre-processing steps.

We have used Python to pre-process the data. The following sets out the detailed steps taken to pre-process both datasets provided.

#### Pre-processing the CL51 dataset.

1. Data is firstly loaded into Python.
2. We notice that missing values in the CSV files are automatically filled with “-999”. This is changed into ‘NaN’ values for ease of future processing.
3. A proper datetime stamp is created based on information already present in the data.
4. Imputation is performed on the “bl\_height” feature to append the missing values. We have chosen to impute the missing values (as opposed to omitting the data with missing values) because we have limited data to work with (i.e. only one week’s worth of observations). We have used the Pandas library’s built in interpolation method (interpolated based on the datetime stamp).

#### Pre-processing the CT model data.

1. Data is again similarly loaded into Python.
2. We note that the “time” field recorded in this dataset is in UTC. We have adjusted this datafield to align the timestamp with the CL51 data which is in AEST. This is done by adding 10 hours to the data in the “time” field.
3. A datetime stamp is then created using the revised “time” field as described above.

#### Aligning the two datasets.

We further note that the granularity of the CL51 dataset datetime stamp is in seconds, whereas the granularity of the CT model output datetime stamp is in hours.

To make the two datasets comparable, we have grouped the CL51 dataset such that it has the same datetime stamp granularity as the CT model outputs (i.e. hours as opposed to minutes). As such in the grouped CL51 dataset, the PBL height at each hour reflects the average PBL height within that hour from the original dataset.

This is then combined with the pre-processed CT model data and used for comparisons.

#### Comparing the CL51 datasets before and after pre-processing

The following graphs show the PBL height as obtained from the CL51 ceilometer in Merriwa and Lidcombe before and after the datasets were imputed.

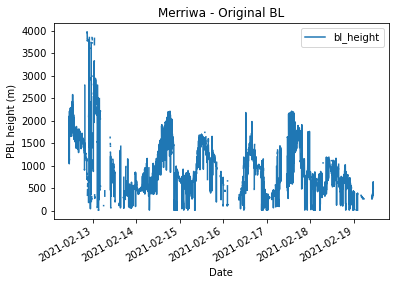


Figure . PBL height in Merriwa - before pre-processing

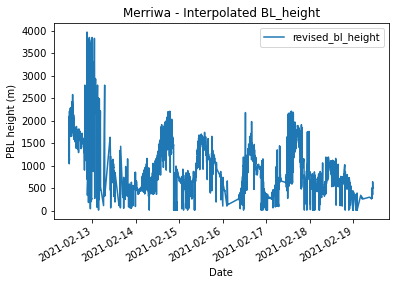


Figure PBL height in Merriwa - after pre-processing

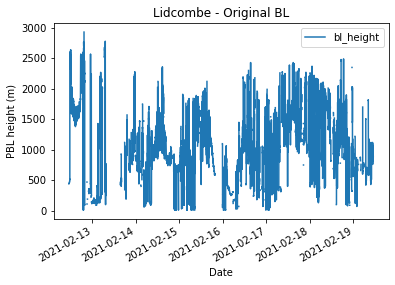


Figure PBL height in Lidcombe - before -pre-processing

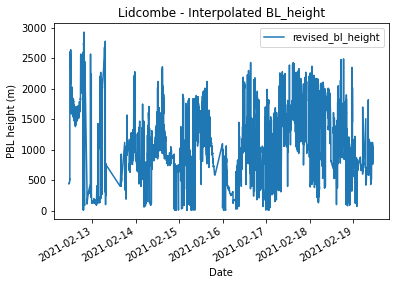


Figure PBL height in Lidcome - after pre-processing

We observe that the pre-processing has successfully filled in the gaps in the missing values in both locations. We further note that based on visual inspection, the method of interpolating the data between datetime stamps is an appropriate in these circumstances given the granularity of the data originally provided.

## Analysis 1 - Comparing the PBL height of both datasets.

In this section we consider a detailed exploratory data analysis of the PBL height as presented in both the CL51 and CT model output datasets.

This section is divided into two parts:

1. Visual comparisons of the CL51 and CTM PBL height data; and

2. Derivation of statistical metrics for comparison of both datasets.

### Visualisation of the CL51 and CTM PBL height data

We firstly denote the CL51 PBL height as the actual PBL height observed in an area. Consequently, the CT modelled PBL height is then denoted as the predicted PBL height.

The following graphs show a comparison between the CL51 (Actual) PBL height and the CT model (Predicted) PBL height at both locations.

