Development of a statistical forecasting model for $PM_{2.5}$ in Macau based on clustering of backward trajectories

Tong Xie[1,*], *Kai Meng* Mok[2], *Ka Veng* Yuen[3], and *Ka In* Hoi[4]

[1]Postgraduate student, Department of Civil and Environmental Engineering, University of Macau, Macau SAR, China
[2]Professor, Department of Civil and Environmental Engineering, University of Macau, Macau SAR, China
[3]Distinguished Professor, Department of Civil and Environmental Engineering, University of Macau, Macau SAR, China
[4]Postdoctoral fellow, Department of Civil and Environmental Engineering, University of Macau, Macau SAR, China

Abstract. A daily $PM_{2.5}$ forecasting model based on multiple linear regression (MLR) and backward trajectory clustering of HYSPLIT was designed for its application to small cities where $PM_{2.5}$ level is easily affected by regional transport. The objective of this study is to investigate the regions that affect the fine particulate concentration of Macau and to develop an effective forecasting system to enhance the capture of $PM_{2.5}$ episodes. By clustering the HYSPLIT 24-hr backward trajectories originated at Macau from 2015 to 2017, five potential transportation paths of $PM_{2.5}$ were found. A cluster based statistical model was developed and trained with air quality and meteorological data of 2015 and 2016. Then, the trained model was evaluated with data of 2017. Comparing to an ordinary model without backward trajectory clustering, the cluster based $PM_{2.5}$ forecasting model yielded similar general forecast performance in 2017. However, the critical success index of the cluster based model was 11% higher than that of the ordinary model. This means the cluster based model has better model performance in $PM_{2.5}$ concentration prediction and it is more important for the health of the public.

1 Introduction

Due to serious adverse effects to respiratory and cardiovascular systems, $PM_{2.5}$ has become a major concern of the public and the Chinese government. In the past decade, China has been experiencing a severe $PM_{2.5}$ pollution due to rapid urban development. Many cities do not comply with the annual $PM_{2.5}$ standard. These cities of nonattainment are mostly distributed in three major city clusters including Beijing – Tianjin – Hebei (BTH), Yangtze River Delta (YRD) and Pearl River Delta (PRD) [1]. Macau a gaming and tourism city located in the southern part of PRD, is having similar situation. In order to protect its citizens, it is important to develop a daily forecasting model of $PM_{2.5}$ for Macau.

Till the third quarter of 2018, this city accommodates a population of 663,400 in an area of 30.8 km^2. Due to the small geographical area, statistical model can be a more attractive alternative to large scale deterministic models when developing operational air quality model [2]. The air quality of Macau is not only governed by the local emissions and the dispersion conditions. Its $PM_{2.5}$ level is also susceptible to the transboundary pollution of neighbouring regions. Therefore, reflecting the regional influence of fine particulates within the forecasting model is necessary. Understanding about the distribution of Macau's upwind cities, where the fine particulates are generated and transported by the atmospheric flow is important.

In view of this, the objective of this study is to develop a cluster based $PM_{2.5}$ statistical forecasting model for performing 1 day ahead forecast of daily averaged $PM_{2.5}$ concentration in Macau. Comparing to deterministic models, the large quantity of historical measured data under a variety of conditions in statistical approaches often have higher accuracy [3]. To achieve this, the historical backward trajectories originated at Macau between 2015 and 2017 are analysed and clustered into several groups by K-means clustering in HYSPLIT. Then, a specific statistical model is developed based on current day local meteorological factors, local time-lagged air pollutant concentrations, hourly $PM_{2.5}$ concentration at mid-night from several upwind cities identified in the classified trajectory clusters. The cluster based $PM_{2.5}$ forecasting model is then trained by using the data of 2015 and 2016, and the dataset of 2017 is used for evaluation of the model. Details of the methodology are described in the following section.

