The dataset looks interesting, and I think it will be fun to model trends in it.

The way it looks now, it sounds like there will be quite a bit of exploratory data analysis in first instance: you'll need to figure out how to represent the weather component efficiently.

Also I expect there will be a lot of seasonality in this sort of data (times within the day, also different times in the year might be systematically different because of seasonality of polluting behavior)

I'd suggest you do not commit yourself to a simple autoregressive process from the get go but rather first figure out what kind of interesting structure is there in the data and only then decide on the model.

The issue of missing data will become an important criterion for deciding on the model as well - as far as i know there are sensible ways to deal with missing data in latent variable models, but it's less clear to me how to go about it with simple AR(p).

Also, worth thinking a bit more precisely about success metrics - I assume you'll want to measure prediction accuracy different horizons in the future… What constitutes success? Maybe also: have people tried to do this sort of thing before and how well did it work?

It sounds like just curating the data and including the additional labeled weather information is a nontrivial task, and the resulting dataset as such is a useful deliverable.

Once you have the data happy to sit together and look at it for a bit and discuss modelling options.

Hope this helps
Let me know if you have questions.

Best wishes
CS