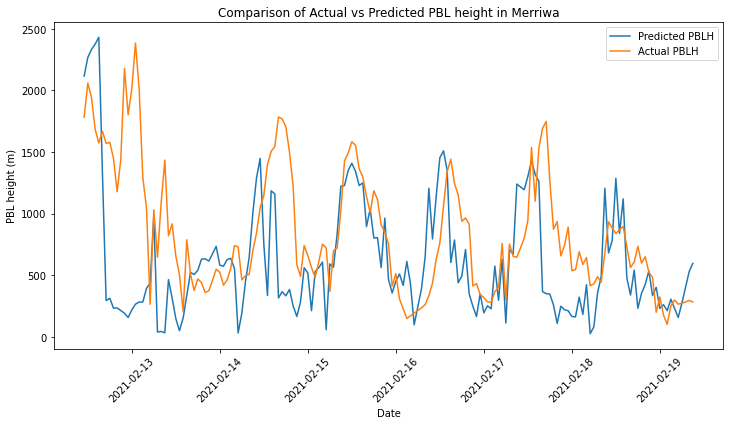


Figure Actual vs Predicted PBL height in Merriwa

Overall, in Merriwa we observe that there is quite a tight overlap between predicted and actual PBL height over the one-week period.

We note however observations in the timeframe from 12/2/2021 to 13/2/2021 deviate significantly. This is due to poor quality readings from CL51 ceilometer which may be as a result of rain during this period.

From midday 13/2/2021 to around midday 17/2/2021 the CT model outputs was in agreement with the actual observed PBL height. There is then another discrepancy in readings from midday 17/2/2021 to 18/2/2021. Unsure as to why this is the case

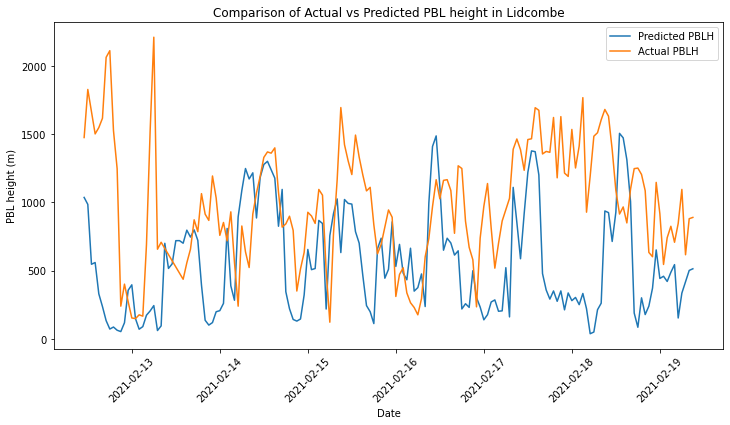


Figure Actual vs Predicted PBL height in Lidcombe

In Lidcombe however, the CT model outputs and CL51 ceilometer readings exhibit significantly more deviations.

We observe that while there is some agreement as to the timing of the peaks and troughs of the PBL height as measured by the CT model and the CL51 ceilometer, their magnitudes are vastly different. For instance, the actual versus predicted PBL heights at 13/2/2021 and again from 17/2/2021 to 18/2/2021 are in total opposition to one another.

It is unclear from our analysis as to what may be the cause of this.

The following graphs compares the PBL height at each hour of the day grouped by the date of the observations.