2 Methodology

2.1 Air quality and meteorological data

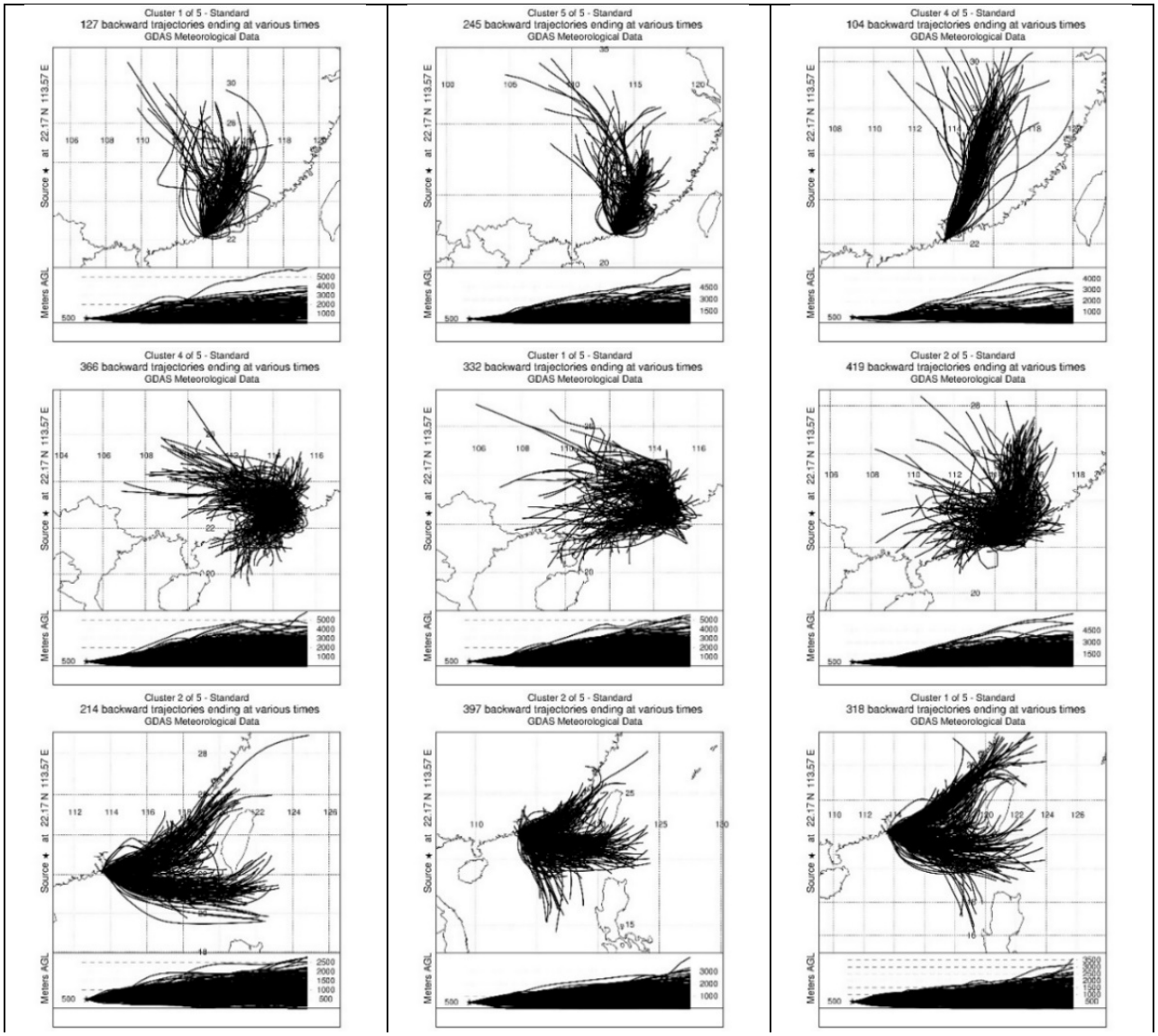
Local air quality data and meteorological data utilized for model development were provided by the Macau Meteorological and Geophysical Bureau. The air quality