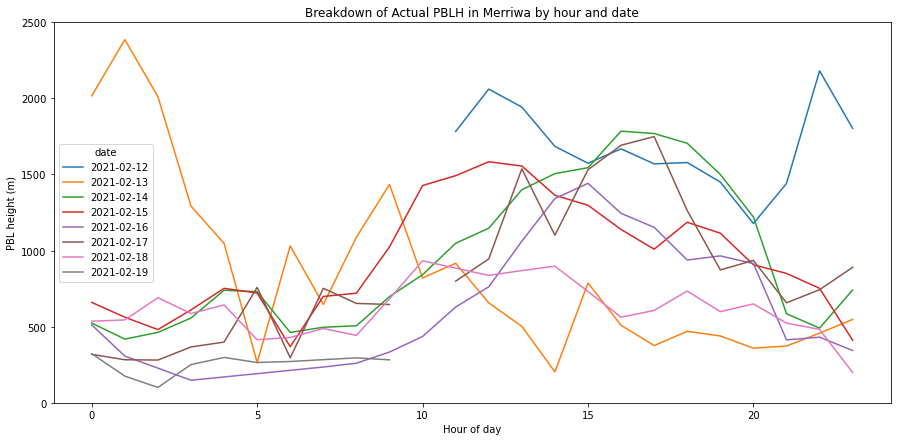


Figure 7 Breakdown of actual PBL height in Merriwa by hour

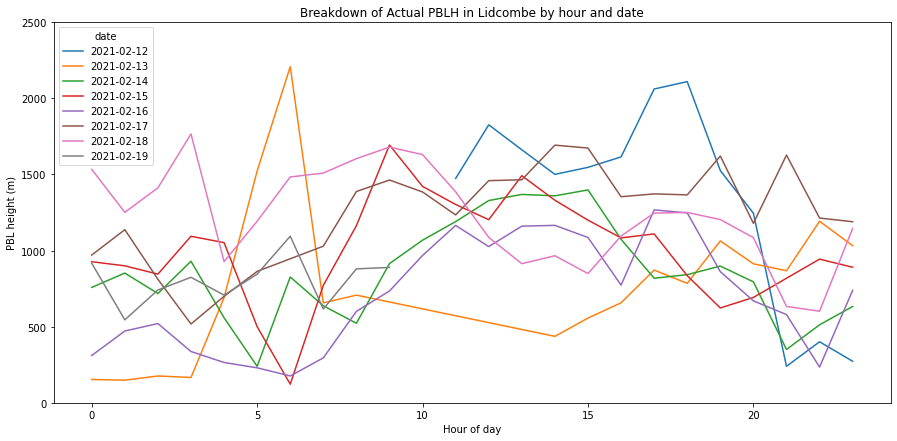


Figure 8 Breakdown of actual PBL height in Lidcombe by hour

The above graphs further highlight the cyclic nature of the PBL height. We observe that the PBL height tends to reach peak height during the middle of the day (at around 3pm) before dropping off again and troughing at around 5 am.

### Statistical metrics

In addition to visual inspections, we have also calculated some select statistical metrics aimed at comparing the fit between the Actual PBL height and the Predicted PBL height.

As a matter of convenience, we will adopt the following notations:

|  |  |
| --- | --- |
| Notation | Description |
|  | Actual PBL height observed at a particular location at time *i.*  This corresponds to dataset obtained from the CL51 ceilometer. |
|  | Predicted PBL height observed at a particular location at time *i.*  This corresponds to dataset obtained from the CT model output. |
|  | The average actual PBL height observed at a particular location based on the CL51 ceilometer data. |
|  | The average predicted PBL height observed at a particular location based on the CT model output. |

In our comparison we have selected 5 measures which are as follows:

|  |  |
| --- | --- |
| Statistical measures | Description of measure |
| WIllmott’s Index of Agreement | A numeric measure that measures how well a model simulates observed data. Simulations that fit the observed data well will exhibit higher index of agreement values.  The formula is given as follows: |
| Root mean squared error (RMSE) | Measures the root of the average squared difference between the actual and predicted PBL height at a given area. This is a proxy measure for the distance between the actual and predicted PBL heights. This metric naturally gives precedence to larger deviations between the actual and predicted PBL heights.  The formula is given as follows: |
| Mean absolute error | Measures the average absolute difference between the actual and predicted PBL height at a given area. This is another proxy measure for the distance between the actual and predicted PBL heights. Contrary to RMSE, this measure is not skewed by larger deviations between the actual and predicted PBL heights.  The formula is given as follows: |
| Mean bias error | Measures the average difference between the actual and predicted PBL height at a given area.  The formula is given as follows: |
| Dynamic Time Warping (DTW) distance | DTW presents an alternative metric for measuring the distance between the actual and predicted PBL heights.  In contrast to the RMSE, DTW attempts to find the best mapping between the actual and predicted PBL heights such that the distance between the mappings is minimised. Such techniques are called time-warping. |

The following sets out the statistical metrics for both locations:

|  |  |  |
| --- | --- | --- |
|  | **Location** | |
|  | **Merriwa** | **Lidcombe** |
| **Willmot's Index of agreement** | 0.64061 | 0.47060 |
| **RMSE** | 597 | 684 |
| **Mean absolute error** | 416 | 540 |
| **Mean bias error** | -233 | -430 |
| **DTW distance** | 91,526 | 107,124 |

### Discussion

The PBL height exhibits a peak and trough type behaviour where peaks are typically around midday and troughs are late at night/early mornings. Overall, there is a fairly reasonable correlation between the predicted and actual PBL over the one-week period. The CT model was able to capture the behaviour of the PBL height in both locations.