* Corresponding author: mb65481@um.edu.mo

© The Authors, published by EDP Sciences. This is an open access article distributed under the terms of the Creative Commons Attribution License 4.0 (http://creativecommons.org/licenses/by/4.0/).

dataset consists of hourly concentrations for $PM_{2.5}$, PM_{10}, SO_2 and NO_2 measured at 5 monitoring stations (http://www.smg.gov.mo/smg/airQuality/c_air_stations.htm) from 2015 to 2017. The meteorological datasets consist of measured data (2015-2016) and forecasted data (2017) at the headquarter of the Macau Meteorological and Geophysical Bureau (22°09′36"N 113°33′54"E). The meteorological dataset contains 5 parameters including the wind speed, wind direction, mean sea level pressure, temperature and relative humidity.

Due to the small area of Macau, its $PM_{2.5}$ level is not only influenced by the local pollution sources. It can be easily affected by the regional transport from its neighbouring upwind cities. Therefore, it is reasonable to adopt the $PM_{2.5}$ concentrations of upwind cities as the model inputs. The data of upwind $PM_{2.5}$ concentrations were obtained from the real-time air quality release platform of the China National Environmental Monitoring Centre (CNEMC). The website reports hourly concentrations of $PM_{2.5}$ measured at 367 cities, with a time lag of 1 hour. However, one difficulty that immediately arises is the choice of the upwind cities, which is addressed in the following subsection.