There are however significant deviations in the quality of the predicted PBL height from the CT model as compared to the actual PBL heights from the CL51 ceilometer in both locations. In particular, we note that in Merriwa the CT model outputs tend to match the actual PBL height much more closely than in Lidcombe.

While we offer no concrete reasons as to why this is the case, we reason that this may be driven in large part by the urban and geographic environment of the two sets of observations. Lidcombe, being an inner suburb within the Sydney Metropolitan area, is subject to constant pollution and disturbances from urban infrastructure that may potentially impact the reading of the ceilometer. Merriwa by contrast, is located in the Hunter Valley where there are little disturbances from urban infrastructures. CITE??

## Analysis 2 - Analysing factors that impact the PBLH

**3. Statistical metrics** (shaohua)

Based on the statistics metric, we have a few observations:

3.1 Combine BLH of Merriwa and Lidcombe data

* External factors of Carbon monoxide, Humidity, Nephelometer, Ozone, Temperature, Wind Direction Sigma Theta and Solar radiation impact both predicted and actual BLH.
* Wind Direction Sigma Theta shows negative correlation with predicted data but positive correlation with actual data.
* Nitrogen Dioxide, Rainfall and Wind Speed (10m) are not correlated with actual BLH, however theses external factors are correlated with predicted BLH

3.2  BLH in Merriwa

* External factors of Carbon monoxide, Humidity, Ozone, Temperature impact both predicted and actual BLH.
* Nitric Oxide, Nitrogen Dioxide, PM2.5 and Wind Speed(10m) are correlated with predicted BLH but not correlated with actual BLH
* Wind Direction(10m) are correlated with actual BLH but not correlated with predicted BLH

3.2  BLH in Lidcombe

* External factors of Carbon monoxide, Humidity, Nephelometer, Ozone, Solar radiation, Temperature and Wind Direction Sigma Theta impact both predicted and actual BLH.
* Nitric Oxide, Nitrogen Dioxide, Rainfall and Wind Speed(10m) are correlated with predicted BLH but not correlated with actual BLH

Maybe do in Sem2 instead of Sem1

1. Based on API data such as PM2.5, Average temperature to see the BLH difference fromCL51 and predicted CTM. For example, The way of PM2.5 effects the difference between two devices.

2. Check the sensitivity factors that affect difference

To be more specific:

Ceilometer -> BLH1

CTM -> BLH2

external factors from gov API -> X

Y -> BLH2 – BLH1 (difference between 2 BLH)

Build a model with X -> Y

Find factors in X which significantly affect the Y. (Ideally, the Y should be 0 all the time).

Models available: Linear (may underfitting), logistic (including interactions), NN(may overfitting). Then based on  what had been found, we could construct some interesting interpolation. For example, when NH3 and CO are high, the BLH1 is dramatically higher than BLH2. But there the air quality did not change much, which indicates the ceilometer does not work well in an environment with NH3 and CO.

[REVISIONS BASED ON DISCUSSION WITH DPIE?]

## Proposal for Semester 2

Semester 2 will focus on the continuation of Analysis 2. In particular:

1. Further in-depth analysis of the factors that drive PBL height;
2. Derivation of model(s) that may be used to predict the PBL height based on such factors; and
3. Analysis of our derived model against: (1) the original PBL height data provided and (2) CT model output provided.

We also aim to provide a report summarizing our findings for DPIE.

## Proposed timeline for Semester 2

|  |  |  |
| --- | --- | --- |
| Task | Timeline in Semester 2 | Doer |
| Further in-depth analysis of factors that drive the PBL height | Weeks 1 – 2 |  |
| Derivation of models to predict the PBL height based on these factors | Weeks 1 – 4 |  |
| Analysis of models developed against CL51 data | Weeks 3 – 5 |  |
| Analysis of models developed against the CT model output | Weeks 3 – 5 |  |
| Consultations with DPIE in relation to our findings | Week 6 |  |
| Further revisions to the modelling based on feedback from DPIE | Weeks 6 – 9 |  |
| Final write-up of report and findings | Weeks 9 – 12 |  |
| Further consultations with DPIE as required | Weeks 10 – 11 |  |
| Final presentation of reports and findings to DPIE and cohort | Weeks 11 – 12 |  |