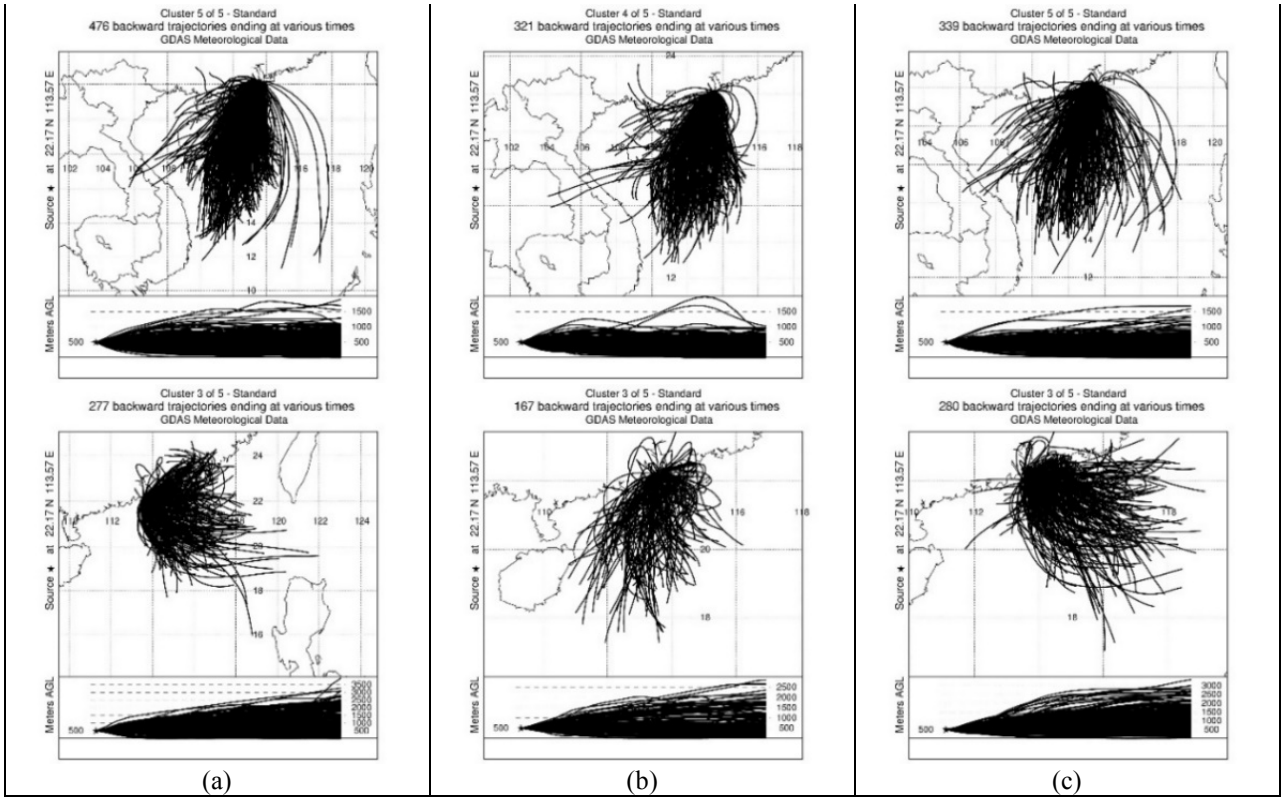


Fig. 1. The HYSPLIT trajectory model computed 24-h backward trajectories originated at 500 meters over Macau at midnight for the period of 2015 (column a), 2016 (column b) and 2017 (column c). (source: ready.arl.noaa.gov/HYSPLIT.php).

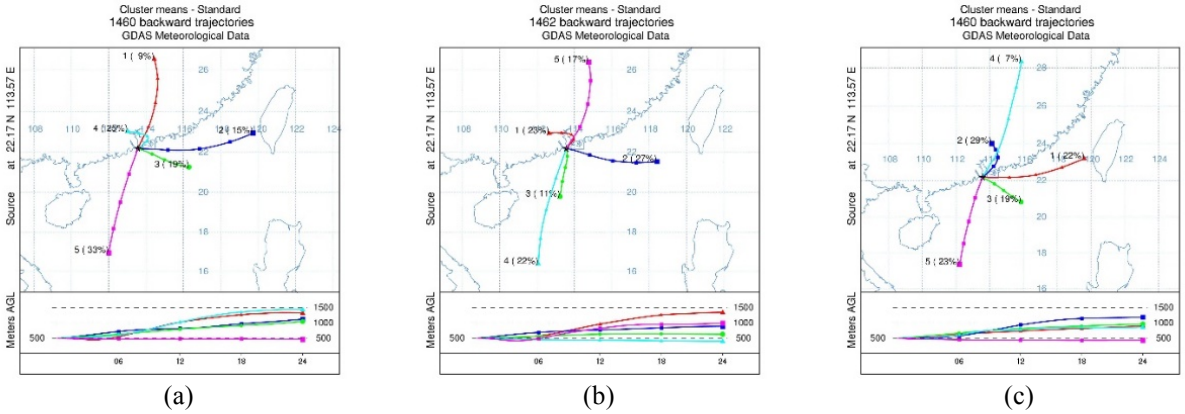


Fig. 2. The mean trajectories of the trajectory clusters computed by the HYSPLIT model for 2015 (a), 2016 (b) and 2017 (c). The squares showed locations of air parcels at 6-h intervals. (source: ready.arl.noaa.gov/HYSPLIT.php).

2.2 Clustering analysis of 24-h backward trajectories

Previous works showed that upwind pollutant concentration of the previous day could influence the performance of air quality forecasting models [4-6], meaning that the choices of upwind cities are important for model development. In this study, the choices for Macau were investigated systematically by using the 24-h backward trajectories with the same starting height of 500 meter for the period between 2015 and 2017. These trajectories were calculated by using the HYSPLIT trajectory model of NOAA. Each trajectory was set to start from Macau at midnight, and the 24-h trajectory end point represented the estimated location of upwind polluted air at midnight on the previous day. All the 24-h backward trajectories for each year between 2015 and 2017 were then classified by K-means into 5 clusters as shown in Fig. 1.

Each column in Fig. 1 shows the clusters of trajectories for a particular year. The general pattern was revealed more clearly by using the mean trajectory of each cluster as in Fig. 2. It shows that the clusters of different years were similar to each other, meaning that this pattern is repeatable over time. The pattern indicates that the regional influence of $PM_{2.5}$ in Macau could be due to (i) the transport of inland fine particulates from the northern part of PRD (north and north-west cluster occurring mostly during winter), (ii) the transport of fine particulates from coastal cities between the Bashi Channel and the Taiwan Strait (northeast cluster

occurring mostly during winter and spring), and (iii) the transport of marine aerosols from the South China Sea (southern clusters occurring mostly during summer).

2.3 Cluster based forecasting model of $PM_{2.5}$

The multi-linear regression (MLR) was shown to be an effective tool for air quality forecasting [7]. Here, the MLR was adopted to develop the cluster based forecasting model of $PM_{2.5}$. The general form of the forecasting model is:

$$z_k = \alpha_0 + \boldsymbol{\alpha}_k \boldsymbol{\tau}_{k-1}^T + \varepsilon_k \quad (1)$$

$$\boldsymbol{\alpha}_k = [\alpha_1, \dots, \alpha_m] \quad (2)$$

$$\boldsymbol{\tau}_{k-1} = [x_{1,k-1}, \dots, x_{m,k-1}] \quad (3)$$

where z_k represents the measured $PM_{2.5}$ concentration of the k^{th} day, which is equal to the spatial average of available daily averaged $PM_{2.5}$ concentrations measured at 5 monitoring stations. It is predicted by the linear combination of the input variables with an intercept term while $\boldsymbol{\alpha}_k$ represents the vector of coefficients corresponding to the linear combination. $\boldsymbol{\tau}_{k-1}$ is the vector of input variables available on the $(k\text{-}1)^{th}$ day. ε_k is the modelling error associated with the forecasting model. Table 1 shows the compositions of input variables adopted to develop the forecasting model:

Table 1. Compositions of input vector $\boldsymbol{\tau}_{k-1}$ of cluster based forecasting model compared to an ordinary model without upwind city input

	$\boldsymbol{\tau}_{k-1}$	Ordinary model		Cluster based model	
		k	*k-1*	*k*	*k-1*
Local input	RH [a]	√			
	PSEA [a]				
	TEMP [a]				
	WSPD [a]				
	U [a]				
	$[NO_2]$ [a]		√		√
	$[SO_2]$ [a]				
	$[PM_{2.5}]$ [a]				
	$[PM_{10}]$ [a]				
	Macau [b]				
Upwind city input	Hongkong [b]				
	Guangzhou [b]				
	Shenzhen [b]				
	Foshan [b]				
	Dongguan [b]				
	Xiamen [b]				

[a] Daily averaged data in Macau.
[b] Measured hourly concentration of $PM_{2.5}$ at the last hour of day *k*-1.
√ means the variable was adopted in the model

In Table 1, the symbols RH, PSEA, TEMP, WSPD and U represent the daily averages of the relative humidity, mean sea level pressure, temperature, wind speed and the north-south component of the wind direction, respectively. These variables were used to reflect the atmospheric stability, the available wind for dilution and the nature (inland or sea) of the trajectory. Besides the meteorological variables, the local daily averaged pollutant concentrations of the $(k\text{-}1)^{th}$ day and the hourly $PM_{2.5}$ concentration at 11:00PM of the $(k\text{-}1)^{th}$ day in Macau were also incorporated to reflect the initial air quality condition of the k^{th} day. Finally, the hourly $PM_{2.5}$ concentration at 11:00PM of the $(k\text{-}1)^{th}$ day for several upwind cities were selected and used to reflect the regional influence of fine particulates.

3 Results and discussion

The model was trained with air-quality and meteorological data of 2015 and 2016. Even though the model involved meteorological inputs of the k^{th} day that were practically not available on the $(k\text{-}1)^{th}$ day, the measured meteorological data of the k^{th} day were still used as the model inputs in order to produce better training results. When the trained models were evaluated against the data of 2017, the meteorological forecasts of the k^{th} day were used instead of the measurements. The model accuracy was evaluated according to the measures shown in Eqns. (4) to (9). In these equations, *P*, *O*, and $\bar{O}$ represent the prediction, the observation, and the annual average of observed $PM_{2.5}$ concentrations, respectively. For the overall accuracy, the coefficient of determination (R^2) describes how well the model variance explains the observation variance. The normalized mean absolute error (*NMAE*) describes the overall magnitude of the forecast errors relative to the mean of the observations. The index of agreement (*IA*) is a dimensionless indicator between zero and one. *IA* = 1 means perfect prediction, while *IA* = 0 means no agreement at all [4]. This index can detect additive and proportional differences between the observations and the predictions. As for the prediction performance during high $PM_{2.5}$ concentrations, the episode detection rate (*EDR*), represents the probability of successful hits under the observed exceedances ($PM_{2.5} > 35\ \mu g\, m^{-3}$), and the false alarm rate (*FAR*) describes the proportion of false alarms under the forecasted exceedances. In this study, the hit or alarm threshold was $35\mu g\, m^{-3}$. This value corresponds to the daily limit of the primary $PM_{2.5}$ standard of US-NAAQS, which is equivalent to the grade I $PM_{2.5}$ standard of China (GB3095-2012). The critical success index (*CSI*) is an integrated measure of *EDR* and *FAR* and it can be treated as a discounted *EDR* and the penalty depends on the *FAR* of the model.

$$R^2 = \frac{\sum_{i=1}^{N}(P_i - \bar{O})^2}{\sum_{i=1}^{N}(O_i - \bar{O})^2} \quad (4)$$

$$NMAE = \frac{\sum_{i=1}^{N}|P_i - O_i|}{\sum_{i=1}^{N} O_i} \quad (5)$$

$$IA = 1 - \frac{\sum_{i=1}^{N}(P_i - O_i)^2}{\sum_{i=1}^{N}(|P_i - \bar{O}| + |O_i - \bar{O}|)^2} \quad (6)$$

$$EDR = \frac{no.of\ hits}{no.of\ observed\ exceedances} \quad (7)$$

$$FAR = \frac{no.of\ false\ alarm}{no.of\ forecasted\ exceedances} \quad (8)$$

$$CSI = \frac{EDR}{1 + FAR\left(\frac{no.of\ forecasted\ exceedances}{no.of\ observed\ exceedances}\right)} \quad (9)$$

Table 2 shows the performance of cluster based $PM_{2.5}$ forecasting model and the ordinary $PM_{2.5}$ forecasting model during training period and evaluation period. The ordinary $PM_{2.5}$ forecasting model is a subset of the cluster based model without including the hourly $PM_{2.5}$ concentrations of the upwind cities. During the training phase, it is noted that the cluster based model outperforms the ordinary model due to the extra model complexity to fit the training data. During the evaluation phase, both the ordinary model and the cluster based model achieve similar general performance (R^2, *NMAE*, and *IA*). However, the critical success index (*CSI*) of the cluster based model is significantly higher than that of the ordinary model, meaning that the inclusion of upwind cities is important.

Table 2. Performance of cluster based PM2.5 forecasting model and ordinary model during training period and evaluation period.

Index	Training period (2015-2016)		Evaluation period (2017)	
	Ordinary model	Cluster based model	Ordinary model	Cluster based model
R^2	0.79	0.82	0.75	0.76
NMAE	22.1%	21.5%	25.0%	25.0%
IA	0.94	0.95	0.93	0.94
EDR	78% (173 days)	80% (177 days)	66% (58 days)	85% (75 days)
FAR	19% (41 days)	20% (44 days)	26% (20 days)	34%(39 days)
CSI	65.5%	66.7%	52.4%	63.4%

Fig. 3 shows the measured daily averaged $PM_{2.5}$ concentrations (green line) versus the corresponding predictions (orange line) by the cluster based model and the predictions by the ordinary model (blue line) in 2017. It is noted that predictions by either the cluster based model or the ordinary model are generally in good agreement with the measurements. However, the cluster based model is more capable to capture the peaks. As the most primary objective of air quality forecast model is to correctly predict the episodes, the development of the cluster based model in this study is successful.

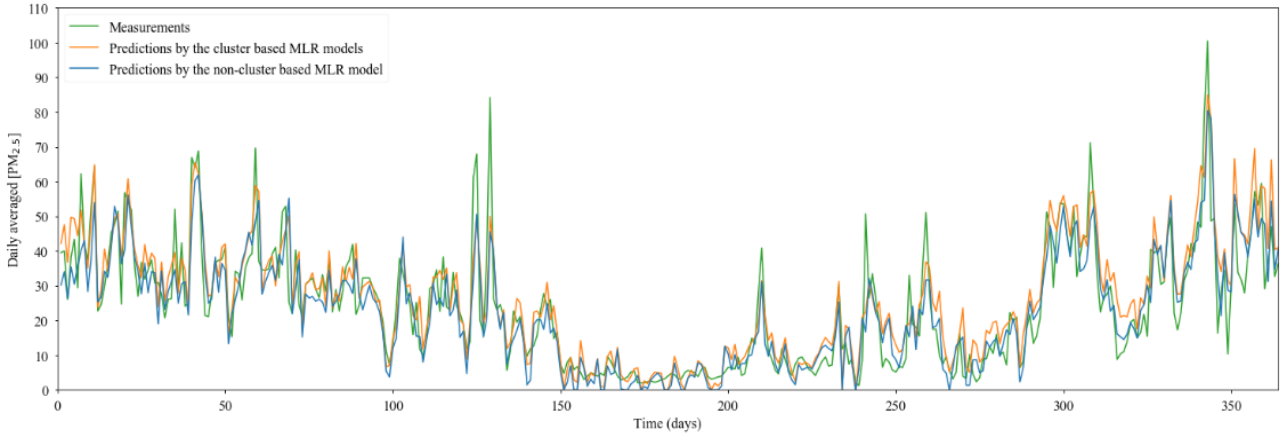


Fig. 3. Plot of measured daily averaged $PM_{2.5}$ concentrations (green line) versus predictions by the cluster based model (orange line) or the predictions by the ordinary model (blue line) in 2017.

4 Conclusions

An empirical $PM_{2.5}$ forecasting model was developed for Macau based on multiple linear regression and backward trajectories clustering of HYSPLIT. The backward trajectories clustering revealed 3 potential regional sources of $PM_{2.5}$ for Macau. A cluster based statistical forecasting model was developed and the regional influence of $PM_{2.5}$ was reflected in the model by using the hourly $PM_{2.5}$ concentrations of the upwind cities before the midnight. Two years of data (2015-2016)

were used for model training and one year of data (2017) was used for model evaluation. It was concluded that the cluster based forecasting model overperformed the ordinary $PM_{2.5}$ forecasting model during the episode days, which are of great concern to the general public.

The authors wish to thank the Macau Meteorological and Geophysical Bureau for providing data and Mr. Wenchao Feng for the technical assistance during the process. The authors gratefully acknowledge the NOAA Air Resources Laboratory (ARL) for the provision of the HYSPLIT transport and dispersion model and/or READY website used in this publication (http://www.arl.noaa.gov/ready.php).

References

1. B. Lv, G. Cobourn, Y. Bai, *Atmos. Environ.* **147,** 209-223 (2016)
2. A. Donnelly, B. Misstear, B. Broderick, *Atmos. Environ.* **103,** 53-65 (2015)
3. Y. Zhang, M. Bocquet, V. Mallet, C. Seigneur, A. Baklanov, *Atmos. Environ.* **60,** 632-655 (2012)
4. G.T. Wolff, P.J. Lioy, *J. Air Pollut. Control Assoc.* **28**, 1034-1038 (1978)
5. G. Cobourn, C. Hubbard, *Atmos. Environ.* **33**, 4663-4674 (1999)
6. G. Cobourn, *Atmos. Environ.* **44**, 3015-3023 (2010)
7. A. Vlachogianni, P. Kassomenos, A. Karppinen, S. Karakitsios, J. Kukkonen, *Sci. Total Environ.* **409**, 1559-1571 (2011)